IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 June 2023, accepted 29 June 2023, date of publication 3 July 2023, date of current version 12 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3291783

RESEARCH ARTICLE

Improved Colorectal Polyp Segmentation Using Enhanced MA-NET and Modified Mix-ViT Transformer

KHALED ELKARAZLE[1,3], **VALLIAPPAN RAMAN**[2], **PATRICK THEN**[1], **AND CASLON CHUA**[3]

[1]Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Kuching, Sarawak 93350, Malaysia
[2]Department of Artificial Intelligence and Data Science, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu 641014, India
[3]Faculty of Science, Engineering and Technology, Swinburne University of Technology, Melbourne, VIC 3122, Australia

Corresponding author: Khaled Elkarazle (kelkaeazle@swinburne.edu.my)

This work was supported in part by the Swinburne University of Technology Sarawak Higher Degree by Research (HDR) Support Fund.

ABSTRACT Colorectal polyps is a prevalent medical condition that could lead to colorectal cancer, a leading cause of cancer-related mortality globally, if left undiagnosed. Colonoscopy remains the gold standard for detection and diagnosis of colorectal neoplasia; however, a significant proportion of neoplastic lesions are missed during routine examinations, particularly diminutive and flat lesions. Deep learning techniques have been employed to improve polyp detection rates in colonoscopy images and have proven successful in reducing the miss rate. However, accurate segmentation of small and flat polyps remains a major challenge to existing models as they struggle to differentiate polypoid and non-polypoid regions apart. To address this issue, we present an enhanced version of the Multi-Scale Attention Network (MA-NET) that incorporates a modified Mix-ViT transformer as the feature extractor. The modified Mix-ViT facilitates ultra-fine-grained visual categorization to improve the segmentation accuracy of polypoid and non-polypoid regions. Additionally, we introduce a pre-processing layer that performs histogram equalization on input images in the CIEL*A*B* color space to enhance their features. Our model was trained on a combined dataset comprising Kvasir-SEG and CVC-ClinicDB and cross-validated on CVC-ColonDB and ETIS-LaribDB. The proposed method demonstrates superior performance compared to existing methods, particularly in the detection of small and flat polyps.

INDEX TERMS Colorectal polyps, colorectal polyps detection, colorectal polyps segmentation, color space, colonoscopy images.

I. INTRODUCTION

Colorectal polyps are neoplastic growths that arise from uncontrolled cellular proliferation in the colon. Current research suggests that the development of these polyps is multifactorial, with lifestyle, dietary habits, and genetic predisposition being among the contributing factors [1], [2], [3].

Estimating the global prevalence of colorectal polyps is challenging because of the low adherence to screening guidelines and the asymptomatic nature of most polyps.

Moreover, the heterogeneity and scarcity of epidemiological data across different regions hamper the comparability and generalizability of the existing studies [1], [2]. While the majority of colorectal polyps are small and asymptomatic, if left undetected and untreated, they may progress to colorectal cancer (CRC), a leading cause of cancer-related mortality, responsible for approximately 1 million deaths annually, with projections indicating a 56% increase in incidence by 2040 as predicted by the International Agency for Research on Cancer (IARC) [4].

Colorectal polyps are precancerous lesions that can transform into colorectal cancer through a series of genetic and epigenetic events. These events affect the growth and differentiation of the colonic epithelial cells, resulting in abnormal proliferation and survival. The risk of colorectal cancer is influenced by the type, size, number, and location of the

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

polyps, as well as the degree of dysplasia or carcinoma in situ. Adenomatous polyps and serrated polyps are the most common types of polyps that can lead to cancer [2], [3], [4].

To date, the most common method to detect and diagnose colorectal polyps has been a multi-step procedure known as colonoscopy. Colonoscopy is a medical procedure that allows for the visual examination of the large intestine (colon) and rectum. During the procedure, a long, flexible tube called a colonoscope is inserted into the rectum. The colonoscope is equipped with a tiny video camera at its tip, which allows for the transmission of images to a monitor for viewing by the physician. If necessary, polyps or other abnormal tissue can be removed through the scope during the procedure. Additionally, tissue samples (biopsies) can be taken for further analysis. Prior to the procedure, patients must undergo bowel preparation to ensure that the colon is free of any residue that may obscure the view. This typically involves following a special diet and taking a bowel-cleansing agent [5], [6].

Colonoscopy is the most reliable method for finding colorectal neoplasia as it allows direct visualization and biopsy of the entire colon and rectum, and it could detect both polypoid and flat lesions that may be missed by other screening modalities. A colonoscope can also remove polyps during the procedure, which can prevent their progression to cancer. Compared with other screening tests, such as fecal occult blood testing, sigmoidoscopy, or computed tomographic colonography, colonoscopy has higher sensitivity and specificity for detecting colorectal neoplasia and can reduce the incidence and mortality of colorectal cancer [7], [8], [9].

Despite being the standard procedure for the diagnosis of colorectal polyps, colonoscopy has an estimated miss rate for lesions ranging from 6% to 28%. This miss rate can be attributed to various factors such as poor bowel preparation, inadequate visualization of certain areas of the colon, and the presence of flat or small polyps [10], [11], [12], [13], [14]. In addition, external factors such as poor training of endoscopists and fatigue due to long working hours may also affect detection accuracy [15], [16].

Some of the restrictions on finding malignant tumors during regular colonoscopy sessions include the availability and accessibility of the procedure, which may vary depending on the health care system and resources, the patient's tolerance and adherence to the bowel preparation process, which may affect the quality and safety of the examination, the endoscopist's skill and experience, which may influence the detection and removal of lesions, and finally the possibility of missing flat or serrated lesions that are more difficult to detect and remove, especially in the proximal colon (the right side of the colon) [17].

As the risk of developing colorectal cancer increases with the number of missed polyps, there has been a growing interest among deep learning researchers in developing colorectal polyp detection models. These models aim to assist endoscopists in identifying and localizing lesions of all sizes and shapes, with the primary objective of improving the detection rate of small polyps that may be overlooked during a colonoscopy session. By incorporating advanced image analysis techniques and machine learning algorithms, these models have the potential to significantly enhance the accuracy and efficiency of colorectal cancer screening [18], [19], [20], [21].

Although these methods may vary in their design and implementation, their primary objective is to accurately localize polyps within colonoscopic images. The process of developing a colorectal polyp detection model typically begins with the collection of a labeled dataset of colonoscopic images. These images then undergo pre-processing steps such as resizing and normalization, as well as data augmentation techniques to enhance the robustness of the model.

Once pre-processed, the data is fed into a deep learning model that has been trained to detect polyps from either static images or real-time video feeds. The format of the model's output is determined by its specific objective; for example, semantic segmentation models generate binary masks indicating the location of polyps, while object detection models produce bounding boxes around detected lesions [15], [16], [17].

In the context of semantic segmentation, the primary objective is to accurately classify each pixel in a given image as either belonging to a polyp or to the background. To train a colorectal segmentation model, a ground truth binary mask indicating the location of the polyp is used as the training label. The model's objective is to produce a predicted mask that accurately represents the estimated location of the polyp within the image. By comparing the predicted mask to the ground truth mask, the model's performance can be evaluated and refined through further training. Some of the most well-known semantic segmentation networks include U-NET [22], U-NET++ [23], DeepLabV3 [24] and Fully Convolutional Networks (FCN) [25].

Although deep learning methods have shown promise in improving the detection rate of polyps in colonoscopic images, their performance on real-life, unedited images is far from perfect due to several challenges. These challenges include the failure to detect small and flat polyps, the variability in the shape and color of polyps, and the presence of external obstructions due to poor bowel preparation [15], [16], [17]. To address these issues, researchers have focused on developing more complex segmentation models with multiple encoders and feature extraction layers to capture as many meaningful features as possible. While these methods have been effective in detecting more obvious polyps, small polyps still pose a significant challenge, with many models failing to accurately detect them. In figure 1, we present several examples of these challenging lesions.

In this work, we address the challenges associated with the detection of small and flat polyps in colonoscopic images by introducing an enhanced version of the multi-scale attention network (MA-NET) [28]. Our enhanced MA-NET design incorporates a modified Mix-ViT transformer [29] as a replacement for the original convolution-based encoder. We choose MA-NET as our preferred decoder due to its

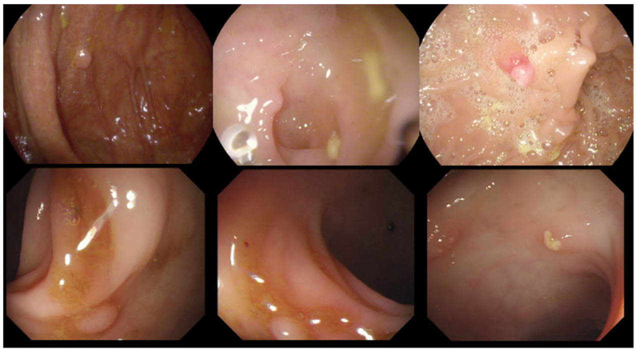


FIGURE 1. Several examples of challenging polyps. Samzples are taken from the ETIS-LaribPolypDB and CVC-ColonDB.

ability to effectively capture and integrate contextual information at multiple scales, leading to improved performance in tasks such as image segmentation, especially on medical images.

For the encoder component of our model, we tweak the design of a pre-trained Mix-ViT by removing the mix token prediction head and replacing the classification head with a global average pooling layer to fit the nature of our problem. We choose the Mix-ViT due to its demonstrated ability to capture ultra-fine-grained features for visual categorization.

We hypothesize that this capability is a crucial yet often overlooked component in the detection of small and flat polyps in colonoscopic images. By incorporating the Mix-ViT architecture into our encoder design, we aim to enhance our model's ability to accurately detect and localize these challenging polyps.

To further enhance the visibility of features such as blood vessels and other non-polypoid features in colonoscopy images, we introduce an additional pre-processing layer that applies Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L component of the images in the CIEL*A*B* color space. This technique enhances the contrast of the images by redistributing the lightness values, resulting in enhanced visibility of fine details and improved image quality.

Our segmentation model is trained and validated on a combined dataset of two publicly available datasets: Kvasir-SEG [30] and CVC-ClinicDB [31]. To evaluate the generalizability and performance of our model on independent samples, we cross-validate our model on two public datasets: ETIS-LaribPolypDB [32] and CVC-ColonDB [33]. To ensure a fair test, none of the samples from the cross-validation datasets were included during training.

We recorded several metrics such as Intersection over Union, Precision, Recall and F1 scores during the cross-validation phase to better understand the performance. We also visualized the predicted mask of our model as another indicator of its accuracy on unseen samples.

Our method outperforms existing state-of-the-art semantic segmentation methods, particularly in the segmentation of small and flat polyps. The main contributions of this study are:

- We present an improved version of the multi-scale attention network (MA-NET) that uses a modified vision transformer, namely Mix-ViT, as the backbone instead of skip connections and convolutional neural networks.
- We modify a pre-trained Mix-ViT transformer by replacing its classification head with a global pooling layer and removing the mix token predictor after pre-training.
- We apply Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L* component of the CIEL* A* B* color space to enhance the visibility of features such as blood vessels, non-polypoid objects and polyps in colonoscopy images.
- We train and validate our model on samples of two publicly available datasets: Kvasir-SEG and CVC-ClinicDB. Moreover, we cross-validate our model on two distinctive datasets: ETIS-LaribPolypDB and CVC-ColonDB to confirm its generalizability and performance on independent samples.

In conclusion, this study aims to address the challenge of detecting small and flat polyps, which, to our knowledge, has been a significant factor in the suboptimal performance of existing colorectal polyp segmentation methodologies. The remainder of this paper is divided into the following sections: Related Works, Proposed Method, Experimental Results, Discussion and Conclusion.

II. RELATED WORKS

The authors in [34] introduced Y-Net, a polyp detection method for colonoscopy images inspired by U-Net. Y-Net comprises two encoders and a single decoder and can be trained on a limited number of samples. Both encoders follow the VGG19 design [35], while the decoder is a custom-built CNN with five deconvolutional blocks and one final convolution block. The first encoder is initialized with ImageNet weights and the second with the Xavier normal initializer. Both use the SELU activation function instead of ReLU. Y-Net was trained and tested on the ASU-Mayo dataset [36] without cross-validation. The authors reported a precision of 87.4%, recall of 84.4%, and F1 score of 85.9%. However, this method did not perform well with reflections, polyp-shaped objects, and flat lesions.

Another model proposed in [37] introduces a polyp segmentation network using a combination of 2D and 3D convolutional layers. The 2D layers extract spatial representation while the 3D layers add a temporal dimension. Initially, features are extracted using the 2D network before the 3D network generates a temporally coherent segmentation mask. An upsampling layer is then added to upscale the predicted mask. The method was trained on a private dataset and tested on both the SUN dataset and Kvasir-SEG, achieving scores of 86.14 for sensitivity, 85.32 for specificity, 93.45 for precision, and 89.65 for F1.

In [38], the authors introduced a custom model for segmenting colorectal polyps by combining a SWIN transformer [39] with EfficientNet [40]. The model includes a multi-dilation convolutional block to refine local features extracted by EfficientNet and global features obtained by the SWIN transformer. These features are then aggregated using a multi-feature aggregation block before constructing a predicted mask using an attentive block. The model was trained on the Kvasir-SEG and CVC-ClinicDB datasets and tested on the CVC-ColonDB, ETIS-Larib, and Endoscene datasets. During evaluation, the authors reported a mean dice coefficient of 0.906, an IoU of 0.842, a mean weighted F-measure of 0.88, and a mean absolute error of 0.001.

In [41], an automatic polyp segmentation model was proposed using the SegNet architecture [42] to segment colonoscopy images. The samples were preprocessed by thresholding red, green, and blue pixels to values between 15 and 50 to filter out non-polyp regions and ease the training process. The preprocessed images were then fed to a SegNet model for segmentation. The CVC-Clinic, CVC-Colon, and ETIS-Larib datasets were used for training and testing, with the study reporting an average IoU of 81.7%.

In [43], a method was presented for detecting polyps from colonoscopy images using two pre-trained models, VGG16 and MobileNet. The authors preprocessed the input images by removing black regions, normalizing RGB values to match the RGB mean of samples in the ImageNet dataset, and resizing all images to a constant size of 224 × 224 pixels. The images were then processed using a multi-resolution sliding window to locate polyps before cropping the region and feeding it to a probability prediction function. The Kvasir-SEG dataset was used for training and the CVC-ClinicDB and ETIS-Larib datasets for cross-validation. On the CVC-ClinicDB dataset, the model achieved precision, recall, and F1 scores of 91.9, 89.0, and 0.90 respectively. On the ETIS-Larib dataset, the model achieved precision of 87.0, recall of 91.0, and an F1 score of 89.0.

In [44], a saliency detection network was introduced to detect polyps from static polyp images. The authors used Neutrosophic theory to decrease the effect of white light reflections caused by colonoscopy light and introduced an image-suppressing technique using a single-value Neutrosophic set (SVNS) to rebuild colonoscopy images without white light reflection. Specular regions were recovered using a dynamic window that searched for non-specular pixels near each specular pixel, using an 8 × 8 window rotated counter-clockwise until all specular regions were recovered. The RGB pixels' average value was used to paint specular pixels in the recovered image. The authors introduced a saliency network known as NeutSS-PLS, inspired by U-Net and DSS, for detection and segmentation. The network had two-level short connections on both sides of the VGG and was trained on the EndoScene and Kvasir-SEG datasets, achieving precision and F1 scores of 92.30 and 92.40 respectively. However, the proposed method struggled to identify polyps near the boundary of colonoscopy images.

In [45], an automatic polyp detection and segmentation system called shuffle-efficient channel attention network (sECA-NET) was introduced to segment colonoscopy images and detect polyps. A CNN was applied to extract the feature map from an input image and a region proposal network (RPN) was developed to predict bounding boxes around polyps in the feature map. A region of interest align (RoiAlign) was applied to extract features from the feature map based on the bounding box of each detected object. Two parallel branches then computed the extracted features for every ROI. In the first branch, features were computed by fully connected layers followed by softmax activation before performing bounding box regression. The second branch concerned mask segmentation, predicting the category of each pixel in the region of interest. The proposed method was trained on the CVC-ClinicDB, ETIS-Larib, and Kvasir-SEG datasets and evaluated on a private cross-validation dataset. The authors reported precision, recall, and F1 scores of 94.9%, 96.9%, and 95.9% respectively.

III. PROPOSED METHOD

A. OVERVIEW

Our proposed method comprises four stages. Initially, we collect training and testing samples from publicly available datasets. We combine Kvasir-SEG and CVC-ClinicDB to create a diverse set of training samples. These datasets provide a range of images that can be used to train our model effectively. We then use ETIS-LaribPolypDB and CVC-ColonDB to test our model. These two distinctive datasets ensure fair testing and allow us to verify the performance of our method on unseen samples.

Prior to training, we pre-process our images by resizing all the samples to a constant dimension of 256 × 256 pixels. This ensures that all images have the same dimensions and can be processed by our model effectively. We then normalize our images so that the pixel values are between 0 and 1. This step is important as it ensures that the pixel values are within a consistent range and can be processed effectively by our model.

Next, we pass the samples to a custom pre-processing layer in which we first convert the images to CIEL*A*B* color space. This color space separates the color information from the luminance information, allowing us to apply a Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L* component of the image. CLAHE is an effective method for enhancing contrast in images and can improve the performance of our model.

Once CLAHE has been applied, we convert the images back to its original RGB format then we pass them to our segmentation model for training. We present the various stages of the proposed method in figure 2.

Our segmentation model is a modified MA-NET model in which the original encoder was replaced with a modified Mix-ViT transformer. The Mix-ViT transformer was initially pre-trained on the ImageNet dataset for image classification. We further fine-tune the transformer's design by removing the mix token predictor and multilayer perceptron layer, which were deemed unnecessary for this study. We selected the Mix-ViT transformer as our encoder to leverage its capability

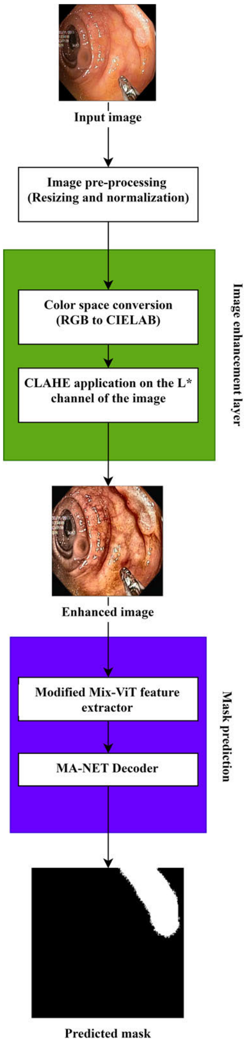


FIGURE 2. An overview of the proposed method. The method consists of three main steps: 1) preprocessing and normalization; 2) image enhancement; 3) feature extraction using the enhanced MA-NET model.

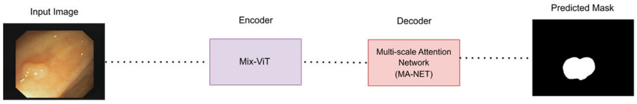


FIGURE 3. An overview of the enhanced MA-NET network. The original convolution-based encoder was replaced by a modified Mix-ViT vision transformer.

to effectively capture ultra-fine-grained features. After training, we cross-validate our model on the testing datasets and observe its performance. Cross-validation allows us to assess how well our model generalizes to new, unseen samples. Figure 3 illustrates a high-level overview of the proposed segmentation model.

B. DATA COLLECTION

In this study, we utilize four public datasets: two for training and validation (Kvasir-SEG and CVC-ClinicDB) and two for testing (ETIS-LaribPolypDB and CVC-ColonDB). The use of public datasets mitigates potential ethical concerns associated with using patient-collected samples. During training we divide our training dataset into 80% training and 20% validation.

TABLE 1. A summary of the training dataset.

Set	Number of Samples (Percentage)
Training	1289 (80%)
Validation	323 (20%)
Total	1612

The Kvasir-SEG dataset comprises 1000 colonoscopy images and ground-truth masks, with polyps of varying shapes, colors, and sizes. Each sample has been manually annotated and verified by experienced gastroenterologists. The resolution of samples ranges from 332 × 487 to 1920 × 1072 pixels. All 1000 samples are used for training and validation.

The CVC-ClinicDB dataset consists of 612 polyp static frames extracted from 31 colonoscopy sequences. All images have a fixed resolution of 384 × 288 pixels. The dataset has been widely used to test and validate various colorectal polyps' segmentation methods Table 1 summarizes the breakdown of our training dataset.

C. PRE-PROCESSING

Pre-processing is a crucial preliminary step that must be performed before training any deep learning models to facilitate the training process [46], [47]. In our study, we combine standard pre-processing steps, such as resizing and normalization, with our novel image enhancement layer. We first resize our images to a constant dimension of 256 × 256 pixels.

We experimented with various image dimensions and found that 256 × 256 pixels yielded the highest values of intersection over union (IoU) and Dice coefficient. The ablation study section presents the analysis of how the image dimensions affect the performance of our model.

Next, we normalize the pixels of our images so that every pixel value is within the range of 0 and 1. For normalization we use Equation 1 to normalize our samples.

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

In this equation x represents the pixel value of an image and x' is the normalized image. We use ImageNet's [48] standard deviation values of [0.485, 0.456, 0.406] and mean values of [0.229, 0.224, 0.225], respectively.

D. IMAGE ENHANCEMENT

After completing the primary pre-processing steps, we convert our resized and normalized images from RGB to CIEL*A*B* color space, which expresses color as three values: L* for perceptual lightness and A* and B* for the four unique colors of human vision. The conversion process consists of several steps. Initially, the RGB values must undergo gamma correction to be linearized, which can be accomplished using Equation 2.

$$I_c = \left(\frac{I_o}{255}\right)^{\frac{1}{\gamma}} \times 255 \tag{2}$$

In this equation, I_c represents the gamma-corrected image, I_o is the original input image, and γ is the gamma factor, which we set to 2.2. The linearized RGB values are then transformed into the XYZ color space via matrix multiplication. The specific matrix (M) used for this transformation depends on the reference white point selected for the conversion. The values of matrix M are derived using Equations 3-5, while Equation 6 is used to obtain the XYZ values.

$$\begin{pmatrix} X_r \\ Y_r \\ Z_r \end{pmatrix} = \begin{pmatrix} \frac{x_r}{y_r} \\ 1 \\ \frac{(1-x_r-y_r)}{y_r} \end{pmatrix} \begin{pmatrix} X_g \\ Y_g \\ Z_g \end{pmatrix} = \begin{pmatrix} \frac{x_g}{y_g} \\ 1 \\ \frac{(1-x_g-y_g)}{y_g} \end{pmatrix} \begin{pmatrix} X_b \\ Y_b \\ Z_b \end{pmatrix} = \begin{pmatrix} \frac{x_b}{y_b} \\ 1 \\ \frac{(1-x_b-y_b)}{y_b} \end{pmatrix} \tag{3}$$

In this equation, $\frac{x_r}{y_r}$, $\frac{x_g}{y_g}$, and $\frac{x_b}{y_b}$ are based on the pre-calculated CIE RGB values described in [49].

$$\begin{pmatrix} \rho_r \\ \rho_g \\ \rho_b \end{pmatrix} = \begin{pmatrix} X_r & X_g & X_b \\ Y_r & Y_g & Y_b \\ Z_r & Z_g & Z_b \end{pmatrix}^{-1} \begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} \tag{4}$$

The matrix $\begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix}$ is derived from the Equal Energy spectrum (E), which consists of tristimulus values that define white pixels in an image [50].

$$(M) = \begin{pmatrix} \rho_r X_r & \rho_g X_g & \rho_b X_b \\ \rho_r Y_r & \rho_g Y_g & \rho_b Y_b \\ \rho_r Z_r & \rho_g Z_g & \rho_b Z_b \end{pmatrix} \tag{5}$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = (M) \begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} \tag{6}$$

Once the XYZ values are obtained, they can be converted into CIELAB values using a non-linear transformation. This involves calculating the L* (lightness) component as a function of the Y value and the reference white point, while the a* (green-red) and b* (blue-yellow) components are calculated as functions of all three XYZ values and the reference white point. The conversion from XYZ to L*A*B* is defined in Equation 7.

$$\begin{aligned} L^* &= 116 f\left(\frac{Y}{Y_n}\right) - 16 \\ a^* &= 500\left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \\ b^* &= 200\left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right) \end{aligned} \tag{7}$$

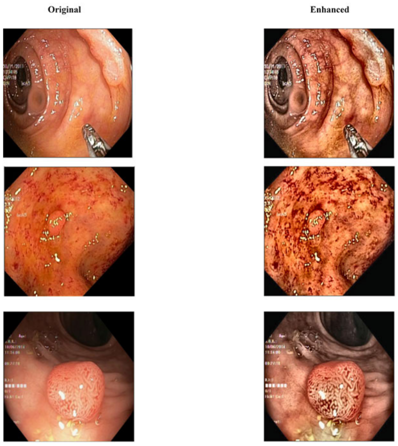


FIGURE 4. Random samples before and after enhancement.

In this equation X_n, Y_n, and Z_n represent a specific white achromatic reference illuminant. We use the standard illuminant D65 values defined in [51]. The resulting CIEL*A*B* values represent colors in terms of perceptual attributes such as lightness and chromaticity. Once the images have been converted, we apply CLAHE on the L* component.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is a variant of Adaptive histogram equalization (AHE) that prevents contrast over-amplification. CLAHE operates on small regions in the image, called tiles, rather than the entire image. The neighboring tiles are then combined using bilinear interpolation to remove the artificial boundaries. CLAHE can be applied to color images, often to the luminance channel. The results of equalizing only the luminance channel of an image outperform equalizing all channels of an RGB image [52], [53], [54]. This is because the CIEL*A*B* color space is designed to approximate human vision. The L* component closely matches human perception of lightness. In the context of colonoscopy images, applying CLAHE on the L component of a CIEL*A*B* image can enhance the contrast and improve the visibility of polypoid and non-polypoid features in an image. In figure 4 we present a comparison of our samples before and after applying CLAHE on the L component of our image.

E. SEGMENTATION MODEL

1) MODIFIED MIX-ViT ENCODER

Convolutional Neural Networks (CNNs) have been the de-facto model for visual data. They use convolution, a "local" operation bounded to a small neighborhood of an image. On the other hand, Vision Transformers (ViT) use self-attention, a "global" operation, since it draws information from the whole image [55]. Recent work has shown that ViT can achieve comparable or even superior performance on image classification tasks [56], [57], [58].

This is because self-attention enables early aggregation of global information and ViT residual connections strongly propagate features from lower to higher layers. In the context

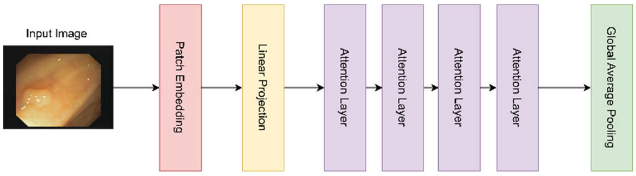


FIGURE 5. An overview of the modified Mix-ViT encoder.

of colorectal polyps semantic segmentation, most polyp segmentation methods use CNNs as their backbone. However, this leads to two key issues when exchanging information between the encoder and decoder: 1) taking into account the differences in contribution between different-level features and 2) designing an effective mechanism for fusing these features. Unlike existing CNN-based methods, some researchers have adopted a transformer encoder, which learns more powerful and robust representations. For instance, in a recent study the authors introduced a model called Polyp-PVT, which utilizes a Pyramid Vision Transformer (PVT) to extract stronger and more robust features for polyp segmentation. Their model was able to effectively reduce noise in the features and greatly enhance its ability to express information [59].

The Mix-ViT transformer was initially developed to address the challenges associated with ultra-fine-grained visual categorization tasks, which involve identifying subcategories within fine-grained objects at a deeper taxonomy level. In our work, we first acquire a Mix-ViT transformer that has already been trained on the ImageNet dataset for classifying 1000 distinct classes.

To adapt the pre-trained transformer to colorectal polyps segmentation, we modify it by replacing its original classification head with a global average pooling layer, which reduces the spatial dimensions of the feature maps while preserving their depth. We also remove the mix token prediction layer, which is not necessary for feature extraction. These modifications allow us to fine-tune the pre-trained transformer for our specific needs while leveraging its ability to extract useful features from images. The modified Mix-ViT encoder is presented in Figure 5.

To create the patches, we represent our input image as $x \in R^{H \times W \times C}$ in which H is the image's height, W is the image's width and C is the number of channels and P represents the patch size. The objective is to create N image patches which can be achieved using Equation 8.

$$N = \frac{HW}{P^2} \tag{8}$$

The input patches undergo a linear transformation via a projection layer, resulting in the generation of patch embedding vectors while preserving their original spatial dimensions. The vectorized patches are then fed to the transformer encoder. Finally, we pass the feature map to a global average pooling layer, prior to feeding the decoder the processed features.

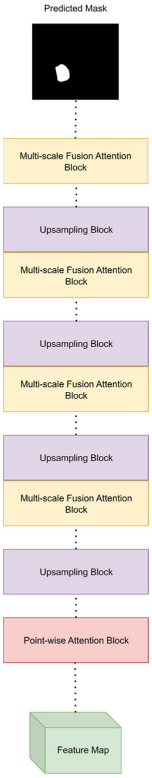

Predicted Mask
Multi-scale Fusion Attention Block
Upsampling Block
Multi-scale Fusion Attention Block
Upsampling Block
Multi-scale Fusion Attention Block
Upsampling Block
Multi-scale Fusion Attention Block
Upsampling Block
Point-wise Attention Block
Feature Map

FIGURE 6. The MA-NET decoder. The default convolution-based downsampling block is replaced with the modified Mix-ViT encoder.

2) ENHANCED MA-NET DECODER

The Multi-scale Attention Network (MA-NET), introduced in [28] has demonstrated exceptional performance in medical image segmentation. This success has inspired us to employ its decoder as an independent module for generating a predicted segmentation mask using the features extracted by the modified Mix-ViT encoder.

MA-NET is designed to capture rich contextual dependencies based on the attention mechanism and has been used for various applications such as liver and tumor segmentation, single image super-resolution, optical flow estimation and correspondence learning.

The original design of MA-NET comprises five key components: Residual blocks, Convolution blocks, Upsampling blocks, Point-wise Attention Block, and Multi-scale Fusion Attention Block. The downsampling module of the network includes several 3 × 3 convolutional layers with a stride of 2 and skip connections between the residual connection blocks and the MFAB blocks. In our study, we replaced this block with our modified Mix-ViT encoder.

The upsampling component of MA-NET is designed with two main blocks: Position-wise Attention Block (PAB) and Multi-scale Fusion Attention Block (MFAB). The PAB models feature interdependencies in spatial dimensions to capture the spatial dependencies between pixels in a global view. In contrast, the MFAB captures channel dependencies between any feature map through multi-scale semantic feature fusion. The fine-tuned decoder is presented in Figure 6.

IV. EXPERIMENTAL RESULTS

A. OVERVIEW

This section presents the methodology and results of our experiments. It includes a description of the testing datasets, in addition sample breakdown, as well as details on the experiment setup and configurations. Additionally, we discuss the evaluation metrics used, and provide a performance analysis and comparative analysis of our results.

B. EXPERIMENT SETUP

In this study, we conducted several experiments to evaluate the performance of our proposed model. To investigate the impact of the proposed image enhancement layer and modified encoder on segmentation accuracy, three additional variants of the proposed method were trained:

- One variant followed to the original design of MA-NET and utilized the proposed image enhancement layer.
- One variant adhered to the original design of MA-NET without the proposed image enhancement layer.
- One variant employed the proposed encoder without the proposed image enhancement layer.

Hyperparameters and configurations such as learning rate, number of epochs, optimizer, loss function, and input size were kept consistent across all the three variants. The proposed method was also compared with existing state-of-the-art segmentation models to examine performance differences in polyp segmentation. Pre-trained models were used in these experiments, with slight modifications made to the model output layer to adapt them to the polyp segmentation task. The modified output layer produces predictions representing two classes: 1) Polyp and 2) Background.

During training, we set the number of epochs to 25 and utilize the Adam optimizer with a learning rate of 0.0001 to optimize our model. Instead of the binary cross-entropy function, we employed the Dice loss as our loss function. All models were trained and tested on a single machine equipped with an RTX4000 GPU and 16GB of RAM. We implemented our model using the PyTorch framework and utilized additional libraries such as NumPy and Sci-Kit learn.

C. TESTING DATASETS

The ETIS-LaribPolypDB dataset, comprises 196 samples with a fixed resolution of 1225 × 996 pixels. Samples are captured in unfiltered settings, with several images being blurry. Due to the limited number of samples, the dataset is primarily used for testing, although some studies have utilized it for training. The CVC-ColonDB dataset consists of 300 colonoscopy images with a constant dimension of 574 × 500 pixels. The dataset is derived from 15 short colonoscopy video frames, with each frame extracted and labeled by professionals. The dataset contains challenging samples with polyps of various shapes and sizes. We use both datasets to cross-validate our model and use them to perform the comparative analysis between our model and existing polyp segmentation methods.

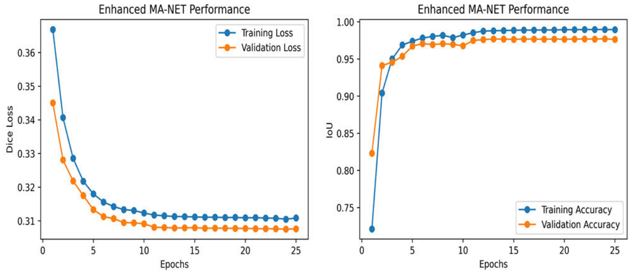


FIGURE 7. Our model's training and validation loss (left) and training and validation IoU scores (right).

D. EVALUATION METRICS

In this section, we present the evaluation metrics used to assess the performance of our deep learning models in the task of semantic segmentation of colorectal polyps. These metrics provide a quantitative measure of the accuracy and reliability of the segmentation masks generated by each model-color space combination. Our analysis considered several commonly used metrics, including precision, recall, F1 score, Dice coefficient, and Intersection-Over-Union (IoU). Precision measures the proportion of true positive predictions among all positive predictions made by the model, while recall measures the proportion of true positive predictions among all actual positive instances in the data.

The F1 score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. The Dice coefficient measures the similarity between two sets, in this case, the predicted segmentation mask and the ground truth mask. The Intersection-Over-Union (IoU) metric measures the overlap between the predicted and ground truth masks as a proportion of their union.

Equations 9 - 13 provide the mathematical formulations for these metrics.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

$$IoU = \frac{TP}{(TP + FP + FN)} \tag{12}$$

$$Dice\ Coefficient = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{13}$$

where TP represents the number of true positives, FP is the number of false positives, FN is the number of false negatives and TN is the number of true negatives.

E. PERFORMANCE ANALYSIS

This subsection presents a performance analysis of our proposed model. Table 2 summarizes the precision, recall, F1 score, IoU, and Dice score obtained by our model when tested on both ETIS-LaribPolypDB and CVC-ColonDB. Additionally, Figures 7 illustrates the losses and accuracies achieved during the training process.

TABLE 2. The obtained scores on both testing datasets.

Metric	Score	Dataset
Precision	0.989	ETIS-LaribPolypDB
Recall	0.996	
F1 score	0.992	
IoU	0.985	
Dice coefficient	0.989	
Precision	0.989	CVC-ColonDB
Recall	0.983	
F1 score	0.985	
IoU	0.973	
Dice coefficient	0.983	

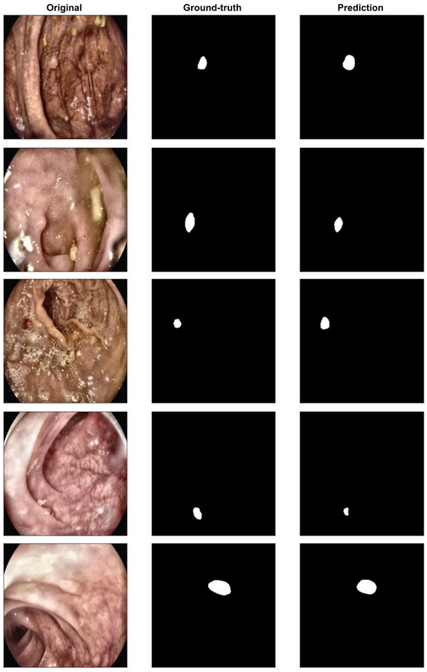


FIGURE 8. Examples of our model predictions on the ETIS-LaribPolypDB.

Figure 8 presents examples of our model's output on the ETIS-LaribPolypDB dataset, while Figure 9 shows several samples of our model's predictions on the CVC-ColonDB dataset. We have chosen to illustrate polyps that are not easily visible to demonstrate the robustness of our method in detecting small polyps.

F. ABLATION STUDY

To further validate the effectiveness of our proposed method in detecting small and flat polyps, we conducted an ablation study. This study involved removing the CLAHE equalization layer while retaining the Mix-ViT encoder, as well as using the original convolution-based encoder of MA-NET with and without the custom CLAHE equalization layer. The results of this ablation study provide insights into the contribution of each component to the overall performance of our model and help us understand how our proposed method is able to effectively capture small and flat polyps. There were no changes to the configuration and hyperparameters of the model while conducting this study. In table 3 we present the scores we obtained during the ablation study on each dataset.

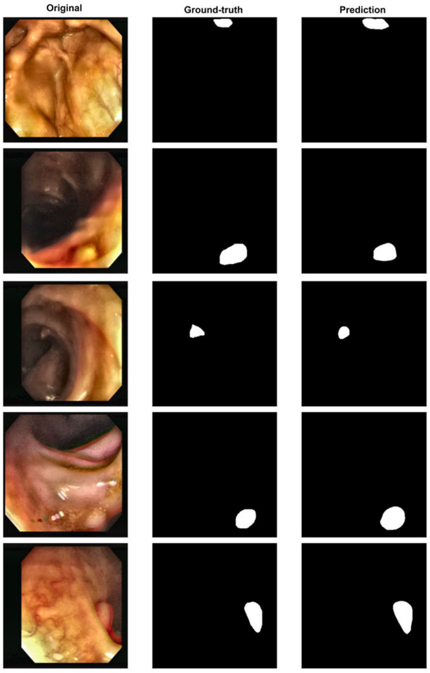


FIGURE 9. Examples of our model predictions on the CVC-ColonDB.

TABLE 3. The scores of the ablation study conducted on our model.

Model	Dataset	IoU	Dice	Precision	Recall	F1
MA-NET	ETIS-DB	0.961	0.974	0.986	0.974	0.979
MA-NET + CLAHE		0.981	0.989	0.990	0.991	0.990
Proposed (No CLAHE)		0.970	0.982	0.978	0.991	0.984
Proposed		**0.981**	**0.989**	**0.991**	**0.996**	**0.990**
MA-NET	CVC-ColonDB	0.952	0.970	0.979	0.972	0.974
MA-NET + CLAHE		0.958	0.974	0.992	0.964	0.977
Proposed (No CLAHE)		0.966	0.979	0.991	0.975	0.982
Proposed		**0.972**	**0.983**	**0.989**	**0.983**	**0.985**

While CLAHE can be applied to images in any color space, we chose to apply it to the L* component of the CIEL*A*B* color space as this approach produced the best image quality in our experiments. To further investigate the relationship between our model's performance and the application of CLAHE to input images in different color spaces, we modified our custom pre-processing layer to equalize images in four color spaces: RGB, HSV, HLS, and CIEL*A*B*. We then trained and tested our model on these equalized samples. Except for the RGB images, the HSV and HLS were converted back to RGB after the correction, similarly to the original implementation. The results of this experiment, presented in Table 4, provide insights into the impact of CLAHE on our model's performance across different color spaces. In addition, we report the results of our method for different image sizes on the validation subset, which consists of 260 images, in Table 5.

TABLE 4. The scores of the ablation study conducted on the image enhancement layer.

Color space	Dataset	IoU	Dice	Precision	Recall	F1
RGB		0.964	0.979	0.980	0.983	0.981
CIEL*A*B*	ETIS-LaribPolypDB	0.985	0.989	0.989	0.996	0.992
HSV		0.944	0.950	0.989	0.920	0.954
HLS		0.957	0.974	0.996	0.961	0.977
RGB		0.956	0.971	0.988	0.967	0.976
CIEL*A*B*	CVC-ColonDB	0.973	0.983	0.989	0.983	0.985
HSV		0.932	0.956	0.989	0.941	0.963
HLS		0.957	0.972	0.988	0.968	0.977

TABLE 5. Comparison of segmentation accuracy for various image dimensions.

Image Size	IoU	Dice	Precision	Recall	F1
128 x 128	0.950	0.967	0.964	0.985	0.973
256 x 256	0.971	0.981	0.987	0.984	0.985
512 x 512	0.954	0.969	0.988	0.965	0.975
1024 x 512	0.892	0.930	0.961	0.928	0.941

TABLE 6. The performance of the proposed method compared to recently proposed models.

Method	IoU	Dice
[60]	0.907	0.949
[61]	0.906	0.948
[62]	0.903	0.947
[63]	0.902	0.946
[64]	0.944	0.899
[37]	0.817	-
[38]	0.842	-
Proposed Method (CVC-ColonDB)	**0.973**	**0.983**
Proposed Method (ETIS-LaribPolypDB)	**0.985**	**0.989**

G. COMPARATIVE ANALYSIS

In this section, we present a comparative analysis of our proposed method and existing state-of-the-art segmentation models. Additionally, we compare the performance of our method with several recently proposed methods to highlight the differences in performance. In table 6 we present the scores of several recently proposed methods compared to ours.

In addition to comparing our method with recently proposed models, we also present a comparative analysis of its performance with four commonly used semantic segmentation models in medical image analysis: UNET, UNET++, DeepLabV3, and Pyramid Scene Parsing Network (PSPNET) [65]. To ensure a fair comparison, we use the same configurations and hyperparameters we used to create our original model. In addition, we utilize pre-trained weights to mitigate the issue of overfitting. The comparison is presented in figure 10. In addition, a quantitative comparison is presented in table 7.

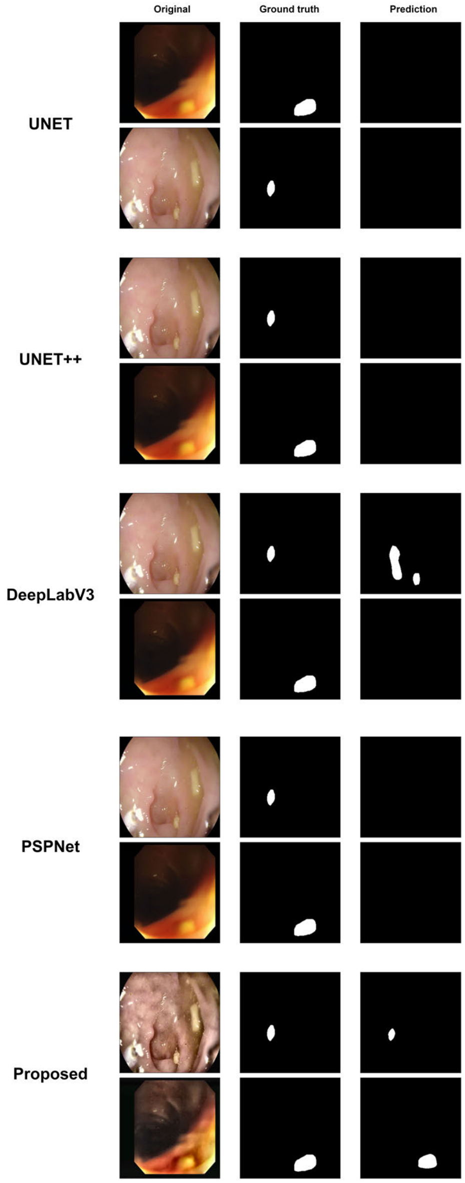


FIGURE 10. Our proposed method performance (bottom row) compared to several well-known semantic segmentation networks.

TABLE 7. Our method compared to several existing semantic segmentation models.

Model	Dataset	IoU	Dice	Precision	Recall	F1
Proposed		**0.985**	**0.989**	**0.989**	**0.996**	**0.992**
UNET		0.975	0.986	0.996	0.978	0.987
UNET++	ETIS-LaribPolypDB	0.976	0.986	0.997	0.978	0.987
DeepLabV3		0.966	0.981	0.979	0.986	0.982
PSPNET		0.964	0.980	0.998	0.965	0.981
Proposed		**0.973**	**0.983**	**0.989**	**0.983**	**0.985**
UNET		0.959	0.975	0.995	0.964	0.978
UNET++	CVC-ColonDB	0.955	0.972	0.997	0.957	0.975
DeepLabV3		0.961	0.976	0.996	0.964	0.979
PSPNET		0.939	0.964	0.999	0.940	0.967

V. DISCUSSION

A. OVERVIEW

In this section, we provide a comprehensive analysis of our experimental results. We compare the performance of our proposed model with other semantic segmentation methods and discuss the factors that may have influenced the outcomes reported in the previous section. This discussion aims to deepen the understanding of our findings and identify potential areas for future research. We evaluate the performance of our enhanced MA-NET model, analyze its strengths and weaknesses, and provide recommendations for future work. In addition to our main findings, we also present a detailed analysis of the results obtained from our ablation study. This analysis provides further justification for the modifications we made to our model and highlights the impact of each individual component on the overall performance.

B. RESULT INTERPRETATION

The results presented in the previous section confirm that our proposed method effectively addresses the challenge of detecting small and flat polyps. We hypothesize that the success of our approach can be attributed to our emphasis on both data and model quality, rather than relying solely on the development of a complex model to detect challenging lesions.

Our analysis indicates that the application of Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L* component of the CIEL*A*B* color space representation of the input image significantly enhances the features of both polypoid and non-polypoid areas. As a result, blood vessels, inner regions of the colon, and polyps become more distinct and easier to identify.

The CIEL*A*B* color space is designed to approximate human vision and is based on the opponent color theory. This theory posits that the human visual system processes color information in terms of opposing pairs: light-dark, red-green, and yellow-blue. The CIEL*A*B* color space separates lightness information from chromatic information, with the L* component representing lightness and the a* and b* components representing the chromatic opponents green-red and blue-yellow, respectively.

We applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L* component of a CIEL*A*B* colonoscopy image to improve its contrast while preserving its chromatic information. CLAHE has helped enhancing local contrast, making small details and textures within the image more apparent without affecting the overall color balance. This contributed significantly to the improved performance of our model.

We observed that applying CLAHE directly to the original RGB colonoscopy image resulted in color distortion since it treats each color channel independently.

The distortions resulted in a significant drop in performance as observed in the results of our ablation study. Moreover, applying CLAHE to the value component of an HSV image or the lightness component of an HSL image can alter the hue and saturation of the image.

In terms of model development, we employed transfer learning to enhance the convergence of our model and expedite the training process. Transfer learning entails utilizing pre-existing knowledge from a related task to augment the performance of a model on a new task. In our study, we initialized the weights of our model using a Mix-ViT model that had been pre-trained on the ImageNet dataset. This approach enabled us to achieve satisfactory performance with a smaller dataset of colonoscopy images and reduced the time required for training. The significance of pre-training is observed in results of the ablation study.

Furthermore, our experiments showed that replacing the original convolution-based encoder of MA-NET with the feature extraction layers of Mix-ViT improved the detection rate of small polyps. By incorporating the feature extraction layers of Mix-ViT into our MA-NET model, we were able to leverage the powerful representational capabilities of this architecture to enhance the detection rate of small and flat polyps.

Our results and experiments indicate that each module in our proposed model plays a critical role in effectively detecting small polyps. By systematically removing and modifying individual modules and evaluating the performance of the resulting model, we were able to demonstrate the importance of each component in achieving high detection accuracy.

The primary implication of our study is that the detection of small and flat polyps, which pose a significant threat due to their high likelihood of developing into colorectal cancer if left undetected, can be improved by placing greater emphasis on the quality of the training dataset and the pre-processing steps. By carefully curating the dataset and applying appropriate pre-processing techniques, it may be possible to enhance the performance of semantic segmentation models in detecting these types of polyps in colonoscopy images.

C. COMPARISON WITH PRIOR STUDIES

An examination of the results presented in Table 6 reveals that the proposed enhanced MA-NET model outperforms several comparable methods by a large margin, despite their more sophisticated design. We hypothesize that the improvement in performance is caused by several factors that are related to both the design of the enhanced MA-NET as well as the training strategy.

In our approach, we address the challenges posed by data disparity and limited training samples by pre-training the encoder before integrating it with the MA-NET decoder. This allows us to leverage existing knowledge and improve the performance of our model even when training data is limited. In contrast, existing methods often focus on training entire segmentation networks from scratch. While this approach can be effective when sufficient training data is available, it can be impractical in situations where publicly accessible data is limited. By pre-training the encoder, our method offers a more feasible solution to this challenge.

In contrast to most existing work, which primarily focuses on the segmentation model with limited attention given to the pre-processing step, our approach emphasizes the importance of both components. While some existing studies have introduced data augmentation methods, there has been little effort to improve or enhance the quality of the data samples themselves. In our work, we hypothesize that there is a relationship between the performance of the model and the quality of the input samples. As such, we have chosen to focus not only on the development of a robust segmentation model but also on introducing an improved pre-processing step to enhance the quality of the input data.

We chose to modify and pre-train the Mix-ViT model for use as our encoder due to its demonstrated superior performance on ultra-fine-grained visual categorization (Ultra-FGVC) tasks. We hypothesize that a model capable of bridging the gap between object categorization and Ultra-FGVC tasks would be particularly effective in detecting small and flat polyps, which often exhibit properties similar to non-polypoid areas. By pre-training the Mix-ViT model and integrating it into our enhanced MA-NET architecture, we aim to leverage its strengths in Ultra-FGVC to improve the detection of these challenging polyps. Based on the results, this approach is superior, predominantly on challenging polyp as compared to existing transformer-based methods.

Furthermore, we compared the performance of our proposed enhanced MA-NET model with four well-established semantic segmentation methods, including the original MA-NET. Our goal is to provide a rigorous evaluation of our approach and demonstrate its advantages over existing methods. The results of this comparison, presented in Tables 6, 7 and Figure 10, confirm the superior performance of our proposed method, particularly in the detection of challenging polyps. There are several key differences between our approach and existing segmentation models that may account for this improved performance.

One of the key differences between our proposed enhanced MA-NET model and the other four semantic segmentation methods is the design of the encoder. In traditional segmentation models, image features are typically extracted using several convolutional layers. In contrast, our approach leverages the attention layer of the Mix-ViT network to extract image features. This allows us to capture more detailed and nuanced information from the input images, which can improve the performance of our model. Although convolutional layers have demonstrated efficacy in feature extraction from images, their capacity to extract polyp features from colonoscopy images, particularly those captured in suboptimal conditions, may be limited. This limitation may arise from several factors including the complexity of visual patterns associated with polyps, variability in their appearance, and the presence of noise or other visual artifacts in the images.

D. MODEL COMPLEXITY ANALYSIS

We evaluate the complexity and stability of our model from different perspectives. Complexity can be measured by the number of layers, operations and feature extraction process.

Stability is assessed by the robustness, generalization, and convergence performance of the model on unseen data samples.

We contrast our model with most of the existing models that adopt multiple branches or encoders to capture multi-scale features and fuse them in different ways. The motivation behind feature fusion is to leverage both global and local information, which theoretically would enhance the segmentation accuracy. In contrast, we focus on simplifying the encoder while maintaining its ability of capturing small and flat polyps.

We adopt a single encoder architecture that reduces the model complexity and computational overhead of feature extraction. Instead of fusing or concatenating features from multiple encoders or branches, our encoder directly produces a single feature map that encompasses both global and local information.

Our encoder is pre-trained on a large and diverse dataset of ultra-fine-grained visual categorization tasks, which enables the model to learn more relevant and transferable features for the polyp segmentation task and improves the model stability and generalization. Compared with multiple encoder or branch architectures, our single encoder architecture achieved superior segmentation accuracy with fewer layers, and operations.

Furthermore, by initializing our encoder with pre-trained weights, we reduced the model's complexity and the training time, as the model converged faster and learned fewer parameters from scratch. This also improved the model performance, especially on small and flat polyps, which are challenging to segment. Furthermore, using pre-training helped us prevent overfitting, which is a common issue in colonoscopy images of small polyps due to the class imbalance.

To further enhance the generalization and stability of our model, we adopted the Dice loss function instead of the conventional binary cross entropy (BCE). The Dice loss function is more suitable for handling class imbalance, which arises when the pixel count of one class is significantly lower than the other.

Despite the superior performances of our model, we acknowledge that the number of parameters can be further reduced. Our model has 20M parameters, which may limit its applicability on low-end devices for real-time predictions. We experimented with different image sizes and found that 256 × 256 pixels yielded the best results for our model. As shown in table 5, other sizes such as 128 × 128, 512 × 512 and 1024 × 1024 pixels led to worse performance. We believe that more research is required to optimize the number of parameters while maintaining the model's ability to capture small polyps.

E. IMPLEMENTATION CHALLENGES

There are several implementation challenges that we have encountered while building our proposed model and we believe such challenges might face future researchers working on the same problem.

A significant challenge that we encountered is the variation of the field of view of the colonoscope across different images, which poses difficulty for image segmentation. To the best of our knowledge, there is no existing solution for this problem in literature as this problem depends solely on the angle of the colonoscope.

A further challenge that stems from the dataset is the inconsistency of the image quality in colonoscopy. The images range from high resolution ones with clear details to blurred, low resolution ones with noise and artifacts. Such low-quality images pose difficulties for accurate polyp segmentation.

In addition, another challenge that needs to be tackled, is the occurrence of specular reflections and blood stains that mislead the model into falsely detecting polyps. These artifacts can result from the illumination and camera settings of the colonoscopy device, or from the bleeding of the mucosa during the procedure. They can affect the contrast and color of the images, making it harder for the model to distinguish between polyps and non-polyp regions.

F. LIMITATIONS AND FUTURE RESEARCH

Despite the superior performance of our method, there are several limitations to our study that warrant further investigation.

One major limitation is that we have not yet tested our method on live stream colonoscopy videos to verify its practicality on real-time datasets. This is an important consideration for the clinical application of our approach, as it is essential to ensure that our method can operate effectively on live data in a clinical setting. Future work should therefore focus on evaluating the performance of our method on live stream colonoscopy videos to assess its practicality and utility in a real-world context.

Despite cross-validating our model on two unseen, public datasets, another limitation of our study is that our method has not been tested on real-life colonoscopy image datasets.

As a result, the performance of our method on unseen, real-world data remains unknown. To address this limitation, future research should focus on evaluating the performance of our method on real-life colonoscopy images to determine its effectiveness in a clinical setting. This will provide valuable insights into the generalizability of our approach and its potential for practical application in the diagnosis and treatment of colon cancer. By rigorously testing our method on real-world data, we can gain a better understanding of its strengths and limitations and identify areas for further improvement.

The final limitation of our study is that our model was sometimes confused by images with strong white light glare. This suggests that addressing the issue of white light reflection is an important consideration for improving the performance of our method, particularly when working with real-life samples. Future research should therefore focus on developing strategies for mitigating the effects of white light glare on our model's performance. This could involve incorporating additional pre-processing steps to reduce glare or developing more robust algorithms that are less sensitive to variations in lighting conditions.

The study demonstrated that the proposed method outperformed similar polyp detection models in detecting small and flat polyps. This suggests that the method effectively learns the complex and subtle features of small, flat polyps and generates accurate and consistent masks. Additionally, the proposed image enhancement layer, which applies contrast limited adaptive histogram equalization (CLAHE) to the L* component of the CIEL*A*B* color space, enhanced the performance of the segmentation model by improving the contrast and visibility of polypoid and non-polypoid regions.

These findings imply that the proposed method can advance early detection of colorectal cancer by improving colorectal polyp detection. However, further studies are required to validate the generalizability and robustness of the method on larger and more diverse datasets and to assess its clinical impact and cost-effectiveness in real-world settings. Moreover, further studies are necessary to better understand the data aspect of the problem. Most existing work has primarily focused on introducing new and complex architectures without considering data quality. This can be addressed by exploring the potential of generative models, such as stable diffusion or generative adversarial networks, to create more samples, enhance existing samples or apply custom augmentation.

VI. CONCLUSION

Colorectal polyps come in various shapes, forms, and colors. A major challenge faced by deep learning researchers has been the misidentification of small and flat polyps. In this paper, we propose an enhanced version of the multi-scale attention network (MA-NET) to segment colorectal polyps from colonoscopy images. We enhance the architecture by replacing the original encoder, based on convolutional layers, with a modified Mix-ViT transformer. Additionally, we introduce a new preprocessing layer that enhances input images by applying Contrast Limited Adaptive Histogram Equalization (CLAHE) to the L* component of images in the CIEL*A*B* color space. This preprocessing step improves the features of input images, making non-polypoid regions more obvious to the model during training. Our model is trained on a combination of two public datasets: Kvasir-SEG and CVC-ClinicDB and tested on two different datasets: ETIS-LaribPolypDB and CVC-ColonDB. Our method has proven effective when tested on several unfiltered colonoscopy images.

REFERENCES

[1] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors," *Przeglad Gastroenterologiczny*, vol. 14, no. 2, pp. 89–103, 2019, doi: 10.5114/pg.2018.81072.

[2] Y. Hao, Y. Wang, M. Qi, X. He, Y. Zhu, and J. Hong, "Risk factors for recurrent colorectal polyps," *Gut Liver*, vol. 14, no. 4, pp. 399–411, Jul. 2020, doi: 10.5009/gnl19097.

[3] J. H. Bond, "Polyp guideline: Diagnosis, treatment, and surveillance for patients with colorectal polyps," *Amer. J. Gastroenterol.*, vol. 95, no. 11, pp. 3053–3063, 2000, doi: 10.1016/S0002-9270(00)02227-9.

[4] World Health Organization. (2022). *Colorectal Cancer*. Accessed: Dec. 20, 2022. [Online]. Available: https://www.iarc.who.int/cancer-type/colorectal-cancer/

[5] C. Williams and R. Teague, "Colonoscopy," *Gut*, vol. 14, no. 12, pp. 990–1003, Dec. 1973, doi: 10.1136/gut.14.12.990.

[6] Y. H. Jeong, K. O. Kim, C. S. Park, S. B. Kim, S. H. Lee, and B. I. Jang, "Risk factors of advanced adenoma in small and diminutive colorectal polyp," *J. Korean Med. Sci.*, vol. 31, no. 9, pp. 1426–1430, 2016, doi: 10.3346/jkms.2016.31.9.1426.

[7] I. A. Issa and M. Noureddine, "Colorectal cancer screening: An updated review of the available options," *World J. Gastroenterol.*, vol. 23, no. 28, pp. 5086–5096, 2017, doi: 10.3748/wjg.v23.i28.5086.

[8] R. Palmier, T. Degand, S. Aho, C. Lepage, O. Facy, C. Michiels, and S. Manfredi, "A colonoscopy quality improvement intervention in an endoscopy unit," *Sci. Rep.*, vol. 12, p. 817, Jan. 2022, doi: 10.1038/s41598-022-04786-y.

[9] L. M. Helsingen and M. Kalager, "Colorectal cancer screening—Approach, evidence, and future directions," *NEJM Evidence*, vol. 1, no. 1, p. e1, Jan. 2022, doi: 10.1056/EVIDra2100035.

[10] S. B. Ahn, D. S. Han, J. H. Bae, T. J. Byun, J. P. Kim, and C. S. Eun, "The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies," *Gut Liver*, vol. 6, no. 1, pp. 64–70, Jan. 2012, doi: 10.5009/gnl.2012.6.1.64.

[11] T. K. Lui and W. K. Leung, "Is artificial intelligence the final answer to missed polyps in colonoscopy?" *World J. Gastroenterol.*, vol. 26, no. 35, pp. 5248–5255, Sep. 2020, doi: 10.3748/WJG.V26.I35.5248.

[12] M. Macari, E. J. Bini, S. L. Jacobs, Y. W. Lui, S. Laks, A. Milano, and J. Babb, "Significance of missed polyps at CT colonography," *Amer. J. Roentgenol.*, vol. 183, no. 1, pp. 127–134, Jul. 2004, doi: 10.2214/ajr.183.1.1830127.

[13] K. Suzuki, D. C. Rockey, and A. H. Dachman, "CT colonography: Advanced computer-aided detection scheme utilizing MTANNs for detection of 'missed' polyps in a multicenter clinical trial," *Med. Phys.*, vol. 37, no. 1, pp. 12–21, Dec. 2009, doi: 10.1118/1.3263615.

[14] J. Lee, S. W. Park, Y. S. Kim, K. J. Lee, H. Sung, P. H. Song, W. J. Yoon, and J. S. Moon, "Risk factors of missed colorectal lesions after colonoscopy," *Medicine*, vol. 96, no. 48, 2017, Art. no. e7468, doi: 10.1097/MD.0000000000007468.

[15] R. Mennigen, J. Kusche, J. Barkun, C. Schreckenberger, and H. Troidl, "Usefulness and limitations of colonoscopy in a proctological clinic," *Surg. Endoscopy*, vol. 2, no. 2, pp. 84–87, 1988, doi: 10.1007/BF00704360.

[16] N. H. Kim, Y. S. Jung, W. S. Jeong, H.-J. Yang, S.-K. Park, K. Choi, and D. I. Park, "Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies," *Intestinal Res.*, vol. 15, no. 3, pp. 411–418, 2017, doi: 10.5217/ir.2017.15.3.411.

[17] S. Chen, S. Lu, Y. Tang, D. Wang, X. Sun, J. Yi, B. Liu, Y. Cao, Y. Chen, and X. Liu, "A machine learning-based system for real-time polyp detection (DeFrame): A retrospective study," *Frontiers Med.*, vol. 9, pp. 1–8, May 2022, doi: 10.3389/fmed.2022.852553.

[18] P. Rasouli, A. D. Moghadam, P. Eslami, M. A. Pasha, H. A. Aghdaei, A. Mehrvar, A. Nezami-Asl, S. Iravani, A. Sadeghi, and M. R. Zali, "The role of artificial intelligence in colon polyps detection," *Gastroenterol. Hepatol. Bed Bench*, vol. 13, no. 3, pp. 191–199, 2020, doi: 10.22037/ghfbb.v13i3.1866.

[19] L. F. Sánchez-Peralta, L. Bote-Curiel, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Deep learning to find colorectal polyps in colonoscopy: A systematic literature review," *Artif. Intell. Med.*, vol. 108, Aug. 2020, Art. no. 101923, doi: 10.1016/j.artmed.2020.101923.

[20] J. Ribeiro, S. Nóbrega, and A. Cunha, "Polyps detection in colonoscopies," *Proc. Comput. Sci.*, vol. 196, pp. 477–484, Jan. 2022, doi: 10.1016/j.procs.2021.12.039.

[21] K. ELKarazle, V. Raman, P. Then, and C. Chua, "Detection of colorectal polyps from colonoscopy using machine learning: A survey on modern techniques," *Sensors*, vol. 23, no. 3, p. 1225, Jan. 2023, doi: 10.3390/s23031225.

[22] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Accessed: Dec. 21, 2022. [Online]. Available: http://lmb.informatik.uni-freiburg.de/

[23] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, in Lecture Notes in Computer Science, vol. 11045, Granada, Spain, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[24] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[25] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017, doi: 10.1109/TPAMI.2016.2572683.

[26] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, H. López-Fernández, Á. Iglesias, J. Cubiella, F. Fdez-Riverola, M. Reboiro-Jato, and D. Glez-Peña, "Deep neural networks approaches for detecting and classifying colorectal polyps," *Neurocomputing*, vol. 423, pp. 721–734, Jan. 2021, doi: 10.1016/j.neucom.2020.02.123.

[27] C. Senore, C. Bellisario, and N. Segnan, "Distribution of colorectal polyps: Implications for screening," *Best Pract. Res. Clin. Gastroenterol.*, vol. 31, no. 4, pp. 481–488, Aug. 2017, doi: 10.1016/j.bpg.2017.04.008.

[28] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-NET: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020, doi: 10.1109/ACCESS.2020.3025372.

[29] X. Yu, J. Wang, Y. Zhao, and Y. Gao, "Mix-ViT: Mixing attentive vision transformer for ultra-fine-grained visual categorization," *Pattern Recognit.*, vol. 135, Mar. 2023, Art. no. 109131, doi: 10.1016/j.patcog.2022.109131.

[30] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*, vol. 2, 2019, pp. 451–462, doi: 10/1007/978-3-030-37734-2.

[31] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/J.COMPMEDIMAG.2015.02.007.

[32] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, Mar. 2014, doi: 10.1007/s11548-013-0926-3.

[33] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, Sep. 2012, doi: 10.1016/j.patcog.2012.03.002.

[34] A. K. Mohammed, S. Yildirim-Yayilgan, I. Farup, M. Pedersen, and O. Hovde, "Y-Net: A deep convolutional neural network to polyp detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–11.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Apr. 2015, pp. 1–14. Accessed: Dec. 23, 2022. [Online]. Available: http://www.robots.ox.ac.uk/

[36] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

[37] J. González-Bueno Puyal, P. Brandao, O. F. Ahmad, K. K. Bhatia, D. Toth, R. Kader, L. Lovat, P. Mountney, and D. Stoyanov, "Polyp detection on video colonoscopy using a hybrid 2D/3D CNN," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102625, doi: 10.1016/j.media.2022.102625.

[38] K.-B. Park and J. Y. Lee, "SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer," *J. Comput. Des. Eng.*, vol. 9, no. 2, pp. 616–632, Apr. 2022, doi: 10.1093/jcde/qwac018.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Virtual Event, 2021, pp. 1169–1179. Accessed: Dec. 24, 2022. [Online]. Available: https://github.com/microsoft/Swin-Transformer

[40] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 10691–10700.

[41] C. Y. Eu, T. B. Tang, and C.-K. Lu, "Automatic polyp segmentation in colonoscopy images using single network model: SegNet," in *Proc. Int. Conf. Artif. Intell. Smart Community*, Singapore: Springer, 2022, doi: 10.1007/978-981-16-2183-3_69.

[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[43] A. Ellahyani, I. E. Jaafari, S. Charfi, and M. E. Ansari, "Fine-tuned deep neural networks for polyp detection in colonoscopy images," *Pers. Ubiquitous Comput.*, vol. 27, no. 2, pp. 235–247, Apr. 2023, doi: 10.1007/s00779-021-01660-y.

[44] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, and Y. Guo, "Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105760, doi: 10.1016/j.compbiomed.2022.105760.

[45] K. Yang, S. Chang, Z. Tian, C. Gao, Y. Du, X. Zhang, K. Liu, J. Meng, and L. Xue, "Automatic polyp detection and segmentation using shuffle efficient channel attention network," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 917–926, Jan. 2022, doi: 10.1016/j.aej.2021.04.072.

[46] S. F. Qadri, H. Lin, L. Shen, M. Ahmad, S. Qadri, S. Khan, M. Khan, S. S. Zareen, M. A. Akbar, M. B. B. Heyat, and S. Qamar, "CT-based automatic spine segmentation using patch-based deep learning," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–14, Mar. 2023, doi: 10.1155/2023/2345835.

[47] M. Ahmad, S. F. Qadri, M. U. Ashraf, K. Subhi, S. Khan, S. S. Zareen, and S. Qadri, "Efficient liver segmentation from computed tomography images using deep learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, May 2022, doi: 10.1155/2022/2665283.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[49] S. Süsstrunk, R. Buckley, and S. Swen, "Standard RGB color spaces," in *Proc. Final Program IS T/SID Color Imag. Conf.*, 1999, pp. 127–134.

[50] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001, doi: 10.1109/34.977559.

[51] *CIE Standard Illuminant D65 | CIE*. Accessed: May 10, 2023. [Online]. Available: https://cie.co.at/datatable/cie-standard-illuminant-d65

[52] Z. Saberi, N. Hashim, A. Ali, P. Boursier, J. Abdullah, and Z. C. Embi, "Adaptive contrast enhancement of satellite images based on histogram and non-linear transfer function methods," *IAENG Int. J. Appl. Math.*, vol. 53, no. 1, pp. 1–9, 2023.

[53] M. U. Khan, M. Safdar, M. F. Mughal, and M. R. Luo, "Performance comparison of uniform color spaces by integrating into a tone mapping operator," in *Applied Sciences in Graphic Communication and Packaging* (Lecture Notes in Electrical Engineering), vol. 477. Singapore: Springer, 2018, pp. 39–45, doi: 10.1007/978-981-10-7629-9_5.

[54] M. F. Hassan, "A uniform illumination image enhancement via linear transformation in CIELAB color space," *Multimedia Tools Appl.*, vol. 81, no. 18, pp. 26331–26343, Jul. 2022, doi: 10.1007/s11042-022-12429-7.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[56] Y. Li, C. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4794–4804, doi: 10.1109/CVPR52688.2022.00476.

[57] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. Al Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, pp. 1–20, 2021, doi: 10.3390/rs13030516.

[58] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356, doi: 10.1109/ICCV48922.2021.00041.

[59] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," 2021, *arXiv:2108.06932*.

[60] Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins. (2022). *ESF-PNet: Efficient Deep Learning Architecture for Real-Time Lesion Segmentation in Autofluorescence Bronchoscopic Video*. Accessed: Feb. 2, 2023. [Online]. Available: http://mipl.ee.psu.edu/

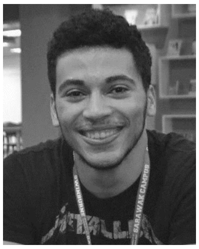
[61] F. Tang, Q. Huang, J. Wang, X. Hou, J. Su, and J. Liu, "DuAT: Dual-aggregation transformer network for medical image segmentation," Dec. 2022, *arXiv:2212.11677*.

[62] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "ColonFormer: An efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80575–80586, 2022, doi: 10.1109/ACCESS.2022.3195241.

[63] E. Sanderson and B. J. Matuszewski, "FCN-transformer feature fusion for polyp segmentation," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13413, 2022, pp. 892–907, doi: 10.1007/978-3-031-12053-4_65.

[64] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13433, 2022, pp. 110–120, doi: 10.1007/978-3-031-16437-8_11.

[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.



KHALED ELKARAZLE is currently pursuing the Ph.D. degree in deep learning with the Swinburne University of Technology, Sarawak Campus. He is an AI Researcher and an Academic Tutor of AI and data science courses with the Swinburne University of Technology. He has a strong interest in computer vision, generative adversarial networks, and facial analysis problems. His research interests include computer vision, generative AI, medical image processing and analysis, and facial recognition. He has published several papers in journals and conferences on these topics.

VALLIAPPAN RAMAN is currently a Professor and the Head of the Department of Artificial Intelligence and Data Science, Coimbatore Institute of Technology. His research interests include the problems of computer vision and pattern recognition, which includes image/video retrieval systems, video semantics, object recognition, and classification systems. Much of his recent research works are artificial intelligence, machine learning, and deep learning in medical imaging. He has also involved in research areas related to health informatics, the Internet of Things (IoT), and data analytics.

PATRICK THEN is currently the Director of the Centre for Digital Futures, Swinburne University of Technology, Sarawak Campus. He is a strong advocate of research and development and commercialization of innovations in big data, data mining, and the Internet of Things. He has established industry collaboration at national and international levels. He has been leading multiple industry-funded projects in research and development in collaboration with prominent ICT partners, such as Sarawak Information Systems Sdn. Bhd. (SAINS), IDS (Malaysia) Sdn. Bhd., Sarawak, and organizations around the world. He has also established partnership between Swinburne and international commercial partners, such as Fusionex International Ltd., U.K., D&J Human Care, South Korea, and Easy Global Market, France. He has won, and has been managing and leading projects worth millions funded by industry and government agencies at national and international level.

CASLON CHUA received the B.Sc., M.Sc., and Ph.D. degrees in computer science from De La Salle University, Manila, Philippines, in 1988, 1993, and 1999, respectively. He is currently the Acting Department Chair of computing technologies with the School of Software and Electrical Engineering, Swinburne University of Technology, Hawthorn, VIC, Australia. His research interests include computing education, data visualization, database systems, human–computer interactions, and software engineering.