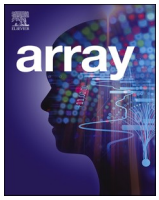
Array 24 (2024) 100370

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array



SAMU-Net: A dual-stage polyp segmentation network with a custom attention-based U-Net and segment anything model for enhanced mask prediction

Radiful Islam [1], Rashik Shahriar Akash [1], Md Awlad Hossen Rony, Md Zahid Hasan [*]

Health Informatics Research Laboratory (HIRL), Department of Computer Science and Engineering, Daffodil International University, Dhaka, 1216, Bangladesh

A B S T R A C T

Early detection of colorectal cancer through the proper segmentation of polyps in the colonoscopy images is crucial. Polyps' complex morphology and varied appearances are the greatest obstacles for the segmentation approaches. The paper introduces SAMU-Net, a novel deep learning-based dual-stage architecture consisting of a custom attention-based U-Net and modified Segment Anything Model (SAM) for better polyp segmentation. In our model, we used the custom U-Net architecture with an attention mechanism to obtain polyp segmentation masks as the first stage. This mask is then used to generate a bounding box input for the second stage that contains the modified Segment Anything Model. The modified SAM relies on the use of High-Quality token-based architecture along with global and local properties to segment polyps accurately, even in cases where the shapes and sizes of polyps are diverse and the polyps have different appearances. The efficiency of SAMU-Net generated from four different datasets of colonoscopy images was examined. Our process produced a dice coefficient score of 0.94, which is very impressive and has a considerable improvement over the existing state-of-the-art polyp segmentation methods. Moreover, the qualitative results also visualize that the SAMU-Net is capable of accurately segmenting polyps of wide ranges, thus, it is a relevant tool for computer-aided detection as well as the diagnosis of colorectal cancer.

1. Introduction

Colorectal cancer (CRC) is a major health concern as it ranked third among all the cancer cases globally and is the second most frequent cause of cancer-related mortality [1]. It was estimated that in 2020, about 1.93 million new cases of CRC were diagnosed worldwide, with 935,000 deaths attributed to the disease [2]. It is also predicted that, by the year 2030, the number of new CRC cases will rise by 60 % to 2.2 million and the number of deaths will go up to 1.1 million per year [3]. The rising number of these cases is a lucid sign pointing to the usefulness of colonoscopy, a standard screening modality for CRC, capable of detecting and removing precancerous polyp [4]. Such growths that might be found in the patient's colon or rectum that can after some time change into cancer. Colonoscopy would be able to reduce the probability of CRC development by detecting and removing polyps [5]. However, colonoscopy has a major shortcoming in the accuracy of polyp's detection and segmentation, thus lowering the chance for the patients with colorectal cancer to be properly diagnosed, not to mention their treatment [6]. Timely detection and treatment are essential to lessen the number of casualties. Despite its importance, the segmentation of polyps during colonoscopy faces several challenges. An example of this is manual polyp segmentation. This traditional method is prone to operator dependency errors, especially for thin and flat polyps [7]. The accuracy of polyp segmentation can also be impaired by different variables, such as the quality of the colonoscopy image, the presence of artifacts, and polyps with various shapes [8]. Besides, there has been a rise in the number of CRC screening, especially among the senior citizens, which has added pressure to the healthcare systems and endoscopists, emphasizing the need for efficient and accurate automated segmentation tools [9].

To counteract these challenges and improve the accuracy in polyp segmentation, there has been growing interest in developing and applying artificial intelligence (AI) technologies in colonoscopy [10]. Computer-aided segmentation systems that are based mostly on algorithms of deep learning have already represented very promising results in increasing accuracy in polyp segmentation. These AI-aided systems could provide accurate polyp boundary delineation and thus help the endoscopist in the complete removal of the polyp and improve the quality of histopathological examination [11–16]. Deep learning-based developments over the past few years have drastically enhanced the potential for medical image analysis. Convolutional Neural Networks have established excellent performance on many medical imaging tasks,

* Corresponding author.
E-mail addresses: radiful15-3837@diu.edu.bd (R. Islam), rashik15-3825@diu.edu.bd (R.S. Akash), awlad15-12208@diu.edu.bd (M.A. Hossen Rony), zahid.cse@diu.edu.bd (M.Z. Hasan).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.array.2024.100370
Received 3 August 2024; Received in revised form 16 October 2024; Accepted 14 November 2024
Available online 16 November 2024

particularly segmentation [17]. Various CNN-based architectures have been brought into the field of polyp segmentation, such as U-Net [18], SegNet, and DeepLab [19]. Among them, U-Net has been particularly noticed for its very good performance in medical image segmentation tasks. The contacting path and the expanding path are the two paths that form the U-Net architecture, linked by skip connections that enable precise localization while context information is maintained. Therefore, the methodology is very appropriate for polyp segmentation, since the accurate delineation of the polyp boundaries is paramount for proper removal and histopathological analysis. While U-Net and variants are able to bring about success, inherently there is still space for further improvements related to polyp segmentation performance [20]. Attention mechanisms have been one of the most powerful techniques that have begun to surface over the past few years for enhancing model performance in deep learning applied in various domains, in particular, medical image analysis. Attention mechanisms allow models to focus on the most relevant parts of the input, possibly improving accuracy and interpretability. Attention mechanisms probably improve the accuracy in polyp segmentation by having the model pay attention to subtle features that outline the boundaries between the polyps and surrounding tissue for segmentation and detection [21]. Another significant development in computer vision is the introduction of the Segment Anything Model (SAM) by Meta AI [22]. SAM is a segmentation system able to generate high-quality object masks directly from points or boxes. Being flexible and having zero-shot properties make it an interesting candidate for the application of medical image segmentation tasks—especially polyp segmentation. Integrating such advanced techniques to develop more robust and accurate systems in this line of polyp segmentation may become a promising avenue.

This paper introduces SAMU-Net, a novel dual-stage polyp segmentation network that integrates a custom, attention-based U-Net together with the Segment Anything Model to enhance mask prediction. We utilize the benefits of both architectures in our methodology to predict more accurate and robust results for the segmentation of polyps. The first stage of SAMU-Net is a modified U-Net architecture that includes custom attention mechanisms. It is a U-Net based on attention-enhanced methods for the prediction of ROI (Region of Interest) and segmentation of polyps. Attention mechanisms let the model concentrate only on small parts in the input that are relevant for its specific task, assuming that this potentially could enhance its capability to detect accurate boundaries of the polyp and be more robust to challenging cases of flat or small polyps. The second stage of SAMU-Net uses the Segment Anything Model to further enhance the quality of the initial segmentation masks generated by the attention-based U-Net. Equipped with powerful segmentation abilities available in SAM, SAMU-Net can thus avoid some traditional CNN-based segmentation method-related issues, such as challenging cases dealing with different polyp shapes and sizes. Our two-stage approach aims to weld the strengths from both models: the capability of the U-Net to learn task-specific features from training data and the flexibility and accuracy of SAM. Only by such a combination will the segmentation of polyps be more accurate and robust for all types of polyps and different image qualities related to colonoscopy.

We conduct detailed experiments on four publicly available colonoscopy image datasets for performance evaluation and compare our approach with the state-of-the-art methods on polyp segmentation tasks. All the experiments are evaluated according to the dice coefficient, IoU, Weighted F-measure, S-measure, and E-measure. Several potential implications for clinical practice could be realized from the development of SAMU-Net. As such, with better accuracy and more robustness in polyp segmentation, it is possible for SAMU-Net to further improve the efficiency of current AI-aided colonoscopy methods. This is expected to achieve more accurate boundary delineation of the polyp and consequently more complete removals. It will then improve the preventive effect against CRC and reduce incarnation and mortality rates of colorectal cancer. Furthermore, the incorporation of SAM into our model is highly likely to yield a much more interactive and flexible polyp-segmentation system. Our Contributions are as follows.

- We propose a novel dual-stage polyp segmentation network SAMU-Net, integrating a custom attention-based U-Net and the Segment Anything Model (SAM).
- We enhanced polyp segmentation accuracy by leveraging an attention-based U-Net to segment and localize polyp regions, providing initial masks precisely.
- We propose SAM to refine segmentation masks, incorporating global contextual information to delineate polyp boundaries accurately.
- To remove unwanted data and enhance the colonoscopy image, we applied morphological operations for specular reflection removal and utilized an image-sharpening kernel.
- We achieve superior performance with a Dice similarity coefficient of 0.94 on Kvasir-SEG and CVC-ClinicDB datasets, surpassing state-of-the-art methods.

2. Related works

In order to automate the process of polyp segmentation, researchers have been developing CAD prototypes. Analyzing the polyp's edge was the backbone of most early polyp segmentation approaches. Modern approaches, on the other hand, rely heavily on convolutional neural networks (CNNs) and pretrained networks. Faysal et al. [23] proposed a methodology that involves using a MultiResUNet with Attention Guidance (AG) and Test-Time Augmentation (TTA) for colorectal polyp segmentation. The model is trained on the Kvasir-SEG dataset using various data augmentation techniques and the Adam optimizer. The results are measured using metrics like the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). This approach outperforms others by integrating attention mechanisms and residual connections to extract vital features, enhancing segmentation accuracy and robustness. Poly-SAM by Li et al. [24] is a foundational vision model designed for polyp segmentation in medical imaging, leveraging the Segment Anything Model (SAM) through transfer learning. The methodology involves fine-tuning SAM using multi-center colonoscopy image datasets to enhance its performance on polyp segmentation tasks. Poly-SAM achieves its best performance on the CVC-300 dataset with a dice similarity coefficient (DSC) of 0.924 and a mean intersection-over-union (mIoU) of 0.882, demonstrating its superior capability in polyp segmentation compared to other models. Dong et al. [25] presents solutions for medical image segmentation tasks, including polyp and surgical instrument segmentation, utilizing advanced transformer-based models. The methodology involves a multi-model fusion approach, integrating Polyp-PVT, Sinv2-PVT, and Transfuse-PVT models with pyramid vision transformers (PVT) as the backbone for feature extraction. This method demonstrates superior performance compared to conventional convolutional networks (ConvNets). The results achieved the best scores of 0.91 in instrument segmentation and 0.83 in polyp segmentation. The proposed solution improves accuracy and generalization in segmentation tasks, enhancing clinical decision-making and potentially reducing surgical risks and missed diagnoses of colorectal polyps. Zijin et al. [26] introduce the Duplex Contextual Relation Network (DCRNet) for automatic polyp segmentation, enhancing performance by capturing contextual relations within individual images and across multiple photos. Methodologically, it employs two parallel modules, the Interior Contextual-Relation Module (ICR) and the Exterior Contextual-Relation Module (ECR), along with a Region Cross-Batch Memory (ROM) to store and utilize embedding features from previous training epochs. The result is a significant improvement in segmentation performance, achieving a Dice score of 85.41 on the PICCOLO dataset and 90.14 on the Kvasir-SEG dataset. DCRNet outperforms state-of-the-art methods like PraNet and ACSNet, with scores of 2.9 in MAE, 84.44 in IoU, 82.05 in F-measure, and 91.49 in Sα on the Kvasir-SEG dataset.

A novel methodology proposed by Zhao et al. [27] for polyp segmentation to improve the accuracy of colorectal cancer detection. The

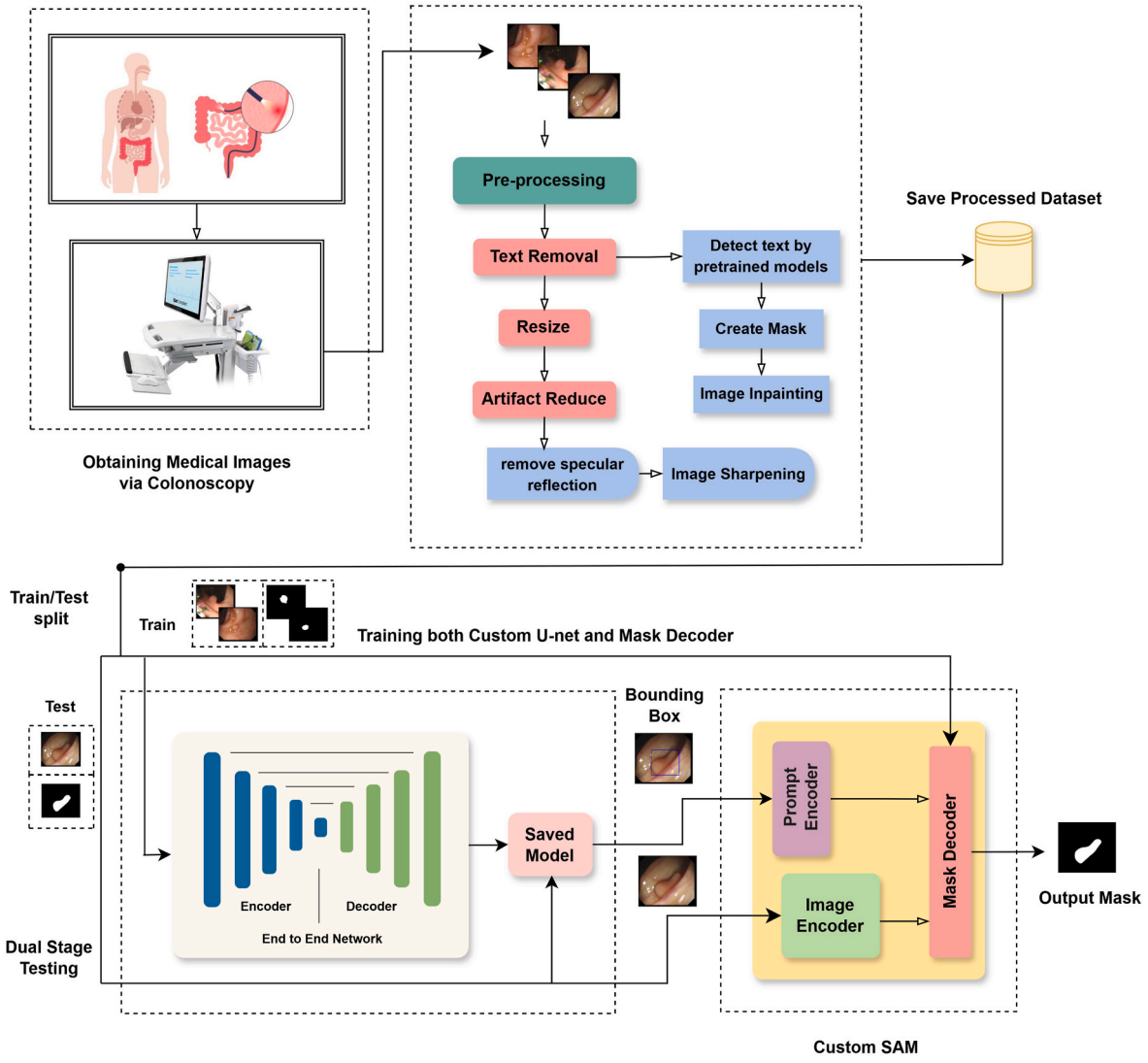


Fig. 1. The comprehensive workflow of SAMU-Net: Custom Attention-Based U-Net, Polyp Detection and Bounding Box Creation, and Segment Anything Model (SAM) for Quality Mask Prediction.

proposed method, MSNet, introduces a multi-scale subtraction network that effectively utilizes complementary information from different feature levels to enhance the perception of polyp areas. This is achieved through a subtraction unit (SU) that captures the difference features between adjacent levels in the encoder and a training-free network "LossNet" for comprehensive supervision across feature layers. MSNet achieves the best scores on several benchmark datasets with a mean Dice coefficient improvement of up to 14.1 % on challenging datasets like ETIS. The network also operates in real-time at approximately 70fps for 352×352 images, making it faster than other methods. Enhanced U-Net by Krushi et al. [28] introduce a model for polyp segmentation. The methodology includes using a combination of pixel-based IoU loss, focal loss, and dice loss to improve the model's learning capability, complemented by the addition of a Selective Feature Enrichment Module (SFEM) and Attention-Gated Context Module (AGCM). This model achieves remarkable results, with the best performance noted as a mean Dice of 88.62 % and mean IoU of 81.30 % on the CVC-300 dataset: Fan et al. [29] present PraNet, a model for accurate polyp segmentation from colonoscopy images. PraNet employs a parallel partial decoder (PPD) to generate high-level semantic maps and integrates reverse attention (RA) modules for enhanced accuracy. The model outperformed state-of-the-art approaches, achieving top scores such as a mean Dice of 0.899 and a mean IoU of 0.849 on the CVC-612 dataset, and a mean Dice of 0.898 and a mean IoU of 0.840 on the Kvasir dataset. Its superior performance, signified by over 7 % improvement in mean Dice across metrics compared to other models, The CPSNet by Wang et al. [30] is an innovative deep learning model designed to segment camouflaged and partially occluded colorectal polyps accurately. It incorporates three key modules: the Deep Multi-Scale-Feature Fusion Module (DMF), the Camouflaged Polyp Detection Module (CDM), and the Multi-Scale Feature Enhancement Module (MFEM). These modules work synergistically to enhance feature extraction, improve boundary localization, and effectively integrate shallow and deep features. CPSNet demonstrates superior performance, achieving a 2.3 % increase in the Dice coefficient on the ETIS-LaribPolypDB dataset compared to previous state-of-the-art methods.

Hao et al. [31] introduced Polyper, a boundary-sensitive method to improve polyp segmentation, especially for small polyps. It employs a specific feature aggregation strategy for small polyp detection. The methodology involves refining the initial segmentation results, mainly focusing on potential boundary extraction. The results exhibited significant enhancements with different encoders, such as ResNet-50 and MiT-B1, and showed marked improvements in mIoU and mDice scores. On datasets like Kvasir and CVC-ClinicDB, Polyper achieved the highest scores, with mIoU and mDice reaching up to 90.57 and 94.49, respectively, when combined with Swin-T encoder. CoInNet by Samir et al.

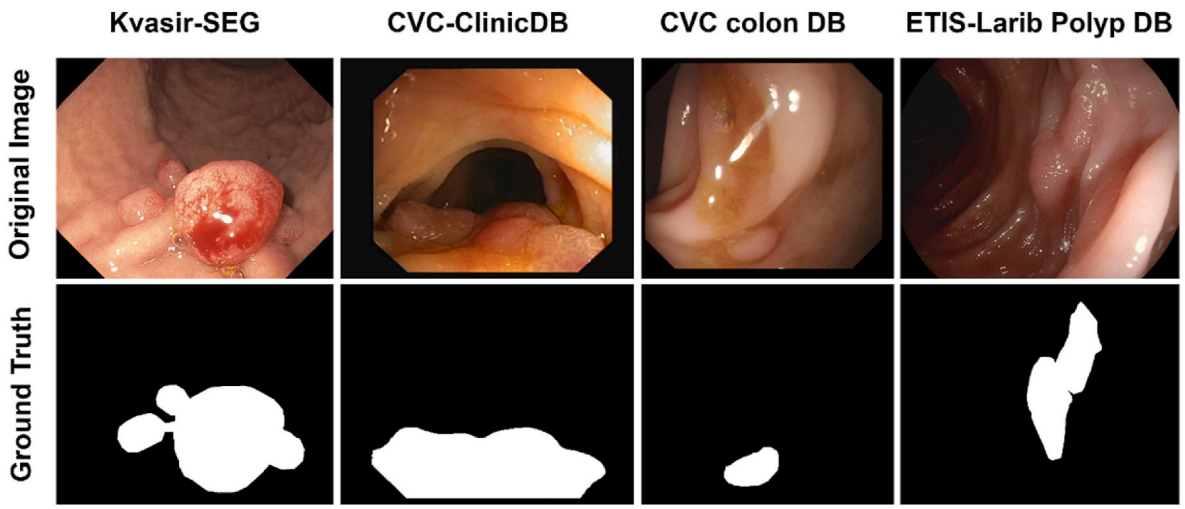


Fig. 2. Representative sample images and corresponding ground truth segmentation masks from each dataset respectively: Kvasir-SEG, ClinicDB, CVC-ColonDB, and ETIS-LaribPolypDB.

[32] is evaluated on five datasets and outperforms thirteen state-of-the-art models, demonstrating superior scores such as a mDice of 93.0 % and a mIoU of 90.25 % on the CVC-ClinicDB dataset, alongside 92.6 % mDice and 87.2 % mIoU on the Kvasir dataset. LightCF-Net by Zhanlin et al. [33] is a novel, lightweight, long-range context fusion network designed for real-time polyp segmentation from colonoscopy videos. This approach results in higher segmentation accuracy and efficiency than other lightweight networks. LightCF-Net demonstrated superior accuracy and precision on the Kvasir-SEG and CVC-ClinicDB datasets. It achieved the highest IoU and DSC values among the models tested, with IoU of 79.02 % and DSC of 88.28 % on Kvasir-SEG and IoU of 68.00 % and DSC of 80.95 % on CVC-ClinicDB.

Junqing et al. [34] propose the IECFNet model to address challenges in polyp segmentation from gastrointestinal endoscopy images, including low contrast boundaries and varied appearances. The methodology involves using an attention encoding-decoding pair to generate saliency maps, an implicit edge-enhanced context attention module for feature aggregation, and a multi-scale feature reasoning module for final predictions. IECFNet outperforms existing methods significantly, particularly with a 7.9 % higher accuracy on the ETIS dataset and 90.7 % Mean dice on the Kvasir Dataset. Guangyu et al. [35] aim to overcome the challenges posed by labor-intensive and expensive annotation processes in polyp segmentation. Their proposed semi-supervised methodology, employing collaborative and adversarial learning, introduces collaborative segmentation networks featuring focused and dispersive extraction modules. Under mutual consistency constraints, two networks are trained to mitigate biases stemming from limited labeled data. Adversarial training, incorporating an auxiliary discriminator, enhances segmentation performance using unlabeled data. The experimental evaluation on Kvasir-SEG and CVC-Clinic DB datasets demonstrates the model's superiority over existing semi-supervised and fully supervised methods. Faysal et al. [36] present a model designed for precise polyp segmentation in medical images to aid in colorectal cancer diagnosis. The methodology involves using a UNet architecture with an InceptionResNetV2 encoder for feature extraction and applying Test Time Augmentation (TTA) to improve segmentation accuracy, achieving the best Kvasir-SEG and CVC-ClinicDB datasets performance. The highest recorded DSC for this model is 0.8706, and a mean IoU of 0.8016. IRv2-Net's superior performance, real-time applicability.

3. Methodology

This study proposes a new dual stage approach for polyp segmentation, a combination of a modified U-Net and a custom zero-shot segmentation model based on SAM (Segment Anything Model). The first stage includes a modified U-Net architecture in the meantime to come up with a primary binary mask from the input image. The binary masks produced from the original images are then applied to create bounding boxes around the polyp ROI(Region Of Interest). In the second stage, the bounding boxes and the original images are feeded to the custom SAM. Mask decoder of SAM is fine-tuned on the same training dataset used on the first stage. This fine-tuning enables SAM to produce higher-quality binary masks based on the segmentation result by Custom U-Net. The bounding box provides contextual information that helps the SAM model accurately locate and segment the polyp regions within the bounding box. By fusing these two stages, our methodology aims to enhance the quality and reliability of polyp segmentation. Fig. 1 illustrates the overall architecture of SAMU-Net.

Table 1
An overview of the datasets used in this study.

Dataset	Number of Images	Image Resolution	Format
Kvasir SEG	1000	Varying	JPG
CVC-ClinicDB	612	384 x 288	PNG
CVC-ColonDB	380	574 x 500	PNG
ETIS-Larib	196	1225 x 966	TIF

3.1. Dataset description

This study employs four separate datasets to assess the effectiveness of the proposed SAMU-Net model for polyp segmentation. Fig. 2 shows the patient's colorectal organ represented by the datasets used in this study and Table 1 provided an overall overview of datasets.

1. Kvasir-SEG Dataset [37]: A fiduciary dataset was generated based on the Kvasir dataset, including images from the gastrointestinal tract. In total, the dataset contains 1000 polyp images with associated ground truth segmentation masks. All images are JPEG and vary in resolution, while their segmentation masks are in PNG format.
2. CVC-ClinicDB [38]: It is also a publicly available dataset, containing a total of 612 images of a resolution of 384 x 288 pixels. They have been extracted from 29 colonoscopy videos, hence providing rich variability regarding the different polyp appearances. There are ground-truth segmentation masks available for every image in the dataset.
3. CVC-ColonDB [39]: CVC-ColonDB includes 380 images from 15 colonoscopy videos. All of the images within this dataset have a

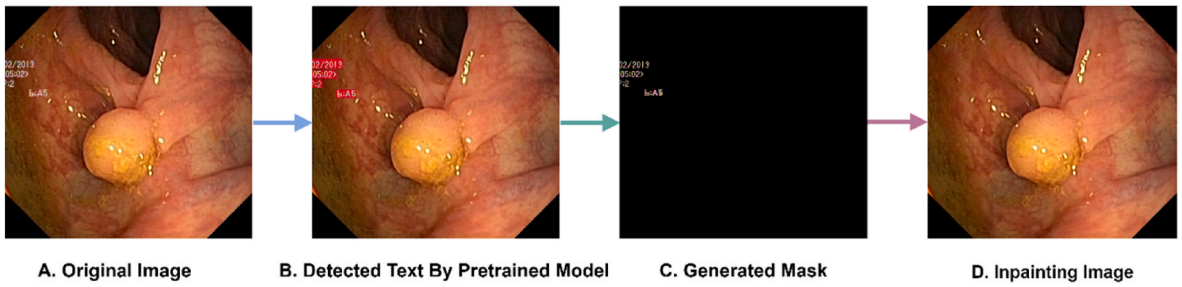


Fig. 3. Text Removal Process by utilizing a pre-trained OCR model and OpenCV's inpainting algorithm to eliminate text from images.

resolution of 574 x 500 pixels and are stored in PNG format. This dataset provides much diversity concerning polyp shapes, sizes, and textures with the ground-truth segmentation masks.

4. ETIS-Larib Polyp DB [40]: ETIS-Larib Polyp DB is constituted by 196 high-resolution colonoscopy images of 1225 x 966 pixels. The reason this dataset is particularly challenging is that all images are high-resolution and the polyps to be detected have a very different appearance. Ground truth segmentation masks are provided for every image.

3.2. Image preprocessing techniques

Image preprocessing is an essential step before they are ready to be inputted in the models. It reduces the time needed for the calculation and increases computational efficiency. The purposes are to generalize and enhance the quality of images while bringing out unwanted distortions in it and emphasizing important features. Various types of artifacts could influence the performance of polyp segmentation in colonoscopy pictures. Colonoscopy images may contain noises, light reflections, and size and aspect ratio mismatch. For seamless handling of segmentation of these images by the models, appropriate pre-processing techniques are required. This section explains how techniques such as scaling, sharpening, and removal of artifacts enhance the quality of images obtained from colonoscopy. As for normalization, images are resized to 256x356 resolution to ensure uniformity across the dataset.

3.2.1. Text removal

The performance of model could be affected by inappropriate text existing in some images in the dataset. In this phase of preprocessing, the use of a pre-trained OCR model is put into action in the removal of text accompanied by openCV's inpainting algorithm. First, obtain the bounding frames by recognizing the text in the image. Second, generate a mask. Third, paint areas with the test. These are the three phases. Keras-OCR is used to find the text's bounding frame. Keras-OCR provides infrastructure for end-to-end training and out-of-the-box OCR models to rapidly develop novel OCR models [41]. The pretrained weights for detectors and recognizers are automatically downloaded. The pretrained model is used in this research because it serves the purpose adequately. Passing an image through Keras-OCR will result in a tuple of the form (word, box). The box contains the coordinates (x, y) of the four corner boxes of the word. After the text is recognized, a mask of the size of the input images is generated containing only the text. This information is then fed into the algorithm to determine where exactly in the image the painting needs to be done. Finally, the obscured regions of the image are inpainted using an inpainting algorithm. We used cv2. INPAINT_NS [42]. This algorithm thus uses fluid dynamics and relies on partial differential equations. The output image does not show any text upon inpainting. The entire process of removing text from an image is illustrated in Fig. 3.

3.2.2. Artifact remove

Several morphological techniques for polyp segmentation are applied in the preprocessing stage to enhance image quality and facilitate accurate segmentation. This section outlines the methods employed, including specular reflection removal and sharpening the images.

3.2.2.1. Specular reflection removal. Specular reflections, caused by the intense light source from the colonoscope reflected in the colon outline, often cause bright spots or patches on colorectal images, obstructing details for segmentation tasks. We employed a robust method based on reflective region thresholding and morphological operations to remove and inpaint those patches. Fig. 4 shows the specular reflection removal process.

Specular regions are identified based on their distinctive intensity range in the grayscale representation of the image. Given a grayscale

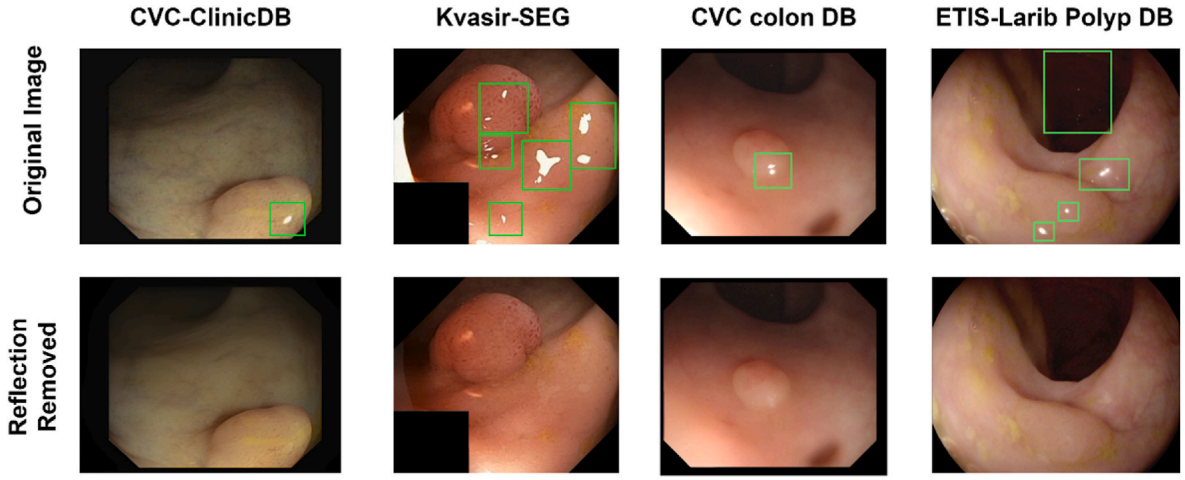


Fig. 4. Specular Reflection Removal Process using reflective region thresholding and morphological operations to eliminate bright spots from colorectal images.

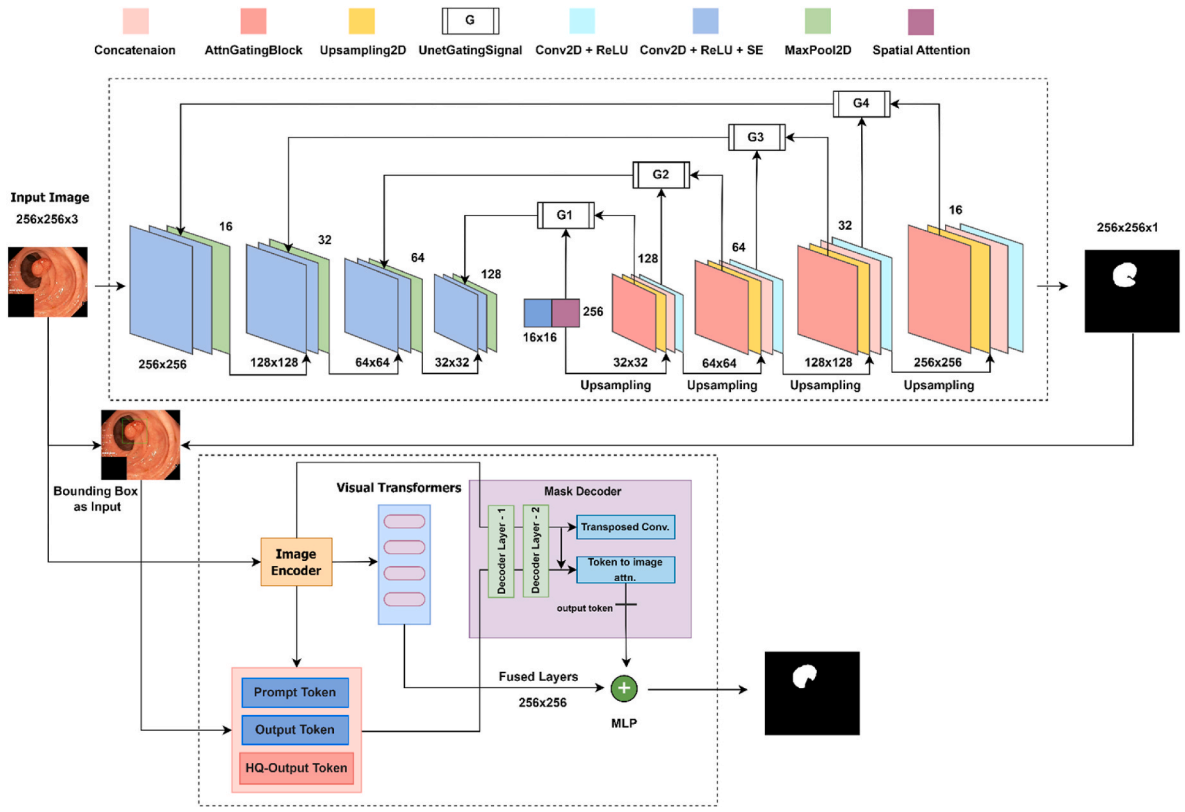


Fig. 5. Architecture of SAMU-Net: Dual-stage Polyp Segmentation Model integrating custom attention-based U-net with advanced attention mechanisms, custom blocks, and high-quality token utilization in custom SAM for refined output Segmentation.

image $I_{\text{gray}}(x,y)$ specular reflections are localized using a thresholding approach:

$$\text{mask}\ (x,y) = \begin{cases} 255, \text{if} T_{min} \le I_{\text{gray}}(x,y) \le T_{max} \\ 0, otherwise \end{cases} \tag{1}$$

Here, T_{min} and T_{max} define the lower and upper bounds of the reflective region threshold, respectively. From our observation, reflective spots have an RGB range of 245–255 for each channel. Based on that observation, Morphological operations are applied to the binary mask to enhance the accuracy of specular reflection removal. A closing operation with a larger kernel size $K \times K$ where $K = 15$, is performed to fill small holes and smooth out the edges:

$$mask_{closed} = close\ (mask, K) \tag{2}$$

where close denotes the morphological closing operation.

The final step involves inpainting, where image I is restored by replacing and brushing the pixels in the specular regions identified by the mask closed with neighboring pixel values. This is achieved using the following operation for seamless inpainting which can be written as following:

$$I_{\text{processed}}(x,y) = \begin{cases} I(x,y), \text{if mask}_{\text{closed}}\ (x,y) = 0 \\ \text{inpaint}(I, \text{mask}_{\text{closed}}, (x,y), 7), \text{if mask}_{\text{closed}}\ (x,y) = 255 \end{cases} \tag{3}$$

where inpaint $I, \text{mask}, (x,y), r$ denotes the inpainting function applied to image pixel coordinates (x,y) and radius r.

3.2.2.2. Image sharpening. The image sharpening technique in this study utilizes a convolution operation with a standard image sharpening kernel or filter. The sharpening kernel that was employed here can be defined as follows:

$$K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{4}$$

This kernel enhances the edges by amplifying the differences between adjacent pixel values, effectively highlighting the edges and details in the image. This operation applied to the input image I with the kernel K can be mathematically represented as:

$$I_{\text{sharpened}}(x,y) = I(x,y) * K \tag{5}$$

where $I_{\text{sharpened}}(x,y)$ denotes the pixel value of the sharpened image at coordinates (x, y) and * represents the convolution operation. The application of this sharpening filter effectively removes blurriness from the input images, enhancing the edge details and contrast.

3.3. Proposed model

The architecture of our proposed model starts with a modified U-Net which is tailored to use an attention mechanism to detect and segment polyp. The network follows the classic encoder-decoder structure of U-Net but introduces significant modifications to improve feature extraction and mask precision. Fig. 5 illustrates the main architecture of our proposed model.

The encoder of Custom U-Net consists of a series of convolutional blocks with squeeze-and-excitation (SE) mechanisms. Each block includes two convolutional layers with 3x3 kernels, batch normalization, and ReLU activation, followed by an SE block that adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. This process is repeated through four stages, where each subsequent stage doubles the number of filters, enhancing the model's capacity to learn complex features. After each stage, a max pooling layer is used to progressively down-sample the

feature maps, thereby capturing multi-scale contextual information.

At the bottleneck of the network, a bridge layer is employed, comprising a convolutional block followed by a spatial attention mechanism. The spatial attention mechanism emphasizes important regions by computing attention maps that capture the spatial dependencies of the feature maps. This configuration aims to refine the most salient features before up-sampling, guiding the model to focus on areas of interest.

The Spatial Attention Block generates a spatial attention map by applying average pooling and max pooling along the channel axis of the input feature maps, concatenating the pooled features, and passing them through a convolutional layer with sigmoid activation. This spatial attention map is multiplied by input feature maps to focus on critical spatial regions.

$$\psi = \sigma(W([\mathrm{AvgP}(x), \mathrm{MaxP}(x)])) \cdot x \tag{6}$$

The input feature maps x undergo average pooling x and max pooling x: The results are concatenated and passed through a convolutional layer W. Finally, a sigmoid activation function σ is applied, producing attention weights highlighting important spatial regions in the input.

Upwards from the deepest part of the custom U-Net where resides the spatial attention block, the decoder blocks are deployed. Each up-sampling step in the decoder is preceded by generating a gating signal through a 1x1 convolution, batch normalization, and ReLU activation. The attention-gating blocks use these gating signals to selectively highlight relevant features from the encoder path, ensuring that only the most critical information is passed through. The up-sampling layers progressively reconstruct the image resolution, with each layer concatenating the up-sampled feature maps with the corresponding attention-weighted encoder outputs. This ensures a rich spatial and contextual information fusion at each decoding stage.

As for the Attention Gating Block, it is designed to highlight salient features passed from the encoder to the decoder. It takes two inputs: the feature maps from the encoder and the gating signal generated in the decoder. The block first performs a 2x2 convolution on the encoder feature maps and a 1x1 convolution on the gating signal. These outputs are summed and passed through a ReLU activation, followed by a 1x1 convolution and a sigmoid activation to generate attention coefficients. These coefficients are up-sampled and multiplied with the original encoder feature maps to focus on the relevant regions.

$$\psi = \sigma\big(W_g(g) + W_x(x)\big) \cdot x \tag{7}$$

where, x Encoder feature maps, g Gating signal from the decoder, W_g, W_x: Convolutional layers for g and x. σ: Sigmoid activation function.

For the loss function of the first stage in Custom U-Net, we employ the Dice loss function. The Dice coefficient D is defined as:

$$D = \frac{2 \cdot |P \cap G|}{|P| + |G|} \tag{8}$$

where P is the set of predicted pixels, and G is the set of ground truth pixels. The Dice loss L is then given by:

$$L = 1 - D = 1 - \frac{2 \cdot \sum (P \cdot G) + \epsilon}{\sum P + \sum G + \epsilon} \tag{9}$$

Here, ϵ is a small constant to avoid division by zero.

3.4. Segment Anything model (SAM)

The second stage of our segmentation network utilizes the Segment Anything Model (SAM). It's a recently released general segmentation model with robust zero-shot segmentation capabilities across a wide range of images of objects. We focus on enhancing SAM's mask decoder, inspired by techniques used in "Segment Anything in High Quality" for edge accurate mask prediction without extensive model retraining. The other two major components that are Image and Prompt encode will remain frozen during the training state.

SAM's default mask decoder uses an output token akin to edge-to-edge object Detection Transformer's (DETR) [43] object query for dynamic MLP (Multi-Layer Perceptron) based mask prediction. In our adaptation of HQ-SAM, we adapt the HQ-Output [44] token alongside SAM's output tokens and prompt tokens and then augment the mask decoder's input. This HQ-Output token is a learnable 1×256 vector that undergoes self-attention with other tokens, enhancing global context integration and refining mask details through token-to-image and image-to-token attention mechanisms across decoder layers.

We adapted global-local fusion in HQ-SAM to enrich feature fidelity. This process unites enriched features from SAM's ViT(Visual Tranformers) encoder stages: early-layer local features capturing edge details, final-layer global features for semantic context, and mask decoder features emphasizing shape information. These features are upsampled to 256×256 and fused via element-wise summation that bolsters HQ-SAM's ability to preserve segmentation details effectively with minimal computational overhead. The encoder Se takes an image I as input and outputs the corresponding features f1:

$$\mathrm{f_I} = S_e(\mathrm{I}) \tag{10}$$

The decoder Sd takes the features $\mathrm{f_i}$ and a set of prompts P as input and outputs the corresponding 2D initial segmentation mask $\mathrm{M_{SAM}}$,

$$\mathrm{M_{SAM}} = S_d(\mathrm{f_I}, \mathscr{P}) \tag{11}$$

The prompts $\mathrm{p} \in \mathrm{P}$ can be points, boxes, texts, and masks. We used bounding boxes as our prompt in (x_1, y_1, x_2, y_2) coordination format.

To enhance the mask prediction, the HQ-Output token THQ was employed which is a learnable 1×256 vector. The tokens (existing T and HQ-Output THQ) undergo a self-attention mechanism:

$$\widehat{T} = \mathit{SelfAttention}(T + T_{\mathrm{HQ}}) \tag{12}$$

We extract multi-scale features from different stages of the ViT encoder as Early-layer features $\mathrm{f_{local}}$ for capturing edge details, Final-layer features $\mathrm{f_{global}}$ for semantic context and Mask decoder features fmask emphasizing shape information.

$$f^{256}_{\mathrm{local}} = \mathrm{Upsample}(f_{local}), f^{256}_{\mathrm{global}} = \mathrm{Upsample}\left(f_{\mathrm{global}}\right), f^{256}_{\mathrm{mask}} = \mathrm{Upsample}\left(f_{\mathrm{mask}}\right) \tag{13}$$

The upsampled features are then fused via element-wise summation:

$$f_{\mathrm{fused}} = f^{256}_{\mathrm{local}} \oplus f^{256}_{\mathrm{global}} \oplus f^{256}_{\mathrm{mask}} \tag{14}$$

where $\oplus$ denotes element-wise summation.

The final segmentation mask MHQ is produced from the fused features using a final convolutional layer:

$$M_{HQ} = conv\left(f_{\mathrm{fused}}\right) \tag{15}$$

To train the model, we use a combination of Dice loss and Binary Cross-Entropy (BCE) loss:

$$\mathit{DiceLoss} = 1 - \frac{2|M_{\mathrm{pred}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{pred}}| + |M_{\mathrm{gt}}|} \tag{16}$$

$$\mathit{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}\left[M_{\mathrm{gt}} \log\left(M_{\mathrm{pred}}\right) + \left(1 - M_{\mathrm{gt}}\right)\log\left(1 - M_{\mathrm{pred}}\right)\right] \tag{17}$$

Where, M_{pred} is the predicted mask, M_{gt} is the ground truth mask and N is the total number of pixels in the mask.

The combined loss is given by:

$$\mathit{TotalLoss} = \lambda_{\mathrm{Dice}} \cdot \mathit{DiceLoss} + \lambda_{\mathrm{BCE}} \cdot \mathit{BCELoss} \tag{18}$$

where λ_{Dice} and λ_{BCE} are weighting factors.

The overall training procedure involves minimizing the total loss

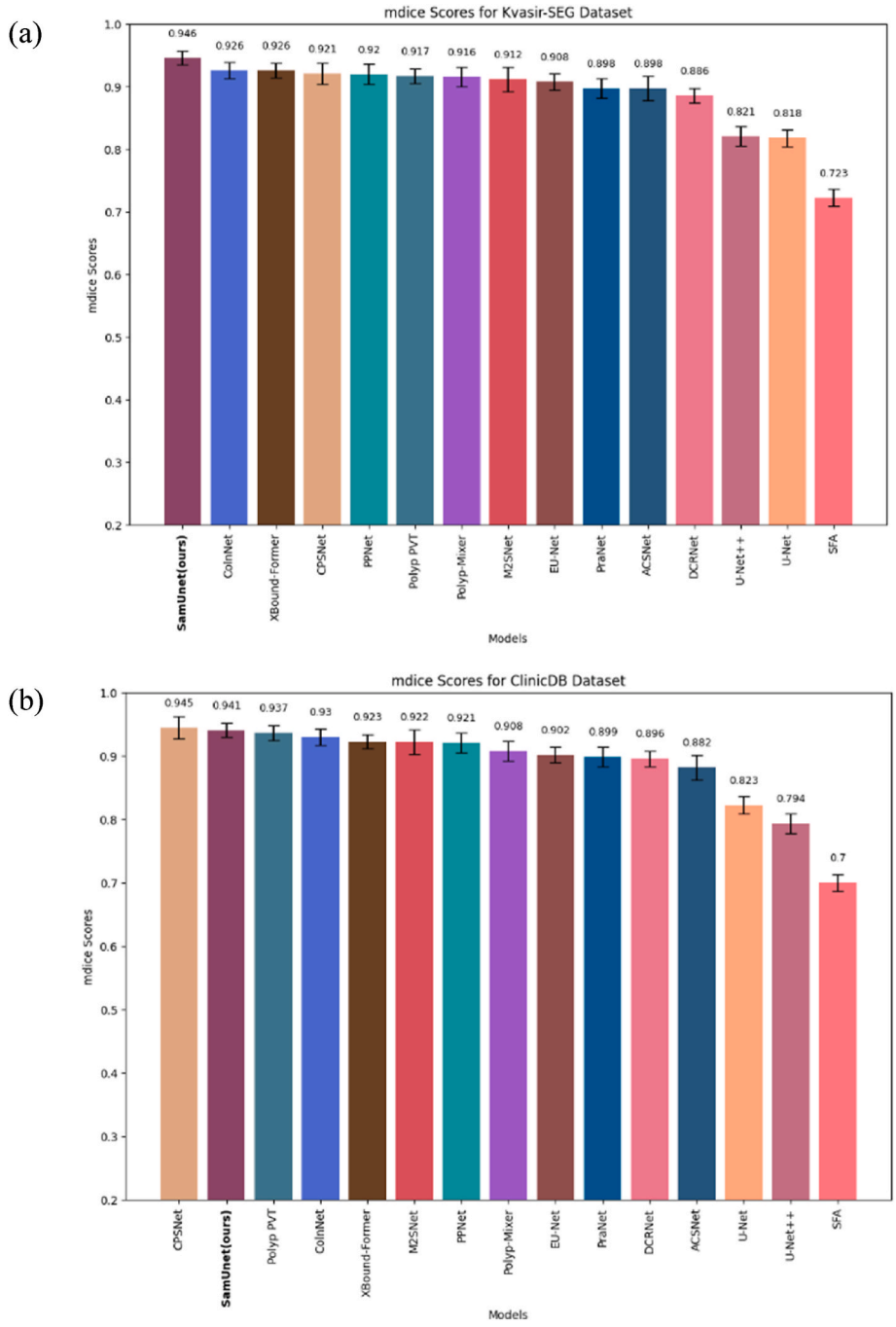


Fig. 6. Evaluate SAMU-Net and Other Models on the Kvasir SEG(a) and ColonDB(b) Datasets Using mDice Score Comparison.

over the training dataset:

$$\theta^{*} = \arg\min_{\theta} \sum_{(I,P,M_{gt}) \in \text{train set}} Total\ Loss(I, P, M_{gt}; \theta) \tag{19}$$

where θ represents the model parameters.

Unlike conventional approaches that may require extensive fine-tuning or additional heavy networks, our method focuses on enhancing mask quality through efficient token learning and feature fusion. This strategy significantly improves segmentation quality.

4. Results

This section discusses the results, including experimental setup, data splitting, a statistical analysis of the segmentation models, and train loss.

4.1. Experimental setup

In our experiments, we implemented the model using the Tensor-Flow, Pytorch framework and the model was trained on a system with an Nvidia A100 GPU(Graphics Processing Unit) and standard 8 core CPU (Central Processing Unit). We used the RMSprop optimizer with a learning rate set to 1e-4. We defined the model with 16 filters for stage one and to ensure reproducibility, we set a constant seed value. The training process iterated for 60 epochs for each stage of SAMU-Net and the model's save checkpoint was threshold for 0.2 validation loss. The dataset was divided into three subsets: 80 % for the training set, 10 % for the validation set, and 10 % for the testing set. The data loading function

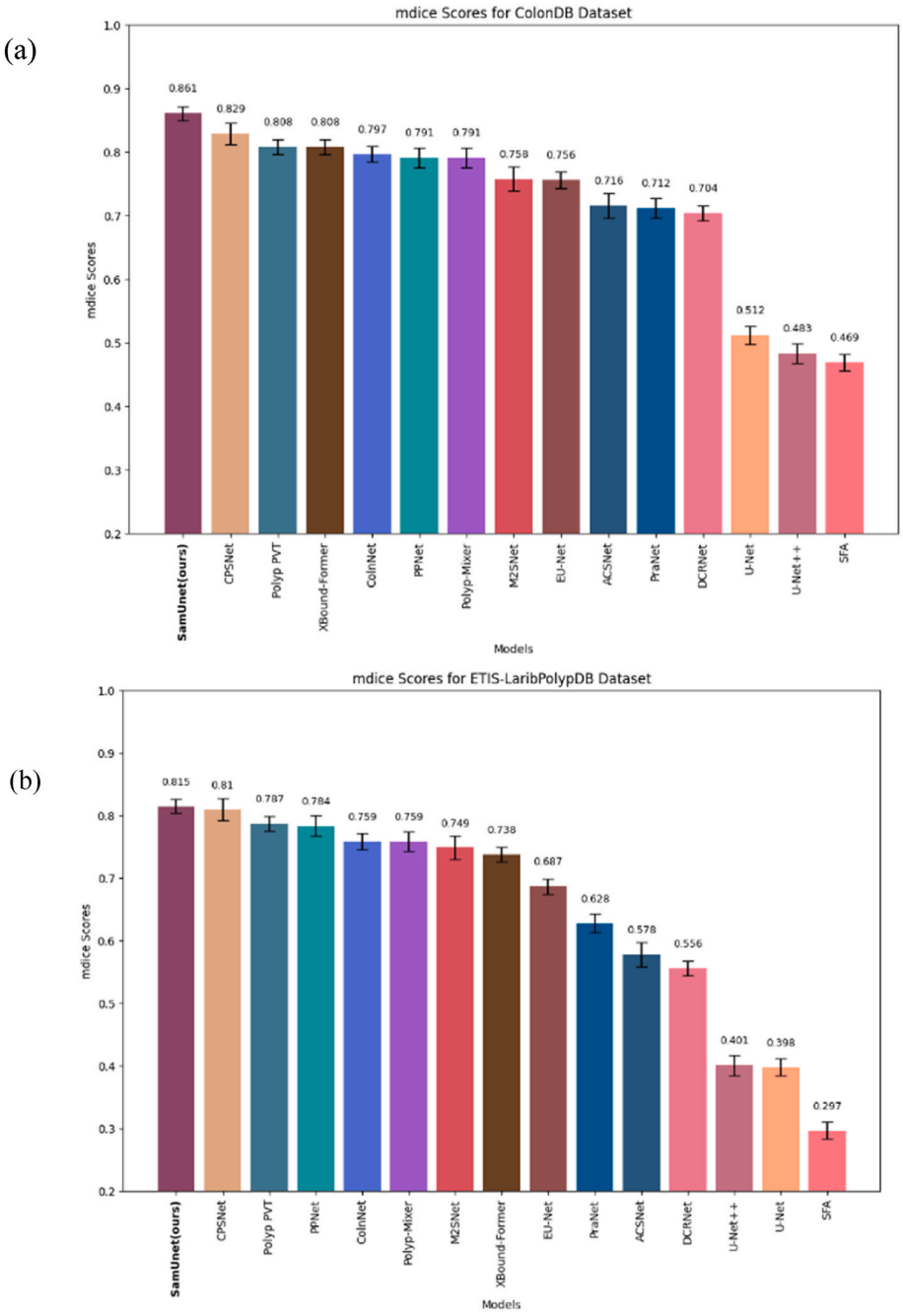


Fig. 7. Evaluate SAMU-Net and Other Models on the ColonDB(a) and ETIS(b) Datasets Using mDice Score Comparison.

was used to handle 3 primary image file formats in four different datasets. Specifically, JPEG images were used for the Kvasir dataset, TIF images for the ETIS-LaribPolypDB datasets, and PNG images for the CVC-ClinicDB and CVC-ColonDB dataset.

4.2. Evaluation metric

We assess the model's efficacy using six commonly used metrics: Dice, IoU, Mean Absolute Error (MAE), Weighted F-measure (F_β^w), S-measure (Sα), and E-measure (Eξ). Two metrics that fall under this category are IoU and Dice, both of which are regional similarity measures that primarily analyze the internal consistency of segmented objects. This report mentions the average value of Dice (mDic) and IoU (mIoU). As an indication of pixel-by-pixel comparison, MAE depicts the average absolute error between the anticipated and true values. Weighted F-measure (F_β^w) considers recall and accuracy thoroughly and eliminates the impact of treating each pixel equally in traditional indicators. S-measure (Sα) is centered around the structural similarity of target possibilities at the level of both regions and objects. To assess the segmentation outcomes on both the pixel and picture levels, the E-measure (Eξ) is employed. As mEξ and maxEξ, we represent the average and maximum values of the E-measure, respectively. Additionally, Precision, Recall, and Accuracy are key metrics that measure the model's ability to correctly identify positive samples and avoid false positives and negatives. Precision reflects the ratio of correctly predicted positive pixels out of all predicted positive pixels, while Recall represents the ratio of correctly predicted positive pixels out of all actual positive pixels. Accuracy measures the overall correctness of the prediction, including both positive and negative samples. The evaluation metrics are represented by Equations (20)–(28).

Table 2
Performance evaluation on Kvasir-SEG dataset.

Study	Model	mIoU	F_β	S_α	mE_ξ	$maxE_\xi$	MAE
Olaf et al. [18] (2015)	U-Net	0.746	0.794	0.858	0.881	0.893	0.055
Zongwei et al. [45] (2018)	U-Net++	0.743	0.808	0.862	0.886	0.909	0.048
Jiacheng et al. [46] (2022)	XBound-Former	0.871	0.939	0.918	0.968	–	**0.016**
Yuqi et al. [47] (2019)	SFA	0.611	0.670	0.782	0.834	0.849	0.075
Ruifei et al. [48] (2023)	ACSNet	0.838	0.882	0.920	0.941	0.952	0.032
Deng et al. [29] (2020)	PraNet	0.840	0.885	0.915	0.944	0.948	0.030
Zijin et al. [26] (2022)	DCRNet	0.825	0.868	0.911	0.933	0.941	0.035
Krushi et al. [28] (2021)	EU-Net	0.854	0.893	0.917	0.951	0.954	0.028
Jing et al. [49] (2023)	Polyp-Mixer	0.864	0.908	**0.932**	0.959	0.967	–
Dong et al. [25] (2024)	Polyp PVT	0.864	0.911	0.925	0.956	0.962	0.023
Xiaoqi et al. [50] (2023)	M2SNet	0.861	0.901	0.922	0.953	–	0.025
Keli et al. [51] (2023)	PPNet	0.878	0.911	0.927	0.949	–	0.024
Wang et al. [30] (2024)	CPSNet	0.868	0.912	0.926	0.960	0.963	0.023
Samir et al. [52] (2023)	ColnNet	0.872	0.939	0.926	**0.97**9	–	0.02
Ours	**SAMU-Net**	**0.882**	**0.962**	0.922	0.96	**0.983**	0.06

Table 3
Performance evaluation on the CVC-ClinicDB dataset.

Study	Model	mIoU	F_β	S_α	mE_ξ	$maxE_\xi$	MAE
Olaf et al. [18] (2015)	U-Net	0.755	0.811	0.889	0.913	0.954	0.019
Zongwei et al. [45] (2018)	U-Net++	0.729	0.785	0.873	0.891	0.931	0.022
Jiacheng et al. [46] (2022)	Xbound-Former	0.875	0.937	0.942	0.974	–	0.008
Yuqi et al. [47] (2019)	SFA	0.607	0.647	0.793	0.840	0.885	0.042
Ruifei et al. [48] (2023)	ACSNet	0.826	0.873	0.927	0.947	0.959	0.011
Deng et al. [29] (2020)	PraNet	0.849	0.896	0.936	0.963	0.979	0.009
Zijin et al. [26] (2022)	DCRNet	0.844	0.890	0.933	0.964	0.978	0.010
Krushi et al. [28] (2021)	EU-Net	0.846	0.891	0.936	0.959	0.965	0.011
Jing et al. [49] (2023)	Polyp-Mixer	0.856	0.902	0.943	0.963	0.968	–
Dong et al. [25] (2024)	Polyp PVT	0.889	0.936	0.949	0.985	0.989	0.006
Xiaoqi et al. [50] (2023)	M2SNet	0.880	0.917	0.942	0.97	–	0.009
Keli et al. [51] (2023)	PPNet	0.878	0.913	0.947	0.969	–	0.008
Wang et al. [30] (2024)	CPSNet	0.900	0.949	**0.954**	0.990	**0.993**	0.006
Samir et al. [52] (2023)	ColnNet	0.887	0.94	0.952	0.987	–	0.006
Ours	**SAMU-Net**	**0.904**	**0.95**	0.951	**0.992**	0.991	**0.006**

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{20}$$

where A is the set of ground truth pixels and B is the set of predicted pixels.

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{21}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|P_i - G_i| \tag{22}$$

where N is the total number of pixels, P_i is the predicted value, and G_i is the ground truth value.

$$F_\beta = \frac{(1+\beta^2)\cdot\text{Precision Recall}}{\beta^2.\text{Precision} + \text{Recall}} \tag{23}$$

$$S_\alpha = \alpha\cdot S_o + (1-\alpha)\cdot S_r \tag{24}$$

where S_o is the object-level similarity and S_r is the region-level similarity, and α is a weight parameter.

$$E_\xi = \frac{1}{N}\sum_{i=1}^{N}\Phi_{FG}(P_i, G_i) + \Phi_{BG}(P_i, G_i) \tag{25}$$

where Φ_{FG} and Φ_{BG} are the foreground and background similarities, respectively, and N is the number of pixels.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{26}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{27}$$

$$Accuracy = \frac{True\,Positives + True\,Negatives}{True\,Positives + True\,Negatives + False\,Positives + False\,Negatives} \tag{28}$$

4.3. Experimental results

The performance of our proposed SAMU-Net model was evaluated on four widely used polyp segmentation datasets and compared with fifteen state-of-the-art models. These include U-Net [Olaf et al., 2015], U-Net++ [Zongwei et al., 2018], SFA [Yuqi et al., 2019], ACSNet [Ruifei et al., 2023], PraNet [Deng et al., 2020], DCRNet [Zijin et al., 2022], EU-Net [Krushi et al., 2021], Polyp-Mixer [Jing et al., 2023], Polyp PVT [Dong et al., 2024], M2SNet [Xiaoqi et al., 2023], PPNet [Keli et al., 2023], CPSNet [Wang et al., 2024], ColnNet [Samir et al., 2023], and XBound-Former [Jiacheng et al., 2022]. Figs. 6–7 and Tables 2–5 shows

Table 4
Performance evaluation on the CVC-ColonDB dataset.

Study	Model	mIoU	F_β	S_α	mE_ξ	$maxE_\xi$	MAE
Olaf et al. [18] (2015)	U-Net	0.444	0.498	0.712	0.696	0.776	0.061
Zongwei et al. [45] (2018)	U-Net++	0.410	0.467	0.691	0.680	0.760	0.064
Jiacheng et al. [46] (2022)	XBound-Former	0.724	0.866	0.855	0.907	–	0.030
Yuqi et al. [47] (2019)	SFA	0.347	0.379	0.634	0.675	0.764	0.094
Ruifei et al. [48] (2023)	ACSNet	0.649	0.697	0.829	0.839	0.851	0.039
Deng et al. [29] (2020)	PraNet	0.640	0.699	0.820	0.847	0.872	0.043
Zijin et al. [26] (2022)	DCRNet	0.631	0.684	0.821	0.840	0.848	0.052
Krushi et al. [28] (2021)	EU-Net	0.681	0.730	0.831	0.863	0.872	0.045
Jing et al. [49] (2023)	Polyp-Mixer	0.706	0.768	0.862	0.893	0.899	–
Dong et al. [25] (2024)	Polyp PVT	0.727	0.795	0.865	0.913	0.919	0.031
Xiaoqi et al. [50] (2023)	M2SNet	0.685	0.737	0.842	0.869	–	0.038
Keli et al. [51] (2023)	PPNet	0.726	0.776	0.865	0.905	–	0.028
Wang et al. [30] (2024)	CPSNet	0.744	0.810	0.870	0.927	**0.930**	0.026
Samir et al. [52] (2023)	CoInNet	0.729	0.789	0.875	0.897	–	0.022
Ours	**SAMU-Net**	**0.756**	**0.872**	**0.881**	**0.936**	0.921	**0.004**

Table 5
Performance evaluation on ETIS-Larib Polyp DB dataset.

Study	Model	mIoU	F_β	S_α	mE_ξ	$maxE_\xi$	MAE
Olaf et al. [18] (2015)	U-Net	0.335	0.366	0.684	0.643	0.740	0.036
Zongwei et al. [45] (2018)	U-Net++	0.344	0.390	0.683	0.629	0.776	0.035
Jiacheng et al. [46] (2022)	XBound-Former	0.650	0.824	0.831	0.850	–	0.031
Yuqi et al. [47] (2019)	SFA	0.217	0.231	0.557	0.531	0.632	0.109
Ruifei et al. [48] (2023)	ACSNet	0.509	0.530	0.754	0.737	0.764	0.059
Deng et al. [29] (2020)	PraNet	0.567	0.600	0.794	0.808	0.841	0.031
Zijin et al. [26] (2022)	DCRNet	0.496	0.506	0.736	0.742	0.773	0.096
Krushi et al. [28] (2021)	EU-Net	0.609	0.636	0.793	0.807	0.841	0.067
Jing et al. [49] (2023)	Polyp-Mixer	0.676	0.711	0.863	0.875	0.884	–
Dong et al. [25] (2024)	Polyp PVT	0.706	0.750	0.871	0.906	0.910	0.013
Xiaoqi et al. [50] (2023)	M2SNet	0.678	0.712	0.846	0.872	–	0.016
Keli et al. [51] (2023)	PPNet	0.716	0.743	0.871	0.885	–	**0.013**
Wang et al. [30] (2024)	CPSNet	0.727	0.769	0.882	**0.918**	0.926	0.014
Samir et al. [52] (2023)	ColnNet	0.69	0.82	0.859	0.898	–	0.024
Ours	**SAMU-Net**	**0.75.8**	**0.854**	**0.901**	0.909	**0.935**	0.015

the performance comparison with other models.

This bar chart in Fig. 6(a) shows the mDice scores of different models on the well-uniformed Kvasir-SEG dataset, highlighting their high performance. Our proposed model scores the highest at 0.946. models like ColnNet, XBound-Former, and CPSNet also achieve strong results, with scores between 0.926 and 0.921. Traditional models such as U-Net and U-Net++ score lower, at 0.818 and 0.821, respectively, while the SFA model has the lowest score of 0.723.

The mDice score of the ClinicDB dataset is shown in Fig. 6(b). Images and masks in this dataset are in a decent shape which help models to achieve good performance. CPSNet leads with a score of 0.945, closely followed by SamUnet at 0.941. Models like Polyp PVT, ColnNet, and XBound-Former also perform robustly, with scores ranging from 0.937 to 0.932. Under performing models are also slightly below acceptable range, lowest score being 0.7 by SFA.

In this Fig. 7(a), the bar chart presents mDice scores of various models on the ColonDB dataset. This dataset's images are noticeably inconsistent due to excessive blurriness and other artifacts which results in variety in mdice scores. Our proposed model achieves a top score of 0.861 in this dataset. Other models like CPSNet, Polyp PVT, and XBound-Former also perform well, with scores above 0.8. early released models such as U-Net and U-Net++ show lower performance, with scores of 0.512 and 0.483, respectively. Fig. 7(b) illustrates mDice scores for multiple models on the ETIS-LaribPolypDB dataset. This chart reveals significant performance variations across different approaches. The dataset's heterogeneous nature likely contributes to the challenge, resulting in diverse mDice scores. While some models struggle, others demonstrate more robust performance. SAMU-Net achieves the highest score of 0.815, showcasing its effectiveness in handling this complex dataset's intricacies. CPSnet, PolypPVT and a few other models show good performance ranging from 0.81 to 0.75.

First, the model performance was tested on Kvasir-SEG and CVC-ClinicDB. From the results shown in Table 2, for the Kvasir-SEG dataset, the proposed SAMU-Net was superior to the rest of the state-of-the-art approaches, with 88.2 % mIoU. This is a pretty good improvement over the next-best model—that is, ColnNet—which obtained an mIoU of 87.2 %. Two more metrics are also improved by SAMU-Net, including an Fβ of 96.2 % and maxEξ of 98.3 %. These are on the CVC-ClinicDB dataset, shown in Table 3, where, with 90.4 %, SAMU-Net surpasses all the baselines and achieves the highest mIoU. Notably, in a model, it equally puts on a competitive Fβ score and mean absolute error of 0.006 but performs slightly below CPSNet.

To test the generalizability of SAMU-Net its performance was also evaluated on two additional datasets: CVC-ColonDB and ETIS-Larib Polyp DB. The results are recorded in Tables 4 and 5, respectively. SAMU-Net shows remarkable perfor-mance on the CVC-ColonDB dataset (Table 4) that significantly outperform all other models across most metrics. It achieves a mIoU of 75.6 %, representing sub-stantial improvement of 1.2 % over the next best model (CPSNet). For the ETIS-Larib Polyp DB dataset (Table 5) our model again demonstrates strong perfor-mance, achieving the highest scores in mIoU (75.8 %), Fβ (87 %), Sα (90.1 %), and *maxE* (93.5 %) that indicate high accuracy in polyp segmentation. It is clear from Tables 2–5 that SAMU-Net is the only approach that consistently performs better than or on par with state-of-the-art methods across most performance metrics on all four datasets. This exemplifies the robustness and generalizability of our proposed model. The superior performance of SAMU-Net particularly on challeng-ing datasets like CVC-ColonDB and ETIS-Larib Polyp DB can be attributed to its dual-stage architecture. That combines a custom attention-based U-Net with the Segment Anything Model. This allows

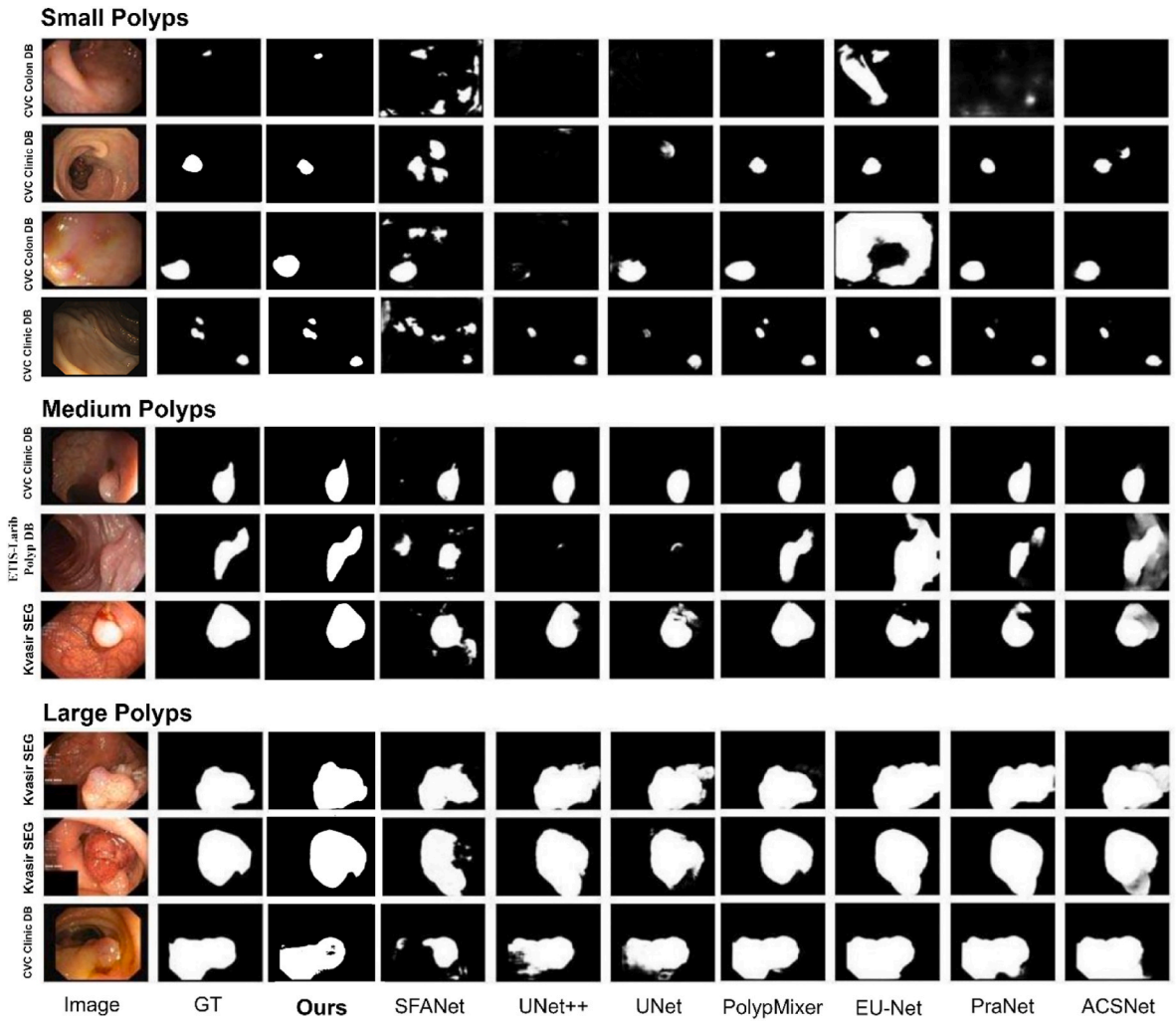


Fig. 8. Qualitative comparison of SAMU-net and state-of-the-art Polyp Segmentation models: Input images, ground truth, SAMU-Net's Segmentations, and Segmentation results from competing models.

Table 6
SAMU-net precision, recall, accuracy across different datasets.

Dataset	Precision	Recall	Accuracy
Kvasir SEG	0.958	0.924	0.973
CVC ClinicDB	0.936	0.938	0.979
CVC ColonDB	0.917	0.903	0.954
ETIS	0.921	0.918	0.941

for better distinction of obscure bounda-ries between polyp regions and normal mucosa, leading to a reduction in false positives and false negatives. Fig. 8 illustrates the qualitative comparison of our proposed model with other state-of-the-art methods for the segmentation task for polyp segmentation in different scenarios.

In addition to the Dice Score, we evaluated the performance of SAMU-Net using Precision, Recall, and Accuracy to provide a comprehensive analysis of its segmentation capabilities across multiple datasets. Table 6 below shows the model's performance on the Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, and ETIS datasets.The Kvasir-SEG dataset achieved the highest Precision of 0.958, highlighting SAMU-Net's ability to minimize false positives during polyp identification. Meanwhile, CVC-ClinicDB exhibited the highest Recall at 0.938, demonstrating the model's effectiveness in capturing true positive regions. The Accuracy values remained consistently high, with CVC-ClinicDB reaching the highest value of 0.979.

Among all light source configurations tested, as well as regarding types of polyp morphology, our approach was very constant in performance, and this allows it to gain superiority in delimiting the edges of the polyp correctly in front of the other models. For Large Polyps, our model stands out in the segmentation results related to polyps occupying a large part of the image. It shows robustness in handling complex cases where polyps have irregular shapes or presence of a variety of internal textures. While some other models need help with over- or under-segmentation, our approach maintains accuracy in these challenging scenarios.In each of these cases, our model—the column titled 'Ours'—maintains very near-accurate segmentation from the ground truth and often surpasses classical approaches such as SFANet, UNet++, UNet, PolypMixer, EU-Net, PraNet, and ACSNet. This comprehensive comparison underscores the effectiveness of our approach in addressing the diverse challenges presented in polyp segmentation tasks. Fig. 9 illustrates the training loss of SAMU-Net on the CVC-ClinicDB and CVC-ColonDB Dataset.

All datasets show that there is an effective early learning process accompanied by a rapid initial drop in training loss, as depicted in Fig. 9 (a). ClinicDB maintained the most stable learning curve, and there was a smooth, continuous takedown in loss during the training phase. ColonDB was found to be with the highest initial loss but experienced a gradual decline to converge eventually with those others. The performance lies in between for the case of Kvasir-SEG and ETIS datasets; however, Kvasir-SEG has the lowest final loss. Much more pronounced

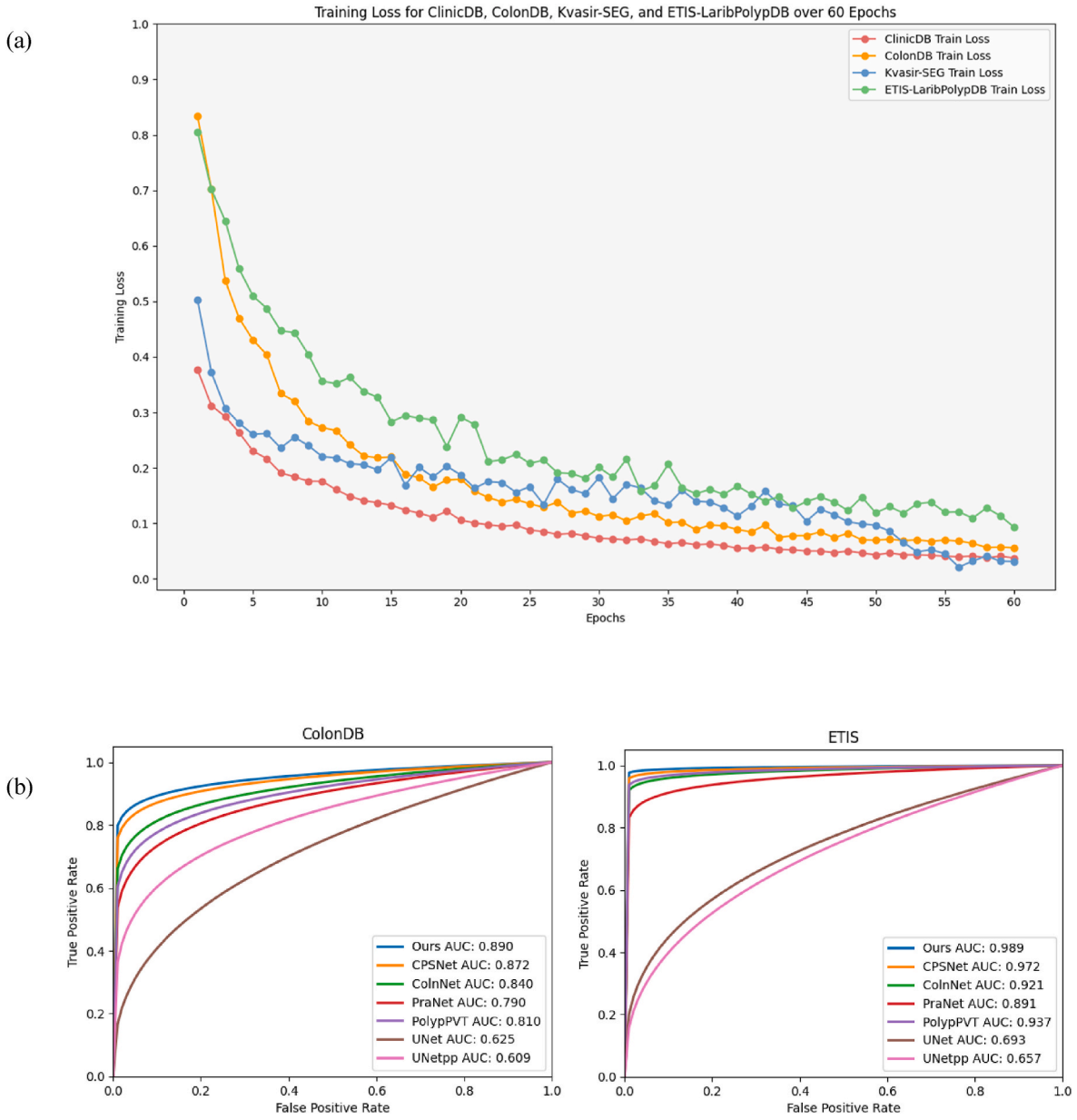


Fig. 9. (a) Train loss curves for SAMU-Net on CVC-ClinicDB, CVC-ColonDB, KvasirSEG, and ETIS Dataset (b) ROC curves on ColonDB and Etis Larib PolypDB for different models.

fluctuations are observed in the case of ETIS. This is evidence that the learning process is much harder for this dataset. For all datasets, it is seen that last epochs converge to low loss values ranging between 0 and 0.1. This comparison is thus interpretative in showing that the SAMU-Net is powerful enough to learn any given polyp segmentation task and can generalize its performance on different datasets while remaining unaffected by the changes in their different characteristics and initial challenges.

The ROC analysis on Fig. 9 (b) shows a boost in performance over the state-of-the-art methods on ColonDB and ETIS datasets. The Area Under the Curve (AUC) values further validate the robustness of our model, with higher AUC value indicate better overall performance. It can be shown from the AUC graphs that our model maintained a higher true positive rate while keeping the false positives at a minimum for more accurate and reliable results inCRC polyp segmentation.

4.3.1. Difficulties Encountered in Polyp Segmentation

Our polyp segmentation framework encountered two major challenging scenarios for a few images. Firstly, If the custom U-Net predicts the polyp ROI (Region of Interest) wrongly, then it causes discrepancy that impacts the subsequent stage where SAM performs segmentation based on the provided bounding box. Due to inherent limitation of not segmenting outside the bounding box for SAM, inaccuracies in the bounding box from stage 1 result in cropped segmentation outputs, thereby compromising the overall accuracy of the polyp segmentation.

Another difficulty arises when multiple bounding boxes overlap in stage 1. SAM attempts to unify overlapping bounding boxes into a single segmentation mask when encountering overlapping bounding boxes. This issue leads to inaccuracies in the unified mask creation, particularly when the overlapping regions do not represent an actual polyp structure predicted by custom U-Net. This issue underscores the complexity of handling overlapping instances during the segmentation process and the need for robust strategies to differentiate and accurately segment intensely cluttered multiple polyps or misplaced bounding boxes due to inaccurate ROI prediction. Fig. 10 shows the difficult cases of our model.

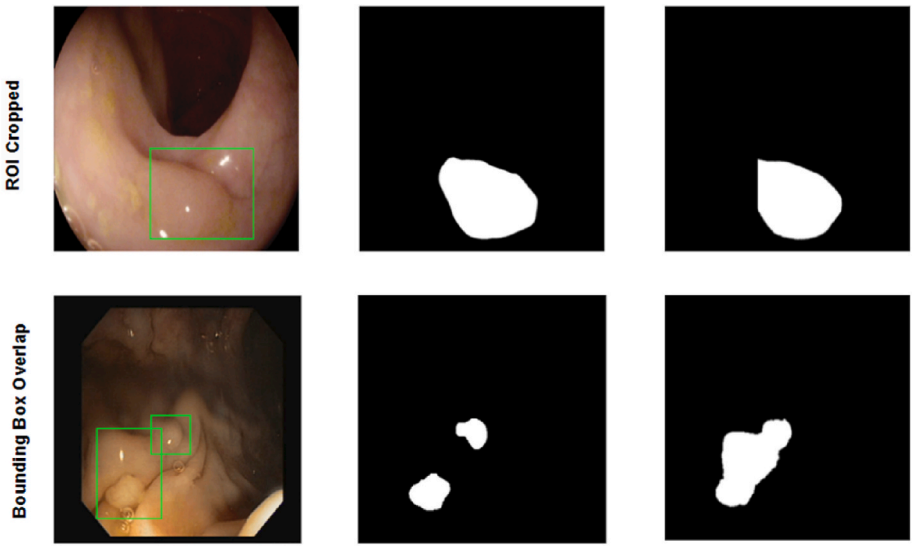


Fig. 10. Difficult Cases of the SAMU-Net model in medical image segmentation.

Table 7
Ablation study results - mean dice scores for different module combinations across multiple Polyp Segmentation datasets.

SL#	Custom U-Net	Custom SAM	SAM	Pre-Processed	mDice on Kvasir SEG	mDice on ClinicDB	mDIce on ColonDB	mDice on ETIS
1	✔			✔	0.853	0.869	0.765	0.703
2		✔		✔	0.918	0.862	0.814	0.742
3	✔		✔	✔	0.804	0.778	0.716	0.644
4		✔			0.906	0.848	0.782	0.730
5	✔		✔		0.795	0.767	0.698	0.629
6	✔	✔		✔	0.946	0.941	0.861	0.815

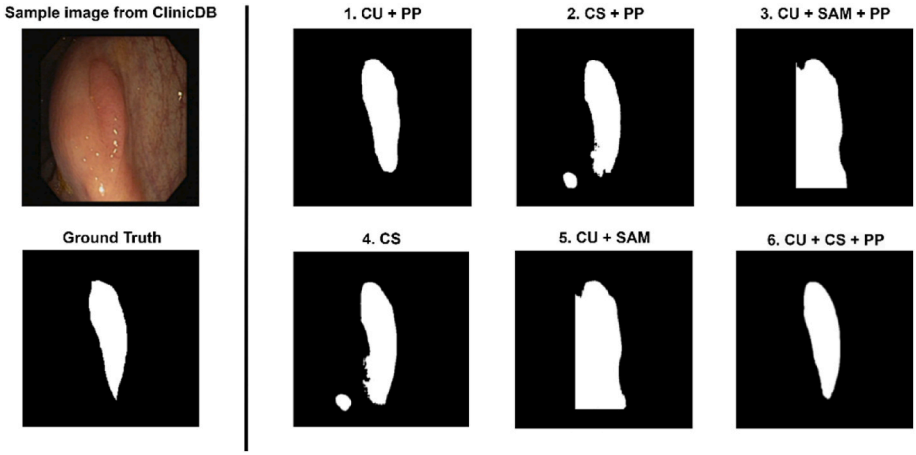


Fig. 11. Comparative analysis of qualitative results across experimental stages: Custom U-net (CU), pre-processed Data (PP), custom SAM (CS), and default segment Anything Model (SAM) with large Model checkpoint.

5. Ablation study

To rigorously evaluate the contributions of each stage within our proposed framework, we conducted comprehensive ablation studies. We systematically assess each stage with a combination to get its impact on the overall model performance. The critical stages analyzed include the Custom U-Net, Custom SAM, SAM and Pre-Processed Data. The results from these experiments are measured across four datasets (Kvasir Seg, CVC ClinicDB, CVC ColonDB, and ETIS Larib Polyp DB), are presented in Table 7. Comparative analysis of qualitative results throughout experimental stages shows in Fig. 11.

This experiment's qualitative and quantitative results illustrate the impact of each stage in mask localization and prediction. In the first setup (1. CU + PP), the absence of the custom SAM for precise mask prediction resulted in a low-quality yet somewhat acceptable binary mask. This performance reduction suggests that the lack of SAM leads to inferior mask quality.

To further evaluate the effectiveness of the custom SAM (CS), we tested the performance of standalone SAM with and without pre-processed images (2. CS + PP and 4. CS). SAM requires some form of input to generate a comparable binary mask, so a single pointer input was provided. However, without proper localization guidance, a poor-quality mask was generated. The presence of pre-processed images only marginally improved the results. In this case, two different polyps

Table 8
Performance comparison of different attention mechanisms.

Attention Mechanism Used	mDice on Kvasir SEG	mDice on ClinicDB	mDIce on ColonDB	mDice on ETIS	Average Training Time (Hours)
Custom Attention (Spatial + Custom Gating Block)	**0.946**	**0.941**	**0.861**	**0.815**	**2.0–2.5**
Conditional Attention	0.900	0.890	0.820	0.760	2.5-3.5
Hierarchical Attention	0.950	0.932	0.852	0.788	5.0–5.5
Channel Attention	0.910	0.899	0.825	0.770	2.0–3.0
Multi-Head Attention	0.870	0.860	0.800	0.730	3.0–4.0
Attention Augmented Convolution	0.922	0.910	0.835	0.780	2.5-3.0
Self-Attention	0.880	0.870	0.810	0.745	1.5-2.0
Dual Attention (Position + Channel)	0.915	0.902	0.828	0.775	3.5-4.0

were predicted and masked because the custom SAM identified a polyp outside the true ROI (Region of Interest), resulting in a false positive in mask prediction. This test indicates that without proper guidance from the custom U- Net, detection can be challenging for the custom SAM.

We also tested the performance without training the mask decoder, leaving all parameters at their default settings in SAM (vit_l checkpoint). In this setup, the custom U-Net was present. However, the absence of HQ-token and a trained mask decoder significantly degraded mask quality (5. CU + SAM). The pre-processed images had a minimal effect in this case (3. CU + SAM + PP). Reintroducing all stages simultaneously provided the most acceptable results in this experiment (6. CU + CS + PP).

We conducted more studies by experimenting with various attention mechanisms to further enhance the stability and generalization ability of SAMU-Net. Our goal was to evaluate the impact of different attention strategies on segmentation performance and training efficiency. Table 8 presents the performance comparison of these attention mechanisms including Dice scores on four widely used polyp segmentation datasets along with the average training time required for each mechanism.

Despite Hierarchical Attention yielding a slightly better Dice score (0.950) on the Kvasir-SEG dataset, it came with a major drawback—an average training time of 5.0–5.5 h, compared to 2.0–2.5 h for our custom attention mechanism. Given that our custom attention also performed robustly across the other datasets, achieving comparable results on CVC-ClinicDB and CVC-ColonDB, we prioritized efficiency for broader applications where computational resources and time are critical factors. Therefore we chose Custom Attention over Hierarchical Attention due to its significantly lower training time without a substantial sacrifice in accuracy. This balance between performance and computational efficiency makes the custom attention mechanism more suitable for practical use cases, where timely and accurate predictions are essential.

6. Discussion

Experimental results show that our proposed dual-stage polyps segmentation architecture, SAMU-Net, has been leading the performance on most benchmark datasets and metrics. This has been successful due to the novel combination of a high degree of customization, attention-based U-Nets, with the Segment Anything Model using the strengths of both methods. Equipped with various attention mechanisms, including squeeze-and-excitation blocks, spatial attention, and attention gating in the Custom U-Net model, it could focus on the most relevant features in the segmentation of polyps. As a result, it can significantly enhance its handling capability with regard to diversified appearances of polyps and challenging imaging conditions. Its good performance on four different datasets is evidence of the robustness and generalizability of SAMU-Net for various polyp types and imaging conditions. Whereas most of the competing models present excellent performance either on some metrics or in some datasets, SAMU-Net generally always performs top-tier across most metrics and all evaluated datasets. Qualitative results show that SAMU-Net can accurately segment polyps of different sizes, shapes, and appearances, even in the presence of poor illumination or complex backgrounds. That means obvious advantages in dealing with more challenging and diversified polyp cases for SAMU-Net are critical to real-world clinical applications. However, it is still clear that there is great space for improvement in model stability and generalization from the Difficulties Encountered in Polyp Segmentation.

7. Conclusion

SAMU-Net is a substantial improvement in automatic polyp segmentation from images of colonoscopy. Our approach integrates an attention-based U-Net model, specifically designed with the Segment Anything Model, to achieve state-of-the-art performance across multiple datasets and metrics. It works very well on generalization, stability, and dealing with hard cases with diverse polyps in size, shape, and appearance. The superior performance by SAMU-Net, especially on datasets never seen before, offers a much greater potential for application in real-world clinical scenarios of computer-aided detection and diagnosis of CRC. With the precise and reliable polyp segmentation provided by SAMU-Net, it is conjectured that clinicians could enhance the effectiveness of screening against CRC and promote its early detection. Future work might focus on further enhancement of the stability of the model across different datasets using other attention mechanisms or architectural modifications and clinical validation studies for investigating in more depth the potential impact on polyp detection rate and diagnosis accuracy in real clinical applications of colonoscopy. In conclusion, SAMU-Net represents a promising step forward in automated polyp segmentation, potentially significantly enhancing the early detection and diagnosis of colorectal cancer.

CRediT authorship contribution statement

Radiful Islam: Writing – original draft, Visualization, Validation, Resources, Methodology, Formal analysis, Conceptualization. **Rashik Shahriar Akash:** Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis. **Md Awlad Hossen Rony:** Writing – review & editing, Validation, Resources, Methodology, Formal analysis, Conceptualization. **Md Zahid Hasan:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

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

Data availability

This study uses four publicly available datasets: Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib Polyp DB.

References

[1] World Health Organization. Colorectal cancer. 2023. https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer. Accessed (11.07.24).

[2] Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Translational Oncology Oct. 2021;14(10):101174. https://doi.org/10.1016/j.tranon.2021.101174.

[3] Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut Jan. 2016;66(4):683–91. https://doi.org/10.1136/gutjnl-2015-310912.

[4] Nachmani R, Nidal I, Robinson D, Yassin M, Abookasis D. Segmentation of polyps based on pyramid vision transformers and residual block for real-time endoscopy imaging. J Pathol Inf Jan. 2023;14:100197. https://doi.org/10.1016/j.jpi.2023.100197.

[5] Korbar B, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inf Jan. 2017;8(1):30. https://doi.org/10.4103/jpi.jpi_34_17.

[6] Dong G, Basu A. Medical image denosing via explainable AI feature preserving loss. arXiv (Cornell University); Jan. 2023. https://doi.org/10.48550/arxiv.2310.20101.

[7] Tomar NK, et al. DDANet: dual decoder attention network for automatic polyp segmentation. In: Lecture notes in computer science; 2021. p. 307–14. https://doi.org/10.1007/978-3-030-68793-9_23.

[8] Wen Y, Zhang L, Meng X, Ye X. Rethinking the transfer learning for FCN based polyp segmentation in colonoscopy. IEEE Access Jan. 2023;11:16183–93. https://doi.org/10.1109/access.2023.3245519.

[9] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell Dec. 2017;39(12):2481–95. https://doi.org/10.1109/tpami.2016.2644615.

[10] Dilmaghani S, Coelho-Prabhu N. Role of artificial intelligence in colonoscopy: a literature review of the past, present, and future directions. Techniques and Innovations in Gastrointestinal Endoscopy Jan. 2023;25(4):399–412. https://doi.org/10.1016/j.tige.2023.03.002.

[11] Wang D, et al. AFP-mask: anchor-free polyp instance segmentation in colonoscopy. IEEE Journal of Biomedical and Health Informatics Jul. 2022;26(7):2995–3006. https://doi.org/10.1109/jbhi.2022.3147686.

[12] Duc NT, Oanh NT, Thuy NT, Triet TM, Dinh VS. ColonFormer: an efficient transformer based method for colon polyp segmentation. IEEE Access Jan. 2022;10:80575–86. https://doi.org/10.1109/access.2022.3195241.

[13] Wang K, Liu L, Fu X, Liu L, Peng W. RA-DENet: reverse attention and distractions elimination network for polyp segmentation. Comput Biol Med Mar. 2023;155:106704. https://doi.org/10.1016/j.compbiomed.2023.106704.

[14] Lu L, Chen S, Tang H, Zhang X, Hu X. A multi-scale perceptual polyp segmentation network based on boundary guidance. Image Vis Comput Oct. 2023;138:104811. https://doi.org/10.1016/j.imavis.2023.104811.

[15] Huang C-H, Wu H-Y, Lin Y-L. HarDNet-MSEG: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS. arXiv (Cornell University); Jan. 2021. https://doi.org/10.48550/arxiv.2101.07172.

[16] Gu Q, Meroueh C, Levernier J, Kroneman T, Flotte T, Hart S. Using an anomaly detection approach for the segmentation of colorectal cancer tumors in whole slide images. J Pathol Inf Jan. 2023;14:100336. https://doi.org/10.1016/j.jpi.2023.100336.

[17] Nogueira-Rodríguez A, et al. Deep Neural Networks approaches for detecting and classifying colorectal polyps. Neurocomputing Jan. 2021;423:721–34. https://doi.org/10.1016/j.neucom.2020.02.123.

[18] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Lecture notes in computer science; 2015. p. 234–41. https://doi.org/10.1007/978-3-319-24574-4_28.

[19] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell Apr. 2018;40(4):834–48. https://doi.org/10.1109/tpami.2017.2699184.

[20] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: a review of theory and applications. IEEE Access 2021;9:82031–57.

[21] Yue G, Li S, Cong R, Zhou T, Lei B, Wang T. Attention-Guided pyramid context network for polyp segmentation in colonoscopy images. IEEE Trans Instrum Meas Jan. 2023;72:1–13. https://doi.org/10.1109/tim.2023.3244219.

[22] Kirillov A, et al. Segment Anything Oct. 2023. https://doi.org/10.1109/iccv51070.2023.00371.

[23] Ahamed Md F, Islam Md R, Nahiduzzaman Md, Chowdhury MEH, Alqahtani A, Murugappan M. Automated colorectal polyps detection from endoscopic images using MultiResUNet framework with attention guided segmentation. Human-centric Intelligent Systems Apr. 2024;4(2):299–315.

[24] Li Y, Hu M, Yang X. Polyp-SAM: transfer SAM for polyp segmentation Apr. 2024. https://doi.org/10.1117/12.3006809.

[25] Dong B, Wang W, Fan D-P, Li J, Fu H, Shao L. Polyp-PVT: polyp segmentation with pyramid vision transformers. CAAI Artificial Intelligence Research Dec. 2023:9150015. https://doi.org/10.26599/air.2023.9150015.

[26] Yin Z, Liang K, Ma Z, Guo J. Duplex contextual relation network for polyp segmentation. In: 2022 IEEE 19th international symposium on biomedical imaging (ISBI); Mar. 2022. https://doi.org/10.1109/isbi52829.2022.9761402.

[27] Zhao X, Zhang L, Lu H. Automatic polyp segmentation via multi-scale subtraction network. In: Lecture notes in computer science; 2021. p. 120–30. https://doi.org/10.1007/978-3-030-87193-2_12.

[28] Patel K, Bur AM, Wang G. Enhanced U-Net: A Feature Enhancement Network for Polyp Segmentation May 2021. https://doi.org/10.1109/crv52889.2021.00032.

[29] Fan D-P, et al. PraNet: parallel reverse attention network for polyp segmentation. In: Lecture notes in computer science; 2020. p. 263–73. https://doi.org/10.1007/978-3-030-59725-2_26.

[30] Wang H, et al. Unveiling camouflaged and partially occluded colorectal polyps: introducing CPSNet for accurate colon polyp segmentation. Comput Biol Med Mar. 2024;171:108186. https://doi.org/10.1016/j.compbiomed.2024.108186.

[31] Shao H, Zhang Y, Hou Q. Polyper: boundary sensitive polyp segmentation. Proc AAAI Conf Artif Intell Mar. 2024;38(5):4731–9. https://doi.org/10.1609/aaai.v38i5.28274.

[32] Jain S, et al. CoInNet: a convolution-involution network with a novel statistical attention for automatic polyp segmentation. IEEE Trans Med Imag Dec. 2023;42(12):3987–4000. https://doi.org/10.1109/tmi.2023.3320151.

[33] Ji Z, et al. LightCF-net: a lightweight long-range context fusion network for real-time polyp segmentation. Bioengineering May 2024;11(6):545. https://doi.org/10.3390/bioengineering11060545.

[34] Liu J, Zhang W, Liu Y, Zhang Q. Polyp segmentation based on implicit edge-guided cross-layer fusion networks. Sci Rep May 2024;14(1). https://doi.org/10.1038/s41598-024-62331-5.

[35] Ren G, Lazarou M, Yuan J, Stathaki T. Towards Automated Polyp Segmentation Using Weakly- and Semi-Supervised Learning and Deformable Transformers Jun. 2023. https://doi.org/10.1109/cvprw59228.2023.00458.

[36] Ahamed Md F, et al. IRv2-Net: a deep learning framework for enhanced polyp segmentation performance integrating InceptionResNetV2 and UNet architecture with test time augmentation techniques. Sensors Sep. 2023;23(18):7724. https://doi.org/10.3390/s23187724.

[37] Jha D, et al. Kvasir-SEG: a segmented polyp dataset. In: Lecture notes in computer science; 2019. p. 451–62. https://doi.org/10.1007/978-3-030-37734-2_37.

[38] Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imag Graph Jul. 2015;43:99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007.

[39] Tajbakhsh N, Gurudu SR, Liang J. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans Med Imag Feb. 2016;35(2):630–44. https://doi.org/10.1109/tmi.2015.2487997.

[40] Silva JS, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg Sep. 2013;9(2):283–93. https://doi.org/10.1007/s11548-013-0926-3.

[41] keras-ocr — keras_ocr documentation. 2019. https://keras-ocr.readthedocs.io/en/latest/ (Accessed 12.04.24).

[42] Bertalmio M, Bertozzi AL, Sapiro G. Navier-stokes, fluid dynamics, and image and video inpainting Aug. 2005. https://doi.org/10.1109/cvpr.2001.990497.

[43] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End object detection with transformers. arXiv (Cornell University); Jan. 2020. https://doi.org/10.48550/arxiv.2005.12872.

[44] Ke L, et al. Segment anything in high quality. arXiv (Cornell University; Jan. 2023. https://doi.org/10.48550/arxiv.2306.01567.

[45] Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: a nested U-net architecture for medical image segmentation. In: Lecture notes in computer science; 2018. p. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1.

[46] Wang J, et al. XBound-former: toward cross-scale boundary modeling in transformers. IEEE Trans Med Imag Jun. 2023;42(6):1735–45. https://doi.org/10.1109/tmi.2023.3236037.

[47] Fang Y, Chen C, Yuan Y, Tong K-Y. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Lecture notes in computer science; 2019. p. 302–10. https://doi.org/10.1007/978-3-030-32239-7_34.

[48] Zhang R, Li G, Li Z, Cui S, Qian D, Yu Y. Adaptive context selection for polyp segmentation. In: Lecture notes in computer science; 2020. p. 253–62. https://doi.org/10.1007/978-3-030-59725-2_25.

[49] Shi J-H, Zhang Q, Tang Y-H, Zhang Z-Q. Polyp-mixer: an efficient context-aware MLP-based paradigm for polyp segmentation. IEEE Trans Circ Syst Video Technol Jan. 2023;33(1):30–42. https://doi.org/10.1109/tcsvt.2022.3197643.

[50] Zhao X, et al. M2SNet: multi-scale in multi-scale subtraction network for medical image segmentation. arXiv (Cornell University); Jan. 2023. https://doi.org/10.48550/arxiv.2303.10894.

[51] Hu K, Chen W, Sun Y, Hu X, Zhou Q, Zheng Z. PPNet: pyramid pooling based network for polyp segmentation. Comput Biol Med Jun. 2023;160:107028. https://doi.org/10.1016/j.compbiomed.2023.107028.

[52] Jain S, et al. CoInNet: a convolution-involution network with a novel statistical attention for automatic polyp segmentation. IEEE Trans Med Imag Dec. 2023;42(12):3987–4000. https://doi.org/10.1109/tmi.2023.3320151.