IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 18, 2022, accepted April 23, 2022, date of publication April 29, 2022, date of current version May 6, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3171238*

# A New Automated Polyp Detection Network MP-FSSD in WCE and Colonoscopy Images Based Fusion Single Shot Multibox Detector and Transfer Learning

**MERYEM SOUAIDI**[1] **AND MOHAMED EL ANSARI**[1,2]

[1]LabSIV, Computer Science Department, Faculty of Sciences, University of Ibn Zohr, Agadir 80000, Morocco
[2]Informatics and Applications Laboratory, Computer Science Department, Faculty of Sciences, University of Moulay Ismail, Meknès 50050, Morocco

Corresponding author: Meryem Souaidi (souaidi.meryem@gmail.com)

This work was supported in part by the Ministry of National Education by Vocational Training; in part by the Higher Education and Scientific Research through the Ministry of Industry, Trade, and Green and Digital Economy; in part by the Digital Development Agency (ADD); and in part by the National Center for Scientific and Technical Research (CNRST) under Project ALKHAWARIZMI/2020/20.

**ABSTRACT** Small polyp region detection in wireless capsule endoscopy (WCE) images is a challenging task in computer vision owing to two major problems: its variation in terms of shape, texture, and size, and the low illumination in the gastrointestinal tract. This study proposes a multiscale pyramidal fusion single-shot multibox detector network (MP-FSSD) to detect small polyp regions in WCE or colonoscopy frames, or both, with respect to the precision-vs-speed trade-off as the base architecture. We investigated deep transfer learning by transferring knowledge to polyp images, thereby enabling the extraction of highly representative features and contextual information from the FSSD. First, an edge-pooling layer was embedded in the shallow part of the network. Subsequently, the feature maps from different layers and scales were transformed to match their sizes. A concatenation module was introduced to integrate the feature maps from different layers, which were delivered to the next layer, followed by downsampling blocks to generate new pyramidal layers. Finally, the feature maps were fed to the multibox detectors to predict the final detection results. Experimentally, we maintained the same hyperparameters for both datasets for a fair comparison. The proposed MP-FSSD network exceeded FSSD by 3.62% in terms of mean average precision (mAP). The testing speed of 62.5 FPS is superior to that of the competitor detection methods. The proposal demonstrates that deep learning has much room for development in the field of gastrointestinal image detection.

**INDEX TERMS** Deep transfer learning, edge pooling, feature maps fusion, image augmentation, polyp, single-shot multibox detector (SSD), wireless capsule endoscopy images (WCE).

## I. INTRODUCTION

Based on recent statistical data, gastrointestinal cancers are the leading cause of death worldwide [1]. Unfortunately, it is estimated that the number of patients affected by this disease has increased considerably in recent years [2], [3]. Adenomatous polyps are one of the most common types of colorectal cancer that occur due to growth of glandular tissue in the colonic mucosa. To detect polyp regions in their early stages and remove them before they become malignant in advance, doctors need to visualize the GI tract directly [4].

Wireless capsule endoscopy (WCE) has become an active research area as an advanced tool for visualization of the gastrointestinal tract [5]–[7]. In contrast to traditional endoscopes, this non-invasive technique enables physicians to explore the GI tract with full visualization from the inside, without pain or sedation. In fact, WCE produces an approximation of 55,000 images per patient [8], [9]. However, only approximately 5% of the frames contained lesions. The large amount of data makes the detection process a tedious task for physicians to manually locate the polyp regions in each WCE frame. Automating the detection of frames containing specific lesions in WCE videos would relieve gastroenterologists of the arduous task of reviewing the entire video before

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

making a diagnosis. Accurate detection of polyp regions is more difficult owing to their complicated characteristics (shape, texture, size, and morphology). Thus, small polyp regions could not be detected by the naked eye. To reach a precise detection process, specialists should meet and commonly agree on the ground truth of the polyp frame. The proposed solution aims to provide an automated computer detection system that has some knowledge of the specialists at hand without requiring their physical presence. Therefore, clinicians can make the right decision by decreasing human error [9], [10]. In recent years, deep learning (DL) has received a tremendous amount of attention in the field of medical image analysis owing to its superior performance in image classification when compared to deep neural networks [11]. A systematic review recently published in 2019 proved that deep-learning-based models match those of healthcare professionals [12]. Currently, deep learning applications have focused on polyp abnormality detection or classification on colonoscopy (not capsule endoscopic images), or both [13]–[15].

Region-based object detection is one of the primary focuses of computer vision. Many detectors based on ConvNets have been proposed to address the trade-off between accuracy and speed in object detection [16]–[18]. Scale variations remain a critical challenge for all the detectors. Several attempts have been made to solve the multiscale object detection problem. Some of these studies have proposed applying a ConvNet to different image scales to generate new scale feature maps that are unreliable. Other object detection methods such as Faster R-CNN [19] and RFCN [20] used a fixed receptive field size by selecting only one scale feature map and creating anchors with different scales for multi-scale object detection, which is an inefficient way to detect small polyp regions in a fast way. As an intuitive solution, the FPN [21] and DSSD [22] architectures have been proposed to fuse features layer by layer. However, this concatenation procedure sacrifices a significant amount of speed. Recent methods adopt the single-shot multibox detector (SSD [23]) as a baseline in their way, owing to its improvement in speed. However, precision vs. speed is still the core trade-off of small-object detection. Furthermore, the contradiction between object recognition and location remains a challenge for object detectors based on ConvNets. Although feature maps can represent more semantic information with translation invariance using deeper ConvNet, the core process is beneficial to object recognition but detrimental to object location. In fact, the location information will not be significantly lost either in the shallow or deeper layers using the fast SSD detector for small or large polyp regions, respectively. However, feature maps of small polyp regions generated by shallow layers lack sufficient semantic information, which may result in performance degradation overtime.

This paper proposes a multiscale pyramidal fusion single-shot multibox detector network (MP-FSSD) to tackle the problems of scale variations and the lack of contextual information for small polyp detection in WCE frames. To achieve this, the proposed method embeds edge pooling into the shallow part of the network. Then, it adds a lightweight feature fusion module to the standard SSD (VGG16 as the backbone) with respect to the precision-vs-speed trade-off as the base architecture. Two feature fusion frameworks, concatenation module and element-sum module, are used, in which features from different layers and scales are projected and concatenated together, followed by a batch normalization layer [24] to normalize the feature values. Finally, down-sampling blocks are applied to generate a new feature pyramid that is fed to the multibox detectors to produce the final detection results. To prove the efficiency of the proposed architecture, the MP-FSSD was evaluated on WCE or colonoscopy polyp datasets, or both. The experimental results indicate that the MP-FSSD model obtains a higher mAP on WCE and colonoscopy polyp datasets than the conventional SSD with a gain of 16.2 and 3.62 points, respectively, especially for small polyp regions with a slight speed drop. Furthermore, the MP-FSSD achieved encouraging results compared to state-of-the-art object detectors based on the VGG network in terms of speed and performance.

The remainder of this paper is organized as follows. Section II describes related studies. Section III describes the proposed MP-FSSD model in detail. Section IV reports the experimental results and compares them with those of other models. The conclusions are presented in Section V.

## II. STATE OF THE ART

Many attempts have been made to address polyp detection tasks. The last few years have seen considerable growth in the investigation of handcrafted features to characterize images that capture attributes of color, texture, shape, and contrast only for WCE polyp classification purposes [5]. Li *et al.* [25] first used a combination of wavelets and uniform LBP along with an SVM as a classifier in [25]. They subsequently improved the traditional SIFT by combining different textural features [26]. In a previous study [5], the T-CWT-based method combined with gamma parameters was proposed to discriminate polyp regions from augmented WCE datasets. The authors of [27] proposed a framework-based pyramid histogram of oriented gradient (PHOG), in which the local polyp shape features are extracted using PHOG, and the local texture features are extracted using FWLBP. Subsequently, different performances metrics are used to evaluate the proposed approach. Owing to an unclear understanding of biological mechanisms, handcrafted features only encode part of the frames and neglect the intrinsic information of the WCE images. Therefore, handcrafted features are unsuitable for polyp WCE images. To overcome this shortcoming, several attempts have been made at colonoscopy polyp abnormality classification using existing deep learning frameworks, such as VGGNet [28], GoogleNet [29] and ResNet [30]. Retraining architectures from scratch in the context of colonoscopies leads to reasonable but insufficient results owing to the limited size of medical datasets. Therefore, the use of pre-trained models with proper fine-tuned configurations leads to very good results in many fields, especially in medical applications [31]. Generally, transfer learning schemes are used to overcome insufficient training samples. Even if the pre-trained model categories are quite different compared to the medical imaging, it has been shown that they can be used in the context of colonoscopy/endoscopy polyp recognition tasks [32], [33]. This is the main motivation for involving transfer learning techniques for wireless capsule endoscopy polyp detection tasks.

To be more accurate, polyp classification as normal or abnormal is beyond the scope of the current systematic report. Some of the published works only performed polyp detection without localization, meaning that they reported systems aimed at predicting whether there are one (or more) polyps in a given video frame, but without indicating the exact location of the polyp. Our main interest in this review is to locate polyps in WCE frames, showing their positions with a square bounding box, whereas the classification is performed once the presence of polyp abnormality is confirmed. Therefore, state-of-the-art methods for the detection, localization, and segmentation of WCE polyps based on deep learning approaches are compared, showing their advantages and disadvantages in identifying the most auspicious trends. Pre-existing domain-specific object detection methods based on deep learning usually can be divided into two categories, the first one is two-stage algorithms based on region proposal such as Faster R-CNN [34]. The other is a one-stage algorithm based on regression, such as YOLO [35] and SSD [36]. In two-stage detectors, region proposals are extracted from the input images using the region-selection algorithm. They are then classified and position-adjusted to output the target detection results. Although this type of algorithm has a high localization and object recognition accuracy, its detection speed is slow, making it difficult to meet the real-time requirements of polyp detection. In contrast, the one-stage algorithms propose predicted boxes from the input images directly without the region proposal step. Therefore, it achieved a high inference speed. Tian *et al.* [37] proposed a single-stage detection and classification approach for five classes of polyp abnormalities. The model is trained in a single process, making the training and reasoning process simple and faster. Tajbakhsh *et al.* [38] proposed a polyp-detection algorithm based on three independent methods of image representation and convolution neural network. Polyp localization is achieved by incorporating various characteristics at multiple scales, such as shape, texture, color, and temporal information. Wang *et al.* [39] proposed an algorithm based on a context enhancement module and cosine ground-truth projection for an accurate polyp detection process. The authors of [40] proposed an algorithm for polyp segmentation from endoscopic images in which they used principal component tracking (PCP) to remove the specular region in the image. Thus, they activated the contour (AC) model to locate the polyp region in each frame. An improved mask R-CNN framework was presented in [41]. The authors used different CNN architectures for the feature extraction backbone network. Subsequently, an integrated method is proposed for polyp detection and segmentation. Zheng *et al.* [42] utilized optical flow and online training to propose a two stages CNN polyp detection algorithm. As a primary step, they used a U-Net network to detect and locate polyps for single-frame target detection. Then, a motion regression model and an effective online training CNN model were established using temporal information and optical flow to track polyps. Ruikai *et al.* [43] presented a regression-based convolutional neural network (CNN) architecture, in which a fast object detection algorithm named ResYOLO was pre-trained and fine-tuned to properly extract the spatial features of intestinal polyps. Subsequently, they optimized the detection results of the ResYOLO output based on temporal information through an efficient convolution operator tracker. The authors of [34] proposed a self-attention-based faster R-CNN architecture for detecting polyps from colonoscopy images. They highlighted polyp saliency regions by performing a contrast enhancement. Then, they integrated the self-attention module over the feature extraction network and adopted a two-stage detection strategy with the pre-generation of region proposals and the post-recognition of polyps to improve the accuracy. Liu *et al.* [44] investigated a single-shot detector (SSD) framework for detecting polyps in colonoscopy videos. ResNet50 and VGG16 models were used as feature extraction backbone networks to evaluate their performance. TASHK *et al.* [45] proposed a polyp detection method in which an improved version of the CNN algorithm was used to locate polyps in an image. They then used DRLSE to automatically segment local polyps. Misawa *et al.* [46] presented a polyp detection system based on YOLOv3. It was trained on 56,668 training images collected from five medical centers and achieved real-time detection with over 90% sensitivity and specificity. A polyp segmentation method for colonoscopy images based on the convolutional neural network was presented by Bagheri *et al.* [47], in which they used the LinkNet network to improve the quality of polyp segmentation. This method uses R and G channels from the RGB and b* channels from the CIE-L*a*b* color space as the input of the network in the design process. Jia *et al.* [48] proposed a two-stage framework based on deep learning for automatic polyp recognition in colonoscopic images.

To address the problems of the SSD algorithm for small polyp detection, many studies have adopted the improvement indicated by the deconvolutional SSD (DSSD) [22]. Zhang *et al.* [49] proposed an enhanced SSD architecture called SSD-GPNet for detecting gastric polyps. Pooling methods were applied to the feature pyramid network to reuse the lost useful information caused by the max-pooling layers. The model takes advantage of the multiresolution features extracted in the feature pyramid architecture by integrating the feature map with the deconvolution of high-level feature maps. Although some attempts have been made to simplify the DSSD architecture and improve the accuracy of small polyp region detection, the addition of deconvolution

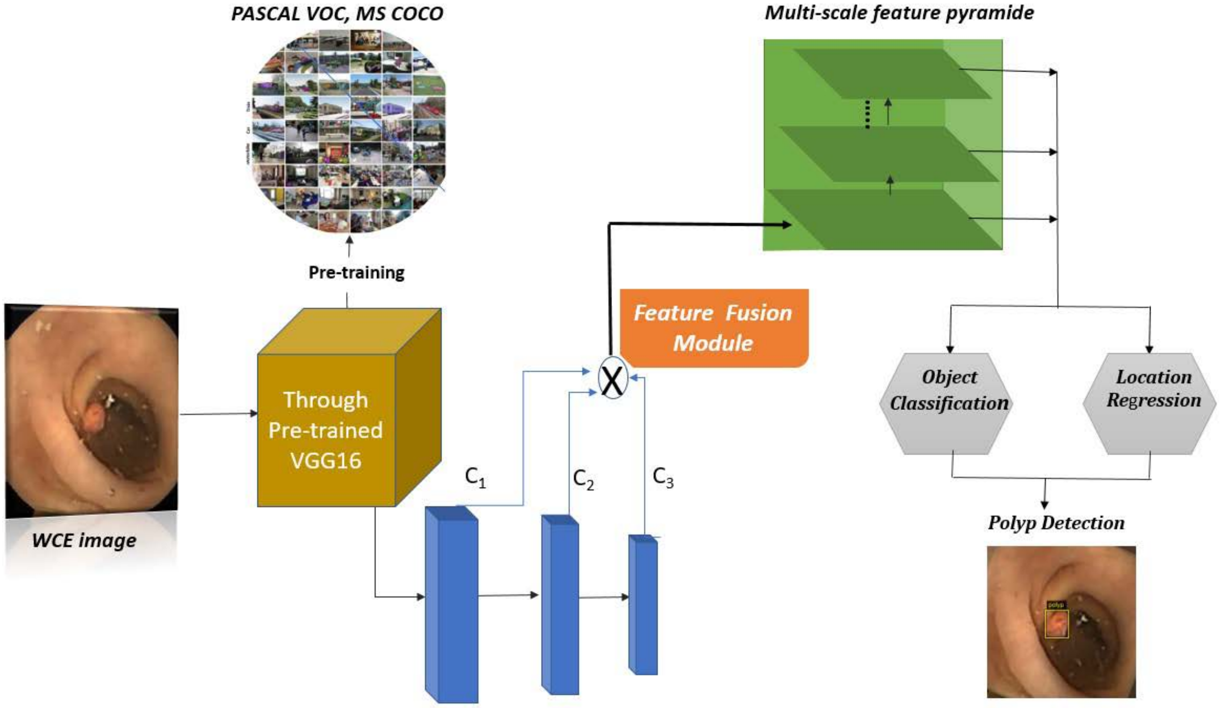


**FIGURE 1.** Flowchart of the proposed MP-FSSD method for small polyp detection in WCE images.

layers leads to excessive computational complexity. In general, DSSD succeeded in raising the accuracy at the expense of speed. Within the same scope, Jeong *et al.* [50] proposed an enhanced version of the SSD model in which they fused feature maps with simple concatenation and deconvolution, making full use of the direction information of the feature maps. These improvements make the Rainbow SSD suitable for small-target detection. Sheping *et al.* [51] proposed an improved SSD object detection algorithm based on a dense convolutional network (DenseNet), feature fusion, and residual prediction module. The original backbone (VGG-16) of the SSD network was replaced with DenseNet-S-32-1 to enhance the feature-extraction ability of the model. Then, a fusion mechanism of multiscale feature layers is performed to enhance the relationships between the levels in the feature pyramid. Finally, a residual block is established before object prediction to further improve the model performance. Remarkably, combining low-level visual and high-level semantic features in the SSD network structure to fully utilize the synthetic information leads to an improvement in performance, which is why it is highly desirable for small polyp detection in WCE images.

## III. MATERIALS AND METHODS

Polyp abnormalities possess different sharp edges, perceptible patterns, and geometries, making them difficult for experts to detect. The architecture of the proposed MP-FSSD system for polyp detection using WCE images is described in detail in this section. We aim to improve the precision of the SSD model and develop an optimization method for the detection results without sacrificing speed. This proposal fully utilizes the relationship between the layers in the feature pyramid without changing the base network. The main structure of the proposed approach is shown in Fig. 1. A data preparation stage is conducted as a pre-processing step, in which the region of interest (ROI) patches are extracted to remove the surrounding black regions in WCE images, providing no useful information. The training phase includes the following two stages: (1) Applying data augmentation process using commonly known geometric methods to solve the data insufficiency problem and to handle over-fitting in deep learning models; (2) The lightweight feature fusion module based single shot detector is pre-trained on the PASCAL VOC [52] and COCO datasets [53] and fine-tuned on the WCE/colonoscopy polyp datasets. The input WCE/colonoscopy image is then fed into a multi-scale pyramidal fusion single-shot multibox detector network (MP-FSSD), which consists of a one-stage feature extraction and classification sub-network for small polyp detection.

### A. SSD ALGORITHM

The Single shot multibox detector (SSD) is based on a forward propagation CNN network (VGG16) and truncated with other convolutional layers at the end [23]. The SSD investigates the pyramidal feature hierarchy in multiple layers within a ConvNet to generate a series of fixed-size bounding boxes and scores. Subsequently, it performs non-maximum suppression to obtain the final predictions. As depicted in

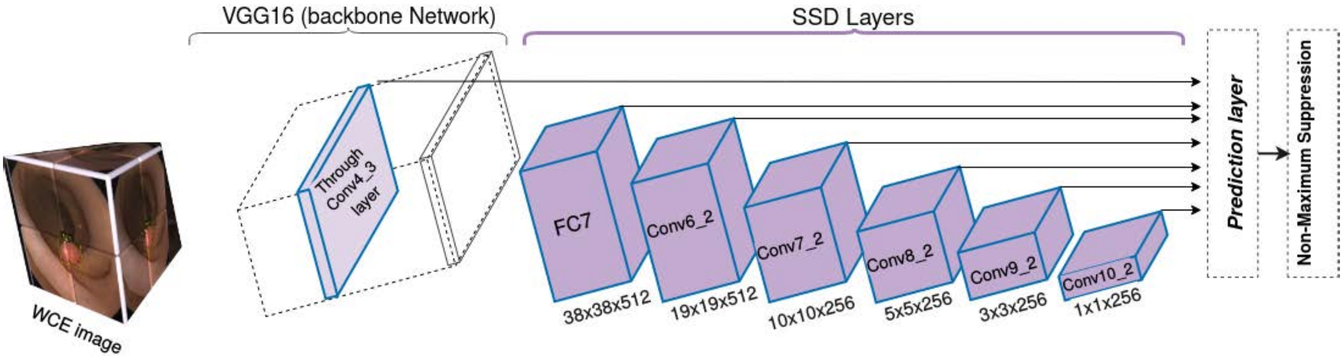


**FIGURE 2.** Framework of the traditional SSD.

Fig.2, the entire SSD network structure is divided into backbone and pyramid networks. The first part is represented by four layers of the VGG-16, and the second part is a simple convNet that is applied to the Conv4_3 layer of VGG to generate five additional layers. In fact, shallower layers are used to predict smaller objects, whereas deeper layers are used to predict larger objects at different scales. This assumption can be effective in accelerating the training process and reducing the prediction burden of the entire model. However, semantic information is regularly missed in the shallower layers despite being more important for small object detection purposes. Therefore, exploiting the semantic information lacking in shallow layers will improve the detection performance of small objects.

### B. MULTI-SCALE FEATURE MAPS PREDICTION

Recently, many detectors based on ConvNets have been presented to deal with the trade-off between accuracy and speed in object detection tasks [13], [51], [54]–[56]. However, scale variations remain a fundamental challenge in the field of computer vision. As depicted in Fig.3, several approaches have been proposed to solve the multiscale object detection problem. In Fig. 3(a), a single-scale feature map is used to create anchors of different scales to detect multiscale objects. This method has been adopted by some two-stage detectors, such as Faster R-CNN [34] and R-FCN [17]. However, the fixed receptive field size has difficulty detecting objects that are too large or small as well as multi-scale objects. Fig. 3(b) applies a ConvNet to the input multiscale images to generate different scale feature maps; however, such a design is incomplete in the sense of multiscale object detection. In Fig. 3(c), bottom-up and top-down architectures have been proven to work well in FPN [21] and DSSD [22], but fusing features layer by layer is not suitable for a fast detection process. Fig. 3(d) presents a popular architecture adopted by the original SSD [23], in which the feature pyramid from bottom to top is used to make predictions. In Fig. 3(e), the features of different layers and scales are concatenated from bottom to top, and the resulting feature map is used to generate a series of pyramid features later. In the same context, we combine the advantages of the feature fusion module presented in the FSSD method [18] by adopting the structure shown in Fig. 3(e) to tackle the problem of scale variations for detecting small polyp regions more effectively. Detailed information on the proposed MP-FSSD architecture is presented in the following sections.

### C. EDGE MAP GENERATION

Holistically nested edge detection (HED) [57], [58] has recently been used by state-of-the-art owing to its excellent performance and computational efficiency. To extract the polyp edges in WCE/colonoscopy images, we adopted the HED method, which performs image-to-image prediction using a deep learning model that deeply supervises nets. The main process of the edge extraction network is illustrated in Fig.4. As the backbone, HED is based on the VGG network which consists of 16 neural network layers. The HED architecture comprises a single-stream deep network with multiple side outputs, in which the side responses are generated for the individual layers. The HED network architecture has five stages, with strides 1, 2, 4, 8 and 16, respectively, and different receptive field sizes resulted in five layers that were selected as the side-outputs and fused by an average pooling layer. De-convolutional layers are adopted to perform the average fusion operation by resizing all side outputs to the same size. It was then fed into the softmax layer to produce the final label map. The side-output capability in producing multilevel edge maps from outlines to details makes edges extraction in a complex background highly feasible. To prove that the HED network architecture efficiently generates perceptually multilevel features and captures the inherent scales of the edge maps. The WCE/colonoscopy images were fed into the pre-trained HED model and a feed-forward operation was performed to produce the side-outputs for the polyp classification network. The training data are denoted as S $= \{(X_n,Y_n),\mathrm{n} = 1,\ldots,\mathrm{N}\}$, where $X_n$ is the raw input image and $Y_n \in \{0, 1\}$ is the corresponding ground truth binary edge map for image $X_n$. Suppose there are K side outputs in the network, where each side output layer is associated with a classifier, in which the corresponding weights are denoted

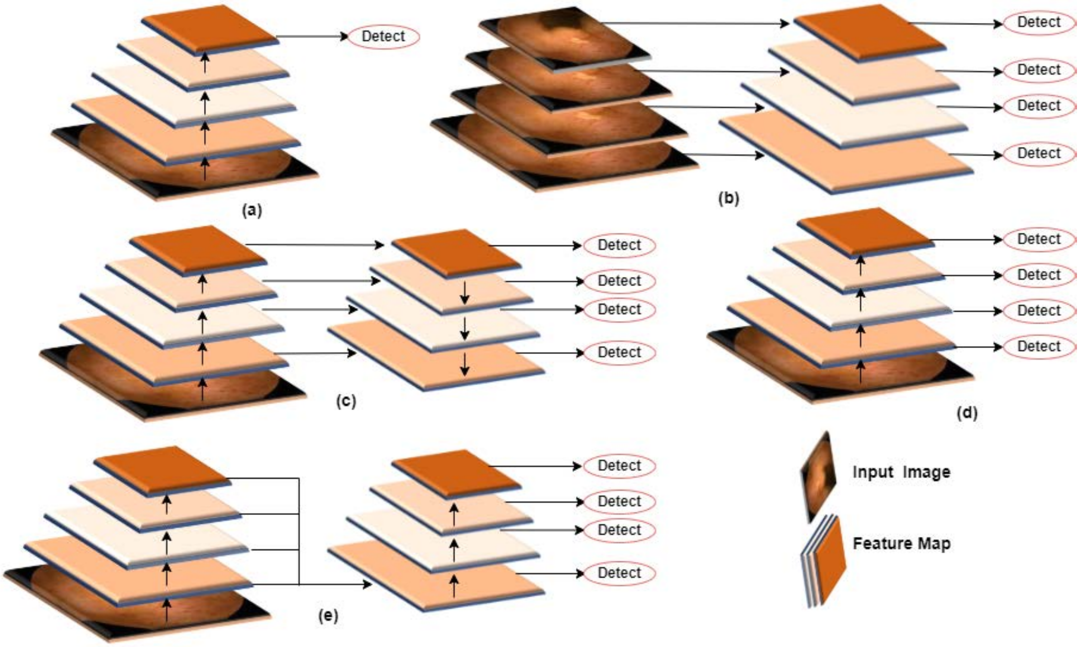


**FIGURE 3.** (a) The topmost feature map is utilized to make predictions, which is adopted by Faster R-CNN and R-FCN (two-stage detectors). (b) Feature maps are generated from the input image pyramids and used independently to make predictions, which is computationally expensive. (c) Features are fused layer by layer from top to bottom, it is adopted by FPN. (d) Use the feature pyramid generated from a ConvNet. (e) Features from different layers and scales are concatenated together first and used to generate pyramidal features similar to FSSD.

as w = $w^{(1)}, \ldots w^{(D)}$, where all parameters of the standard network layer are donated as W. The objective function of the side outputs is defined as:

$$L_{side}(W, w) = \sum_{d=1}^{K} \alpha_d l_{side}^{(d)}(W, w^{(d)}) \tag{1}$$

where the $d$th loss function for the side-output $l_{side}^{(d)}$ is a class-balanced cross-entropy loss function:

$$\begin{aligned} l_{side}^{(d)}(W, w_{(d)}) = -\beta \sum^{Y^+} logPr(y_i = 1|X; W, w^{(d)}) \\ -(1-\beta) \sum^{Y^-} logPr(y_j = 0|X; W, w^{(d)}) \end{aligned} \tag{2}$$

where $\beta$ = |Y-|/|Y|and 1 - $\beta$ = |Y + |/|Y|.|Y − | and |Y+| denote the edge and non-edge ground-truth label sets, respectively. Pr($y_j$ = 1|X;W,$w^{(d)}$) = $\sigma a_j^{(d)}$ ∈ [0, 1] was computed using a sigmoid function σ(.). Edge map from each side output layer is obtained by $\hat{Y}_{side}^{(d)} = \sigma(\hat{A}_{side}^{(d)})$, where $\hat{A}_{side}^{(d)} = \{a_j^{(d)}, j =, \ldots |Y|\}$ are the activations of the side output of layer d. To directly utilize the side-output predictions, a "weighted-fusion" layer is added to the network and simultaneously learns the fusion weight during training, which is minimized by a standard (backpropagation) stochastic gradient descent. In the testing phase, the edge map predictions were produced by the side-output layers and weighted-fusion layer:

$$(\hat{Y}_{fusion}, \hat{Y}_{side}^{(1)}, \hat{Y}_{side}^{(D)}) = HED(X, (W, w, h)) \tag{3}$$

where X donates the testing image and HED refers to the HED model.

As depicted in Fig.4. The polyp abnormality edge was marked in the center-right region of the input WCE image. The detailed edges in the side outputs are not easily discerned at a high level (e.g., the 3rd and 4th layers), and the polyp edge is merged with the background. Otherwise, the side outputs at low levels (e.g., the 1st and 2nd layers) include more details, in which the edges of polyp regions can be well detected. In this proposal, the second side-output layer is selected to emphasize the feature maps of the MP-FSSD network for polyp abnormality detection from the WCE images. Most information on WCE polyp regions incorporated in the shallow part of the network will be lost after deeply passing through the network layers. Therefore, the critical challenge is how to integrate polyp regions into a deep network. Meanwhile, the low-level side outputs contain rich detailed edges that are suitable for low-level fusion, whereas the high-level side outputs are less. To integrate the polyp regions, an edge pooling layer was embedded in the shallow part of the MP-FSSD network. Bilinear interpolation is applied to resize the polyp region map to perform the edge pooling operation so that the feature maps of the detection network and polyp region maps have the same spatial dimensions for further processing. On top of the network, after the

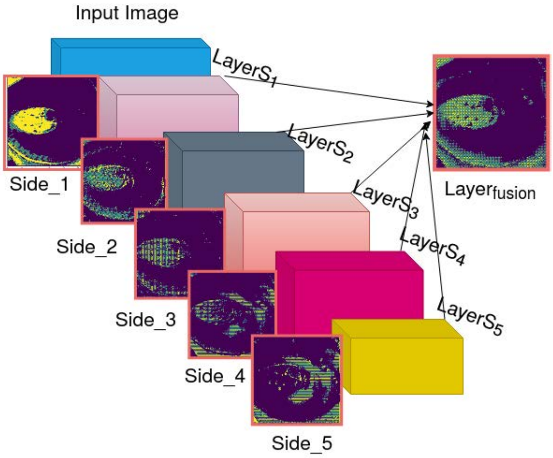


**FIGURE 4.** The edge exaction network is based on holistically-nested edge detection (HED), which consists of five side outputs and a fused output.

feature fusion model, newly added layers are used to detect polyps.

### D. LOW-LEVEL EDGE POOLING

Edge pooling was investigated [57] to integrate polyp region information into the shallow part of the MP-FSSD network. The same VGG16 (backbone network) was used for both edge extraction and SSD networks to accelerate the training process and reduce complexity. The polyp region maps produced by the $2^{nd}$ side-output induced from the edge extraction network and feature maps of the Conv4_3 layer from the VGG16 network are two inputs of the edge pooling layer. Owing to the sparse characteristics, the $2^{nd}$ side-output edge maps are resized to further match the *conv*4_3 size using bilinear interpolation. To produce a single output, the edge-pooling layer takes small rectangular blocks from the two emphasis layers and subsamples them. A weighting scheme was adopted to leverage the feature maps corresponding to polyp regions. The max pooling process takes the maximum values of the blocks to obtain the strongest response and passes them to the next layer. Consequently, the top invariant features are selected, leading to a faster convergence rate and improved performance. Therefore, the feature maps corresponding to the polyp regions are enhanced. Output of edge pooling $Ep_{i,j,m}$ for position(i,j) of the $k^{th}$ channel is defined as follows, from which max pooling takes the maximum value,

$$Ep_{i,j,m} = \max_{Y}\{I_m * ((1+\beta) * S_k)\} \tag{4}$$

where $I_m$ is the feature map, $S_k$ is the $k^{th}$ channel of the polyp region map, Y is the sliding window, and $*$ is the product of the corresponding positions of the two matrices. A parameter $\beta$ was used to control the weight of the polyp region map, which was set to 0.3 as in [57].

### E. FEATURE MAPS FUSION MODULE

ConvNets prove their capabilities in extracting semantic information of pyramidal feature hierarchies from low to high levels. A conventional SSD extracts features from different layers and considers them at the same level. Subsequently, it builds detectors directly on them. For small objects, SSD mainly uses the features of shallow layers to make predictions that lack semantic information. Consequently, the local detailed features and global semantic features are not well captured by the traditional SSD architecture design, and they perform poorly on small objects. Thus, slightly restructuring the feature maps is a promising strategy for improving the precision of a ConvNet object detector. As stated previously, the main purpose of the proposed approach is to improve the precision and speed of the SSD by fully utilizing the relationship between the layers in the feature pyramid without changing the basic backbone network.

As depicted in Fig. 5, the feature fusion SSD [18] network adds new layers, Conv6_2, Conv7_2, Conv8_2, and Conv9_2, for object classification and location regression. According to the analysis in [18], a feature map with a spatial size smaller than 10 px × 10 px has little information to merge. We constructed multiscale feature layers based on the conventional SSD. To ensure precision and detection speed, the MP-FSSD network uses its feature fusion module layers Conv1_Fu, Conv2_Fu, Conv3_Fu, Conv4_Fu, Conv5_Fu, and Conv6_Fu for small polyp detection. The target sizes of the resulting feature layers are 38 × 38, 19 × 19, 10 × 10, 5 × 5, 3 × 3, and 1 × 1, which are the same as those of the original SSD. The two parts of the MP-FSSD are described in detail below. Therefore, to make full use of semantic information and textural features, a top-down MP-FSSD is designed to introduce semantic information into shallow layers. As shown in Fig. 6, the entire process is as follows: First, to perform the edge pooling layer, bi-linear interpolation down-sampling is applied to the $2^{nd}$ side-output edge map of the HED to match the Conv4_3 size. The feature fusion block consists of the regenerated edge pooling layer, 1 × 1 convolution to compress the feature map channels of the FC7 and Conv6_2 layers, and bi-linear interpolation up-sampling to resize the feature maps of the FC7 and Conv6_2 layers to the same size as conv4_3. Then, simple concatenation is used to integrate deep features with shallow features. After the above treatments, the channel of the fused feature remains unchanged, but a single channel contains richer semantic information. Subsequent experiments demonstrated that these steps enrich the semantic information of shallow features and improve model performance in small polyp regions. To trade off the precision and speed, MP-FSSD does not use de-convolution for small objects in the deeper layers to reduce the decrease in speed and uses the same VGG16 network as the backbone for both the HED and FSSD networks. It should be noted that before concatenating the feature maps, a normalization is inevitable. This is because the feature values in the layers are significantly different in

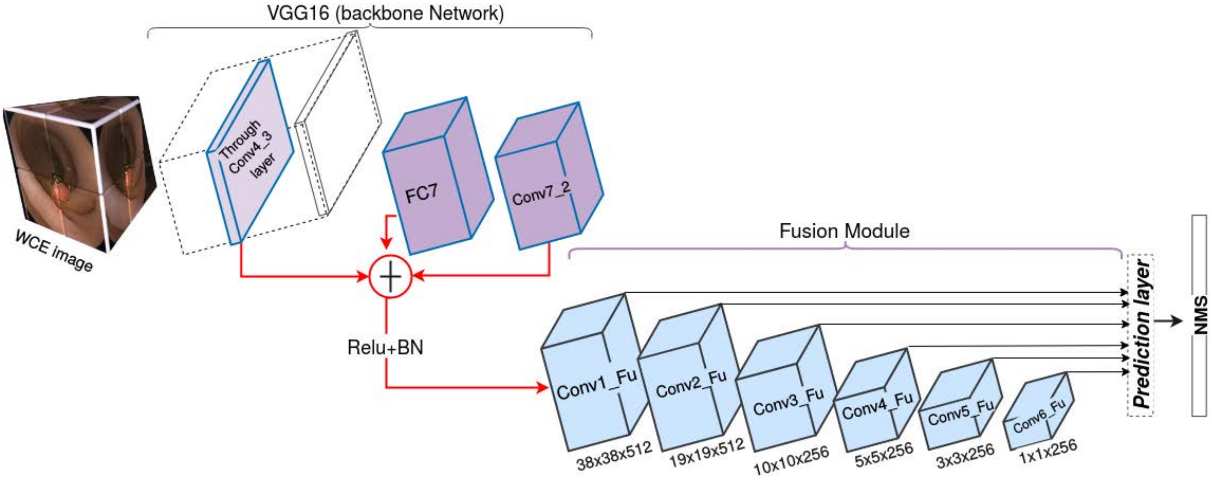


**FIGURE 5.** A framework of the FSSD network.

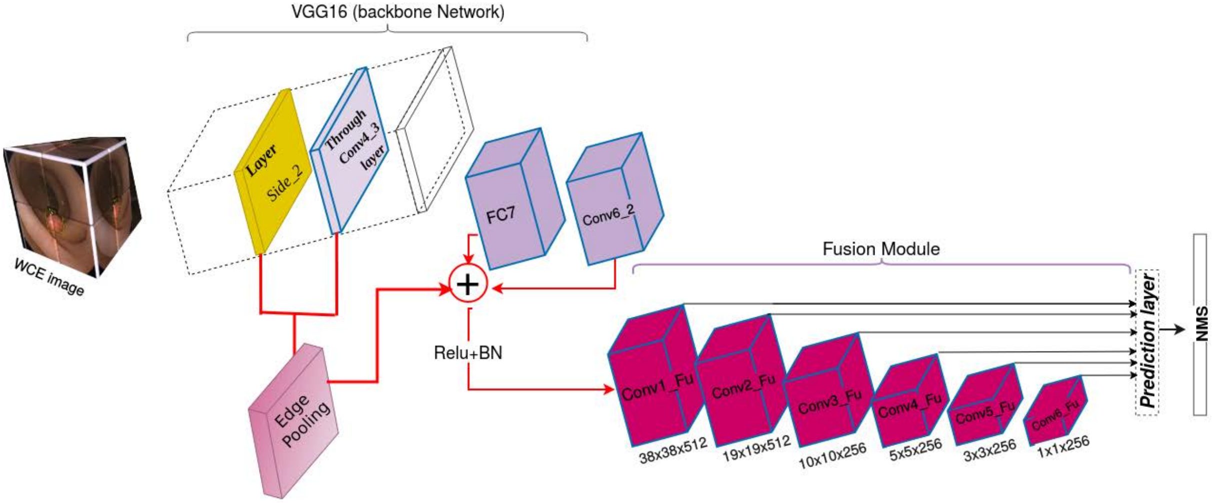


**FIGURE 6.** MP-FSSD network with the feature fusion module and an edge pooling layer.

scale. A batch normalization operation was applied to each filter before concatenation. As depicted in Fig. 6, the network structure of the MP-FSSD can be divided into two parts: the backbone network VGG-16 for polyp edge map generation and feature extraction, and the feature fusion module for polyp region detection. The structure of the model is shown in Fig. 7. The input size of the WCE/colonoscopy images was 300 × 300. The proposed MP-FSSD network selects Conv4_3 of VGG-16, and the output edge maps side_2 of the HED to generate new edge pooling, which is properly fused with FC7 and Conv6_2 of VGG-16.

## IV. EXPERIMENTS

### A. DATASETS

The first dataset is a WCE dataset from PillCam©COLON 2 polyps, which consists of 120 pedunculated polyps and 181 normal WCE frames within a single patient VCE session, as shown in Fig. 8. The original images have a resolution of 256 × 256 pixels. By pre-processing, we increased the size of the training dataset while avoiding overfitting, as described in this section. As a result, the new dataset included 1250 polyp patches and 1864 normal patches, respectively. As mentioned in this section, we pre-processed the training dataset to increase its size and avoid overfitting. As a result, there were 1250 polyp patches and 1864 normal patches in the revised dataset. To provide ground truths, the bounding boxes of the polyp patches were manually labeled and annotated as positive and negative samples. After that, they were reviewed and corrected by a trained expert. The second database is CVC-ClinicDB [60]. It consists of images containing various types of polyps that were extracted from colonoscopy videos. The dataset was selected from the 25 colonoscopy videos. The researchers selected 29 sequences that contained at least one polyp in every frame from the 25 videos. Finally, a set of frames is selected for each sequence. Moreover, it contains the ground truth of these polyps, which consists of masks corresponding to the region covered by the polyp in the frame. The CVC-ClinicDB dataset comprises 612 polyp images of size 576 × 768. In addition to the frames, a ground truth was created by experts by manually defining a mask on the

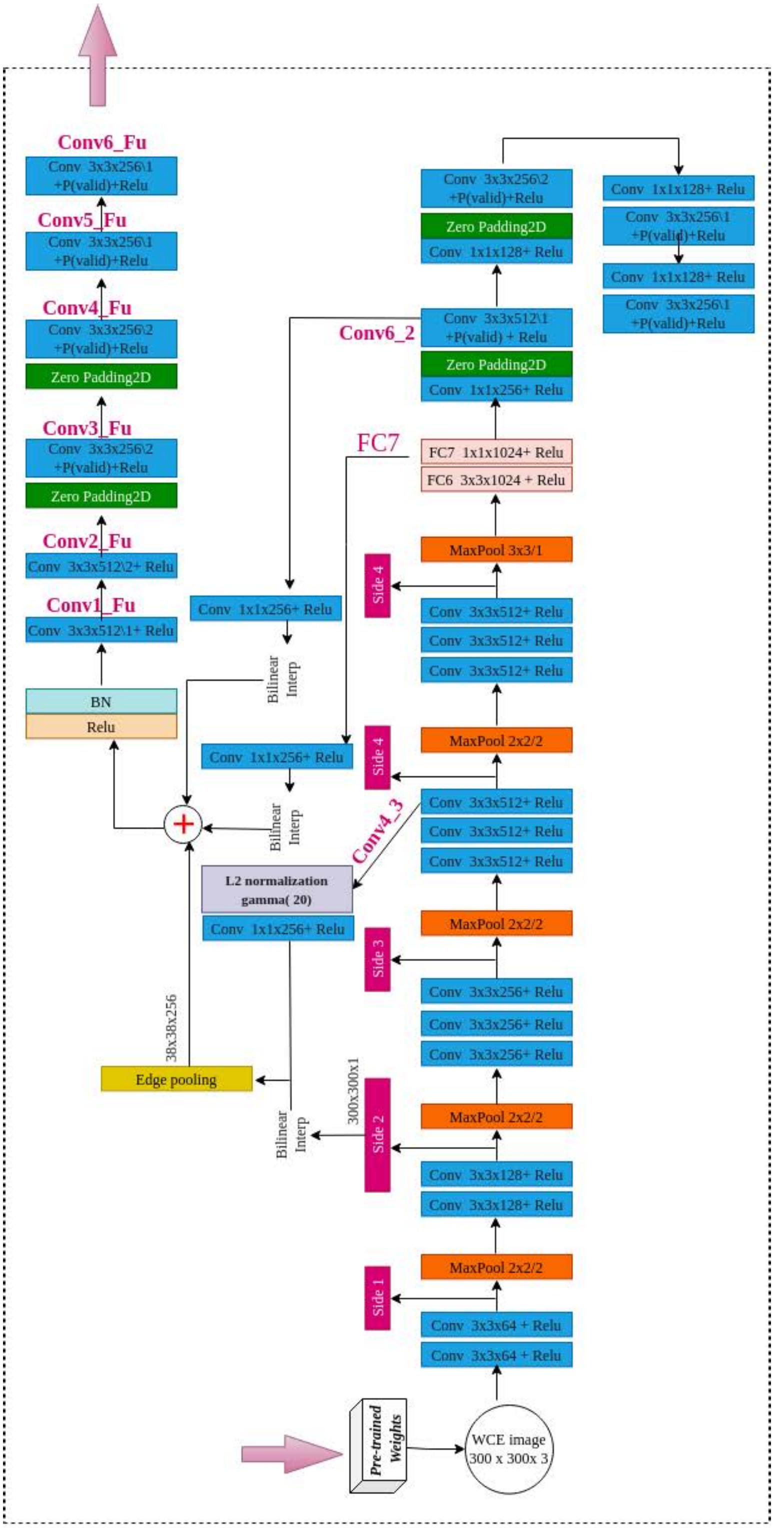


**FIGURE 7.** Detailed structure of the MP-FSSD network.

region covered by the polyp. To assess the impact of image pre-processing on polyp detection results, the ground truth bounding box of the colonoscopy dataset was labeled based on the ground truth for specular highlights provided by the experts.

In this study, we used the annotated ETIS-Larib [61] dataset for colonoscopy polyp detection. It comprises 196 polyp images of various sizes and appearances, generated from 34 colonoscopy videos. At least one polyp was present in all the 196 images. Ground truths of the polyp

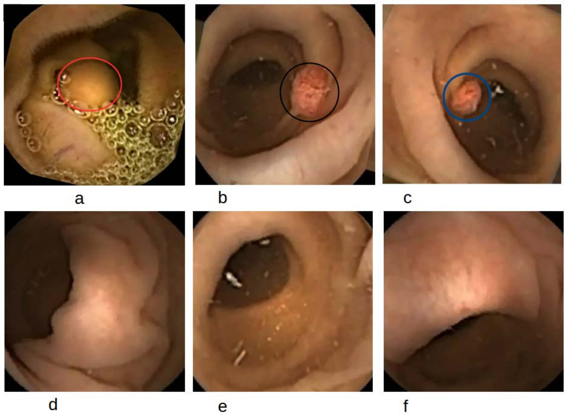


**FIGURE 8.** Example of WCE polyp images (a, b, c) and normal images (d, e, f).

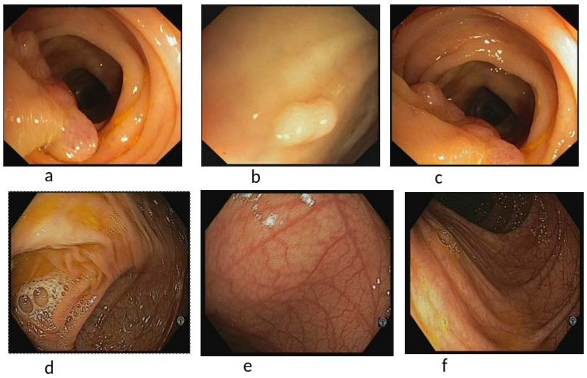


**FIGURE 9.** Example of colonoscopy polyp images (a, b, c) and normal images (d, e, f).

regions were annotated by skilled video endoscopists from the corresponding associated clinical institutions. The CVC-ClinicDB [62] and ETIS-Larib [63] colonoscopy datasets were used in the automatic polyp detection sub-challenge at MICCAI 2015. For a fair performance comparison with the challenge results, 196 images from the ETIS-LARIB dataset were used for testing purposes. Figures 8 and 9 show examples of images, including the endoscopy/colonoscopy datasets used in this study. The split data set proportions were 70% for training, 10% for validation, and 20% for model testing. The training and testing process was performed using 5 fold cross validation [64], in which each group's model obtained the corresponding correct rate. The resulting rates were then averaged to estimate the precision of the target detection algorithm. The validation set refers to checking the state and convergence of the model after each epoch was completed. It does not contribute to the gradient descent, but only adjusts the hyperparameters such as the number of iterations, and learning rate. The validation set determines which group of hyperparameters has a good performance and adopts them according to the five group performances in the models. It can also be used to monitor whether the model has been fitted to determine the time when the training stops. Finally, the generalization ability and performance of model detection and classification were determined during the testing process. To reach the input size required by the source-pre-trained SSD network, it is necessary to re-scale the normal and abnormal images to 300 × 300 pixels.

## B. EVALUATION METRICS

The mean average precision (mAP) is the most commonly employed metric for evaluating target detection accuracy. The mAP is adopted as the criterion of detection precision and it is defined as the average of the average precision (AP) of all object categories, which is an indicator related to the IoU threshold. In our experiments, we used the most commonly used threshold IoU = 0.5. It is formulated by Eq. 5.

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{5}$$

where Q is the number of queries in the set, and q is the query for average precision.

Precision and recall can be defined as:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \tag{6}$$

where TP denotes true positives, that is, IoU $>$ 0.5, FP indicates false positives, and FN indicates false negatives.

Indicatively, given a set of ground truth bounding boxes annotated by the experts in all frames, and a set of predicted bounding boxes produced by the network with a confidence score above a specific threshold. The true positive, false positive and false negative bounding boxes are defined as follows:

**A true positive (TP)** is a bounding box of a polyp region that is predicted by the network with an IoU $>$ 0.5 with ground truth bounding boxes.

**A false positive (FP)** is a predicted bounding box that does not overlap with any ground truth bounding box or overlaps with ground truth bounding boxes with IoU $<$ 0.5.

**A false negative (FN)** is a ground-truth bounding box that does not overlap with any predicted bounding boxes, or overlaps with predicted bounding boxes with IoU $<$ 0.5.

**The IoU** is the ratio of the overlapping area divided by the area of the union of the ground truth and the predicted bounding boxes.

**FPS** refers to the detection speed and indicates the number of frames transmitted per second. With a high FPS value, more frames per second occur; therefore, the display effect is smoother and clearer.

## C. EXPERIMENTAL SETUP

### 1) EXPERIMENTAL ENVIRONMENT

We performed the experiments using the Colab Pro Plus solution provided by Google, which has a maximum RAM of 52Gb and a disk of 166,83Gb. All experiments were conducted using TensorFlow 1.15, Tensorflow-GPU 1.15, Cuda 8.0.61-1, CuDNN6.0, Keras 2.0.5, Python 3.5, h5py 2.10.0, NumPy 1.16.3, and OpenCV 3.1. Based on the small polyp region size, the same aspect ratio falling within a range of 1-2 according to the ground truth bounding boxes, was maintained for both datasets. We applied NMS with a confidence threshold of 0.05, Jaccard overlap of 0.4 per class and keep the top 200 detections per image.

#### 2) NETWORK TRAINING

In the context of this study, the model ability is represented by providing the location of the polyp within a given image. We employed the commonly known evaluation metrics in the field of medical imaging and investigated them with state-of-the-art to further compare MP-FSSD polyp detector performances to other models based on SSD networks for polyp detection targets. The outputs of the proposed model are the four rectangular coordinates (x, y, w, h) of the detected bounding box. More useful information from the WCE polyp image is preserved in the square center region. Therefore, it is necessary to remove the surrounding black regions that contain no useful information, as this may degrade the performance of WCE polyp detection and increase the computation time. Training involves choosing a set of data augmentation strategies [65]. However, this proposal chooses to investigate some popular augmentation methods used in the recent literature [66], [67]. In particular, we apply geometric methods that alter the geometry of the resulting RoI image by mapping individual pixel values to new destinations. Given their success in related works, we investigated flipping, rotation by 270°, and cropping schemes. All the images were resized to $300 \times 300 \times 3$. Training a model on a large-scale object detection dataset such as COCO, and using it directly to detect small objects, results in a domain-shift problem. Therefore, we pretrained model on the PASCAL VOC 2007 and 2012 datasets. Then, we fine-tuned the model on WCE or colonoscopy datasets, or both, separately using the Adam optimizer with an initial learning rate of 0.01, beta_1 = 0.9, beta_2 = 0.999, weight decay 0.00, and epsilon 1e-08. The learning rate schedule using Keras is 0.001 if epoch $< 10$, 0.0001 if epoch $< 50$, and 0.00001 otherwise. The learning rate decay policy is slightly different from the original SSD with a drop of 0.5 and epochs_drop of 10. The batch size used was 32, which is beyond the GPU's memory capacity. The MP-FSSD model was used for the training process, with a total of 100 and 500 steps per epoch. The detection performance was mainly evaluated using the mean average precision (mAP). Other indicators, such as frame per second (FPS) and parameters, will also help us further evaluate the model performance. The MP-FSSD training objective was to minimize the weighted sum of the smooth L1 loss [68], [69]. We used the same loss function as that used in the conventional SSD [23]. The hyper-parameter $\alpha$ was set to 1 by cross-validation and neg_pos_ratio to 3. More details regarding SSD_Loss can be found in [51].

### *D. RESULTS AND DISCUSSION*

#### 1) ABLATION STUDY ON WCE/COLONOSCOPY DATASETS

In this section, we describe an ablation study on the WCE or colonoscopy datasets, or both, and analyze some important design factors that affect the experimental results of the main structure of the MP-FSSD detector. The experimental results were divided into two parts, as listed in Table 1. In the first, training and testing were performed on the WCE images. In the latter, the models were trained on the WCE and colonoscopy joint training set and tested on Clinic-DB test set. An edge pooling layer is embedded into the shallow part of the detection network to integrate the polyp edge regions into the MP-FSSD network. A critical issue is which side_output maps of holistically nested hed networks can be passed as deep as possible to the top MP-FSSD network without deteriorating detection performance. From Table 1, the two-level side_output maps surpassed the other side_outputs {1-3-4-5} and side_outputs-Fu in terms of (mAP) metrics with a gains of 3.62%, 0.73%, 1.38%, 1.9%, and 0.86%, respectively, on the WCE dataset and a gain 3.24%, 1.19%, 4.27%, 6.27%, and 1.56%, respectively, for the joint training set WCE/colonoscopy datasets. The experimental results prove that the side_outputs2 feature maps contain rich detailed edges that are suitable for low-level fusion within the edge-pooling layer. To reflect the effect of the series of actions that we added to the conventional FSSD, the models were run on both the WCE and colonoscopy datasets with different settings, and their evaluations are recorded in Table 2. We maintained the VGG16 backbone structure for all the emphasis MP-FSSD model settings listed in Table 2.

##### *a: FUSION BLOCK*

This consists of applying concatenation or element-wise summation. Using a simple concatenation to fuse features, the model can obtain 88.42% mAP (row 1), whereas element-wise summation can only achieve 86.73% (row 2). Concatenation is better than element-wise summation with a margin of 1.69 points. Therefore, we chose concatenation as the feature fusion method of the MP-FSSD.

##### *b: BN*

Normalizing the feature map values of different layers to perform feature fusion blocks is another critical issue in many recent approaches. L2 normalization was used in the traditional SSD model to scale the feature map from conv4_3. To use a simple and efficient way to scale the feature maps, we add a batch normalization layer after the concatenation process. The results in Table 2 (rows 4 and row 5) show that using the batch normalization layer to re-scale the feature maps can bring us about 0.79% mAP improvement considering the WCE dataset, and an improvement of 0.67% mAP (row 12 and row 13) from the joint training set of WCE and CVC-ClinicDB datasets.

##### *c: FUSION LAYERS*

To perform the pyramid feature fusion block, the range of layers to be fused is listed in (column 9) of Table 2. We conducted different feature fusion layers and compared their direct impact on MP-FSSD performance in terms of (mAP) metrics. While we fused the feature maps (conv3_4, FC7,

**TABLE 1.** Results of the ablation study of the MP-FSSD using one side output of the holistically nested hed model. BN means that a batch normalization layer is added after the feature concatenation. The mAP is measured on the WCE and CVC-ClinicDB test sets.

| Dataset | Methods | Edge Pooling | Feature Fusion | mAP |
|---|---|---|---|---|
| WCE | MP-FSSD | Side_Ouput1 | Concat+BN | 89.78% |
| | MP-FSSD | Side_Ouput2 | Concat+BN | 93.4% |
| | MP-FSSD | Side_Ouput3 | Concat+BN | 92.67% |
| | MP-FSSD | Side_Ouput4 | Concat+BN | 92.02% |
| | MP-FSSD | Side_Ouput5 | Concat+BN | 91.15% |
| | MP-FSSD | Side_Ouput-Fu | Concat+BN | 92.54% |
| WCE + CVC-ClinicDB | MP-FSSD | Side_Ouput1 | Concat+BN | 88.32% |
| | MP-FSSD | Side_Ouput2 | Concat+BN | 91.56% |
| | MP-FSSD | Side_Ouput3 | Concat+BN | 90.37% |
| | MP-FSSD | Side_Ouput4 | Concat+BN | 87.29% |
| | MP-FSSD | Side_Ouput5 | Concat+BN | 85.86% |
| | MP-FSSD | Side_Ouput-Fu | Concat+BN | 90% |

and conv6_2), the mAP on the WCE test set (row 1) was 88.42% and 85.13% for the colonoscopy test dataset (row 9). It is interesting if we replace the conv4_3 with an edge pooling layer, the mAP is increased to 93.4% on the WCE test set (row 8) and 91.56% on the colonoscopy test dataset (row 16), which means that the embedded edge pooling into the feature fusion block has more benefit to the final system performance.

#### d: PRE-TRAINED VGG OR SSD

For the training procedure, pre-training a model on a large-scale object detection dataset, such as COCO, may help detect small objects directly. However, the domain-shift problem is inevitable. The results in Table 2 indicate that after removing the COCO datasets for pretraining purposes, the module performance was further improved. We used the VGG16 trained on the PASCAL VOC 2007, PASCAL VOC 2012, and COCO datasets as a pre-trained model. Then, we fine-tuned the model on the WCE and the joint training set of WCE/colonoscopy datasets before reconstructing the MP-FSSD detector. As depicted in Table 2, the model performances increase from 91.16% mAP to 93.4% mAP and from 89.62% to 91.56%, which is an improvement of 2.24 points and 1.94 points on WCE or Colonoscopy datasets, or both if the original VGG network is pre-trained on PASCAL VOC 07+12 rather than taking the original SSD model trained on PASCAL VOC 07+12+COCO datasets as a pre-trained model and fine-tuning on polyp datasets.

#### e: OPTIMIZER

Adaptive algorithms such as Adam have good convergence speed, whereas algorithms such as SGD generalize better. As can be seen in Table 2, optimizing the training process of the model using the Adam optimizer improves the performance by 0.86% mAP (92.54% (row 7) vs. 93.4% (row 8)) on the WCE dataset, and by 1.42% mAP (90.14% (row 15) vs. 91.56% (row 16)) on the CVC-ClinicDB dataset compared with the traditional SGD algorithm. This proves its efficiency in reducing the loss of information when deeply passing through several layers and accelerating the convergence of the model. The traditional SGD algorithm was no longer used and the relevant training process of the model was optimized.

### 2) SSD RESULTS ON WCE AND CVC-ClinicDB DATASETS

The results of some state-of-the-art detector-based SSD models and the proposed MP-FSSD model on both the WCE and CVC-ClinicDB test sets are shown in Table 3. To reflect the effects of different actions added to conventional SSD. The SSD models were run as indicated in their original studies, and their evaluations are presented in Table 3. The (mAP) of the conventional SSD with VGG16 was 77.2% and 75% for the WCE and WCE + colonoscopy datasets, respectively. After changing the backbone structure, which was replaced with ResNet-101, the (mAP) is improved to 81.65% and 79.63% on both training sets, respectively. The effectiveness of the feature fusion method is also shown in the results in Table 3. By adding feature fusion modules using concatenation with VGGNet as the backbone network, FSSD300 and FSSD500 were increased in terms of (mAP) by 11.58% and 11.5% compared to SSD300 and SSD500 models on WCE test set, respectively, because fused feature layers contain rich details and semantic information. By replacing the backbone network VGGNet with DenseNet-S-32-1, DF-SSD300 model improves the detection performance by 2.53% and 2.73% compared with FSS300 (91.24% vs. 89.78%) and (89.11% vs. 86.38%) on WCE or colonoscopy test sets, or both, respectively. It also exceeded FSSD500 by 2.53% and 2.11% mAP (89.11% vs. 88.71%) and (89.11% vs. 87%) on both test sets, respectively the detection speed decreases by half. The lightweight feature pyramid L_SSD replaces the original VGG16 with ResNet-101 to perform feature fusion, which contains rich detail and semantic information. The mAP of ResNet-101 was further improved to 89.98% mAP and 86.63% mAP on WCE or colonoscopy test sets, or both, slightly better than FSSD300 (89.78% mAP and

**TABLE 2.** Results of the ablation study on WCE/colonoscopy datasets. BN means that a batch normalization layer is added after the feature concatenation. pre-trained VGG means that a pre-trained VGG16 is adopted to initialize the model. Pre-trained SSD means that the FSSD is optimized from a well-trained SSD model. Edge pooling represents the fusion layer of the side_output2 of hed and the conv4_3 of VGG16. The options of fusion blocks represent which ones we choose to merge, it includes edge_pooling/conv4_3, fc7, and conv6_2. The mAP is measured on the WCE or CVC-ClinicDB datasets, or both.

| Dataset | Method | Backbone | Pre-trained VGG | Pre-trained SSD | Edge Pooling | Fusion Block | BN | Fusion layers | Optimizer | mAP |
|---|---|---|---|---|---|---|---|---|---|---|
| WCE | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | × | Concat | ✓ | Conv4 + FC7 +Conv6 | Adam | 88.42% |
| | MP-FSSD | VGG16 | Pascal Voc-07 | × | × | Elem-sum | ✓ | Conv4 + FC7 +Conv6 | Adam | 86.73% |
| | MP-FSSD | VGG16 | × | ✓ | × | Concat | × | Conv4 + FC7 +Conv6 | SGD | 78.12% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12+Coco | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 90.6% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12+Coco | × | ✓ | Concat | × | Edge Pool + FC7 +Conv6 | Adam | 89.81% |
| | MP-FSSD | VGG16 | × | ✓ | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 91.16% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | SGD | 92.54% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 93.4% |
| WCE+CVC-ClinicDB | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | × | Concat | ✓ | Conv4 + FC7 +Conv6 | Adam | 85.13% |
| | MP-FSSD | VGG16 | Pascal Voc-07 | × | × | Elem-sum | ✓ | Conv4 + FC7 +Conv6 | Adam | 83.27% |
| | MP-FSSD | VGG16 | × | ✓ | × | Concat | × | Conv4 + FC7 +Conv6 | SGD | 81.46% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12+Coco | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 88.35% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12+Coco | × | ✓ | Concat | × | Edge Pool + FC7 +Conv6 | Adam | 87.68% |
| | MP-FSSD | VGG16 | × | ✓ | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 89.62% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | SGD | 90.14% |
| | MP-FSSD | VGG16 | Pascal Voc-07+12 | × | ✓ | Concat | ✓ | Edge Pool + FC7 +Conv6 | Adam | 91.56% |

**TABLE 3.** SSD results on WCE and CVC-ClinicDB datasets. Pre-train means that a pre-trained backbone is adopted to initialize the model or it is initialized from scratch. For a fair comparison, the speed (FPS) of the SSDs and FSSDs, and (mAP) performances are tested using google Colab pro+ GPU.

| Dataset | Methods | Backbone | Pre-train | FPS | mAP |
|---|---|---|---|---|---|
| WCE | SSD300 | VGG16 | ✓ | 46 | 77.2% |
| | SSD300 | ResNet-101 | ✓ | 47.3 | 81.65% |
| | SSD500 | VGG16 | ✓ | 19 | 79.45% |
| | SSD500 | ResNet-101 | ✓ | 20 | 84.95% |
| | FSSD300 | VGG16 | ✓ | 65.9 | 89.78% |
| | FSSD500 | VGG16 | ✓ | 69.6 | 88.71% |
| | DF-SSD300 [51] | DenseNet-S-32-1 | ✓ | 11.6 | 91.24% |
| | L_SSD [54] | ResNet-101 | ✓ | 40 | 89.98% |
| | MP-FSSD(ours) | ResNet-101 | ✓ | 45.6 | 91.13% |
| | MP-FSSD(ours) | VGG16 | ✓ | 62.57 | **93.4%** |
| WCE+CVC-ClinicDB | SSD300 | VGG16 | ✓ | 46 | 75.6% |
| | SSD300 | ResNet-101 | ✓ | 47.3 | 79.63% |
| | SSD500 | VGG16 | ✓ | 19 | 78.5% |
| | SSD500 | ResNet-101 | ✓ | 20 | 81.02% |
| | FSSD300 | VGG16 | ✓ | 65.9 | 86.38% |
| | FSSD500 | VGG16 | ✓ | 69.6 | 87% |
| | DF-SSD300 [51] | DenseNet-S-32-1 | ✓ | 11.6 | 89.11% |
| | L_SSD [54] | ResNet-101 | ✓ | 40 | 86.63% |
| | MP-FSSD(ours) | ResNet-101 | ✓ | 45.6 | 89.72% |
| | MP-FSSD(ours) | VGG16 | ✓ | 62.57 | **91.56%** |

86.38% mAP) for both test sets, respectively. As shown in Table 3 (rows 9–10 and 19–20), when we combine our proposed improvements and adopt a better network framework, the proposed MP-FSSD algorithm exceeds FSSD, DF-SSD and L_SSD models with VGG16 by 3.62 points (93.4% vs. 89.78%), 2.16 points (93.4% vs. 91.24%) and 3.42 points (93.4% vs. 89.98%) on WCE dataset. The running time was also evaluated for both the WCE and CVC-ClinicDB test datasets, as shown in Table 3 (column 5). The detection speed of the MP-FSSD method based on VGG16 and ResNet-101 is 62.57 FPS and 45.6 FPS respectively, slower than the FSSD model owing to the pooling layer embedded into the feature fusion module. However, MP-FSSD using the VGG16 backbone still achieved real-time detection, compared with DF-SSD300 and L-SSD with 11.6 FPS and 40 FPS, respectively. By replacing the backbone network VGGNet with ResNet-101, MP-FSSD300 model improved the (mAP) performance by 5.18% compared with FSSD300 (91.13% vs. 89.78%). It can be seen from the obtained results that MP-FSSD has a strong feature reuse and extraction ability. In addition, they proved that the integration of an edge pooling layer into the fusion block of the MP-SSD is more effective than other SSDs methods.

### 3) COMPARISON WITH THE STATE-OF-THE-ART METHOD

In fact, WCE dataset acquisition still presents a challenge, owing to the lack of large and publicly available annotated datasets. To verify and evaluate the performance of

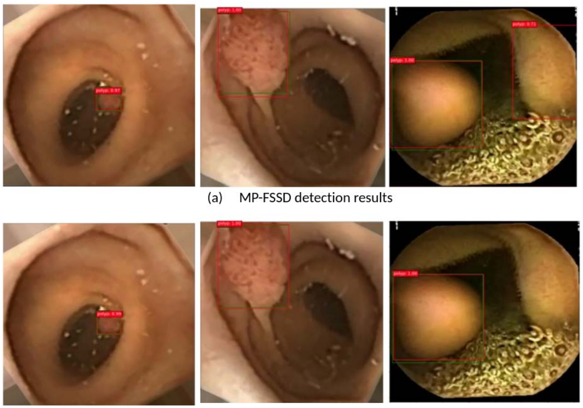
**TABLE 4.** WCE or colonoscopy test detection results, or both.

| Training Dataset | Methods | Testing Dataset | Backbone Network | Pre-train | Input Size | Sensitivity | Specificity | Prec |
|---|---|---|---|---|---|---|---|---|
| WCE images | MP-FSSD (ours) | WCE images | VGG16 | ✓ | 300 × 300 | × | × | 93.4%(mAP) |
| WCE images +CVC-ClinicDB | MP-FSSD (ours) | CVC-ClinicDB | VGG16 | ✓ | 300 × 300 | × | × | 91.56%(mAP) |
| WCE images +CVC-ClinicDB | MP-FSSD (ours) | Etis-Larib | VGG16 | ✓ | 300 × 300 | × | × | 90.02%(mAP) |
| CVC-ClinicDB + Kvasir | Dulf et al., 2021 [2] | Etis-Larib | Inceptionv3 | ✓ | 288 × 384 | 98.13% | 99.73% | × |
| CVC-ClinicDB + ETIS-LARIB | Shin et al., 2018 [4] | Etis-Larib | Inception Resnet | ✓ | 768 × 576 | × | × | 92.2% |
| SUN+ PICCOLO+ CVC-ClinicDB | Ishak et al., 2021 [35] | Etis-Larib | YOLOv3 | ✓ | 448 × 448 | × | × | 90.61% |
| CVC-ClinicDB | Liu et al., 2021 [70] | Etis-Larib | ResNet-101 | ✓ | 384 × 288/1225 × 966 | × | × | 77.80% |
| GIANA 2017 | Wang et al., 2019 [39] | Etis-Larib | VGG16 | ✓ | 1225 × 996 | × | × | 88.89% |
| Colonoscopy images | Ozawa et al., 2020 [71] | Etis-Larib | VGG16 | × | 300 × 300 | 92% | × | × |
| CVC-ClinicDB | Qadir et al., 2021 [72] | Etis-Larib | Resnet34 | ✓ | 512 × 512 | × | × | 86.54% |
| CVC-ClinicDB | Pacal and Karaboga, 2021 [73] | Etis-Larib | CSPDarkNet53 | ✓ | 384 × 288 | × | × | 91.62% |
| CVC-ClinicDB | Shen et al., 2021 [34] | Etis-Larib | VGG16 | × | 224 × 224 | × | × | 91.49% |
| Colonoscopy images | Nogueira-Rodríguez et al., 2021 [13] | Etis-Larib | YOLOv3 | ✓ | 416 × 416 | 72.61% | 83.04% | × |

MP-FSSD, the performances of the trained model on the WCE and colonoscopy datasets were evaluated on the ETIS LARIB dataset to quantitatively compare MP-FSSD with other state-of-the-art models. As proven previously, the proposed MP-FSSD model provided the best performance for the WCE and CVC-ClinicDB datasets. Therefore, the model was evaluated using the publicly available Etis-Larib dataset. This study aimed to examine the effect of increased training data on polyp detection and to evaluate its success. It is commonly known, the one-stage detection SSD algorithm is one of the most popular target detection algorithms with high accuracy and speed. For FSSD300 with VGG16 backbone, the mAP and FPS were 89.78% and 65.9 on the WCE test set, respectively. Compared with this, MP-FSSD has a 3.62% mAP gain. However, owing to the embedded edge pooling layer, the feature fusion block reduces the speed to 62.57 frames per second (Table 3 (row 10)). Even if the holistically nested edge detection network was applied on the same VGG16 backbone, it did not affect the complexity of the model structure. As reported in Table 4, it can be seen from the metric values that training and testing the MP-FSSD model on only the WCE dataset appear to exceed other models in terms of (mAP) metric. This may be explained by the differences in nature, texture, and illumination acquisition conditions for both WCE and colonoscopy images. In addition, the color of the polyp is not homogeneous across different polyp frames within a patient and is highly variable across different examinations from patients. The model trained on the WCE + CVC-ClinicDB dataset is slightly the same as some competitor models on the Etis-Larib dataset in terms of success and exceeds the others. In general, the MP-FSSD shows higher accuracy and faster speed because the structure of the feature fusion module is simpler and does not increase the detection time.

### 4) VISUALIZATION

Fig. 10 presents some detection examples of the FSSD (row 1) and their analogs (row 2) of the MP-FSSD model on WCE polyp datasets. Compared to FSSD, the MP-SSD model showed relative improvement in small polyp localization from the normal intestinal mucosa. (Lightweight version of the SSD algorithm) does not correctly predict the ground truth of small polyp regions, but MP-FSSD showed an obvious improvement. In addition, the small polyp regions depend more on their surroundings. Thus, in contrast to large polyps, the position information of small polyp regions is more likely to be lost during the detection process. We note that the FSSD model only detects smaller objects from shallow layers, such as conv4_3, whose receptive field is too small to observe the object's context information. The problem is that some polyp regions are similar in appearance to the surrounding normal mucosa, which may affect the performance of the identification of smaller objects. As a solution, MP-FSSD adds embedded edge pooling into the fusion block, which can capture scene contexts and differentiate polyp edge regions from the normal mucosa. From Fig. 10 (image 3 of row 2), we can observe a concrete case that benefits from the feature fusion module of the MP-FSSD detector.

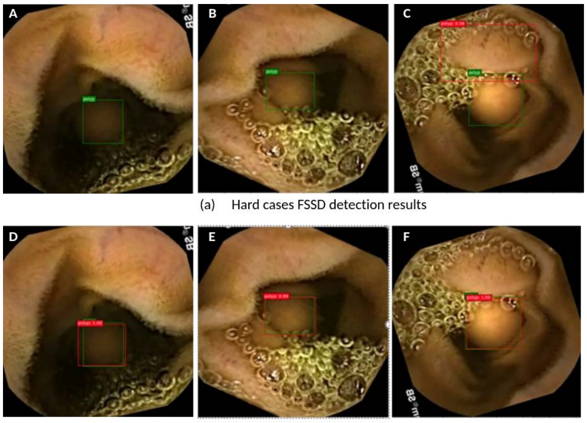


(a) MP-FSSD detection results

(b) MP-FSSD detection results

**FIGURE 10.** Comparison between the FSSD300 and the proposed MP-FSSD300 on the WCE dataset. True bounding boxes with IoU of 0.5 or higher with the bounding predicted boxes are drawn in green and red colors, respectively. (a) Results of the FSSD. (b) Results of the proposed method.

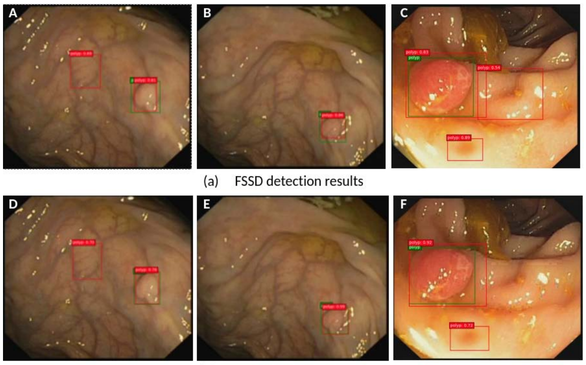
Several artifacts can reduce the diagnostic yield. In reality, the actual extent of the lumen visualized by capsule endoscopy is limited by air bubbles, food, and other debris, which hinders detection. Fig. 11 shows representative detection examples of polyp identifications of the FSSD and MP-FSSD models on WCE images. The upper row in Figs. 11 (A–B) corresponds to the failure identification of the FSSD detector. Fig. 11 (C) shows the false-positive identification in the normal mucosa areas. It can be difficult to accurately detect polyps owing to their complicated characteristics (color, texture, contrast, and size). This problem adds to the difficulty in identification. The difference in appearance between the polyp regions and normal mucosa was not obvious in the WCE images. In addition, the WCE frames captured in the case of insufficient light produce poor



(a) Hard cases FSSD detection results

(b) Hard cases MP-FSSD detection results

**FIGURE 11. Representative examples of polyp identifications on the WCE test dataset for FSSD and MP-FSSD models. The first row shows failure cases of the FSSD compared with MP-FSSD: A–B false negative polyp cases; C false positive identification; D–E–F true positive identifications of MP-FSSD model.**



(a) FSSD detection results

(b) MP-FSSD detection results

**FIGURE 12. Comparison between the FSSD300 and the proposed MP-FSSD300 on the colonoscopy test dataset. True bounding boxes with IoU of 0.5 or higher with the bounding predicted boxes are drawn in green and red colors, respectively. (a) Results of the FSSD. (b) Results of the proposed method.**

pixels, which will also hinder the processing of the blurred image. However, the proposed MP-FSSD algorithm can reasonably classify and detect small polyps and accurately distinguish them from the surrounding normal areas, limited by different debris. Compared to FSSD, the bottom row in Fig. 11 (D–E–F) shows some true-positive identification cases of MP-FSSD in different normal mucosa areas.

To prove the effectiveness of the proposed MP-FSSD model not only in WCE polyp localization cases, but also in colonoscopy image detection (CVC-ClinicDB and ETIS-Larib). Fig. 12 shows a graphical representation of the FSSD (row 1) and MP-FSSD (row 2) detection examples on the CVC-ClinicDB and ETIS-Larib test sets. Polyp detection consists of the identification of polyp regions contained in colonoscopy video frames and the rejection of regions containing normal, blurry tissues, and others showing feces or water jet sprays to clean the colon. Furthermore, colonoscopy frames may contain many distractors. Unfortunately, the rejection of the region showing feces was not reached by the FSSD detector, as depicted in Fig. 12(C), compared to MP-FSSD (Fig. 12(F)). However, both detectors succeeded in detecting a hard polyp case in which the ground truth was not annotated by experts (Figs. 12 (C) and 12 (F)).

## V. CONCLUSION

This work is aimed at the need for effective polyp abnormalities localization and detection in both WCE and colonoscopy datasets. Several approaches have been published and regularly updated in the field of colonoscopy. However, research on the use of deep learning for polyp detection from WCE frames is limited owing to the absence of standard and public datasets, which has pushed the researchers to use their datasets in most cases. Besides the problem of medical ethics, many other reasons may affect the results of state-of-the-art approaches and lead to subjective performance. In this paper, we propose a deep polyp detector (MP-FSSD), an enhanced version of the FSSD to model the visual appearance of small polyp regions that contain an edge extraction network and a polyp detection network using the same VGG16 backbone. Therefore, a slight drop in the detection speed FPS of the FSSD model with remarkable improvement in terms of (mAP) measure. An efficient feature fusion module was applied to the SSD framework to combine the embedded pooling layer with different feature layers and generate new pyramid feature maps. The experimental results prove that feature maps from different layers can be fully fused by simple concatenation rather than an element-wise summation. In the previous qualitative analysis, after adding the feature pyramid module, the polyp detection mAP was greatly improved. The effectiveness of this module was also confirmed using the annotated ETIS-Larib dataset. Although the MP-FSSD network uses a pre-trained dataset and is fine-tuned on WCE/colonoscopy datasets to perform polyp detection, it can achieve advanced performance on three testing datasets with real-time processing speed and more compact models. Moreover, MP-FSSD shows good detection effects for small polyp regions compared with FSSD in rejecting normal parts under specific circumstances. First, all the feature maps of different scales are fused once in the topmost feature map in the multiscale pyramid module to obtain more semantic and rich features. Second, we used only one backbone for both the edge extraction and polyp detection networks, as well as one horizontal connection to reduce the amount of repetitive computation, which shortens the detection time.

In the future, to improve the performance on WCE or colonoscopy datasets, or both, MP-FSSD will be enhanced using more powerful backbone networks such as DenseNet [74]. To further improve the robustness and effectiveness, it is worthwhile to explore a filtration strategy to regularize the edge-pooling layer. Because the proposed system detects polyps based on the entire image, the influence of the background (feces, debris, and other circumstances) cannot be completely avoided.

## REFERENCES

[1] A. Garrido, R. Sont, W. Dghoughi, S. Marcoval, J. Romeu, G. Fernandez-Esparrach, I. Belda, and M. Guardiola, "Automatic polyp detection using microwave endoscopy for colorectal cancer prevention and early detection: Phantom validation," *IEEE Access*, vol. 9, pp. 148048–148059, 2021.

[2] E.-H. Dulf, M. Bledea, T. Mocan, and L. Mocan, "Automatic detection of colorectal polyps using transfer learning," *Sensors*, vol. 21, no. 17, p. 5704, 2021.

[3] S. Charfi, M. El Ansari, and I. Balasingham, "Computer-aided diagnosis system for ulcer detection in wireless capsule endoscopy images," *IET Image Process.*, vol. 13, no. 6, pp. 1023–1030, 2019.

[4] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[5] M. Souaidi, S. Charfi, A. A. Abdelouahad, and M. El Ansari, "New features for wireless capsule endoscopy polyp detection," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–6.

[6] A. Ellahyani, I. E. Jaafari, S. Charfi, and M. E. Ansari, "Detection of abnormalities in wireless capsule endoscopy based on extreme learning machine," *Signal, Image Video Process.*, vol. 15, no. 5, pp. 877–884, Jul. 2021.

[7] S. Charfi and M. E. Ansari, "Computer-aided diagnosis system for colon abnormalities detection in wireless capsule endoscopy images," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 4047–4064, Feb. 2018.

[8] M. Souaidi, A. A. Abdelouahad, and M. E. Ansari, "A fully automated ulcer detection system for wireless capsule endoscopy images," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, pp. 1–6.

[9] M. Souaidi, A. A. Abdelouahed, and M. El Ansari, "Multi-scale completed local binary patterns for ulcer detection in wireless capsule endoscopy images," *Multimedia Tools Appl.*, vol. 78, no. 10, pp. 13091–13108, May 2019.

[10] S. Charfi and M. El Ansari, "A locally based feature descriptor for abnormalities detection," *Soft Comput.*, vol. 24, no. 6, pp. 4469–4481, Mar. 2020.

[11] M. Souaidi and M. El Ansari, "Automated detection of wireless capsule endoscopy polyp abnormalities with deep transfer learning and support vector machines," in *Proc. Int. Conf. Adv. Intell. Syst. Sustain. Develop. (AI2SD)* Cham, Switzerland: Springer, 2020, pp. 870–880, 978-3-030-90633-7.

[12] X. Liu, "A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," in *Artificial Intelligence in Lung Cancer Pathology Images*. Amsterdam, The Netherlands: Elsevier, vol. 1, no. 6, pp. e271–e297, 2019.

[13] A. Nogueira-Rodríguez, R. Domínguez-Carbajales, F. Campos-Tato, J. Herrero, M. Puga, D. Remedios, L. Rivas, E. Sánchez, Á. Iglesias, J. Cubiella, F. Fdez-Riverola, H. López-Fernández, M. Reboiro-Jato, and D. Glez-Peña, "Real-time polyp detection model using convolutional neural networks," *Neural Comput. Appl.*, pp. 1–22, 2021.

[14] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and Ø. Hovde, "Y-net: A deep convolutional neural network for polyp detection," 2018, *arXiv:1806.01907*.

[15] X. Chen, K. Zhang, S. Lin, K. F. Dai, and Y. Yun, "Single shot multibox detector automatic polyp detection network based on gastrointestinal endoscopic Images," *Comput. Math. Methods Med.*, vol. 2021, Nov. 2021, Art. no. 2144472.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[17] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*.

[18] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[20] D. Jifeng, L. Yi, H. Kaiming, and S. Jian, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst* vol. 29. Red Hook, NY, USA: Curran Associates, 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/577ef1154f3240ad5b9b413aa7346a1e-Paper.pdf

[21] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[22] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 21–37.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[25] B. Li and M. Q.-H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10952–10958, Sep. 2012.

[26] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 529–535, Apr. 2016.

[27] P. Sasmal, M. Bhuyan, Y. Iwahori, and K. Kasugai, "Colonoscopic polyp classification using local shape and texture features," *IEEE Access*, vol. 9, pp. 92629–92639, 2021.

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

[29] 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.

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

[31] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image deep mining on a very large-scale radiology database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1090–1099.

[32] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, and I. Balasingham, "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Feb. 2017.

[33] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[34] B.-L. Chen, J.-J. Wan, T.-Y. Chen, Y.-T. Yu, and M. Ji, "A self-attention based faster R-CNN for polyp detection from colonoscopy images," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103019.

[35] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets," *Comput. Biol. Med.*, vol. 141, Feb. 2021, Art. no. 105031.

[36] C. Hong-Tae, L. Ho-Jun, H. Kang, and S. Yu, "SSD-EMB: An improved SSD using enhanced feature map block for object detection," *Sensors*, vol. 21, no. 8, p. 2842, 2021.

[37] Y. Tian, L. Z. Pu, R. Singh, A. D. Burt, and G. Carneiro, "One-stage five-class polyp detection and classification," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 70–73.

[38] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 79–83.

[39] D. Wang, N. Zhang, X. Sun, P. Zhang, C. Zhang, Y. Cao, and B. Liu, "AFP-Net: Realtime anchor-free polyp detection in colonoscopy," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 636–643.

[40] P. Sasmal, Y. Iwahori, M. Bhuyan, and K. Kasugai, "Active contour segmentation of polyps in capsule endoscopic images," in *Proc. Int. Conf. Signals Syst. (ICSigSys)*, May 2018, pp. 201–204.

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

[42] H. Zheng, H. Chen, J. Huang, X. Li, X. Han, and J. Yao, "Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained CNN," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 79–82.

[43] R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognit.*, vol. 83, pp. 209–219, Nov. 2018.

[44] M. Liu, J. Jiang, and Z. Wang, "Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network," *IEEE Access*, vol. 7, pp. 75058–75066, 2019.

[45] A. Tashk and E. Nadimi, "An innovative polyp detection method from colon capsule endoscopy images based on a novel combination of RCNN and DRLSE," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2020, pp. 1–6.

[46] M. Misawa, S. Kudo, Y. Mori, K. Hotta, and K. Ohtsuka, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967, 2021.

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

[48] X. Jia, X. Mai, Y. Cui, Y. Yuan, X. Xing, H. Seo, L. Xing, and M. Q.-H. Meng, "Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction," *IEEE Trans. Automat. Sci. Eng.*, vol. 17, no. 3, pp. 1570–1584, Jul. 2020.

[49] X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, and J. Si, "Real-time gastric polyp detection using convolutional neural networks," *PLoS ONE*, vol. 14, no. 3, Mar. 2019, Art. no. e0214133.

[50] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*.

[51] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.

[52] 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. 98–136, Jan. 2015.

[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[54] W. Ma, X. Wang, and J. Yu, "A lightweight feature fusion single shot multibox detector for garbage detection," *IEEE Access*, vol. 8, pp. 188577–188586, 2020.

[55] Q. Yin, W. Yang, M. Ran, and S. Wang, "FD-SSD: An improved SSD object detection algorithm based on feature fusion and dilated convolution," *Signal Process., Image Commun.*, vol. 98, Oct. 2021, Art. no. 116402.

[56] X. Lu, J. Ji, Z. Xing, and Q. Miao, "Attention and feature fusion SSD for remote sensing object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[57] J.-Y. He, X. Wu, Y.-G. Jiang, Q. Peng, and R. Jain, "Hookworm detection in wireless capsule endoscopy images with deep learning," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2379–2392, May 2018.

[58] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[59] O. We. (1962). *WEO Clinical Endoscopy Atlas*. [Online]. Available: http://www.endoatlas.org/

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

[61] O. We. (2015). *ETIS-Larib Polyp DB*. [Online]. Available: https://polyp.grand-challenge.org/EtisLarib/

[62] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Cham, Switzerland: Springer, 2017, pp. 29–41.

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

[64] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *J. Amer. Statist. Assoc.*, vol. 79, no. 387, pp. 575–583, Sep. 1984.

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

[66] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*.

[67] R. Mash, B. Borghetti, and J. Pecarina, "Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2016, pp. 113–122.

[68] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2020.

[69] L. Wang, F. Chen, and H. Yin, "Detecting and tracking vehicles in traffic by unmanned aerial vehicles," *Automat. Construct.*, vol. 72, pp. 294–308, Dec. 2016.

[70] X. Liu, X. Guo, Y. Liu, and Y. Yuan, "Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102052.

[71] T. Ozawa, S. Ishihara, M. Fujishiro, Y. Kumagai, S. Shichijo, and T. Tada, "Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks," *Therapeutic Adv. Gastroenterol.*, vol. 13, Mar. 2020, Art. no. 1756284820910659.

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

[73] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104519.

[74] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention-based densenet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, Jul. 2020.

**MERYEM SOUAIDI** received the Ph.D. degree in computer science from the Faculty of Sciences, University Ibn Zohr, Agadir, Morocco, in 2020. Her research interests include image processing, machine learning, biomedical image processing, deep learning, and computer vision.

**MOHAMED EL ANSARI** received the Ph.D. degree in computer science from Sidi Mohamed Ben Abdellah University, Fès, Morocco, in 2000. He was a Postdoctoral Fellowship with ENSEIRB-MATMECA, France, from 2001 to 2002. He was an European Research Consortium for Informatics and Mathematics (ERCIM) Postdoctoral Fellowship with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, from 2002 to 2003. He was appointed as a Professor with the University of Ibn Zohr, Agadir, Morocco, from 2004 to 2020. He worked as a Visiting Professor with the Institut National des Sciences Appliquées (INSA) of Rouen, in 2007, 2008, and 2009, Rouen, France, the Ecole Centrale de Nantes, in 2011 and 2012, Nantes, France, and the University of Jean Monnet, in 2013 and 2014, Saint-Etienne, France. He was a Visiting Fullbright Scholar with the University of Nevada, Reno (UNR), USA, in 2010. In 2020, he moved to the Faculty of Sciences, Meknès, Moulay Ismail University, where he is currently working as a Professor with the Department of Computer Science. His research interests include image processing, computer vision, and artificial intelligence.