Contour-Aware Polyp Segmentation in Colonoscopy Images Using Detailed Upsampling Encoder-Decoder Networks

NGOC-QUANG NGUYEN, DUC MY VO, AND SANG-WOONG LEE, (Senior Member, IEEE)
Pattern Recognition and Machine Learning Laboratory, Gachon University, Seongnam 13120, South Korea
Corresponding author: Sang-Woong Lee (slee@gachon.ac.kr)

This work was supported by the Gyeonggi-do Regional Research Center (GRRC) Program of Gyeonggi Province through the Analysis of Behavior Based on Senior Life Log under Grant GRRC-2017-B01.

ABSTRACT Colorectal cancer has become one of the most common cause of cancer mortality worldwide, with a five-year survival rate of over 50%. Additionally, the potential of some common polyp types to progress to colorectal cancer is considered high. Colonoscopy is the most common method for finding and removing polyps. However, during colonoscopy, a significant number of polyps is missed as a result of human error mistakes. Thus, this study was primarily motivated by the need to obtain an early and accurate diagnosis of polyps detected in colonoscopy images. In this paper, we propose a new polyp segmentation method based on an architecture of multi-model deep encoder-decoder networks called MED-Net. Not only does this architecture obtain multi-level contextual information by extracting discriminative features at different effective fields-of-view and multiple image scales, it also can substantially do upsample more correctly to produce better prediction. It is also able to capture more accurate polyp boundaries by using multi-scale effective decoders. Moreover, we also present a complementary strategy for improving the method’s segmentation performance based on a combination of a boundary-aware data augmentation method and an effective weighted loss function. The purpose of this strategy is to allow our deep learning network to sequentially focus on poorly defined polyp boundaries, which are caused by the non-specular transition zone between the polyp and non-polyp regions. To provide a general view of the proposed method, our network was trained and evaluated on four well-known dataset CVC-ColonDB, CVC-ClinicDB, ASU-Mayo Clinic Colonoscopy Video Database, and ETIS-LaribPolypDB. Our results show that our MED-Net significantly outperforms state-of-the-art methods.

INDEX TERMS Polyp segmentation, hierarchical collaborative representation-based classification, local ternary patterns, deep convolutional neural network, colorectal segmentation.

I. INTRODUCTION
According to the report of the American Cancer Society [50], the number of newly diagnosed colorectal cancer cases in the United State was approximately 143460 in 2012 and this number is increasing quite rapidly every year. Besides, in the study provided by the Spanish Association against Cancer, between 28,500 and 33,800 new cases are diagnosed each year and around 20,000 and 14,000 in men and women, respectively, being detected [46]. More detailedly, colorectal cancer is fourth leading cause of cancer deaths worldwide. Thus, the objective of early colorectal cancer detection is to increase the survival rate and improve cancer control systems [47].

In the last two decades, colonoscopy technology has made noteworthy contributions to the development of diagnosis systems of colorectal diseases [59]. Colonoscopy is used to screen patients to detect inflamed tissue and abnormal growths such as tiny polyps. Since the potential of some common polyp types to progress to colorectal cancer is considered high, colonoscopy is a useful method in that it enables physicians to excise polyps and analyze them to detect signs of colorectal cancer. Moreover, according to the danger of polyps which is frequently fatal if discovered in the late stages. Besides, colorectal cancer researchers claim that these polyps frequently do not cause symptoms due to their size.
Thus, colonoscopy is highly recommended as one of the best diagnostic and therapeutic tools for detecting and preventing colorectal cancer. However, the sensitivity and specificity are not 100% and physicians can therefore miss 17% to 28% of colorectal polyps during colonoscopy [50]. Although some patients have more than one diagnostic test, the polyps in their colon are not detected. Undetected polyps are usually smaller than 9 mm in diameter, a size that endoscopists cannot easily see and detect. An additional reason why some polyps are not detected is that they are located in the danger area of the left colon or even behind a fold. They may also be too flat and blurred in appearance to allow them to be observed visually. Hence, more than 60% of colorectal cancer cases are attributed to missed polyps [37]. These polyps can develop to malignant tumors and therefore missing them can increase the risk of the patient developing cancer. Besides, the field of semantic segmentation has recently made remarkable contributions to the development of scene understanding and object recognition [57]. Thus, the objective of this study was to develop an automatic polyp segmentation tool that can support clinical endoscopists to detect tiny and flat polyps more effectively.

Computer-aided polyp segmentation systems can reduce the endoscopic miss rates of colorectal polyps and help clinical endoscopists localize the most complex colon polyps [23]. Such a system also supports the diagnosis procedure and can reduce the operation cost significantly. Based on this computer-aided system, more reliable classification and segmentation methods have been developed for the assessment of different polyp types [28]. These methods can automatically segment polyps regions into different categories and their performance rate is high. Therefore, many researchers have developed fast and precise polyp segmentation algorithms for providing early indications of colorectal cancer. However, their results are still far from meeting clinical requirements. The failures are related mainly to the appearance of polyps which have a large variety of sizes and shapes. Moreover, the non-specular transition zone between a polyp and its surrounding area does not show a significant change in texture or color that would enable endoscopists to distinguish it from all other non-polyp regions. Furthermore, the boundaries of wrinkles and folds in the colon are similar to those of polyps citehoerter2020artificial. In order to solve these main problems, we focused on building a deep convolutional neural network to generate discriminative features that are focused on polyp boundary regions and tiny polyps.

Over the last few years, deep fully convolutional neural networks (FCNNs) [12], [36], [42], [57] have led to dramatic developments in semantic segmentation research can be used to recognize and understand the content of an image at the pixel level [65], [66]. (FCNNs) [36] have also led to a dominant research direction for improving polyp segmentation because of their computational efficiency for discriminative feature extraction. They can automatically generate segmentation maps for all polyps of any sizes. Ronneberger et al. [45] improved FCNNs by introducing a new deep learning network, namely’s U-Net, which integrates an FCNN into an encoder-decoder architecture. This network achieved noteworthy results for segmenting biomedical images. Chen et al. [13] proposed investigating contour information to simultaneously segment and separate clustered objects. Because the contour features are emphasized, the objects are separated much better, and thus, the segmentation performance is significantly improved [21]. Their results show the important role of a model that learns contours in cell, tumor or organ segmentation. These approaches inspired us to find a better solution by utilizing the advantages of a model that explicitly learns polyp boundaries to achieve accurate polyp segmentation. However, the polyp segmentation task has its own challenges that are related to a polyp’s nature and environment. These challenges motivated us to develop a novel multi-model deep encoder-decoder network, called MED-Net that is focused on tackling two main issues for polyp segmentation.

The first issue of polyp segmentation concerns polyps’ wide variety of shapes, sizes, colors, and textures. First, the polyp size has a direct relation with miss rates of colorectal examinations, because doctors usually cannot easily evaluate small adenomas that are smaller than 9 mm in diameter. In addition, the estimation of the physical size of polyps is not easy to estimate, owing to the fact that it depends on the distance between the polyps and the colonoscopy camera [18]. If the doctor accidentally misses a small polyp, as mentioned above it can later develop into dangerous colorectal cancer. Second, colon polyps can appear in different shapes, for example, as radially pretubarant or flat polyps. Flat polyps are frequently attached to the colon wall, which makes them difficult to detect, because their boundaries are not clear in many cases. Finally, and not least important, polyps may appear in different shades and colors because of the uncontrolled illumination from the camera moving in the colon. These differences render computer-aided polyp detection algorithms considerably less effective in real environments. Thus, our proposed MED-Net is designed such that its inputs are multi-resolution images. It consists of a cascade architecture of dilated convolutions and includes an effective decoder module. This network architecture was inspired by a previous deep encoder-decoder network [16], the DeepLabv3+ [14] network and our previous work [38] with the advantage of upsampling method by Tian et al. [54] for segmenting objects. First, the cascade architecture of dilated convolutions is used at the end of our network to extract multi-scale context information in local regions without requiring an increased number of training parameters [67]. This architecture can also effectively recover detailed information related to polyp boundaries that are lost when the data pass through many convolution and pooling layers. Second, because the size and shape of polyp can appear in different images, we apply multi-scale input images to the same deep convolution neural network to search multi-scale objects [56]. These two techniques in combination can enlarge the receptive field size without resulting in a loss of important information.
Finally, the decoder module gradually recovers sharp polyp boundaries.

The second issue of polyp segmentation concerns the poorly defined polyp boundaries caused by the non-specular transition zone between the polyp and non-polyp regions. Thus, the improved U-Net network proposed by Chen et al. [13] for gland segmentation inspired us to find a better segmentation method for polyps by exploring the complementary features of their boundaries. However, gland contours are very distinct and can be easily predicted and segmented by Chen et al.’s improved U-Net network, whereas polyp boundaries are poorly defined and can render this network considerably less effective in the colon environments. For this reason, instead of masking polyp boundaries by fixed contours to explore the complementary information, as mentioned in [13], we randomly mask polyp boundaries and their neighboring regions, which do not differ widely. These masked regions provide important information for segmenting polyps precisely. Moreover, our proposed augmentation technique allows hundreds of boundary-aware polyp patterns to be generated randomly from each training polyp image. Thus, instead of extracting only one contour for each training image, as presented in [13], our deep learning network can focus on extracting the most important features in many different parts of the polyp boundaries for each training image. In addition, so that our proposed network is focused on extracting useful features for segmenting polyp boundaries, we propose using two boundary-aware loss functions for training our polyp boundary segmentation method. The first loss function is a new adaptive weighted loss function for focusing on learning the important parts of a polyp such as its boundaries. The second loss function is an attention-based loss function that allows the network to focus more on the most relevant local regions that include polyps and their boundaries. This function is computed based on minimizing the Euclidean distance between the polyp ground truth and its predicted regions. As a result, the predicted polyp boundary can coincide with its ground truth. After the training processes, our encoder-decoder network can achieve a considerably better accuracy rate because of the effectiveness of these loss functions.

When it comes to the third issue of polyp segmentation, recovering the prediction resolution is significantly considered as a big obstacle for researchers to achieve more accurate predictions. Furthermore, by observing the clear and meticulous result, clinical doctors can give wiser and better decision to increase the surviving rate of victims. In order to provide better prediction mask, instated of giving prediction at each pixel based on a local receptive field and parameters are shared at different spatial locations, Liu et al. [35] proposed the fully-connected fusion which used fc layers to yield different properties. Besides, the former are location sensitive since predictions at different spatial locations. Hence, fc layers have the ability to suit to these locations. Moreover, this predictions which are produced at each location is made with global information. Fully convolutional networks (FCNNs) [36] have achieved great success in dense pixel prediction applications such for semantic segmentation task. Although the sharing convolutional computation mechanism makes training and inference computationally very efficient, it still has drawbacks due to several stages of strided convolutions [60]. The later might cause to lose fine image structure information and poor prediction. To overcome the shortcoming of simple upsampling method, Chen et al. [14] applies atrous convolutions to achieve large receptive fields while still maintaining a higher-resolution feature map. The encoder part of DeepLabv3+ is used to extract rich information features which are meaningful for medical image especially in colorectal cancer segmentation and the decoder fuses low-level features to capture the fine-grained information lost by convolution and pooling operations in CNNs. To produce best performance, Chen et al. [14] reduced the overall strides of its encoder by four times and fused features of downsample ratio = 4. As a result, this method can be a catalyst for lower performance of a combination of features to be aggregated in the decoder. By dint of applying the simple bilinear upsampling, Chen et al. [14] has limited capability in recovering the pixel-wise prediction accurately. The former does not take into account the correlation among the prediction of each pixel since it is data independent [54]. Furthermore, bilinear is oversimple and has an inferior upper bound in terms of reconstructing. Besides, because of massive computation in both encoder and decoder path of the common networks usually give a slow inference [4]. Therefore, these flaws inspired us to propose a better method in helping doctor purpose. First, we adopted DeepLabv3+ as encoder which employs multiple atrous convolution [24], [41] to extract features in numerous resolution images and later we combined with a new technique of upsampling, and finally, with two loss functions to produce best performance.

We tested our proposed algorithm and its competitors on a challenging dataset to evaluate the performance of our polyp segmentation method. The results of extensive experiments show that our algorithms significantly outperform state-of-the-art algorithms. In particular, our paper presents several novel properties of our proposed method:

- We propose a novel encoder-decoder architecture which can recover the full-resolution prediction by applying data-dependent upsampling method, namely MED-Net, to extract the most useful visual features from multi-scale image inputs.
- We introduce a new boundary-focused data augmentation method for randomly generating a high number of boundary-aware polyp patterns from each training image. This method contributes to the improvement of MED-Net.
- We propose a new adaptive weighted loss function to boost the segmentation performance of MED-Net.
- We present an attention-based loss function that allows the network to focus more on the polyps and their boundaries. The combination of the loss functions leads to a
better performance, because the network can focus on learning iteratively polyp boundaries.

The remainder of this paper is organized as follows. In Section II, we briefly present some related state-of-the-art algorithms for polyp segmentation, which motivated our research. In Section III, we describe our proposed method of boundary-focused data augmentation and the processing of MED-Net. Besides, in this Section we also present about our two novel loss function. In Section IV, the experimental results obtained on a challenging databases are presented. We conclude this paper, outlining our intentions for future work, in Section V. The discussion also presented in Section VI for our shortcomings.

II. RELATED WORKS
In this section, we briefly review and discuss the state-of-the-art algorithms of colorectal cancer segmentation and related fields.

Recently, computer-aided diagnosis (CAD) has been rapidly developed to assist doctors in diagnosing patients faster and more accurately at many hospitals. In general, CAD is used to provide objective results to facilitate medical image diagnosis. One of the major CAD applications is for segmenting precisely cancer tumors, organs, and polyps. The CAD system can also be incorporated into the diagnostic process of polyp detection and segmentation to decrease inter-observer variation, effectively provide biopsy recommendations, and reduce unnecessary false-positive biopsies. There has been a number of research efforts in the field of polyp segmentation using CAD systems [6].

Conventional CAD systems use image processing methods for segmenting polyps automatically. Bernal et al. [8] employed a watershed algorithm to segment and classify polyp candidate regions. Jia [29] used a K-means clustering method to segment and localize polyp contours. For obtaining accurate boundaries and segmentations of polyps in each image, Breier et al. [11] presented two different approaches based on active contours and active rays, which can achieve a high segmentation performance. They also used the Chan-Vese segmentation method to improve their segmentation performance. Ganz et al. [20] proposed utilizing the contour information to segment polyps precisely. They used the Hough transform technique to detect the region of interest (ROI) before using an ultrametric contour map (UCM) for polyp segmentation. To increase the sensitivity of their method, they applied an ellipse fitting algorithm to extract the polyp boundaries in each image. This method can achieve both a high specificity and high sensitivity on some challenging datasets where the polyp is present in the center of the image. An ellipse fitting algorithm was also effectively used by Hwang et al. [26] to handle the false positive problem by setting valid polyp boundaries. Besides, Akbari et al. [2] proposed the polyp segmentation method which contained two main stages. In the first stage, the authors proposed candidate regions of probable polyp with FCNN-8S [2] network. Then in the second stage, Akbari et al. [2] used Otsu thresholding and select the largest connected component to segment polyp regions among all candidate regions.

Some researchers have developed polyp segmentation methods based on extracting features from image patches. Each patch is then classified into polyp and non-polyp classes. Tajbakhsh et al. [52] developed a new algorithm to extract oriented patches based on edge maps and classify them into polyp and non-polyp classes using a two-stage random forest classifier. To address the wide variety of polyp shapes, Tajbakhsh et al. [52] presented an ensemble of three convolution neural networks to classify input patches based on color, texture, and shape features, respectively.

Over the last few years, FCNNs [36] have been shown to be among the best algorithms for improving semantic segmentation, because of their computational efficiency for dense prediction. Most state-of-the-art methods for semantic image segmentation using FCNNs are based on the idea of adding convolutional layers at the end of networks instead of using any fully-connected layers. Chen et al. [13] combined an FCNN with a fully connected conditional random field to sharpen object boundaries and improve segmentation performance. However, one of the drawbacks of FCNNs is the low-resolution output responses. To improve for higher resolution images predictions, we improved on the FCNN by combining a cascade architecture of dilated convolutions with a deep network architecture for multi-resolution image inputs. Our network can obtain multi-scale information, recover the spatial resolution, and increase the field-of-view. Chen et al. [14] improved their network by using convolution filters and pooling operations at multiple rates and multiple effective fields-of-view to extract better multi-scale contextual information. They then used the second network to refine the segmentation regions, gradually recover the spatial information, and sharpen object boundaries. In the field of polyp segmentation, Zhang et al. [63] combined DCNN with a random forest classifier to segment potential polyp regions. Akbari et al. [2] also used an FCNN to find potential polyp candidates, and then segmented polyp regions by using the Otsu thresholding method which improves segmentation accuracy. In the first stage of post-processing they use Otsu thresholding method to change probability map which was resulted from FCNN-8S into a binary image and then find the largest connected component and consider it as the most probable location of polyp in the colonoscopy image.

More recently, in order to understand a scene, each piece of visual information has to be associated with an entity while considering the spatial information. An image segmentation can be done using Generative Adversarial Networks (GAN) [44]. GAN has shown outperformed results in many generative tasks to replicate the real-world rich content such as image field. It is inspired by theory: two models, a generator and a critic, are competing with each other while making each other stronger at the same time. A discriminator model estimates the probability which is given by sample coming from the real dataset. It works as a critic and is optimized to
tell the fake samples from the real ones. Meanwhile, a generator outputs synthetic samples given a noise variable input. It is trained to capture the real data distribution so that its generative samples can be as real as possible, or in other words, can trick the discriminator to offer a high probability. GAN-based segmentation methods have been proposed in the literature. In [40], the authors replace the traditional discriminator with a fully convolutional multiclass classifier. The classifier assigns to each input image pixel one label that corresponds to a semantic class or to fake/real mark. In this way, they use unlabeled images during the training process. Zhang et al. [64] used GAN to build a Task Driven Generative Adversarial Network (TD-GAN) to achieve simultaneous synthesis and parsing for unseen real X-ray images. By suing the strength of GAN their entire model pipeline does not ask for any annotations from the X-ray image domain.

III. PROPOSED METHOD

In this section, we describe our approach for polyp segmentation in detail. The entire training process of MED-Net is illustrated in Figure 1. Our approach includes three basic steps. First, we present novel augmentation methods that are able to increase effectively the amount of training images by generating additional images from those in the original limited training dataset. In this step, we also implement our proposed new boundary-focused data augmentation method for randomly generating a large number of boundary-aware polyp patterns. This method contributes to the improvement of polyp boundary segmentation. In the second step, all the augmented images are used to train in MED-Net to extract the most discriminative deep features of polyps from multi-scale image inputs. In the final step, we present an adaptive weighted loss function and an attention-based loss function that can effectively improve the segmentation performance of MED-Net.

A. BOUNDARY-FOCUSED DATA AUGMENTATION

Deep neural nets are powerful machine learning systems that tend to work well when trained on massive amounts of data [43]. Therefore, data augmentation is an effective technique for improving the performance of modern image segmentation. However, according to specific requirements of medical imaging field, the limitation of database is a huge wall for researcher to overcome [22], [55]. More specifically, for an automated polyp segmentation method using deep learning networks, data augmentation is an essential step for improving segmentation performance. Because the endoscopy procedures involving color calibration and controlling the camera’s movement are not consistent, the appearance of endoscopy images differs significantly across laboratories. The data augmentation step brings endoscopy images into an extended space that can cover all their variances. Since access to data is restricted because of privacy concerns, polyp segmentation networks were frequently trained with insufficient training dataset. The cancer classification performance was hindered by this lack of training data. Recent studies have demonstrated the effectiveness of data augmentation by generating additional data using the original limited training dataset for increasing the amount of training data. By augmenting training data, we can also reduce the over-fitting problem of training models. In this study, we used mainly geometric augmentation techniques including reflection, random cropping, translation, rotation, and elastic distortion which is introduced by Wong et al. [61]. Since the color of endoscopy images significantly varies across laboratories as a result of technicians’ varying technical skills, we applied an effective color constancy method, namely, gray world, which assumes the scene in an image, on average, is a neutral gray and the source of average reflected color is the color of the light. For this reason, the illuminant color cast can be effectively estimated by computing the average color and comparing it to gray values. In this algorithm, the illuminant colors are computed by using the mean of each color channel of the image.
Algorithm 1 Boundary-Focused Augmentation Algorithm

Input: Input image \( I \);
Polypl region \( S \);
Circle region \( S_{c} \);
Circle center \( C(x_{C}, y_{C}) \);
Minimum radius \( R_{\text{max}} \) from \( C(x_{C}, y_{C}) \);
Circle radius \( r_{i} \), \( r_{i} \subseteq \text{Randint} (R, R_{\text{max}}) \) with \( R_{\text{max}} \) is the shortest distance from circle center \( C(x_{C}, y_{C}) \) to the polyp boundary;

Output: Processed image \( I^{*} \);

While True do
\[ C(x_{C}, y_{C}) \subseteq \text{Randint} (S) ; \]
if \( R \leq R_{\text{max}} \) then
\[ r_{i} \subseteq \text{Randint} (R, R_{\text{max}}) \]
\[ S_{c} \leftarrow (x_{C}, y_{C}, r_{i}) ; \]
\[ I(S_{c}) \leftarrow \text{Randint}(0,128) ; \]
\[ I^{*} \leftarrow I ; \]
return False
end

segmentation methods are not able to segment the polyp boundaries. This results in a poor segmentation performance. Moreover, as previously mentioned the boundaries of wrinkles and folds in the colon are similar to those of a polyp and a polyp can be partly overlapped and hidden by these wrinkles and folds. To address the boundary appearance problem and improve the learning ability of our encoder-decoder network, we present a new boundary-focused data augmentation approach that can be combined with most existing deep convolutional neural networks to boost classification performance. In the training dataset, every image generates a number of augmented images by randomly selecting a circle region of an arbitrary diameter and an arbitrary circle center for each of its augmented images, provided that this center is located in the polyp region and the circle does not cover the whole polyp region. Each pixel within the circle region is set to a random value. Some typical examples of the results of the boundary-focused data augmentation approach are shown in Figure 2.

The entire procedure of the boundary-focused data augmentation method is shown in Alg. 1. In particular, this method is aimed to randomly select a circle region \( S_{c} \) in the polyp region \( S \) and set the values of all the pixels in this region randomly which is illustrated in the Figure 3.

In the initial step, we find the contour of the polyp in label images. Once the contour is found by Bradski Et al’s method [9], the 2D continuous function \( f(x, y) \) which is called raw moment of \((p+q)\) is defined in Equation 1.

\[
M_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^{p} y^{q} f(x, y) dx dy. \tag{1}
\]

However, for reducing the noise purpose, we apply \( p, q \in \{0,1,2,3\} \) to adapt to the scalar(greyscale) image with with pixel intensities \( I(x, y) \), the Equation 1 turns into Equation 2:

\[
M_{pq} = \sum_{x} \sum_{y} (x^{p} y^{q} I(x, y)). \tag{2}
\]

The central moment of the polyp are defined as the following:

\[
\mu_{pq} = \sum_{x} \sum_{y} (x - \bar{x})^{p} (y - \bar{y})^{q} f(x, y). \tag{3}
\]

The polyp center coordination is showed as \((\bar{x}, \bar{y})\), which is calculated in Equation 4.

\[
(\bar{x}, \bar{y}) = \left( \frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right). \tag{4}
\]

\[
D = \sqrt{(x_{C} - \bar{x})^{2} + (y_{C} - \bar{y})^{2}} - \sqrt{(x_{m} - \bar{x})^{2} + (y_{m} - \bar{y})^{2}}
\times \left\{ \begin{array}{ll}
0 < D: I(x_{C}, y_{C}) \in S, \\
0 \geq D: I(x_{C}, y_{C}) \notin S,
\end{array} \right.
\tag{5}
\]

where \((x_{m}, y_{m})\) is presented for the point with lowest moments started from the centroid inside polyp region \( S \).

When it comes to the second stage, a circle center \( C(x_{C}, y_{C}) \) is randomly chosen in the polyp area with the condition, that is showed in Equation 5. Subsequently, the distance \( R \) from circle center \( C(x_{C}, y_{C}) \) to the polyp boundary is computed. If distance \( R \) is smaller than the allowed maximum distance \( R_{\text{max}} \) which is determined from \( C(x_{C}, y_{C}) \) to the nearest point, a circle radius \( r_{i} \) ranging from \( R \) to \( R_{\text{max}} \) is randomly selected. The selected circle region \( S_{c} \) is defined by circle center \( C(x_{C}, y_{C}) \) and circle radius \( r_{i} \). In the third step, each pixel in \( S_{c} \) of the raw images is assigned a random value ranging from 0 to 128 due to the value of greyscale image. Nevertheless, if the distance \( R \) is longer than max distance \( R_{\text{max}} \), this process is repeated at the beginning of the first step till find the proper centroid.

B. FEATURE EXTRACTION USING A MULTI-MODEL DEEP ENCODER-DECODER NETWORK

We propose MED-Net, which was inspired by recent research on semantic segmentation [25], [63], DeepLabv3+ [14] and Upsampling method of Lee and Carlberg [54], Tian et al. [32]. The objective of this study was to build an ensemble of
deep encoder-decoder networks (DEDNs) to train and can distinguish the most discriminative features for the polyp segmentation task [58]. Moreover, the network is also able to give accurate predictions by applying newest upsample method presented by Tian et al. [54]. A DEDN ensemble is a learning paradigm in which many DEDNs are jointly employed to solve a specific problem. Our study demonstrated that an ensemble of multiple deep encoder-decoder networks significantly outperforms single simple deep encoder-decoder networks. This is because our proposed DEDN ensemble competitive advantages that are useful for increasing the prediction accuracy rate.

First, we can apply multi-scale input images to our network ensemble in which each scale is passed through at least one DEDN. This ensemble not only expands the receptive field in the original image to cover better global features but also extracts better multi-scale local features.

Second, our ensemble of DEDNs constitutes a reliable technique to increase the segmentation performance. Because each deep training model presents several local minima, multiple training processes of different DEDNs can improve the distribution of errors in each class. Thus, in combination their outputs lead to an improved performance on the overall task.

Third, in this study, we employed the encoder part of the state-of-the-art DeepLab model, namely, the DeepLabv3+ [14]. Chen et al. [14] demonstrated in the Pascal VOC 2012 challenge that the DeepLabv3+ encoder-decoder model yields a state-of-the-art performance [19]. Later we combine it with the decoder part which is known as DUpSample [54], the network is shown in Fig 4. This is because its encoder module can capture rich contextual information from several parallel atrous convolution layers with different rates and its decoder module is able to recover effectively missing boundaries caused by the pooling or convolutions with striding operations in the encoder module [31], [34]. The main advantage of the new upsampling layer lies in that with a relatively lower resolution feature map [54]. Therefore, our DEDN achieve even better segmentation performance by significantly reducing computation complexity. Let $G \in \mathbb{R}^{h \times w \times c}$ is the final product of the encoder part and $Y \in \{0, 1, \ldots, a\}^{h \times w}$ be the label map, in which $h$ and $w$ is the high and the weight, respectively. While, $c$ is the number of channels in final output and $a$ denotes for the number of classes. However, in this study, we highly focus on binary segmentation, therefore, $Y \in \{0, 1\}^{h \times w}$. Precisely, $G$ is typically of a factor of 16 or 32 in spatial size of the ground-truth $Y$ [54]. However, semantic segmentation task strongly requires per-pixel prediction, $G$ needs to be upsampled to the spatial size of $Y$ before going to the computing process of the training loss function. In other words, by integrating DUpSampling method we can make the component DEDN avoid overly reducing the overall strides of the encoder and it also significantly reduces the computation time and memory footprint of the semantic segmentation method.

Unlike in our previous research studies [38], in which we fully applied the architecture of Chen et al. [14], we created a new DEDN by merging two state-of-the-art approaches in segmentation. In addition, instead of using pre-trained deep learning models to extract discriminative features, in our study we could train every single component DEDN using its pre-trained model and our augmented training dataset. This is because we could successfully generate sufficient augmented images for training these networks and thus avoid the overfitting problem. After these training processes, our models can extract considerably better features than their pre-trained models. The training steps are explained in detail in below.

In the data augmentation step, by means of random cropping, rotation, elastic distortion, and translation, hundreds of new augmented training images are generated from each of the original training images. This new dataset is used for training each component network with differently scaled input images. Each DEDN model is trained with one of three types of scaled input images. Moreover, the fully connected layers from each model are disconnected so that each network can accept an input image of an arbitrary size. Each input image is a rescaled image from the augmented dataset.

Each single component DEDN encoder module can summarize the features of a scaled image input in the low-level feature map taken from the output of the last block. This feature map is then passed through four parallel atrous convolution layers with different rates to capture multi-scale context information and generate convolutional feature maps at multiple scales. Notably, not only can the encoder output feature map contain 256 channels and rich semantic information, but it also extracts features at an arbitrary resolution by applying both the atrous convolution and multiple resolution input dataset. This method boosted the performance of the model which is going to be demonstrated detailedly in Session IV.
With regard to decoder path, instead of applying the decoder that’s belonged to Chen et al. [14]’s work, we used data-dependent upsampling [54] which can fill the weaknesses of the former. Besides, the component DEDN feature map of each branch is rescaled such that its resolution is the same as that of the other feature maps. All DEDN feature maps are fused into a shared feature map. Since each polyp in an image is prominent in different feature maps, we use a maxout layer to obtain the competitive and dominant features from all the DEDN feature maps and integrate these features into the shared feature map. The entire architecture of our feature extraction model is illustrated in Figure 1. Data-dependent Upsampling method is used to recover the pixelwise segmentation prediction from the raw outputs of the convolutional decoder. Interestingly, this method can avoid overly decreasing the overall strides of the encoder, significantly reducing the massive computation and memory footprint. Besides, before merging features this method allows the decoder to downsample them to the lowest resolution of feature maps. By applying data-dependent upsampling, the huge computation was reduced remarkably and it can also extend the design space of feature aggregation to make better feature aggregation. The following step is to ensemble the DUpsampling into the encoder-decoder framework for an end-to-end trainable system. While DUpsampling can be performed with a convolution operation by kernel 1 × 1, incorporating directly it into the framework encounters difficulties in optimization.

So far, we have shown that DUpsampling method has been adopted to our MED-Net can be used to replace the bilinear upsampling in semantic segmentation which is less effective in making the training to converge. The next step is to incorporate the DUpsampling into the encoder-decoder framework, resulting in an end-to-end trainable system. Besides, deep learning neural networks are nonlinear methods which offer increased flexibility and can affect in proportion to the amount of training data available. Nevertheless, a drawback of this flexibility is that they learn via a stochastic training algorithm, therefore, they are sensitive to the specifics of the training data. In other words, the neural network may find a different set of weights each time they are trained, which in turn to produce different predictions. Generally, a network usually have a high variance and it can make a bad effect when trying to develop a final model. Therefore, one of successful approaches to reduce the variance of neural networks is to train multiple models which strongly motivated us to make MED-Net. Moreover, differences in random initialization, random selection of mini-batches, differences in hyperparameters, or different outcomes of non-deterministic implementations of neural networks are often enough to cause different members of the ensemble to make partially independent errors [5]. The weights of multiple networks can be averaged, to hopefully result in a new single model that has better overall performance than any original model. Likewise, ensemble method is promising to average these points in weight space, and use a network with these averaged weights, instead of forming an ensemble by averaging the outputs of networks in model space [15], [27]. The results are in predictions of out experiments show that they are better than any single networks. According to the vary of datasets in our training strategies the variance is significantly diminished.

We train each network with a stochastic gradient algorithm running on a GeForce GTX 1080 Ti, utilizing the Tensor-Flow [1] distributed machine learning system. We fine-tune the networks by using the models utilizing RMSProp with a weight decay of 0.001. We use a learning rate of 0.007, decayed every two epochs at an exponential rate of 0.94. The training process for each component network takes 350,000 steps. Direct training of our encoder-decoder network with a conventional loss function may not result in a good segmentation performance. Multi-level contextual features extracted from our network can be trained by minimizing the overall loss $L_{\text{apt}}$, which is a combination of an adaptive weighted loss $L_{\text{apt}}$ and an attention-based loss $L_{\text{att}}$.

### C. ADAPTIVE WEIGHTED LOSS FUNCTION

We use a new adaptive weighted loss function inspired by [3], [62], to boost the segmentation performance of the MED-Net. First, we compute the soft-max value in the feature channel $k$ at the pixel position $z \in \Omega$ with $\Omega \subset \mathbb{Z}^2$:

$$p_k(z) = \frac{\exp(u_k(z))}{\sum_{i=1}^{K} \exp(u_k(z))}. \quad (6)$$

where $K$ is the number of classes and $u_k(z)$ is the activation function. The adaptive weighted loss $L_{\text{apt}}$ is computed by

$$L_{\text{apt}} = - \sum_{z \in \Omega} \omega(z)p_{gt}(z)\log(p_l(z)). \quad (7)$$

where $\omega : \Omega \to \mathbb{R}$ denotes the weight map, $l : \Omega \to \{1, \ldots, K\}$ is the true label of each pixel, and $p_{gt}(z)$ denotes the true probability of current pixel $z$. We aim to build a weight map $\omega$ to give high weights to only the pixels that belong to the polyp in a training image or belong to the polyp boundary in an augmented training image. As a result, the network can consider these pixels more important than other pixels in the background. Thus, the weight map can be computed as

$$\omega(z) = \omega_o(z) + \omega_e(z). \quad (8)$$

where $\omega_o : \Omega \to \mathbb{R}$ is the basic weight map to create a balance between the polyp and the background and $\omega_e : \Omega \to \mathbb{R}$ denotes the boundary-focused weight map, which can be computed as

$$\omega_e(z) = \begin{cases} \frac{S}{S_{po}} & \text{if } z \in S_{po}, \\ 0 & \text{if } z \notin S_{po}, \end{cases} \quad (9)$$

where $n_e$ is a regularization parameter, $S_{po}$ denotes the polyp area, and $S$ is the entire image area. In our experiments, we set $n_e = 0.5$.
D. ATTENTION-BASED LOSS FUNCTION

To provide better segmentation results, we also aim to measure the similarity between a predicted polyp mask and a ground truth polyp mask. This measurement allows the network to focus more on the polyps and their boundaries. In particular, in these binary masks, pixel values in the polyp area are set to one and those in the background are set to zero. We denote the predicted value of pixel \(z\) by \(u_{\text{pred}}(z)\) and the true value of pixel \(z\) by \(u_{\text{true}}(z)\). The attention-based loss \(L_{\text{att}}\) is computed as

\[
L_{\text{att}} = \frac{K_{\text{att}}}{N} \sum_{i=1}^{N} (u_{\text{pred}}(z) - u_{\text{true}}(z))^2. \tag{10}
\]

where \(N\) is the total number of pixels in the polyp region, \(z\) is the present pixel position in the polyp region, and \(K_{\text{att}}\) is a regularization parameter. By using this loss function, the difference between the predicted and the true polyp boundary can be iteratively reduced. Moreover, the combination of the above loss functions can lead to a better segmentation performance, because the network is focused focusing on learning iteratively polyp boundaries.

E. ABBLATION STUDIES

In order to analyze the effect of each contribution to the segmentation performance, we performed an ablation study where different parts of the system were removed - for instance removing one or both of two in three main contributions. To make equal to all ablation experiments we separated the CVC-ColonDB [48] data set in to training and testing part with the ratio 80 % and 20 %, respectively.

For instance, in the first experiment, not only did we adopt the Deeplabv3+ [14] as the main network architecture with the backbone is Xception [17], and kept the softmax cross entropy loss was for logits of each scale was also used. While the augmentation method was kept during the process. Strikingly, all experiments were applied the Boundary-focused data augmentation method always tend to get higher Dice score by comparing to others. The Table 1 has shown that while the single component contribution slightly impact the the performance of the model, the combination of two contributions always showed the better one.

Furthermore, we also performed four experiments to show the process of choosing the number of component networks. To analyze the performance of each network we kept all the remaining contributions and the quantity of dataset for training and test ratio was also remained unchanged as the former experiments. The result has shown in the Table 2, the MED-Net with three component networks significantly outperformed the three of networks.

IV. EXPERIMENTS RESULTS AND ANALYSIS

In this section, we demonstrate the effectiveness of our proposed polyp segmentation methods. We used three databases to evaluate our methods and compared our results with those of state-of-the-art algorithms.

A. DATASETS

In order to demonstrate the performance of our proposed approach, we used well-known datasets from the MICCAI 2015 polyp detection challenge [52]. For comparison purposes, we separated the datasets into a training and a testing dataset according to the recommendation of the MICCAI challenge guidelines: CVC-CLINIC for training and ETIS-Larib for testing. Furthermore, we also report the results from a second publicly available dataset (CVC-ColonDB). In addition, to show the effectiveness of our method when applied to additional datasets, we also used many datasets to meet different goals : for example, we used a combination of CVC-ColonDB and CVC-ClinicDB, which contains 912 images with associated polyps, to compare our method with that of Li et al. [33].

The datasets are briefly described in the following paragraphs.

- CVC-ClinicDB [7] contains 612 images, where all images show at least one polyp. The segmentation labels obtained from 25 colorectal videos by selecting 29 sequences from these videos. In this database, these images are in three channels with format of tiff, dimension of 384 × 288 pixels and their corresponding label maps.
- CVC-ColonDB [48] contains 379 frames from 15 different colonoscopy sequences where each sequence shows at least one polyp each. These frames were selected in order to maximise the visual difference between them. In this dataset, images are appeared in RGB and the resolution is 574 × 500 pixels. Hence, this dataset can cover as many types of polyp appearances as possible.
- ETIS-LaribPolypDB [8] contains 196 images, which are generated from 34 videos. The size of each image is 1225 × 966 pixels. This dataset contains 44 different polyps with various sizes and shapes. This dataset is one of the most challenging dataset in polyp field. Not only

| Dataset | Augmentation | MED-ColonDB | Loss Function | Dice |
|---------|--------------|-------------|---------------|------|
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.812 |
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.765 |
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.761 |
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.875 |
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.792 |
| CVC-ColonDB | ✓ | ✓ | ✓ | 0.821 |

| Network | Dice |
|---------|------|
| MED-Net 1 component network | 0.841 |
| MED-Net 2 component networks | 0.873 |
| MED-Net 4 component networks | 0.711 |
| MED-Net | 0.908 |
does it lack of quantity, it also includes a vast of polyp appearance.

- ASU-Mayo Clinic Colonoscopy Video Database [53] contains 20 short colonoscopy videos, which has 10 videos have a unique polyp inside (positive shots) and the other 10 videos have no polyps (negative shots). These video are selected so as to display maximum variation in colonoscopy procedures with two types of resolutions.

**B. CALCULATION METRICS**

We used the Jaccard index, also known as the Intersection over Union (IoU), as the main metric to evaluate our approach’s performance. Furthermore, in order to provide a general view of the effectiveness of our method, we also employed Dice score to describe our results.

We calculated the mean IoU parameter which known as each per-class IoU was computed over a validation/test set by the Equation 11.

\[
IoU = \frac{PR \cap GT}{PR \cup GT}. \quad (11)
\]

where PR represents the binary mask produced by the segmentation model and GT the ground truth mask, while \(\cap\) denotes a set of an intersection and \(\cup\) a union set between PR and GT.

In addition, we also report the results of three common segmentation evaluation metrics: mean pixel precision, mean pixel recall, and mean accuracy.

We apply precision as a measurement refers to a state of strict exactness — how consistently the pixel is strictly exact and represents the proportion of the negative values produced by the segmentation method and the real negatives of GT in the Equation 12.

\[
Precision = \frac{TP}{TP + FP}. \quad (12)
\]

In addition, we use recall metric which is also known as Sensitivity to show the fraction of pixels that have been retrieved over the total amount pixel of ground truth in the Equation 13.

\[
Recall = \frac{TP}{TP + FN}. \quad (13)
\]

Accuracy metric is one of the most common evaluated metric which is used to simply report the percent of pixels in the image which were correctly classified was adopted and presented in Equation 14.

\[
Accuracy = \frac{TN + TP}{TN + TP + FN + FP}. \quad (14)
\]

Notably, the Dice coefficient is a statistic used for comparing the similarity of prediction images and label images which is the quotient of likeness and ranges between 0 and 1. Which is presented in Equation 15.

\[
Dice = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}. \quad (15)
\]

If a pixel of a polyp was correctly segmented, it was counted as true positive (TP). Every pixel segment as belonging to a polyp that fell outside a polyp label was counted as false positive (FP). Finally, every polyp pixel that was not been detected was counted as false negative (FN) and the remaining pixels were described as true negative (TN).

**C. COMPARISON WITH OTHER STATE-OF-THE-ART APPROACHES**

In every comparison experiment, we upsampled and downsampled the training dataset to feed it into three different training phases. Not only did we apply our proposed augmentation method, but we also trained the network using other effective methods. For instance, a cropping method was used; that is, we cropped from the center to remove the black parts generated by the camera that exist at image corners. The rotation method was applied using random degrees from 0 to 360. We conducted very many experiments, the results of which showed that using grayscale images for both the training and the testing phase is always better than using RGB images. Therefore, we proceeded to use all the datasets. Subsequently, we formatted the datasets in TFRecords format to optimize the model. Furthermore, all training processes were executed on TensorFow with GPU GTX-1080Ti.

1) RESULT ON THE ETIS-LARIB DATASET

First, in terms of the most challenging dataset, which is provided by ETIS laboratory and Lariboisiere Hospital-APHP. In order to challenge the previous accomplishment on ETIS-Larib dataset, Brandao et al. [10]’s methodology follows the same data guidelines and restrictions with which is given in the e 2015 MICCAI sub-challenge. In the training phase, the author used 4664 images and its corresponding labels which contains at least one polyp and later they use ETIS-Larib [8] to evaluate the performance of model. In the beginning, as mentioned above we turned all images in the CVC-ColonDB dataset [48] and extracted frames in ASU-Mayo Clinic Colonoscopy Video Database [53] into grayscale images. Once this process have been done, these images were resized from 384 \times 288 pixels to 500 \times 400 pixels and 250 \times 200 pixels, respectively. Subsequently, all three sets were carried into three component networks for training. Table 3 shows the achievement of our model in terms of polyp segmentation. This table shows that our MED-Net has outperformed 3 FCNN networks and our previous work. More specially, proposed model achieved both the highest precision and the highest recall among the models. The experimental results show again that the fused method, because of its ability to aggregate multi-scale contextual information, outperforms all the other approaches. Moreover, examples of three FCNNs results for polyp segmentation are depicted in Figure 5. The figure 5 shows that our model can detect the tumor boundary optimally, a result that the other models can not achieve. Finally, in the testing phase, we transformed these images from RGB image in ETIS-Larib [8] to grayscale and
TABLE 3. Comparison of proposed method and three fully convolutional neural networks in terms of mean pixel precision and recall for the ETIS-Larib dataset [8].

| Networks                  | Mean precision (%) | Mean recall (%) |
|---------------------------|--------------------|-----------------|
| FCNN-AlexNet [30]         | 27.89              | 35.54           |
| FCNN-GoogLeNet [51]       | 25.83              | 37.82           |
| FCNN-VGG [49]             | 70.23              | 54.20           |
| U-Net [45]                | 88.87              | 86.83           |
| DeepLabv3+ [14]           | 87.98              | 86.25           |
| CDED-Net [38]             | 92.93              | 90.87           |
| Proposed                  | 93.82              | 90.93           |

resized it to the same size with images in the training stage so that they would be compatible with the model.

2) RESULT ON THE CVC-ColonDB DATASET
Second, we appraised our proposed approach performance on the CVC-ColonDB dataset. This database contains images all of which show polyps of different shapes and all of them were annotated by expert clinicians. In addition, we employed the methods presented in [2] and [63] to analyze the comparative effectiveness of our method. Whereas the method presented by Akbari et al. [2] applies a post-processing procedure on the probability map produced by the network to improve the results, we did not apply any post-processing during the testing phase. Besides, we also used two well-known encoder-decoder networks, one was U-Net [45] which is known as a standardization in medical field and the second network is one of state-of-the-art network, DeepLabv3+ [14]. In the training step, the training data selection which was used by Akbari et al. [2], we selected 200 images for training purpose arbitrarily and the remain of the dataset was used for testing. After that, again we upsampled these training images from 384 × 288 pixels to 500 × 400 pixels, and then we downsampled them to create a smaller dataset, the size of the images in which was exactly 250 × 200 pixels. Table 4 shows that our approach outperformed the other two methods completely. Our model achieved a Dice score of 0.908, which is 0.012 better than that of our previous approach. This is because our method can capture multi-level and multi-scale features, which is the advantage of using multiple-model deep encoder-decoder networks.

FIGURE 5. Comparison with FCNN networks predictions [10] in the ETIS-Larib dataset [8].

V. CONCLUSION
The segmentation of tumour epithelium in histopathology is an essential beginning step for biomarker assessment, tumour quantification, and prognosis determination in colorectal cancer. In this paper, we proposed a novel deep learning approach that uses a multi-model deep encoder-decoder network, called MED-Net, for colorectal polyp segmentation, in the MED-Net we present a new component networks which are applied with DUpsample method which is known as one of the most powerful approach in segmentation task to robust the performance of the model. In addition, a new data augmentation method that can be applied in polyp segmentation approaches was presented. Furthermore, we described the use of a new adaptive weighted loss function and an attention-based loss function to boost the performance of the model. Our method outperformed the state-of-the-art polyp segmentation methods on three datasets. The key advantage of the proposed method over existing methods is that it employs an ensemble of encoder-decoder networks trained to extract visual features from multi-scale images.

TABLE 4. Comparison of proposed method with FCNN-8S combined with post-processing and the method in [63].

| Networks                  | Accuracy | Dice  | Specificity | Recall | Precision | Dice  |
|---------------------------|----------|-------|-------------|--------|-----------|-------|
| Zhang et al. [63]         | 0.975    | 0.701 | 0.988       | 0.757  |
| Akbari et al. [2]         | 0.977    | 0.810 | 0.993       | 0.748  |
| U-Net [45]                | 0.832    | 0.759 | 0.856       | 0.648  |
| DeepLabv3+ [14]           | 0.942    | 0.862 | 0.945       | 0.764  |
| Nguyen et al. [39]        | 0.984    | 0.889 | 0.915       | 0.772  |
| CDED-Net [38]             | 0.980    | 0.896 | 0.991       | 0.792  |
| Proposed                  | 0.990    | 0.908 | 0.982       | 0.798  |

TABLE 5. Comparison of proposed method with others in CVC-ClinicDB [7].

| Networks                  | Accuracy | Specificity | Recall | Precision | Dice  |
|---------------------------|----------|-------------|--------|-----------|-------|
| Li et al. [33]            | 0.970    | 0.770       | 0.990  | 0.900     | 0.830 |
| U-Net [45]                | 0.857    | 0.845       | 0.868  | 0.885     | 0.810 |
| DeepLabv3+ [14]           | 0.988    | 0.862       | 0.883  | 0.917     | 0.862 |
| CDED-Net [38]             | 0.987    | 0.942       | 0.962  | 0.950     | 0.891 |
| Proposed                  | 0.992    | 0.954       | 0.953  | 0.958     | 0.913 |

3) RESULT ON THE CVC-ClinicDB DATASET
Third, we evaluated the performance of our model on the CVC-ClinicDB dataset [7]. From this database we selected 430 images (80% in total dataset) randomly for training, and the remaining 182 images were used as the test set: thus, there is no intersection between the training and the test set. This training and testing strategy also were applied to Chen et al. [14]’s, Li et al. [33]’s and Ronneberger et al. [45]’s network. Table 5 shows that the performance of our model is significantly higher than the competitors in all scores, especially in Dice score which is the main. While our method could clearly segment the the tumor in these testing images, the comparison method could not do locate the tumor position perfectly.
Our proposed method consists of two training stages. In the first training stage, strong data augmentation methods are used to improve the segmentation performance of the deep learning networks. To achieve a better segmentation performance, despite the challenging problem that only a limited number of colon cancer databases area accessible, we not only apply a novel augmentation method, but also other effective methods. MED-Net is trained to extract multi-context information from multi-scale training images. It was able to extract both global and local features of colorectal cancer polyps appropriately, and it could thus greatly improve the performance of different types of methods for detecting malignant polyps. The best training model was built by combining three training models with different resolutions, using the majority voting strategy. Moreover, we introduced two loss functions which strongly supported MED-Net in terms of achieving ideal results.

Our experimental results demonstrate the superiority of the proposed method to state-of-the-art polyp segmentation approaches. Our future work will be focused on improving the IoU measure by adopting distinctive deep learning methods, such as batch normalization and residual networks, which may provide a good tool that will further improve the segmentation performance of our method. Furthermore, we also plan to apply our approach to a real-time system.

VI. DISCUSSION
Our proposed approach is still erroneous, but we are attempting to improve on existing research results by means of a variety of methods. In colorectal polyp segmentation, a good classification algorithm and high-quality data complement each other. When segmenting a high resolution image we must upsample the image and then downsample it to recover the original resolution. Thus, we intend to explore the convolutional neural network structure with the objective of devising a network with wider and deeper polyps. Especially, we will highly focus on the encoder part to extract most of meaningful features. The training time is also one of our concerns due to the encoder-decoder structure. There is a trade-off between time and performance, therefore, the training time is also a drawback in this study due to the number of component networks in MED-Net and the discrimination of encoder-decoder structure.

REFERENCES
[1] M. Abadi, “TensorFlow: A system for large-scale machine learning,” in Proc. OSDI, vol. 16, 2016, pp. 265–283.
[2] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S. M. Reza Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, “Polyp segmentation in colonoscopy images using fully convolutional network,” 2018, arXiv:1802.00368. [Online]. Available: http://arxiv.org/abs/1802.00368
[3] M. Arshad and O. Abdalghani, “On estimating the location parameter of the selected exponential population under the LINEX loss function,” Brazilian J. Probab. Statist., vol. 34, no. 1, pp. 167–182, Feb. 2020.
[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[5] Y. Bengio, A. Courville, and I. J. Goodfellow, Deep Learning: Adaptive Computation and Machine Learning (Adaptive Computation and Machine Learning), Cambridge, MA, USA: MIT Press, 2016.
[6] J. Bernal, J. M. Nanez, and F. J. Sánchez, “Polyp segmentation method in colonoscopy videos by means of nsa-dova energy maps calculation,” in Proc. Workshop Clin. Image-Based Procedures. Cham, Switzerland: Springer, 2014, pp. 41–49.
[7] J. Bernal, F. J. Sánchez, G. Fernández-Esparrrach, D. Gil, C. Rodríguez, and F. Vilarino, “NMD-DA maps for accurate polyp highlighting in colonoscopy: Validation vs. Saliency maps from physicians,” Comput. Med. Imag. Graph., vol. 43, pp. 99–111, Jul. 2015.
[8] J. Bernal, J. Sánchez, and F. Vilarino, “Towards automatic polyp detection with a polyp appearance model,” Pattern Recognit., vol. 45, no. 9, pp. 3166–3182, Sep. 2012.
[9] G. Bradski and A. Kaehler, Learning OpenCV: Computer Vision With the OpenCV Library, Newton, MA, USA: O’Reilly Media, 2008.
[10] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, “Fully convolutional neural networks for polyp segmentation in colonoscopy,” Proc. SPIE Comput-Aided, Diagnosis, vol. 10134, Dec. 2017, Art. no. 101340F.
[11] M. Breier, S. Gross, A. Behrens, T. Stiehle, and T. Aach, “Active contours for localizing polyps in colonoscopie abbi image data,” Proc. SPIE Int. Opt. Photon., vol. 7963, Oct. 2011, Art. no. 79632M.
[12] T. Duc Bui, J. Shin, and T. Moon, “3D densely convolutional networks for volumetric segmentation,” 2017, arXiv:1709.03199. [Online]. Available: http://arxiv.org/abs/1709.03199
[13] H. Chen, X. Qi, L. Yu, and P.-A. Heng, “DCAN: Deep-contour-aware networks for accurate gland segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2487–2496.
[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 801–818.
[15] Y. Chen, H. Chang, J. Meng, and D. Zhang, “Ensemble neural networks (ENN): A gradient-free stochastic method,” Neural Netw., vol. 110, pp. 170–185, Feb. 2019.
[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” 2014, arXiv:1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078
[17] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1251–1258.
[18] 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.
[19] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes challenge: A retrospective,” Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–138, Jan. 2015.
[20] M. Ganz, X. Yang, and G. Slabaugh, “Automatic segmentation of polyps in colonoscopic narrow-band imaging data,” IEEE Trans. Biomed. Eng., vol. 59, no. 8, pp. 2144–2151, Aug. 2012.
[21] A. Hajmancouchehri and H. Mahvash Mohammadi, “Detection of frames containing polyp using CNN and polyp segmentation using FCN,” IET Comput. Vis., Mar. 2020, doi: 10.1049/iet-cvi.2019.0300.
[22] C. Han, L. Rundo, R. Araki, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi, “Infinite brain mr images: Pagan-base data augmentation for tumor detection,” in Neural Approaches to Dynamics of Signal Exchanges. Singapore: Springer, 2020, pp. 291–303.
[23] C. Hassan, M. B. Wallace, P. Sharma, R. Maselli, V. Craviotto, M. Spadaccini, and A. Repici, “New artificial intelligence system: First validation study versus experienced endoscopists for colorectal polyp detection,” Gut, vol. 69, no. 5, pp. 799–800, May 2020.
[24] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, “A real-time algorithm for signal analysis with the help of the wavelet transforms,” in Wavelets, Berlin, Germany: Springer, 1990, pp. 286–297.
[25] F. Huang, J. Zhang, C. Zhou, Y. Wang, J. Huang, and L. Zhu, “A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction,” Landslides, vol. 17, no. 1, pp. 40–54, Jan. 2020.
[26] S. Hvam, J. Oh, W. Tavanapong, J. Wong, and P. C. D. Groen, “Polyp detection in colonoscopy video using elliptical shape feature,” in Proc. IEEE Int. Conf. Image Process. (ICIP), vol. 2, Dec. 2007, p. 465.
[27] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. Gordon Wilson, “Averaging weights leads to wider optima and better generalization,” 2018, arXiv:1803.05407. [Online]. Available: http://arxiv.org/abs/1803.05407

99506
VOLUME 8, 2020
[28] X. Jia, X. Mai, Y. Cui, Y. Yuan, X. Xing, H. Seo, L. Xing, and M. Q.-H. Meng, “Automated polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction,” IEEE Trans. Autom. Sci. Eng., early access, Jan. 30, 2020, doi: 10.1109/TASE.2020.2964827.

[29] Y. Jia, “Polyps auto-detection in wireless capsule endoscopy images using improved method based on image segmentation,” in Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO), Dec. 2015, pp. 1631–1636.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[31] D. Kucharski, P. Kleczek, J. Jasorek-Konjakowska, G. Dyduch, and M. Gorgoń, “Semi-supervised nests of melanocytes segmentation method using convolutional autoencoders,” Sensors, vol. 20, no. 6, p. 1546, 2020.

[32] K. Lee and K. T. Carlberg, “Model reduction of dynamical systems on non-manifolds using deep convolutional autoencoders,” J. Comput. Phys., vol. 404, Mar. 2020, Art. no. 108973.

[33] Q. Li, G. Yang, Z. Chen, B. Huang, L. Chen, D. Xu, X. Zhou, S. Zhong, H. Zhang, and T. Wang, “Colorectal polyp segmentation using a fully convolutional neural network,” in Proc. 10th Int. Congr. Image Signal Process., Biomed. Infornat. (CISP-BMEI), Oct. 2017, pp. 1–5.

[34] P. Liu, Q. Dou, Q. Wang, and P.-A. Heng, “An encoder-decoder neural network with 3D squeeze-and-excitation and deep supervision for brain tumor segmentation,” IEEE Access, vol. 8, pp. 34029–34037, 2020.

[35] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8759–8768.

[36] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[37] F. Mahmood, Z. Yang, R. Chen, D. Borders, W. Xu, and N. J. Durr, “Polyp segmentation and classification using predicted depth from monocular endoscopy,” Proc. SPIE Int. Soc. Opt. Photon., vol. 10950, May 2019, Art. no. 1095011.

[38] N.-Q. Nguyen and S.-W. Lee, “Robust boundary segmentation in medical images using a consecutive deep encoder-decoder network,” IEEE Access, vol. 7, pp. 33795–33808, 2019.

[39] Q. Nguyen and S.-W. Lee, “Colorectal segmentation using multiple encoder-decoder network in colonoscopy images,” in Proc. IEEE 1st Int. Conf. Artif. Intell. Knowl. Eng. (AKIE), Sep. 2018, pp. 208–211.

[40] Y. Ouali, C. Hudelet, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency training,” 2020, arXiv:2003.09005. [Online]. Available: http://arxiv.org/abs/2003.09005

[41] G. Papandreou, I. Kokkinos, and P.-A. Savalle, “Modeling local and global deformations in deep learning: Epithomic convolution, multiple instance learning, and sliding window detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 390–399.

[42] H.-C. Park, Y.-J. Kim, and S.-W. Lee, “Adenocarcinoma recognition in endoscopy images using optimized convolutional neural networks,” Appl. Sci., vol. 10, no. 5, p. 1650, 2020.

[43] Y. S. Perl, C. Pallacivini, I. P. Ipina, M. L. Kringelbach, G. Deco, H. Laufs, and E. Tagliazucchi, “Data augmentation based on dynamical systems for the classification of brain states,” in Proc. IEEE 1st Int. Conf. Med. Image Anal., vol. 482, pp. 123–138, May 2019.

[44] L. Zhang, S. Dolwani, and X. Ye, “Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons, in Proc. Annu. Conf. Med. Image Understand. Anal. Cham, Switzerland, Springer: 2017, pp. 707–717.

[45] Y. Zhang, S. Miao, T. Mansi, and R. Liao, “Unsupervised X-ray image segmentation with task driven generative adversarial networks,” Med. Image Anal., vol. 62, May 2020, Art. no. 101664.

[46] 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.

[47] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, “Conditional random fields as recurrent neural networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1529–1537.

[48] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, “Multi-scale deep context convolutional neural networks for semantic segmentation,” World Wide Web, vol. 22, no. 2, pp. 555–570, Mar. 2019.

[49] Colorectal cancer facts & figures 2011–2013, American Cancer Society, Atlanta, GA, USA, 2011.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

[51] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, and A. Jemal, “Colorectal cancer statistics, 2017,” Cancer J. Clinicians, vol. 67, no. 3, pp. 177–193, May 2017.

[52] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polycytes in WCE images for early diagnosis of colorectal cancer,” Int. J. Comput. Assist. Radiol. Surgery, vol. 9, no. 2, pp. 283–293, Mar. 2014.

[53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556
DUC MY VO received the Bachelor of Engineering degree in automation from the University of Transport and Communication, Vietnam, in 2006, the master’s degree from the Asian Institute of Technology, Thailand, in 2009, and the Ph.D. degree from the Laboratory of Prof. Dr. Andreas Zells, University of Tübingen, Germany, in 2015. His thesis projects were focused on addressing the use of RGB-D images for six important tasks of mobile robots, such as face detection, face tracking, face pose estimation, face recognition, person detection, and person tracking. These topics have widely been researched in recent years, because they provide mobile robots with abilities necessary to communicate with humans in natural ways. After one year of research with the Vietnam National Satellite Center, he joined the Computer Vision and Multimedia Laboratory, Chosun University, South Korea, in 2016. In 2017, he moved to the Pattern Recognition and Machine Learning Laboratory, Gachon University, South Korea. His current research interests include development of action recognition, face recognition, semantic segmentation, and biomedical imaging.

SANG-WOONG LEE (Senior Member, IEEE) received the B.S. degree in electronics and computer engineering and the M.S. and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1996, 2001, and 2006, respectively. From June 2006 to May 2007, he was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University. From September 2007 to February 2017, he was a Professor with the Department of Computer Engineering, Chosun University, Gwangju, South Korea. He is currently an Associate Professor with the Department of Software, Gachon University. His current research interests include face recognition, computational aesthetics, machine learning, bioinformatics, and medical imaging analysis.