IEEE *Access*

Multidisciplinary : Rapid Review : Open Access Journal

Deep Convolutional Neural Networks for WCE Abnormality Detection: CNN Architecture, Region Proposal and Transfer Learning

LIBIN LAN[ID]**[1], CHUNXIAO YE[1], CHENGLIANG WANG[2], AND SHANGBO ZHOU**[ID]**[2]**
[1]Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400030, China
[2]College of Computer Science, Chongqing University, Chongqing 400030, China

Corresponding author: Chunxiao Ye (yecx@cqu.edu.cn)

ABSTRACT Wireless capsule endoscopy (WCE) plays an important role in the diagnosis of gastrointestinal diseases. However, it is very time-consuming and fatiguing for a physician to review a large number of WCE images. Some methods to address this problem have recently been presented. However, these methods generally employ classification algorithms to discriminate abnormal from normal images, which do not localize, recognize, or detect abnormal patterns in abnormal images. We sought to identify a better method for the WCE abnormal pattern detection. In this paper, convolutional neural networks (CNNs) are used to implement detection function, and several methods are also adopted to boost the performance of WCE abnormality detection from aspects of the CNN architecture, region proposal, and transfer learning. First, we present a deep cascade network, namely, CascadeProposal, trained end-to-end to generate a small number of region proposals with high-recall by a region proposal rejection module and to simultaneously detect abnormal patterns using a detection module. Second, we use a multiregional combination (MRC) method to obtain good coverage of the regions of interest and employ the salient region segmentation (SRS) method to capture accurate region locations. Third, we use the dense region fusion (DRF) method for object boundary refinement. Fourth, we introduce negative category (Neg) and transfer learning (TL) strategies into our CNNs to obtain a better model performance. The extensive experiments are performed on our WCE image dataset of more than 7k annotated images. A final mean average precision (mAP) of 70.3% and a better mAP of 72.3% can be achieved via CascadeProposal with ZF and Fast R-CNN with VGG-16 networks, respectively, using MRC+Neg+TL method in the training stage and MRC+DRF+SRS method in the testing stage. The comprehensive results demonstrate that our method is efficient and effective for WCE abnormality detection with high-localization accuracy.

INDEX TERMS Convolutional neural networks, medical image analysis, region proposal, transfer learning, wireless capsule endoscopy, WCE abnormality detection.

I. INTRODUCTION

Wireless capsule endoscopy (WCE) was introduced in 2000 by Given Imaging Incorporation [1]. They [1] reported the development of a new type of endoscopy, namely, wireless capsule endoscopy, which for the first time allows painless endoscopic imaging of the entire small bowel. Compared with traditional diagnostic procedures

(gastroscopy, small-bowel endoscopy, and colonoscopy, respectively), the endoscopy capsule is small enough to be swallowed (11 × 30 mm) and has no external wires, fiber-optic bundles or cables, thus causing no discomfort during the internal gastrointestinal examination. Additionally, the capsule endoscope is propelled by peristalsis through the gastrointestinal tract and does not require a pushing force to propel it through the bowel. These advantages make WCE a promising diagnostic tool for gastrointestinal (GI) diseases [1]–[4]. However, capsule endoscopy produces

---

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.
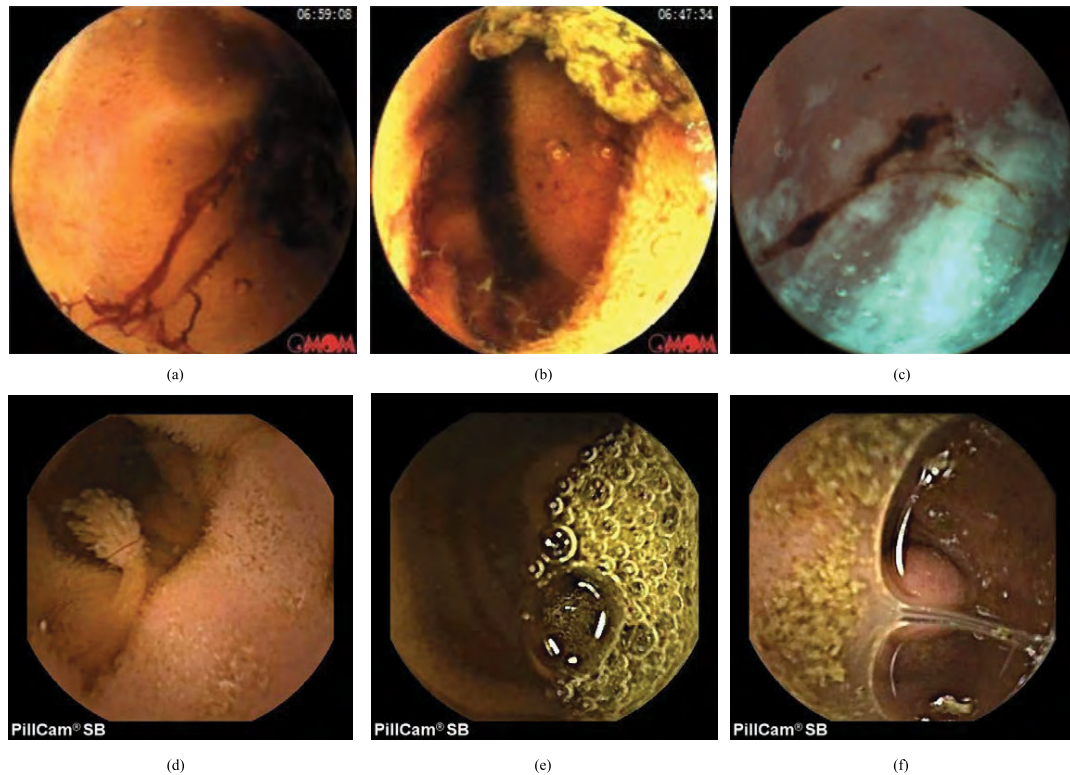
FIGURE 1. Images captured by WCE and examples of abnormal patterns. (a) Active bleeding. (b) Active bleeding and undigested residue. (c) Inactive bleeding. (d) Polyp. (e) Bubbles. (f) Tumor.

approximately 50k images for a patient during one examination, which makes it time-consuming and unacceptable for an experienced physician to review all these images to identify the abnormalities in the GI tract [5]–[9]. Furthermore, localizing and recognizing abnormal patterns in abnormal images is cumbersome, as abnormal images usually account for less than 5% of the entire set [5]. Therefore, it is extremely difficult for a physician to detect abnormal patterns in so many WCE images. An abnormal image generally includes one pattern, such as bleeding, polyp, or tumor. Occasionally, other abnormal patterns, such as bubbles or undigested residue, can be simultaneously captured by capsule endoscopy. Each pattern is shown in Fig. 1. Some reported methods differentiate abnormal images from normal images only for one specific abnormal pattern, e.g., [6] only for bleeding, [7] and [9] only for polyp, [8] only for ulcers, or [10] only for tumors, etc. Others detect multiple abnormal patterns, e.g., [5] for bleeding, polyp, and ulcers, or [11] for polyp and ulcers. Whether the above methods address the detection of single or multiple pattern abnormalities, these detection tasks substantially adopt classification algorithms to discriminate abnormal images from normal images, which do not localize or recognize abnormal patterns in abnormal WCE images. This motivated us to develop a generic detection solution to localize and recognize abnormal patterns in WCE via suitable methods.

In this paper, we introduce an effective scheme integrating various methods to accurately detect multiple abnormal patterns. These methods include CNN architecture design, the use of top region proposal and dense region fusion methods, and the introduction of transfer learning and negative category methods, etc. The several above-mentioned methods also reflect our main contributions. The motivation behind this proposal is to effectively localize and recognize abnormal objects. First, we redesign the network architecture based on Fast R-CNN [12] to reach a higher recall and improve the generalization capacity. We call the modified network CascadeProposal. Second, for the region locations obtained by the final trained CNN model, we use the dense region fusion method to improve the detection and localization accuracy. Third, we introduce negative category strategy into our model training to obtain better model performance. Furthermore, we use the model-based transfer learning strategy to fine-tune the network and yield the best performance results. These methods have been trained and tested on our WCE image dataset with more than 7k annotated images, achieving desirable results. It is noteworthy that the two concepts of abnormal patterns and abnormal objects are equivalent in this paper and that region (object) proposal and region (object) proposal generation are also equivalent. We refer to the abnormal pattern as the object.

After acquiring WCE images by capsule endoscopy, it can be seen that image characteristics, such as texture, color, and shape, are different from natural scene images, e.g., images from the Pascal VOC [13] and ImageNet [14] datasets. We observed that the shape of abnormal patterns in WCE

images is generally petechial, zonal, and blocky. Specifically, the bleeding abnormal pattern is amorphous, similar to water. In addition, other features (e.g., color and texture) are also rich and distinctive. Moreover, there is generally one abnormal pattern (e.g., either bleeding or polyp) in each WCE image. It is also possible that multiple abnormal patterns exist in one WCE image (e.g., bleeding and undigested residue). Inspired by the above observations, we aim to generate high-quality object locations by adopting both current top-performing region proposal methods, such as Selective Search (SS) [15], [16] and EdgeBoxes (EB) [17], [18], and salient region segmentation methods, such as region contrast (RC) [19] and Otsu [20], which is also our other contribution.

The main research questions in this paper can be summarized by the following three aspects: 1) For WCE images, how can more accurate region localization of abnormal patterns be generated? 2) How can CNN architecture be used to obtain higher recall and better generalization capacity? 3) Can we find other methods to improve the detection accuracy?

The rest of this paper is organized as follows. Section II introduces related work and discusses their differences with our proposed method. Section III describes our method, including the network architecture, region proposal generation, and implementation details. The experimental results and analysis are illustrated and discussed in Section IV. In Section V, we compare and discuss the performance of various trained models and elaborate on the major findings and significance of our work. Finally, conclusions are presented in Section VI.

II. RELATED WORK

This section reviews and discusses the related: (A) traditional methods applied to WCE image detection; (B) deep learning applied to medical image analysis; and (C) the current top region proposal methods.

A. TRADITIONAL METHODS APPLIED TO WCE IMAGE DETECTION

In this paper, we consider detection or classification methods, which are different from deep learning [21], as traditional methods. Many methods have been reported for the recognition and detection of WCE abnormal patterns. In [7], a texture feature extraction method is presented to identify polyp WCE images. Experiments on a small dataset verified the effectiveness of this method. Yuan et al. [8] use locality-constrained linear coding, superpixel, and saliency map methods, etc. to detect ulcer images. The results achieved promising accuracy. In [5], a coding method called saliency and adaptive LLC is proposed to detect multiple abnormal images such as bleeding, polyp, and ulcers. Yuan et al. [6] propose a words-based color histogram method for bleeding detection and a two-stage saliency map extraction method to localize bleeding regions. Yuan et al. [22] present a novel discriminative joint-feature topic model with dual constraints to classify multiple abnormalities in WCE images. These above-mentioned methods are in fact classification methods.

Moreover, Karargyris and Bourbakis [11] employ a synergistic scheme to address the detection of polyp and ulcers. But the method fails to increase accuracy due to the irregularity of abnormal patterns (e.g., round or elongated). Also, other studies have already been undertaken on automatic abnormal image detection in WCE videos [23]–[30]. For more details, readers can refer to the above-listed literatures. However, as previously stated, these methods are essentially classification or segmentation methods, without localizing or recognizing abnormal patterns.

Therefore, in this paper, our main aim is to localize, recognize and detect abnormal patterns. Since the traditional methods of WCE abnormal pattern detection are essentially classification methods and do not implement detection function (i.e., the goal of object detection [17], [31] is not only to determine whether an object exists in an image, but if so, where in the image does it occur and what is the object). These methods can be used prior to our method to directly discriminate abnormal frames from the whole WCE video. Our method can then be used to detect the WCE abnormal pattern, which can speed up the detection of WCE abnormal frames.

B. DEEP LEARNING APPLIED TO MEDICAL IMAGE ANALYSIS

In recent research, deep learning has been employed for medical image processing and analysis and achieved remarkable results. Some successful application fields include classification of skin cancer [32] using transfer learning [33], mitosis detection in breast histology image [34], identification of cancer metastases [35], thoraco-abdominal lymph node detection and interstitial lung disease classification [36], lung pattern classification [37], infant brain image segmentation [38], polyp detection [39], skin lesion segmentation [40], cancer prediction [41], WCE abnormal image classification [42], a baseline dataset for gastrointestinal (GI) disease detection [43], and brain MRI segmentation [76], [77], etc. These fields mainly focus on medical image classification, localization, and segmentation.

In short, deep learning has made much success in medical image analysis [44]–[54]. Also, Ker et al. [55] review the recent success of applying deep learning to medical image analysis. But to the best of our knowledge, there is little work focusing on WCE abnormal pattern detection. Additionally, since training a deep CNN from scratch is difficult for a small number of medical images, transfer learning is typically used to fine-tune the CNN to ensure a quick convergence and performance improvement [32], [39].

Therefore, in this paper, our scheme is apt to utilize the top detection algorithms, such as ZF [56]-and VGG-16 [57]-based CNN architectures to detect the WCE abnormal patterns, and simultaneously adopt transfer learning strategy to fine-tune our CNN network to ensure a quick convergence. Training is performed on our WCE image dataset using various methods for abnormal pattern detection, and promising results are achieved.

C. THE CURRENT TOP REGION PROPOSAL METHODS

Currently, there is extensive literature on region proposal methods. Comprehensive surveys and comparisons of region proposal methods can be found in [58] and [59]. According to the analyses in [58] and [59] and the self-characteristics and experimental results of WCE images, we integrate Selective Search, EdgeBoxes, and Objectness [60], [61] to generate region proposals, which is called a multiregional combination (MRC) method. Furthermore, we study the possibility of using a salient region segmentation method to generate region proposals. For a more detailed comparisons and analysis of salient region detection and segmentation methods, we refer readers to [62] and [63]. According to the analyses in [62] and [63] and our experiment results, and considering good trade-off between the computational efficiency and algorithm performance, we employ RC [19] salient region detection and SaliencyCut [19] segmentation and Otsu [20] to generate region proposals.

Selective Search (SS) uses hierarchical grouping strategies and various color spaces with different invariance properties and similarity measures such as color, texture, size, and fit, which makes it stable, robust, and class-independent for determining object location, where object types range from rigid to non-rigid and even include amorphous objects. EdgeBoxes (EB) relies on one simple observation: the number of contours that are wholly enclosed by a bounding box is indicative of the likelihood of the box containing an object. Objectness (OB) uses four image cues to generate a closed boundary and thereby predicts any potential object location. The four image cues are multiscale saliency, color contrast, edge density, and superpixel straddling. In this paper, we abbreviate the multiregional combination methods including SS, EB, and OB and the salient region segmentation method as MRC and SRS, respectively.

As pointed out by Hosang *et al.* [59], for object (WCE abnormal pattern) detection, improving region localization accuracy is as important as improving recall. So, one of our main aims is to achieve more accurate region localization via various methods. These methods are listed in Subsection III-C. Inspired by [64]–[67], we use an analogous method to re-rank the region proposals generated by MRC method.

III. METHODS
A. OVERVIEW OF OUR PROPOSED APPROACH
In this paper, a novel WCE abnormal pattern detection scheme integrating various methods is proposed, as depicted in Fig. 2. We elaborate on how the detection framework works separately from the training and testing stages.

1) TRAINING STAGE
The training pipeline is shown at the top of Fig. 2. Initially, we utilize the region proposals generated by the MRC method to train the RPR module of the whole CascadeProposal network. Next, the re-ranked region proposals yielded by RPR are forwarded to the Detection module of the Cascade-Proposal network, and then, the trained Detection model is produced. Finally, according to Algorithm 1, the two trained
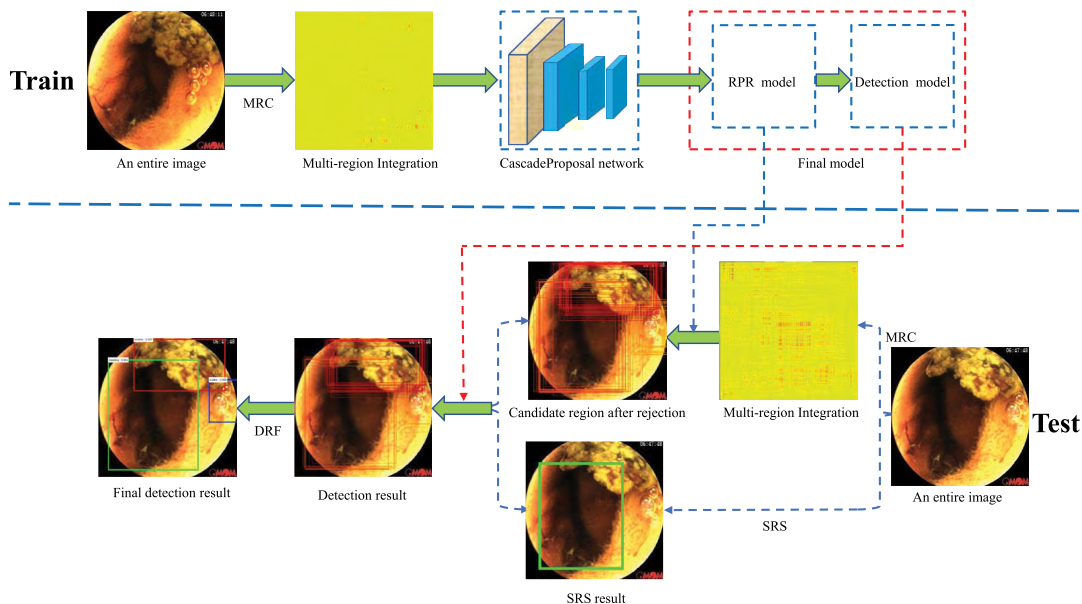

FIGURE 2. Overview of our approach. This workflow includes training and testing stages. During the training stage, SS, OB, OE and their combination are used to generate category-independent region proposals, and then, the RPR module is trained by using these region proposals, and the detection module is trained by using the re-ranked region proposals generated via the RPR module. During the testing stage, region proposals generated by MRC are used to test via both the RPR and detection models. Salient region proposals via SRS are input into the Detection model together. Region proposals yielded by the detection module, whose overlap ratio is higher than 0.9 between them, will be fused into a final region proposal.

Algorithm 1 CascadeProposal Network Joint Training Process. The RPR and Detection Modules Will Form a Unified Network After All Steps

step 1. Learned from scratch or fine-tuned for initializing all layers in Steps 2 and 3.
step 2. Train the RPR module from scratch or fine-tune the module via the ImageNet pretrained model to generate re-ranked region proposals.
step 3. Train the Detection module from scratch or fine-tune the module via the ImageNet pretrained model using region proposals from Step 2.
step 4. Initialize by Step 3 and fine-tune unique layers of the RPR module for region proposals sharing Conv1 to Conv5 layer feature trained in Step 3.
step 5. Fine-tune the Detection module for abnormal pattern detection using region proposals from Step 4, keeping the shared convolutional layers fixed.
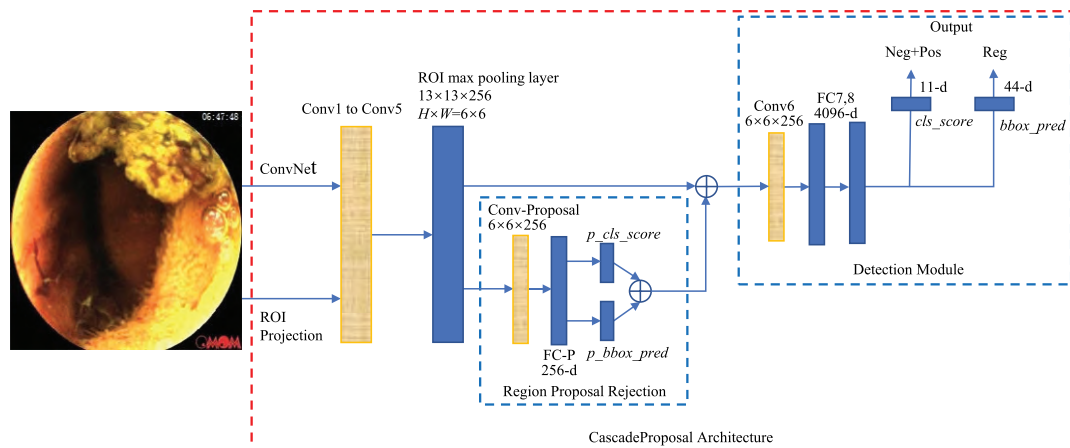step 6. Output the unified network jointly trained in Steps 4 and 5 as the final model.

FIGURE 3. Our detection network architecture. Following Fast R-CNN, an input image and a set of regions of interest (ROIs) are input into a fully convolutional network. This architecture of Fast R-CNN can be extended to reject region proposals that have a higher IoU than a given threshold by including the region proposal rejection module. This module has two sibling output layers to yield softmax probability estimates over 10 object classes (5 positive categories and 5 negative categories) plus a catch-all background class and per-class bounding-box positions. The final detection network, similar to RPR, has also two sibling output vectors per ROI: softmax probabilities over 10 object classes and per-class bounding-box regression offsets.

models, i.e., RPR and Detection models, form a unified network, i.e., CascadeProposal network which is used to test images. The steps of the training pipeline are summarized as follows.

1) MRC is used to generate the region proposals on the trainval set. Details are given in Subsection III-C. MRC denotes a combination of SS, EB, and OB.
2) The RPR module is trained via these region proposals generated by MRC, and then, the re-ranked region proposals are obtained through implementing greedy non-maximum suppression (NMS) [68] on the regions generated by RPR based on their scores, aiming to achieve a high recall.
3) The Detection module of the CascadeProposal network is trained via the re-rank region proposals.
4) A unified network, i.e., CascadeProposal network, is finally produced through combining the RPR and Detection modules. The procedure of joint training is given in Subsection III-D.
5) The negative category and the model-based transfer learning strategy are utilized for the training process of the CascadeProposal network. Details about the

negative category and transfer learning are given in Subsections III-D and III-E, respectively.

The WCE abnormal pattern detection scheme is composed of two modules: RPR and Detection modules. We unify the two modules into a new network CascadeProposal and train the entire network with features shared through a developed Algorithm 1, as shown in Subsection III-D, and we compare the performance of each trained CNN model under the above-mentioned several methods, including being fine-tuned by transfer learning or trained from scratch. The results are given in Subsection IV-D.

2) TESTING STAGE

The test pipeline is shown at the bottom of Fig. 3. As in the training stage, we first perform MRC on the test set to generate the region proposals. Next, these region proposals are fed into the RPR model to generate candidate region proposals after rejection. The salient region segmentation method is also performed on the test set to generate salient region proposals. Finally, the two region proposals are input into the Detection model to output a discrete probability

distribution over each category, i.e., the accuracy of each category via a score and a bounding-box prediction. The steps of the test pipeline are described below.

1) Multiregional proposals are generated by MRC.
2) Candidate region proposals are obtained via the RPR model, and simultaneously, salient region proposals via SRS are combined and input together into the Detection model.
3) The dense region fusion method is performed on the detection results generated by step 2 to improve the detection performance and localization accuracy.
4) The final detection results are output, including the bounding-box and the score of each category.

B. CASCADEPROPOSAL ARCHITECTURE

Following Fast R-CNN [12], Faster R-CNN [64], Hypernet [65], DeepProposal [66] and Deepbox [67], the CascadeProposal architecture is shown in Fig. 3. We take the ZF network as a baseline. The final classification layer is replaced with the two sibling layers, a box-classification layer (*cls_score*) and a box-regression layer (*bbox_pred*), and is retrained on our WCE image dataset. Our network has two modules: region proposal rejection (RPR) and Detection modules. Each module has two sibling output layers, similar to Fast R-CNN. The RPR is a convolutional network that generates re-ranked region proposals by NMS. The Detection module is a softmax classifier and a bounding-box regressor, which output per-class label and bounding-box regression offsets. As in Fast R-CNN, the ROI max pooling layer uses max pooling to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent of $H \times W$, which is configured by setting $H = 6$ and $W = 6$, where H and W are layer hyper-parameters that are independent of any particular ROI. ROI max pooling works by dividing the $h \times w$ ROI window into an $H \times W$ grid of sub-windows of approximate size $h/H \times w/W$ and then max-pooling the values in each sub-window into the corresponding output grid cell, where $h \times w = 13 \times 13$ in this network. Inspired by [65]–[67], our RPR module closely follows on the top of the ROI max pooling layer, which differs from Faster

R-CNN. The RPN module of Faster R-CNN is added to the bottom of the ROI max pooling layer. The object proposals (i.e., ROIs) generated by the RPR module are then input into the Detection module. We call the redesigned Fast R-CNN network CascadeProposal because the network consists of two analogous cascading deep convolutional networks based on ZF to generate region proposals.

Similar to Fast R-CNN, the CascadeProposal network takes two data inputs, a list of images and a set of ROIs in these images, and simultaneously has two sibling outputs, probability estimates over objection classes and bounding box regression offsets. The entire image is processed with five convolutional layers and two max pooling layers to produce a convolutional feature map. Then, the ROI max pooling layer extracts a fixed-length feature vector from each feature map.

Both the RPR and Detection network configurations evaluated in this paper are outlined in Table 1 and Table 2. All configurations follow the generic design presented in Fig. 3. Conv1 to Conv5 layers are shareable between the RPR and Detection networks.

C. REGION PROPOSAL GENERATION

In this paper, several methods, including MRC, SRS, DRF, and RPR, etc., are used to generate region proposals. We give a more detailed description on how each method outputs region proposals and on which part of the network the region proposals are fed into. The process of each method for generating the proposals is depicted below.

1) MULTIREGIONAL COMBINATION (MRC)

Multiregional combination integrating SS, EB, and OB methods are performed on all images to generate preliminary region proposals. Then, these proposals are loaded into the region proposal rejection (RPR) module, which yields re-ranked region proposals by NMS [68]. After NMS, we select the top-N ranked region proposals to train the Detection network based on their scores. For each image, MRC generates approximately 3k candidate boxes. Based on comprehensive surveys and comparisons of various object proposal methods provided in [58] and [59] and

TABLE 1. The RPR network with ZF configurations. Input and output parameters, filter size, stride, and number of weight parameters of each layer are listed in detail. "Input" and "Output" denote the number of filters (*d*) and the size of the feature map (*w* × *h*). *h* × *w* is the size of the ROI window, where *h* × *w* = 13 × 13 Conv1 to Conv5 layers are shareable convolutional layers with the detection network.

Layer	$Input_{d@w \times h}$	$Size_{filter}$	Stride	$Output_{d@w \times h}$	Num_{params}
Conv1	$3@224 \times 224$	7×7	2	$96@110 \times 110$	$96 \times (7 \times 7 \times 3 + 1)$
Pool1	$96@110 \times 110$	3×3	2	$96@55 \times 55$	0
Conv2	$96@55 \times 55$	5×5	2	$256@26 \times 26$	$256 \times (5 \times 5 \times 96 + 1)$
Pool2	$256@26 \times 26$	3×3	2	$256@13 \times 13$	0
Conv3	$256@13 \times 13$	3×3	1	$384@13 \times 13$	$384 \times (3 \times 3 \times 256 + 1)$
Conv4	$384@13 \times 13$	3×3	1	$384@13 \times 13$	$384 \times (3 \times 3 \times 384 + 1)$
Conv5	$384@13 \times 13$	3×3	1	$256@13 \times 13$	$256 \times (3 \times 3 \times 384 + 1)$
ROI	$256@13 \times 13$	$(h/6 + 1) \times (w/6 + 1)$	$(h/6) \times (w/6)$	$256@6 \times 6$	0
Conv-P	$256@6 \times 6$	3×3	1	$256@6 \times 6$	$256 \times (3 \times 3 \times 256 + 1)$
FC-P	$9216@1 \times 1$	1×1	1	$256@1 \times 1$	$256 \times (9216 + 1)$
p_cls_score	$256@1 \times 1$	1×1	1	$11@1 \times 1$	$11 \times (256 + 1)$
p_bbox_pred	$256@1 \times 1$	1×1	1	$44@1 \times 1$	$44 \times (256 + 1)$

TABLE 2. The Detection network with ZF configurations. Input and output parameters, filter size, stride, and number of weight parameters of each layer are listed in detail. "Input" and "Output" denote the number of filters (d) and the size of the feature map ($w \times h$). $h \times w$ is the size of the ROI window, where $h \times w = 13 \times 13$ Conv1 to Conv5 layers are shareable convolutional layers with the RPR network.

Layer	$Input_{d@w \times h}$	$Size_{filter}$	Stride	$Output_{d@w \times h}$	Num_{params}
Conv1	$3@224 \times 224$	7×7	2	$96@110 \times 110$	$96 \times (7 \times 7 \times 3 + 1)$
Pool1	$96@110 \times 110$	3×3	2	$96@55 \times 55$	0
Conv2	$96@55 \times 55$	5×5	2	$256@26 \times 26$	$256 \times (5 \times 5 \times 96 + 1)$
Pool2	$256@26 \times 26$	3×3	2	$256@13 \times 13$	0
Conv3	$256@13 \times 13$	3×3	1	$384@13 \times 13$	$384 \times (3 \times 3 \times 256 + 1)$
Conv4	$384@13 \times 13$	3×3	1	$384@13 \times 13$	$384 \times (3 \times 3 \times 384 + 1)$
Conv5	$384@13 \times 13$	3×3	1	$256@13 \times 13$	$256 \times (3 \times 3 \times 384 + 1)$
ROI	$256@13 \times 13$	$(h/6+1) \times (w/6+1)$	$(h/6) \times (w/6)$	$256@6 \times 6$	0
Conv6	$256@6 \times 6$	3×3	1	$256@6 \times 6$	$256 \times (3 \times 3 \times 256 + 1)$
FC7	$9216@1 \times 1$	1×1	1	$4096@1 \times 1$	$4096 \times (9216 + 1)$
FC8	$4096@1 \times 1$	1×1	1	$4096@1 \times 1$	$4096 \times (4096 + 1)$
cls_score	$4096@1 \times 1$	1×1	1	$11@1 \times 1$	$11 \times (4096 + 1)$
$bbox\_pred$	$4096@1 \times 1$	1×1	1	$44@1 \times 1$	$44 \times (4096 + 1)$

the characteristics of the WCE image, MRC may be more suitable for non-rigid and amorphous object detection. The performance of each method is compared by experiments, as shown in Subsection IV-B.

2) REGION PROPOSAL REJECTION (RPR)

The RPR network includes a convolutional layer, a fully connected layer and two output layers. The RPR is added on top of the ROI max pooling layer, which differs from the RPN of Faster R-CNN added on the bottom of the ROI layer. This module is a convolutional network and has two sibling output layers for each candidate region (ROI) generated by MRC. One is the p_cls_score layer, which yields a softmax probability estimate over each object. The other is the p_bbox_pred layer, which outputs bounding-box offsets using four real-valued numbers. The RPR re-ranks region proposals by implementing greedy non-maximum suppression (NMS [68]) based on their scores.

3) SALIENT REGION SEGMENTATION (SRS)

The motivation for using SRS is the same as in [69] and [70] which comes from the fact that segmentation-related cues are empirically known to often aid in object detection. However, different from semantic segmentation-aware in [69] and [70], which uses CNN features to yield the segmentation target image, we directly adopt RC and Otsu to generate object proposals on the auxiliary object detection task. The object proposals generated by SRS are forwarded to the Detection model in the testing stage. The experimental results show that using SRS can improve the detection and localization accuracy.

4) DETECTION NETWORK

Following [64]–[67], the Detection network includes a convolutional layer, two fully connected layers and two output layers. These region proposals generated by SRS and RPR, are fed into the Detection network. For the two output layers, the cls_score layer outputs 11 scores, and the $bbox\_pred$ layer yields 44 bounding-box regressions. When a score threshold is given, region proposals yielded by the Detection network, whose overlap ratio is higher than a fixed value, will be fused into a final proposal. Dropout is used in the fully connected layers 7 and 8 with a rate of 0.5.

5) DENSE REGION FUSION (DRF)

Dense regions are some similar candidate region proposals around object region, with a high IoU overlap ratio and a short distance between the region proposals' center coordinates, generated by the Detection module. Whether some regions are dense or not will be determined by (1), which calculates the distance between the center coordinates of two regions and the IoU of two regions, where R denotes ROIs (i.e., region proposals). The dense regions are fused by calculating the mean values of two regions' center coordinates, their width and height, and their scores.

$$F(R_1, R_2) = \begin{cases} 0, & Dis(R_1, R_2) > \alpha \, \& \, IoU(R_1, R_2) < \beta \\ 1, & Dis(R_1, R_2) < \alpha \, \& \, IoU(R_1, R_2) > \beta \end{cases}$$
(1)

D. TRAINING AND LOSS FUNCTION

Our main goal is to design an end-to-end network that includes both region proposal rejection (RPR) and Detection modules and then to optimize the algorithm with back propagation. The main steps of training can be depicted as follows. First, the RPR module yields re-ranked region proposals. Second, the re-ranked proposals are used to train the Detection network. Finally, the network outputs the final trained model. It is noteworthy that the network can be either learned from scratch or fine-tuned from the ImageNet pre-trained model. Following [64]–[67], we design a 6-step training process as shown in Algorithm 1. Similar to Faster R-CNN [64], for step 2 and step 3, the RPR and Detection modules are trained independently and do not share convolutional layers. After fine-tuning of step 4 and step 5, both networks share the same convolutional layers and form a unified network.

We observed that for some outstanding object detection algorithm such as R-CNN [68], [71], SPP-net [72], Fast R-CNN [12] and Faster R-CNN [64], directly setting all

regions with an IoU lower than 0.5 to the background category is not reasonable. Therefore, we set these regions with an IoU higher than 0.5 to the positive category, the regions with an IoU overlap with ground truth region in the interval of [0.2, 0.5) to the negative category, and other regions with an IoU lower than 0.2 to the background. We assign a positive label to positive and negative categories and a negative label to the other regions, respectively, to train our redesigned network. That is the main idea of a negative category strategy introduced into our CNNs training. We use the multi-task loss function defined in [12] to optimize the softmax classifier and bounding-box regression, as shown in (2).

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{reg}(t^u, v) \qquad (2)$$

where p is a discrete probability distribution per ROI. $p = (p_0, \ldots, p_k, \ldots, p_{2K})$, over $2K+1$ categories. K denotes the object classes. For parameter t^u and v, readers can refer to [68]. As usual, the background classes are labeled with $u = 0$, and L_{reg} is ignored. For negative categories, it is meaningless to adjust the term L_{reg}. Therefore, we set $u = 0$ to ignore L_{reg}. The hyper-parameter $\lambda = 2$ is apt for better region localization. The above parameters are used in both the RPR and Detection modules.

During training, we use the ImageNet [14] pretrained model to initialize and fine-tune the parameters across all basic layers between both modules. This procedure is known as transfer learning [33]. In addition, we also train the network from scratch.

E. TRANSFER LEARNING
We utilize model-based transfer learning to train our whole network, i.e., initialize the network from the pre-trained model. In addition, we compare the performance of the trained models fine-tuned via the transfer learning strategy and learned from scratch. The CascadeProposal network can be either learned from scratch or fine-tuned from the ImageNet [14] pre-trained model. When using the ImageNet pretrained model to fine-tune the network, we set the learning rate to 0.001. When learned from scratch, all the parameters of the CNN models are initialized with random Gaussian distributions and trained for 40k iterations using stochastic gradient descent with a mini-batch of size 128, starting with a learning rate of 0.01. Training convergence can be observed within 40k iterations, and the caffemodel can be obtained. The other hyper-parameters are the same as those used in the off-the-shelf ImageNet pre-trained model, including a momentum of 0.9, a weight decay of 0.0005, and gamma of 0.1.

F. IMPLEMENTATION DETAILS
For the RPR module, each region proposal generated by RPR is scored and adjusted. The number of these proposals is the maximum for an image, and some proposals highly overlap with each other. To reduce redundancy, we perform greedy non-maximum suppression [68] on the proposals based on their scores. Concretely, we fix the IoU threshold for NMS

at 0.7, which leaves us approximately 800 proposals per image from approximately 3k proposals generated by MRC. According to the analysis in [64], NMS does not harm the ultimate detection accuracy but substantially reduces the number of proposals. After NMS, we use the top-200 ranked proposal regions to train the Detection module but evaluate different numbers during the testing stage. We use the evaluation method of Pascal VOC2007 [13].

For the object regions generated by the Detection module, we adopt a discriminant function of the dense region (cf. (1)) to fuse multiple regions into region. Concretely, the regions with a distance lower than 0.1 and an IoU higher than 0.9 between regions are fused into one region.

For the network architecture, we use the CascadeProposal with ZF [56], [64] and Fast R-CNN with VGG-16 [12], [57] to train our dataset. For CascadeProposal, the number of filters and the size of the feature map of the "Conv1 to Conv5" layers are the same as [12]. The parameters of the other unique layers of both the RPR and Detection modules are listed in Table 1 and Table 2. For Fast R-CNN, we follow [12] to train the network based on multiregional combination, negative category, and transfer learning.

In this paper, Selective Search, EdgeBoxes, and Objectness are used to generate region proposals using default parameter configurations, except the parameters used to fix the number of proposals. As in Fast R-CNN [12], we train and test both the RPR and Detection networks on images of a single scale of $s = 240$ pixels for ZF and $s = 600$ pixels for VGG-16. The other hyper-parameters for training and fine-tuning of both networks via a pretrained deep model are a base learning rate of 0.001, a momentum of 0.9, a weight decay of 0.0005, and gamma of 0.1. We also use data augmentation by flip and rotation. All images are horizontally flipped with a probability 0.5, and each image is rotated randomly between $0°$ and $359°$ during training.

IV. EXPERIMENTS AND RESULTS
A. WCE IMAGE DATASET
Following the Pascal VOC2007 detection task, our WCE image dataset contains a total of 7,381 annotated images, which come from a combination of raw images of patients provided by Chongqing Jinshan Science & Technology (Group) Co., Ltd. and the online open-access images atlas from Given Imaging Incorporation. The first dataset from Jinshan has 7,200 images, and the other dataset from Given has 181 images. The WCE images from Jinshan and Given are 256×240 pixels and 576×576 pixels, respectively. For this detection task, four sets of images are provided: train, test, trainval and val sets. This dataset is annotated under the guidance of clinicians. It consists of five object classes, including undigested residue, bleeding, bubbles, tumor and polyp. We failed to obtain WCE abnormal images of other types, such as ulcers and Crohn's disease. The data are randomly allocated to 50% for the trainval set and 50% for the test set. The train set and val set account for 25% of trainval
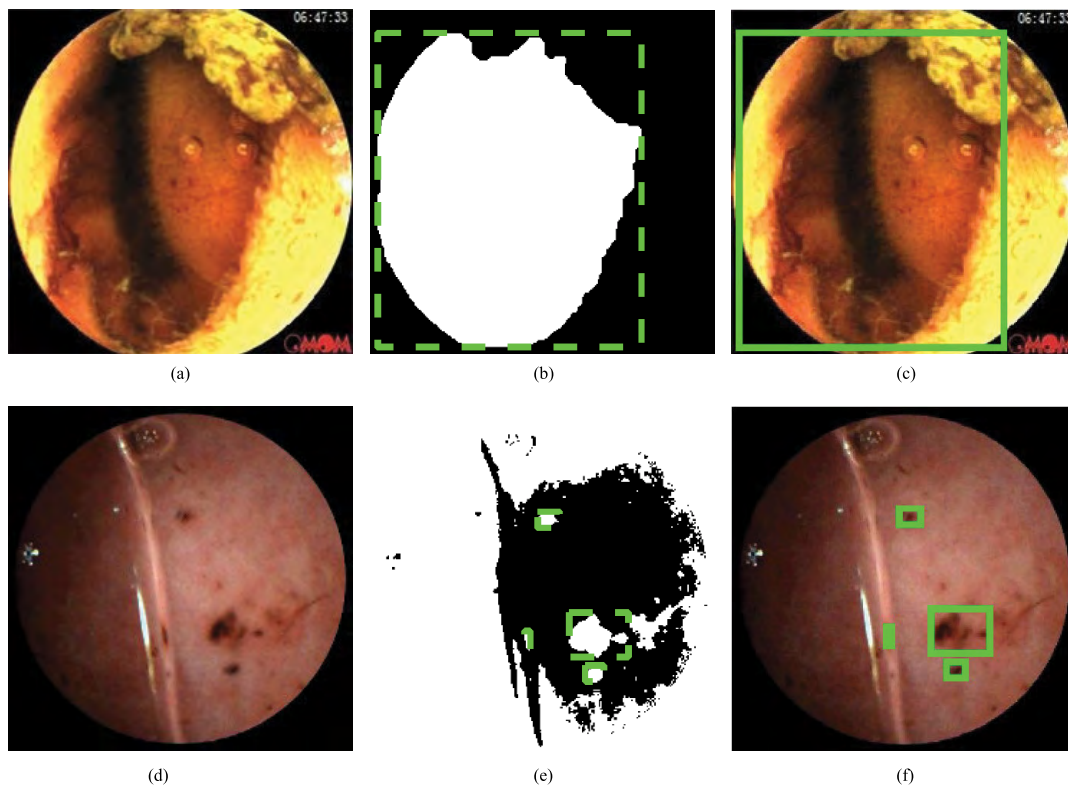
FIGURE 4. Illustrations of salient region segmentation-based region proposal generation. (a) and (d) Original image. (b) and (e) Binary maps yielded by RC and Otsu, respectively. The salient objects resulting from binary maps (b) and (e) are shown in (c) and (f).

set, respectively. The trainval set is the union of the train and val sets. To easily remember for use in this paper, we denote the WCE image dataset as WCE2017.

The basic criteria of image annotation are that the bounding box covers as much of the object (abnormal pattern) region as possible and that scattered homogeneous objects are separately annotated, including petechial, zonal, and blocky objects. We also refer to the Pascal VOC2007 annotation guidelines. Following the guidelines on categorization, bleeding includes active bleeding and inactive bleeding, which is shown in the left column of Fig. 4. Statistics on evaluating the detection task are shown in Table 3. Table 3 summarizes the number of objects (abnormal pattern) and images (containing at least one object of a given class) for each class and image set. In total, there are 7,381 images, containing 5 object categories, 3,690 trainval images, 3,691 test images, and 11,653 annotated objects. The distributions of images and objects by class are approximately equal across the trainval and test sets.

B. LOCALIZATION PERFORMANCE

As already explained in the above relevant section, we utilize Selective Search, EdgeBoxes, and Objectness i.e., multiregional combination (MRC) to obtain a relatively accurate object localization based on the non-rigid and amorphous characteristics of the WCE abnormal pattern. Additionally, salient region segmentation based on the saliency map

TABLE 3. Statistics of the main image sets. Following Pascal VOC2007, the statistics of abnormal patterns are used in the evaluation.

class	train		val		trainval		test	
	img	obj	img	obj	img	obj	img	obj
residue	225	244	274	291	499	535	494	534
bleeding	1801	1840	1797	1832	3598	3672	3603	3680
bubble	637	788	633	783	1270	1571	1263	1549
tumor	5	5	7	7	12	12	17	17
polyp	10	18	12	23	22	41	24	42
total	2678	2895	2723	2936	5401	5831	5401	5822

computed by region contrast (RC) [19] and SaliencyCut [19] segmentation methods and Otsu [20] can obtain a better object localization. The visual experimental results are shown in Fig. 4. Following [69], [70], we adopt RC-based SaliencyCut and Otsu methods to generate object proposals on the auxiliary task of WCE abnormality detection. In the whole detection system, the salient region is used to generate class-agnostic region proposals. For WCE abnormal pattern detection, our observation is that object proposals generated by SRS applied to WCE abnormal pattern detection not only yield accurate object localization but improve detection accuracy.

There also exist other region proposal methods, such as Region Proposal Networks (RPNs) [64] and Bing [73]. We compare our MRC method to RPNs and Bing, and the qualitative results are shown in Fig. 5. Furthermore, the quantitative results are presented and discussed in
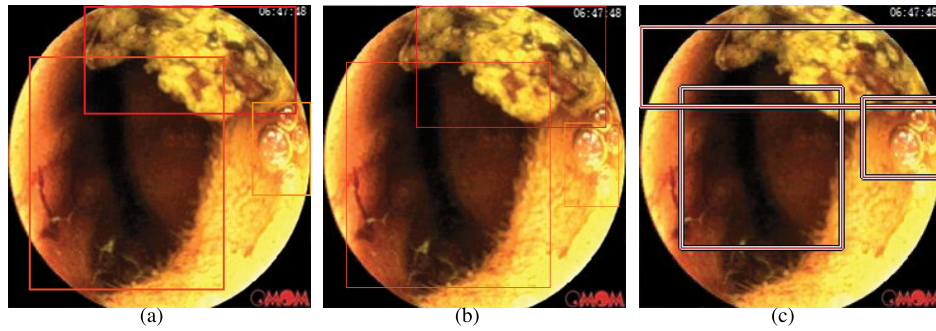
FIGURE 5. A visual illustration of the abnormal pattern localization obtained by each method. The methods are from left to right: (a) Our method, (b) RPNs, and (c) Bing. Our method has a visibly better localization accuracy. (a) and (b) Display the detection results yielded by the CascadeProposal detector in our method and the RPNs, respectively.
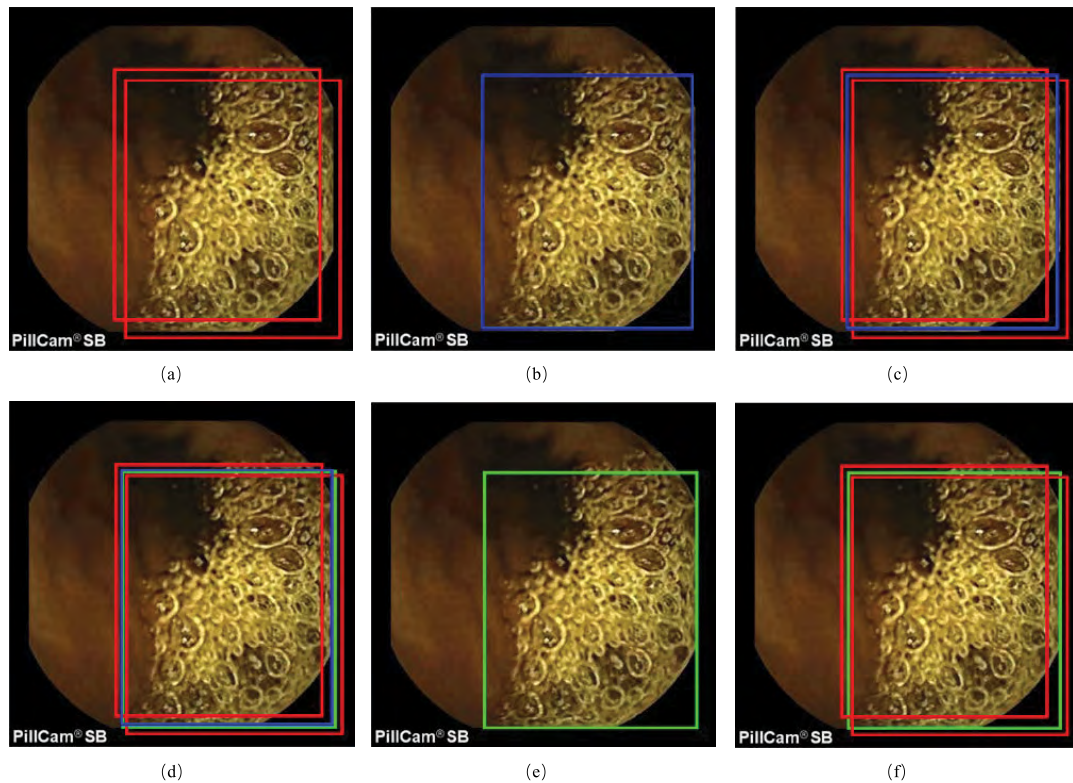

FIGURE 6. A simplified illustration of dense region fusion. Top to bottom, left to right: (a) Two region proposals by the detection model, (b) The result of dense region fusion, and (c) Illustration of combining (a) and (b). (d) Illustration of combining (c) and the ground-truth region box. (e) The ground-truth region box. (f) Illustration of combining (a) and (e).

Subsections IV-C and IV-D. It can be seen from Fig. 5 that the region proposals generated by our method seem to have a better localization accuracy and a higher recall than the other two methods. One reason may be that the methods utilize several limited fixed window sizes (e.g., RPNs using 9 anchors and Bing using 36 quantized target window sizes) to generate object proposals, which does not fulfill the requirement of non-rigid and amorphous object detection. However, for the region proposals, RPNs perform significantly better than Bing.

For dense region fusion (DRF), the experimental results are obtained by using the above-mentioned discriminant function and fusion method described in Subsubsection III-C5. The results are shown in Fig. 6. We take two region proposals (red) generated by the Detection model as an example. It can be seen from Fig. 6 that there is a significant offset between the region proposals and the ground truth region (green). The generated region proposal does not provide desirable localization. However, there is only a small offset between the region proposal (blue) obtained by the DRF

and the ground truth region proposal (green). The fused region is approximately similar to the ground truth region. Concretely, we use $\alpha = 0.1$ and $\beta = 0.9$ in (1) to obtain the fusion results. The detailed quantitative analysis is shown in Subsections IV-C and IV-D.

C. DETECTION PERFORMANCE

We evaluate our proposed methods on the WCE2017 dataset. We utilize the model-based transfer learning strategy to train three detection networks on our WCE2017 image dataset by using the ImageNet pretrained model. The first network is CascadeProposal with ZF, the second is Fast R-CNN with ZF or VGG-16, and the final is Faster R-CNN with ZF or VGG-16. We also train our CascadeProposal network from scratch. Table 4 shows the detection results using our CascadeProposal network, either trained from scratch or fine-tuned from the off-the-shelf pretrained model. For training and testing, we use various region proposal methods to generate region proposals. We adopt the same region generation algorithm in the training and testing stages to maintain consistency between the training and testing proposals, which serves as a baseline for comparisons with other methods.

It can be seen from Table 4 that we reach the following conclusions. 1) Using the region proposal method of MRC+SRS+DRF at the testing stage, we can achieve a mean average precision (mAP) of 70.3% under the final model trained by the MRC+Neg+TL method, surpassing any other single region generation algorithm by a significant margin, which indicates that the trained model has a better generalization performance than other models trained by a single region proposal method. 2) Using the MRC+SRS and MRC+SRS+DRF for abnormal pattern detection, we can achieve mAP values of 69.5% and 70.3%, correspondingly. The mAP values outperform MRC by 1.1 and 1.9 points, respectively, which shows that DRF and SRS methods can improve the detection and localization accuracy. 3) When using MRC, we achieve mAP values of 68.4% and 64.3%

TABLE 4. Detection results of the WCE2017 test set (trained on trainval). The detection networks are fast R-CNN with ZF baseline (rows 1-4), Faster R-CNN with ZF (rows 5-6) and CascadeProposal with ZF (rows 7-12), respectively. However, various proposal methods are used for training and testing. The results of the bold-faced rows serve as a baseline for comparison with other methods for training and testing. The acronyms MRC, DRF, SRS, Neg, and TL denote multiregional combination, dense region fusion, salient region segmentation, negative category, and transfer learning, respectively. The used metric is mAP.

training stage region proposal		testing stage region proposal		mAP
method	#boxes	method	#proposals	
SS+TL	**1000**	**SS**	**1000**	**0.618**
EB+TL	**1000**	**EB**	**1000**	**0.589**
OB+TL	**1000**	**OB**	**1000**	**0.507**
Bing+TL	**800**	**Bing**	**800**	**0.563**
RPN	2000	RPN(stage1)	300	0.686
RPN	2000	RPN(stage2)	300	0.721
MRC+Neg+TL	**200**	**RPN**	**600**	**0.639**
MRC	**200**	**MRC**	**200**	**0.626**
MRC+Neg	200	MRC	200	0.643
MRC+Neg+TL	**200**	**MRC**	**200**	**0.684**
MRC+Neg+TL	200	MRC+SRS	200	0.695
MRC+Neg+TL	**200**	**MRC +SRS+DRF**	**200**	**0.703**

by using the trained models based on MRC+Neg+TL and MRC+Neg, respectively. The mAP of the trained model based on MRC+Neg+TL outperforms the other mAP value based on MRC+Neg by 4.1 points. This result proves that the potential of transfer learning from natural images to medical images could be beneficial for WCE abnormal pattern detection with limited available medical data and annotations. 4) Compared with the trained model from scratch based on MRC, the detection accuracy rate of the trained model based on MRC+Neg increases by 1.7 points by adopting the same region proposal generation method (MRC). It shows that the network trained by adding negative categories can improve detection performance. 5) The mAP of the model trained only from scratch based on MRC outperforms other single region proposal methods such as SS, EB, OB, and Bing by 0.8, 3.7, 11.9 and 6.3 points, respectively. This demonstrates that the multiregional combination has advantages over the single region proposal generation method for object detection. 6) When using MRC and RPN in the testing stage, we can achieve mAP values of 68.4% and 63.9% by using the trained models based on MRC+Neg+TL, respectively. The mAP of MRC outperforms the PRN by 4.5 points. This result demonstrates that MRC may have a better localization accuracy. 7) Using RPN during the testing stage, we can achieve mAP values of 70.3% and 72.1% under the final models trained by the CascadeProposal using MRC+Neg+TL and Fast R-CNN using RPN (i.e., Faster R-CNN) proposals, respectively. The detection performance of Fast R-CNN using RPN proposals surpasses our methods by 1.8 points, which indicates that the final model trained by Faster R-CNN has a better generalization performance. However, Faster R-CNN uses 2000 proposals to train model. Also, the detection performance of our trained model surpasses the Faster R-CNN at stage1 by 1.7 points. So, to some extent, the performance of our method may not be lower than that of Faster R-CNN.

From what has been discussed above, we find that our CascadeProposal model has a better generalization capability relative to Fast R-CNN with ZF baseline, but Faster R-CNN with ZF has a slight advantage over our network. These experimental results in Table 4 verify our previous hypothesis that multiregional combination, dense region fusion, salient region segmentation, a negative category, transfer learning, and a redesigned network architecture (i.e., CascadeProposal) can improve the detection performance.

Additionally, the detection performance of various trained model based on CascadeProposal with ZF architecture for each category is detailed in Table 5. Each detection method corresponds to the trained model in Table 4.

Furthermore, we also perform a set of experiments on the Fast R-CNN with VGG-16 network [12], [57]. The data are shown in Table 6, and the best mAP of 72.3% can be achieved by using our method. It is notable that this experiment only shows the performance of the trained model using a transfer learning strategy. Faster R-CNN with VGG-16 still has a slight advantage over Fast R-CNN with VGG-16 based on the MRC+Neg+TL method.
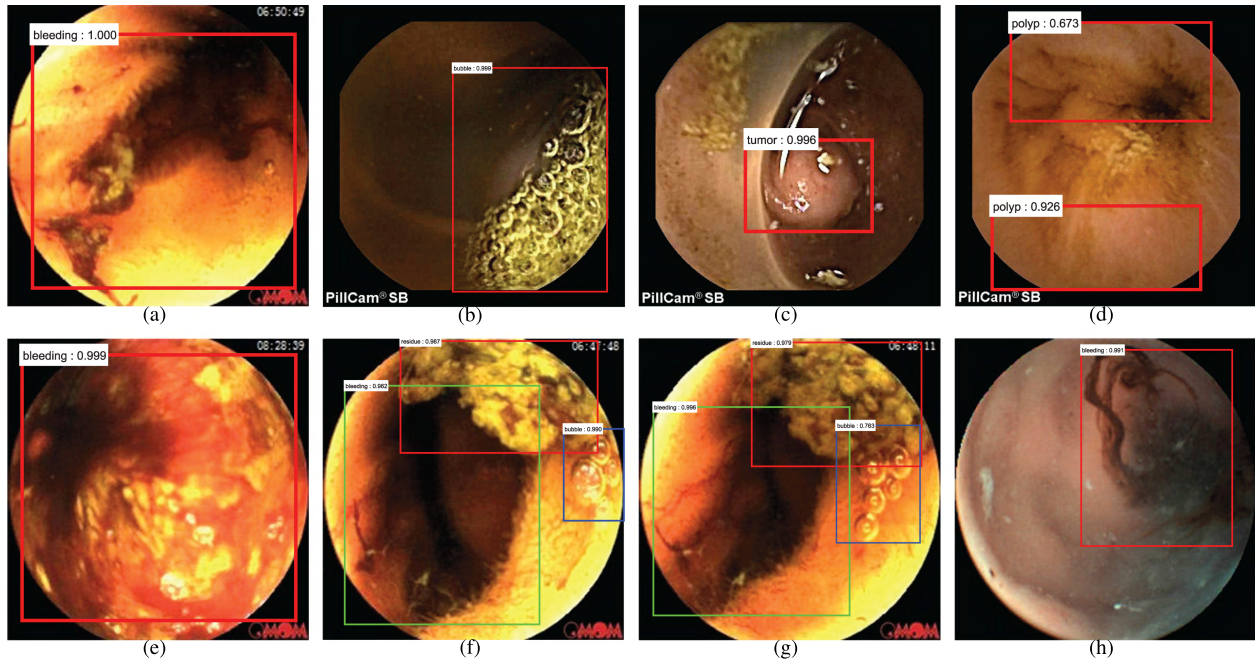
FIGURE 7. Some selected examples of WCE abnormal pattern detection results on the WCE2017 test set using the CascadeProposal architecture. The training data is WCE2017 trainval. Each output box is associated with a class label and a softmax score. A score threshold of 0.6 is used to display these images. An IoU threshold of 0.3 is used to reduce redundant boxes.

TABLE 5. Results on the WCE2017 test set corresponding to various fast R-CNN with ZF baseline, faster R-CNN with ZF and CascadeProposal with ZF models in Table 4. The used metric is mAP.

method	mAP	residue	bleeding	bubble	tumor	polyp
SS	0.618	0.827	0.908	0.523	0.470	0.361
EB	0.589	0.800	0.907	0.496	0.354	0.333
OB	0.507	0.760	0.907	0.330	0.199	0.337
Bing	0.563	0.845	0.907	0.474	0.411	0.309
RPN(stage1)	0.686	0.808	0.906	0.695	0.618	0.403
RPN(stage2)	0.721	0.812	0.907	0.733	0.605	0.548
RPN	0.639	0.850	0.906	0.536	0.592	0.311
MRC	0.626	0.845	0.907	0.576	0.492	0.309
MRC	0.643	0.846	0.908	0.586	0.512	0.323
MRC	0.684	0.851	0.907	0.629	0.513	0.520
MRC+SRS	0.695	0.864	0.910	0.668	0.516	0.518
MRC+ SRS+DRF	0.703	0.872	0.912	0.679	0.523	0.530

Fig. 7 shows some results of the WCE2017 test set using the CascadeProposal network.

D. RECALL RESULTS

When using detection proposals for detection, it is important to have a good coverage of the objects of interest in the test image, since missed objects cannot be recovered in the subsequent classification stage. Thus, it is common practice to evaluate the quality of proposals based on the recall of the ground truth annotations [59]. We employ the common and primary recall metrics mentioned in [58] and [59] to compare the recall results between our proposal method and other top proposal methods [15]–[18], [60], [61]. Hosang *et al.* [59] refer to two primary metrics to evaluate detection proposals that are referred to as Recall-to-IoU and Recall-to-Proposal in this paper. For a fixed number of proposals (e.g., number of proposals = 50, 200, 500 in this paper), the Recall-to-IoU metric is the fraction of ground truth annotations covered as the intersection over union (IoU) threshold is varied, as shown in Fig. 8. It can be seen from Fig. 8 that our proposal generation method (i.e., multiregional combination plus region proposal rejection module) performs well and has the highest recall. For high IoU thresholds (e.g., 0.8 to 0.9), our method still achieves promising detection results compared with other region proposal generation methods. Hosang *et al.* [59] note that although recall at an IoU threshold of 0.5 has been

TABLE 6. Results on the WCE2017 test set using fast R-CNN with VGG-16 (rows 1-4,7-8) and faster R-CNN with VGG-16 (rows 5-6), where the proposals in rows 4-6 are generated by RPN with ZF. The used metric is mAP.

| method		mAP	residue	bleeding	bubble	tumor	polyp
training stage	testing stage						
SS+TL	SS	0.683	0.828	0.907	0.623	0.583	0.472
EB+TL	EB	0.662	0.822	0.907	0.601	0.569	0.411
OB+TL	OB	0.601	0.806	0.906	0.538	0.366	0.389
Bing+TL	Bing	0.611	0.749	0.906	0.546	0.548	0.305
RPN	RPN(stage1)	0.739	0.824	0.907	0.730	0.693	0.541
RPN	RPN(stage2)	0.699	0.812	0.908	0.684	0.691	0.401
MRC+Neg+TL	RPN	0.682	0.821	0.906	0.622	0.665	0.397
MRC+Neg+TL	MRC +SRS+DRF	0.723	0.898	0.913	0.687	0.598	0.518
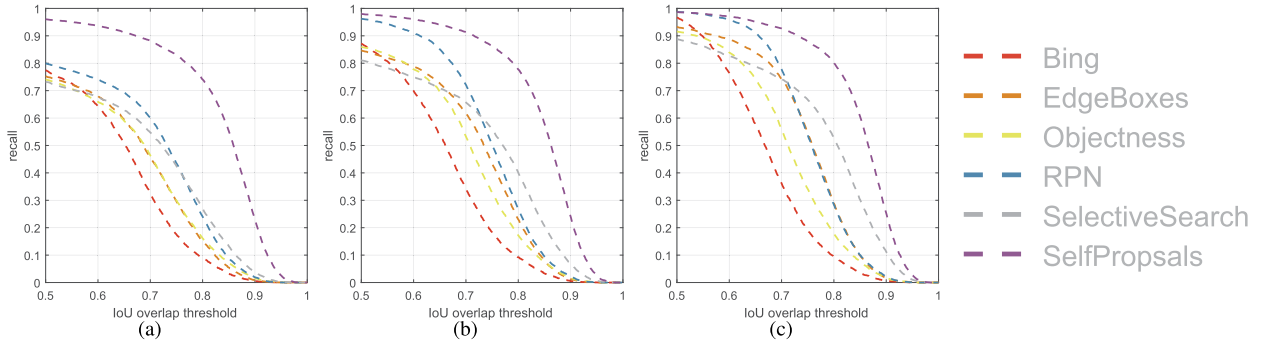
FIGURE 8. Recall versus IoU threshold on the WCE2017 test set. (a) 50 region proposals. (b) 200 region proposals. (c) 500 region proposals.
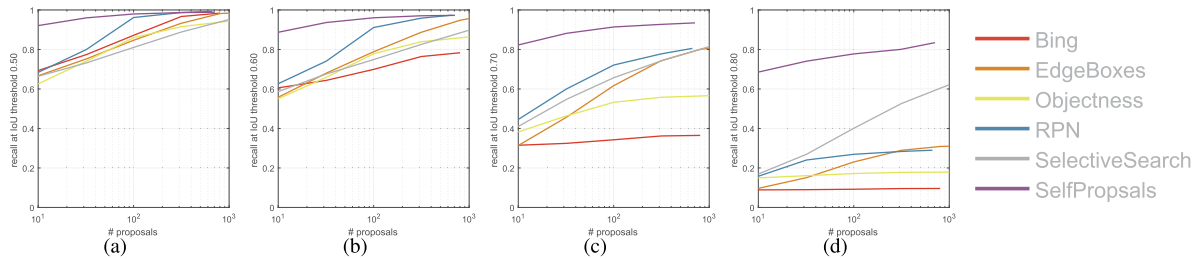

FIGURE 9. Recall versus number of proposals on the WCE2017 test set. Using IoU = 0.5, 0.6, 0.7, and 0.8, we illustrate the curves of Recall-to-Proposal for various region proposal generation methods.

traditionally used to evaluate object proposals, it is not a good metric for predicting detection performance. Our region proposal generation method achieves good results across a variety of IoU thresholds, which is desirable in practice and plays an important role in the performance of object detectors [59], [65].

For a fixed IoU threshold (e.g., IoU = 0.5, 0.6, 0.7, and 0.8), the Recall-to-Proposal is the proposal recall as the number of proposals is varied, as shown in Fig. 9. The plots show Recall-to-Proposal for various methods. Hosang *et al.* [59] show that there is a correlation between detector performance and recall at different overlap thresholds. More concretely, there is a strong correlation at an IoU range of approximately 0.6 to 0.8. An object proposal with an IoU threshold of 0.5 is too loose to fit the ground truth object, which usually leads to the failure of following object detectors [65]. In fact, Hosang *et al.* [59] note that recall at an IoU threshold of 0.5 is only weakly correlated with detection performance. It can be seen from Fig. 9 that our method achieves better recalls than other methods at an IoU range of 0.6 to 0.8, which indicates our method can serve as a good predictor for detector performance.

E. TRAINING AND TESTING TIME

We use the Caffe framework [74] and a Nvidia TitanXp GPU to train the various CNNs. The entire training and testing times of Fast R-CNN with ZF, Faster R-CNN with ZF, and CascadeProposal with ZF are shown in Table 7, and those of Fast R-CNN with VGG-16 and Faster R-CNN

with VGG-16 are shown in Table 8. The CascadeProposal with ZF and Fast R-CNN with VGG-16 networks based on MRC+Neg+TL are approximately 57 minutes and 5.2 hours, respectively. The mean testing times per image with CascadeProposal with ZF and Fast R-CNN with VGG-16 models are approximately 89.6 ms and 278.6 ms, respectively. It is noteworthy that the time reflects only the detection time, excluding the proposal generation time. Our detection system

TABLE 7. Runtime comparison between various models trained by using fast R-CNN with ZF (rows 1-4), faster R-CNN with ZF (rows 5-6) and CascadeProposal with ZF (rows 7-11).

method	training time(min)	method	testing rate (ms/img)
SS	24	SS	22.4
EB	24	EB	33.8
OB	25	OB	26.7
Bing	24	Bing	21.6
RPN	210	PRN(final)	63.1
MRC+Neg+TL	57	RPN	27.1
MRC	40	MRC	91.1
MRC+Neg	57	MRC	90.4
MRC+Neg+TL	57	MRC	89.3
MRC+Neg+TL	57	MRC+SRS	89.5
MRC+Neg+TL	57	MRC+DRF+SRS	89.6

TABLE 8. Runtime comparison between various models trained by using fast R-CNN with VGG-16(rows 1-3,5) and faster R-CNN with VGG-16 (rows 4).

method	training time(h)	method	testing rate(ms/img)
SS+TL	3	SS	86.3
EB+TL	3	EB	89.7
OB+TL	3	OB	88.1
RPN	15	RPN with ZF(final)	130.1
MRC+Neg+TL	5.2	MRC+DRF+SRS	278.6

takes a total of 89.6 ms for both the RPR and Detection modules, which take approximately 86 ms and 5 ms, respectively.

V. DISCUSSION

In this section, we compare and discuss the performances of the various models trained by the methods in the previous sections. We explain the major findings and significance in terms of the results of our work. We also discuss and analyze future work on WCE abnormal pattern detection.

For the CascadeProposal architecture, it can be seen from Table 4 and Table 5 that using the model trained by the MRC+Neg+TL method, we can achieve a final mAP of 70.3%. This model outperforms single region proposal methods such as OB under the Fast R-CNN with ZF baseline by 19.6 points, which shows that the best model performance can be achieved using the MRC+Neg+TL method. For Fast R-CNN with VGG-16 architecture, we can achieve a better mAP of 72.3% at the price of a higher computational cost. This network does not exhibit a good speed-accuracy trade-off. For Faster R-CNN with ZF and VGG-16, we can reach the best mAP values of 72.1% and 73.9%, respectively, at the price of more training time. It can be seen from Table 7 and Table 8 that using the CascadeProposal architecture, we can reach a good trade-off between computational efficiency and model performance. By combining Table 5, Table 6, Table 7, and Table 8, we compare and analyze the detection accuracy and the running time among CascadeProposal with ZF, Fast R-CNN with ZF or VGG-16, and Faster R-CNN with ZF or VGG-16 networks. CascadeProposal trained by MRC+Neg+TL achieves competitive results, with a final mAP of 70.3% while requiring in total approximately 57 minutes for the whole training time. However, Fast R-CNN with VGG-16 with a better mAP of 72.3% takes approximately 5.2 hours, and Faster R-CNN with VGG-16 with the best 73.9% mAP takes 15 hours; these mAP values are slightly higher than that of CascadeProposal. Furthermore, it can also be visualized from 4, 5, and 6 that CascadeProposal achieves a promising detection performance.

In our work, we utilize CNNs to detect WCE abnormal patterns. Our work is different from traditional detection or classification methods [5]–[10], [22] applied to WCE image detection. As Section II notes, the traditional methods are generally classification methods rather than detection methods. Our scheme mainly aims to localize, recognize and detect WCE abnormal patterns. Therefore, we believe that this work will be helpful in clinical practice and can provide direct detection results of WCE abnormal patterns for gastroenterologists. Moreover, through a review of published papers, we find that there are only a few reports that have used CNNs to address the problem of WCE abnormal pattern detection. Therefore, this work is the first step toward using deep learning for WCE abnormality detection, and this may be a guide for researchers in subsequent studies. There has also been some recent research employing deep learning to medical image processing and analysis. These studies include classification tasks [32], [36], [42]–[49] and detection tasks

[34], [35], [75]. Deep learning has shown great potential for medical image analysis [32]–[40], [42]–[49], [55], [75].

In this study, we find that the performance of the CascadeProposal network fine-tuned by the ImageNet [14] pretrained model generally surpasses the model trained from scratch. This conclusion conforms with [36] and [39], which demonstrate that CNN models pre-trained via transfer learning outperform or perform as well as CNN models trained from scratch. However, training a deep CNN model is not an easy task; it can be time-consuming and complicated and is often faced with overfitting or failure to converge. Therefore, we suggest using a transfer learning strategy to address the problem of medical image classification and detection and that following studies should focus on transfer learning.

VI. CONCLUSION AND FUTURE WORK

In this work, we report a good method using deep learning (referring in particular to CNNs) for abnormal pattern detection in abnormal WCE images. Several methods are integrated, including a deep cascade network architecture (CascadeProposal), multiregional combination, salient region segmentation, dense region fusion, a negative category, and transfer learning. Our overall experimental results proved that the above-mentioned methods significantly improve the performance of the trained models. Concretely, we achieve a final mAP of 70.3% and a better mAP of 72.3% via CascadeProposal with ZF and Fast R-CNN with VGG-16 networks, respectively, using MRC+Neg+TL method in the training stage and MRC+DRF+SRS method in the testing stage. This work is the first step toward using deep learning for WCE abnormality detection and provides a guideline for subsequent studies. Our method will help physicians directly determine the accurate localization of abnormal patterns.

Since the WCE abnormal pattern detection task has a real-time need. So, we will try to develop and design state-of-the-art real-time detector to this detection task later. Also, it is well known that the WCE produces a large number of redundant and irrelevant frames for a patient in one examination. How to eliminate redundant and locate critical informative frames is an urgent demand. Therefore, our future research will focus on some effective WCE video summarization methods.

REFERENCES

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, p. 417, May 2000.

[2] B. S. Lewis and P. Swain, "Capsule endoscopy in the evaluation of patients with suspected small intestinal bleeding: Results of a pilot study," *Gastrointestinal Endoscopy*, vol. 56, no. 3, pp. 349–353, Sep. 2002.

[3] D. Hartmann *et al.*, "A prospective two-center study comparing wireless capsule endoscopy with intraoperative enteroscopy in patients with obscure GI bleeding," *Gastrointestinal Endoscopy*, vol. 61, no. 7, pp. 826–832, Jun. 2005.

[4] D. G. Adler, M. Knipschield, and C. Gostout, "A prospective comparison of capsule endoscopy and push enteroscopy in patients with GI bleeding of obscure origin," *Gastrointestinal Endoscopy*, vol. 59, no. 4, pp. 492–498, Apr. 2004.

[5] Y. Yuan, B. Li, and M. Q.-H. Meng, "WCE abnormality detection based on saliency and adaptive locality-constrained linear coding," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 149–159, Jan. 2017.

[6] Y. Yuan, B. Li, and M. Q. H. Meng, "Bleeding frame and region detection in the wireless capsule endoscopy video," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 2, pp. 624–630, Mar. 2016.

[7] 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.

[8] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 2046–2057, Oct. 2015.

[9] 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.

[10] B. Li and M. Q.-H. Meng, "Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 323–329, May 2012.

[11] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 10, pp. 2777–2786, Oct. 2011.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[14] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[15] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[16] K. E. A. Van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1879–1886.

[17] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[18] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1558–1570, Aug. 2015.

[19] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[20] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[22] Y. Yuan, X. Yao, J. Han, L. Guo, and M. Q.-H. Meng, "Discriminative joint-feature topic model with dual constraints for WCE classification," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2074–2085, Jul. 2017.

[23] R. Kumar, Q. Zhao, S. Seshamani, G. Mullin, G. Hager, and T. Dassopoulos, "Assessment of Crohn's disease lesions in wireless capsule endoscopy images," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 2, pp. 355–362, Feb. 2012.

[24] O. Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, M. Alizadeh, and H. A. Soleimani, "Informative and uninformative regions detection in WCE frames," *J. Adv. Comput.*, vol. 3, no. 1, pp. 12–34, 2014.

[25] M. Alizadeh, O. H. Maghsoudi, K. Sharzehi, H. R. Hemati, A. K. Asl, and A. Talebpour, "Detection of small bowel tumor in wireless capsule endoscopy images using an adaptive neuro-fuzzy inference system," *J. Biomed. Res.*, vol. 33, no. 5, pp. 419–427, 2017.

[26] O. H. Maghsoudi, "Superpixel based segmentation and classification of polyps in wireless capsule endoscopy," in *Proc. IEEE Signal Process. Med. Biol. Symp.*, Philadelphia, PA, USA, Dec. 2017, pp. 1–4.

[27] M. T. Coimbra and J. P. S. Cunha, "MPEG-7 visual descriptors—Contributions for automated feature extraction in capsule endoscopy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 628–637, May 2006.

[28] M. Mackiewicz, J. Berens, M. Fisher, and D. Bell, "Colour and texture based gastrointestinal tissue discrimination," in *Proc. IEEE Acoust. Speech and Signal Process.*, Toulouse, France, May 2006, p. 2.

[29] M. Mackiewicz, J. Berens, and M. Fisher, "Wireless capsule endoscopy color video segmentation," *IEEE Trans. Med. Imag.*, vol. 27, no. 12, pp. 1769–1781, Dec. 2008.

[30] P. Szczypinski, A. Klepaczko, M. Pazurek, and P. Daniel, "Texture and color based image segmentation and pathology detection in capsule endoscopy videos," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 396–411, Jan. 2014.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229

[32] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[33] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[34] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, 2013, pp. 411–418.

[35] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck. (2016). "Deep learning for identifying metastatic breast cancer." [Online]. Available: https://arxiv.org/abs/1606.05718

[36] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[37] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 76–84, Jan. 2017.

[38] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, Mar. 2015.

[39] N. Tajbakhsh *et al.*, "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.

[40] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017.

[41] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Comput. Methods Programs Biomed.*, vol. 153, pp. 1–9, Jan. 2018.

[42] A. K. Sekuboyina, S. T. Devarakonda, and C. S. Seelamantula, "A convolutional neural network approach for abnormality detection in wireless capsule endoscopy," in *Proc. 14th IEEE Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 1057–1060.

[43] K. Pogorelov *et al.*, "KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. 8th ACM Int. Conf. Multimedia Syst.*, 2017, pp. 164–169.

[44] J.-S. Yu, J. Chenm Z. Q. Xiang, and Y.-X. Zou, "A hybrid convolutional neural networks with extreme learning machine for WCE image classification," in *Proc. IEEE Int. Robot. Biomimetics*, Dec. 2015, pp. 1822–1827.

[45] Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. J. Deng, "Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network," in *Proc. IEEE Int. Digit. Signal Process.*, Jul. 2015, pp. 1274–1278.

[46] X. Jia and M. Q.-H. Meng, "Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and CNN features," in *Proc. 39th IEEE Eng. Med. Biol. Soc.*, Jul. 2017, pp. 3154–3157.

[47] P. Li, Z. Li, F. Gao, L. Wan, and J. Yu, "Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images," in *Proc. IEEE Multimedia Expo*, Jul. 2017, pp. 1518–1523.

[48] 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.

[49] S. Seguí *et al.*, "Generic feature learning for wireless capsule endoscopy analysis," *Comput. Biol. Med.*, vol. 79, no. 1, pp. 163–172, Dec. 2016.

[50] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Med. Phys.*, vol. 44, no. 4, pp. 1379–1389, 2017.

[51] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images," in *Proc. 38th IEEE Eng. Med. Biol. Soc.*, Orlando, FL, USA, Aug. 2016, pp. 639–642.

[52] A. Mohammed *et al.*, "Sparse coded handcrafted and deep features for colon capsule video summarization," in *Proc. 30th IEEE Comput.-Based Med. Syst.*, Thessaloniki, Greece, Jun. 2017, pp. 728–733.

[53] M. Hajabdollahi *et al.* (2018). "Multiple abnormality detection for automatic medical image diagnosis using bifurcated convolutional neural network." [Online]. Available: https://arxiv.org/abs/1809.05831

[54] S. Xu, H. Wu, and R. Bie, "CXNet-m1: Anomaly detection on chest X-rays with image-based deep learning," *IEEE Access*, vol. 7, pp. 4466–4477, 2018. doi: 10.1109/ACCESS.2018.2885997.

[55] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[56] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[57] K. Simonyan and A. Zisserman. (2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[58] J. Hosang, R. Benenson, and B. Schiele. (2014). "How good are detection proposals, really?" [Online]. Available: https://arxiv.org/abs/1406.6962

[59] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Sep. 2016.

[60] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 73–80.

[61] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[62] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[63] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. (2017). "Salient object detection: A survey." [Online]. Available: https://arxiv.org/abs/1411.5878

[64] 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.

[65] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2016, pp. 845–853.

[66] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "DeepProposal: Hunting objects by cascading deep convolutional layers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 2578–2586.

[67] W. Kuo, B. Hariharan, and J. Malik, "DeepBox: Learning objectness with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2015, pp. 2479–2487.

[68] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2014, pp. 580–587.

[69] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware CNN model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1134–1142.

[70] S. Gidaris and N. Komodakis, "LocNet: Improving localization accuracy for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2016, pp. 789–798.

[71] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[73] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3286–3293.

[74] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[75] P. Rajpurkar *et al.* (2017). "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning." [Online]. Available: https://arxiv.org/abs/1711.05225

[76] Y. Deng, F. Bao, X. Deng, R. Wang, Y. Kong, and Q. Dai, "Deep and structured robust information theoretic learning for image analysis," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4209–4221, Jul. 2016.

[77] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 1006–1012, Aug. 2017.

LIBIN LAN received the B.Sc. degree from the Chongqing University of Posts and Telecommunications, in 2008, and the M.Sc. degree from Chongqing University, Chongqing, China, in 2011, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include deep learning, image processing, medical image analysis, and computer vision.

CHUNXIAO YE received the B.Sc. degree from Sichuan University, in 1995, and the M.Sc. and Ph.D. degrees from Chongqing University, in 2002 and 2005, respectively, all in computer science, where he is currently a Professor in computer science. He has published over 50 papers His research interests include information security, especially access control and privacy protection, pattern recognition, and cloud computing. He is also a Senior Member of China Computer Federation and a member of IET.

CHENGLIANG WANG is currently a Professor and the Ph.D. degree Supervisor with the College of Computer Science, Chongqing University. He has published more than 80 writings, including more than 20 articles in SCI, EI, and he hold six national patents. His research interests include smart control for complex systems and the theory and application of artificial intelligence. He is a Senior Member of the China Computer Federation and a member of the America Association of Computing Machinery. He is a Visiting Scholar with the George Institute of Technology. He is a Reviewer of some international Journals, such as the *Journal of Algorithms and Computational Technology*, the IEEE Transactions on Parallel and Distributed Systems, and the IEEE Transactions on Systems, Man, and Cybernetics, Part B.

SHANGBO ZHOU received the B.Sc. degree in mathematics from the Guangxi National College, in 1985, the M.Sc. degree in mathematics from Sichuan University, in 1991, and the Ph.D. degree in circuit and system from Electron Science and Technology University. From 1991 to 2000, he was with the Chongqing Aerospace Electronic and Mechanical Technology Design Research Institute. Since 2003, he has been with the College of Computer Science and Engineering, Chongqing University, where he is currently a Professor. His current research interests include artificial neural networks, physical engineering simulation, image processing, and nonlinear dynamics.

• • •