IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 June 2023, accepted 11 July 2023, date of publication 19 July 2023, date of current version 27 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297097

RESEARCH ARTICLE

A Two-Stage Method for Polyp Detection in Colonoscopy Images Based on Saliency Object Extraction and Transformers

ALAN CARLOS DE MOURA LIMA[1], LISLE FARAY DE PAIVA[2], GERALDO BRÁZ JR.[1], JOÃO DALLYSON S. DE ALMEIDA[1], ARISTÓFANES CORRÉA SILVA[1], MIGUEL TAVARES COIMBRA[3], (Senior Member, IEEE), AND ANSELMO CARDOSO DE PAIVA[1]

[1]NCA, Federal University of Maranhão, São Luís 65085-580, Brazil
[2]UFR SC, University of Burgundy, 71200 Le Creusot, France
[3]INESC TEC, Faculty of Sciences, University of Porto, 4200-465 Porto, Portugal

Corresponding author: Alan Carlos de Moura Lima (alanlima@nca.ufma.br)

This work was supported in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, under Grant 001; in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil; in part by Fundação de Amparo á Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA), Brazil; and in part by National Funds through the Portuguese funding agency, FCT—Fundacao para a Ciencia e a Tecnologia, within project PTDC/EEI-EEE/5557/2020.

ABSTRACT The gastrointestinal tract is responsible for the entire digestive process. Several diseases, including colorectal cancer, can affect this pathway. Among the deadliest cancers, colorectal cancer is the second most common. It arises from benign tumors in the colon, rectum, and anus. These benign tumors, known as colorectal polyps, can be diagnosed and removed during colonoscopy. Early detection is essential to reduce the risk of cancer. However, approximately 28% of polyps are lost during this examination, mainly because of limitations in diagnostic techniques and image analysis methods. In recent years, computer-aided detection techniques for these lesions have been developed to improve detection quality during periodic examinations. We proposed an automatic method for polyp detection using colonoscopy images. This study presents a two-stage polyp detection method for colonoscopy images using transformers. In the first stage, a saliency map extraction model is supported by the extracted depth maps to identify possible polyp areas. The second stage of the method consists of detecting polyps in the extracted images resulting from the first stage, combined with the green and blue channels. Several experiments were performed using four public colonoscopy datasets. The best results obtained for the polyp detection task were satisfactory, reaching 91% Average Precision in the CVC-ClinicDB dataset, 92% Average Precision in the Kvasir-SEG dataset, and 84% Average Precision in the CVC-ColonDB dataset. This study demonstrates that polyp detection in colonoscopy images can be efficiently performed using a combination of depth maps, salient object-extracted maps, and transformers.

INDEX TERMS Colonoscopy images, deep learning, depth maps, polyp detection, saliency objects, transformers.

I. INTRODUCTION

Gastrointestinal (GI) diseases are prevalent worldwide, causing high mortality and requiring special attention from healthcare systems. Worldwide, there were approximately 10 million deaths related to gastrointestinal diseases in the 2020s [1]. Colorectal cancer is one of the most common gastrointestinal tract diseases and the second deadliest type of cancer. More than 1.9 million people worldwide were diagnosed with this type of cancer, with more than 930,600 deaths in 2020 [2].

Most reported cases of colorectal cancer appear as benign tumors in the colon and rectum. These benign tumors are called polyps and need to be detected as early as possible because they can become cancerous tissues over the years. If they are detected at an early stage, there is a 90% chance

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

of survival over the next five years. However, if they are detected at an advanced stage, the survival rate drops to 10% [3]. The detection and removal of polyps are mainly performed using a colonoscope, which can film and examine the inside of the colon [4]. One of the purposes of this examination (colonoscopy) is to inspect the colon mucosa in real-time to detect and remove polyps (polypectomy). During the colonoscopy, images and videos are taken for further evaluation if necessary [5].

With the advancement of computer-aided diagnostic (CAD) tools, the medical field has been able to improve polyp localization rates by delineating the polyp area using segmentation and detection techniques during or after the examination, making physicians' evaluation even more accurate [6]. According to [7], although CAD tools have brought benefits regarding the polyp localization task, the accurate diagnosis of these lesions is quite challenging because approximately 28% of polyps are not found during a colonoscopy. These challenges arise because of differences in size, texture, angulation of the polyps, presence of organic material, and specular reflections caused by the light from the colonoscopy camera illumination.

However, in recent years, several successes in polyp detection have been achieved, mainly using deep learning-based methods, such as transformers-based architectures. These methods have gained ground in computer vision because they can capture the global context in an image compared with Convolutional Neural Networks (CNN)-based architectures that only capture local receptive fields [8]. Successful examples in application areas such as natural language processing and computer vision motivate their usage for in-body imaging systems, given their ability to learn complex representations and patterns from large-scale data [9].

In our work, this aspect becomes more apparent when we use images associated with the representation of the relief extracted from the mucosa. The idea is to detect areas of high relief that are analyzed in more detail by the detection algorithm, ensuring a greater chance of distinguishing the region containing a polyp from the colon wall. Using transformers for polyp detection can also help identify the majority of lesions. Its power is mainly due to the multi-head attention layers, which can discriminate global features invariably from the vectors of pixels. Also, the transformer models are more flexible and effective because they have dynamic weights. Furthermore, the bipartite matching component can significantly reduce the number of false positives.

This paper describes a two-stage method for the detection of colorectal polyps in colonoscopy images. Accurate detection of polyps in colonoscopy images is a crucial step in the early diagnosis of colorectal cancer. It can alert physicians to the presence or evolution of lesions by drawing bounding boxes around the polyps, for example. The main contribution of this paper is a polyp detection method with high sensitivity and low false positive rates based on a new representation of the colonoscopic images generated by salient object extracted maps.

The rest of this paper is organized as follows: Section II describes the existing work on colonoscopy image segmentation and detection using deep learning techniques. The materials and method used and the proposed methods are described in Section III. The experiments and results are reported in Section IV. Finally, Section V presents the conclusions, limitations, and future research directions.

II. RELATED WORK

Recently, much work has been published on colorectal polyp detection using colonoscopy imaging and deep learning techniques. Here, we summarize recent work on this topic. As the salient object extraction step is crucial to our detection method, we will also describe some methods for polyps region identification in images. In the literature, these methods are commonly referred to as segmentation methods. In our proposal, we used this method to identify regions with a higher probability of polyps occurrence preliminarily.

Regarding the segmentation of colorectal polyps, some work such as [10] and [11] presented architectures using an encoder-decoder CNN with squeeze and excitation blocks. Other works have presented post-processing techniques aimed at false negatives reduction. A ResUnet++ architecture with the addition of attention blocks in the encoder is used in [11] followed by two post-processing steps. On [12], a U-Net [13] and MobileNetV2 [14] were combined for image feature extraction.

In [10] a CNN SE-Resnext-50 [15] was used as an encoder, and two decoders worked in parallel, where one segments the possible area containing the polyp and the other segments the polyp contours. Finally, a simple two-layer U-net performs a regression of the result of the first U-net on the result of the second U-net. A polyp segmentation framework that uses a pyramid view transformer backbone as an encoder to extract explicit and more robust features was proposed in [8]. The following two proposed models use attention-based methods for polyp segmentation: [16] presents CaraNet, which improves small object segmentation using axial reverse attention and a channel-wise feature pyramid, and [17] propose a decoder that leverages multiscale features of hierarchical vision transformers to address the lack of learning local contextual relations among pixels.

As our proposal aims to detect colorectal polyps, we highlight works as [18] that utilized a Fully Convolutional Neural Network (MDeNetplus) trained with RGB images embedded in Gaussian masks extracted from ground truth images to enhance the training images.

In addition, [19] used a SegNet [20] network with an encoder-decoder architecture for semantic segmentation. Furthermore, [21] used the ResNet-101 [22] and YOLO-v2 [23] networks with a DarkNet19 [23] backbone, respectively, to validate their experiments. The proposal of [24] was an encoder-decoder CNN architecture with residual blocks with squeeze and excitation to identify polyp regions

in colonoscopy images. In [25] and [26] were presented quality enhancement techniques for polyps detection in colonoscopy images using VGGNet [27] and YOLO-v3 [28], respectively.

In [29] proposed a model based on convolutional networks and transformers, called COTR, for polyp detection using a ResNet-18 [22] CNN for feature extraction and transformer encoder and decoder layers for coding, feature recalibration, and region identification. Moreover, [30] used a YOLO-v5 [31] architecture with an auto-attention mechanism for polyp detection in colonoscopy images, combining training images into mosaics to increase image variability and enhance channels with information-rich features. In [32] DC-SSDNet was proposed, a version of DenseNet that improves the feature extraction capacity for small polyp detection and achieves a high precision with less computational time.

Although the presented related works are essential and have significant advantages in the fields of detection and segmentation of polyps in the gastrointestinal tract, some of them have limitations, such as the use of private datasets that difficult the reproducibility of the work. Also, these private datasets may represent a partial range of variation in the population. Some studies also used repeated image sequences in training and testing, which can lead to overfitting and negatively affect the ability of the models to detect polyps in different images. The manual separation of images to create the CNN training dataset can be time-consuming, and there may be human errors in labeling the images, leading to inaccurate predictions. Finally, we state that the feature modification process in the pre-processing phase can introduce bias in the dataset and negatively impact the model's accuracy.

To address the limitations of the existing work, we propose an automatic deep learning-based method for extracting the key features from colonoscopy images. Our method aims to segment regions with a higher probability of containing a polyp, reduce the examination area, classify these regions as polyps, and annotate them with bounding boxes to assist the specialist. Our proposed method uses transformers to effectively classify the regions containing polyps and accurately detect polyps with different sizes, contrast, shape, location, and quantity.

III. MATERIALS AND METHOD

This section presents the materials and methods used in this study. The proposed method consists of five steps, detailed in Figure 1. In the first step, pre-processing techniques were applied to the datasets. In the second step, the depth maps are extracted. In the third step, the resulting depth maps are used to train a transformer to extract salient object maps. In the fourth step, the salient object maps obtained in the second step were processed into binary masks, and SGB images were created by combining the binary masks, green channel, and blue channel. Finally, in the fifth step, an object detection technique was applied to extract the polyp regions.

A. MATERIALS

This study used four public datasets of colonoscopy images and their ground truth images. The main datasets are: CVC-ClinicDB [33] with 612 images; CVC-ColonDB [34] with 300 images; ETIS-LaribPolypDB [35] composed of 196 high-resolution images; and Kvasir-SEG [36] with 1,000 images with different resolutions. Due to the wide variety of colonoscopy devices and image acquisition issues, the images in these datasets may differ in lighting, quality, and angulation.

It is worth noting that the CVC-ClinicDB dataset includes images from 29 different patients, meaning that multiple images of the same polyp can be found, taken from different angles and positions. Using these images in the training and validation stages of the model can lead to overfitting. In the experiments conducted in this study, the same polyps were not used in either the training or validation stages. This peculiarity in image repetition was not observed in the other datasets used in this study. As mentioned earlier, the images in these datasets had a variety of resolutions. Therefore, digital image-processing techniques must be applied to standardize these images.

B. RESIZING AND DATA AUGMENTATION

Initially, all images were resized to 640 $\times$ 640 pixel resolution while maintaining the *width* x *height* ratio by applying zero padding and preserving the image format. This image resolution was defined because of the limitations of the development environment.

We use the following data augmentation approaches [37]: translations of up to 30% of the total image size with a randomly chosen value; rotations of a randomly chosen angle in the range between 0° and 90°; and scaling with a randomly chosen scaling factor in the interval [−0.5, 1.8]. We also applied horizontal and vertical flipping.

The satisfactory results obtained in the work of [37] are due to the careful selection of parameters that capture the different scenarios in which polyps can be found during colonoscopy examination. These parameters contributed to the accuracy and robustness of the results.

C. EXTRACTION OF DEPTH MAPS

After pre-processing the images, their depth maps were extracted using a monocular depth estimation-based architecture. For this purpose, a pre-trained Dense Prediction Transformer (DPT) model [38] was used. It was trained for 60 epochs, with a batch size of 16, and an image dataset of more than 1.4 million images.

Figure 2 shows a sample of 102.tif file from the CVC-ClinicDB dataset with the original image, ground truth, and results obtained after applying the DPT model.

The observed results indicate that the DPT model is promising for depth map extraction of salient polyps compared to the provided ground truth.

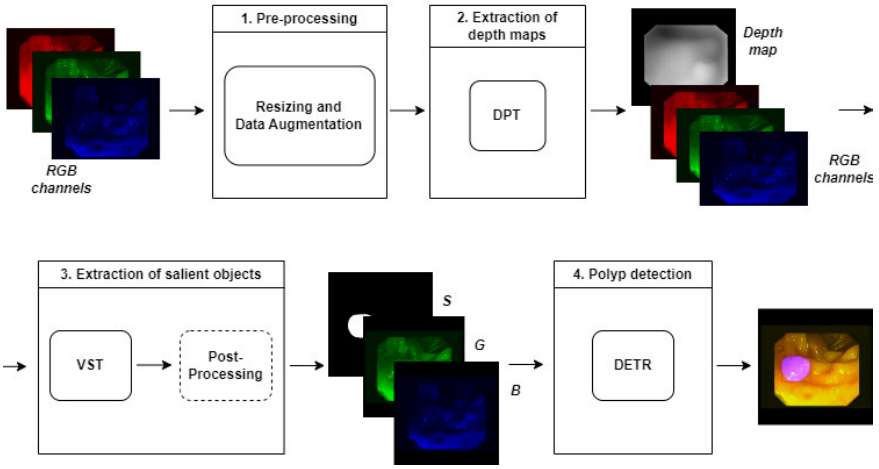

FIGURE 1. Steps involved in the proposed method.

(a) Original. (b) GT. (c) Depth map.

FIGURE 2. DPT model application example. (a) original image (b) ground truth (GT). (c) extracted depth map.

D. EXTRACTION OF SALIENT OBJECTS

The Visual Saliency Transformer (VST) architecture [39] was used to extract salient objects and depth geometric information associated with polyp regions. For VST training, we need depth maps (extracted in the previous step), RGB images, and their respective ground truth images. The encoder uses a variant of the ViT model called T2T-ViT [40] as its backbone for feature map extraction. The T2T-ViT model employs the Tokens-to-Token (T2T) transformation, which aims to reduce the length of the input sequence of tokens.

First, before the input images are sent to the encoder, they are resized from 640 × 640 pixels to 384 × 384 pixels (values suggested by the VST architecture due to memory processing needs). At the encoder, tokens (T) of various levels are generated from a sequence of patches (T') of the input image with size $l = h \times w$. In the end, T is reshaped to a 2D image to recover spatial structures. This operation is called restructuring (Eq. 1) [39].

$$T = MLP(MultiHeadAttention(T')), \tag{1}$$

where MultiHeadAttention [41] and MLP [42] are a Multi-Head Attention layer and an original MLP (Multi-Layer Perceptron) network, respectively.

After the restructuring process, the image is split into $k \times k$ patches with s overlapping, and p zero-padding is utilized to pad image boundaries. The patch length sequence size ($l_o = h_o \times w_o$) is obtained in Equation 2, as demonstrated in the study by Liu et al. [39].

$$l_o = \left\lfloor \frac{h + 2p - k}{k - s} + 1 \right\rfloor \times \left\lfloor \frac{w + 2p - k}{k - s} + 1 \right\rfloor, \tag{2}$$

where w, h, p, k, and s are the image width, image height, zero-padding, patch size, and overlapping size, respectively.

The T2T transformation is applied three times, with the patch sizes set to $k = [7, 3, 3]$, the overlap sizes set to $s = [3, 1, 1]$, and the padding sizes set to $p = [2, 1, 1]$ at each iteration. Initially, the embedding dimensionality c is set to 64, which is then transformed to a higher dimensionality d of 384, following the recommendations in the cited study [39]. The RGB images and depth maps follow a similar flow in a standard encoder transformer to extract tokens from the patches.

Subsequently, the resulting tokens are sent to the Cross Modality Transformer [39]. There, linear projection patterns are generated in T_r^{ε} and T_d^{ε}, computing the attention between the query matrix Q from the RGB images and the key matrix K from the depth maps.

In the Cross Modality Transformer, linear projection patterns are made in T_r^{ε} and T_d^{ε}, generating matrices Query (Q), Key (K) and Value (V), so that the attention can be calculated between the matrices Q RGB and K from the depth map. The result of this operation follows the typical path of an encoder transformer with a Feed-Forward Neural Network (FNN), residual connections, and a normalization layer.

In the decoder path, the RT2T (reverse of T2T module) module is employed to reverse the process implemented by the T2T module. Initially, a projection of the input patch tokens is performed, reducing their dimensions from $d = 384$ to $c = 64$ [39]. Subsequently, a linear projection is used to expand the embedding dimension from c to ck^2. At this stage, each token is interpreted as an image patch of size $k \times k$, and neighboring patches are overlaid with a stride of s.

Finally, the RT2T module was used three times to upsample the original images. At each stage of upsampling, the

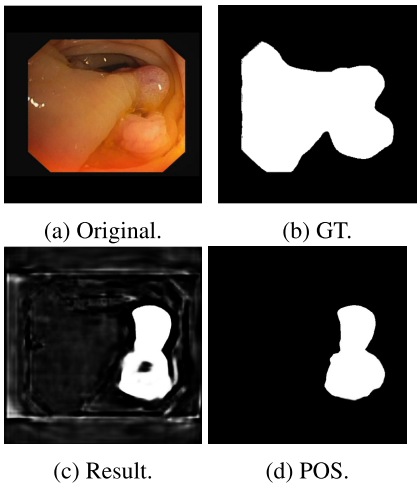

FIGURE 3. CVC-ClinicDB examples after the model inference for salient object extraction. (a) original image (b) ground truth (GT). (c) results of the first inference. (d) after applying the post-processing.

FIGURE 4. Samples of the results of RGB standard images after merging the G and B channels with the ground truth, on the CVC-ColonDB dataset.

FIGURE 5. Sample of the result of the image in the SGB pattern after the union of the G and B channels with the saliency map, in the CVC-ClinicDB dataset.

inverse values of those used in the encoder flow are used to define the size of the resulting patch: $k = [3, 3, 7]$, $s = [1, 1, 3]$, and $p = [1, 1, 3]$ [39]. Thus, the length of the patches was gradually increased to $H \times W$, corresponding to the original size of the input image.

E. POST-PROCESSING OF SALIENT OBJECT EXTRACTION (POS) AND SGB IMAGE CREATION

The saliency map extraction step results in some regions that are not part of the polyp area, but correspond to other high-relief structures in the colon wall. We apply a thresholding technique to eliminate some disposable pixels, using an empirically defined threshold value of 4, which was chosen empirically after performing some tests. This resulted in a more visible polyp area. However, one or more holes are also formed.

Thus, we applied a hole-filling algorithm that first detects all contours in the image using the Canny operator [43]. Next, it selects the largest contours and fills all objects whose contours are selected. Finally, a morphological opening operation is performed using a 10×10 elliptical structuring element. The size and shape of the structuring element were chosen because of the oval shape of the holes and because it is a value that does not affect the reconstruction of the resulting masks.

Figure 3 shows a sample of 22.tif file from the CVC-ClinicDB dataset after inference of the salient object extraction step.

Next, we created SGB images formed by three channels: the saliency map (S), green (G), and blue (B). The decision to remove the red (R) channel, apart from being made after analyzing the results of several tests, is mainly related to the fact that the texture of the colon wall is a mixture of red colors, which confuses the detection algorithm and gives more prominence to this region than to the polyp itself [44].

To emphasize the choice of only blue and green channels, we highlight the work of [45] where, after performing experiments with the RGB channels of colonoscopy images. This study found that these two channels can efficiently discriminate the surfaces of the colon wall between normal and abnormal and are less susceptible to interference caused by differences in illumination during image acquisition. Thus, we propose this three-channel combination to create a representation of a colonoscopy image with appropriate feature sets capable of visually discriminating a polyp from the colon wall.

The provided expert masks (ground truth) and RGB images were used to train the supervised model. We applied preprocessing techniques to both the masks and RGB images to ensure accuracy. The decision to use expert masks rather than saliency maps was made to avoid any form of data leakage. Since the dataset used in the initial phase of our method is identical to the one used in the secondary phase, we ensure that the training of each stage utilizes the same information, resulting in consistent learning and prevention of cross-contamination. Importantly, these images from the training subset were not evaluated during the VST stage, given that they were used to train the VST itself.

Figure 4 shows graphical representations of the SGB images from the CVC-ColonDB dataset. The ground truth (GT) replaces the R-channel value to generate an SGB image representation.

For the evaluation of the supervised model, a different treatment was applied than for the training images because instead of using the ground truth, the images resulting from the extraction process of salient objects (saliency map - SM) were previously applied only to these test images. Figure 5 shows a graphic representation of the results of the SGB images in the CVC-ClinicDB dataset.

F. POLYP DETECTION

The next step in the proposed method is responsible for detecting the polyps in the colonoscopy images using the DEtection TRansformer (DETR) architecture (Fig. 6).

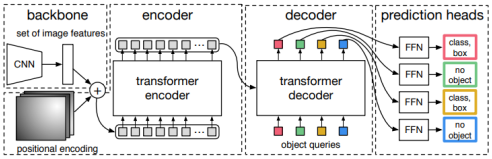

FIGURE 6. DETR architecture [46].

DETR was chosen as the standard detector because it has an easy-to-understand architecture with only three modules and provides consolidated object detection results, making it suitable for complex tasks such as polyp detection. It can also handle data with complex shapes, object variations, and crowded scenes [47]. It uses ResNet with 50 depth layers (ResNet-50 [22]) as a backbone for extracting the feature maps from the input images $x \in \mathbb{R}^{H\times W\times 3}$ (three channels). ResNet was chosen because, for the transformers model, the residual blocks present in ResNet can extract key features for training the model. The initialization weights of the ResNet are the result of previous training with ImageNet.

The result of the ResNet is a flattened feature map that is added to a representative vector containing the position of each pixel in the image (positional encoding). This vector is computed using the sine function (Eq. 3).

$$PE_{(i,2j)} = sin(i/10000^{2j/d_{model}}),$$
$$PE_{(i,2j+1)} = cos(i/10000^{2j/d_{model}}), \quad (3)$$

The resulting vector is a sequence sent to a transformer encoder with 6 Attention layers, where each layer is formed by a Multi-head Attention block and an FNN. The resulting vector from the encoder is sent to the decoder, which also consists of six Attention layers.

IV. RESULTS

The purpose of this section is to present and discuss the results obtained after performing experiments using our polyp detection method. For the detection method to achieve satisfactory results, the first stage of our method must extract the main regions containing polyps with high precision, thereby reducing the scope area by extracting salient objects.

Due to the importance of the salient object extraction method, we first present the experiments and results obtained in this step. Finally, we present the experiments and results obtained after applying the polyp detection method to the images resulting from the first stage of extraction combined with the green and blue channels.

The experiments were conducted on a computer with the following specifications: Intel (R) Core (TM) i7-7700K with 16 GB of RAM, CPU clock frequency of 3.60 GHz, and NVIDIA GeForce GTX 1080 Ti GPU with 12 GB of memory. The programming language Python and the machine learning library Pytorch were used to implement the proposed method.

A. SALIENT OBJECT EXTRACTION

To evaluate the performance of the salient object extraction algorithm, this study used Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) metrics to calculate the region with the highest possibility of polyps occurring.

In this stage, three experiments were conducted, each one employing a training and validation data split of 80% and 20%, respectively, and a data augmentation ratio of 1:20. In Experiment 1, training was carried out with 1,496 images from the Kvasir-SEG, CVC-ColonDB, and ETIS-LaribPolypDB datasets, while testing was performed on 612 images from the CVC-ClinicDB dataset. Experiment 2, on the other hand, utilized 850 images from the Kvasir-SEG dataset for training and 150 for testing. In Experiment 3, the Kvasir and CVC-ClinicDB datasets were used for training, the ETIS-LaribPolypDB dataset for validation, and the model was tested on the CVC-ColonDB dataset. Those experiments were trained on the VST model over 100 epochs with a batch size of 6 and an initial learning rate of 0.0001, employing an AdamW optimizer.

We chose the AdamW optimizer because it is an extension of the stochastic gradient descent, which has recently been widely used in deep learning tasks with transformers. It is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure because it can iteratively update the model weights based on training data [48]. In addition, the AdamW optimizer can assign different learning rates to various parameters, resulting in consistent update magnitudes even with unbalanced gradients. This property is critical for correctly training a transformer architecture because the gradients of the Attention modules are highly unbalanced [49].

1) EXPERIMENT 1

After training the salient object extraction architecture, we evaluated the final model on the 612 images from the CVC-ClinicDB dataset, and the results are presented in Table 1. The resulting polyp masks, generated after model inference, had saliency information beyond the polyp region, so we applied a post-processing technique to delimit only the region containing the polyp.

After applying the post-processing step, we performed a new evaluation of the resulting salient object-extracted images. The values obtained from the new metrics are shown in Table 1. The results were superior to those obtained before this step for all metrics because more pixels are now classified as belonging to the region bounded by the ground truth.

Figure 7 provides a visualization of the progression of training and validation losses throughout the training phase in Experiment 1.

Among the results evaluated using the CVC-ClinicDB database (Sec. II), our method was the only one to present values for precision and recall metrics. However, for papers that presented IoU and Dice metrics, [17] performed better with 89.9% and 94.3% versus 86.6% and 90.9% of ours for

TABLE 1. Comparison of the performance of the proposed method with the literature. The best results are in bold.

Work	Test dataset (Samples)	PRE	REC	IoU	Dice
[8]	CVC-ClinicDB (612)	-	-	0.889	0.937
[16]	CVC-ClinicDB (612)	-	-	0887	0.936
[17]	CVC-ClinicDB (612)	-	-	**0.899**	**0.943**
Proposed method (before POS)	CVC-ClinicDB (612)	0.899	0.878	0.802	0.852
Proposed method (after POS)	CVC-ClinicDB (612)	**0.947**	**0.891**	0.866	0.909
[8]	Kvasir-SEG (1,000)	-	-	0.864	**0.917**
[10]	Kvasir-SEG (1,000)	0.942	0.915	**0.917**	-
[11]	Kvasir-SEG (1,000)	0.822	0.875	0.832	0.850
[12]	Kvasir-SEG (1,000)	-	-	0.816	0.897
Proposed method (before POS)	Kvasir-SEG (1,000)	0.913	0.919	0.831	0.853
Proposed method (after POS)	Kvasir-SEG (1,000)	**0.952**	**0.937**	0.879	0.899
[17]	CVC-ColonDB (300)	-	-	0.745	**0.825**
Proposed method (before POS)	CVC-ColonDB (300)	0.834	0.832	0.716	0.743
Proposed method (after POS)	CVC-ColonDB (300)	0.960	0.848	**0.765**	0.811

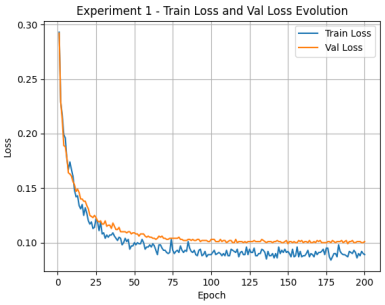

FIGURE 7. Train and validation loss evolution of Experiment 1.

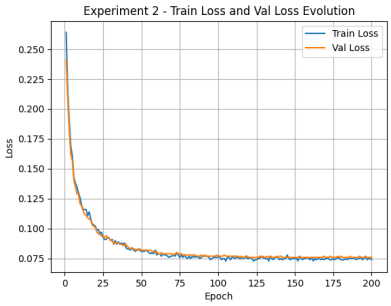

FIGURE 8. Train and validation loss evolution of Experiment 2.

IoU and Dice, respectively. However, it is important to note that [17], which is also based on transformers, uses images of the same polyp (image sequences) from the CVC-ClinicDB database in the training and testing phases, which may bias the results.

Concerning the experiments where CVC-ColonDB was used as the test dataset, we can also compare our method with the work of [17]. Although their method achieved a better performance in the Dice metric with 0.825 against our 0.811, our method outperformed theirs in the IoU metric, scoring 0.765 against 0.745.

2) EXPERIMENT 2

We repeated the method described in Experiment 1 using the Kvasir-SEG dataset (Experiment 2). The results before and after post-processing are presented in Table 1.

For Experiment 2, Figure 8 demonstrates how the training and validation losses evolved during the training phase.

3) EXPERIMENT 3

In Experiment 3, we also repeated the method described in Experiment 1, but now using CVC-ColonDB as a test dataset. The results before and after post-processing are presented in Table 1.

In Figure 9, we graphically present the changes in training and validation losses throughout the training phase for Experiment 3.

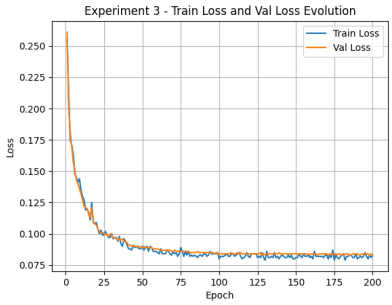

FIGURE 9. Train and validation loss evolution of Experiment 3.

Table 1 presents a comparative analysis between the results of Experiment 1, Experiment 2, and Experiment 3 before and after post-processing, alongside related work in the fields of polyp segmentation and salient object extraction.

In all experiments, we conducted a significance test [50] to compare the results obtained before and after post-processing. We used the paired t-test [51] to evaluate the significant impact of the post-processing stage on the results. This test allows for a direct comparison of means between the same groups under two different conditions. The significance level for the test was set at $\alpha = 0.05$. Our results from Experiment 1, Experiment 2, and Experiment 3 demonstrated statistical significance in PRE, REC, and DSC metrics. This finding not only strengthens the argument for the beneficial impact of our post-processing stage on polyp detection

but also underlines the effectiveness and robustness of our method.

Specifically related to the results presented on the Kvasir-SEG dataset and described in Section II, [10] stands out, which achieved a higher IoU than the proposed method, obtaining 91.7% compared to 87.9%, respectively. This result ensures that related work can correctly detect more pixels in the resulting segmented mask relative to the ground truth. However, our study achieved a higher precision and recall (95.2% and 93.7% versus 94.2% and 91.5%) [10]), which ensures that our model generated fewer false positives. Regarding the Dice metric, [8] show a higher value of 91.7%, compared to 89.9% in our work, highlighting the work [8], which uses transformers. The results of these experiments indicate that this strategy is effective for extracting the region of interest (salient object), leading to better detection results in the later steps.

B. POLYP DETECTION

Next, we present the results obtained after conducting the experiments using the proposed method. The results of each experiment are evaluated by the values obtained in the metrics average precision (AP), recall (REC), precision (PRE), and f-score (F1).

We performed three experiments using the salient object extractor built into the experiments described in the previous section. Therefore, it is necessary to train the DETR architecture. The three experiments performed with the DETR architecture were run with the following configuration: batch size of 2, AdamW optimizer [48] with learning rate of 0.0001, score threshold of 0.5, and IoU threshold of 0.5. The training was performed over 200 epochs. The choice of 200 epochs is because the transformers architecture requires more training time to converge. During training, we found that the model typically converged by the 200th epoch.

Experiment 1 used 1,330 images for training from the Kvasir-SEG and CVC-ColonDB datasets, 196 images for validation from ETIS-LaribPolypDB, and 612 images for testing from the CVC-ClinicDB dataset. Experiment 2 used 700 images for training, 150 for validation, and 150 for testing, all from the Kvasir-SEG dataset. Experiment 3 used 1,612 images for training from the Kvasir-SEG and CVC-ClinicDB datasets, 196 images for validation from ETIS-LaribPolypDB, and 300 images for testing from the CVC-ColonDB dataset. All of the experiments were performed using a standard RGB and our proposed SGB representation as inputs to the DETR architecture.

Table 2 presents results obtained in Experiment 1, Experiment 2, and Experiment 3, along with a comparison with the related work in the field of polyp detection.

Comparing the results of the experiments with RGB and SGB images, we verified that the experiment with SGB images achieved better results in all evaluated metrics.

Comparing the results of our work with those of the literature when CVC-ClinicDB was used as the dataset for model validation, such as in [19], [21], [26], and [32], works [21] and [26] obtained the highest value of metric F1 (94. 0%) compared to all related works and the proposed method (93.2%), with the images in SGB. However, [21] used private datasets to perform training. In terms of the AP metric, our method outperformed the method proposed in [32], achieving 92.6% compared to 92.2%. For the CVC-ColonDB dataset, [29] achieved the highest value for metric F1 (92.6%).

The paper [24] used the Kvasir-SEG dataset in its methods. Our method performed better than all others using Kvasir-SEG, achieving an F1 of 94.0% with SGB images, compared to 90.7% in [30]. Compared to [24], our method achieved an average precision of 92.6% with SGB images, compared to 81.6% in related work.

The works [29] and [30] used transformers-based architectures in their methods. DETR is a variation of the architecture used in [29]. Although there are only two papers, they show that transformers-based approaches achieve higher values for the metrics analyzed compared to other papers that do not use transformers but on the same datasets.

Among the results obtained after the execution of the proposed method, it was observed that there was a significant improvement in the results when using SGB images compared to the RGB standard. The positive aspects highlighted allowed the proposed method to achieve good results in the detection of lesions in the gastrointestinal tract compared with the literature. Despite some limitations, such as the exaggerated presence of illumination, reflections, internal structures, and organic materials, this study can contribute to the development of robust automatic methods.

To further validate the efficacy of our proposed method, we conducted two additional experiments, comparing the effectiveness of using the estimated depth map as a channel directly in the detection against our proposed method of saliency map extracted maps, and integrating the S channel from the first step with the original RGB channels of the images.

In the first experiment, we replaced the S channel with the estimated depth map (D channel) produced using the DPT architecture. This depth map was directly combined with the G and B channels of the RGB images, forming the DGB images, and these combined channels were used to train the DETR model. The purpose of this experiment was to investigate the value of the saliency map by comparing the detection performance using the direct depth map versus the saliency map.

Our results demonstrated in Table 3 show that while the depth map provides valuable spatial information for polyp detection, the use of the saliency map outperformed the depth map in our detection method. The superior performance of the saliency map could be attributed to its ability to better highlight the regions of interest within the images, making it easier for the detection method to identify the polyps.

In the second experiment, we integrated the S channel from the first step with the original RGB channels of the images, forming the SRGB images. This experiment aimed

TABLE 2. Comparison of the performance of the proposed method with the related work. The best results are in bold.

Work	Test dataset (samples)	AP	REC	PRE	F1
[25]	Custom	0.914	-	-	-
[19]	CVC-ClinicDB (612)	-	0.882	0.931	0.906
[21]	CVC-ClinicDB (612)	-	0.902	0.983	**0.940**
[26]	CVC-ClinicDB (612)	-	0.915	**0.966**	**0.940**
[32]	CVC-ClinicDB (612)	**0.922**	0.922	0.910	0.884
Proposed method	CVC-ClinicDB (612) - RGB	0.895	0.923	0.883	0.898
Proposed method	CVC-ClinicDB (612) - SGB	0.916	**0.944**	0.920	0.932
[24]	Kvasir-SEG (1,000)	0.816	-	-	-
[30]	Kvasir-SEG (1,000)	-	0.899	0.915	0.907
Proposed method	Kvasir-SEG (1,000) - RGB	0.911	0.927	0.895	0.911
Proposed method	Kvasir-SEG (1,000) - SGB	**0.926**	**0.931**	**0.951**	**0.940**
[18]	CVC-ColonDB (300)	-	0.910	0.883	0.896
[29]	CVC-ColonDB (300)	-	**0.935**	0.916	**0.926**
Proposed method	CVC-ColonDB (300) - RGB	0.819	0.861	0.907	0.883
Proposed method	CVC-ColonDB (300) - SGB	**0.842**	0.879	**0.932**	0.905

TABLE 3. Results of polyp detection method using DETR and images with the R and G channels complemented by the estimated depth maps (DGB images).

	Experiment 1	Experiment 2	Experiment 3
AP	0.768	0.884	0.813
REC	0.822	0.892	0.858
PRE	0.730	0.913	0.890
F1	0.773	0.908	0.873

TABLE 4. Results of polyp detection method using DETR and images with the R, G and B channels complemented by the saliency maps (SRGB images).

	Experiment 1	Experiment 2	Experiment 3
AP	0.902	0.923	0.839
REC	0.943	0.929	0.871
PRE	0.920	0.941	0.930
F1	0.931	0.935	0.899

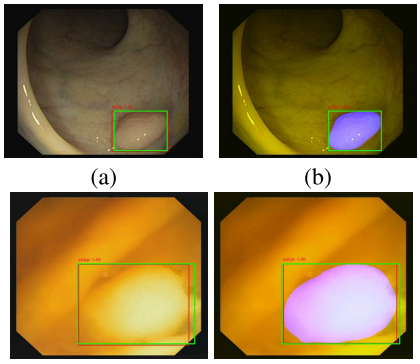

FIGURE 10. Results obtained with CVC-ClinicDB images. (a) 127.tif file in RGB. (b) 127.tif file in SGB. (c) 152.tif file in RGB. (d) 152.tif file in SGB.

to investigate the effect of including the entire color information in the model. However, the results indicated that the performance with the SGB channels was superior in all metrics. This demonstrates that including the original R channel did not contribute significantly to the polyp detection performance, and the S channel provides a more accurate and efficient representation for this task. This finding further underlines the effectiveness of our proposed method of extracting and utilizing saliency information from colorectal polyps. The results of this experiment are shown in Table 4.

Now, we present some examples of successful lesion detection using the proposed model. For ease of understanding, when there is a green bounding box, it represents the expert annotation provided by the dataset. If the bounding box is red, it represents the region detected by the proposed model.

The detector proposed in this paper proved to be efficient, hitting several regions with polyps even though there was a great diversity among the elements in the images. In some situations, the model maintained its detection characteristics even with the exaggerated presence of illumination, reflection, internal structures, and organic material.

Figure 10 shows the results obtained for the 127.tif (10a and 10b) and 152.tif (10c and 10d) files (CVC-ClinicDB dataset). Related to 127.tif file, we can see that in the RGB (10a) and SGB (10b) images, there is an almost complete overlap between the bounding boxes, which is better in the SGB image. For the 152.tif file, the model performed better using the SGB image (10d) than with the RGB image (10c). This case was an image of reduced quality. The DETR model showed better results with more overlap between bounding boxes in the SGB image.

Although our method was able to detect most of the polyps present in the datasets, there were some detection errors. Most polyp detection errors are related to the detector misinterpretation of other regions of the colon wall as polyps. There are also situations in which image acquisition problems are interpreted as lesions. Examples of image acquisition problems include specular reflection, high brightness, shadows, and the presence of organic material in the colon. In addition, the structure of some polyps can significantly interfere with the evaluation of the detector because some are very small or flat, making them difficult to see even with the naked eye.

The results are shown in Figure 11 (files 71.tif (11a and 11b) and 200.tif (11c and 11d) of the CVC-ClinicDB dataset). For 71.tif file, when using the RGB image (11a), the model cannot obtain the region of either of the two polyps correctly, generating a false positive. With the SGB image (11b), the model can only correctly identify one of the two polyps in the image. In the 200.tif file of the CVC-ClinicDB dataset,

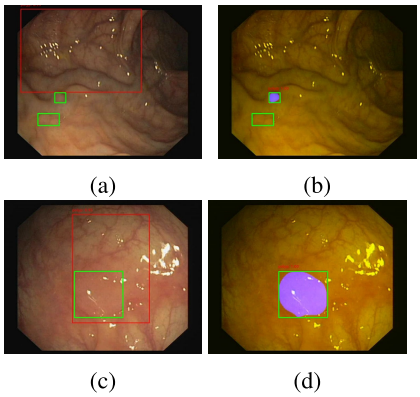

FIGURE 11. Results obtained with CVC-ClinicDB images. (a) 71.tif file in RGB. (b) 71.tif file in SGB. (c) 200.tif file in RGB. (d) 200.tif file in SGB.

in RGB (11c), a large region containing the polyp was classified as belonging to the polyp class, creating regions of true positives and regions of false positives. The overlay was complete in the SGB image (11d).

V. CONCLUSION

In this study, we proposed a two-stage method for polyp detection in the gastrointestinal tract using colonoscopy images. First, extraction is performed using a transformer architecture (VST) to identify salient objects, supported by depth maps to locate possible areas containing colorectal polyp lesions. Once these regions were extracted, the suspicious regions were processed to delineate only the area containing the polyp. The resulting extracted images were added to the green and blue channels of the image dataset used in this paper to create SGB images. These SGB and RGB images were sent to a transformers-based polyp detection architecture, the DETR.

The obtained results motivated us to conclude that our proposed method is adequate for extracting the best areas containing polyps and detecting them with high accuracy, showing the advantages of using SGB images. Furthermore, our method can contribute to the development of a CADx system that can assist physicians in the diagnosis and treatment of colorectal cancer.

In addition to our primary experiments, we conducted two complementary experiments to further verify the effectiveness of our method. The first supplementary experiment compared the direct use of the estimated depth map with our approach of using a saliency map. The second experiment evaluated the effect of including the original color information in SRGB images. In these additional investigations, our method demonstrated superior performance, underscoring its effectiveness and robustness in polyp detection.

Although the results were satisfactory, there are ways to improve the proposed method to reduce its limitations and increase its efficiency. It is clear that for the successful detection of polyps in SGB images, it is necessary that the salient object extraction step must comply with a good performance. In addition, the use of an architecture based on transformers, although it is responsible for an improvement in the results, requires a high processing power due mainly to a large number of parameters in the model, since the use of the bipartite match has a cubic time complexity in relation to the number of bounding boxes ground truth [47] and by processing an image as a text vector.

As suggestions for future work, we can consider applying the method to other colonoscopy image datasets to increase the training step. Also, we suggest analyzing other geometric feature techniques to add more details to the salient object extraction step. Other possibilities for improvement include the use of colonoscopy examination video datasets to validate the polyp detection process in real-time; the use of a parameter optimization technique to configure the model, such as particle swarm optimization, genetic algorithm, Bayesian optimization, tree-structured Parzen estimator; the use of an image quality enhancement technique aimed to remove problems in the acquisition stage; and the use of other post-processing techniques to further reduce the number of false positives.

REFERENCES

[1] Organização Mundial da Saúde. (2020). *Colorectal cancer. International Agency for Research on Cancer*. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[3] L. T. Thu Hong, N. Chi Thanh, and T. Q. Long, "Polyp segmentation in colonoscopy images using ensembles of U-Nets with EfficientNet and asymmetric similarity loss function," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Oct. 2020, pp. 1–6.

[4] F. Deeba, F. M. Bui, and K. A. Wahid, "Computer-aided polyp detection based on image enhancement and saliency-based selection," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101530.

[5] S. K. Ratuapli and H. E. Vargas, "Colonoscopy in liver disease," *Clin. Liver Disease*, vol. 4, no. 5, pp. 109–112, Nov. 2014. [Online]. Available: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/cld.433

[6] E. Neri, P. Bemi, L. Faggioni, and C. Bartolozzi, "MSCT of the abdomen: Colon, rectum and CT colonography," in *Information Processing in Medical Imaging*. Springer, 2015, pp. 327–338.

[7] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer-aided polyp detection system for colonoscopy videos," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2015, pp. 327–338.

[8] 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*.

[9] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667102622000717

[10] Y. Fang, D. Zhu, J. Yao, Y. Yuan, and K.-Y. Tong, "ABC-Net: Area-boundary constraint network with dynamical feature selection for colorectal polyp segmentation," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11799–11809, May 2021.

[11] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2029–2040, Jun. 2021.

[12] M. V. L. Branch and A. S. Carvalho, "Polyp segmentation in colonoscopy images using U-Net-MobileNetV2," 2021, *arXiv:2103.15715*.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

[16] A. Lou, S. Guan, and M. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," *J. Med. Imag.*, vol. 10, no. 1, pp. 81–92, Feb. 2023.

[17] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6211–6220.

[18] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101897.

[19] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, X. Yang, L. Li, J. He, X. Yi, J. Liu, and X. Liu, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nature Biomed. Eng.*, vol. 2, no. 10, pp. 741–748, Oct. 2018.

[20] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*.

[21] J. Y. Lee, J. Jeong, E. M. Song, C. Ha, H. J. Lee, J. E. Koo, D.-H. Yang, N. Kim, and J.-S. Byeon, "Real-time detection of colon polyps during colonoscopy using deep learning: Systematic validation with four independent datasets," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, May 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

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

[25] Z. Qian, Y. Lv, D. Lv, H. Gu, K. Wang, W. Zhang, and M. M. Gupta, "A new approach to polyp detection by pre-processing of images and enhanced faster R-CNN," *IEEE Sensors J.*, vol. 21, no. 10, pp. 11374–11381, May 2021.

[26] L. Cai, R. Beets-Tan, and S. Benson, "An improved automatic system for aiding the detection of colon polyps using deep learning," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[29] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 1757–1761.

[30] J. Wan, B. Chen, and Y. Yu, "Polyp detection from colorectum images by using attentive YOLOv5," *Diagnostics*, vol. 11, no. 12, p. 2264, Dec. 2021.

[31] A. Bochkovskiy, C.-Y. Wang, H.-Y. Liao, F. Zhu, and R. Manduchi, "Yolov5: Improved real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2020, pp. 722–723.

[32] M. Souaidi, S. Lafraxo, Z. Kerkaou, M. El Ansari, and L. Koutti, "A multiscale polyp detection approach for GI tract images based on improved DenseNet and single-shot multibox detector," *Diagnostics*, vol. 13, no. 4, p. 733, Feb. 2023.

[33] 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," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[34] 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.

[35] 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. Surgery*, vol. 9, no. 2, pp. 283–293, Mar. 2014.

[36] 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 Model.* Cham, Switzerland: Springer, 2020, pp. 451–462.

[37] L. F. Sánchez-Peralta, A. Picón, F. M. Sánchez-Margallo, and J. B. Pagador, "Unravelling the effect of data augmentation transformations in polyp segmentation," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 15, no. 12, pp. 1975–1988, Dec. 2020.

[38] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168.

[39] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4702–4712.

[40] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–14.

[42] H. Taud and J. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*. Cham, Switzerland: Springer, 2018, pp. 451–455, doi: 10.1007/978-3-319-60801-3_27.

[43] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–8, no. 6, pp. 679–698, Nov. 1986.

[44] M. Bagheri, M. Mohrekesh, M. Tehrani, K. Najarian, N. Karimi, S. Samavi, and S. M. R. Soroushmehr, "Deep neural network based polyp segmentation in colonoscopy images using a combination of color spaces," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 6742–6745.

[45] D.-C. Cheng, W.-C. Ting, Y.-F. Chen, and X. Jiang, "Automatic detection of colorectal polyps in static images," *Biomed. Eng., Appl., Basis Commun.*, vol. 23, no. 5, pp. 357–367, Oct. 2011.

[46] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[47] M. Lin, C. Li, X. Bu, M. Sun, C. Lin, J. Yan, W. Ouyang, and Z. Deng, "DETR for crowd pedestrian detection," 2020, *arXiv:2012.06785*.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[49] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than CNNs?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[50] D. R. Cox, "Statistical significance tests," *Brit. J. Clin. Pharmacol.*, vol. 14, no. 3, pp. 325–331, 1982.

[51] H. Hsu and P. A. Lachenbruch, "Paired T test," in *Wiley StatsRef: Statistics Reference Online*, vol. 1. Hoboken, NJ, USA: Wiley, 2014, pp. 1–5.

ALAN CARLOS DE MOURA LIMA was born in Teresina, Piauí, in June 1987. He received the B.A. degree in computer science from the State University of Piauí (UESPI), in 2010, and the master's degree in computer science from the Federal University of Maranhão (UFMA), in 2019, where he is currently pursuing the Ph.D. degree in electrical engineering. His major field of study is computer science.

He has also authored or coauthored several publications. He is currently an Assistant Professor of informatics with the Federal Institute of Maranhão (IFMA), Rosário, Maranhão, Brazil. His research interests include medical images, image processing, machine learning, and artificial intelligence. He has received the Best Master's Thesis Terezinha Rêgo 2019 Award, FAPEMA, for his work. He holds a Computer Technician Certification from the Federal Institute of Piauí (IFPI), in 2008.

LISLE FARAY DE PAIVA was born in São Luís, Maranhão, Brazil, in 2000. She received the B.S. degree in computer science from the Federal University of Maranhão, in 2022. She is currently pursuing the joint master's degree in medical imaging and applications with the University of Bourgogne, the University of Cassino, and the University of Girona.

She has experience in the field of computer science, working mainly on the following topics: computer vision, machine learning, deep learning, and medical image processing. From 2018 to 2021, she was a Student Researcher with the Applied Computer Group.

GERALDO BRÁZ JR. received the degree in computer science, the master's degree in electrical engineering with an emphasis in computer science, and the Ph.D. degree in electrical engineering with an emphasis in computer science from the Federal University of Maranhão, in 2005, 2007, and 2014, respectively. He is currently an Associate Professor I with the Federal University of Maranhão, permanent in the Graduate Programs of Master in Computer Science (PPGCC/UFMA) and the Doctorate in Computer Science UFMA-UFPI Association. He has been holding CNPq Research Productivity Scholarship Level 2, since 2019. He has experience in the field of computer science, working mainly on the following topics: computer vision, machine learning, deep learning, and medical image processing.

JOÃO DALLYSON S. DE ALMEIDA received the degree in computer science and the master's and Ph.D. degrees in electrical engineering from the Federal University of Maranhão (UFMA), in 2007, 2010, and 2013, respectively. He is currently an Associate Professor I with UFMA. He coordinates the Vision and Image Processing Laboratory (VipLab-UFMA). He has experience in computer science, working mainly on the following topics: image processing, machine learning, ophthalmic medical images, and time series.

ARISTÓFANES CORRÉA SILVA received the B.S. degree in computer science and the master's degree in electrical engineering from the Federal University of Maranhão (UFMA), in 1995 and 1997, respectively, and the Ph.D. degree in informatics from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2004. Since 2008, he has been a Founding Member and the Leader of the Applied Computing Group, Federal University of Rio de Janeiro. He is currently a Full Professor with UFMA, working in research and teaching computer science, with a focus on the following subjects: image processing, computational vision, and applied artificial intelligence. His awards and honors include the 2010 and 2012 First Place FAPEMA Award in the Senior Research Category.

MIGUEL TAVARES COIMBRA (Senior Member, IEEE) was one of the founders of IT Porto, in 2007, its coordinator, from 2015 to 2019, and the Founder of the Interactive Media Group at this institute. He was the Director of the Master in Medical Informatics of the University of Porto, from 2014 to 2016, and was the Co-Founder of IS4H—Interactive Systems for Healthcare, a spin-off company of the University of Porto, in 2013. He has been a member of the Executive Board of the Faculty of Sciences, University of Porto, since 2019, the current Coordinator of the TEC4Health Line of INESC TEC, and the past Chair of the Portugal Chapter of the IEEE Engineering and Medicine Society (2017–2021). He is currently an Associate Professor (with Aggregation) with the Computer Science Department, Faculty of Sciences, University of Porto. He leads and participates in various projects involving engineering and medicine, namely cardiology and gastroenterology, with current and past collaborations with hospitals in Portugal, Brazil (Pernambuco, Paraíba, Minas Gerais, and São Paulo), Germany, and Sweden. The nearly 16 years of experience in biomedical signal processing and interactive systems for healthcare have led to the development and deployment of systems for the collection and analysis of auscultation signals, echocardiography image processing for rheumatic fever screening, monitoring of stress and fatigue of firefighters in action, endoscopy signal analysis for cancer detection, computer-assisted decision systems for capsule endoscopy, and quantification of 3D motion patterns for epilepsy, among others. He has more than 130 scientific publications, 25 of which in high-impact scientific journals (17 IEEE TRANSACTIONS) and has attracted and managed more than 2M€ in research funding, over a total of 15 research projects acting as a PI of the project (ten projects) or a co-PI of its institution (five projects).

ANSELMO CARDOSO DE PAIVA received the B.S. degree in civil engineering from the State University of Maranhão, in 1990, and the master's degree in civil engineering and the Ph.D. degree in informatics from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 1993 and 2001, respectively. In 2018, he was a Visiting Scholar with the Engineering Mechanics Department, Porto University Faculty of Engineering (FEUP), with a Postdoctoral Fellowship from CAPES. He is currently a Full Professor with the Federal University of Maranhão (UFMA), a Founding Member, and the Leader of the Applied Computing Group, Federal University of Rio de Janeiro. His research interests include image processing, computational vision, virtual and augmented reality, and natural language processing. His awards and honors include the 2011 and 2013 First Place FAPEMA Award in the Senior Research Category.