Multi-scale Features Fusion for the Detection of Tiny Bleeding in Wireless Capsule Endoscopy Images

FENG LU*, Huazhong University of Science and Technology, China
WEI LI*, The University of Sydney, Australia
SONG LIN and CHENGWANGLI PENG, Huazhong University of Science and Technology, China
ZHIYONG WANG, The University of Sydney, Australia
BIN QIAN and RAJIV RANJAN, Newcastle University, United Kingdom
HAI JIN, Huazhong University of Science and Technology, China
ALBERT Y. ZOMAYA, The University of Sydney, Australia

Wireless capsule endoscopy is a modern non-invasive Internet of Medical Imaging Things that has been increasingly used in gastrointestinal tract examination. With about one Gigabyte image data generated for a patient in each examination, automatic lesion detection is highly desirable to improve the efficiency of the diagnosis process and mitigate human errors. Despite many approaches for lesion detection have been proposed, they mainly focus on large lesions and are not directly applicable to tiny lesions due to the limitations of feature representation. As bleeding lesions are a common symptom in most serious gastrointestinal diseases, detecting tiny bleeding lesions is extremely important for early diagnosis of those diseases, which is highly relevant to the survival, treatment, and expenses of patients. In this paper, a method is proposed to extract and fuse multi-scale deep features for detecting and locating both large and tiny lesions. A feature extracting network is firstly used as our backbone network to extract the basic features from wireless capsule endoscopy images, and then at each layer multiple regions could be identified as potential lesions. As a result, the features maps of those potential lesions are obtained at each level and fused in a top-down manner to the fully connected layer for producing final detection results. Our proposed method has been evaluated on a clinical dataset that contains 20,000 wireless capsule endoscopy images with clinical annotation. Experimental results demonstrate that our method can achieve 98.9% prediction accuracy and 93.5% F_1 score, which has a significant

*Both authors contributed equally to this research.

Authors' addresses: Feng Lu, National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Huazhong University of Science and Technology, Wu Han, China, 430074, lufeng@hust.edu.cn; Wei Li, The Australia-China Joint Research Centre for Energy Informatics and Demand Response Technologies, Centre for Distributed and High Performance Computing, School of Computer Science, The University of Sydney, J12/1 Cleveland St, Darlington, NSW, 2008, Australia, weiwilson.li@sydney.edu.au; Song Lin, slin@hust.edu.cn; Chengwangli Peng, m202073502@hust.edu.cn, National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wu Han, China, 430074; Zhiyong Wang, School of Computer Science, The University of Sydney, J12/1 Cleveland St, Darlington, NSW, 2008, Australia, zhiyong.wang@sydney.edu.au; Bin Qian, B.Qian3@newcastle.ac.uk; Rajiv Ranjan, raj.ranjan@ncl.ac.uk, School of Computing, Newcastle University, 1 Science Square, Newcastle Helix, Newcastle upon Tyne, NE4 5TG, United Kingdom; Hai Jin, National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Lab, Cluster and Grid Computing Lab, School of Computer Science and Technology, Huazhong University of Science and Technology, Wu Han, China, 430074, hjin@hust.edu.cn; Albert Y. Zomaya, Centre for Distributed and High Performance Computing, School of Computer Science, The University of Sydney, Darlington, NSW, 2008, Australia, albert.zomaya@sydney.edu.au.

XXXX-XXXX/2020/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

performance improvement of up to 31.69% and 22.12% in terms of recall rate and F_1 score, respectively, when compared to the state-of-the-art approaches for both large and tiny bleeding lesions. Moreover, our model also has the highest AP and the best medical diagnosis performance compared to state-of-the-art multi-scale models.

CCS Concepts: • **Computer systems organization** → *Sensors and actuators*; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → Neural networks.

Additional Key Words and Phrases: Wireless Capsule Endoscopy, deep learning, bleeding lesion detection

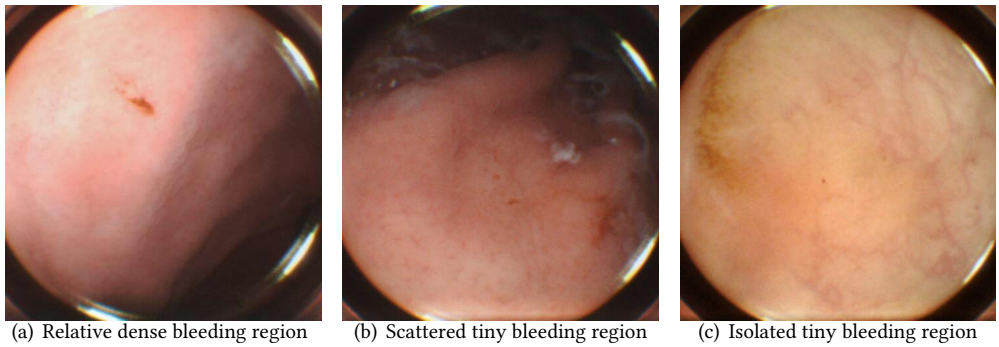
ACM Reference Format:
Feng Lu, Wei Li, Song Lin, Chengwangli Peng, Zhiyong Wang, Bin Qian, Rajiv Ranjan, Hai Jin, and Albert Y. Zomaya. 2020. Multi-scale Features Fusion for the Detection of Tiny Bleeding in Wireless Capsule Endoscopy Images. 1, 1 (August 2020), 20 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the rapid development of non-invasive Internet of Medical Imaging Things, *Wireless Capsule Endoscopy* (WCE) has been increasingly used for examining *Gastrointestinal* (GI) tracts [3]. Compared to the traditional endoscopy approaches, WCE can provide painless GI tract examination to patients, particularly desirable for those who are aged or pain intolerance. Despite the undeniable success of WCE, the vast volume of high resolution images collected from each examination imposes a significant challenge for the doctors when conducting manual assessment and diagnosis. In general, it often takes 8 - 10 hours for each single WCE examination where more than 60,000 images are produced, a basic collection rate of 2 frames per second. Bleeding lesions (a common symptom of GI diseases) [15] are critical to the early detection of serious diseases, which is extremely important to the survival, treatment, and expenses of a patient [5]. However, it is highly difficult and time-consuming for clinicians to accurately identify bleeding lesions in these images by naked eyes, not to mention those tiny bleeding lesions that are unevenly distributed, easily distorted by background or system noise (as shown in Fig. 1).

Cloud computing serves as a feasible approach for automatic image processing where with abundant computation resource and universal examination models, the uploaded images can be processed in a fast fashion. However, concern lies within lesion detection and comes at two folds. Firstly, data transmission and storage expenses are costly. According to a recent statistical study [13], there are 2,232 tertiary hospitals in China with each performing approximately 50-80 cases of gastroscopy examination every day, counting up to 100 terabytes of raw images waiting to be processed per day. To transmit and store all these data in the central server is cost-intensive, especially for only few images that contain lesions in a WCE examination. Secondly, protection of patients' information is a critical issue for medical image process. Data transmission and storage on the cloud poses great threat to the patients' privacy issue in this manner. To address these issues, edge computing [28] can be leveraged for the design of highly secured lesion detection system. Edge computing brings the computation close to where the data is created and is efficient in reducing the bandwidth consumption, thus reduce the overall system latency. Also, the data can be physically isolated and managed within the edge networks after collection, making it ideal for privacy-aware medical image processing.

Nowadays, deep learning-based methods, e.g., convolutional neural networks (CNN), have been proposed for GI tract examination of WCE images [16, 34, 35] due to the great success of deep learning in computer vision applications, including object detection and classification. However, detecting and localizing tiny bleeding lesions in the WCE-based GI tract examination has not been well studied and remains a challenging topic. Most existing lesion region detection methods have not made full use of the multi-scale features extracted from CNNs. Similarly, the typical



(a) Relative dense bleeding region (b) Scattered tiny bleeding region (c) Isolated tiny bleeding region

Fig. 1. WCE images with tiny bleeding lesions

object detection methods, like Faster-RCNN [27], only utilize the top-level feature map. They limit their capability in detecting tiny bleeding lesions. Multi-scale feature methods without multi-scale feature fusion like *Single Shot MultiBox Detector* (SSD) [22] cannot combine semantic and context information for detection. Multi-scale feature fusion methods, such as *Feature Pyramid Networks* (FPN) [21], detect objects by fusing feature maps from the top down. However, existing multi-scale feature fusion methods often introduce background noise while they fuse the whole feature maps. Besides, for the methods based on FPN, the high-level feature maps cannot use context information from the low-level feature maps in the feature pyramid.

Considering the aforementioned concerns of model design for small target detection, we aim to develop a holistic edge analytic framework for detection and localization of bleeding lesions in WCE images, assisting the doctors in making early diagnosis of serious gastric diseases. In this work, we propose a deep multi-scale feature fusion network for sensing both large and tiny bleeding lesions in WCE images of various sizes. In summary, the key contributions of this work are as follows:

- We propose a multi-scale feature extraction and fusion network for accurate detection and localization of bleeding lesions.
- A multi-scale regional proposal network is designed to generate proposals from feature maps at different layers. Feature fusion of our design only needs to perform on proposed candidates rather than the whole feature maps.
- A top-down feature fusion network integrates shallow features with high-level semantic features to obtain a feature map of the highest resolution while ensuring the positioning sensitivity of detecting small objects.
- Our proposed method was evaluated by an approved clinical trial dataset. The experimental results showed that, compared with the latest bleeding lesion detection algorithm, our approach could achieve performance improvement of up to 31.69% and 22.12% in terms of recall rate and F_1 score, and has a prediction accuracy of 97.8%. Compared with the state-of-the-art multi-scale feature fusion detection algorithms, our approach can achieve performance improvement of up to 4.88% on average precision (AP).

The rest of the paper is organized as follows. In Section 2, we present a brief literature review on bleeding detection, object detection, and small object detection. In Section 3, we provide the technical details of our proposed method for detecting and locating bleeding lesions. In Section 4, we conduct multiple experiments to evaluate the performance of our design and discuss the results. Finally, we conclude and discuss the work in Section 5.

2 RELATED WORK

Our work is substantially related to both bleeding lesion detection in WCE images and object detection. To allow a better understanding of the defining characteristics of small object detection, we divide this section into three parts, namely, bleeding lesion detection, small object detection, and multi-scale feature fusion for small object detection.

2.1 Bleeding Lesion Detection

The existing lesion bleeding detection approaches mostly utilize a pipeline consisting of feature extraction and classification. For example, Cui et al. [4] proposed six color features in HSI color space and used the Support Vector Machine (SVM) classifier for detecting bleeding sites. In [26], Local features from different levels are combined to train a neural network cell-classifier for bleeding classification. Several approaches were further proposed to extract features from different color spaces such as RGB, HSV, YIQ with block histograms and to use various classifiers such as k-Nearest Neighbor (KNN) or Artificial Neutral Network (ANN) [7, 9, 10, 31]. In [8], the regions of interest (ROIs) in a particular composite color space Y.I/Q are selected and a novel method was proposed to combine the local features of ROIs. In [34], the authors designed a CNN-based feature extractor instead of the traditional feature extraction method for endoscopy image lesion detection. Moreover, there are also studies on combining handcrafted features and CNN-based features for gastrointestinal bleeding detection [17]. These results further demonstrate that the schemes based on CNN outperform the traditional methods based on color and texture. For example, to classify digestive organs in WCE images, [35] utilized DCNN to learn layer-wise hierarchy models from real WCE images, and achieved better performance than traditional classification methods. In [16], an eight-layer CNN was constructed for bleeding detection in WCE images, and it outperforms the solutions using handcrafted features.

2.2 Small Object Detection

Object detection is a well-studied topic in computer vision and its evolution is summarized in Fig. 2. Small object detection is a type of object detection methods, where targets are often of small sizes. It has attracted increasing research attention due to detecting small objects is a very challenging and difficult task in the real world [14, 19].

Deep learning is commonly used for small object detection. The existing solutions mainly focus on two aspects, context information and proposal generation. Context plays an essential role in small object detection. FPN [21] demonstrates the importance of context for small object detection. It makes full use of context and high-level semantic information to detect objects at different scales by constructing feature pyramids. Generating object proposals or bounding boxes in an image is a prerequisite for accurate classification or segmentation. By using CNN instead of sliding windows to generate proposals, the region proposal network proposed in Faster-RCNN [27] greatly improved the computational accuracy and efficiency.

Besides, to make the proposals more suitable to detect small objects, the feature maps of different resolutions are used in the network to detect small, medium, and large types of objects [18]. Multi-scale features methods, such as *a unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection* (MS-CNN) [1], SSD [22] were proposed to improve performance in detecting small objects. They often suffer from the low detection rate as no features fusion is performed during the detection.

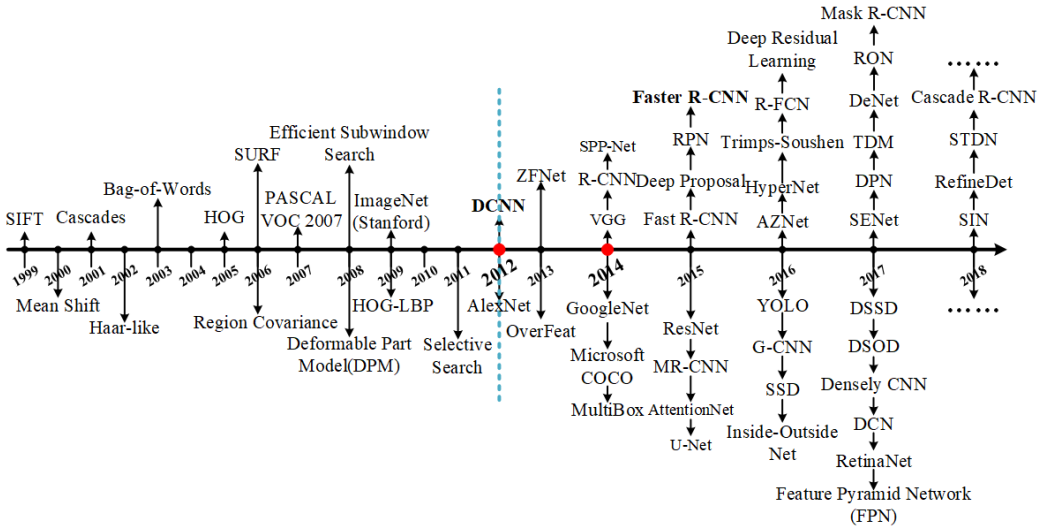


Fig. 2. A timeline summary of research on object detection.

2.3 Multi-scale Feature Fusion Models

Nowadays, most small objection detection methods use multi-scale feature fusion and perform well. In 2017, a well-known multi-scale feature fusion method, FPN [21], is proposed. Since then, many variants, such as FFSSD [2], RefineDet [32], M2Det [33], and DeFusionNet [30] were proposed to combine context and semantic information through feature fusion. Some of them improve the performance of object detection. RefineDet is a method with a feature pyramid structure and shows better efficiency and accuracy on VOC dataset. Some methods made a marked improvement on a specific topic or goal. DeFusionNet is such a method for defocus blur detection.

Existing feature fusion technologies based on feature pyramid structure improve the performance in detecting small targets. However, they still have difficulties dealing with WCE images as the bleeding lesions are variable in size, irregular in shape, and random in distribution. This is mainly due to the existing approaches often fuse the whole feature map to reduce the efficiency of the model and introduce background noise. Furthermore, high-level feature maps used in most current methods are based on FPN, such as DFPN [20] and Augfpn [11]. They cannot combine low-level feature maps with context information, which show a low performance on localizing the object.

3 PROPOSED METHOD

The core of our bleeding detection model is to fuse the proposed multi-scale features generated from the feature pyramid instead of the whole feature map. As illustrated in Fig. 3, the overall design of our proposed model for bleeding detection and localization consists of a backbone network and two branch networks. The backbone network extracts the basic features of a WCE image, while the two branch networks select and fuse the regional features of feature maps. More specifically, a typical feature extracting network, *residual network* (ResNet) [12], is first used as our backbone network to extract the basic features of an input image. Then, the multi-scale regional proposal branch network selects the candidate targets of bleeding regions from the feature maps of different layers. After that, the features of candidate regions are fused through the top-down feature fusion branch network. Finally, the feature map with the largest scale is sent to the subsequent network

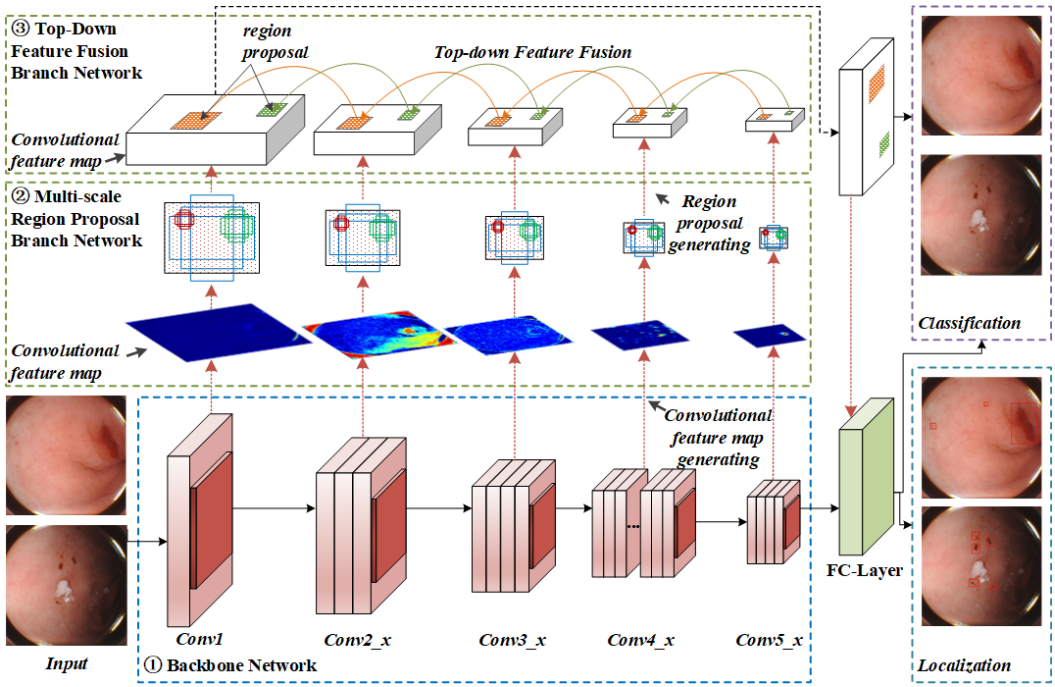


Fig. 3. Architecture overview of our proposed bleeding detection and localization method

to produce target bleeding regions. Since the branch networks and the backbone network are independent of each other, they can be run in parallel to speed up the model training process in edge nodes.

3.1 Backbone Network

To restrain network degradation and reduce parameters while improving accuracy, we employ ResNet101 as our backbone network in this work. Many other feature extraction networks, such as VGG16, VGG19, and ResNet50, are also available for the same purpose. VGG16 is a DCNN that uses smaller convolution kernels to extract deep features, while ResNet is a deeper network than VGG [12]. By using a residual network structure and introducing a building block and a bottleneck structure, the deeper network can thus extract deep features and detect objects more accurately. To provide an intuitive understanding of our choice, a simple comparison between VGG and ResNet is provided in Table. 1.

Although the main function of the convolutional layer is extracting features, we hierarchically combine the convolutional layers in ResNet to extract features more efficiently. As shown in Table. 1, the ResNet model is divided into five layers, from Conv1 to Conv5_x. Different layer involves different number of bottlenecks, e.g., Conv2_x contains 3 bottlenecks, while Conv4_x has 23. Each bottleneck is a combination of convolutional layers, *Batch Normalization* (BN) layer, and *rectified linear units* (ReLU) layer. The purpose of using BN and ReLU layer is to accelerate the convergence of the training process, reduce overfitting, and improve the performance of the network model. As a DCNN with 101 layers, the features extracted from ResNet101 will become more and more abstract with the deepening of the network layers. Besides, the max-pooling method is adopted in

Table 1. The simple comparison of VGG and ResNet from architecture.

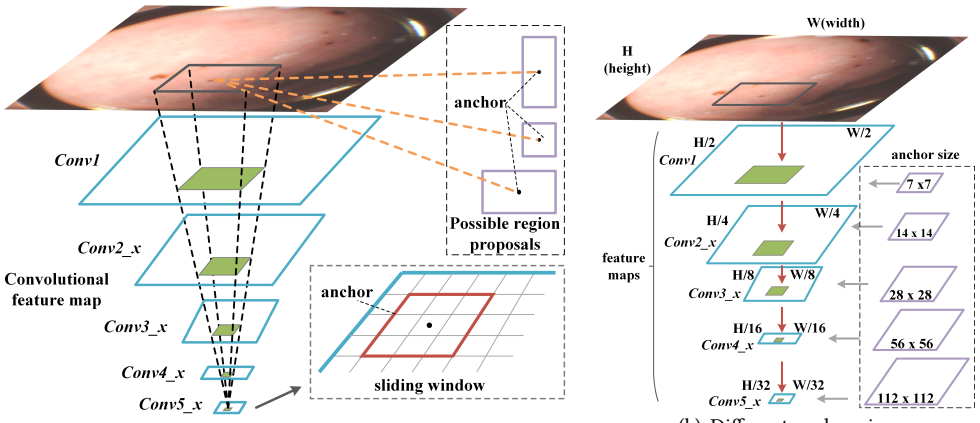
Layer name	VGG			ResNet		
	16-layer	19-layer	Output size	50-layer	101-layer	Output size
Conv1	$\begin{bmatrix}3\times3, & 64\\ 3\times3, & 64\end{bmatrix}\times1$	$\begin{bmatrix}3\times3, & 64\\ 3\times3, & 64\end{bmatrix}\times1$	224×224	$\begin{bmatrix}7\times7, & 64\end{bmatrix}\times1$	$\begin{bmatrix}7\times7, & 64\end{bmatrix}\times1$	112×112
	2×2 max pool,stride=2			3×3 max pool,stride=2		
Conv2_x	$\begin{bmatrix}3\times3, & 128\\ 3\times3, & 128\end{bmatrix}\times1$	$\begin{bmatrix}3\times3, & 128\\ 3\times3, & 128\end{bmatrix}\times1$	112×112	$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$	56×56
	2×2 max pool,stride=2			2×2 max pool,stride=2		
Conv3_x	$\begin{bmatrix}3\times3, & 256\\ 3\times3, & 256\\ 3\times3, & 256\end{bmatrix}\times1$	$\begin{bmatrix}3\times3, & 256\\ 3\times3, & 256\\ 3\times3, & 256\\ 3\times3, & 256\end{bmatrix}\times1$	56×56	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times4$	28×28
	2×2 max pool,stride=2			2×2 max pool,stride=2		
Conv4_x	$\begin{bmatrix}3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\end{bmatrix}\times1$	$\begin{bmatrix}3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\end{bmatrix}\times1$	28×28	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times23$	14×14
	2×2 max pool,stride=2			2×2 max pool,stride=2		
Conv5_x	$\begin{bmatrix}3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\end{bmatrix}\times1$	$\begin{bmatrix}3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\\ 3\times3, & 512\end{bmatrix}\times1$	14×14	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$	7×7
	2×2 max pool,stride=2			average pool		
	FC-4096 FC-4096 FC-1000	FC-4096 FC-4096 FC-1000	1×1	FC-1000	FC-1000	1×1
	softmax			softmax		

the pooling layer to highlight the main texture features and minimize the interference of image background information.

3.2 Multi-scale Region Proposal Branch Network

An RPN is a network that can directly identify candidate target regions from the feature map. Compared with the traditional methods always use sliding window and searching algorithm, RPN is more efficient and flexible. As shown in Fig. 4(a), to generate the region proposals, the small convolutional network (kernel size set to 3*3) slides on the feature map like a sliding window with one pixel sliding step. The central position of each sliding window is called an anchor. Then, the anchor will be mapped back to the original image, so that the multiple candidate region proposals are generated based on different scales and aspect ratios. Here, for positioning the bleeding area in advance, the proposal is represented by the vector (x,y,h,w), respectively giving the coordinates of the anchor, height and width of the proposal. The ReLU is used as the activation function. Similar as the methods of S. Ren [27], the nine possible region proposals generated by three scales (i.e., 128*128, 256*256, 512*512) and three aspect ratios (i.e., 1:1, 1:2, 2:1), which are always considered to cover all possibilities, are used in our branch network. In this way, the selective search is no long needed, and the RPN can greatly improve the proposed region generating speed.

Afterward, the candidate region is classified according to a certain score. That means, when the score reaches or exceeds a threshold, the proposal is considered to have a target and retained. Otherwise, the proposal contains no useful information. In our branch network, the Intersection-over-Union (IoU) parameter, representing the overlapping rate of the proposed region and the marked area, is used to make the evaluation. To be more specific, if the IoU value of a proposal is higher than 0.7, which indicates that the region proposal contains the target of interest and needs



(a) Region proposal generating procedure for feature map (b) Different anchor size

Fig. 4. Multi-scale region proposal branch network

to be retained. If a proposal has an IoU value less than 0.7, the proposal is discarded. As known, 0.7 is a common choice for IoU used in RPN with good performance.

One of the important issues for developing an RPN-based branch network to generate candidate region proposals is to specify the size of the proposal center on the anchor (anchor size). Most methods for generating proposals are targeted for one layer's feature map and using a single-scale template. This approach has been used in [27] for generating a feature map from Conv5_x to cover all possibilities. However, as shown in Table. 1, different convolutional feature maps have different size. The feature map size (output size) of Conv1 is 112*112, but the size of Conv5_x is 7*7 due to the role of the pooling layer. Consistent with this, the mapping area of a one-pixel point in the different feature maps represents different areas in the original image.

Therefore, we design a multi-scale region proposal branch network in which the anchor size is negatively related to the feature map's size to generate candidate region proposals. With the deepening of the feature layers, the feature map will become smaller, but each pixel area in the original image will be larger. This also makes the anchors become larger. For example, if nine anchors of (14*14, 14*28, 28*14, 28*28, 28*56, 56*28, 56*56, 56*112, 112*56) are used in the feature map of Conv2_x, the sizes of these nine anchors in the feature map of Conv3_x turn into (28*28, 28*56, 56*28, 56*56, 56*112, 112*56, 112*112, 112*224, 224*112). With nine anchors used in every level of the feature pyramid, we have 45 anchors in total in our approach.

The other issue is the potential overlapping of the generated proposals. When an anchor slides on the feature map, a large number of proposals are created, and each proposal is highly likely overlapping its neighbors. Thus, a Non-Maximum Suppression (NMS) algorithm is adopted to remove the invalid or duplicated proposals. Through the multi-scale region proposal branch network, the target candidate region for the bleeding area of different feature map can be chosen, which cannot only reduce the workload of feature fusion, but also highlight the target feature to improve the detection accuracy.

3.3 Top-down Feature Fusion Branch Network

In general, shallow features extracted from low-level feature maps have more contextual information, while deep features extracted from high-level feature maps always contain more semantic information. Comparing the feature map of the Conv2_x and Conv5_x, as shown in Fig. 5, it is

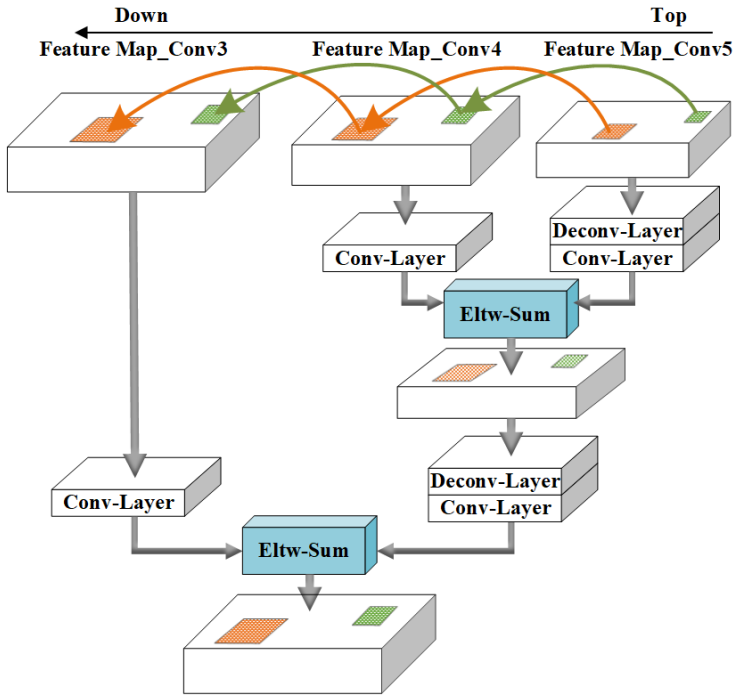
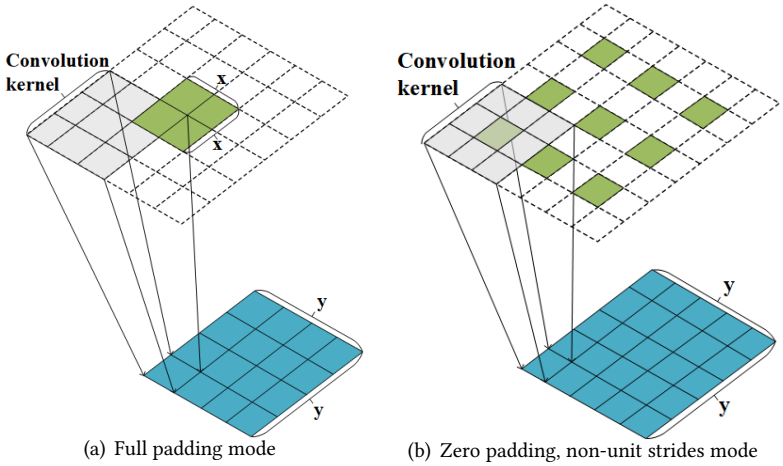


Fig. 5. Procedure of the top-down feature fusion approach used in Conv3, Conv4 and Conv5 layer

easy to observe that some of the contextual information is lost along with the growth of the layers. The reason lies in the convolution of multiple layers, especially the pooling layer. To indicate the features belonging to the target object or the background, the maximum pooling method is selected in the pooling process. However, there is a catch that the features of the small objects are often "lost" when they are passing through multiple pooling layers. With the growth of the convolutional layers, the loss of such features becomes more serious. Thus, the branch network should be designed to fuse the context information and semantic information of the different feature maps.

The most common method of fusing feature maps is to employ the feed-forward method from the low-level to the high-level layers. By doing so, the shallow features are added to the deep features and both can be represented in the final fusion feature map is output at high-level layers. However, since the high-level feature map is coarse-grained and insensitive to position, this feed-forward method is not suitable for our bleeding detection and localization network. It is notable that the low-level feature map is fine-grained and the localization information of the target region can be obtained more accurately from this layer. Therefore, a top-down feature fusion approach is used in our branch network.

Because of the pooling layer, the size of the feature map is decreasing layer by layer, as shown in Table. 1. Hence, to fuse the features of different convolutional layers, the size of the feature map should be enlarged from the high-level layer to the low-level layer. As shown in Fig. 5, the deconvolution layer is used in this top-down feature fusion approach to enlarge the size of the feature map [24, 25] and to fuse the features layer by layer. Furthermore, for the features fusion of Conv5 and Conv4 in Fig. 5, the feature map of Conv5 is firstly enlarged by the deconvolution layer. Then the features both from Conv5 and Conv4 are uniformly processed though the convolutional



(a) Full padding mode (b) Zero padding, non-unit strides mode

Fig. 6. Deconvolutional operation

layer and the following BN layer. After that, the features are superimposed together by an Eltw-Sum module and get the first fused feature map (Fm_Conv5_Conv4). Here, the Eltw-Sum module is used for the dot product of the pixel matrix. Afterward, a new round of fusion for Fm_Conv5_Conv4 and Conv3 is started. Hereby, all of the feature maps of different layers are fused from top to down. Finally, the last feature map is a 1*1 convolutional layer, which can be used for dimensionality reduction and feature reorganization. Through this feature fusion model, background noise interference can be reduced, while context and semantic information can be profitably fused.

The common methods of size enlargement used in the deconvolutional operation are shown in Fig. 6, in which, the input is the green part of the above and the output is the blue map below. For example, the input is a 2*2 green map and output is a 4*4 blue map in Fig. 6(a), and the input is a 3*3 map and output is a 5*5 map in the Fig. 6(b). Assuming the input size is x, the convolution kernel size is k, the sliding step size is s, and the deconvolution output result size is y. Then, for the full padding mode in Fig. 6(a),

$$y = (x - 1) * s + k. \tag{1}$$

For zero padding, non-unit strides mode in the Fig. 6(b),

$$y = \lfloor \frac{2 * x + 1 - k}{s} \rfloor. \tag{2}$$

After all, this top-down converged network module should be combined with the multi-scale region proposal branch network. Before the top-down feature fusion, the region proposals in different layers are searched on the feature map by the multi-scale RPN network. By then, the candidate target regions from different layers are subjected to feature fusion, ignoring non-target areas. This cannot only greatly shorten the processing time of feature fusion steps, but also improve the efficiency and accuracy of the whole model. This is because through fusing the candidate target features from different layers, a more expressive and comprehensive feature map for recognizing can be obtained. Thus, the accuracy of the detection for tiny bleeding lesions can be improved.

3.4 Classification and Localization

After the operations of the above branch networks, we get the feature map of the most fine-grained level with the candidate bleeding regions. Then the feature map is sent to the Full Connection Layer (FC-Layer) to convert high-dimensional features into one-dimensional features for achieving classification. The next steps are to choose the proper bleeding regions by classification method, correct the position by bounding box regression algorithm, and eliminate the redundant regions by non-maximum suppression algorithm again.

Softmax for Classification. As a typical logistic regression classifier that widely used in classification problems, the softmax classifier is a function whose output is the probability of a category [23]. The form of the softmax function is usually given by the following formula:

$$L = \frac{1}{N}\sum_{i} -log(\frac{e^{f_{yi}}}{\sum_j e^{f_j}}) \tag{3}$$

To let the i-th input feature with the label y_i, where N is the number of training data, and f_j denotes the j-th element of the vector of class scores f. We use the softmax function to calculate the score of each proposal. After that, we sort the scores of all the proposals and select the scores higher than 0.8 as the final result.

Bounding box Regression for Localization. It is well known that, in the RPN-based branch network, the localization information for each proposal is recorded by the vector (x,y,h,w). However, the initial region of the proposal through the multi-scale region proposal network operation is less accurate. The actual localization of the bleeding area in the WCE image needs to be further adjusted. To complete this task, the bounding box regression used in [6] is adopted in our model to train the network. The method performs well when the estimated coordinates of the proposal are close to the real coordinates.

Non-Maximum Suppression. During the detection processes including region proposal network, a large number of candidate regions are generated at the same target position. These candidate regions always overlap with each other. Although the invalid and duplicate proposals have been removed by the NMS algorithm in the multi-scale region proposal branch network, redundancy still exists. In this case, non-maximum value suppression has been used to find the best target region frame and eliminate the redundant bounding box. Non-maximum suppression algorithms have been widely used in computer vision applications, such as edge detection, face detection, and target detection.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Dataset and Preprocessing

To evaluate our model, 20,000 clinical WCE images are collected, including 15,000 normal images (negative samples) and 5,000 images containing hemorrhagic lesions (positive samples). These samples are taken from multiple patients. This is because images involving lesions tend to account for only a small proportion of the total images. In particular, for a single GI tract examination of one patient, there are only a few dozen or fewer images that can be accurately marked as a sample of hemorrhagic lesion. At the same time, the diversity of data sources can overcome the singleness of train samples and improve the robustness of the model.

Due to the complex physiological environment of the GI tract, the WCE images always contain noises or lack of clarity. To improve the sample quality, Wavelet Transform is used to denoise and enhance the images. In addition, in order to balance the number of positive and negative samples, 90°, 180°, or 270° rotation, and horizontal inversion of positive samples are randomly performed here. After amplification, 15,000 positive samples are obtained through random selection. Then,

the 30,000 WCE images obtained after augmentation are split into three: training, validation, and testing, and the partition ratio are 3:1:1.

All WCE samples are JPEG color images with a resolution of 480*480. Each positive sample is annotated by professional with LabelImg, and generate XML file following the PASCAL VOC format. The location of the bleeding region is completely recorded in the XML file using the coordinate of the bleeding region's upper left and lower right corner. Here, the origin is in the upper left corner of the whole image. Besides, the experimental hardware environment includes NVIDIA TESLA-P100 GPU, Intel Xeon CPU e7-4850 V4, 32GB memory, and 3TB hard disk. Software environments include Ubuntu14.04, Caffe deep learning framework, CUDA8.0, and cuDNN 6.0. It is worthy to note that the computational power of the machine in the clinic environment is better than the one used in our experiments. Furthermore, the stochastic gradient descent (SGD) method is adopted to accelerate the convergence of the training process, which has a total of 80,000 iterations. At this point, the learning rate is reduced once after every 20,000 iterations. The momentum value is set as 0.9, and the weight attenuation value is 0.0001.

4.2 Evaluation Metrics

We used five common performance metrics, sensitivity (recall), specificity, precision, accuracy, and F1 score to evaluate the performance of the models. Since multiple hemorrhagic lesions may exist in each image, Average Precision (AP) is also adopted to assess the model's efficiency in detecting all bleeding lesions in an image.

$$AP = \frac{1}{N} \sum_{r} Pinterp(r). \tag{4}$$

Where N represents the given rank number, and Pinterp(r) represents the region under the P-R (Precision-Recall) curve at different r values.

4.3 Comparison with the State of the Art

To fully evaluate our bleeding detection model's performance, we conducted extensive experiments to compare with some well-known models, including the typical bleeding detection networks, classic object detection models, and multi-scale feature object detection models.

4.3.1 ***Typical Bleeding Detection Networks.*** We choose three typical algorithms used for bleeding detection as the benchmarks, including DCNN-8 [16], VGG-16 [29], and ResNet-101 [12]. DCNN-8 and VGG-16 are image classification algorithms with 8-layer and 16-layer structures, respectively. Their structure is not deep enough to extract the semantic information from images. ResNet-101 is a popular classification algorithm with 101-layer and has been successfully applied in many applications. Table. 2 shows the sensitivity, specificity, precision, accuracy, and F_1 scores of the four models for our experiments. It can be seen that our model outperforms the benchmark algorithms in most terms. Compared with the latest bleeding detection model (DCNN-8), our method's sensitivity is improved by 31.69%, and the F_1 score is increased by 22.12%. Besides, Table. 2 also shows that the overall performance of ResNet-101 with is better than VGG-16 and DCNN-8. This implies that the capacity of containing more feature layers can be attributed to the detection of tiny lesions.

Table. 2 also shows that the specificity and precision of VGG-16 and ResNet-101 are higher than our model. The reason lies in that the VGG-16 and ResNet-101 do not make full use of each feature layer in the network but only use the last layer. Accordingly, they tend to ignore images that contain very few tiny bleeding regions, resulting in higher specificity and lower sensitivity. Comparatively, our model tends to identify all suspected bleeding lesions in the WCE images.

Table 2. Results statistics of bleeding detection work.

Method	Sensitivity	Specificity	Precision	Accuracy	F1 score	Localization
Our method	0.9890	0.8730	0.8862	0.9310	0.9348	Yes
DCNN-8	0.7510	0.7890	0.7807	0.7700	0.7655	No
VGG-16	0.7400	0.9550	0.9426	0.8457	0.8291	No
Resnet-101	0.8310	0.9460	0.9389	0.8885	0.8817	No

It sometimes misdiagnoses negative samples (that do not have a lesion) as positive, resulting in higher sensitivity but lower specificity. Fig. 7 shows the results of our method correctly detect the tiny bleeding regions from the images. In Fig. 8, we show some misdiagnosed samples. In medical applications, the missing diagnosis could easily lead to catastrophe. Therefore, within the range of acceptable specificity, the missed diagnosis rate should be reduced as much as possible, which means the sensitivity should be as high as possible.

4.3.2 ***Classic Object detection Models.*** In this section, we choose Faster-RCNN [27] as our benchmark. As an widely used RPN in the field of target detection, the Faster-RCNN can be directly applied in our case. Table. 3 shows that our approach outperforms Faster-RCNN in all aspects. Furthermore, as shown in Fig. 9, the P-R (precision-recall) curve of our approach also outperforms Faster-RCNN. This indicates that the overall performance of our model is superior to the Faster-RCNN in detecting small lesions.

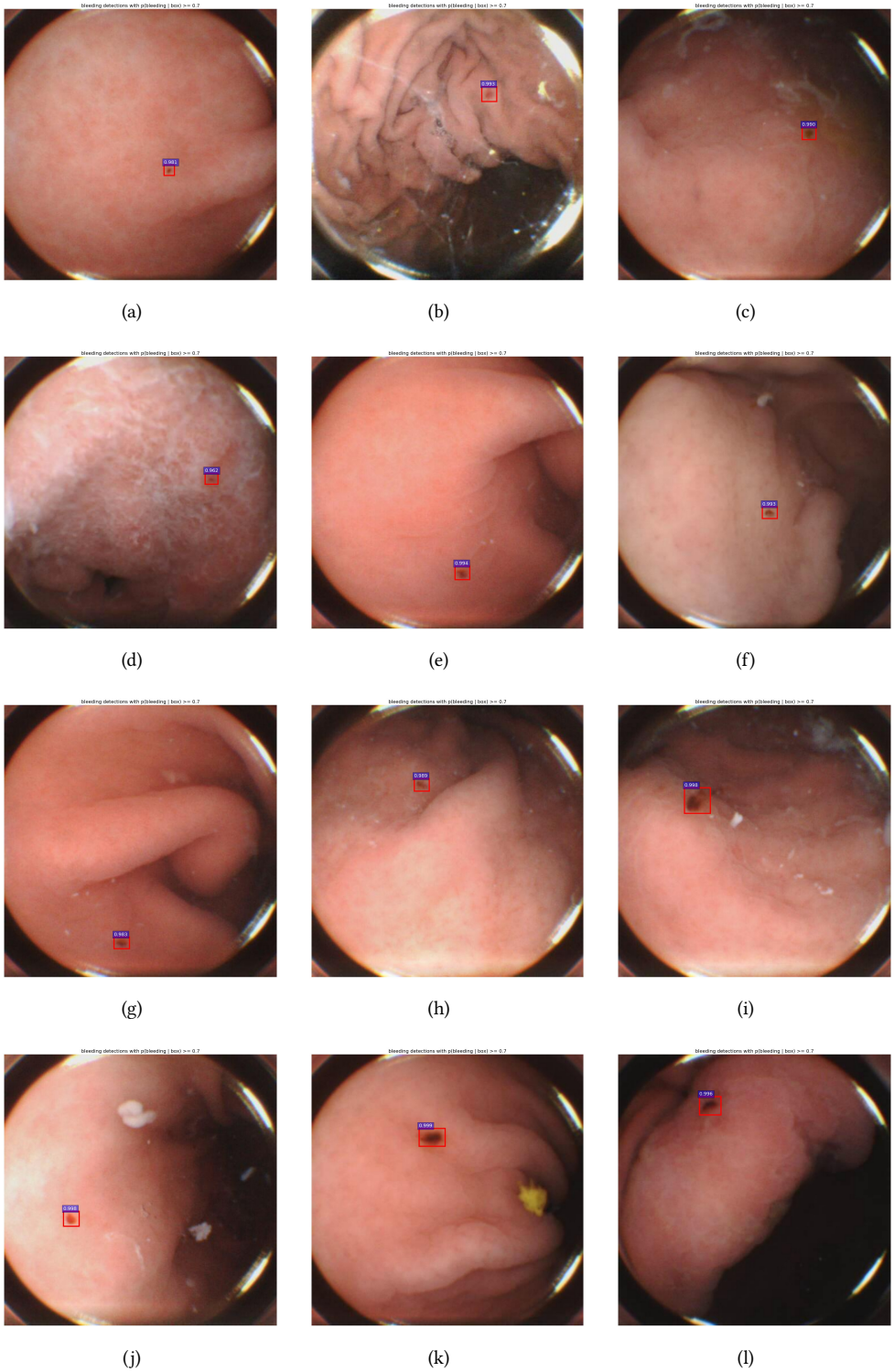
The poor performance of Faster-RCNN is mainly caused by its target detection template is single and fixed in size. In our experiments, only the images containing large hemorrhage lesions can be correctly classified by Faster-RCNN, but for those images containing only small hemorrhagic lesions are often classified as a negative sample. As shown in Fig. 10, our model detects 8 bleeding regions, while the Faster-RCNN only detects 4. This also reflects the significant improvement in AP value, a 22.27% performance difference between our approach and Faster-RCNN.

In the experiments, the large bleeding regions are often accompanied by small ones. As shown in Fig. 11(a), the Faster-RCNN cannot identify the largest hemorrhagic lesions, and also miss some small hemorrhagic lesions. In this case, our model detects the hemorrhagic lesions in different scales as shown in Figure 11(b). This is because our model employs a multi-scale region proposal branch network to enrich its capability of identifying different size lesions in the same image.

Table 3. Results statistics of classic object detection work.

Method	Sensitivity	Specificity	Precision	Accuracy	F1 score	AP
Our method	0.9890	0.8730	0.8862	0.9310	0.9348	0.7520
Faster RCNN	0.9530	0.8450	0.8601	0.8990	0.9042	0.6150

4.3.3 ***Multi-scale Feature Object Detection Models.*** We selected four representative multi-scale feature methods as our performance benchmarks, which are SSD [22] (a fast and popular multi-scale feature method), M2Det [33] (a one-stage, multi-scale feature fusion method), RefineDet [32] (a high-performance multi-scale feature fusion method), and FPN [21] (the classic multi-scale feature fusion method). Table. 4 shows the performance comparison between our proposed method and the benchmarks. Our approach has the highest AP and the best overall performance in medical diagnoses. Moreover, we achieve a comparative performance of the state-of-the-art algorithms in terms of sensitivity, accuracy, and F1 score.



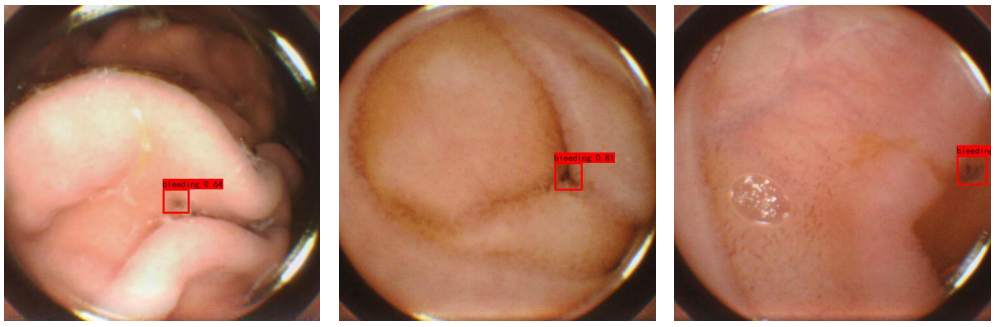
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 7. The samples of successful detection of tiny bleeding lesions



(a) (b) (c)

Fig. 8. Samples of misdiagnosis due to the high similarity of the labelled regions and actual bleeding lesions

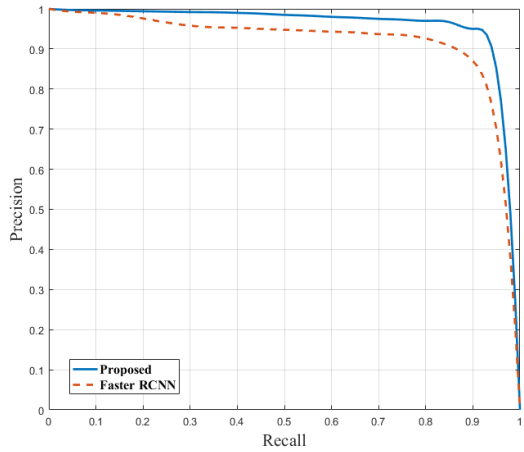


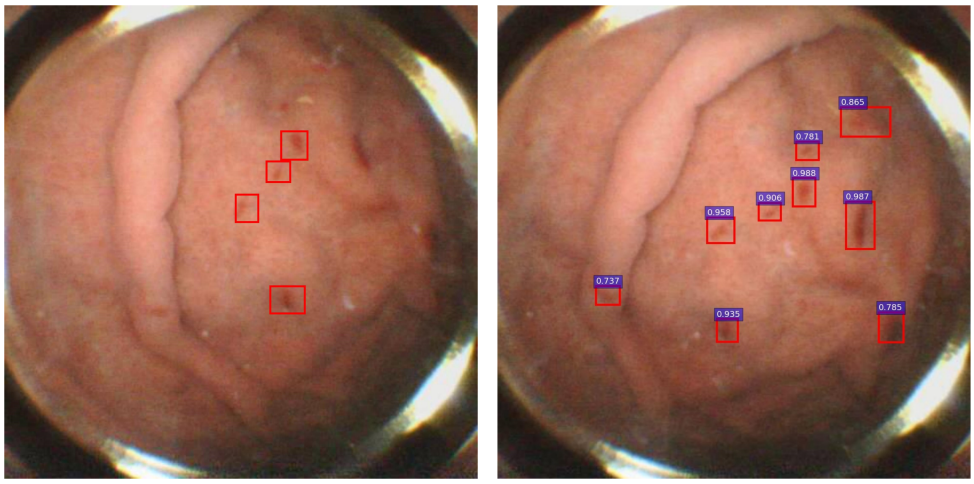
Fig. 9. P-R curve comparing our model with Faster-RCNN

Among them, SSD is a one-stage multi-scale feature method without feature fusion. Without RPN, it cannot combine context information and semantics information, performing poorly in our experiments. M2Det is a one-stage multi-scale feature fusion method. Since it is difficult to get a precise localization in one step, the AP is very low. RefineDet and FPN showed good performance in our experiments since they fuse the whole feature map. However, they introduce background noise during the fusion and ignore the context information available in low-level feature maps. As a result, they often missed the lesions in different sizes.

Our method addresses these issues by fusing the proposal regions in different feature maps in a top-down manner. As shown in Table. 4, our method's AP is 4.88% better than the best benchmark. As shown in Fig.12, our approach can detect all the lesions in a WCE image while the benchmark algorithms failed to do so.

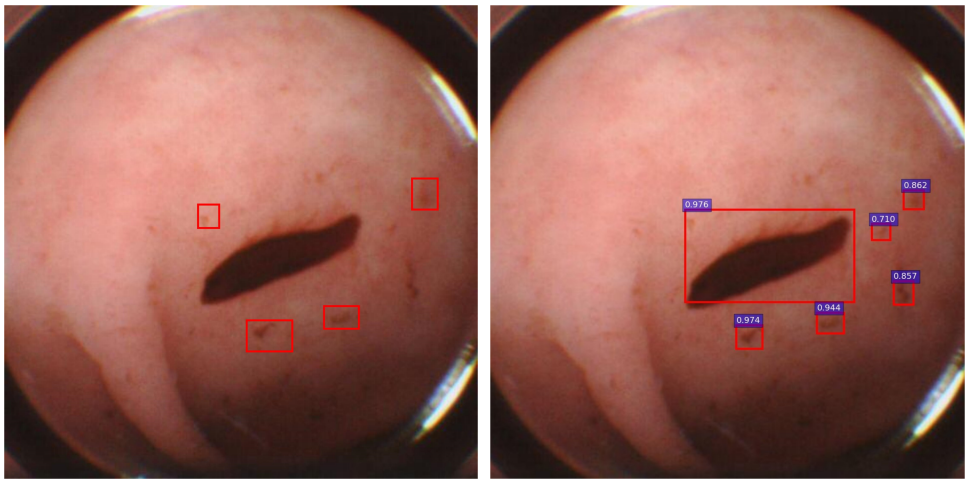
4.4 Ablation Study

4.4.1 ***Baseline***. To investigate the influence of basic feature extraction network on the detection performance of our model, two representative network models, VGG-16 [29] and ResNet-101 [12],



(a) Detection result of RCNN (b) Detection result of our model

Fig. 10. Detection comparison of tiny bleeding lesions



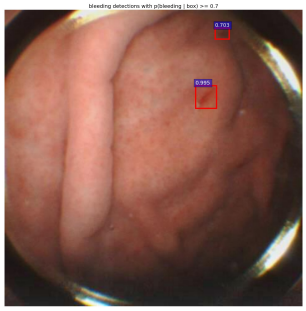
(a) Detection result of Faster-RCNN (b) Detection result of our model

Fig. 11. Detection comparison of multi-scale bleeding lesions

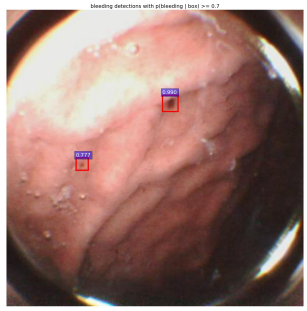
are used for comparative experiments. The results is also given in Table. 5. The results show that the ReNet-101 is superior to VGG-16 not only in overall performance (F_1 score), but also in terms of sensitivity and accuracy. This indicates that the feature extraction of our backbone network can greatly affect detection performance. Because the network model level of ResNet-101 is deeper than that of VGG-16, the feature extraction ability for small lesions of ResNet-101 is superior to that of VGG-16. Here, ResNet-101 is slightly less in specific and accuracy than VGG-16. The reason is the same as the tendency to find more images containing suspected small bleeding lesions.

Table 4. Results statistics of multi-scale feature object detection work.

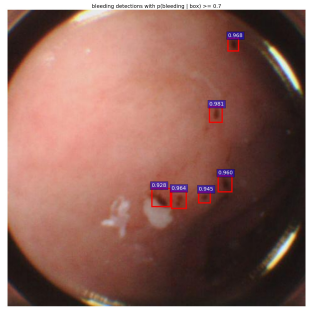
Method	Sensitivity	Specificity	Precision	Accuracy	F1 score	AP
Our method	0.9890	0.8730	0.8862	0.9310	0.9348	0.7520
SSD	0.7440	0.8950	0.8770	0.8200	0.8050	0.6300
M2Det	0.8800	0.8950	0.8930	0.8870	0.8860	0.4940
FPN	0.9590	0.9340	0.9360	0.9470	0.9470	0.7170
Refinedet	0.9890	0.8250	0.8500	0.9070	0.9140	0.6270



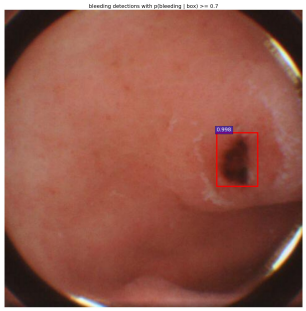
(a) Case1 for small bleeding lesions



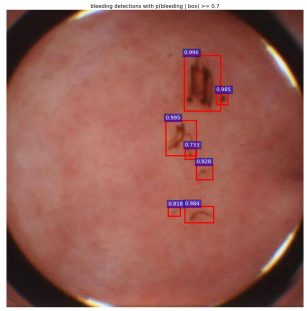
(b) Case2 for a small bleeding lesion



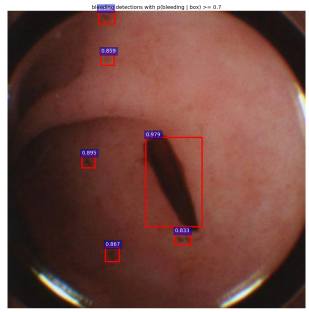
(c) Case3 for a small bleeding lesion



(d) Case for a medium bleeding lesion



(e) Case for medium and small lesions



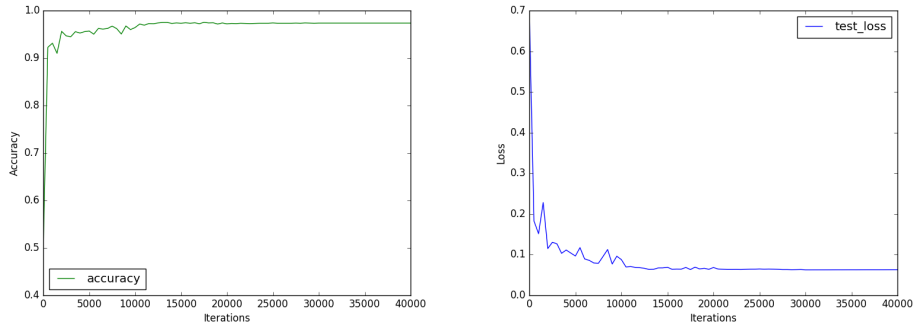
(f) Case for large and small lesions

Fig. 12. Some detection results of multi-scale bleeding lesions

Table 5. Results statistics of backbone network based on VGG-16 and Resnet-101.

Backbone Network	Sensitivity	Specificity	Precision	Accuracy	F1 score
VGG-16	0.7400	0.9550	0.9426	0.8457	0.8291
Resnet-101	0.8310	0.9460	0.9389	0.8885	0.8817
Difference	0.091(12.30%)	-0.0090(0.94%)	-0.0037(0.39%)	0.0428(5.06%)	0.0526(6.34%)

4.4.2 ***Convergence of Iterations.*** The convergence of our model's iteration is shown in Fig. 13. When the number of iterations exceeds 20,000, the model tends to stable convergence, the loss value in the test process is less than 0.05, and the final accuracy rate is stable at 0.989. Besides, it can be



(a) Accuracy vs iterations (b) Loss vs iterations

Fig. 13. Convergence of iterations

seen from the accuracy-iterations and loss-iterations curves that there are no loss oscillation, non-convergence and gradient explosion in the model. In terms of the number of iterations, although the number of initial iterations is set to be 80,000, the model can reach a stable convergence state at 20,000. This clearly indicates that our model has a high robustness.

In real-world applications, the shooting frame rate of WCE is 2 frames per second. That is, it takes 0.5s to shoot a WCE image. In our experiments, when the number of candidate region proposals is set to 300 for faster turnaround time, the average detection processing time is 0.07s per image. When we set the proposals' number to 2000 for performance, the average time is 0.29s per image. Also, the approach reaches its best performance. The processing time includes pre-processing steps. The size of a WCE image is not large, pre-processing steps like the wavelet transformation consume little time and effort. Even under the best performance setting, the processing time of our approach is just nearly half of the time of shooting an WCE image. This allows our approach to better fit into real-world practices.

5 CONCLUSION AND FUTURE WORK

This paper presented a multi-scale feature extraction and fusion-based method for detecting bleeding lesions in WCE images. Our proposed model consists of a backbone network, multi-scale region proposal branch network, top-down feature fusion branch network, and FC-layer to classify and localize the bleeding lesions. The experimental results on a clinical dataset with 20,000 WCE images demonstrate that our proposed model can improve up to 31.69% and 22.12% in sensitivity and overall performance, which outperforms existing bleeding detecting models. Our model can also achieve the highest AP and get the best medical diagnosis performance than state-of-the-art multi-scale models. Besides, comparing the backbone network and convergence of iterations also thoroughly verifies that our model has favourable feature learning and generalization ability.

In our future research, we will focus on simplifying the architecture of the model without losing the ability of feature fusion, and adapting to different target detection. On this basis, the model can be applied to other diseases of the GI tract, such as early detection of gastric cancer in WCE images, or early warning of biliary and pancreatic diseases in endoscopic retrograde cholangiopancreatography (ERCP) images.

ACKNOWLEDGMENTS

This work is supported by Hubei Provincial Development and Reform Commission Program "Hubei Big Data Analysis Platform and Intelligent Service Project for Medical and Health". The authors would also like to thanks for the long-term cooperation and support of Ankon Technologies (Wuhan) Co., Ltd.

REFERENCES

[1] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*. Springer, 354–370.

[2] Eli Chen, Oren Haik, and Yitzhak Yitzhaky. 2021. Online Spatio-Temporal Action Detection in Long-Distance Imaging Affected by the Atmosphere. *IEEE Access* 9 (2021), 24531–24545.

[3] Gastone Ciuti, Arianna Menciassi, and Paolo Dario. 2011. Capsule endoscopy: From current achievements to open challenges. *IEEE Reviews in Biomedical Engineering* 4 (2011), 59–72.

[4] Lei Cui, Chao Hu, Yuexian Zou, and Max Q-H Meng. 2010. Bleeding detection in wireless capsule endoscopy images by support vector classifier. In *Proceedings of the 2010 IEEE International Conference on Information and Automation*. IEEE, 1746–1751.

[5] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2010), 1627–1645.

[7] Yanan Fu, Mrinal Mandal, and Gencheng Guo. 2011. Bleeding region detection in WCE images based on color features and neural network. In *Proceedings of 2011 IEEE International Midwest Symposium on Circuits and Systems*. IEEE, 1–4.

[8] Tonmoy Ghosh, SK Bashar, Shaikh Anowarul Fattah, Celia Shahnaz, and Khan A Wahid. 2014. A feature extraction scheme from region of interest of wireless capsule endoscopy images for automatic bleeding detection. In *Proceedings of 2014 IEEE International Symposium on Signal Processing and Information Technology*. IEEE, 000256–000260.

[9] Tonmoy Ghosh, Syed Khairul Bashar, Md Samiul Alam, Khan Wahid, and Shaikh Anowarul Fattah. 2014. A statistical feature based novel method to detect bleeding in wireless capsule endoscopy images. In *Proceedings of 2014 International Conference on Informatics, Electronics & Vision*. IEEE, 1–4.

[10] T Ghosh, SA Fattah, C Shahnaz, AK Kundu, and MN Rizve. 2015. Block based histogram feature extraction method for bleeding detection in wireless capsule endoscopy. In *Proceedings of 2015 IEEE Region 10 Conference*. IEEE, 1–4.

[11] Chaoxu Guo, Bin Fan, Qian Zhang, Shiming Xiang, and Chunhong Pan. 2020. Augfpn: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12595–12604.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[13] National health commission of the People's Republic of China. 2017. Statistical Communique of National Health Commission. http://www.nhc.gov.cn/guihuaxxs/s10748/201708/d82fa7141696407abb4ef764f3edf095.shtml.

[14] Peiyun Hu and Deva Ramanan. 2017. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 951–959.

[15] Dimitris K Iakovidis and Anastasios Koulaouzidis. 2015. Software for enhanced video capsule endoscopy: Challenges for essential progress. *Nature Reviews Gastroenterology & Hepatology* 12, 3 (2015), 172–186.

[16] Xiao Jia and Max Q-H Meng. 2016. A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. In *Proceedings of 2016 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 639–642.

[17] Xiao Jia and Max Q-H Meng. 2017. Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and CNN features. In *Proceedings of 2017 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3154–3157.

[18] Hongyang Li, Yu Liu, Wanli Ouyang, and Xiaogang Wang. 2017. Zoom out-and-in network with recursive training for object proposal. *arXiv preprint arXiv:1702.05711* (2017).

[19] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual Generative Adversarial Networks for Small Object Detection. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* (2017), 1951–1959.

[20] Yixiong Liang, Zhihong Tang, Meng Yan, and Jianfeng Liu. 2018. Object detection based on deep learning for urine sediment examination. *Biocybernetics and Biomedical Engineering* 38, 3 (2018), 661–670.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2117–2125.

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *Proceedings of European Conference on Computer Vision*. Springer, 21–37.

[23] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks.. In *Proceedings of The International Conference on Machine Learning*. PMLR, 507–516.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3431–3440.

[25] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1520–1528.

[26] Chee Khun Poh, That Mon Htwe, Liyuan Li, Weijia Shen, Jiang Liu, Joo Hwee Lim, Kap Luk Chan, and Ping Chun Tan. 2010. Multi-level local feature classification for bleeding detection in wireless capsule endoscopy images. In *Proceedings of 2010 IEEE Conference on Cybernetics and Intelligent Systems*. IEEE, 76–81.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*. Curran Associates, Inc., 91–99.

[28] M. Satyanarayanan. 2017. The Emergence of Edge Computing. *Computer* 50, 1 (Jan 2017), 30–39. https://doi.org/10.1109/MC.2017.9

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[30] Chang Tang, Xinzhong Zhu, Xinwang Liu, Lizhe Wang, and Albert Zomaya. 2019. Defusionnet: Defocus blur detection via recurrently fusing and refining multi-scale deep features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2700–2709.

[31] Yixuan Yuan, Baopu Li, and Max Q-H Meng. 2016. Bleeding frame and region detection in the wireless capsule endoscopy video. *IEEE Journal of Biomedical and Health Informatics* 20, 2 (2016), 624–630.

[32] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. 2018. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4203–4212.

[33] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. 2019. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9259–9266.

[34] Rongsheng Zhu, Rong Zhang, and Dixiu Xue. 2015. Lesion detection of endoscopy images based on convolutional neural network features. In *Proceedings of 2015 International Congress on Image and Signal Processing*. IEEE, 372–376.

[35] Yuexian Zou, Lei Li, Yi Wang, Jiasheng Yu, Yi Li, and WJ Deng. 2015. Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network. In *Proceedings of 2015 IEEE International Conference on Digital Signal Processing*. IEEE, 1274–1278.