Biomedical Signal Processing and Control 70 (2021) 103019

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

A self-attention based faster R-CNN for polyp detection from colonoscopy images

Bo-Lun Chen [a,c,*], Jing-Jing Wan [b], Tai-Yue Chen [a], Yong-Tao Yu [a], Min Ji [a]

[a] Department of Computer Science, Huaiyin Institute of Technology, Huaiyin 223001,China
[b] Department of Gastroenterology, The Affiliated Huai'an Hospital of Xuzhou Medical University, the Second People's Hospital of Huai'an, Huaian 223002, China
[c] Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland

ARTICLE INFO

Keywords:
Colorectal cancer
Polyp detection
Contrast enhancement
Feature extraction network

ABSTRACT

At present, the incidence rate of colorectal cancer (CRC) is increasing year by year. It has always affected people's physical and mental health and quality of life. How to improve the detection ability of polyp plays a key role in colonoscopy. In order to solve these problems, in this paper, we first enhance the contrast of the input image by well distinguishing the foreground from the background in order to improve the saliency of the polyp regions. Then, we feed the enhanced image into an improved Faster R-CNN architecture comprised three processing modules for feature extraction, region proposal generation, and polyp detection, respectively. In order to further improve the quality, as well as the feature abstraction capability of the feature maps produced by the feature extraction network, we append an attention module to attend to the useful feature channels and weaken the contributions of the helpless feature channels. The experimental results demonstrate that the accuracy of the proposed polyp detection network is greatly improved compared with the existing algorithms, and the network not only can accurately identify polyps of varying sizes and conditions in single polyp images, but also can achieve excellent performance in handling multiple polyp images. This paper will be greatly helpful to alleviate the missed diagnosis of clinicians in the process of endoscopic examination and disease treatment, as well as providing effective assistance for the early diagnosis, treatment and prevention of the CRC, which is also of great significance to the clinical work of physicians.

1. Introduction

Colorectal cancer (CRC) is known to be a kind of common malignant tumor in the world. As the people's daily living standards improve and the dietary habits change, the incidence and mortality rates of the colorectal cancers have been increasing, which causes seriously dangers to the health and life quality of the human beings. The CRC has been a dominant public health issue caused by the high incidence and mortality rates. On the basis of statistics, the CRC is treated as the second and third dominant cause of deaths on males and females, respectively. Specifically, the percentage of the CRC incidence among the youngsters has increased significantly year by year [1,2].

Colonoscopy acts a vital role in the capturing of the CRC, and the detection ability of polyps plays an important part in colonoscopy [3,4]. In addition, some studies show that the risk of interphase CRC will be degraded by three percentages when the detection rate of adenoma (ADR) increases by one percentage [5]. In the past half century, the mortality and incidence rates of CRC in adults have dropped sharply (51% and 32% respectively), primarily due to the CRC capturing and deletion of the adenomatous polyps [6]. However, the poly detection rate varies greatly due to the factors of polyps and the skill level of endoscopists. Moreover, in some cases, even if the polyp is in the field of vision, the polyp may still be missed, and the missed diagnosis rate is as high as 27%. It can be seen that the polyp that cannot be recognized in the field of vision is an important problem [7]. In the medical industry, the previous prediction of early colon cancer is based on statistical analysis, or diagnosis according to the patient's case, or assisted by the second observer to increase the detection rate of polyps (PDR), but such a strategy for improving the detection rate of adenomas (ADR) is still controversial, and there is a lack of automatic detection auxiliary mechanism.

Currently, medical industries have integrated more advanced techniques, including artificial intelligence and sensing technology to conduct intelligent medical services in real sense and boost the

* Corresponding author.
E-mail address: chenbolun1986@163.com (B.-L. Chen).

https://doi.org/10.1016/j.bspc.2021.103019
Received 30 May 2021; Received in revised form 23 July 2021; Accepted 30 July 2021
Available online 11 August 2021

prosperity and development of medical industry. As for the latest breakthrough of artificial intelligence, particularly the advancement of deep learning, the diagnosis of polyps with computer aided means during colonoscopy has been paid more and more attention [8].

From the perspective of deep learning, this paper discusses how to build an expert system by using the case report and attribute characteristics of the patients, so as to upgrade the recognition rate of the polyps. It will be very helpful to lower the missing diagnosis in the process of endoscopic examination and disease treatment, as well as providing effective assistance for the early diagnosis, treatment, and prevention of the CRC. The main contributions of this paper include the following: (1) a contrast enhancement strategy is leveraged as the pre-processing for improving the saliency of the polyp regions and (2) an attention module is integrated into the feature extraction backbone for promoting the feature representation quality by emphasizing the channel-wise informative features.

2. Related works

In recent years, the automatic detection of polyps has been the focus of many researchers, and the research on polyp detection also emerges in an endless stream. Traditionally, the detection and classification of anatomy in medical images are usually divided into two phases. The first phase detects the region of interest (ROI), and the second phase classifies the detected ROI. Tian et al. used a single-stage detection and classification model to classify polyps into five categories, and trained the model in a one-process manner, making the training and reasoning process simpler and faster [9]. Among them, Hwang et al. developed a new approach for polyp region detection based on ellipse shape features. This method does not use the texture features of the image, but uses ellipse target detection which is suitable for almost all small colon polyp shape features. Firstly, watershed image segmentation and ellipse fitting algorithm are used to determine whether the ellipse is suitable for segmentation in the video frame of colonoscopy, and then non-polyp regions are filtered from candidate regions by matching curve direction, curvature, margin and intensity to identify polyp regions [10]. Tajbakhsh et al. achieved more accurate polyp localization by learning various features of polyps on multiple scales, such as color, texture, shape and temporal information, and proposed a polyp detection algorithm based on independent three-way image representation and convolutional neural network. This method first extracts the features of different types of polyps by a group of convolutional neural networks near the candidate regions of polyps, then fuses the features, and finally judges the existence of polyps by the fused features [11]. Next, Tajbakhsh et al. proposed a novel method by using computer-aided detection for the detection of mixed polyps. The algorithm first uses contextual properties to delete non-polyp edges from the edge graph, then uses a voting scheme based on shape information to locate polyp candidate regions in the improved edge graph, and assigns a probability confidence value to each generated candidate region, so as to determine the polyp position [12]. Wang et al. used context and cosine ground truth projection to improve the performance of detection model [13]. Sasmal et al. proposed an efficient algorithm framework for polyp segmentation from endoscopic images. In this algorithm, principal component tracking (PCP) is used to remove the specular region in the image, and active contour (AC) model is used to locate the polyp region in each frame. The algorithm has a good effect on the polyp region with better illumination in the image [14].

Bernal et al. compared different polyp detection methods, including the traditional method based on manual feature selection and the method based on machine learning. From the final results, it can be seen that the method using machine learning for polyp detection has superior performance in both static frame and full sequence set, and the response time of some algorithms is short, which can be used in practical clinical applications [15]. Similarly, Mo et al. conducted research and comparison on a large number of polyp detection algorithms, and found that the method based on deep learning is in the leading position in the algorithm performance, among which the Faster R-CNN algorithm has higher polyp detection performance and achieved satisfactory results, which can be used in clinical practice [16].

Billah et al. combined color wavelet features and convolutional neural network features in video frames, and then detected polyps by linear support vector machine [17]. Qadir et al. improved Mask R-CNN, using different CNN architectures as feature extraction backbone network to detect and segment polyps. In the algorithm design, the performance improvement of each feature extractor is analyzed by adding additional polyp images to the training set. Finally, an integrated method is proposed for polyp detection and segmentation [18]. Qadir et al. also improved CNN algorithm and proposed a semi-automatic colon polyp labeling framework based on fully convolutional neural network. In this algorithm, CNN network only needs ground truth of several frames in the video for training, and pre-processing and post-processing steps are carried out by using data enhancement strategy, morphological operation and Fourier descriptor [19]. Pozdeev et al. first classified the endoscopic images according to their global features to determine whether there were polyps, and then used convolutional neural network to segment the polyps [20]. Zheng et al. proposed a CNN polyp detection algorithm by utilizing optical flow and online training. The algorithm first uses a single frame target detection or segmentation network, such as U-Net, to preliminarily detect and locate polyps, and then uses temporal information and optical flow to track polyps. Next, a motion regression model and an effective online training CNN model are established [21].

Ruikai et al. designed a convolutional neural network architecture based on regression, which first extracts the spatial features of intestinal polyps through the pre-trained ResYOLO model, and then optimizes the detection results of the ResYOLO output based on temporal information through an efficient convolution operator tracker [22]. Liu et al. studied the single scan detector (SSD) framework, which is a single-stage method. It uses feedforward CNN to generate a set of boundary frames of fixed size for each object from different feature maps, and takes ResNet50, VGG16, etc. as feature extraction backbone networks for performance evaluation [23]. Zhang et al. proposed a gastric polyp detection model with the architecture of SSD (ssdgpnet). The model takes advantage of the multi-resolution features extracted in the feature pyramid architecture, reuses the information discarded by the maximum pooling, and stitches these data as additional features with the output features to enhance the effect of classification and detection. At the same time, in the feature pyramid, the underlying feature map is integrated with the deconvolution of the high-level feature map, making the relationship between layers more clear, effectively increasing the number of feature channels [24]. Bagheri et al. combined the segmentation method formulated with a convolutional neural network with LinkNet to improve the quality of polyp segmentation. The method used color space combination as the input of the network in the design process, and the results showed that it achieved good segmentation effect [25]. TASHK et al. suggested a polyp detection method by using R-CNN and DRLSE. The algorithm first improves CNN algorithm to locate the polyps in the image, and then uses DRLSE to segment the local polyps automatically [26]. Jia et al. proposed a two-stage pyramid feature prediction algorithm based on deep learning for automatic polyp recognition in colonoscopy images, and achieved good performance [27].

Due to the lack of polyp detection images, Henriksen et al. found that the performance of polyp detection algorithm can be effectively improved through data enhancement and optimization of training set [28]. Due to the lack of polyp detection images, Bardhi et al. used the open source image enhancement Library in Python for image enhancement, and combined convolutional neural network with self-encoder to detect colon polyps [29]. Yu et al. proposed an offline and online 3D depth learning algorithm based on 3D-FCN, which can learn more representative spatiotemporal features from colonoscopy videos, and use the specific information of input video to solve the problem of

limited sample data set [30]. Younghak et al. proposed a CNN polyp automatic detection method based on region candidate box. In the algorithm design process, considering the lack of polyp detection image, the image enhancement strategy was used to enhance the data set, and then a deep CNN structure was used as a transfer learning strategy, two effective post-learning methods, named automated false positive and offline learning, are proposed, which are combined with the detection system with the assistance of region candidate box to realize the automatic detection of polyps [31]. Ma et al. developed a polyp detection framework using the bootstrap method by enhancing the training data through time sequence consistency. The algorithm selects samples with time sequence consistency from the test video to fine-tune the model, so as to upgrade the accuracy of the model constructed on the small sample training data [32].

In the process of polyp detection, CNN model is easily affected by small disturbance and noise, which will miss polyps in adjacent frames and produce a large number of false positives. Qadir et al. proposed a two-stage polyp detection algorithm, including a CNN based target region of interest extraction network and a false positive filtering unit. The false positive filtering unit integrates the two-way temporal information obtained from the region of interest into a group of consecutive frames, and uses the temporal correlation between the image frames in the video to predict the polyp position. The algorithm has a comprehensive performance improvement in sensitivity, accuracy and specificity [33]. Ruiz et al. suggested a polyp detection pipeline based on dense hough transform using pixel-level direction and curvature features. This algorithm can not only solve the occlusion issues in the middle frame, but also reduce false positives by converting the region histogram into the opposite color for representation [34].

Jha et al. took multiple SE blocks to build a Unet-like model. They added several unique skip connections to provide more semantic information for ColonSegNet to generate more accurate result[35]. Qadir et al. adopted Gaussian masks to weaken the effect of outer edges. It forced F-CNN to reject many false-positive proposals with strong edges and thus improved the performance on flat and small polyps[36]. Xu et al. proposed a feature enhancement module which combined temporal and spatial information to refine a feature map from the original feature extractor. The temporal information helped the network to distinguish the relationship between consecutive frames and made better results[37]. Billah et al. took multiple techniques to extract feature maps, such as color wavelet and convolutional neural network, which were fed into mutual information model for dimension reduction. It can provide more diversification of feature map and make a wise choice[38]. Yang et al. utilized LCDH to obtain a color texture feature on positive regions, and then used it to generate visual words by a codebook acquisition method. At testing stage, the author adopted NVLLC and SVM algorithms to make final decision. It had greatly promoted the development of WCE polyp detection[39].

3. Methodology

The architecture of the designed polyp detection system is described in detail in this part. The structure of the polyp detection system is shown in Fig. 1. Concretely, first, contrast enhancement is performed on the input image to enhance the contrasts between the foreground and the background to improve the saliency of the polyp regions. Then, the enhanced input image is fed into an improved Faster R-CNN network, which is made of a feature extraction subnetwork, an RPN, and a classification subnetwork, to conduct polyp detection. Specifically, to further improve the quality of the feature semantics provided by the feature extraction network, we append an attention module to the feature extraction network to focus on the informative feature channels and weaken the importance of the helpless channels.

3.1. Contrast enhancement

On the surface, the color of the polyp and the stomach itself looks very similar, they show very weak contrasts. They presented very weak contrast so that we may not obtain good features during feature extraction later. Therefore, the first step of the experiment is to enhance the contrast of the ground truth of gastroscopic polyp images.

There are many algorithms for contrast enhancement and are relatively mature. In this study, we refer to the Tone mapping method proposed by Liang[40] to achieve the contrast enhancement of polyp pictures gastroscopic polyps image. Fig. 2 shows the effect of contrast enhancement. In this way, we artificially added a feature to the training of the latter model to improve the discrimination.

In the tone mapping process, we first compute the average brightness of the scene according to the current scene, then choose the suitable brightness domain based on the average brightness, and finally conduct a mapping from the whole scene to this brightness domain to obtain the desired result. At present, there are many methods that can be used to compute the average brightness of the whole scene. In this paper, we use log-average brightness to measure it. The calculation formula is as follows:

$$\overline{L_w} = \frac{1}{N} exp\left(\sum_{x,y} log(\delta + L_w(x,y))\right) \tag{1}$$

$L_w(x,y)$represents the pixel brightness at location (x,y), N denotes the amount of pixels in the scene,δ denotes a very small number for dealing with the case of pure black pixels.

$$L(x,y) = \frac{\alpha}{\overline{LW}} LW(x,y) \tag{2}$$

The above formula is used to map the brightness domain. α* **MERGEFORMAT is determine the brightness tendency of the entire scene. In order to meet the range that the computer can display, the

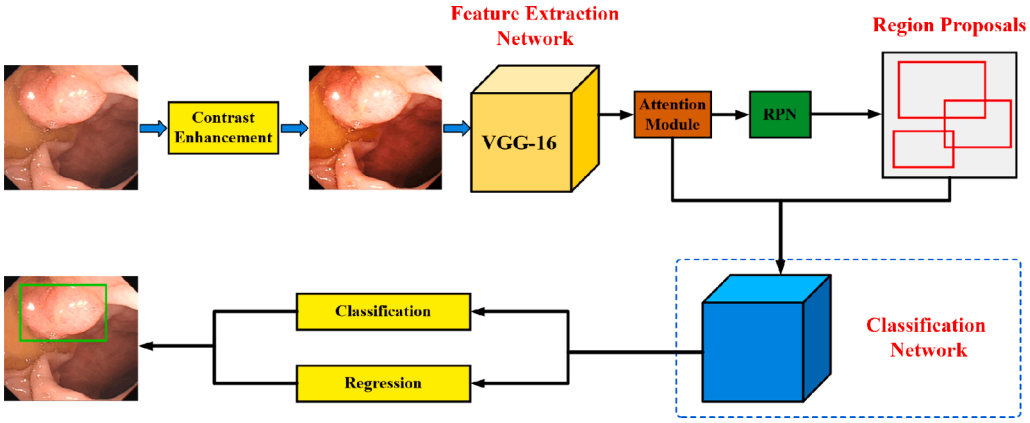

Fig. 1. Illustration of the structure of the designed polyp detection system.

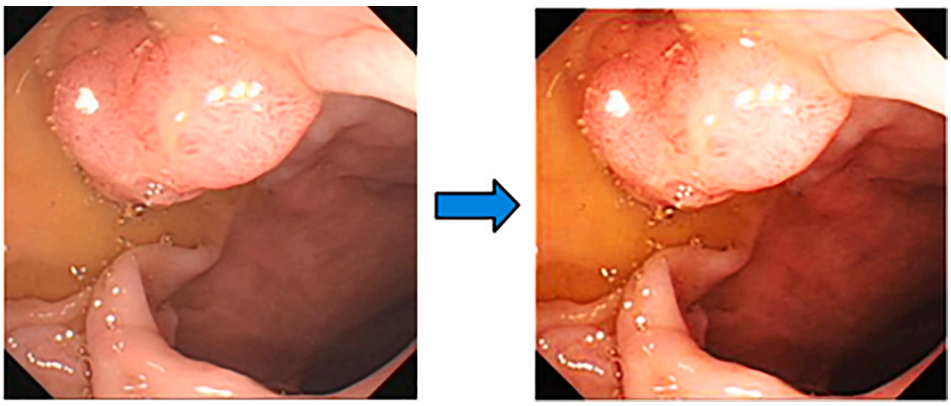

Fig. 2. (a) The raw input image, and (b) the contrast enhanced image.

brightness range must be remapped to the [0,1] interval. The brightness of the [0,1] interval can be simply obtained by the following formula.

$$L_d(x,y) = \frac{L(x,y)}{1+L(x,y)} \tag{3}$$

However, the results obtained in this way are not always satisfactory, so it is generally extended to the following formula. The parameter $L_{polyp}\backslash^*$ *MERGEFORMAT* in the formula is used to control the exposure in the scene. $L_d(x,y)\backslash^*$ *MERGEFORMAT* denotes the pixel brightness value at location (x,y) after the mapping.

$$Ld(x,y) = \frac{L(x,y)(1+\frac{L(x,y)}{L_{polyp}^2})}{1+L(x,y)} \tag{4}$$

3.2. Feature extraction network

In this paper, we leverage VGG-16 to design the feature extraction backbone network. Specifically, the VGG-16 comprises 13 convolutional layers separated into five network stages by four max-pooling layers to extract different-level and different-scale feature maps. Rather than directly feeding the feature maps generated by the feature extraction network into the RPN to produce region proposals, we append an attention module to recalibrate the feature maps to enhance the feature representation capability, as well as generating high-quality region proposals, by explicitly emphasizing the informative feature channels and weaken the saliencies of the less significant ones. The architecture of the attention module is presented in Fig. 3.

For the feature map input to the feature attention module, first, the global average pooling operation is applied channel by channel to transform the input feature map into a channel descriptor. In this way, the statistical properties of a channel can be well obtained with a global perspective. Next, two fully-connected layers are connected to the formed channel descriptor to further exploit channel-wise interdependencies. Specifically, the first fully-connected layer is modulated by the rectified linear unit (ReLU) and the second by the sigmoid function. The second probability-form fully-connected layer provides an attention descriptor, each of whose elements reflects the saliency and importance of the associated channel of the input feature map. This attention descriptor is used as a weight regulator for modifying the input feature map. Finally, the input feature map is recalibrated by multiplying the attention descriptor channel by channel to generate the informative feature emphasized feature map, which is further fed into the RPN to conduct region proposal generation.

3.3. Region proposal network

As shown in Fig. 4, the RPN is made of two parallel branches for conducting classification and regression, respectively. The classification branch functions to classify the generated region proposals into the foreground and the background possibly containing the instances of polyps. The regression branch, therefore, functions to modulate the original anchors centered at a location to regress a fine-grained bounding box of a polyp encapsulated in the anchor. As illustrated in Fig. 4, the classification branch and regression branch can be formulated with two parallel convolutional layers having the kernel sizes of 1×1 performed over the input feature map.

For obtaining region proposals, we predesign a group of anchors having different aspect ratios and varying scales centered at every location of the feature map input to the RPN. In our implementation, an anchor is formulated as a rectangle. Let denote the amount of anchors at a location as k. Specifically, we determine three aspect ratios and four scales, resulting in $k = 12$ different anchors at a location. Totally, for the input feature map having a spatial size of $H \times W$, there are $H \times W \times k$ anchors deployed on the input feature map. If an anchor is formulated as a four-tuple (x, y, w, h), respectively, depicting the location, the width, and the height, a location should produce 4 k values by the regression branch for modulating the location, the width, and the height of an anchor, respectively. Meanwhile, the classification branch should

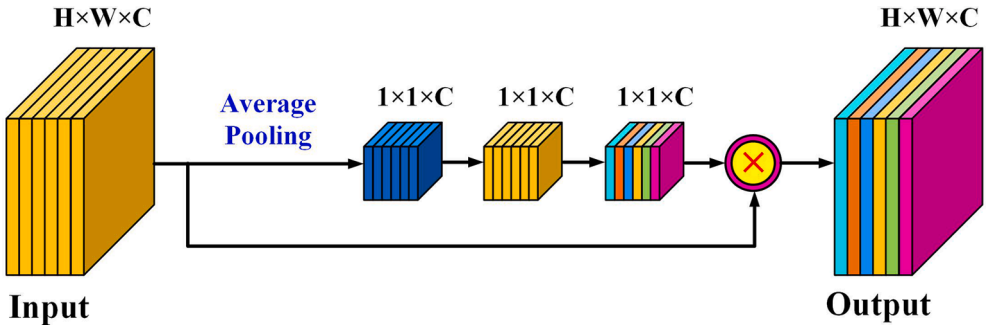

Fig. 3. Illustration of the structure of the attention module.

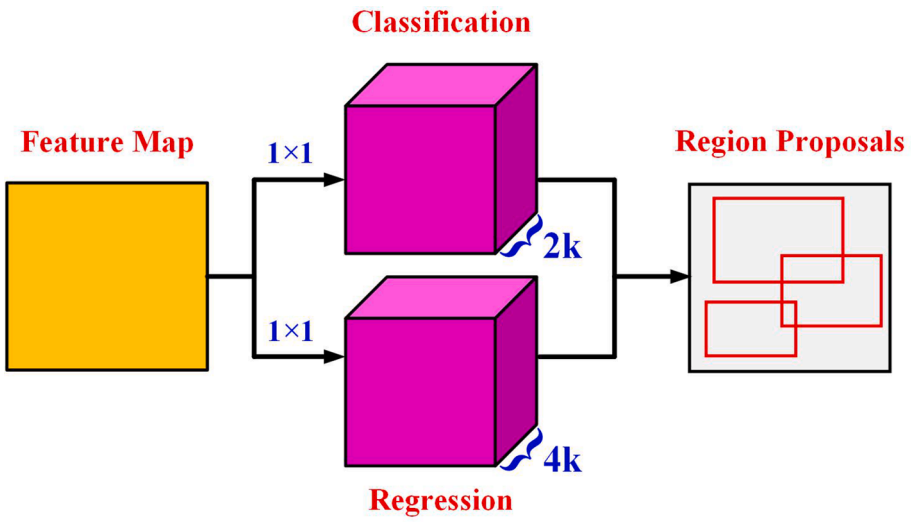

Fig. 4. Illustration of the architecture of the region proposal network.

output 2 k scores at each position for estimating the probabilities of the objectness of each of the adjusted anchors.

In order to generate non-redundant and high-quality region proposals, firstly, the positive anchors confirmed by the classification branch are modulated with the corresponding parameters suggested by the regression branch. Next, these positive anchors are grouped according to their objectness scores in a high-to-low order. Just the first N anchors are used for further verifications. Finally, non-maximum suppression is carried out on the N selected anchors to abandon those redundant overlapped anchors corresponding to the same polyp. Those anchors remained after non-maximum suppression are determined be to the region proposals for polyp recognition.

3.4. Classification network

As presented in Fig. 1, the classification network leverages the region proposals provided by the RPN and the recalibrated feature semantics output by the attention module as the inputs. That is, the feature regions constrained by the region proposals on the feature map are leveraged as the input features fed to the classification network for polyp recognition. However, the sizes of the region proposals are different from each other, which results in problems to handle such different-size region proposals towards poly recognition. For solving this problem, we carry out a region of interest pooling (ROI-pooling) operation on the region proposals to transform them into the same size to ignore their size variations. To this end, we equally partition a region proposal into $N_w \times N_h$ sub-blocks with respect to the height and width sides of the region proposal. Next, for every sub-block, we carry out the max-pooling operation on the feature semantics in the sub-block to select the most salient feature. Finally, the values from all the sub-blocks obtained through max-pooling construct a fixed-size feature map. The fixed-size feature map is used as the ingredient fed into the classification network.

The classification network comprises two convolutional layers, two fully-connected layers, and two parallel sibling branches for, respectively, region proposal classification and fine bounding box regression. The classification branch is a binary softmax classification layer for, respectively, distinguishing the polyp and the background. Once a region proposal is predicted to encapsulate a polyp, the regression branch outputs a tetrad (x, y, w, h), denoting the fine-grained bounding box of the polyp in the input image domain based on the region proposal.

3.5. Loss function

For training the entire network, we define ta multi-task loss function combining a classification loss item L_{cls} and a regression loss item L_{reg} as follows:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{reg} \tag{5}$$

where λ_1 and λ_2 are the regularization factors functioning to adjust the raitos of the two loss items. The classification loss item L_{cls} is defined to be the Focal Loss as follows:

$$L_{cls} = \sum_i \left[-p_i^*(1-p_i)^2 log(p_i) - (1-p_i^*)p_i^2 log(1-p_i) \right] \tag{6}$$

where p_i* is the ground-truth probability prediction for region proposal i, specifically, the value is 1 if the region proposal contains a polyp and is 0 otherwise. p_i is the predicted value, denoting the probability of region proposal i containing a polyp. The regression loss term L_{reg} is formulated with a scale invariant parameterization scheme as follows:

$$L_{reg} = \sum_i \sum_{t \in \{d_x, d_y, d_w, d_h\}, t^* \in \{d_x^*, d_y^*, d_w^*, d_h^*\}} L_1(t - t^*) \tag{7}$$

where $\{d_x, d_y, d_w, d_h\}$ denote the predicted parameters for regressing region proposal i, and $\{d_x$*, d_y*, d_w*, d_h*$\}$ denote the ground-truth regression parameters for the region proposal. They are calculated as follows:

$$\begin{aligned} d_x &= (x - x_r)/w_r, \ d_y = (y - y_r)/h_r \\ d_w &= log(w/w_r), \ d_h = log(h/h_r) \end{aligned} \tag{8}$$

$$\begin{aligned} d_x^* &= (x^* - x_r)/w_r, \ d_y^* = (y^* - y_r)/h_r \\ d_w^* &= log(w^*/w_r), \ d_h^* = log(h^*/h_r) \end{aligned} \tag{9}$$

where (x, y, w, h), (x_r, y_r, w_r, h_r), and (x*, y*, w*, h*), respectively, mean the predicted bounding box, the region proposal, and the ground-truth bounding box. $L_1(\bullet)$ is the smooth L_1 function, which is formulated as follows:

$$L_1(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geqslant 1 \end{cases} \tag{10}$$

4. Results and discussions

4.1. Dataset

Aiming to evaluate the accuracy of the algorithm, we collected 1000 colonoscopy samples, each of which has the same size of 224×224 pixels. In the process of experiment, we adopted a 5-fold cross validation

to divide the data set. The 1000 images were randomly divided into 5 parts. Each time, 800 samples were determined randomly to form the training set of the model, while the rest 200 samples were leveraged for evaluating the model. Fig. 5(a) shows some polyp sample images. Fig. 5 (b) illustrates the ground truths using bounding boxes.

4.2. Network configuration

The proposed method was trained by stochastic gradient descent (SGD) and backpropagation in an end-to-end way on a cloud computing platform configured with eight 16-GB GPUs, a 16-core CPU, and a 64-GB memory. In our implementation, the RPN and the classification network depend on features from the feature extraction backbone network. Meanwhile, the output of the RPN also serves as parts of the classification network. In this regard, first, we constructed the RPN alongside the feature extraction network. Before carrying out training, the network layers were randomly initialized by drawing parameters from a zero-mean Gaussian distribution with a standard deviation of 0.01. We organized the training samples into batches, each of which had two images on a GPU, and they were trained for 1000 epochs. Specifically, the initial learning rate was set to be 0.01 for the first 800 epochs and reduced to be 0.001 for the rest 200 epochs. When the RPN was constructed, its network parameters were fixed and used to further construct the classification network by using the region proposals generated by the RPN. Likewise, before model training, the network layers were randomly initialized by drawing parameters from a zero-mean Gaussian distribution with a standard deviation of 0.01. The region proposals were batched with 50 samples per batch on a GPU and they were trained for 800 epochs. Specifically, we configured the initial learning as 0.01 for the first 600 epochs and modulated it to 0.001 for the rest 200 epochs. When the classification network was constructed, we finally jointly fine-tuned the entire network for another 200 epochs with the network parameters determined in the previous stages. Similarly, the initial learning rate of 0.001 was used in the first 150 epochs and the learning rate of 0.0001 was applied to the last 50 epochs.

4.3. Polyp object detection

For a test image, first, we enhanced its contrasts using the Tone mapping method to better distinguish the textures of the polyps from the background. Then, the contrast-enhanced image was fed into the feature extraction network to generate high-level features, which were further used as the input to the RPN to generate region proposals that possibly contained polyps. Finally, the generated region proposals were classified by the classification network to recognition the polyps.

For quantitatively evaluating the polyp detection performance, the following three evaluation metrics were leveraged: precision, recall, and F-score. Specially, precision evaluates the ability to distinguish false targets, that is, it is represented as the proportion of the true positives over all the detection results. Recall assesses the ability to recognize true targets, that is, it is represented as the proportion of the true positives over the ground truths. In other words, the higher the values of the precision and recall metrics, the better the performance of the polyp detection method. F-score, therefore, provides an overall evaluation by comprehensively taking into consideration the precision and the recall metrics. The three metrics are formally formulated as follows:

$$precision = \frac{TP}{TP + FP} \tag{11}$$

$$recall = \frac{TP}{TP + FN} \tag{12}$$

$$F - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{13}$$

where *TP*, *FN*, and *FP* are the numbers of the true positives, false negatives, and false positives, respectively. The quantitative evaluation performances on the polyp detection results are shown in Table 1. As reflected in Table 1, the proposed method obtained a very competitive polyp detection accuracy on the test set with a precision of 0.943, a recall of 0.925, and an F-score of 0.934, respectively. The advantageous performance benefitted from the following three aspects: first, by carrying out contrast enhancement on the raw input image, the textures of the polyp regions were reasonably highlighted and better distinguished from the background. Thus, through such preprocessing, the quality of the input image was improved to serve for the polyp detection task. Second, by integrating the attention module over the feature extraction network, the feature maps fed into the RPN were significantly enhanced to concentrate on the useful and informative features and weaken the

Table 1
Performance of polyp detection between different algorithms. (The optimal value of each index is bold.)

Methods	Precision	Recall	F-score
Faster R-CNN	0.916	0.897	0.906
Yolo-v4	0.895	0.876	0.885
CNN	0.908	0.889	0.898
Dense hough transform	0.902	0.885	0.893
MDeNetplus	0.912	0.901	0.906
mRMR-Ensemble Classifier	0.934	0.916	0.925
LCDH	0.892	0.873	0.882
Ours	**0.943**	**0.925**	**0.934**

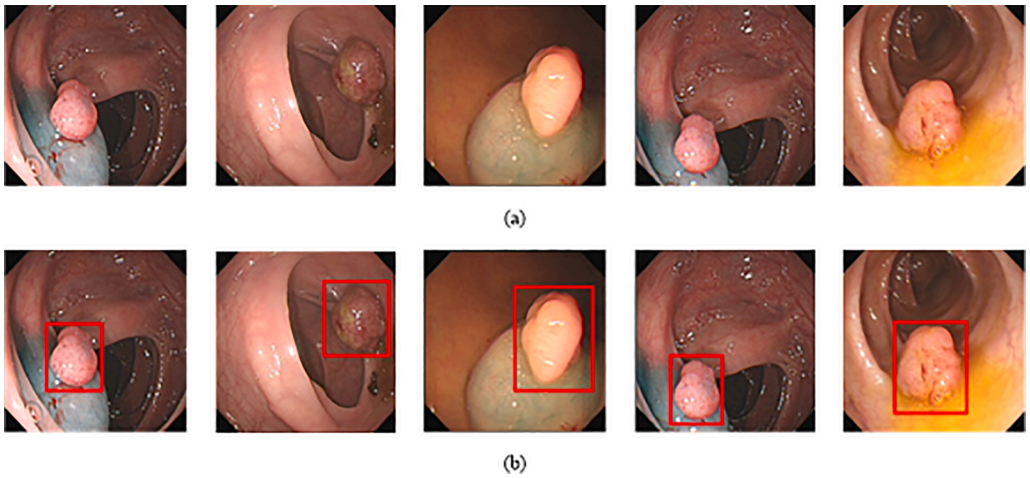

Fig. 5. The polyp object detection dataset. (a) Polyp image samples, (b) ground-truths with bounding boxes.

influences of the less useful ones. Thus, through such feature map recalibration, the robustness and the encoding capability of the extracted features were greatly improved to serve for the generation of high-quality region proposals. Finally, by adopting a two-stage detection framework following the Faster R-CNN architecture, the polyp detection performance was further improved with the assistance of high-quality features and region proposals.

Specifically, for the polyps showing less salient contrasts to their backgrounds, the proposed method can still obtain reasonable detection accuracy because of the preprocessing of contrast enhancement and the integration of the attention module. In addition, for the images containing multiple adjacent polyps, the proposed method performed effectively to correctly recognize and accurately locate them by using the region proposal based two-stage detection framework. However, for some polyps showing extremely low contrasts to their backgrounds and some polyps of extremely small sizes, the proposed method failed to correctly detect them. In the whole, the proposed method can effectively handle colonoscopy images of different image qualities and containing polyps of varying conditions.

For visual inspections, Figs. 6-8 present three subsets of polyp detection results from the test set. As shown by the images containing single polyps in Fig. 6, the polyps of different sizes, particularly the small-size polyps, and texture conditions were correctly detected. Specifically, as shown by the images in Fig. 7, the polyps showing low contrasts with the background were also accurately located in the images and their bounding boxes were correctly delineated.

In addition, due to the multiple polyps in the colorectal, there are multiple polyps in some images. Fortunately, our algorithm can effectively deal with this situation in the process of polyp detection. The detection results of multiple polyps in the test set are shown in Fig. 8. As illustrated in Fig. 8, the multiple polyps showing varying sizes and conditions were detected with promising accuracies. However, for some polyps with extremely small sizes and terribly low contrasts to their surroundings, our proposed method failed to correctly recognize them. In the whole, our proposed model behaved effectively on detecting polyps with different conditions.

4.4. Comparisons with state-of-the-art methods

For further assessing the performance of the proposed polyp detection method, we experimented a group of comparative evaluations with some state-of-the-art models. The following seven models were leveraged for performance evaluations: Faster R-CNN, Yolo-v4, CNN [33], dense hough transform [34], MDeNetplus[36], mRMR + Ensemble Classifier[38], and LCDH[39]. Faster R-CNN and CNN adopted a two-stage framework to conduct object detection by first generating a group of dense region proposals and then classifying the region proposals into the true targets and the false alarms. Yolo-v4 adopted a one-stage framework that accomplished feature abstraction and object recognition by a single-forward network. The dense hough transform method behaved effectively in dealing with occluded polyps and reducing the number of false alarms through the dense hough transform technique. MDeNetplus was based on the concept of multi-layer feature fusion through the feedback skip connections between different layers of the decoders. It improves the accuracy by merging different semantic features iteratively and hierarchically. mRMR-Ensemble Classifier mainly extracted color wavelet features and convolution neural network features from endoscopic video frames, which were further dimension-reduced through Minimum redundancy maximum relevance (nRMR) and classified using an ensemble classifier. LCDH mainly used a bag-of-visual-words representation to realize feature encoding on the basis of the histogram of local color difference. The recognition of polyps was finalized with the combination of a spatial matching pyramid model and a support vector machine.

For fair comparisons, these methods were constructed by using the same training set and examined on the same test set. Table 1 details the quantitative evaluation results on the test set obtained by these methods based on the precision, recall, and F-score metrics.

As reported in Table 1, Faster R-CNN, MDeNetplus, and mRMR-Ensemble Classifier performed superiorly over the other models. For example, the mRMR-Ensemble Classifier obtained an accuracy promotion by about 0.043 with regard to the F-score metric compared with the LCDH. Specifically, the superior performance of the Faster R-CNN benefitted from the implementation of the two-stage detection framework, which behaved better than that of the one-stage framework used in Yolo-v4 when handling multiple tightly adjacent polyps. In addition, by comprehensively taking into account multi-layer features of different semantics for improving the feature representation quality and by fusing different-order deep features and color wavelet features for feature complement, the MDeNetplus and mRMR-Ensemble Classifier also achieved competitive performances towards polyp detection. In contrast, by only leveraging low-order image features, the feature encoding capability of the LCDH was quite limited, thereby resulting in a relative lower detection performance. Comparatively, our proposed method outperformed the compared methods with regard to all the quantitative evaluation metrics. Specifically, an overall performance improvement of about 0.049 with regard to the F-score was obtained compared with the one-stage deep learning model Yolo-v4 and an overall performance enhancement of about 0.052 with respect to the F-score was obtained compared with the hand-crafted feature-based

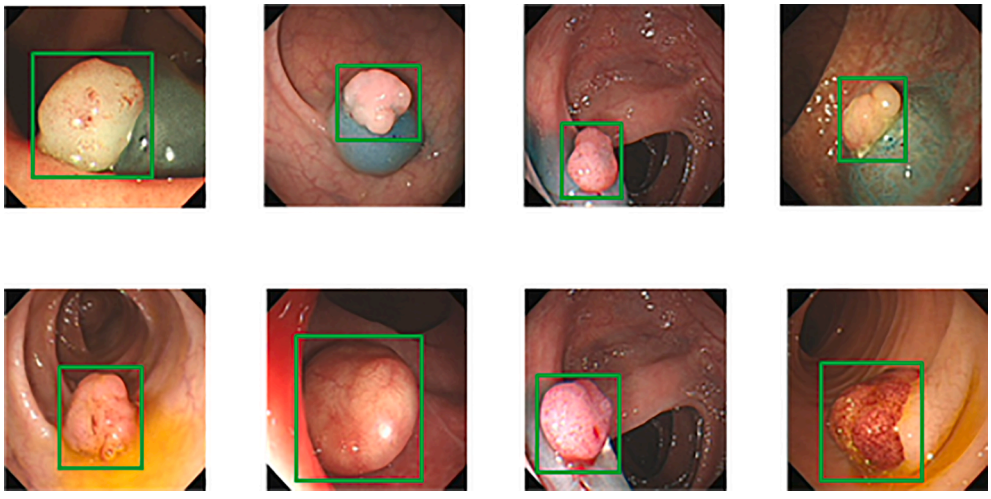

Fig. 6. A subset of single polyp detection results.

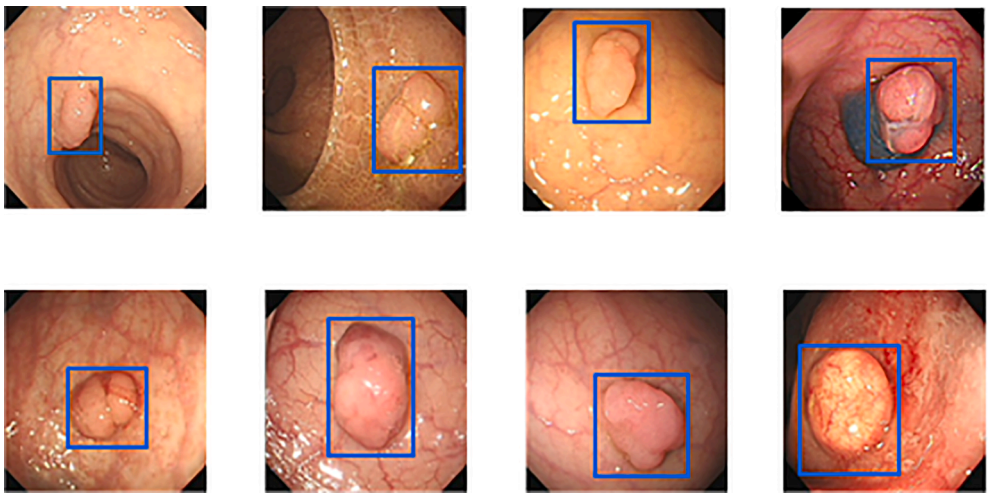

Fig. 7. A subset of detected single polyps showing low contrasts to the background.

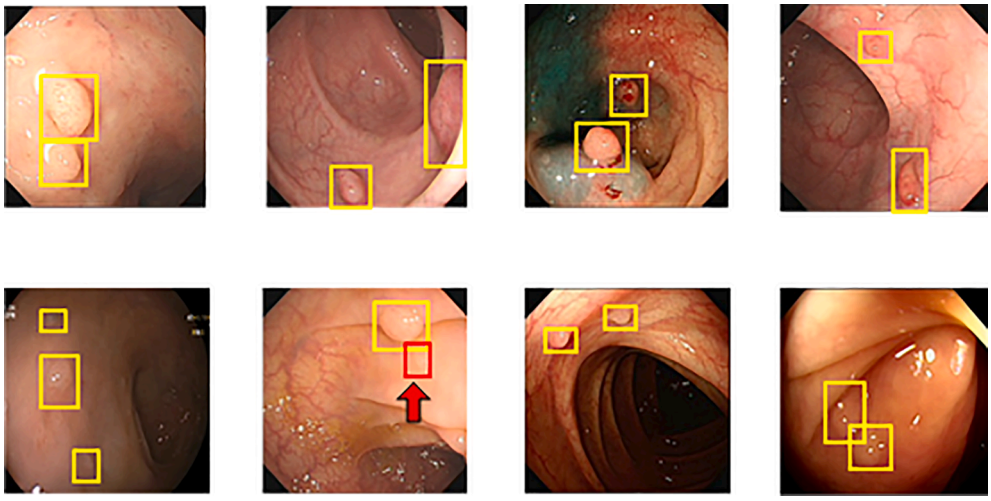

Fig. 8. A subset of multiple polyp detection results.

model LCDH.

The advantageous performance of our proposed method was due to the effective preprocessing strategy, the feature enhancement module, and the anchor-based two-stage detection architecture. Via comparative studies, we confirmed that our proposed method provided a suitable and effective solution to polyp detection from colonoscopy images.

5. Conclusion

CRC has become one of the diseases that seriously endanger the healthy life of the Chinese people. In this paper, we proposed a self-attention based Faster R-CNN architecture for detecting polyps from colonoscopy images. By performing contrast enhancement on the input image, the saliencies of the polyp regions were positively highlighted. By integrating the self-attention module over the feature extraction network, the quality and the representation capability of the output feature map were significantly improved. By adopting a two-stage detection strategy with the pre-generation of region proposals and the post-recognition of polyps, the detection accuracy is reasonably upgraded. Experimental results showed that this method not only can achieve a high detection rate for single polyp images, but also can effectively realize the correct recognition and accurate location of multiple polyps. Comparative studies also demonstrated the advantageous performance and applicability of the proposed model in poly detection tasks.

CRediT authorship contribution statement

Bo-Lun Chen: Conceptualization, Methodology, Validation, Writing - original draft, Funding acquisition. **Jing-Jing Wan:** Conceptualization, Investigation, Data curation, Writing - original draft. **Tai-Yue Chen:** Conceptualization, Validation, Software, Writing - review & editing. **Yong-Tao Yu:** Conceptualization, Software, Writing - review & editing, Funding acquisition. **Min Ji:** Conceptualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under grant No. 61602202 and 62076107, Natural Science Foundation of Jiangsu Province under contracts No. BK20160428 and Natural Science Foundation of Education Department of Jiangsu Province under contract No. 20KJA520008. Six talent peaks project in Jiangsu Province (Grant No. XYDXX-034). China Scholarship

Council also supported this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2021.103019.

References

[1] American cancer Society, Cancer Facts and Figures: 2017, Atlanta, Georgia, 2017.

[2] C.E. Bailey, C.-Y. Hu, Y.N. You, B.K. Bednarski, M.A. Rodriguez-Bigas, J. M. Skibber, S.B. Cantor, G.J. Chang, Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975–2010[J], JAMA surgery 150 (1) (2015) 17, https://doi.org/10.1001/jamasurg.2014.1756.

[3] D.K. Rex, C.R. Boland, J.A. Dominitz, et al., Colorectal cancer screening: recommendations for physicians and patients from the US Multi-Society Task Force on Colorectal Cancer[J], The American journal of gastroenterology. 112 (7) (2017) 1016.

[4] C.A. Doubeni, D.A. Corley, V.P. Quinn, C.D. Jensen, A.G. Zauber, M. Goodman, J. R. Johnson, S.J. Mehta, T.A. Becerra, W.K. Zhao, J. Schottinger, V.P. Doria-Rose, T. R. Levin, N.S. Weiss, R.H. Fletcher, Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: a large community-based study[J], Gut 67 (2) (2018) 291–298.

[5] 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, C.P. Quesenberry, Adenoma detection rate and risk of colorectal cancer and death[J], New England Journal of Medicine 370 (14) (2014) 1298–1306.

[6] C. Burke, V. Kaul, H. Pohl, Polyp resection and removal procedures: insights from the 2017 Digestive Disease Week[J], Gastroenterology & hepatology 13 (19 Suppl 2) (2017) 1.

[7] N. Mahmud, J. Cohen, K. Tsourides, et al., Computer vision and augmented reality in gastrointestinal endoscopy[J], Gastroenterology report 3 (3) (2015) 179–184.

[8] P.-J. Chen, M.-C. Lin, M.-J. Lai, J.-C. Lin, H.-S. Lu, V.S. Tseng, Accurate classification of diminutive colorectal polyps using computer-aided analysis[J], Gastroenterology 154 (3) (2018) 568–575.

[9] Y. Tian, L.Z.C.T. Pu, R. Singh, et al., One-stage five-class polyp detection and classification[C], in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 70–73.

[10] S. Hwang, J.H. Oh, W. Tavanapong, et al., Polyp detection in colonoscopy video using elliptical shape feature[C], in: 2007 IEEE International Conference on Image Processing, IEEE, 2007, pp. II-465–II-468.

[11] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks[C], in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 79–83.

[12] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information[J], IEEE Transactions on Medical Imaging 35 (2) (2016) 630–644.

[13] D. Wang, N. Zhang, X. Sun, et al., AFP-Net: Realtime Anchor-Free Polyp Detection in Colonoscopy[C], in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2019, pp. 636–643.

[14] P. Sasmal, Y. Iwahori, M.K. Bhuyan, et al., Active contour segmentation of polyps in capsule endoscopic images[C], in: 2018International Conference on Signals and Systems (ICSigSys), IEEE, 2018, pp. 201–204.

[15] J. Bernal, N. Tajkbaksh, F.J. Sanchez, B.J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Cordova, C. Sanchez-Montes, S.R. Gurudu, G. Fernandez-Esparrach, X. Dray, J. Liang, A. Histace, Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge[J], IEEE Transactions on Medical Imaging 36 (6) (2017) 1231–1249.

[16] X. Mo, K. Tao, Q. Wang, et al., An efficient approach for polyps detection in endoscopic videos based on faster R-CNN[C], in: 201824th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3929–3934.

[17] Billah M, Waheed S, Rahman M M. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features[J]. International journal of biomedical imaging, 2017.

[18] H.A. Qadir, Y. Shin, J. Solhusvik, et al., Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better?[C], in: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), IEEE, 2019, pp. 1–6.

[19] H.A. Qadir, J. Solhusvik, J. Bergsland, L. Aabakken, I. Balasingham, A Framework With a Fully Convolutional Neural Network for Semi-Automatic Colon Polyp Annotation[J], IEEE Access 7 (2019) 169537–169547.

[20] A.A. Pozdeev, N.A. Obukhova, A.A. Motyko, Automatic analysis of endoscopic images for polyps detection and segmentation[C], in: 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), IEEE, 2019, pp. 1216–1220.

[21] H. Zheng, H. Chen, J. Huang, et al., Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained CNN[C], in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 79–82.

[22] R. Zhang, Y. Zheng, C.C.Y. Poon, D. Shen, J.Y.W. Lau, Polyp Detection during Colonoscopy using a Regression-based Convolutional Neural Network with a Tracker[J], Pattern Recognition 83 (2018) 209–219.

[23] M. Liu, J. Jiang, Z. Wang, Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network[J], IEEE Access 7 (2019) 75058–75066.

[24] X.u. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, J. Si, S. Diciotti, Real-time gastric polyp detection using convolutional neural networks [J], PLoS ONE 14 (3) (2019) e0214133, https://doi.org/10.1371/journal.pone.0214133.

[25] M. Bagheri, M. Mohrekesh, M. Tehrani, et al., Deep Neural Network based Polyp Segmentation in Colonoscopy Images using a Combination of Color Spaces[C], in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 6742–6745.

[26] A. Tashk, E. Nadimi, An Innovative Polyp Detection Method from Colon Capsule Endoscopy Images Based on A Novel Combination of RCNN and DRLSE[C], in: 2020 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2020, pp. 1–6.

[27] Jia X, Mai X, Cui Y, et al. Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction[J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(3): 1570-1584.

[28] F.L. Henriksen, R. Jensen, H.K. Stensland, et al., Performance of data enhancements and training optimization for neural network: A polyp detection case study[C], in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2019, pp. 287–293.

[29] O. Bardhi, D. Sierra-Sosa, B. Garcia-Zapirain, et al., Automatic colon polyp detection using Convolutional encoder-decoder model[C], in: 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2017, pp. 445–448.

[30] L. Yu, H. Chen, Q.i. Dou, J. Qin, P.A. Heng, Integrating Online and Offline Three-Dimensional Deep Learning for Automated Polyp Detection in Colonoscopy Videos [J], IEEE J Biomed Health Inform 21 (1) (2017) 65–75.

[31] Y. Shin, H.A. Qadir, L. Aabakken, J. Bergsland, I. Balasingham, Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches[J], IEEE Access 6 (2018) 40950–40962.

[32] Y. Ma, X. Chen, B. Sun, Polyp Detection in Colonoscopy Videos by Bootstrapping Via Temporal Consistency[C]. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020.

[33] H.A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, Y. Shin, Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video[J], IEEE Journal of Biomedical and Health Informatics 24 (1) (2020) 180–193.

[34] L. Ruiz, L. Guayacán, F. Martínez, Automatic polyp detection from a regional appearance model and a robust dense Hough coding[C]2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), IEEE (2019) 1–5.

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

[36] H.A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, I. Balasingham, Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction[J], Medical Image Analysis 68 (2021) 101897, https://doi.org/10.1016/j.media.2020.101897.

[37] J. Xu, R. Zhao, Y. Yu, Q. Zhang, X. Bian, J. Wang, Z. Ge, D. Qian, Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit[J], Biomedical Signal Processing and Control 66 (2021) 102503, https://doi.org/10.1016/j.bspc.2021.102503.

[38] M. Billah, S. Waheed, Minimum redundancy maximum relevance (mRMR) based feature selection from endoscopic images for automatic gastrointestinal polyp detection[J], Multimedia Tools and Applications 79 (33-34) (2020) 23633–23643.

[39] J. Yang, L. Chang, S. Li, X. He, T. Zhu, WCE polyp detection based on novel feature descriptor with normalized variance locality-constrained linear coding[J], International Journal of Computer Assisted Radiology and Surgery 15 (8) (2020) 1291–1302.

[40] Liang Z, Xu J, Zhang D, et al. A hybrid l1-l0 layer decomposition model for tone mapping[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4758-4766.