IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 13, 2020, accepted August 19, 2020, date of publication August 27, 2020, date of current version September 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019888

A Study on Wireless Capsule Endoscopy for Small Intestinal Lesions Detection Based on Deep Learning Target Detection

ZHIGUO XIAO AND LI NIAN FENG
College of Computer Science and Technology, Changchun University, Changchun 130022, China
Corresponding author: Li Nian Feng (cculinianfeng@126.com)

This work was supported in part by the Project of the Education Department of Jilin Province under Grant 2019LY505L28, in part by the Jilin Science and Technology Development Plan Project under Grant 20200404221YY, and in part by the School Level Training Program under Project 2018JBC05L11.

ABSTRACT Wireless capsule endoscope (WCE) has been verified in clinical medicine for many years. However, the detection process needs experienced doctors to read the film manually for a long time. In addition, the cost of the endoscope itself leads to a high cost of WCE detection and overall cycle is long. New research method based on deep learning technology with robustness and high accuracy can reduce the detection cost and benefit the public. According to the characteristics of small intestine lesion, a method focuses on labeling and feature detection which can optimize the process by analyzing small intestine WCE image and experimental comparison. Based on the YOLOv3 detection network, retaining the original basic feature of detection network, an improved one is further optimized and effectively verified. Finally, the redundant images are filtered out by comparing the Hash value of images, presenting the final concise detection results for doctors. Starting from image labeling, the design of deep learning network structure for the image of small intestine digestive tract endoscope is studied, which can effectively improve intelligent detection computer-aided clinical application of WCE, with higher accuracy and lower missing detection rate than manual detection.

INDEX TERMS Detection network, wireless capsule endoscope, small intestine lesions, YOLOv3.

I. INTRODUCTION

According to the statistics report in 2016 released by the American Cancer Association, the incidence of digestive tract cancer in the United States is about 305000, including 132000 women and 173000 men; the death toll caused by digestive tract diseases has reached 153000, including 64000 women and 89000 men [1]. China has a high incidence rate of digestive tract cancer which takes 42% in the world and increases by 10 million annually, and the number of people in need is 20 million every year. The common clinical detection methods are mainly mechanical push in endoscopy, such as gastroscopy and enter scope. But this method must be to insert the endoscope into the body from the patient's mouth or anus, which will cause injury and pain to the patient's body, and even a psychology of fear, so they choose to give up the examination and delay the cure. Meanwhile, both methods are difficult to reach the small intestine, so it is impossible to carry out the examination. Some non-invasive imaging methods, such as ultrasonic imaging and computed tomography, will also be used for gastrointestine diseases, but for the low imaging resolution, which is easy to affect or even mislead the diagnosis results of doctors. WCE just makes up for the defects, which has the advantages of safety, painless, noninvasive and so on [2]. At the same time, it can also go deep into areas of the small intestine where are difficult to reach by the traditional ways, and truly realize the detection of the whole digestive tract. WCE is a milestone for curing clinical digestive tract diseases and achieving good results.

The length of human small intestine is 5-7 meters, where needs 2-8 hours from swallowing WCE into the body to discharge. With frequency of 2 frames per second, each patient has at least 50000-80000 images. Reading the number of images is an arduous and time-consuming task, and it takes an average of two hours for a professional doctor to read the WCE image of a patient. However, the number

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

of pathological images is very small, accounting for about 5% - 10%, and the doctor may miss a diagnosis. Therefore, in order to reduce missed diagnosis, doctors usually have to check repeatedly leading to labor intensity and inefficiency. The cost of WCE is mainly composed of a capsule and a physician, who reads the images. In continuous decline price of capsule endoscopy, the cost of manual reading is gradually increasing, resulting in the overall cost high which makes it far from popular in China. The early diagnosis and treatment is of great significance in clinical medicine, and the potential market is huge in China. With the development of artificial intelligence (AI) technology, automatic detection of small intestine lesions has become an inevitable trend. At present, the automatic detection can be roughly divided into two types: traditional machine learning method and deep learning method [3].

The core idea of traditional machine learning methods can be summarized as follows: extracting features manually and adopting appropriate classifiers [4]. For the detection of small intestine lesions, the extraction features mainly start from the color features and texture features of small intestine images. In classifiers, SVM (support vector machine), KNN (k-nearest neighbor), MLP (multi-layer perception) are commonly used, among which SVM classifier is very popular among many researchers [5]–[8].

In deep learning method, Teng Zhou *et al.* firstly cut the frame, and then correct the intensity of the frame before rotation. A deep Convolution Neural Network was established to train these frames by GoogLeNet to measure quantitatively the presence and degree of small intestine lesions, and to evaluate mucosal atrophy and other causes in real time [9]. Haya Alaskar *et al.* used AlexNet and GoogLeNet to extensively evaluate small intestine ulcers or non-ulcers. Moreover, the images containing ulcer objects are analyzed to evaluate the efficiency of CNNs [10]. Some scholars have researched the small intestinal bleeding recognition, using different network structures for experimental comparison, correspondingly changing the classifiers, and the experiment also achieved good results [11]–[13].

From above, people are trying to use intelligent detection technology to detect the WCE image, but the speed and accuracy need to be improved. There are many methods applied in the research of lesion detection of WCE image. The deep learning method has been applied to medical image analysis in recent years with the development and application of image detection technology, which has been paid more attention to.

II. RELATED WORK

A. DEVELOPMENT AND RESEARCH STATUS OF TARGET DETECTION

In recent five years, scholars have divided target detection into single-stage and double-stage according to whether there is a candidate frame in the generation stage, which studied the improvement and combination of target detection algorithm, and the method of improving and optimizing the mainstream target detection algorithm. Girshick *et al.* proposed a target detection algorithm based on deep learning, R-CNN, which can be called a milestone in target detection [14]. Based on R-CNN, He *et al.* and Girshick *et al.* proposed Spatial Pyramid Pooling Net (SPP-Net) and Fast R-CNN algorithm respectively. Two methods only need to send the image into the depth network once, and then map all candidate frame on a certain layer of the network, which improves the detection speed [15], [16]. Aiming at the problem that the selective search time of the network is long, Ren *et al.* proposed the Faster R-CNN algorithm, which added the Region Proposal Network (RPN), instead the traditional candidate frame generation methods such as selective search, realized the end-to-end training of the network and improved the network computing speed [17]. Dai *et al.* from Microsoft Asia Research Institute found that the network layer after ROI pooling no longer has translation invariance, and the number of layers after ROI pooling will directly affect the detection efficiency. A Region Based Fully Convolutional Network (RFCN) was proposed to solve this contradiction by position sensitive score maps [18]. In addition, there are also methods to improve and optimize the performance of double-stage target detection algorithm from the perspective of sample post-processing, such as Non Maximum Suppression (NMSnon) algorithm, soft-NMS and softer-NMS [19], [20]. Aiming at the low efficiency of the double-stage target detection algorithm, YOLO (you only look once) V1 omits the candidate frame extraction branch in the algorithm, and directly implements feature extraction, candidate frame classification and regression in the same depth convolution network without branches, which increases from 7 frames/s in Fast-CNN to 45 frames/s [21]. In order to improving the detection speed, Liu *et al.* proposed SSD (single shot multibox detector), which, to a large extent, balances the detection speed and detection accuracy of the single-stage target detection algorithm [22]. Aiming at the problem of low accuracy of YOLOv1, Redmon and Farhadi proposed YOLOv2 as an improved model [23]. Based on YOLOv2, combined with the newly designed residual network, Darknet-53 and the Feature Pyramid Networks (FPN), Seferbekov *et al.* carried out multi-scale fusion prediction, and YOLOv3 network was designed. These changes make YOLOv3 reach the accuracy equivalent to SSD with 1/3 time [24]. In addition, the single-stage detection networks include RON, DSSD, TridentNet and DES512 [25]–[28].

The double-stage target detection algorithm with high accuracy often consumes more computing resources, which leads to a decrease of efficiency, while the opposite is a single-stage target detection algorithm with high efficiency. Therefore, how to improve and optimize the mainstream target detection algorithm and achieve the optimal balance between accuracy and efficiency is a key problem. The existing target detection algorithms have been applied in many aspects, such as auto driving, smart city, and intelligent medical. The detection of small intestinal lesions based on WCE is a typical target detection problem. In this paper, we want to

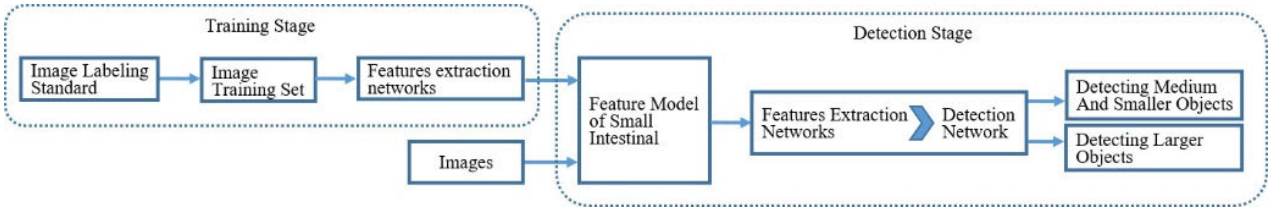


FIGURE 1. Structure of detection method for small intestinal lesions.

explore the efficient performance of different target detection algorithm on small intestine WCE image.

B. RESEARCH STATUS OF WCE FOCUS DETECTION BASED ON TARGET DETECTION METHOD

In recent years, with the development of machine learning, image-based detection of small intestinal lesions has been paid more and more attention. Some scholars use Convolutional Neural Network to identify small intestinal bleeding [11]–[13]. Panpeng Li *et al.* used different network structures for experimental comparison, and made corresponding changes in the classifiers and corresponding expansion in the amount of data. The final accuracy rate was about 98.87% [11]. Jia X *et al.* started with classifiers and compared the experimental results by different classifiers, and the final average value reached 99.00% [13]. Aoki Tomonori *et al.* focused on the most common small-bowel abnormalities, erosions and ulcerations, trained a deep convolutional neural network (CNN) system based on a Single Shot Multibox Detector. They developed and validated a new system based on CNN to automatically detect erosions and ulcerations in WCE images. The sensitivity, specificity, and accuracy of the CNN were relatively high, but the speed of the method needs to be further improved [27]. In order to locate the alimentary tract of capsule endoscopy, Mackiewicz *et al.* extracted the images of intestinal wall and divided them into regular blocks. Color, texture and motion feature vectors were extracted from each small block. Finally, the classifier was trained according to these vector features, and the classifier output of continuous frames was coded and Hidden Markov Model (HMM) was used to divide the whole alimentary tract into different organs [28]. Mahdi Alizadeh *et al.* computed 32 features incorporating four statistical measures(contrast, correlation, homogeneity and energy) calculated from co-occurrence metrics. Then, mutual information was used to select features with maximal dependence on the target class and with minimal redundancy between features. Finally, a trained classifier, adaptive neuro-fuzzy interface system was implemented to classify endoscopic images into tumor, healthy and unhealthy classes. Such techniques are valuable for accurate detection characterization and interpretation of endoscopic images [29]. Lecheng Yu *et al.* obtained features by converting the source image into LAB color space. By using the Quadratic Support Vector Machine (Q-SVM) to classify ulcer images and non-ulcer images, the accuracy rate was 90.39%, and the sensitivity was 99.00% [30]. Reference proposed a novel method based on textural features (such as Gabor filters, local binary pattern, and Haralick) in HSV color space, Fisher score test, and neural networks to detect and differentiate regions such as bleeding, tumor, and other types of gastric diseases including Crohn's. The experimental results indicate that this method is a rare study that can detect multiple gastric lesions [31]. Xiao Wu *et al.* proposed a segmented parallel region detection method, with multiscale double matched filter to detect the position of tubular structure, to identify potential areas with hookworm and obtain high detection sensitivity [32].

Compared with the traditional machine learning method, the method based on deep learning has higher detection and recognition ability in WCE small intestinal lesions image detection. Most of the existing small intestinal lesions detection methods focus on the feature extraction and detection of one kind of lesions, and the detection results are relatively good. For patients with a variety of small intestinal lesions, in order to achieve more accurate diagnosis of which diseases the patient has at the same time, it is necessary to use a variety of disease detection models for automatic identification, so there are not many complete detection methods for this purpose.

III. METHOD FOR MULTI SORT SMALL INTESTINE LESIONS BASED ON WCE

In view of the problems existing in the auxiliary diagnosis of small intestinal diseases and the current situation of detection based on machine learning, a new method based on WCE image were designed. The overall framework is shown in Fig. 1:

The detection method is divided into training stage and application detection stage. In the training stage, according to the set image labeling standard, the manual labeling of multiple lesions images is carried out to form the image training set. The image training set is used to train the detection network repeatedly to find the feature model with the best performance. In the detection stage, the detection network extracts the features of the input image through the feature model, and realizes the feature detection by the network. In this method, the detection network and image labeling standards need to be combined, which shows the advantages of this method in the detection of lesions based on WCE images.

A. IMAGE LABELING BASED ON THE CHARACTERISTICS OF SMALL INTESTINE LESIONS

The reason why deep learning can make a breakthrough depends on the complex network structure and massive data. It often leads to over fitting of models and poor training quality because of few sufficient training samples. The existing unsupervised learning is not mature enough to be applied to the medical image detection. By the semi supervised learning method, fine-tuning training adding to the pre training model is applied to the labeling method to improve the effect of labeling and detection, which is more recognized in medical image detection.

Image labeling involves two key points: one is extracting visual features. Good features can reflect the visual content information of the image to the greatest extent, which is related to the effect and accuracy of subsequent semantic mapping. It is difficult to describe an image by sole features such as pixels, contours and regions and adopt more effective image features in the process of semantic mapping; the other is semantic mapping. Discriminant Models and Generative Models have their own advantages when completing the mapping task from visual features to semantics. They improve the quality of image semantic label generation based on high-quality visual feature modeling which helps to overcome the semantic gap [33], [34]. The prominent feature of deep learning method is the ability of autonomous learning high quality visual features, which is a basic guarantee for various machine learning methods. The powerful visual expression ability of CNN can learn more advanced visual features, compared with the traditional feature extraction method [35]. By extracting the high-level fusion visual features of multi instance, CNN can not only learn better visual features, but also have more advantages in the fusion expression of multi instance.

Low level features mainly refer to some outstanding visual features extracting from the image by certain methods, and an image can be represented not only by the features of the global association of the image, but also by the features of the local salient information. This paper discusses the common low-level and high-level visual features, analyzes the advantages and disadvantages and the scope of application, so as to realize the representation, selection and integration of multiple features in image semantic mapping.

The reason why deep learning can make a breakthrough depends largely on the complex network structure and massive data support. It is difficult to provide enough training samples for most applications, which often results in model over fitting and poor training quality. The existing unsupervised learning is not mature enough to be applied to the deep learning in medical image detection. The semi supervised learning method is more convincing in medical image detection, adjusting slightly based on the pre training model to improve the effect of labeling and detection.

The feature extraction of small intestine lesion is different from that of common image objects. The tissue feature around the small intestine is a part of the lesion, which must be included in the whole lesion labeling and can also be extended to the whole WCE image labeling. For example, the characteristic of small intestine tumor is similar to other tumor in the digestive tract. If there is no tumor peripheral information characteristic, even experienced doctors can hardly locate the tumor which part of the digestive tract is. In the process of image labeling, the area size of the tissue around the lesion is the key. If too much size is selected, it will affect the extraction of the lesion features and the accuracy of image detection. If too small size is selected, the detection accuracy seems to be very high, but there will be over fitting phenomenon. Appropriate size of the surrounding lesion tissue information can improve the detection accuracy and inhibit the over fitting phenomenon. As shown in the Fig. 2 below:

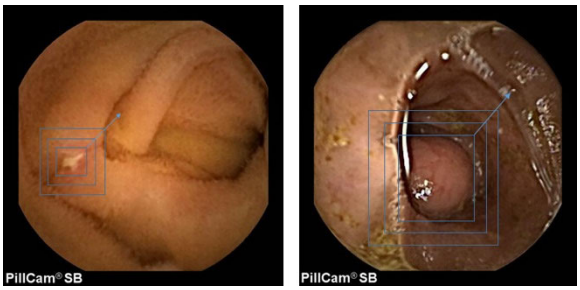


FIGURE 2. Small intestine tumor labeling diagram.

In the process, the small intestine tumor and small intestine ulcer in Fig. 2 are analyzed. The innermost labeling box only labels the tumor and ulcer area. In the proportion of the same length and width, the edge area of the tumor is also included in the labeling area. According to the experimental results, with the expansion of the area of the lesion, the detection accuracy is improving, while it is expanded to a certain range, the detection accuracy begins to decrease. The same group but two types of lesions were labeled with different areas, the training results are shown in Fig. 3:

According to the experimental observation, with the expansion of the area of the lesion, the detection accuracy is improving, while it is expanded to a certain range, the detection accuracy begins to decrease. After several labeling, it is found that the labeled area has a certain impact on the training results. With the expansion of the labeled area, the detection accuracy of the training model is constantly improved. The identified over fitting phenomenon is decreasing with the expansion of the marked area, and it still exists when it expands in the same proportion to a certain range, as shown in Fig. 4:

Fig. 4 presents the detection results relationship between the marked area of small intestinal tumor and ulcer. It is the best to expand the labeled length and width to about 1.33 times the original lesions. If the lesions show irregular state, then the expansion area should be appropriately reduced or not expanded. After summarizing the labeling experiments of different areas for varied lesions, this rule is universal. Because the labeling of image is carried out by experts by

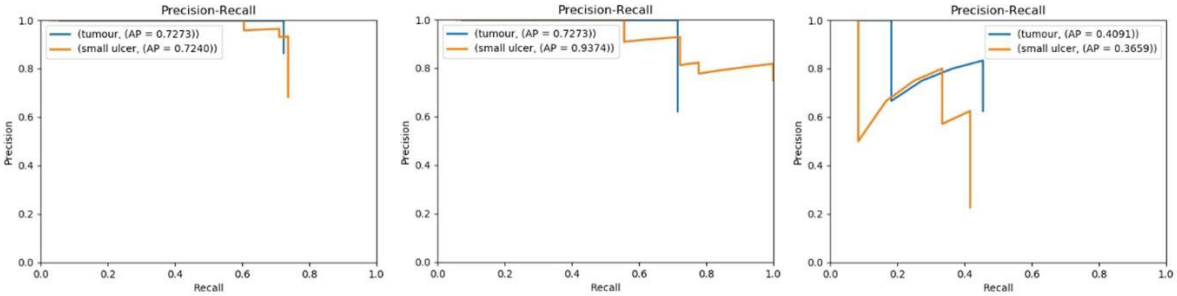


FIGURE 3. Comparison of training results with different labeling. (a) Only the lesion area. (b) 133% Expansion than the lesion area. (c) 200% Expansion than the lesion area.

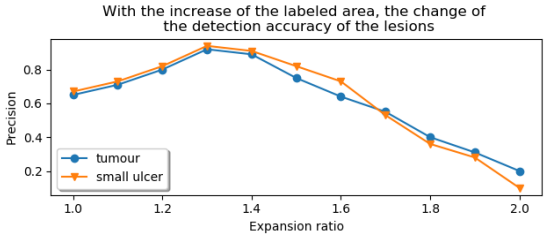


FIGURE 4. Relationship between labeled area and training results of small intestine lesion.

hand, it forms a convention. The research and experimental data of this paper are also based on this convention.

B. DETECTION NETWORK OF SMALL INTESTINE LESIONS BASED ON YOLOv3

The lesion detection based on WCE can be summed up as the target detection in machine vision. Before 2012, target detection mainly used manual features Haar, HOG, LBP, and machine learning methods AdaBoost, SVM, DPM [36], [37]. In 2012, Krizhevsky and Hinton [38] improved CNN to DCNN, and won the championship in the ILSVRC-2012 competition. CNN is valued again. Since then, the speed and accuracy of target detection are greatly improved based on CNN.

Small intestine lesion detection can use a variety of target detection networks that have been verified to be effective, including Faster R-CNN, Cascade R-CNN, SSD, YOLOv3, M2Det, etc. In the perspective of implementation, detection network can be divided into two parts: Basic Feature Extraction Network and Detection Network. According to different applications, the detection network can be replaced as a whole, or optimized to some extent. In the data set, the small intestine and small intestine lesions were divided into 14 types, including three types of normal small intestine sites (descending duodenum, duodenal bulb and normal small intestine) and 11 types of lesions. According to the above-mentioned methods, 3120 images were labeled manually. Due to the particularity and application of WCE, this paper selects the network which has outstanding performance in target detection in recent years for comparative study, as shown in Table 1 below:

TABLE 1. Results comparison of typical target detection network in small intestinal lesions.

Method	Train(COCO)		Train(Ours)	
	mAP	FPS	mAP	FPS
Faster R-CNN(VGG16)	59.1	6	90.2	7.1
Cascade R-CNN	60.9	7.1	90.5	7.5
SSD300	43.1	43	82.6	45
YOLOv3	57.9	19.8	89.6	21.3
M2Det	56.6	18	86.3	19.2

The data is divided into two groups: one is COCO data set, the other is self-labeled data set. In this paper, YOLOv3 is selected as the basic network of small intestine lesions detection and optimized. It is an important development direction of the target detection network that the pyramid feature model is constructed firstly, and the bottom feature map and the top feature map are effectively fused to achieve better detection effect.

The mainstay of YOLOv3 is Darknet53 to extract the features of the input image. The size of the input image is transformed into 416*416*3. Through a series of convolution processes, the input image features are extracted, and the height and width of the input image are continuously compressed. By continuous down sampling, the number of channels is increasing, and a series of feature layers are obtained. There are three kinds of output of the YOLOv3, which detect the large, medium and small objects, and the output sizes are 13*13*75, 26*26*75 and 52* 52*75 respectively. By continuous up sampling in network structure and stacking with the upper layer as the process of building pyramid, features are obtained and integrated.

The network sets three kinds of prior frames for each subsampling scale. Larger prior frames (116 × 90), (156 × 198) and (373 × 326) are applied on the smallest 13*13 feature map (with the largest receptive field), which are suitable for detecting larger objects. The medium prior frames (30 × 61), (62 × 45) and (59 × 119) are applied on the medium 26*26 feature map (medium receptive field), which

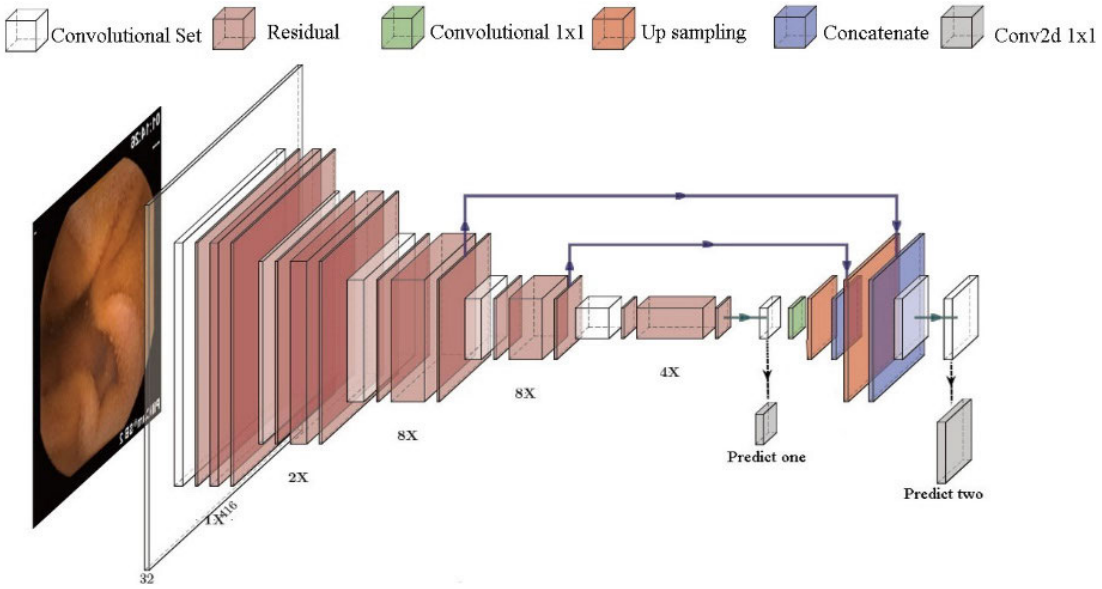


FIGURE 5. Improved detection network of small intestine lesions.

are suitable for detecting medium size objects. Smaller prior frames (10 × 13), (16 × 30), (33 × 23) are applied on the larger 52*52 feature map (smaller receptive field), which are suitable for detecting smaller objects. The total number of prior frames was 10627. The Yolo series networks use Kmeans algorithm to cluster three groups and 9 anchors totally. Different sizes objects will be assigned to different prediction layers by these three anchors, which is suitable for that objects detection in the scene is evenly distributed on the scale. If the anchor calculation strategy based on the default Kmeans algorithm is followed and most of the lesion areas are medium and large-sized objects, the remaining branches for detecting small-sized objects are not well trained, or there is no training at all, which wastes the network [39]. We use kmeans to cluster all the labeled bbox data according to the width and height, and distance = 1-iou. We get 9 anchors sizes needed in the Yolov3 [40], and the anchor scale is rounded to [55, 30 39, 33 28, 24 43, 38 54, 48 107, 80 172, 138 216, 187 288, 284]. The size of anchors is distributed in the range from medium to large objects in the original Yolov3 detection network, and the small objects detection part is rarely used. We classified the labeled bbox data into 6 categories according to the width and height, and obtained the scale of anchors [38, 31 48, 39 65, 49 105, 79 162,134 180, 285], and all the anchors scales were distributed in the original network for detecting medium and large objects. Based on the size features of anchors, the feature extraction network darknet53 of Yolov3 is retained, and the overall network structure is pruned to optimize Yolov3. The network structure is shown in Fig. 5:

According to the size of small intestine lesions, the labeled size of samples and the detection requirements, the output of small objects is removed. The output of the last layer of Darknet53 is convoluted (13, 13, 57), which is used to detect the medium and large target in the capsule endoscope image. The last layer is convoluted and three times of up sampling, (26, 26, 57) is obtained, which is used to detect the small target. Automatically generated prior frames only correspond to these two feature outputs, and the total number is reduced to 2535. The output layer is changed from three of YOLOv3 to two, and the shapes are (13, 13, 57), (26, 26, 57), respectively. The final dimension is 57 because of free small intestine lesion data set. In the original YOLOv3 framework, there are three prior frames for each feature layer, each prior frame has its own 19 parameter outputs, including 14 categories of small intestine digestive tract and lesions (3 normal digestive tract, 11lesions and 14 totally), plus 1output whether there is an object in the prior box, and 4 outputs adjusting the prior box parameters. The total outputs are 19, so the final dimension is 57.

The advantage of the network structure design is that after three times up-sampling feature fusion operations, we can make better use intermediate layer about the color and edge feature extraction in the feature extraction network. The detection accuracy and adaptability were improved, and the output was reduced in the detection phase of the network, and the prior frame is compressed to 1/4 of the original, but it can make more accurate prediction, improve the overall detection speed and reduce the utilization rate of video memory.

IV. EXPERIMENTS

Based on PillCam SB, 3120 normal and lesion images of small intestine were manually labeled. The data is from the First Hospital of Peking University, which is labeled by Dr. Suo and Dr. Li. Of the 3120 images, 620 were normal and 2500 were pathological. The number of training set,

verification set and test set is divided according to the conventional proportion, that is, 6:2:2.

Before training, it is necessary to adjust the ratio of positive and negative samples to improve the category imbalance. As for the problem of category imbalance, the paper takes the sample selected in this experiment as an example. Because the sample image is divided into positive and negative samples, the average distribution ratio of positive samples (Road) and negative samples (Non-Road) is 1:3. The network will undoubtedly focus on the feature learning of negative samples according to this ratio, thus reducing the recognition accuracy of positive samples [33], [34]. Because the random oversampling method is directly readopted for a few sample classes, there will be many duplicate samples in the training data set, which is easy to cause the model over fitting problem. Therefore, SMOTE method is taken, which randomly select a nearest neighbor $\widehat{x_i}$ (a sample of a few classes) from x_i(every minority sample), and then select a point on the line between x_i and $\widehat{x_i}$ as a few samples of new synthesis. For each neighbor $\widehat{x_i}$, according to the following formula to build a respectively new sample $\widehat{x_{new}}$.

$$\mathrm{x_{new}} = \mathrm{x_i} + \mathrm{rand}(0, 1) \times (\widehat{x_i} - \mathrm{x_i}) \tag{1}$$

For each x_i, it calculates distance of all samples with Euclidean as the standard to get k, nearest neighbor. According to sample imbalance ratio, the sample rate N is determined by setting sampling scale. For each minority sample x_i, several samples $\widehat{x_i}$ are randomly selected from their k-nearest neighbors.

The computer used in the experiment is equipped with a GTX 2080Ti graphics card, with Ubuntu 16.04 system and a detection network developed by the Python framework.1200 epochs trainings are conducted on the self-owned data, and the results are as follows:

In the Fig. 6, precision is the ratio that the actual number of positive samples accounts for the number of positive samples considered by the network. Recall is the ratio that real positive samples identified by the network accounts for the proportion of actual positive samples. F1 is the harmonic average of positioning Precision and Recall. AP is the area under the PR curve, and mAP is the average AP. Classification and val Classification represent the classification performance and classification test evaluation performance of the training process respectively. It can be observed from the graph that the performance of the method in the initial classification is not very good. With the increase of training times, the classification performance tends to be stable. At 950 epochs, the training results began to stabilize, and the highest value of mAP is 0.935. The test results are as follows:

It shows results that can recognize from large lesions to small lesions, and the detection frames does not only select the lesions area, but also includes the surrounding tissue of the lesions as well. Small continuous lesions are also tagged by the combination of big and small labeling. In the detection process, multiple lesion areas are merged to form a large

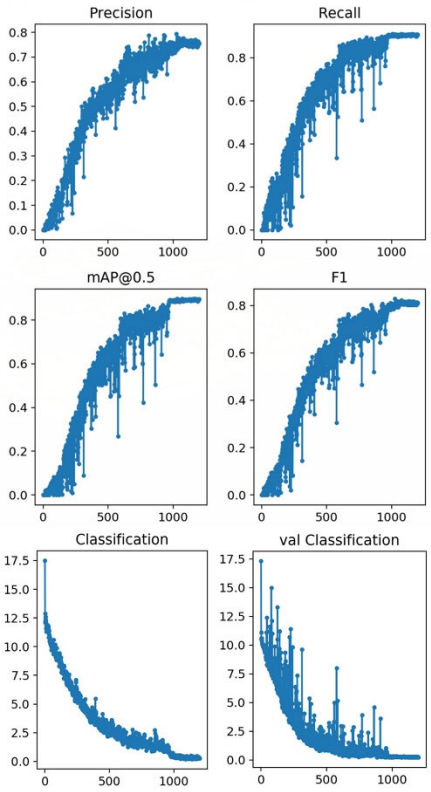


FIGURE 6. Training result of small intestine lesion.

lesion detection frame. Based on this labeling, the area of the network input size corresponding to the original size has been increased to more than 16*16. Therefore, in the network design, deleting the small target detection output, up sampling and fusing the feature layer is unified from data input to data output, which is used in the most important medium target detection output. From the results, the existing lesions detection reflects the advantages of labeling methods and network structure, the robustness has been enhanced, and over fitting phenomenon reduced.

In the existing model, some candidate frames of the target object are predicted. Generally, the cross area IOU between the candidate frames and the real value determines whether the frame is a positive sample, that is, the candidate frame to be retained. The disadvantage of setting to 0.7 is that it is inevitable to miss some candidate frames, especially small targets. At the same time, because the number of positive samples is too few, it is easy to over fit. In this paper, IOU is set to 0.5.

In the deep learning, the data samples used in the training process have a great impact on the training results, which is obvious for the single target features in simple background and is easy to learn in the training process, and the recognition rate is high [18]. However, due to the complexity of the background and the existence of many food residues and other interference factors, the feature learning is relatively difficult and the recognition rate is low, which needs a variety

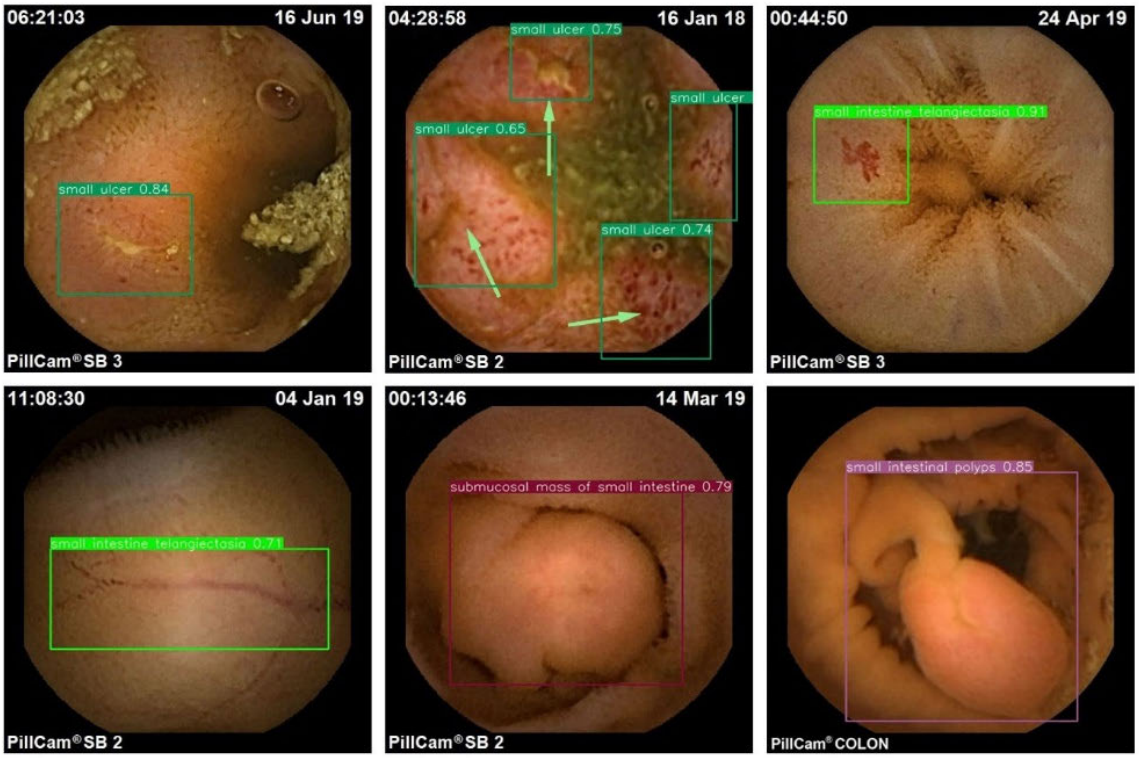


FIGURE 7. Detection results of small intestine lesions based on improved YOLOv3.

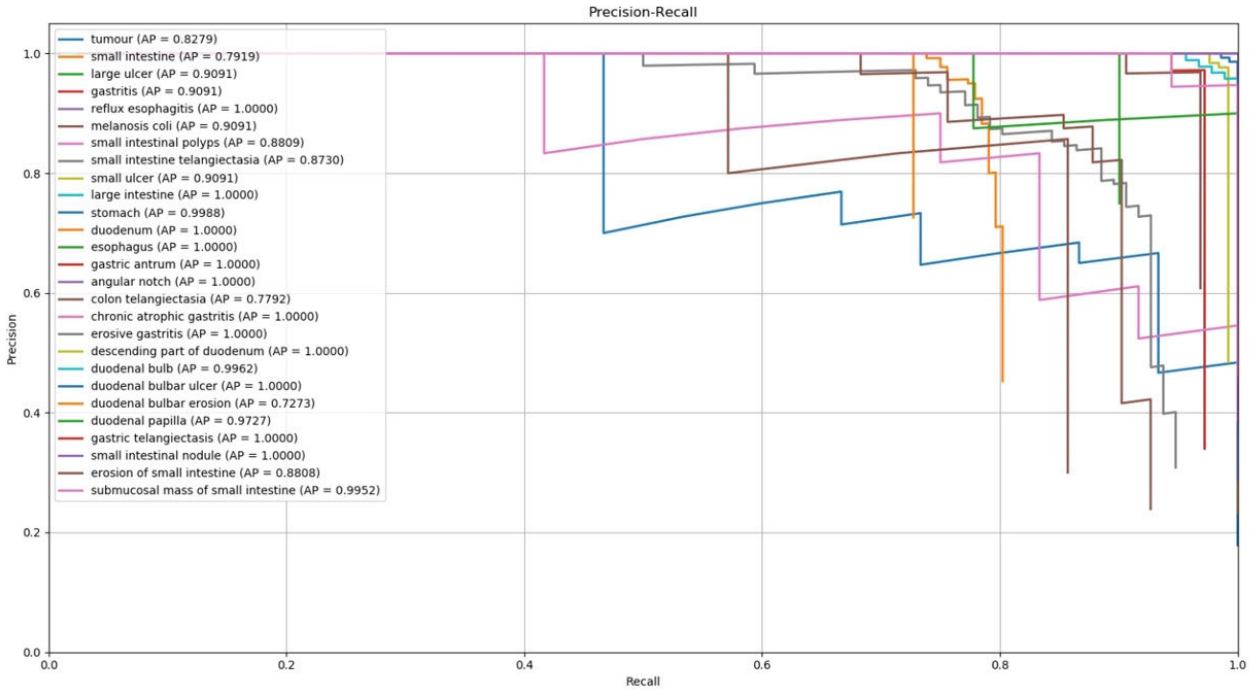


FIGURE 8. Training results after expanding data samples.

of samples and representative. Therefore, only the images of small intestine are not practical and the paper expands the training and test samples. It includes all images of digestive tract organs and related common lesions into the original training set. There are 27 categories totally and the results of training and testing are shown in Fig. 8:

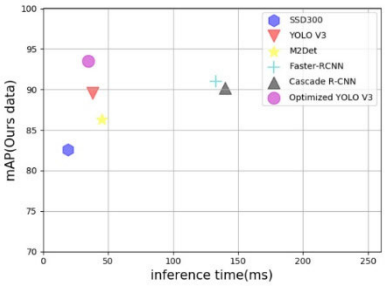


FIGURE 9. Results of improved network detection.

From precision recall, AP values of most categories are greater than 0.9, and three categories are lower than 0.8. The method is suitable for most lesions detection based on WCE images. As showing in Fig. 8, AP values of ''Duodenal Bulbar Erosion'', ''Colon Telangiectasia'' and ''Small Intestine'' are the lowest because their data are obviously less than other categories and are not rich enough.

V. CONCLUSION

In the case of the same hardware, the results are shown in Fig. 9, where data set is labeled by the paper.

Compared with YOLOv3, the image detection is simplified and the speed is improved. Because the improved network integrates the features of two intermediate layers, the edge and color extraction ability of the image has been significantly improved. Applicating this feature layer as the main lesion detection, combined with the existing labeling methods, the detection accuracy has also been improved by 4%. Compared with SSD model, the speed is slightly slower, but it has obvious advantages in detection accuracy. Compared with other network models, the speed and accuracy are obviously improved.

WCE usually takes pictures of the whole digestive tract at the speed of 2 frames per second, and a large number of images will be generated in the whole process, among which there are a large number of images with high similarity. In order to assist the film reading, we can filter out the images with high similarity through the algorithm. The Difference Hash Algorithm is the most sensitive in the experiment, compares the fingerprints of the two images and calculates the Hamming Distance, which is how many bits value are different between two 64 bit hashes. The fewer the different bits, the more similar the picture. The best filtering is about 17, which can effectively filter out 50% without affecting the reading quality.

This paper proposed an improved YOLOv3 for small intestine lesion detection and an image manual labeling method by fusing basic feature of intermediate layer and deleting the output of detected small size feature, where the lesion detection ability of medium size feature is strengthened, so that the detection can be successfully trained. The model effectively solves the problem that the general detection network fails to make full use of the intermediate layer and the detection speed is slow. The experimental shows that the method of artificial labeling improves the adaptability of detection as well as the optimized detection network, and is equally excellent in the detection of small-sized objects. Compared with the current mainstream target detection network, its extracted features are more recognizable and have higher accuracy on their own data sets. In the next research, we can continue to increase the amount of image data, further improve the detection accuracy and reduce the over fitting of detection results. In addition, it can also be applied to the detection of lesions in the whole digestive tract, improve the applicability and the computer-aided clinical technology.

REFERENCES

[1] L. Rebecca Siegel, D. Kimberly Miller, and A. Jemal, ''Cancer statistics,'' *CA: A Cancer J. Clinicians*, vol. 1, Jan. 2016.

[2] J.-Y. Yeh, T.-H. Wu, and W.-J. Tsai, ''Bleeding and ulcer detection using wireless capsule endoscopy images,'' *J. Softw. Eng. Appl.*, vol. 7, no. 5, pp. 422–432, 2014.

[3] V. S. Charisis, L. J. Hadjileontiadis, C. N. Liatsos, C. C. Mavrogiannis, and G. D. Sergiadis, ''Capsule endoscopy image analysis using texture information from various colour models,'' *Comput. Methods Programs Biomed.*, vol. 107, no. 1, pp. 61–74, Jul. 2012.

[4] 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.

[5] P. Szczypiński, 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.

[6] B. Li and M. Q.-H. Meng, ''Wireless capsule endoscopy images enhancement via adaptive contrast diffusion,'' *J. Vis. Commun. Image Represent.*, vol. 23, no. 1, pp. 222–228, Jan. 2012.

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

[8] 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.

[9] T. Zhou, G. Han, B. N. Li, Z. Lin, E. J. Ciaccio, P. H. Green, and J. Qin, ''Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method,'' *Comput. Biol. Med.*, vol. 85, pp. 1–6, Jun. 2017.

[10] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, ''Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images,'' *Sensors*, vol. 19, no. 6, p. 1265, Mar. 2019.

[11] 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 Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1518–1523.

[12] G. Xie and L. Wang, ''Periodic stabilizability of switched linear control systems,'' *Automatica*, vol. 45, no. 9, pp. 2141–2148, Sep. 2009.

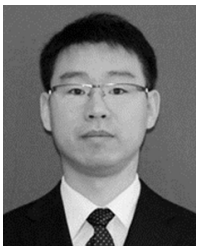
[13] X. Jia and M. Q.-H. Meng, ''A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images,'' in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 639–642.

[14] 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.*, Columbus, OH, USA, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[15] 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, doi: 10.1109/TPAMI.2015.2389824.

[16] R. Girshick, F. Iandola, T. Darrell, and J. Malik, ''Deformable part models are convolutional neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 437–446, doi: 10.1109/CVPR.2015.7298641.

[17] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," 2018, *arXiv:1803.00676*. [Online]. Available: http://arxiv.org/abs/1803.00676
[18] S. Li, B. Shi, Y. Liu, L. Zhang, and X. Wang. *Multi Vehicle Target Recognition Method Based on Improved YOLOv2 Model*. Progress in Laser and Optoelectronics. [Online]. Available: http://kns.cnki.net/kcms/detail/31.1690.TN.20191106.1156.022.html
[19] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5562–5570, doi: 10.1109/ICCV.2017.593.
[20] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," 2018, *arXiv:1809.08545*. [Online]. Available: http://arxiv.org/abs/1809.08545
[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-02.
[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[25] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "RON: Reverse connection with objectness prior networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5244–5252, doi: 10.1109/CVPR.2017.557.
[26] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: http://arxiv.org/abs/1701.06659
[27] T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, and T. Tada, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357–363, Feb. 2019.
[28] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.
[29] M. Alizadeh and O. Haji at al, "Detection of small bowel tumor in wireless capsule endoscopy images using an adaptive neuro-fuzzy inference system," *J. Biomed. Res.*, vol. 31, no. 5, pp. 419–427, 2017.
[30] L. Yu and P. C. Yuen, and J. Lai, "Ulcer detection in wireless capsule endoscopy images," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012.
[31] O. H. Maghsoudi, M. Alizadeh, and M. Mirmomen, "A computer aided method to detect bleeding, tumor, and disease regions in wireless capsule endoscopy," in *Proc. IEEE Signal Process. Med. Biol. Symp. (SPMB)*, Dec. 2016, pp. 1–6.
[32] X. Wu, H. Chen, and T. Gan, "Automatic hookworm detection in wireless capsule endoscopy images," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1741–1752, Jul. 2016.
[33] J. Yu, L. Wang, and G. Xie, "Disturbance rejection of switched systems subject to actuator saturation," *Trans. Inst. Meas. Control*, vol. 32, no. 6, pp. 607–616, 2011.
[34] K. Liu, G. Xie, and L. Wang, "Consensus for multi-agent systems under double integrator dynamics with time-varying communication delays," *Int. J. Robust Nonlinear Control*, vol. 22, no. 17, pp. 1881–1898, 2011.
[35] C. Wang, G. Xie, L. Wang, and M. Cao, "CPG-based locomotion control of a robotic fish: Using linear oscillators and reducing control parameters via PSO," *Int. J. Innov. Comput. Inf. Control*, vol. 7, no. 7B, pp. 4237–4249, 2011.
[36] S. M. S. Erfani, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognit.*, vol. 58, pp. 121–134, Oct. 2016.
[37] A. J. Wyner and M. Olson, "Explaining the success of adaboost and random forests as interpolating classifiers," *J. Mach. Learn. Res.*, vol. 18, no. 48, pp. 1–33, 2017.
[38] A. Krizhevsky and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, no. 2, pp. 1097–1105.
[39] P. Hurtik, V. Molek, J. Hula, M. Vajgl, P. Vlasanek, and T. Nejezchleba, "Poly-YOLO: Higher speed, more precise detection and instance segmentation for YOLOv3," 2020, *arXiv:2005.13243*. [Online]. Available: http://arxiv.org/abs/2005.13243
[40] L. Shuo and G. Yuhai, "Detection method of illegal vehicles based on optimized YOLOv3 algorithm," *J. Chongqing Univ. Technol. (Natural Sci.)*, to be published. [Online]. Available: http://kns.cnki.net/kcms/detail/50.1205.t.20200318.1917.006.html



ZHIGUO XIAO was born in Changchun, China, in 1977. He has presided over four provincial projects, participated in six national and provincial projects, published more than ten articles, and holds five patents for utility models and seven software copyrights. He has made great achievements in image processing and virtual reality technology. He has made deep research both in theory and technology. At the same time, he was deeply loved by his classmates during his school life. His main research interests include artificial intelligence and campus security technology.

LI NIAN FENG is the Dean of the Computer Institute of Changchun University, the Secretary-General of the Jilin Provincial Computer Professional Teaching Steering Committee, the Director of the Jilin Provincial Computer Society, a member of the China Computer Society, and the head of provincial experimental teaching demonstration. He has hosted or participated in more than 20 provincial and ministerial scientific research projects, published more than 20 articles on research results, holds 14 licensed patents and six software copyrights, and received one Second Prize of the Jilin Provincial Science and Technology Award and three academic achievement awards for natural sciences in Jilin Province. His main research interests include machine learning, image processing, and campus security technology.