Comparative analysis and application of deep learning polyp detection in colonoscopy

1st Bing Sun
School of Logistics Engineering
Shanghai Maritime University
Shanghai, China
hmsunbing@163.com

2nd Wei Zhang
School of Electrical Engineering
Shanghai Dianji University
Shanghai, China
viweizhang@163.com

3rd Xiaoyu Xie
School of Logistics Engineering
Shanghai Maritime University
Shanghai, China
1174688476@qq.com

2022 China Automation Congress (CAC) | 978-1-6654-6533-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CAC57257.2022.10055481

***Abstract*—Colonoscopy is the most common method for early diagnosis of colorectal cancer. The traditional detection method mainly relies on manual detection by doctors, and the detection efficiency is low. By contrast, the deep learning method has high detection precision and accuracy, which is an ideal tool for the disease detection of colonoscopy. Based on this, this paper compares the six models of YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, Faster RCNN with and without "fpn" structure, and finds out the most effective colonoscopy polyp detection among the six models. Comparing and integrating various factors, it is found that the YOLOv5l model has the best comprehensive detection effect, and its mAP value is generally higher than 98.5%, which is the largest mAP value among the above six models, and the detection time is relatively short, about 0.018s per image. The balance of accuracy and rapidity in colonoscopy can be achieved.**

***Index Terms*—Colonoscopy image, polyps detection, deep learning, YOLOv5, Faster RCNN**

I. Introduction

Colorectal cancer is the general name of colon cancer and rectum cancer, which frequently occurs in the rectum and colon, is a very common digestive tract malignant tumor[1]. According to research data in 2020, the incidence and fatality rate of colorectal cancer in China ranked second and third among all malignant tumors.

The early stage of colorectal cancer is in the form of polyps, which can be seen in Figure 1. Timely detection and removal of polyps is the most effective treatment for early colorectal cancer. And with the clinical application and popularity of the intestinal endoscopy examination technology [2-3], intestinal endoscopy examination become one of the most main means of polyp diagnosis, the detection means to a large extent depends on the doctor itself, so there is a certain extent, the possibility of residual and the fault detection. On the other hand, intestinal endoscopy generally requires the cooperation of more than one doctor, which has low detection efficiency and high detection cost. It is difficult to be widely popularized, which is not conducive to the early diagnosis and treatment of colorectal cancer patients.

This work was supported in part by the National Natural Science Foundation of China under Grant 61873161, 52271321, Shanghai Rising-Star Program under Grant 20QA1404200 and Natural Science Foundation of Shanghai under Grant 22ZR1426700.

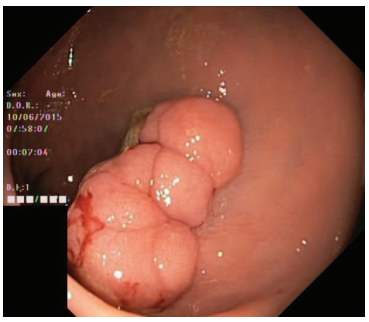


Fig. 1: Polyp image.

In 1959, Ledley et al. introduced Bayes theorem and Boolean algebra into the clinical diagnosis of lung cancer for the first time, and computers began to enter the medical field. Ledley et al. pioneered computer-aided diagnosis. With the development of computer software and hardware, electronic engineering, mathematics, statistics and other related contents, computer-aided diagnosis technology has also been promoted to a large extent, especially in the field of medical imaging [4-5].

Qadir et al. [6] divided the experimental work into two stages and proved that two-way time information was helpful to estimate the location of polyps and accurately predict FPs. Compared with traditional false positive learning methods, this method provides an overall performance improvement in sensitivity, accuracy and specificity, thus achieving better results on the video data set. Ciobanu et al. [7] successfully detected 10 categories of objects in real-time playback of colonoscopy video using the retrained Mobilenet deep learning model. Kwon et al. [8] proposed a weakly supervised learning method for histological localization by training two different types of data sets. The accuracy of histological classification and lesion location was improved by this method. Kora et al. [9] proposed the U-NET model and applied it to the

combination of data enhancement and patch extraction in colonoscopy images to detect polyps. In addition, the model is scalable and adaptable, and can be upgraded to other forms of disease detection models in the future.

This paper mainly analyses six models, namely YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, Faster RCNN with and without "FPN" structure, to extract features from colonoscopy image data sets. The new weights obtained from training were applied to colonoscopic polyp recognition. By comparing the obtained mAP value, test time, and colonoscopic polyp recognition effect, the best comprehensive effect of the six models in colonoscopic polyp recognition was determined.

II. Target Detection Algorithm Based on Colonoscopy Image

In this paper, the newer detection algorithms YOLOv5 and Faster RCNN in "single-stage detection" and "two-stage detection" are respectively selected to train the existing colonoscopy image data set and record the obtained data. Several models used were compared to explore the most suitable model for colonoscopy image detection.

A. Single-stage target detection–YOLOv5

YOLOv5 series algorithm is a relatively new algorithm in the current YOLO series algorithm. It adds some new improvement ideas based on YOLOv4, which greatly improves its calculation speed and accuracy, and provides an excellent model for the subsequent training of our own data set. In the official code of YOLOv5, four versions of the target detection network are presented, namely, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Among the four models, YOLOv5s model has the smallest network, the fastest speed and the lowest AP accuracy. The other three networks, based on YOLOv5s, have been deepened and widened, and AP accuracy has improved, but speed has also slowed. Since the same network has different training effects on different data sets, we still need to train colonoscopy images to obtain the most suitable training model and training weight for colonoscopy image detection.

This section will take YOLOv5s as an example to briefly introduce the improvement of YOLOv5 over previous YOLO series. The differences among the four YOLOv5 models are also introduced. Fig. 2 shows the network structure diagram of YOLOv5s [10]. From the structure diagram, it can be seen that YOLOv5s network consists of four parts: input, Backbone, Neck and output.

The backbone of YOLOv5 integrates some new ideas in other detection algorithms, mainly including Focus structure and CSP structure. Focus structure mainly plays the role of slicing. Take YOLOv5s as an example. After the image of 608*608*3 is input into the Focus structure, Focus structure will slice it to obtain the feature map of 304*304*12, and the feature map of 304*304*12 will undergo convolution operation of 32 convolution kernels. Eventually it becomes 32 feature maps of 304*304. Different from YOLOv4, YOLOv5 not only uses the original CSP1_X structure in the Backbone

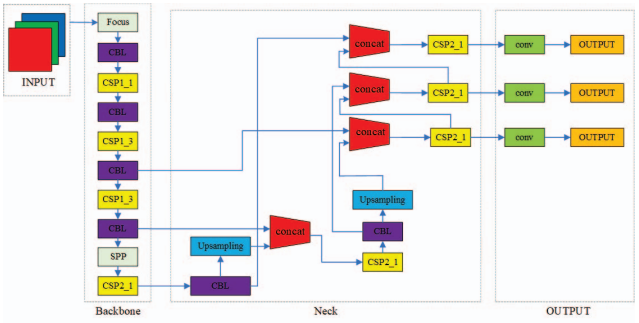


Fig. 2: YOLOv5s network structure diagram.

network, but also uses another CSP structure - CSP2_X in Neck. The network structure diagram of CSP1_X structure and CSP2_X structure is shown in Figure 3.

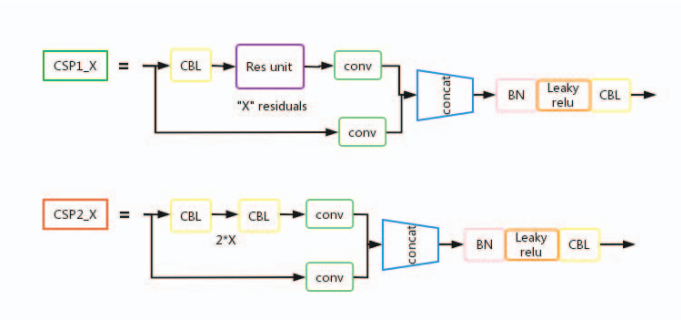


Fig. 3: Network structure diagram of CSP structure.

The Neck structure of YOLOv5 is the same as that of YOLOv4, both of which adopt FPN+PAN structure. The difference is that the Neck structure of YOLOv4 only adopts ordinary convolution operation, while the Neck structure of YOLOv5 uses the CSP2_X structure to enhance the ability of network feature fusion.

At the output end, YOLOv5 mainly improves the loss function of Bounding Box into CIOU_Loss during training. CIOU_Loss takes into account three important geometric factors that should be considered by the regression function of the target frame: overlap area, center distance and aspect ratio, which greatly improves the speed and accuracy of the regression function of the target frame. CIOU_Loss calculation formula is shown in Formula 1.

$$CIOU\_\text{Loss} = 1 - \left(IOU - \frac{\text{Dis}2^2}{\text{Dis}C^2} - \frac{v^2}{(1-IOU)+v} \right) \quad (1)$$

where IOU represents the intersection coefficient of the target box and the prediction box, Dis2 represents the Euclidean distance between the target frame and the center point of the prediction frame. DisC represents the diagonal distance between the target box and the smallest enclosing rectangle of the prediction box. V is a parameter to measure the consistency of the aspect ratio of the target frame and the

prediction frame, and the calculation formula is shown in Formula 2 below:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{p}}{h^{p}}\right)^2 \quad (2)$$

The difference between the four YOLOv5 network structures mainly lies in the depth and width of the network structure. The control of network structure depth is realized by CSP structure. There are two kinds of CSP structures in YOLOv5, CSP1_X and CSP2_X, among which CSP1_X is used in the Backbone and CSP2_X is used in Neck. The depth of each CSP structure in the four network structures is different. In Backbone, the first CSP1 of YOLOv5s uses a residual component, which is CSP1_1. In the first CSP1 of YOLOv5m, two residual components are added based on the original residual component, which increases the depth of the network, so it is CSP1_2. In YOLOv5l, a residual component is added at the same position, and in YOLOv5x, a residual component is added further. The principle of the second and third CSP1 is the same. In Neck, the first CSP2 structure of YOLOv5s uses a group of convolution, so it is CSP2_1. However, two groups of convolution are used in YOLOv5m, three groups are used in YOLOv5l and four groups are used in YOLOv5x, so they are CSPP2_2, CSP2_3 and CSP2_4 respectively. The rest of the CSP2 structure works the same way. It can be seen that the four network structures of YOLOv5 are constantly deepening and increasing the capability of network feature extraction and feature fusion.

B. Two-stage target detection–Faster RCNN

In 2016, Ross B. Girshick proposed Faster RCNN based on RCNN and Fast RCNN. Structurally, the improvement of the overall performance of Faster RCNN, especially detection speed, is largely attributed to its integration of feature extraction, two-stage target detection proposal extraction, bounding box regression, and classification into a network. Its advantages are mainly reflected in the realization of an end-to-end target detection framework and the improvement of detection accuracy and speed, and it only takes about 10ms to generate the suggestion box.

The working steps of Faster RCNN are mainly divided into three steps. Firstly, the selective search method is used to generate 1000 to 2000 candidate frames on an image. Secondly, the image is input into the network for the calculation to obtain the corresponding feature graph, and the candidate box generated by the SS algorithm is mapped to the feature graph to obtain the corresponding feature matrix. Finally, each feature matrix is scaled to a 7*7 feature map through the ROI pooling layer, and then the feature map is flattened out through a series of fully connected layers to obtain the prediction results.

ResNet50 network was used for feature extraction of the Faster RCNN model. Compared with other networks, the ResNet network has three breakthroughs. First, it has a super deep network structure. Second, the residual module is presented. The structure of residual structure is shown in Figure 4 below. Third, the Batch Normalization method has been used to accelerate training instead of the dropout method. The purpose of Batch Normalization is to make the number of feature maps meet the distribution rule of 0 mean and 1 variance, which improves the convergence speed and accuracy of the network.

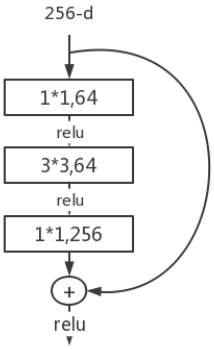


Fig. 4: Schematic diagram of residual structure.

III. Colonoscopy image training and testing

A. Preparation of colonoscopy image data set

The adoption of online under the white light 756 colonoscopy image data set on the depth of two learning neural network training, including the definition of high and low resolution image complete with polyps, polyp of incomplete, polyp size larger and small polyp, intestinal environment clean and intestinal environment of more general. It provides as many situations as possible for neural network learning, ensures that the trained neural network can deal with more situations, assist doctors in the examination, and reduce the rate of missed detection.

B. The markup of the data set

After learning the knowledge of colonoscopy polyps through consulting relevant materials, labelimg was used to label all the images in the collected data set, as shown in Fig. 5. The left is the unlabeled image and the right is the labeled image.

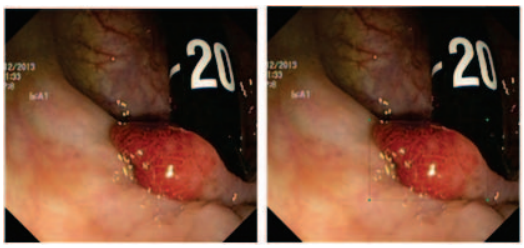


Fig. 5: Comparison before and after image labeling.

TABLE I: Environment configuration for YOLOv5

Environment configuration	Version Requirements
graphics card	NVIDIA GeForce RTX2080Ti
Python Version	3.8
Pytorch Version	1.7.1
CUDA Version	11
Support packages	matplotlib $\geq$ 3.2.2 numpy $\geq$ 1.18.5 opencv-python $\geq$ 4.1.2 Pillow PyYAML $\geq$ 5.3.1 scipy $\geq$ 1.4.1 torchvision $\geq$ 0.8.1 tqdm $\geq$ 4.41.0

C. Environment configuration

Table 1 shows the environment configuration for YOLOv5.

A Faster RCNN environment configuration with "FPN" Compared to the YOLOv5 environment configuration, the Faster RCNN environment configuration with "FPN" has only a few different support packages required for the program to run, so just add the missing support packages on top of the existing environment.

B. Faster RCNN environment configuration without "FPN" Because the PyTorch version required for Faster RCNN without "FPN" is not the same as the PyTorch version required for Faster RCNN with "FPN", it needs to recreate an environment with PyTorch version 1.10.0.

D. Neural network model training for colonoscopy detection

1) YOLOv5 model training and validation: The process of training YOLOv5 series models is shown in Fig. 6. Take YOLOv5s for example.

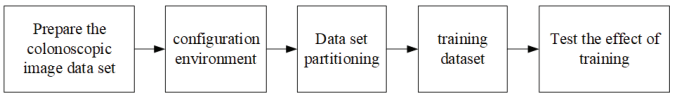


Fig. 6: YOLOv5 series training and test flow chart.

Step 1: Prepare the colonoscopic image data set.

Step 2: Configure the environment.

Step 3: The division of data set.

Step 4: Train the data set. The epochs is set as 100. The batch-size is set as 4. The workers is set as 4. After setting, we can train the YOLOv5s model.

Step 5: Test the training effect.

The training process of other YOLOv5 models is similar to that of YOLOv5s, only the weight file path initialized in step 4 and the model file path need to be replaced with corresponding. Modify the corresponding model file and continue the training. Repeat the fifth step to copy and paste the best weights into "weights". Finally, change the path of the weight to verify the testing effect of the new model.

2) Faster RCNN model training and validation: A. Contains the FPN module

The general steps are conducted as follows.

Step 1: Data set processing.

Step 2: Environment configuration.

Step 3: Data set division.

Step 4: Train the data set. Set the number of "detection target category number (excluding background)" to 1. Set the "total epoch number of training" to 100 rounds. Set the value of "Batch Size for training" to 4. Once set up, train the Faster RCNN network model.

Step 5: Test the training results. We can select the best training weight according to the training effect.

B. Not including "FPN" module

The process is basically the same, just take care of the environment configuration.

IV. Comparison and analysis of training results

A. Common indicators

1) learning rate: In deep learning and statistics, the learning rate refers to the tuning parameter in the optimization algorithm, which can determine the step size of each iteration so that the loss function converges to the minimum value. When the learning rate is too small, the function converges too slowly, and we need to spend a lot of time finding the minimum value. However, when the learning rate is too high and the step size is too large, the convergence speed will be too fast and the minimum value will be missed, and even the oscillation of the function will diverge.

2) mAP: The full name of mAP is Mean Average Precision, which is the Average of all categories of Average accuracy (AP). It is mainly related to the following parameters:

(1) True positives (TP): when IoU $>$ 0.5, the number of test enclosures was correctly classified as positive examples when the coincidence rate between the predicted enclosures and the real enclosures was high.

(2) False positives (FP): Refers to the number of False positives when IoU $\leq$ 0.5, or the number of extra enclosures that can detect the same Ground Truth. In other words, when the coincidence rate between the predicted enclosures and the real enclosures is low, they are wrongly classified as positive examples.

(3) False negatives (FN): The number of undetected Ground truths, that is, the number incorrectly classified as negative cases.

(4) Precision: Precision refers to the ratio of the actual number of positive samples in the predicted sample to the number of all positive samples detected, namely:

$$P = \frac{TP}{TP + FP} \tag{3}$$

(5) Recall: Recall refers to the ratio of the actual number of positive samples in the predicted samples to the number of all tested samples, namely:

$$R = \frac{TP}{TP + FN} \tag{4}$$

Among them, precision and recall can form a P-R curve. By calculating the area under the curve, the average accuracy can be calculated, and then the mAP can be obtained by averaging

it. In the process of program output, this is shown in Figure 7, where the second line refers to the AP value calculated when IoU is 0.5. It is the Pascal VOC evaluation index adopted in this paper.

```
Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.290
Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.545
Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = 0.278
Average Precision  (AP) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.000
Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.107
Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.438
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=  1 ] = 0.228
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 10 ] = 0.366
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.386
Average Recall     (AR) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.063
Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.226
Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.525
```

Fig. 7: COCO Evaluation Result.

B. The training results

After repeated tests, the optimal parameters for the training effect were selected as shown in Table 2 below. The total mAP value obtained by YOLOv5 series training is shown in Fig. 8.

TABLE II: Setting parameters

parameter	set value
batch-size	4
epoch	100
Ir(Initial learning rate)	0.005

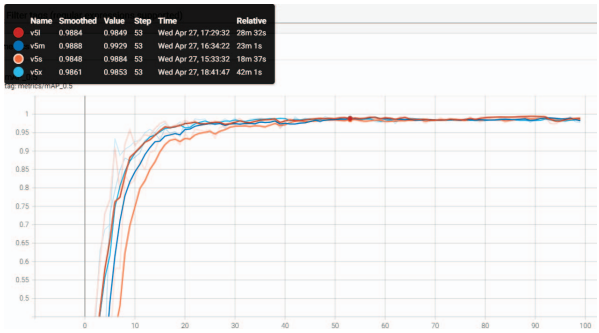


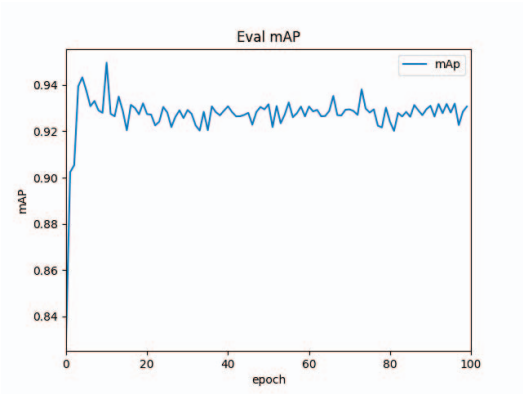
Fig. 8: Total mAP results obtained by YOLOv5 series.

As can be seen from the comparison figure, YOLOv5l model ranks second in reaching stability speed among the four models, and the stable value of mAP is generally the highest, which is above 98.5% and up to 99.4%. MAP values of other networks are as follows: YOLOv5s network has the slowest stabilizing speed, which is generally higher than 96.5% and the maximum value is about 98.9%. The stability speed of YOLOv5m network ranked third, which was generally higher than 98.0% and the maximum value was about 99.1%. YOLOv5x network has the fastest stabilization speed, which is generally higher than 98.2% after stabilization, and the maximum is about 98.9%. It can be seen from Figure 9 that 100 rounds of training consume time: YOLOv5x > YOLOv5l > YOLOv5m > YOLOv5s.

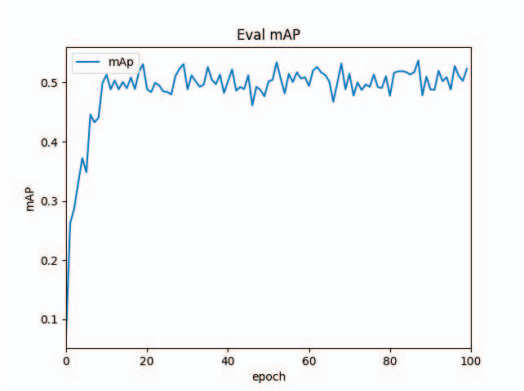
Figure 10(a) is the mAP curve obtained from 100 training rounds of Faster RCNN with ”FPN” structure. It can be seen

Fig. 9: YOLOv5 series training time consumed in 100 rounds.

from the figure that the mAP value of the Faster RCNN network containing ”FPN” structure is generally higher than 92.0%, and the maximum value is about 95.0%. Figure 10(b) is the mAP obtained after 100 rounds of training for Faster RCNN without ”FPN” structure. As can be seen from the figure, the mAP value of the Faster RCNN network without ”FPN” structure is generally lower than 52%, and the highest value only reaches 53.7%.



(a) mAP curve with ”FPN” structure



(b) mAP curve without ”FPN” structure

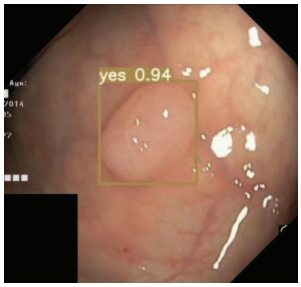
Fig. 10: mAP curve of Faster RCNN.

In summary, the YOLOv5l test results are generally the best. However, the Faster RCNN network mAP without ”FPN” structure is too low to meet the requirements completely, so it can be excluded.

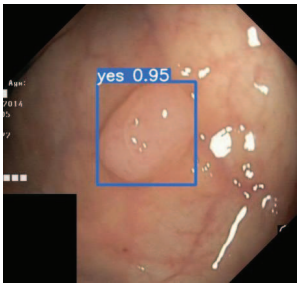
C. Simulation results

Since the mAP value of the Faster RCNN network without ”FPN” structure was too low, we did not test it. Figure 11 and Figure 12 below show the test results of the same polyp

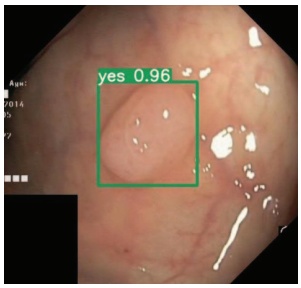
image by the YOLOV5 series and Faster RCNN method. And the test time is sorted out as shown in Table 3.



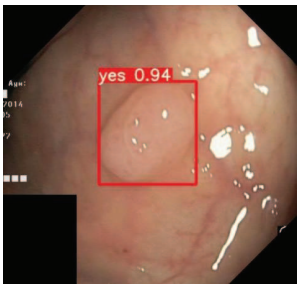
(a) YOLOv5s Test Results



(b) YOLOv5m test results



(c) YOLOv5l test results



(d) YOLOv5x test results

Fig. 11: Test results of YOLOV5 series.

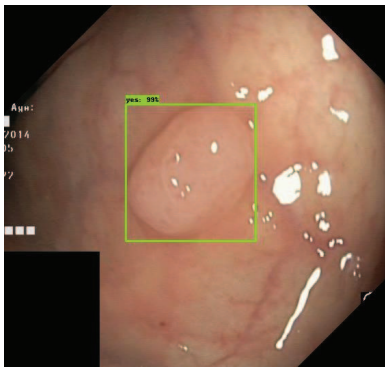


Fig. 12: Test results of Faster RCNN.

TABLE III: Comparison of test results

network model	Test accuracy	Test time(s)
YOLOv5s	0.94	0.013
YOLOv5m	0.95	0.016
YOLOv5l	0.96	0.018
YOLOv5x	0.94	0.020
Faster RCNN	0.99	0.0348

As can be seen from the above table, the Faster RCNN model has the highest test accuracy but the longest test time, while the YOLOv5l model has 0.03 lower accuracy than the Faster RCNN model, but the test time is only half of Faster RCNN model, and YOLOv5l model has the largest mAP value. To take into account that a long detection time may bring physical discomfort to patients during colonoscopy, YOLOv5l model with the largest mAP value, relatively short test time and relatively high test accuracy was selected as the most suitable target detection model for colonoscopy among the models used in this experiment.

V. Conclusion

In this paper, six neural network models: YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, and Faster RCNN with and without FPN structure are analyzed and tested for intestinal disease detection. Considering the feasibility of the operation and the patient's experience, the YOLOv5l model was selected as the most suitable target detection model for colonoscopy. The neural network effects of this comparative experiment are all good, and the following factors can be considered in later research: (1) The combination of other updated neural network models and colonoscopy image detection, such as the Mask RCNN model and YOLOx model, can be considered. (2) Collect more colonoscopy images and continuously expand the data set. (3) Determine the stage of colorectal cancer by recognizing the size, surface condition and intestinal bleeding of polyps.

References

[1] W. Wang, P. Yin, Y. Liu, J. Liu, L. Wang, J. Qi, J. You, L. Lin, S. Meng, F. Wang and M. Zhou, "Mortality and years of life lost of colorectal cancer in China, 2005-2020: findings from the national mortality surveillance system," Chinese Medical Journal. 2021, vol. 134, pp. 1933–1940.

[2] Z. Xiao and L. N. Feng, "A Study on Wireless Capsule Endoscopy for Small Intestinal Lesions Detection Based on Deep Learning Target Detection," IEEE Access, vol. 8, pp. 159017-159026, 2020.

[3] S. Yang, C. Lemke, B. F. Cox, I. P. Newton, I. Näthke and S. Cochran, "A Learning-Based Microultrasound System for the Detection of Inflammation of the Gastrointestinal Tract," IEEE Transactions on Medical Imaging, vol. 40, no. 1, pp. 38-47, Jan. 2021.

[4] A. J. Larrazabal , N. Nieto, V. Peterson, D. H. Milone, E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," Proceedings of the National Academy of Sciences, 2021, vol. 117, pp. 12592–12594.

[5] Y. Yan, X. J. Yao, S. H. Wang, and Y.-D. Zhang, "A Survey of Computer-Aided Tumor Diagnosis Based on Convolutional Neural Network," Biology, vol. 10, no. 11, p. 1084, Oct. 2021, doi: 10.3390/biology10111084.

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

[7] A. Ciobanu, M. Luca, T. Barbu, V. Drug, A. Olteanu and R. Vulpoi, "Experimental Deep Learning Object Detection in Real-time Colonoscopies," 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2021, pp. 1–4.

[8] J. Kwon and K. Choi, "Weakly Supervised Attention Map Training for Histological Localization of Colonoscopy Images," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Mexico, 2021, pp. 3725–3728.

[9] P. Kora, B. Haneesha, D. Sahith, S. P. Grace, K. Benny Jasper, K. Swaraja and K. Meenakshi, "Automatic Segmentation of Polyps using U-Net from Colonoscopy images," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 855–859.

[10] C. Li, Y. Cao and Y. Peng, "Research on Automatic Driving Target Detection Based on YOLOv5s," Journal of Physics: Conference Series, 2022, vol. 2171, 012047.