RESEARCH ARTICLE

ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation

NGUYEN THANH DUC[1], NGUYEN THI OANH [1], NGUYEN THI THUY [2], TRAN MINH TRIET [3,4], (Member, IEEE), AND DINH VIET SANG [1]

[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 10000, Vietnam
[2]Faculty of Information Technology and Software Engineering Laboratory, University of Science, VNU-HCM, Ho Chi Minh 70000, Vietnam
[3]University of Science, VNU-HCM, Ho Chi Minh 70000, Vietnam
[4]John Von Neumann Institute, Viet Nam National University Ho Chi Minh City, Ho Chi Minh 70000, Vietnam

Corresponding author: Dinh Viet Sang (sangdv@soict.hust.edu.vn)

ABSTRACT Identifying polyps is challenging for automatic analysis of endoscopic images in computer-aided clinical support systems. Models based on convolutional networks (CNN), transformers, and their combinations have been proposed to segment polyps with promising results. However, those approaches have limitations either in modeling the local appearance of the polyps only or lack of multi-level feature representation for spatial dependency in the decoding process. This paper proposes a novel network, namely ColonFormer, to address these limitations. ColonFormer is an encoder-decoder architecture capable of modeling long-range semantic information at both encoder and decoder branches. The encoder is a lightweight architecture based on transformers for modeling global semantic relations at multi scales. The decoder is a hierarchical network structure designed for learning multi-level features to enrich feature representation. Besides, a refinement module is added with a new skip connection technique to refine the boundary of polyp objects in the global map for accurate segmentation. Extensive experiments have been conducted on five popular benchmark datasets for polyp segmentation, including Kvasir, CVC-Clinic DB, CVC-ColonDB, CVC-T, and ETIS-Larib. Experimental results show that our ColonFormer outperforms other state-of-the-art methods on all benchmark datasets. Our code is available at: https://github.com/ducnt9907/ColonFormer.

INDEX TERMS Polyp segmentation, deep learning, encoder-decoder network, hierarchical multi-scale CNN, computer-aided diagnosis.

I. INTRODUCTION

Colorectal cancer (CRC) is among the most common types of cancer worldwide, causing over 694,000 fatalities each year [1]. The most common cause of CRC is colon polyps, particularly adenomas with high-grade dysplasia. According to a longitudinal study [2], every 1% increase in adenoma detection rate is linked to a 3% reduction in the risk of colon cancer. As a result, detecting and removing polyps early is critical for cancer prevention and treatment. Therefore, colonoscopy is regarded as the standard tool for detecting colon adenomas and colorectal cancer. In practice, over-burdened healthcare systems, particularly in low-resource

settings, may result in shorter endoscopy durations and more missed polyps. According to a literature review, the proportion of colon polyps missing during endoscopies could range from 20 to 47 percent [3]. This may lead to high associated risk factors in patients. Therefore, research in developing computer-aided tools to assist endoscopists in endoscopy procedures is an essential need.

Advancements in artificial intelligence and deep learning have changed the landscape of such systems. Attempts have been made to develop learning algorithms to deploy in computer-aided diagnostic (CAD) systems for the automatic detection and prediction of polyps, which could benefit clinicians in detecting lesions and lower the miss detection rate [4]–[6]. Deep neural networks have shown great potential in assisting colon polyp detection in several retrospective

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

investigations and diagnoses. A CAD system can support endoscopists in improving lesion detection rates, optimizing strategies during endoscopy for high-risk lesions, and increasing clinics' capacity while preserving diagnostic quality [7], [8].

Despite progress in machine learning and computer vision research, automatic polyp segmentation remains a challenging problem. Polyps are caused by abnormal cell growth in the human colon, meaning their appearances have strong relationships with the surroundings. Images of polyps come in various shapes, sizes, textures, and colors. In addition, the boundary between polyps and their surrounding mucosa is not always apparent during colonoscopy, especially in different lighting modes and in cases of flat lesions or unclean bowel preparation. These cause a lot of uncertainty for the learning models for polyp segmentation.

In recent years, the most widely used methods for image segmentation in general and polyp segmentation, in particular, are based on Convolutional Neural Networks (CNNs). Most segmentation models use a UNet-based architecture containing an encoder and a decoder, which are often built up from convolutional layers. Despite being widely used for segmentation tasks with impressive performance, CNNs pose certain limitations: They can only capture local information while ignoring spatial context and global information due to the limited receptive field. Furthermore, it was shown that CNNs act like a series of high-pass filters and favor high-frequency information.

Transformer [9] is a recently proposed deep neural network architecture that models the global interactions among input components using attention mechanisms. While initially designed to tackle natural language and speech processing problems, Transformers have significantly impacted computer vision in recent years. In contrast to CNNs, self-attention layers in Transformers work as low-pass filters, and they can effectively capture long-range dependency. Therefore, combining the strengths of convolutional and self-attention layers can increase the representation power of deep networks. Very recently, there has been fast-growing interest in using Transformers for semantic image segmentation [10]–[13]. These methods use well-known encoder-decoder architectures wherein Transformers and CNNs are combined in various settings. The works in [10]–[12] proposed Transformer-CNN architectures, in which a Transformer is used as an encoder, and a traditional CNN is used as a decoder. The hybrid architecture of Transformers and CNN has been proposed in [13], in which the decoder is a traditional CNN or a Transformer, while the encoder is a combination of CNN and Transformer layers.

Inspired by these approaches for modeling multi-scale and multi-level features, we propose a new Transformers-based network called ColonFormer. The main design of our ColonFormer also contains a transformer encoder and a CNN decoder, but our approach differs from the models mentioned above in several ways. In ColonFormer, the encoder is a hierarchically structured lightweight Transformer for learning multi-scale features. The decoder is a hierarchical pyramid structure capable of learning from heterogeneous data containing feature maps extracted from encoder blocks at different scales and subregions. Besides, a refinement module is proposed for further improving the segmentation accuracy on hard regions and small polyps.

Our main contributions include:

- A novel deep neural network, namely ColonFormer, that integrates a hierarchical Transformer and a hierarchical pyramid CNN in a unified architecture for efficient and accurate polyp segmentation;
- An improved refinement technique using a newly proposed residual axial attention module for feature fusion and smoothing aiming at improving the segmentation accuracy;
- A set of experiments on five standard benchmark datasets for polyp segmentation (Kvasir, CVC-Clinic DB, CVC-ColonDB, CVC-T, and ETIS-Larib) and comparisons of the effectiveness of ColonFormer to current state-of-the-art methods.

The rest of the paper is organized as follows. We provide a brief review of related works in Section II. The ColonFormer architecture is described in Section III. Section IV presents our experiments and results. Finally, we conclude the paper and highlight future works in Section V.

II. RELATED WORK

In this section, we briefly review common methods and techniques that have been developed for polyp segmentation. First, we review CNN architectures and their variants, especially UNet models, in medical image segmentation. Then we investigate the attention mechanism as a promising technique that boosts the capability of a deep neural network in learning feature representation. Finally, the Vision Transformer and its applications in polyp segmentation and medical image processing are investigated.

A. CONVOLUTIONAL NEURAL NETWORKS

CNNs are one of the most widely used deep neural network architectures, especially in computer vision. A deep CNN extracts features on multiple layers with increasing levels of abstraction. Low-level features with high resolutions represent spatial details, while high-level features with low resolutions represent rich semantic information. CNNs are especially powerful in image analysis as they can extract highly informative and valuable features.

UNet [14] is a pioneering CNN architecture for medical image segmentation. UNet consists of an encoder and a decoder. The encoder includes convolutional, pooling layers for feature extraction, and the decoder uses upsampling (or deconvolutional) and convolutional layers for yielding the final segmentation prediction. Later works attempted to improve UNet by introducing skip connections, which alleviate information loss caused by stacking multiple convolutional layers. However, retaining information from low-level may introduce noisy signals that degrade the performance.
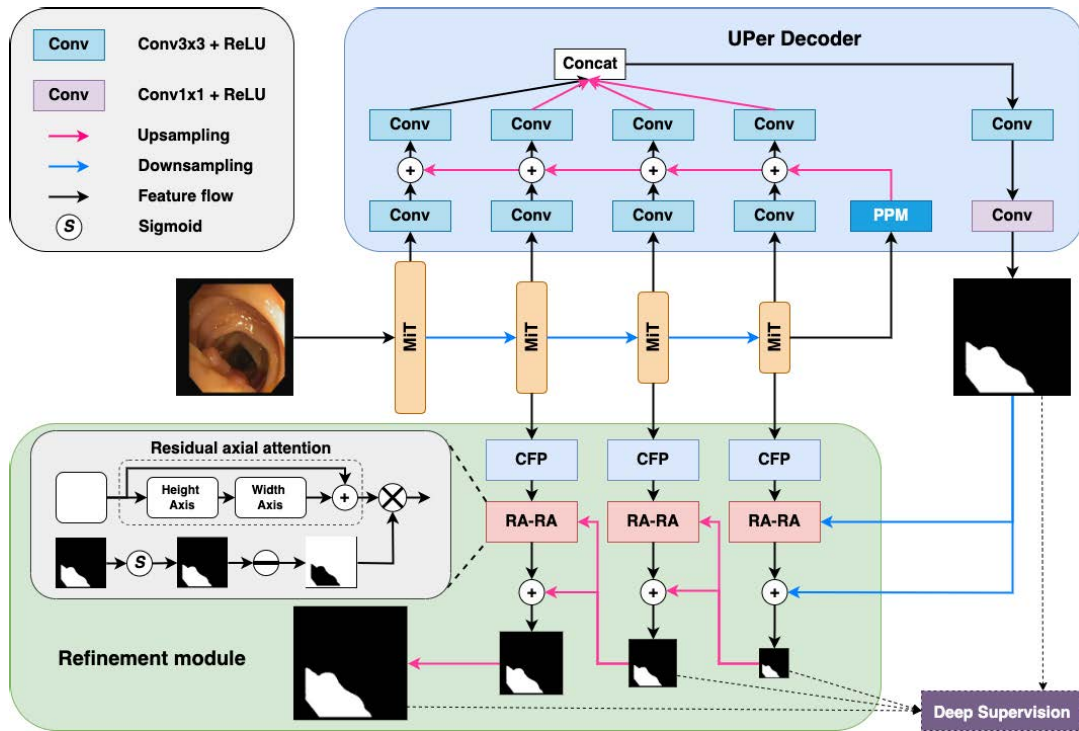
FIGURE 1. Overall architecture of our ColonFormer contains three components: an encoder, a decoder, and a refinement module. The encoder is based on the Mix Transformer backbone. The decoder starts with a pyramid pooling module (PPM), where its outputs are combined layer-wise with the output feature maps of the encoder at multiple levels to produce a global map. The refinement module aims to gradually refine the boundary of the global map to yield the final accurate segmentation. Besides this predicted output, the global map and two intermediate maps are also passed into the training loss in a deep supervision manner. Before calculating the training loss, all refined maps are upsampled back to the original image input size.

UNet variants such as UNet++ [15] and DoubleUnet [16] have achieved stellar results on segmentation benchmarks. UNet++ is constructed as an ensemble of nested UNets of varying depths, which partially share an encoder and jointly learn using deep supervision. DoubleUNet stacks two UNet blocks and uses ASPP [17], and SE blocks [18] to enhance the feature representation.

UNet encoders often use an existing pretrained architecture, also known as the backbone. Widely used backbones include VGG [19], MobileNet [20], ResNet [21], DenseNet [22] and so on. PraNet [23] uses Res2Net as the backbone, while AG-CuResNeSt [24] uses ResNeSt. Meanwhile HarDNet-MSEG [25], NeoUNet [26] and BlazeNeo [27] use HarDNet, an improvement of DenseNet to extract features.

B. ATTENTION MECHANISM

The attention mechanism is a widely used technique to help deep neural networks learn better feature representations, especially on highly variant inputs. Oktay *et al.* [28] proposed an Attention Gate module for UNet, which helps the model focus on necessary information while preserving computational efficiency. AG-ResUNet++ [29] integrates the attention gates with the ResNet backbone to improve UNet++ [15] for polyp segmentation. PraNet [23] uses the

Reverse Attention module [30], which enforces focus on the boundary between a polyp and its surroundings. In general, most CNNs and neural networks can benefit by adding attention modules. However, even with these attention mechanisms, CNNs are limited by the locality of convolution operations. This limitation makes them difficult to model natural long-range spatial dependencies between input segments.

C. VISION TRANSFORMER

Transformer [9] is a highly influential deep neural network architecture, originally proposed to solve natural language processing and similar problems. While the original Transformer architecture is not very well suited for image analysis, attempts have been proposed to leverage its advantages for computer vision through some modifications. Vision Transformer (ViT) [32] was the first successful application of Transformers for computer vision. ViT splits an image into patches and processes them as sequential tokens. This method greatly reduces computational costs and allows Transformers to work with large images feasibly.

A major issue of ViT is that it requires extensive datasets for training to remain effective while being severely limited when trained on small datasets. Such property hinders its usage in problems such as medical image analysis, including polyp segmentation, where data is typically scarce.
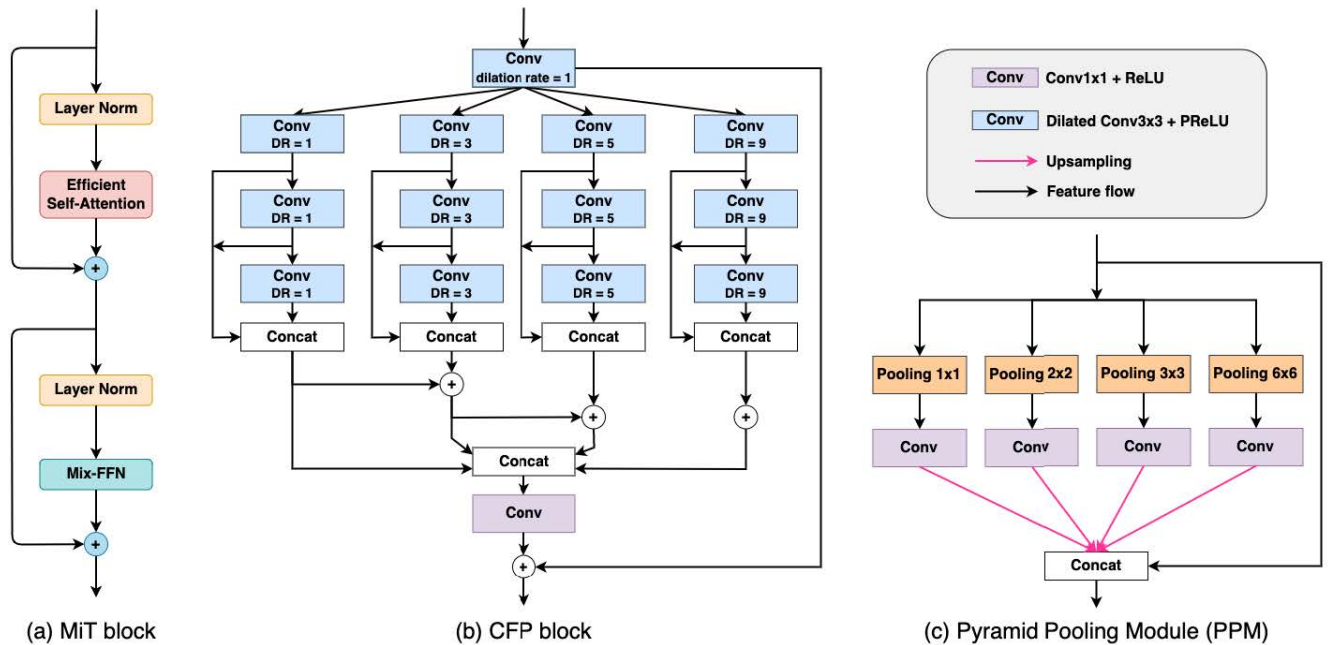
FIGURE 2. Architecture of three neural blocks used in our ColonFormer. The left block (a) is the Mix Transformer block [31]. The middle block (b) is the channel-wise feature pyramid block, where DR stands for dilation rate. The pyramid pooling module is shown on the right (c).

The Kvasir dataset, for example, contains just 1000 images and their corresponding ground truth, despite being the largest public image dataset of the gastrointestinal tract for polyp segmentation.

Recent works have attempted to further enhance ViT in several ways. DeiT [33] introduces a data-efficient training strategy combined with a distillation approach, which helps improve the performance when training on small datasets. Swin Transformer [10] redesigned the encoder for Transformers. The Swin Transformer encoder computes self-attention among a collection of adjacent patches within a sliding window. Patches are merged every few blocks, reducing the number of tokens and forming a multi-resolution token hierarchy similar to convolutional blocks. SegFormer [31] is another hierarchical Transformer design, where patches are merged with overlap and preserving local continuity around patches. The authors also introduced Efficient Self-Attention, a modified attention mechanism for reducing computational complexity, and Mix-FFN for better positional information.

Both TransUNet [34] and TransFuse [35] models have been developed based on Transformers for polyp segmentation and yielded promising results. TransUNet uses a Transformer-based network with a hybrid ViT encoder and upsampling CNN decoder. The Hybrid ViT stacks the CNN and Transformer together, leading to high computational costs. TransFuse addressed this problem by using a parallel architecture. Both models use the attention gate mechanism [28] and a so-called BiFusion Module. These components make the network architecture large and highly complex.

While there have been promising results in using Transformers to develop networks for polyp segmentation, there is plenty of room for improvement in this direction. Most notably, reduced network size and latency can greatly benefit downstream applications. In addition, improved accuracy and robustness can also be achieved with more optimized architectures. This paper seeks to design a Transformer-based architecture that achieves these goals.

III. COLONFORMER

Fig. 1 depicts the overall architecture of our proposed network, ColonFormer. The network consists of a hybrid encoder, a decoder, and a refinement module. We will describe each component in detail in the following sections.

A. ENCODER

A hierarchically structured model that can extract coarse-to-fine features at multi-scale and multi-level is desired for semantic segmentation. Our model uses Mix Transformer (MiT) proposed in [31] as the encoder backbone. MiT is a hierarchical Transformer encoder that can represent both high-resolution coarse and low-resolution fine features. Assume $X \in \mathbb{R}^{H \times W \times C}$ denotes the input image. MiT generates the CNN-like multi-level features F_i. The hierarchical feature map F_i has the resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where $i \in \{1, 2, 3, 4\}$ and C_i is ascending. The hierarchical feature representation is brought by the overlapped patch merging. After several Transformer blocks (Fig. 2a), a kernel with a stride smaller than kernel size is used to divide the feature map into overlapping patches. Such an overlapping patch

merging process ensures the local continuity around those patches.

Like other Transformer blocks, MiT blocks contain three main parts: Multi-head Self-Attention (MHSA) layers, Feed Forward Network (FFN), and Layer Norm. The MHSA is improved into Efficient Self-Attention, where the number of keys is decreased by a factor of R to reduce the computational complexity of self-attention layers. Another reason that we decide to choose MiT is the Mix-FFN. Instead of using the positional encoding (PE) as ViT, a 3×3 convolution kernel is integrated into FFN. Since the resolution of PE is fixed, it can not utilize the positional information of the pre-trained dataset like ImageNet when the resolution of the test images differs from the training ones. In such cases, ViT [32] suggests interpolating the PE, which can lead to a drop in accuracy. In contrast, arguing that convolutional layers are adequate for providing location information for Transformer, MiT directly uses a 3×3 convolutional layer for positional encoding. MiT has a series of variants, from MiT-B1 to MiT-B5, with the same architecture but different model's sizes. We name the variants of our model as **ColonFormer-XS**, **ColonFormer-S**, **ColonFormer-L**, **ColonFormer-XL**, **ColonFormer-XXL**, corresponding to different MiT backbone scales from MiT-B1 to MiT-B5, respectively. According to ablation study described later in Section IV-D, we found that ColonFormer-S and ColonFormer-L achieve the best results. Therefore, we mostly use ColonFormer-S and ColonFormer-L for comparison with other state-of-the-arts in all experiments except where it is specified otherwise.

B. DECODER

In order to further capture global context information, the feature maps extracted from the final block of the encoder are first processed by a Pyramid Pooling Module (PPM) [36] before being passed through the decoder blocks. The PPM simultaneously produces multi-scale outputs of the input feature map via a pyramid of pooling layers. The resulting feature maps, which form a hierarchy of features containing information at different scales and sub-regions, are then concatenated to produce an efficient prior global representation. Fig. 2c depicts the Pyramid Pooling Module in detail.

ColonFormer uses a decoder architecture inspired by UPerNet [37], which we denote as UPer Decoder. The decoder gradually fuses the prior global map produced by the PPM with multi-scale feature maps yielded by the MiT backbone. We suppose that applying convolutional layers to the feature maps of the MiT backbone is necessary since such layers can condense the information by emphasizing the coherence between neighboring elements, thus enhancing the resulting semantic map.

C. REFINEMENT MODULE

The decoder's outputs are further processed by a refinement module to achieve more precise and complete prediction maps. The refinement module consists of Channel-wise Feature Pyramid (CFP) module [38] (Fig. 2b) and our novel

Reverse Attention module enhanced by a new residual axial attention block for incremental correction of polyp boundary [30], [39].

In the parallel reverse attention network architecture [23], the global map is derived from the deepest CNN layer, so it does not have many structural details and hence can present only rough locations of the polyp pixels. The proposed strategy to recover precise location and label is to exploit complementary regions and details sequentially by removing previously estimated polyp regions from high-level side-output features, where the current estimation is up-sampled from the deeper layer. By using Reverse Attention, a coarse saliency map is guided to sequentially discover complement object regions and details by erasing the current predicted salient regions from side-output features. The current prediction is upsampled from its deeper layer. This erasing approach can refine the imprecise and coarse estimation into an accurate and complete prediction map. It was shown that self-attention layers in the MiT backbone work like low-pass filters. Therefore, we argue that using convolutional layers is essential for the refinement module since such layers favor high-frequency components and can provide richer edge information for the boundary correction.

Inspired by CaraNet [40], we use Channel-wise Feature Pyramid (CFP) to extract features from the encoder in multi-scale views. As depicted in Fig. 2b, the CFP module has $K = 4$ branches with different dilation rates that allow them to capture information at multiple scales. However, a direct concatenation of all branches could lead to some unwanted checkerboard or gridding artifacts that significantly impact the quality of the following boundary correction. In order to avoid this issue, the CFP module combines these branches step by step to build a final accurate feature map to correct the polyp boundaries.

CaraNet [40] also enhanced the Reverse Attention module by an axial attention block, which is a straightforward generalization of self-attention that naturally aligns with the multiple dimensions of the tensors. This module is supposed to filter the necessary information for the refinement process. However, axial attention may not always be good for the network since it can accidentally eliminate important edge information. Therefore, we propose to relax this mechanism using an additional residual connection, which allows the network to omit the axial attention layers when required and thus facilitates the learning process. The novel refinement module is called Residual Axial Reverse Attention (RA-RA). We experimentally found that utilizing the RA-RA module up to the finest feature map does not help refine the polyp boundary better. Hence, we propose to use just three RA-RA blocks, as shown in Fig. 1. The effectiveness of the RA-RA module is investigated in detail in Section IV-D.

D. LOSS FUNCTION

ColonFormer uses a compound loss combining the weighted focal loss and weighted IoU loss to train the model. The weighted focal loss is a distribution-based loss that treats

every pixel individually. In contrast, the weighted IoU loss is a region-based loss that considers the relationships between neighboring pixels.

Some image pixels can be easy to be learned and classified. However, some pixels, such as those on the edge regions, may be harder to learn. Thus, the model should pay more attention to challenging samples. In other words, some image pixels may be more important than others in contributing to the learning process. We represent the importance of pixel (i, j) by a weight β_{ij}. As suggested in [41], the weight β_{ij} for pixel (i, j) is defined as the difference between the center pixel and its neighbors:

$$\beta_{ij} = \left| \frac{\sum_{m,n \in \mathcal{N}_{ij}} g_{mn}}{|\mathcal{N}_{ij}|} - g_{ij} \right| \quad (1)$$

where $\mathcal{N}_{ij}$ represents the area of size 31×31 surrounding the pixel (i, j), and $g_{ij} \in \{0, 1\}$ is the true label of the pixel (i, j). A large value of β_{ij} indicates a pixel with considerable distinction from its vicinity, i.e., pixels at polyp edges. Such a weighting scheme enforces the model to focus more on the boundary regions.

Assume that p_{ij} is the prediction probability of a pixel (i, j) belonging to the polyp class. Let us define q_{ij} as:

$$q_{ij} = \begin{cases} p_{ij}, & \text{if } g_{ij} = 1 \\ 1 - p_{ij}, & \text{otherwise} \end{cases} \quad (2)$$

As polyp segmentation is a problem with highly imbalanced data, the focal loss is employed to deal with class imbalance during training. It integrates a modulating term in order to focus learning on hard pixels. The weighted focal loss is then defined as follows:

$$\mathcal{L}_{wfocal} = -\frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (1 + \lambda \beta ij)\alpha(1 - q_{ij})^{\gamma} \log(q_{ij})}{\sum_{i=1}^{H} \sum_{j=1}^{W} (1 + \lambda \beta_{ij})} \quad (3)$$

where α, γ are tunable hyperparameters.

The weighted IoU loss is defined as follows:

$$\mathcal{L}_{wiou} = 1 - \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (g_{ij} * p_{ij}) * (1 + \lambda \beta_{ij})}{\sum_{i=1}^{H} \sum_{j=1}^{W} (g_{ij} + p_{ij} - g_{ij} * p_{ij}) * (1 + \lambda \beta_{ij})} \quad (4)$$

where λ is a hyperparameter to adjust the impact of importance weights β_{ij}.

The total loss of our ColonFormer is calculated as:

$$\mathcal{L}_{total} = \frac{\mathcal{L}_{wfocal} + \mathcal{L}_{wiou}}{2} \quad (5)$$

The total loss in Eq. (5) is applied to train our model for multi-scale outputs as shown in Fig. 1. The final loss is the sum of all total losses computed at different output levels. Note that each output is upsampled back to the original size of the image's ground truth before the losses are evaluated.

IV. EXPERIMENTS

A. DATASETS

We perform experiments on the five popular benchmark datasets for polyp segmentation: Kvasir [47], CVC-Clinic DB [48], CVC-Colon DB [49], CVC-T [50], and ETIS-Larib Polyp DB [51]. Details of these datasets are described as follows:

- **The Kvasir dataset** is collected using endoscopic equipment at Vestre Viken Health Trust (VV), Norway. Images are carefully annotated and verified by experienced gastroenterologists from VV and the Cancer Registry of Norway. The dataset consists of 1000 images with different resolutions from 720×576 to 1920×1072 pixels.
- **The CVC-ClinicDB dataset** is a database of image frames extracted from colonoscopy videos. The dataset consists of 612 images with a resolution of 384×288 pixels from 31 colonoscopy sequences. The dataset was used in the training stages of the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection Challenge in Colonoscopy Videos.
- **The CVC-ColonDB dataset** is provided by the Machine Vision Group (MVG). The dataset consists of 380 images with a resolution of 574×500 pixels from 15 short colonoscopy videos.
- **The CVC-T dataset** is the test set of a more extensive dataset called Endoscene. CVC-T consists of 60 images obtained from 44 video sequences acquired from 36 patients.
- **The ETIS-Larib dataset** contains 196 high resolution (1226×996) colonoscopy images.

B. EXPERIMENT SETTINGS

We implement ColonFormer using the PyTorch framework. For a fair setting and comparison, we use the same parameters as [31] for the MiT backbone: kernel size $K = 7$, stride $S = 4$, padding size $P = 3$, and $K = 3, S = 2, P = 1$ to produce features with the same size as the non-overlapping process. Based on experiments in [41], [53], we use $\lambda = 5$, $\alpha = 0.25$ and $\gamma = 2$ for the losses in Eq. (3) and Eq. (4). Training is performed using Google Colab on virtual machines with 16GB RAM and an NVIDIA Tesla P100 GPU. Input images are resized to 352×352 for testing. In order to increase the model's robustness w.r.t image sizes, the training images are consequently scaled with a factor of $\{0.75, 1, 1.25\}$, respectively, and fed to the model for learning. None of the data augmentation techniques is used in the training phase.

We use six experiment setups to evaluate our method; each setup is described in detail below:

- **Experiment 1**: We use the same split as suggested in [23], where 90% of the Kvasir and ClinicDB datasets are used for training. The remaining images in the Kvasir and CVC-ClinicDB datasets and all images from CVC-ColonDB, CVC-T, and ETIS-Larib are used for testing.

TABLE 1. Performance comparison of different methods on the Kvasir, ClinicDB, ColonDB, CVC-T, and ETIS-Larib test sets. All results of ColonFormer are averaged over five runs.

Method	Kvasir		CVC-ClinicDB		CVC-ColonDB		CVC-T		ETIS-Larib	
	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU
UNet [14]	0.818	0.746	0.823	0.750	0.512	0.444	0.710	0.627	0.398	0.335
UNet++ [15]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
SFA [42]	0.723	0.611	0.700	0.607	0.469	0.347	0.297	0.217	0.467	0.329
PraNet [23]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
HarDNet-MSEG [25]	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613
CaraNet [40]	0.918	0.865	0.936	0.887	0.773	0.689	0.903	0.838	0.747	0.672
TransUNet [34]	0.913	0.857	0.935	0.887	0.781	0.699	0.893	0.824	0.731	0.660
TransFuse-L* [35]	0.920	0.870	**0.942**	**0.897**	_0.781_	0.706	_0.894_	_0.826_	0.737	0.663
ColonFormer-S (Ours)	**0.927**	**0.877**	_0.932_	0.883	**0.811**	_0.730_	_0.894_	_0.826_	_0.789_	_0.711_
ColonFormer-L (Ours)	_0.924_	_0.876_	_0.932_	_0.884_	**0.811**	**0.733**	**0.906**	**0.842**	**0.801**	**0.722**

TABLE 2. Performance comparison of different methods on 5-fold cross-validation of the CVC-ClinicDB and Kvasir datasets. All results are averaged over 5 folds.

Dataset	Method	mDice	mIOU	Recall	Precision
ClinicDB	UNet [14]	-	0.792	-	-
	MultiResUNet [43]	-	0.849	-	-
	ResUNet++ [44]	0.815 ± 0.018	0.736 ± 0.017	0.832 ± 0.018	0.830 ± 0.020
	DoubleUNet [16]	0.920 ± 0.018	0.866 ± 0.025	0.922 ± 0.027	0.928 ± 0.017
	DDANet [45]	0.860 ± 0.014	0.786 ± 0.017	0.858 ± 0.023	0.892 ± 0.014
	ColonSegNet [46]	0.817 ± 0.020	0.873 ± 0.024	0.926 ± 0.025	0.933 ± 0.014
	HarDNet-MSEG [25]	0.923 ± 0.020	0.873 ± 0.024	0.926 ± 0.025	0.933 ± 0.014
	PraNet [23]	0.933 ± 0.012	0.884 ± 0.015	0.940 ± 0.005	0.937 ± 0.016
	ColonFormer-S (Ours)	**0.948 ± 0.002**	**0.904 ± 0.004**	**0.958 ± 0.003**	**0.941 ± 0.004**
	ColonFormer-L (Ours)	_0.947 ± 0.002_	_0.903 ± 0.003_	_0.956 ± 0.002_	_0.942 ± 0.005_
Kvasir	UNet [14]	0.708 ± 0.017	0.602 ± 0.010	0.805 ± 0.014	0.716 ± 0.020
	ResUNet++ [44]	0.780 ± 0.010	0.681 ± 0.008	0.834 ± 0.010	0.799 ± 0.010
	DoubleUNet [16]	0.879 ± 0.018	0.816 ± 0.026	0.902 ± 0.027	0.894 ± 0.039
	DDANet [45]	0.860 ± 0.005	0.791 ± 0.004	0.876 ± 0.015	0.892 ± 0.018
	ColonSegNet [46]	0.676 ± 0.037	0.557 ± 0.040	0.731 ± 0.088	0.730 ± 0.080
	HarDNet-MSEG [25]	0.889 ± 0.011	0.831 ± 0.011	0.892 ± 0.015	0.926 ± 0.014
	PraNet [23]	0.883 ± 0.020	0.822 ± 0.020	0.897 ± 0.020	0.906 ± 0.010
	ColonFormer-S (Ours)	**0.924 ± 0.008**	**0.875 ± 0.010**	**0.941 ± 0.010**	**0.927 ± 0.008**
	ColonFormer-L (Ours)	_0.917 ± 0.006_	_0.865 ± 0.007_	_0.932 ± 0.007_	_0.926 ± 0.008_

- **Experiment 2**: 5-fold cross-validation on the CVC-ClinicDB and Kvasir datasets. Each dataset is divided into five equal folds. Each run uses one fold for testing and four remaining folds for training.
- **Experiment 3**: Cross-dataset evaluation with 3 training-testing configurations:
  1) CVC-ColonDB and ETIS-Larib for training, CVC-ClinicDB for testing;
  2) CVC-ColonDB for training, CVC-ClinicDB for testing;
  3) CVC-ClinicDB for training, ETIS-Larib for testing.

The first experiment compares our ColonFormer model with state-of-the-art CNN-based and Transformer-based networks using the same widely-used dataset configuration as suggested in [23]. The second experiment compares Colon-Former's learning ability to several recent polyp segmentation methods. Finally, the last experiment provides deeper insights into the generalization capability of ColonFormer and other benchmark models.

We use the Adam optimizer and cosine annealing scheduler with a learning rate of 1e-4. ColonFormer is trained in 20 epochs with a batch size of 8. The checkpoint of the last epoch is used for evaluation. Except for the second experiment with 5-fold cross-validation, we train ColonFormer five times, and the ColonFormer's results are averaged over five runs.

In addition, we perform a series of ablation studies to evaluate the effectiveness of each component in the proposed ColonFormer. All ablation studies are performed on the dataset configuration for Experiment 1.

C. COMPARISON WITH BENCHMARK MODELS

Table 1 describes the comparison results for Experiment 1. ColonFormer generally outperforms the benchmark models on most datasets. Notably, both ColonFormer-S and

TABLE 3. Performance comparison of different methods on cross-dataset configurations. All results are averaged over five runs.

| Train	Test	Method	mDice	mIOU	Recall	Precision
CVC-ColonDB + ETIS-Larib	CVC-ClinicDB	ResUNet++ [44]	0.406	0.302	0.481	0.496
		ColonSegNet [46]	0.427	0.321	0.529	0.552
		DDANet [45]	0.624	0.515	0.697	0.692
		DoubleUNet [16]	0.738	0.651	0.758	0.824
		HarDNet-MSEG [25]	0.765	0.681	0.774	0.863
		PraNet [23]	0.779	0.689	0.832	0.812
		ColonFormer-S (Ours)	**0.851**	**0.771**	**0.853**	*0.896*
		ColonFormer-L (Ours)	*0.847*	*0.770*	*0.844*	**0.902**
CVC-ColonDB	CVC-ClinicDB	ResUNet++ [44]	0.339	0.247	0.380	0.484
		DoubleUNet [16]	0.441	0.375	0.423	0.639
		DDANet [45]	0.476	0.370	0.501	0.644
		ResNet101-Mask-RCNN [52]	0.641	0.565	0.646	0.725
		ColonSegNet [46]	0.582	0.268	0.511	0.460
		HarDNet-MSEG [25]	0.721	0.633	0.744	0.818
		PraNet [23]	0.738	0.647	0.751	0.832
		ColonFormer-S (Ours)	**0.816**	**0.731**	**0.809**	**0.881**
		ColonFormer-L (Ours)	*0.804*	*0.723*	*0.794*	*0.877*
CVC-ClinicDB	ETIS-Larib	ResUNet++ [44]	0.211	0.155	0.309	0.203
		ColonSegNet [46]	0.217	0.110	0.654	0.144
		DDANet [45]	0.400	0.313	0.507	0.464
		ResNet101-Mask-RCNN [52]	0.565	0.469	0.565	0.639
		DoubleUNet [16]	0.588	0.500	0.689	0.599
		PraNet [23]	0.631	0.555	0.762	0.597
		HarDNet-MSEG [25]	0.659	0.583	0.676	0.705
		ColonFormer-S (Ours)	*0.723*	*0.635*	*0.797*	*0.731*
		ColonFormer-L (Ours)	**0.760**	**0.673**	**0.859**	**0.734**

(a) ROC curves
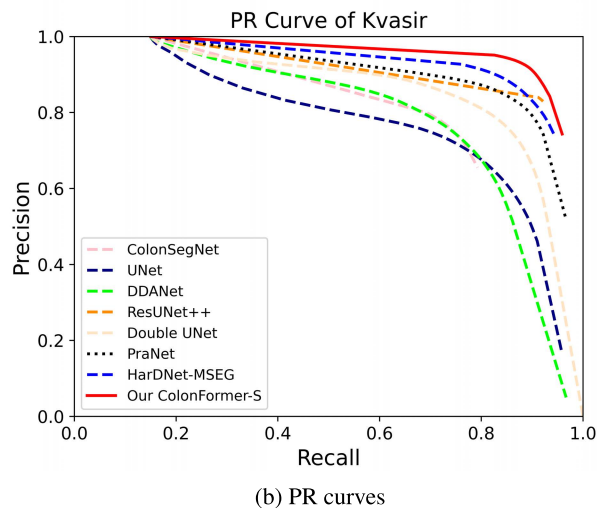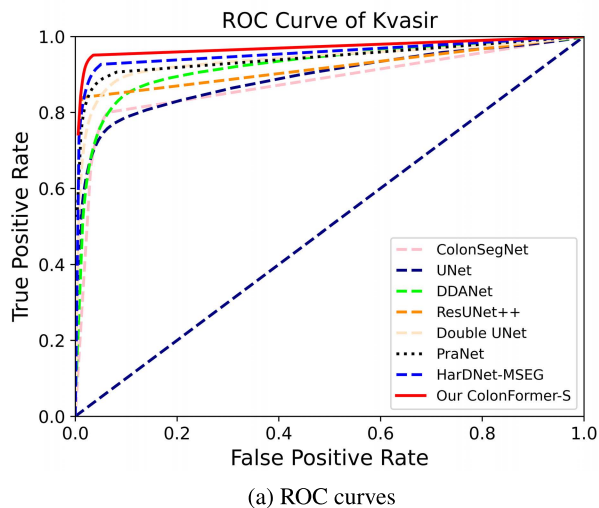
(b) PR curves

FIGURE 3. ROC and PR curves on the 5-fold cross-validation on the Kvasir-SEG dataset. All the curves are averaged over 5 folds.

ColonFormer-L outperform the second-best TransFuse-L* by 3% in mDice and 2.7% in mIOU on the ColonDB dataset. Compared to the second-best CaraNet on the ETIS-Larib dataset, ColonFormer-S achieves an improvement of 5.2% in mDice, and 4.8% in mIOU, while ColonFormer-L achieves an improvement of 6.4% in mDice and 5.9% in mIOU. The high capacity of ColonFormer-L seems more suitable for high-resolution images in the ETIS-Larib dataset. However, both ColonFormer-S and ColonFormer-L achieve roughly 1% lower metrics against TransFuse-L* on the CVC-ClinicDB dataset, whose images obtain very low resolution.
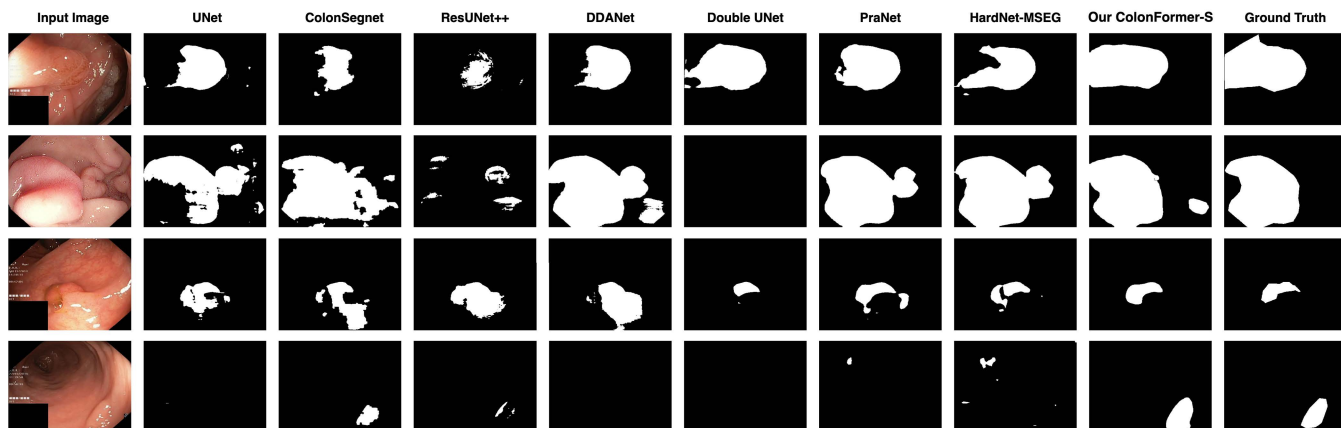
FIGURE 4. Qualitative result comparison of different models test on the first fold in the 5-fold cross-validation experiment on the Kvasir dataset.

Table 2 describes the comparison results for Experiment 2. We report both the average value and the standard deviation for each metric, which reflects the models' stability. One can see that both ColonFormer-S and ColonFormer-L outperform all other state-of-the-art models in mDice, mIOU, precision, and recall on both datasets. Notably, our ColonFormer is the most stable model on both datasets, achieving the lowest standard deviation for each evaluation metric.

Qualitative results for Experiment 2 are shown in Fig. 4. ColonFormer-S demonstrates much fewer wrongly predicted pixels in segmentation results than other models. ColonFormer-S also produces better ROC and PR curves than the benchmark models, as depicted in Fig. 3.

Table 3 describes the comparison results for Experiment 3. Overall, both ColonFormer-S and ColonFormer-L significantly outperform benchmark models on cross-dataset metrics. For the first configuration, ColonFormer-S outperforms the second-best PraNet by $7 - 8\%$ on all metrics. In the second configuration, ColonFormer-S continues to achieve a 5.8% improvement in precision, 7.8% improvement in mDice, and 9.4% improvement in mIoU over PraNet. For the third configuration, ColonFormer-L again shows its suitability to the ETIS-Larib dataset achieving a 10.1% improvement in mDice and 18.3% in recall over the second-best HardNet-MSEG. These are highly significant improvements, showing that our ColonFormer can generalize very well to new unseen data. Some result samples for this experiment are shown in Fig. 5. Similar to Fig. 4, one can see that ColonFormer yields better segmentation results than other state-of-the-arts.

Table 4 compares ColonFormer with other benchmark models in terms of size and computational complexity. One can see that our ColonFormer-S obtains competitive size and computational complexity compared to the most lightweight CNN-based models such as PraNet [23], and HardNet-MSEG [25]. Our ColonFormer-L is larger than most CNN-based neural networks but still more efficient than other Transformer-based methods in terms of GFlops.

TABLE 4. Number of parameters and GFLOPs of different methods.

Method	Parameters (M)	GFLOPs
PraNet [23]	32.55	13.11
HarDNet-MSEG [25]	33.34	11.38
CaraNet [40]	46.64	21.69
TransUNet [34]	105.5	60.75
TransFuse-L* [35]	-	-
SegFormer-B3 [31]	47.22	33.68
SegFormer-B3-Uper	46.61	20.99
ColonFormer-S (Ours)	33.04	16.03
ColonFormer-L (Ours)	52.94	22.94

D. ABLATION STUDIES
1) EFFECTIVENESS OF THE UPER DECODER
We firstly compare the original SegFormer-B3 [31] with MLP Decoder and another model called SegFormer-B3-Uper that replaces the original MLP decoder with the UPer Decoder. Both models use the MiT-B3 backbone in terms of the encoder.

Results are shown in the first two rows of Table 5. Both network versions show similar metrics across the test datasets, with slight variations of roughly 1%. However, one can see from Table 4 that UPer Decoder is also significantly less costly, requiring only 20.99 GFLOPs as opposed to MLP Decoder (33.68 GFLOPs). These results compel us to choose the UPer decoder for ColonFormer, which alleviates the high computational cost incurred with the use of the Transformer-based backbone.

2) EFFECTIVENESS OF THE REFINEMENT MODULE
We evaluate the performance of SegFormer-B3-Uper-ARA with the A-RA Refinement Module as in [40], and our ColonFormer-L with the adjusted Refinement Module as described in Section III-C. Results are shown in Table 5. Overall, incorporating the Refinement Module yields improvement across all datasets. Our ColonFormer-L also
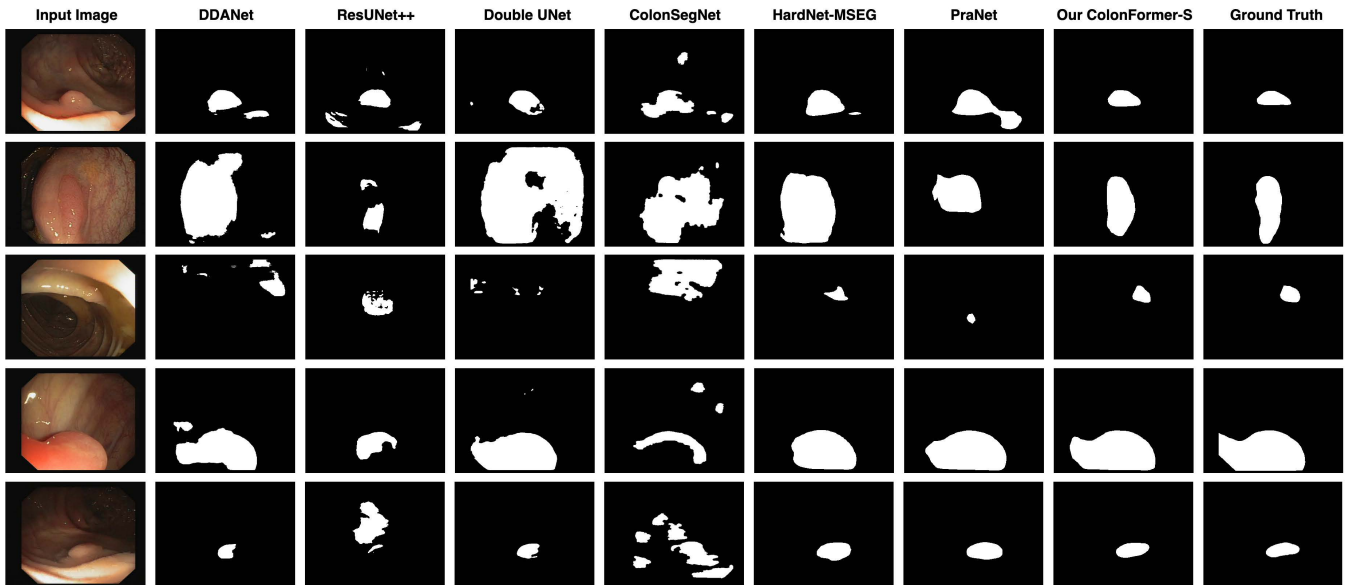
FIGURE 5. Qualitative result comparison using CVC-Colon for training and CVC-Clinic for testing.

TABLE 5. Ablation study on the effectiveness of different components. All results are averaged over five runs.

Method	Uper	A-RA	RA-RA	Kvasir		CVC-ClinicDB		CVC-ColonDB		CVC-T		ETIS-Larib	
				mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU
SegFormer-B3 [31]	—	—	—	0.920	0.866	*0.925*	*0.876*	0.806	0.726	0.905	0.840	*0.786*	*0.707*
SegFormer-B3-Uper	✓	—	—	0.916	0.864	0.924	0.876	*0.811*	0.731	0.900	0.832	0.784	*0.707*
SegFormer-B3-Uper-ARA	✓	✓	—	*0.922*	*0.872*	0.922	0.875	**0.812**	**0.734**	*0.903*	*0.837*	0.787	0.704
ColonFormer-L (Ours)	✓	—	✓	**0.924**	**0.876**	**0.932**	**0.884**	*0.811*	*0.733*	**0.906**	**0.842**	**0.801**	**0.722**

TABLE 6. Evaluation metrics for different variations of the MiT backbone. All results are averaged over five runs.

Method	Backbone	Kvasir		CVC-ClinicDB		CVC-ColonDB		CVC-T		ETIS-Larib	
		mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU
ColonFormer-XS	MiT-B1	0.913	0.859	0.926	0.876	0.784	0.700	0.879	0.808	0.758	0.679
ColonFormer-S	MiT-B2	**0.927**	**0.877**	**0.932**	*0.883*	*0.811*	0.730	0.894	0.826	0.789	0.711
ColonFormer-L	MiT-B3	*0.924*	*0.876*	**0.932**	**0.884**	*0.811*	*0.733*	**0.906**	**0.842**	**0.801**	**0.722**
ColonFormer-XL	MiT-B4	0.920	0.870	0.923	0.875	**0.814**	**0.735**	*0.905*	*0.840*	*0.795*	*0.715*
ColonFormer-XXL	MiT-B5	0.920	0.872	0.924	0.876	0.802	0.724	0.899	0.831	0.776	0.700

yields superior performance than SegFormer-B3-Uper-ARA on the Kvasir, CVC-ClinicDB, CVC-T, and most significantly the ETIS-Larib datasets, while slightly underperforming on the CVC-ColonDB dataset.

3) EFFECTIVENESS OF THE MIT BACKBONE

The Mix Transformer (MiT) [31] backbone has several variations ranging from MiT-B0 to MiT-B5. Accordingly, our ColonFormer also have different variations, including ColonFormer-XS, ColonFormer-S, ColonFormer-L, ColonFormer-XL, ColonFormer-XXL, respectively. Table 6 shows our comparison between all variations of Colon-Former. Overall, ColonFormer-S and ColonFormer-L yield the best average results across our test datasets.

V. CONCLUSION

This paper proposes a novel deep neural network architecture called ColonFormer for colon polyp segmentation. Our model leverages both the advantages of Transformers and CNNs architectures to learn a powerful multi-scale hierarchical feature representation. We also enhance the reverse attention with axial attention by relaxing it with a residual connection. The refinement module allows the network to incrementally correct the polyp boundary from a coarse global map produced by the decoder. Our extensive experiments show that ColonFormer significantly outperforms existing state-of-the-art models on popular benchmark datasets.

In future works, we will investigate lightweight or sparse self-attention layers to reduce the computational complexity.

In addition, other types of architectures for combining Transformers and CNNs can also be exploited.

REFERENCES

[1] J. Bernal *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Feb. 2017.

[2] D. A. Corley, C. D. Jensen, A. R. Marks, W. K. Zhao, J. K. Lee, C. A. Doubeni, A. G. Zauber, J. de Boer, B. H. Fireman, J. E. Schottinger, V. P. Quinn, N. R. Ghai, T. R. Levin, and C. P. Quesenberry, "Adenoma detection rate and risk of colorectal cancer and death," *New England J. Med.*, vol. 370, no. 14, pp. 1298–1306, 2014.

[3] A. Leufkens, M. G. H. Van Oijen, F. P. Vleggaar, and P. D. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[4] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.

[5] G. Zhou, X. Liu, T. M. Berzin, J. R. G. Brown, L. Li, C. Zhou, Z. Guo, L. Lei, F. Xiong, Y. Pan, and P. Wang, "951e—A real-time automatic deep learning polyp detection system increases polyp and adenoma detection during colonoscopy: A prospective double-blind randomized study," *Gastroenterology*, vol. 156, no. 6, p. 1511, May 2019.

[6] S. Kudo, Y. Mori, M. Misawa, K. Takeda, T. Kudo, H. Itoh, M. Oda, and K. Mori, "Artificial intelligence and colonoscopy: Current status and future perspectives," *Digestive Endoscopy*, vol. 31, no. 4, pp. 363–371, Jul. 2019.

[7] P.-J. Chen, M.-C. Lin, M.-J. Lai, J.-C. Lin, H. H.-S. Lu, and V. S. Tseng, "Accurate classification of diminutive colorectal polyps using computer-aided analysis," *Gastroenterology*, vol. 154, no. 3, pp. 568–575, 2018.

[8] R. Bisschops, J. E. East, C. Hassan, Y. Hazewinkel, M. F. Kaminski, H. Neumann, M. Pellisé, G. Antonelli, M. B. Balen, E. Coron, G. Cortas, M. Iacucci, M. Yuichi, G. Longcroft-Wheaton, S. Mouzyka, N. Pilonis, I. Puig, J. E. van Hooft, and E. Dekker, "Advanced imaging for detection and differentiation of colorectal neoplasia: European society of gastrointestinal endoscopy (ESGE) guideline—Update 2019," *Endoscopy*, vol. 51, no. 12, pp. 1155–1179, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[10] 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)*, Oct. 2021, pp. 10012–10022.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[12] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[13] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[14] 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.

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[16] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[17] L.-C. 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.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–14.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[21] 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.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Jun. 2017, pp. 4700–4708.

[23] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273.

[24] D. V. Sang, T. Q. Chung, P. N. Lan, D. V. Hang, D. Van Long, and N. T. Thuy, "AG-CUResNeSt: A novel method for colon polyp segmentation," 2021, *arXiv:2105.00402*.

[25] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "HarDNet-MSEG: A simple encoder–decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS," 2021, *arXiv:2101.07172*.

[26] P. N. Lan, N. S. An, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and V. D. Sang, "NeoUNet: Towards accurate colon polyp segmentation and neoplasm detection," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2021, pp. 15–28.

[27] N. S. An, P. N. Lan, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and D. V. Sang, "BlazeNeo: Blazing fast polyp segmentation and neoplasm detection," *IEEE Access*, vol. 10, pp. 43669–43684, 2022.

[28] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[29] N. B. Hung, N. T. Duc, T. Van Chien, and D. V. Sang, "AG-ResUNet++: An improved encoder–decoder based method for polyp segmentation in colonoscopy images," in *Proc. Int. Conf. Comput. Commun. Technol. (RIVF)*, Aug. 2021, pp. 1–6.

[30] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. ECCV*, 2018, pp. 234–250.

[31] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

[32] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021. [Online]. Available: https://dblp.org/db/conf/iclr/iclr2021.html

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021, pp. 10347–10357.

[34] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[35] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 14–24.

[36] 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. 2881–2890.

[37] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. ECCV*, 2018, pp. 418–434.

[38] A. Lou and M. Loew, "CFPNET: Channel-wise feature pyramid for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1894–1898.

[39] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.

[40] A. Lou, S. Guan, H. Ko, and M. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," 2021, *arXiv:2108.07368*.

[41] J. Wei, S. Wang, and Q. Huang, "F³Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI*, vol. 34, 2020, pp. 12321–12328.

[42] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 302–310.

[43] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.

[44] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 1–17.

[45] N. K. Tomar, D. Jha, S. Ali, H. D. Johansen, D. Johansen, M. A. Riegler, and P. Halvorsen, "DDANet: Dual decoder attention network for automatic polyp segmentation," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 307–314.

[46] 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.

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

[48] 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.

[49] 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.

[50] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *J. Healthcare Eng.*, vol. 2017, pp. 1–9, Jul. 2017.

[51] 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.

[52] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor CNN always perform better?" in *Proc. 13th Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, May 2019, pp. 1–6.

[53] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

NGUYEN THI OANH received the Ph.D. degree in computer science from Nancy 2 University, France, in 2010. She is the (co)author of scientific papers and participated in research and development projects funded by different national and international associations such as NAFOSTED, NICOP, AFORS, VINIF, and HUST. Her current research interests include image/video representation for content-based image retrieval, semantic segmentation, and human action recognition. Further, she is a member of the Vietnamese Association for Pattern Recognition (VAPR). She also joins the Program Committee and a Reviewer of many national and international conferences and journals such as *IET Computer Vision*, IEEE ACCESS, *Journal of Science and Technology*, ACML, PAKDD, KSE, SOICT, and MAPR.
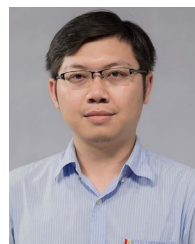

NGUYEN THI THUY received the Ph.D. degree in computer science from the Graz University of Technology, Austria, in 2009. She has more than ten years of research experience in computer vision, machine learning, and pattern recognition. She is the author or coauthor of more than 70 scientific papers and patents. She has been a principal investigator and key member of a number of research projects in computer vision, machine learning, and applications. She also joins the Program Committee and a reviewer of a number of international conferences and journals.

TRAN MINH TRIET (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Science, VNU-HCM, in 2001, 2005, and 2009, respectively. In 2001, he joined the University of Science. He was a Visiting Scholar with the National Institutes of Informatics (NII), Japan, from 2008 to 2010, and the University of Illinois at Urbana-Champaign (UIUC), from 2015 to 2016. His research interests include cryptography and security, computer vision and human–computer interaction, and software engineering. He is currently the Vice President of the University of Science, VNU-HCM, and the Director of the John Von Neumann Institute, VNU-HCM. He is also the Membership Development, Student Activities Coordinator of IEEE Vietnam. He is also a member of the Advisory Council for Artificial Intelligence Development of Ho Chi Minh City, and the Vice President of the Vietnam Information Security Association (VNISA), South Branch.

NGUYEN THANH DUC received the bachelor's degree in information and communication technology from the Hanoi University of Science and Technology (HUST). His research interests include computer vision and deep learning. He achieved Third Prize in Vietnam National Olympiad in Informatics, in 2016 and 2017.

DINH VIET SANG received the Ph.D. degree in computer science from the Dorodnitsyn Computing Centre of the Russian Academy of Sciences (CCRAS), in 2013. He is currently working at the Faculty of Computer Science, School of Information and Communication Technology (SoICT), Hanoi University of Science and Technology (HUST), Vietnam. He is now the Deputy Managing Director of the International Research Center for Artificial Intelligence (BK.AI), HUST. He has more than ten years of research experience in computer vision and machine learning and has published more than 50 publications. His research interests include computer vision, machine learning, and deep learning. He is the first NVIDIA Deep Learning Institute (DLI) Ambassador in Vietnam.

• • •