scientific reports

Check for updates

OPEN

A colonic polyps detection algorithm based on an improved YOLOv5s

Jianjun Li[1,2], Jinhui Zhao[2]✉, Yifan Wang[1], Jinhui Zhu[3], Yanhong Wei[1], Junjiang Zhu[1], Xiaolu Li[1], Shubin Yan[2] & Qichun Zhang[4]

Colon cancer is a prevalent malignancy, substantially it prevented most effectively from killing patients through early endoscopic detection. With the rapid development of artificial intelligence technology, the early diagnosis rate of colonic polyps achieves greater clinical efficacy for colon cancer by applying target detection algorithms to colonoscopy images. This paper presents two outcomes achieved through the application of the improved YOLOv5s algorithm with annotated microscopy images of clinical cases and publicly available polyp image data: (1) enhancement of the C3(Cross Stage Partial Networks) module with multiple layers to C3SE(Cross Stage Partial Networks with Squeeze-and-Excitation) via the attention mechanism SE (squeeze-and-excitation) and (2) fusion of higher-level features utilizing BiFPN (the weighted bi-directional feature pyramid network). Experimental comparisons are performed based on a new image dataset of colonic polyps among more than 6 target detection algorithms to validate the better detection capability. The tests indicate that the YOLOv5s + BiFPN and YOLOv5s-1st-2nd-C3SE models exhibit enhancements of detection capability compared to the YOLOv5 algorithm according to the main indicators of the mAP, accuracy, and recall. The YOLOv5s + SEBiFPN model demonstrate a substantial improvement over the YOLOv5s algorithm, and establishing a benchmark technology for advancing computer-assisted diagnostic systems is feasible.

Keywords Target detection, Feature fusion, Digestive endoscopy, Convolutional neural network

It has been reported that China has a high incidence of cancer, particularly colorectal cancer, which ranks among the highest in the world. Symptoms include indigestion, abdominal pain, nausea, vomiting, loss of appetite, weight loss, and bloody stools. In 2020 alone, colorectal cancer ranked third in incidence and second in mortality worldwide. That same year, colorectal cancer was the third most common cancer in China, with approximately 555,000 new cases, representing a 7.4% increase compared to the previous year[1]. Colorectal cancer generally progresses through four stages, from early to advanced. In the first stage, known as early colorectal cancer, patients have a survival rate of over 90% with timely treatment. The second and third stages are considered intermediate colorectal cancer, with survival rates ranging from 50 to 70%. By the fourth stage, or advanced colorectal cancer, the survival rate drops to just 10–20%. Therefore, early screening for colorectal cancer is extremely important[2].

Some common gastrointestinal diseases, as shown in Fig. 1, can undergo malignant transformation to varying degrees, eventually leading to colorectal cancer. Colitis is an inflammatory bowel disease[3] characterized by inflammation of the colon mucosa and intestinal wall. It can be classified into two types based on symptoms: ulcerative colitis and Crohn's disease. Ulcerative colitis is an autoimmune disease that typically begins in the rectal area of the colon, with inflammation confined to the mucosal layer of the intestine and gradually spreading to the muscular layer. Clinical manifestations include diarrhea, abdominal pain, bloody stools, and discomfort during defecation. Ulcerative colitis may also be associated with other conditions such as anal fistulas and perianal abscesses. Figure 1e shows an actual image of colitis.

[1]College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China. [2]Zhejiang-Belarus Joint Laboratory of Intelligent Equipment and System for Water Conservancy and Hydropower Safety Monitoring, College of Electrical Engineering, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China. [3]Department of Hepato-Pancreato-Biliary (HPB) Surgery, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China. [4]School of Creative and Digital Industries, Buckinghamshire New University, High Wycombe HP11 2JZ, UK. ✉email: jhzhao2009@zju.edu.cn

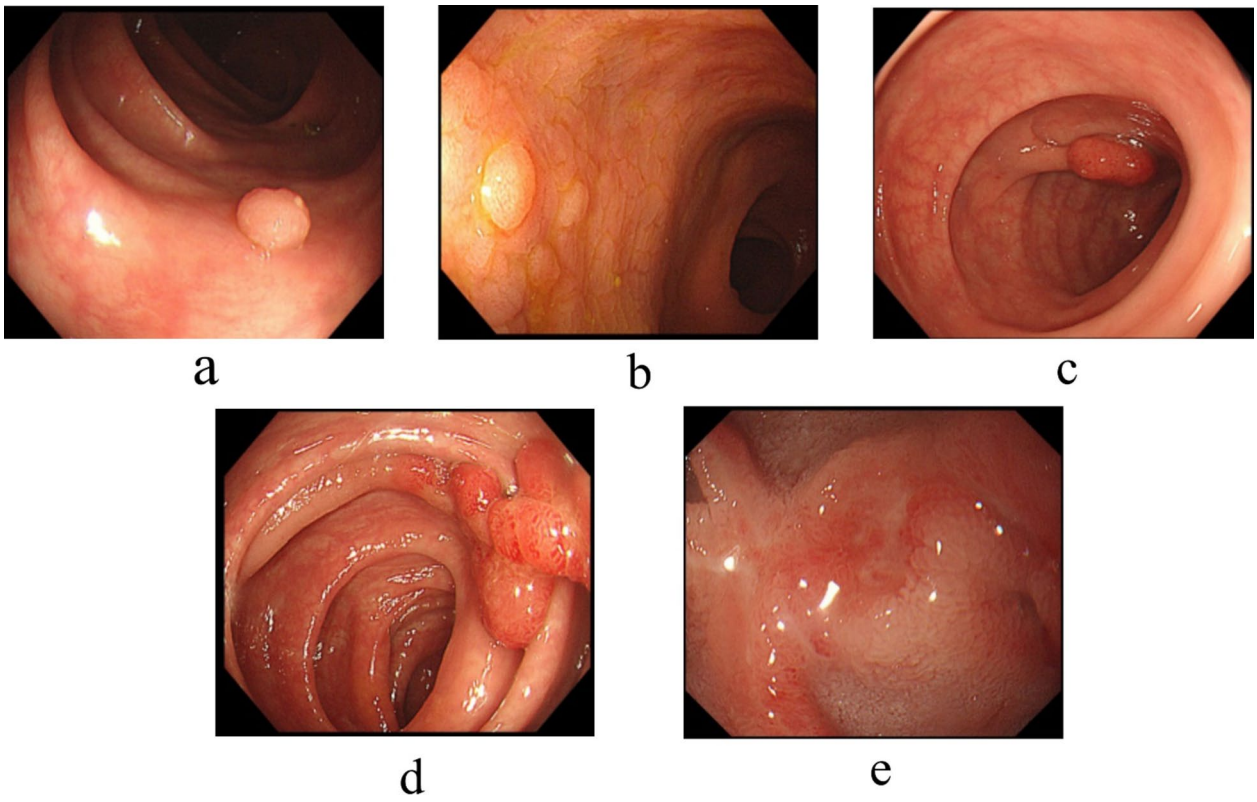

Fig. 1. Real images of common digestive diseases. (**a**) Arti polyps. (**b**) Non-tipped polyps. (**c**) Tipped polyps. (**d**) Multiple polyps. (**e**) Colitis.

A polyp is an abnormal growth on the surface of body tissues. Based on its pathology, it can be classified into adenomatous polyps and non-adenomatous polyps[4–7]. Most colorectal polyps resemble mushrooms or cauliflower, featuring short stalks or stems that connect to the normal mucosal lining of the intestines. Other polyps have a flatter shape (flat polyps) or appear like carpets (sessile polyps). Figure 1a–d show different shapes of polyps in colonoscopy images. Polyps are a type of benign tumor, but over time, some polyps have the tendency to become malignant. If left untreated, they may develop into colorectal cancer. Early diagnosis of colorectal cancer primarily relies on colonoscopy, an effective method that allows direct observation of lesions and pathological analysis. However, its diagnostic accuracy heavily depends on the physician's expertise and operational skills, with fatigue or inexperience potentially leading to missed or incorrect diagnoses. Computer vision-based assisted diagnostic technologies leverage deep learning algorithms to analyze colonoscopy images in real time, enhancing the sensitivity and accuracy of lesion detection, precisely localizing affected areas, reducing the workload of physicians, and significantly improving the detection of small or flat polyps. Compared to traditional methods, these technologies provide an efficient and reliable solution for early colorectal cancer screening, paving the way for the intelligent development of medical diagnostics.

To address these challenges, Angermann et al.[8] proposed a real-time polyp detection method that uses active learning algorithms to improve detection accuracy and efficiency. This method employs frame-based features to identify polyps in videos and utilizes an active learning framework based on classifiers to progressively enhance the classifier's performance, allowing for more accurate identification and localization of polyps. However, the sensitivity and detection accuracy of this method are relatively low. Misawa et al. utilized CNNs to identify lesions in colonoscopy images[9]. The dataset comprised videos from 73 patients, including 73 colonoscopy video segments. However, the experimental results were less than ideal. Although the sensitivity reached 90.0%, the specificity and accuracy were relatively low, at 63.3% and 76.5%, respectively. In reference[10], Peter Klare from Germany researched a novel computer-assisted polyp detection system. The study recruited 30 participants from a hospital in Germany and collected 280 colonoscopy images of polyp detection events to evaluate the system's performance and accuracy. The experimental results indicated that the automated polyp detection system demonstrated high accuracy and robustness in clinical applications, with strong performance in detecting small polyps. However, the system's detection performance still needs improvement in complex scenarios to meet more challenging and practical medical needs. Liu et al.[11] utilized SSD (Single Shot MultiBox Detector) for the localization of polyp images. SSD is an algorithm known for its high precision and speed, but it has limited capability in recognizing small-sized objects. Nisha et al.[12] designed a Dual-Path Convolutional Neural Network (DP-CNN) to classify polyps and normal bowel tissues in colonoscopy images. After training and testing on the CVC-ColonDB dataset (which contains 380 images), the system achieved an accuracy of 99.6% and a recall rate of 99.2%, although the dataset size was relatively small. Nogueira-Rodríguez et al.[13] designed a

deep learning model for real-time polyp detection, based on the YOLOv3 (You Only Look Once) architecture and supplemented it with a post-processing step using object tracking algorithms. The model demonstrated high prediction performance for sessile and pedunculated polyps, but it showed lower performance for flat polyps.

The primary aim of this work is to enhance the detection capability of colonic polyps using an improved YOLOv5s algorithm, thereby advancing early colorectal cancer diagnosis and providing a benchmark for computer-assisted diagnostic systems. The novelty lies in the integration of the SE (Squeeze-and-Excitation) attention mechanism to enhance the C3(Cross Stage Partial Networks) module C3SE(Cross Stage Partial Networks with Squeeze-and-Excitation) and the use of BiFPN (Bi-directional Feature Pyramid Network) to optimize multi-scale feature fusion. Motivated by the need to address challenges like missed detection of small targets and poor performance in complex scenarios, the study validates its improvements on a newly constructed colonic polyp image dataset. Experimental results demonstrate significant gains in mAP, accuracy, and recall, highlighting the feasibility of this approach for advancing intelligent medical diagnostics.

The improved YOLOv5s algorithms

YOLOv5s network structure and algorithm principle

As is well known, the YOLO family of algorithms has been widely applied to numerous target detection tasks in medical imaging[14]. YOLOv5[15] is one algorithm in the YOLO family, with various target detection network models for different image input sizes and datasets[16].

The network architecture of YOLOv5 is depicted in Fig. 2, which provides insight into its four fundamental components: input, backbone, neck, and head[17].

At the input stage, Mosaic data augmentation combines four input images into one to enhance dataset diversity, while adaptive image scaling adjusts input dimensions for improved small and large object detection. Anchor box dimensions are optimized using k-means clustering, enhancing detection accuracy, as shown in Fig. 2. The backbone network, based on CSPDarknet53 (Fig. 3), employs CSP modules and the Focus structure to optimize feature extraction, reduce computation, and retain semantic information. The neck network integrates Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) architectures, enabling multi-scale object detection by fusing features across different layers. Finally, the head network processes feature maps from the neck to produce multi-scale predictions for small, medium, and large objects using C3 modules at various levels (Fig. 2), supporting applications like lesion detection in medical imaging.

SE attention mechanism module

The attention mechanism is a weighted summation process that calculates the final output by aggregating the weights assigned to various input components. This mechanism allows a model to enhance its performance by focusing more on critical components during the input data processing. As a result, attention mechanisms have become essential in neural network design[18].

For example, in target detection tasks, the attention mechanism helps the model concentrate on the most relevant regions for detection while automatically ignoring irrelevant ones.

The SE (Squeeze-and-Excitation) attention mechanism[19] is a popular lightweight approach widely used in convolutional neural networks. Its core procedure involves compression and excitation operations to determine the importance of each channel, thereby enabling the network to capture more refined features[20], as shown in Fig. 4.

The two phases that comprise the SE attention mechanism's concrete implementation are as follows.

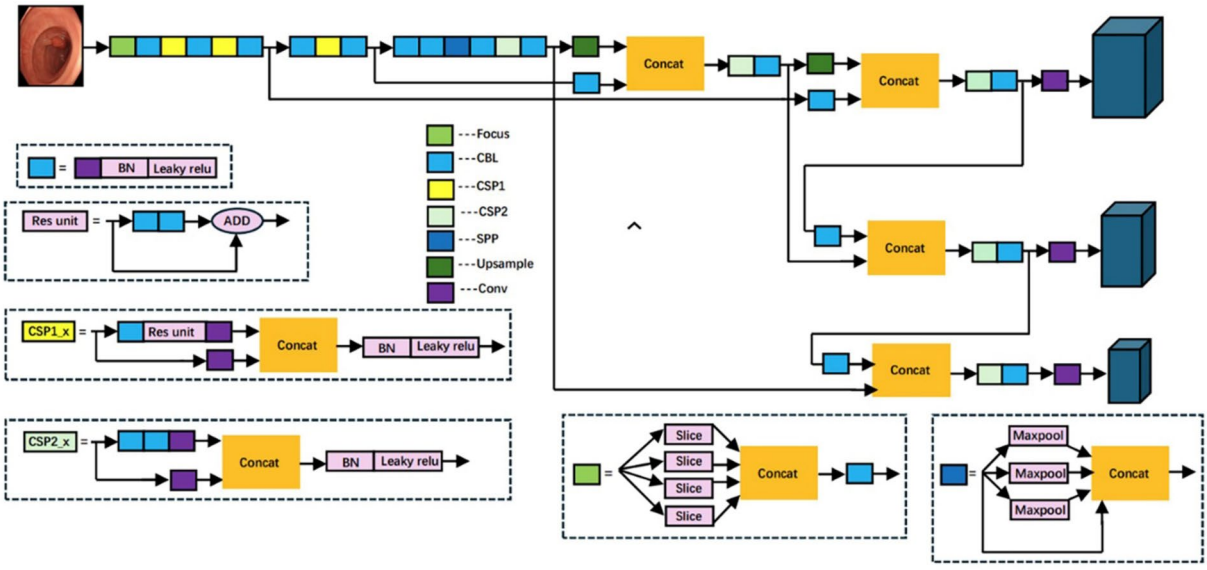

Fig. 2. YOLOv5 network structure diagram.

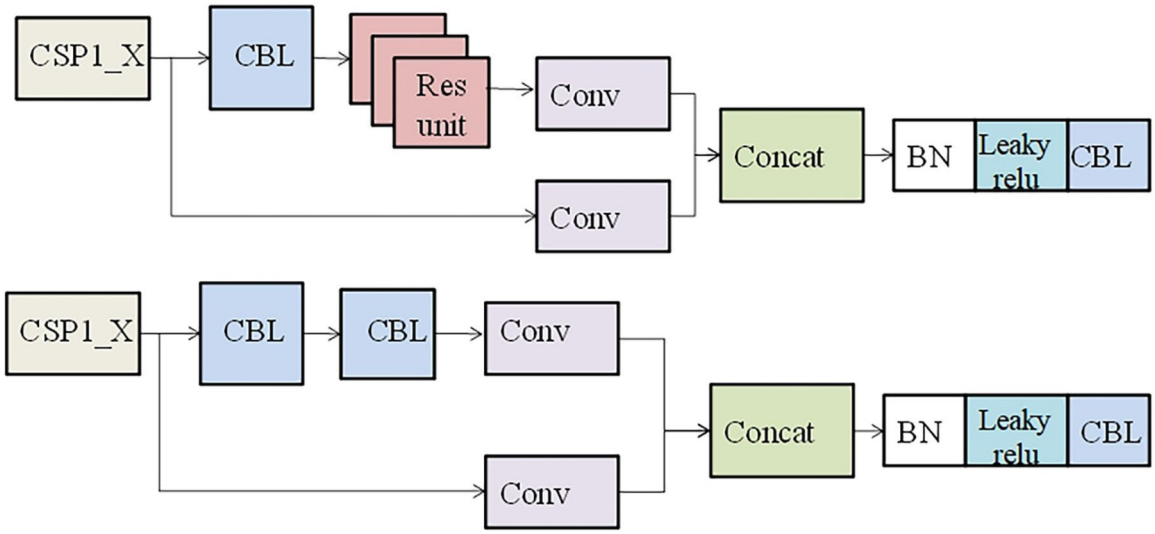

Fig. 3. CSPDarknet53 structure diagram.

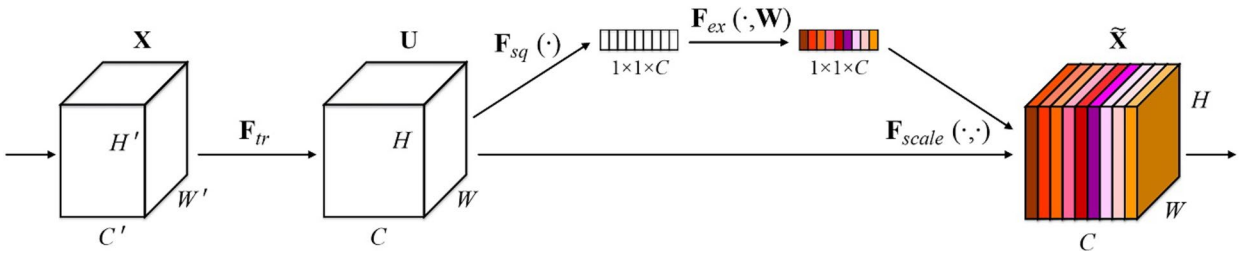

Fig. 4. Compression, excitation blocks.

Squeeze

The compression mechanism pools the global average of the feature maps for each channel, generating a $1 \times 1 \times C$ scalar that represents the global features of all channels through convolutional compression of the feature maps. This process is mathematically expressed in Eq. 1.

$$Z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{w} u_c(i,j) \tag{1}$$

Excitation

The excitation operation enables the model to learn the weights of each feature channel. It consists of two fully connected layers: the first layer has $C \times SERatio$ neurons, while the second layer contains C neurons. Given an input with dimensions $1 \times 1 \times C$, the output remains $1 \times 1 \times C$ after passing through both fully connected layers. Here, SERatio represents the ratio of the output feature vector size in the Squeeze layer to the number of channels in the input feature map. This design effectively simplifies the mathematical expression in the second Eq.

$$s = F_{ex}(z,W) = \sigma\ (g(z,W)) \ = \sigma\ (W_2 \delta\ (W_1 z)) \tag{2}$$

It is possible to apply the SE attention mechanism to either the residual block or the convolutional layer. Implementing the SE attention mechanism allows the model to dynamically prioritize the importance of individual channels, thereby improving its generalization capability. Additionally, due to its limited number of parameters, the SE attention mechanism can be easily integrated into existing convolutional neural networks.

Multi-scale feature fusion network

The multiscale feature fusion network is a neural network architecture designed to improve performance in computer vision tasks by integrating feature information across different scales. A multiscale feature fusion network typically uses layers of varying depths to extract features at distinct scales. These extracted features are then combined through a fusion operation. The specific fusion method, which may include addition, multiplication, concatenation, etc., depends on the task requirements and desired performance. To further boost

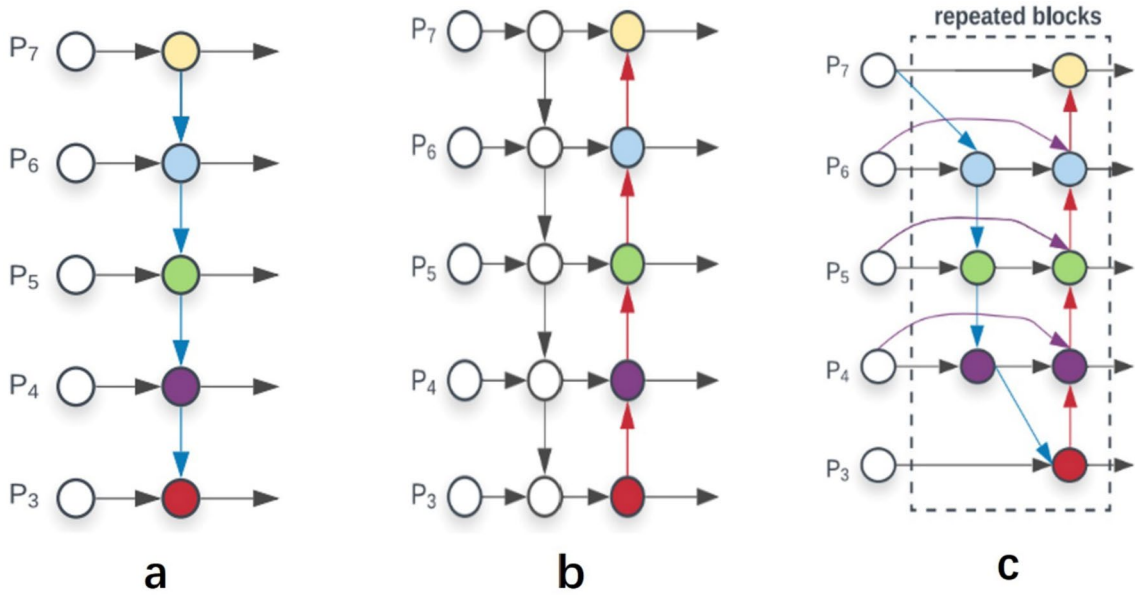

Fig. 5. Different forms of multi-scale feature fusion networks.

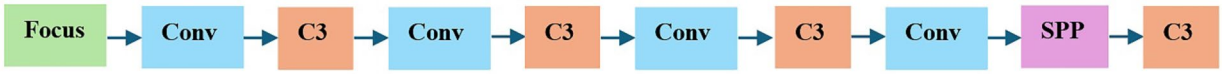

Fig. 6. Backbone network structure.

the network's performance, multiscale feature fusion networks often incorporate additional modules, such as attention mechanisms and aggregation operations.

Multi-scale feature fusion networks have the advantage of fully utilizing feature information at various scales, enhancing the model's receptive field and expressive capability, which leads to improved performance across different computer vision applications. Figure 5 illustrates three standard design patterns for multiscale feature fusion networks: (a) the FPN (Feature Pyramid Network)[21], (b) the PANet (Path Aggregation Network)[22], and (c) the BiFPN (Bidirectional Feature Pyramid Network)[23]. FPN and PANet are used in the Neck network of the YOLOv5 algorithm, as discussed in the previous section. In 2020, Tan et al.[23] proposed the Weighted Bidirectional Feature Pyramid Network (BiFPN) for the EfficientDet network, as shown in Fig. 5c. BiFPN and PANet differ in the following ways: (1) Network Structure: BiFPN builds upon the FPN and introduces multiple bidirectional connections to achieve multi-level feature map fusion. In contrast, PANet performs path aggregation of feature maps at various scales to achieve multi-scale feature map fusion. (2) Feature Fusion: BiFPN establishes bidirectional connections between different layers, allowing information to flow freely and adaptively optimizing feature fusion across levels by learning connection weights. PANet, however, aggregates feature information across scales using path aggregation of feature maps. (3) Training Method: BiFPN requires training of connection weights to enhance the network's adaptability to target detection tasks. In contrast, PANet does not require additional training and instead aggregates feature maps of varying scales directly.

Improvements based on the attention mechanism (SENet-yolov5s network design)

Figure 6 illustrates the structure and parameters of the backbone network as defined in the YOLOv5s.yaml file. To mitigate the impact of background noise and other interfering factors, the attention mechanism module has been introduced, with the layers of the C3 module upgraded to C3SE. A key consideration in developing a fusion network is selecting the appropriate layer for integrating the attention mechanism module to optimize detection performance.

The backbone structure of YOLOv5s contains four C3 layers. To assess the impact of integrating attention mechanisms at different locations, a comparative experiment was conducted by modifying various C3 layers into C3SE. The performance of the modified network was then compared with the original to determine whether the attention mechanism fusion at different C3 layers consistently enhances detection performance and to identify the configuration that yields the best results.

BiFPN-based model improvement

Although top-down and bottom-up feature fusion is performed in the Neck of the YOLOv5 network, the fusion methods described above do not consider the relative importance of different input features. BiFPN

not only introduces a better feature aggregation technique but also addresses the issue of input features having varying resolutions and contributing differently to the final feature fusion. To resolve this, BiFPN incorporates a weighting network based on an attention mechanism that adaptively learns the weights of each input feature layer and adjusts the aggregation process to enhance detection performance. This weighted network takes each input feature layer as input and outputs the corresponding importance weights of the feature layers. These weights are then fused with the features through quick normalization, with the fusion formula as follows:

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_j \omega_j} \cdot I_i \tag{3}$$

In this formula, O represents the fused feature map, d ω enotes the learnable weights, I is the original feature map, and ϵ is a very small value (0.0001) added to avoid numerical instability. The concept of BiFPN is applied to enhance the neck region of YOLOv5, focusing primarily on the following two aspects:

(1) For feature weight fusion, this work introduces a new module, BiFPN_Concat, to replace the previous Concat operation. The detailed structure of the BiFPN_Concat module is shown in Fig. 7. This module defines a trainable weight parameter ω for each feature layer to be fused, and the weight parameters are then normalized to obtain the corresponding normalized weights for each feature map across different layers. Each feature layer is then multiplied by its corresponding normalized weight, and the results are summed to generate the new fused feature layer output. This completes the feature fusion process. By learning the weight parameters of each feature layer, the contribution of each layer can be adaptively adjusted to enhance the algorithm's ability to detect colorectal polyps.

(2) Network connection: Since YOLO has three detection layers, a BiFPN with three nodes is introduced into the algorithm and used only once. At this stage, the outputs from the three detection heads, $P3$, $P4$, and $P5$, are:

$$P_3^{out} = Conv\left(\frac{\omega_1 \cdot P_3^{in} + \omega_2 \cdot \mathrm{Resize}\left(P_4^{td}\right)}{\omega_1 + \omega_2 + \epsilon}\right) \tag{4}$$

$$P_4^{out} = Conv\left(\frac{\omega_3 \cdot P_4^{in} + \omega_4 \cdot P_4^{td} + \omega_5 \cdot \mathrm{Resize}\left(P_3^{out}\right)}{\omega_3 + \omega_4 + \omega_5 + \epsilon}\right) \tag{5}$$

$$P_5^{out} = Conv\left(\frac{\omega_6 \cdot P_5^{in} + \omega_7 \cdot Resize\left(P_4^{out}\right)}{\omega_6 + \omega_7 + \epsilon}\right) \tag{6}$$

where P_4^{td} is calculated by the formula:

$$P_4^{td} = Conv\left(\frac{\omega_8 \cdot P_4^{in} + \omega_9 \cdot \mathrm{Resize}\left(P_5^{in}\right)}{\omega_8 + \omega_9 + \epsilon}\right) \tag{7}$$

Here, ω represents the trainable weight parameter, $Resize$ denotes the up-sampling or down-sampling operation, and Conv refers to the convolution process. P_3^{td}, P_4^{td}, and P_5^{td} are the third, fourth, and fifth feature maps of the image, respectively, following the bottom-up pathway. P_3^{out}, P_4^{out}, and P_5^{out} are the output feature maps of $P3$, $P4$, and $P5$.

The neck structure of the YOLOv5 P3 detection head is now connected to the feature map of layer 4. The P4 detection head is linked to the feature maps of layers 18, 6, and 13, while the P5 detection head is connected to the feature map of layer 9. Through additional layers of feature aggregation, the model can concentrate on critical feature levels by assigning different training weights to various feature map levels, resulting in more accurate feature representations.

Experimental design

Experimental data sets

The images used in our study are sourced from publicly available databases and are randomly extracted from these datasets, including[Kvasir-SEG] https://datasets.simula.no/downloads/kvasir-sessile.zip, [CVC-ClinicDB] https://datasets.simula.no/downloads/kvasir-sessile.zip, [CVC-ClinicDB] https://www.dropbox.com/s/p5qe9eotetjnbmq/CVC-ClinicDB.rar?dl=0, and [LDPolypVideo] https://github.com/dashishi/LDPolypVideo-Benchmark,

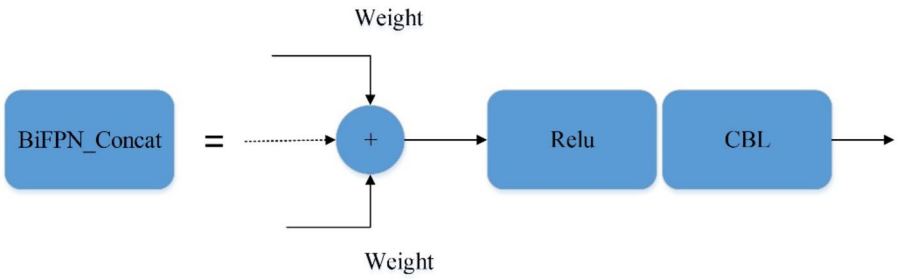

Fig. 7. Introduction to the BiFPN_Concat module.

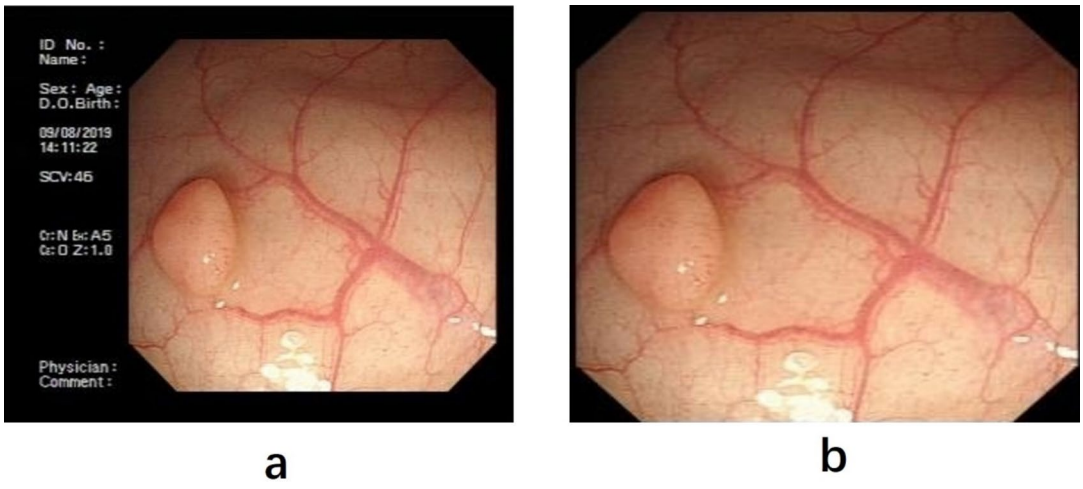

Fig. 8. Images of the actual stomach and intestine.

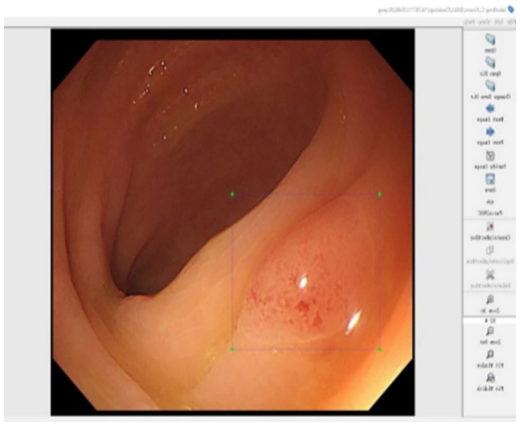

Fig. 9. Image annotation.

all of which are freely accessible for research purposes. The datasets were carefully selected and consist of real image frames extracted from thousands of digestive endoscopy videos. Figure 8(a) depicts the image screen of the actual gastric acquired image. However, owing to the varying pixel resolutions of the acquired images and the left side containing undesirable information and a black border, the image is cropped and scaled to 640 × 640 resolution, as seen in Fig. 8b.

This image dataset utilized for modeling the intestinal illnesses, LabelImg software can be used for annotating these images, as shown in Fig. 9. Figure 10 illustrates how the < size > tag specifies the image's width and height.

The image dataset is made up of 609 photographs of diseased intestines and 2000 photographs of normal intestines. The total 2609 images are separated into training and validation sets in an 8:2 ratio. The remainder, 195 photographs from the collected colonoscopy videos are set as the test set, including 175 images of diseased intestines and 20 images of healthy intestines respectively. As shown in Fig. 11, according to the distribution of X and Y of the Ground Truth Box (real labelled box), it is obvious that the GT Box is mainly concentrated in the middle of the image. According to the distribution of width and height of GT Box, it can be seen that the targets of colonic polyps in the dataset are of different sizes, with extremely large detection targets as well as tiny targets. It is obvious that to build a high-precision detection model is very difficult.

Indicators for model evaluation

The built models are assessed by using the evaluation metrics listed below, which are commonly used in most medical image models.

Confusion matrix

The confusion matrix is a standard method for evaluating classification performance. It is an N × N matrix, with 'N' denoting the number of categorization categories. For a binary classification problem, the confusion matrix is shown in Table 1, where true positive (TP) is the number of positive cases that the model correctly predicts as positive, false negative (FN) is the number of positive cases that the model incorrectly predicts as negative, false

```
<?xml version="1.0"?>
- <annotation>
    <folder>Original</folder>
    <filename>1.png</filename>
    <path>E:\CVC-ClinicDB_datasets\CVC-ClinicDB_PNG_datasets\CVC-ClinicDB_PNG_datasets\Original\1.png</path>
  - <source>
      <database>Unknown</database>
    </source>
  - <size>
      <width>384</width>
      <height>288</height>
      <depth>3</depth>
    </size>
    <segmented>0</segmented>
  - <object>
      <name>Polyp</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
    - <bndbox>
        <xmin>74</xmin>
        <ymin>152</ymin>
        <xmax>180</xmax>
        <ymax>256</ymax>
      </bndbox>
    </object>
  - <object>
      <name>Polyp</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
    - <bndbox>
        <xmin>183</xmin>
        <ymin>208</ymin>
        <xmax>281</xmax>
        <ymax>278</ymax>
      </bndbox>
    </object>
</annotation>
```


Fig. 10. XML file content.

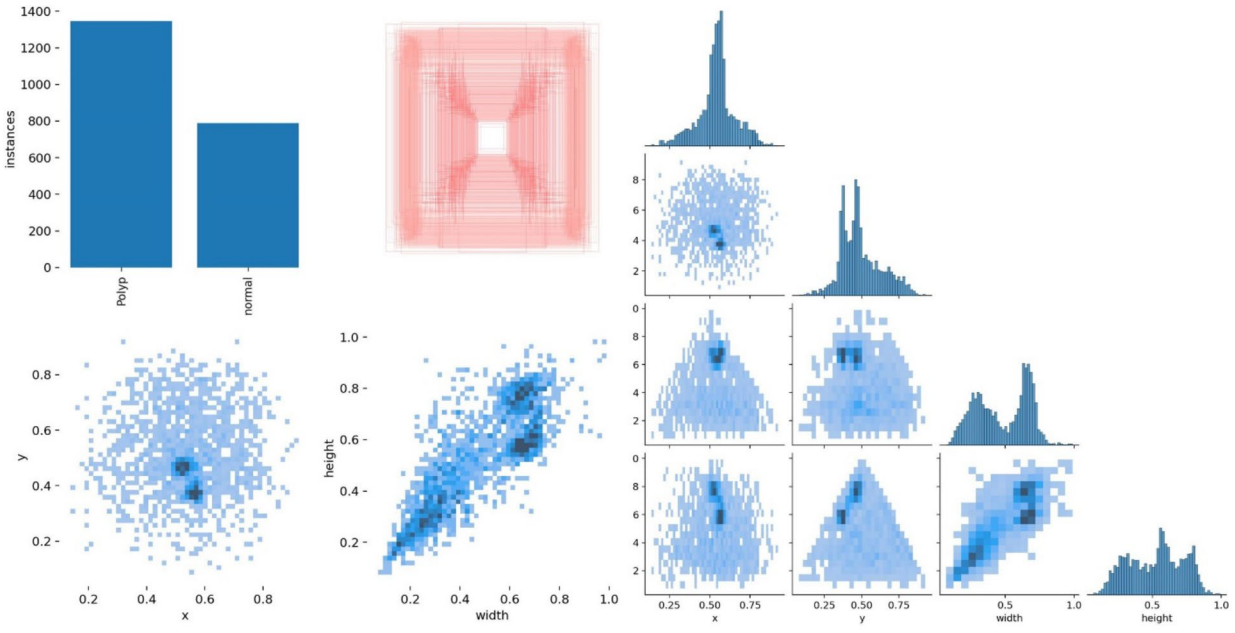

Fig. 11. Dataset distribution and bounding box characteristics for colonic polyp detection.

Confusion matrix		Genuine category	
		Positive sample	Negative sample
Type of projection	Positive sample	TP	FP
	Negative sample	FN	TN

Table 1. Confusion matrix.

Methods	Map (%)	Precision (%)	Recall (%)	FPS (Hz)
YOLOv5s	95.8	93.55	94.11	26
YOLOv5s-1th-2th-C3SE	96.8	95.94	94.45	26
YOLOv5s-1th-3th-C3SE	95.3	92.25	92.71	25
YOLOv5s-1th-4th-C3SE	95.8	95.82	92.09	26
YOLOv5s-2th-3th-C3SE	96.2	94.49	94.39	26
YOLOv5s-2th-4th-C3SE	96.1	95.02	94.89	25
YOLOv5s-3th-4th-C3SE	96.1	95.67	92.63	24

Table 2. Comparative experimental data.

positive (FP) is the number of negative cases that the model incorrectly predicts as positive, and true negative (TN) is the number of negative cases that the model correctly predicts as negative. The indicators of accuracy, recall, and precision can be calculated through the confusion matrix.

Accuracy is defined as the proportion of correctly predicted samples out of the total samples. In general, higher precision generally leads to better model performance. It is defined as:

$$A = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

Recall is defined as the ratio of correctly predicted positive cases to the total number of positive cases. It is also known as sensitivity. It is defined as:

$$R = \frac{TP}{TP+FN} \tag{9}$$

Precision is defined as the number of negative cases properly predicted as negative ones by the model di-vided by the total number of negative cases. Defined as:

$$T = \frac{TP}{TP+FP} \tag{10}$$

mAP (mean average precision)

This paper uses the mean Average Precision (mAP), giga floating-point operations per second (GFLOPs), and frames per second (FPS) to evaluate the model's performance. mAP is used to assess the accuracy of the model, and its calculation formula is as follows:

$$mAp = \sum {}^{P_A}\big/_{N} \tag{11}$$

where P_A represents the area under the curve formed by precision on the x-axis and recall on the y-axis, and N denotes the total number of detection classes. mAP@0.5 indicates the average precision (AP) for each class calculated at an IoU threshold of 0.5, followed by averaging the AP values across all classes. mAP@0.5:0.95 refers to the computation of mAP for IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05, with the final mAP being the average of these values.

Experimental results and analysis

Because the original YOLOv5s has four C3 layers in its backbone structure, a comparison experiment is designed to reveal the performance improvement when different C3 layers are upgraded to C3SE and the original network. Additionally, the experiment includes fusing the attention mechanisms of the C3 layers at different locations.

It is shown from the result list in Table 2 that the model has the highest indexes when the first and second C3 layers are upgraded to C3SE; specifically, the mAP.5 is 1% higher, precision is 2.39% higher, and recall is 0.34% higher than that of the original YOLOv5s model. Furthermore, the detection performance of the YOLOv5s-1st-2nd-C3SE model exceeds the others. Therefore, the introduction of the SE module can increase the model's feature extraction ability to some extent, thereby improving the model's ability to detect polyps.

The next step, in addition to the YOLOv5s-1st-2nd-C3SE model, i.e., the YOLOv5s with the first and second C3 layers upgraded to C3SE, is to build a new model by integrating BiFPN into YOLOv5s as described in the BiFPN-based model improvement subsection. Furthermore, the YOLOv5s-SEBiFPN model is created by coupling the YOLOv5s-1st-2nd-C3SE model with this improvement. Finally, the test experiment is executed

Methods	Map (%)	Precision (%)	Recall (%)	FPS (Hz)
YOLOv5s	95.8	93.55	94.11	26
YOLOv8n	96.1	93.96	95.02	44
YOLOv5s + Bifpn	96.5	94.05	95.21	27
YOLOv5s-1th-2th-C3SE	96.8	95.94	94.45	26
YOLOv5s + SEBifpn	97.4	94.85	95.61	25
FasterRCNN	0.941	0.9296	0.9005	6
SSD	0.952	0.9162	0.9425	40

Table 3. Comparison of the results of the performance indicators of each model.

Methods	Map (%)			Precision (%)			Recall (%)			FPS (Hz)		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
YOLOv5s	95.7	96.1	95.4	93.4	94.2	92.8	94.0	94.5	93.5	26.1	27.0	25.0
YOLOv8n	96.2	96.5	95.8	94.0	94.3	93.7	95.1	95.5	94.6	43.9	44.5	43.0
YOLOv5s + Bifpn	96.4	96.7	96.0	94.1	94.3	93.9	95.3	95.3	94.6	27.0	28.0	26.5
YOLOv5s-1th-2th-C3SE	96.7	97.0	96.5	95.9	96.2	95.5	94.5	95.0	94.0	26.2	27.5	25.8
YOLOv5s + SEBifpn	97.3	97.6	97.2	94.8	95.2	94.7	95.6	94.9	95.7	24.9	26.0	23.5
FasterRCNN	0.94	0.945	0.935	0.92	0.925	0.915	0.90	0.910	0.890	6.0	6.5	5.5
SSD	0.951	0.955	0.947	0.916	0.92	0.912	0.943	0.946	0.938	39.8	40.5	39.0

Table 4. Comparison of performance indicators after 5-fold cross-validation for each model.

in the same way as above, including the Faster R-CNN, SSD, and original YOLOv5s models. The performance indicators' findings are displayed in Table 3.

Since the training conditions of the seven models are consistent, the indicators result listed in Table 3 reveals that the improved YOLOv5s model outperforms the original YOLOv5s model in terms of mAP, accuracy, and recall values, with 0.7%, 0.5%, and 1.1% increments, respectively. Comparatively, the coupled improvement model, YOLOv5s-SEBiFPN, increases the performance indicators more, with mAP, accuracy, and recall increasing by at least 1.6%, 1.3%, and 1.5%, respectively.

Another experiment is performed by using the means of k-fold cross-validation to identify differences in the models' random performance. As is shown in Table 4, YOLOv5s + SEBifpn demonstrates exceptional stability, particularly in mAP, with a variation of only 0.4%, and in precision, with a variation of merely 0.5%, reflecting its consistency in accuracy and recall. Compared with other models, YOLOv5s + SEBifpn exhibits the highest stability in mAP, showcasing superior precision and robust recall capabilities, making it well-suited for high-accuracy and stability-demanding tasks such as intestinal detection. Although its FPS is slightly lower (ranging from 24.9 Hz to 26.0 Hz), its outstanding accuracy is reliable performance in practical applications. YOLOv5s + SEBifpn offers significant advantages in terms of both high precision and stability, making it particularly suitable for scenarios requiring rigorous accuracy and robustness.

Figure 12 is collected from the results of YOLOv5s and YOLOv5s-SEBiFPN models on the test set to depict the visualization of detection targets, with a red border representing the confidence of the prediction box's category, as indicated in Eq. (12).

$$P_r\left(class_i?objet\right) * P_r\left(object\right) \cdot \ IoU = P_r\left(class_i\right) * IoU \tag{12}$$

where $P_r\left(class_i?objet\right)$ is the probability of being a specific class if there is an object, and $P_r\left(object\right) \cdot \ IoU$ is $P_r\left(object\right)$ multiplied by IoU (Intersection over Union).

In Fig. 12a–d are typical results from the original YOLOv5s model, and images (e), (f), (g), and (h) are corresponding results from the YOLOv5s + SEBiFPN model. The sub-images (c) and (g), (d) and (h) show that the confidence level of the YOLOv5s-SEBiFPN model increases. The sub-images (a) and (e) show that the YOLOv5s-SEBiFPN model not only improves in confidence level but also enhances the detection capabilities for tiny and medium-sized polyps. However, it can also be seen that the polyp obscured by the intestinal wall in (b) is undetected due to a failure in detecting polyps in complex and diverse scenes.

Figure 13 shows YOLOv5s-1s-2nd-C3SE algorithm is employed to identify the (a) and (b) images on the upper side. The improved YOLOv5s + SEBifpn algorithm is utilized to identify the (d) and (e) images on the lower side. It is observed that the YOLOv5s-1s-2nd-C3SE algorithm cannot precisely identify certain flat-shaped polyps. This phenomenon can occur since certain flat-shaped objects may exhibit varying dimensions or angles within the image, which may prove difficult for the attention mechanism to accurately detect. In contrast, the YOLOv5s-1th-2nd-C3SE algorithm exhibits superior performance over YOLOv5s + Bifpn. For instance, Fig. 13c remains undetected in YOLOv5s + Bifpn, whereas, Fig. 13(f) is detected in YOLOv5s-1th-2nd-C3SE. This discrepancy indicates that the YOLOv5s-1th-2nd-C3SE algorithm is superior to the YOLOv5s + Bifpn algorithm when dealing with flat polyps.

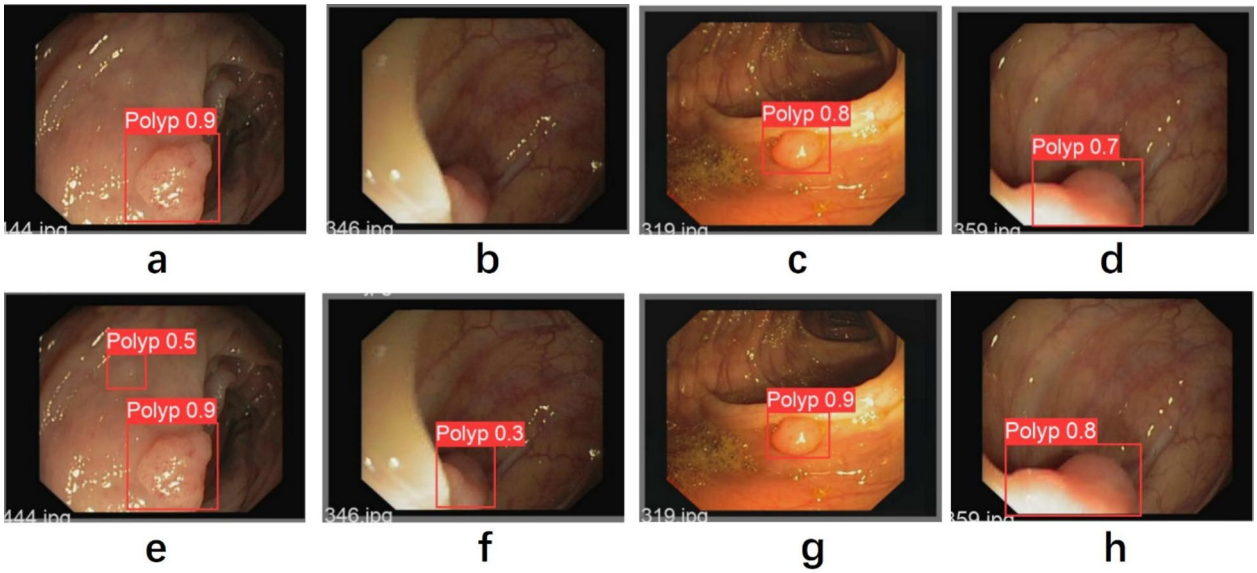

Fig. 12. YOLOv5s + SEBifpn algorithm vs. YOLOv5s algorithm visualisations.

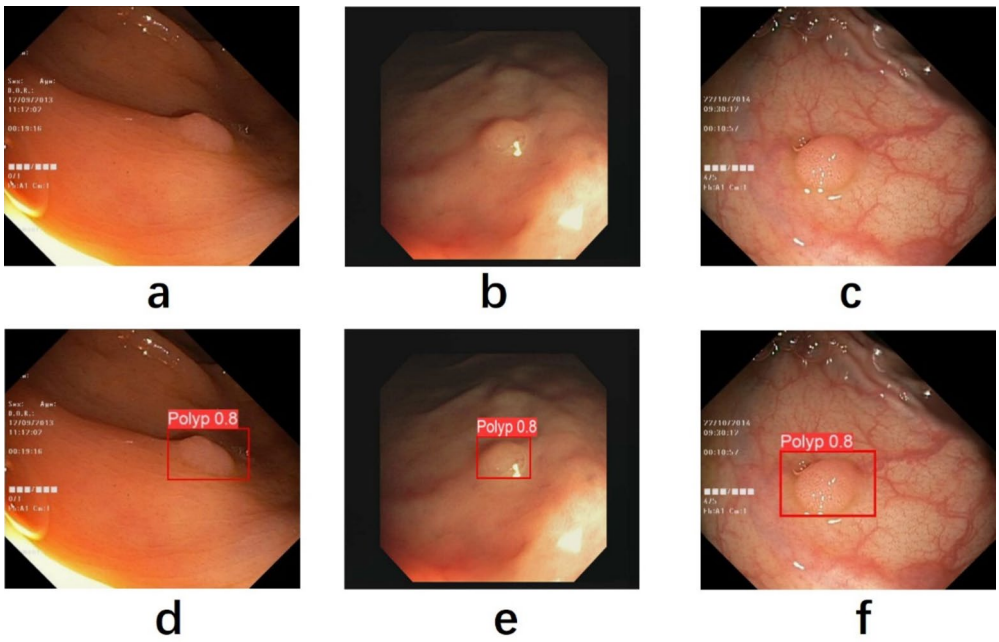

Fig. 13. YOLOv5s + SEBifpn, YOLOv5s-1st-2nd-C3SE, YOLOv5s + Bifpn visualisation comparison.

Figure 13 is exhibited to reveal the performance of SE, BiFPN, and SE + BiFPN modified YOLOv5s models, respectively, selected from the undetected results of the three models. The sub-images (a) and (b) represent the weak performance of the YOLOv5s-1st-2nd-C3SE model. The sub-images (d) and (e) correspond to the YOLOv5s + SEBiFPN model. The sub-image (c) is from the undetected set of the YOLOv5s + BiFPN model, with the same undetected result as in sub-images (a) and (b). The sub-image (f) is from the YOLOv5s + SEBiFPN model. This discrepancy indicates that the YOLOv5s + SEBiFPN algorithm may be slightly superior to the YOLOv5s + BiFPN model when dealing with flat polyps. Perhaps some flat-shaped targets exhibit varying dimensions or angles in the images, making it difficult for the attention mechanism to detect them accurately.

Finally, Fig. 14 illustrates typical recognition images from the YOLOv5s + SEBiFPN model. In each sub-image of Fig. 14, prediction boxes, confidence scores, and classes are tagged, except for the normal intestinal images. Specifically, Fig. 15 displays the number of images with similar confidence scores in the test set (195 images). For instance, the number of images with a confidence score of 0.9 that tested positive for intestinal polyps is about 118, which is 60.8% of the total, while the ratio for images with confidence scores between 0.7 and 0.8 is 25.3%.

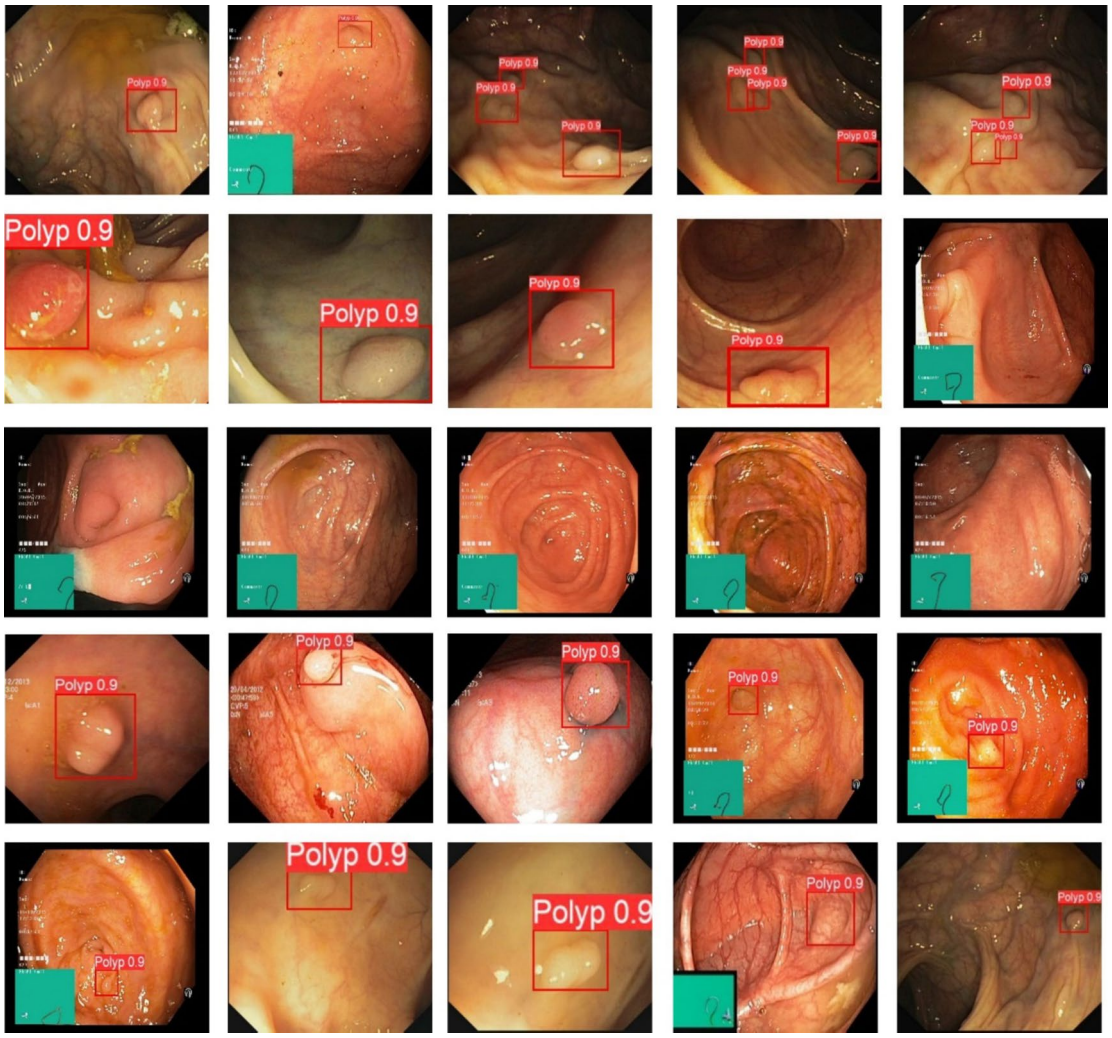

Fig. 14. YOLOv5s + SEBifpn algorithm detection visualisation.

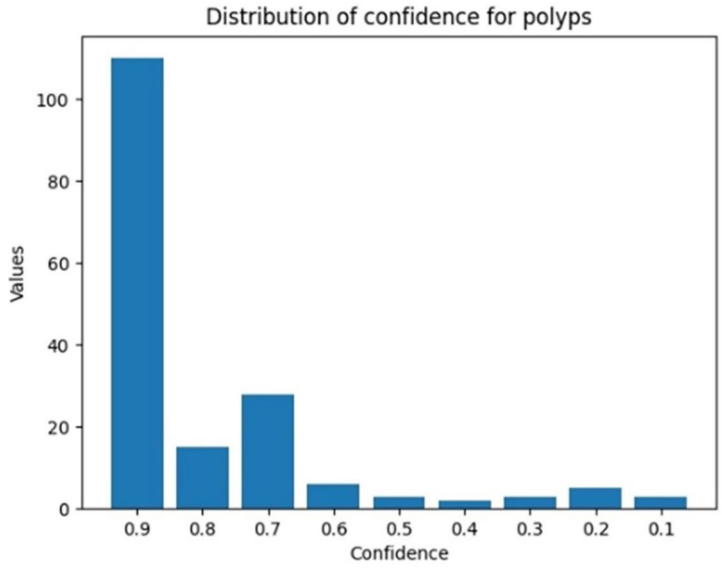

Fig. 15. Distribution of confidence levels comparison.

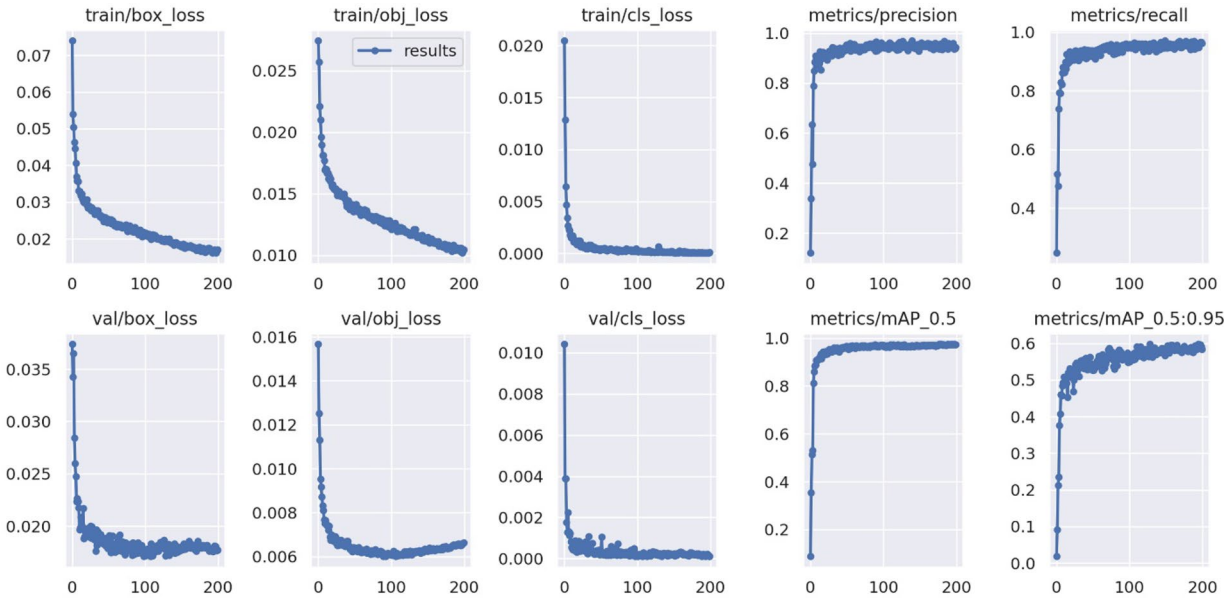

Fig. 16. Training results of YOLOv5s + SEBifpn algorithm.

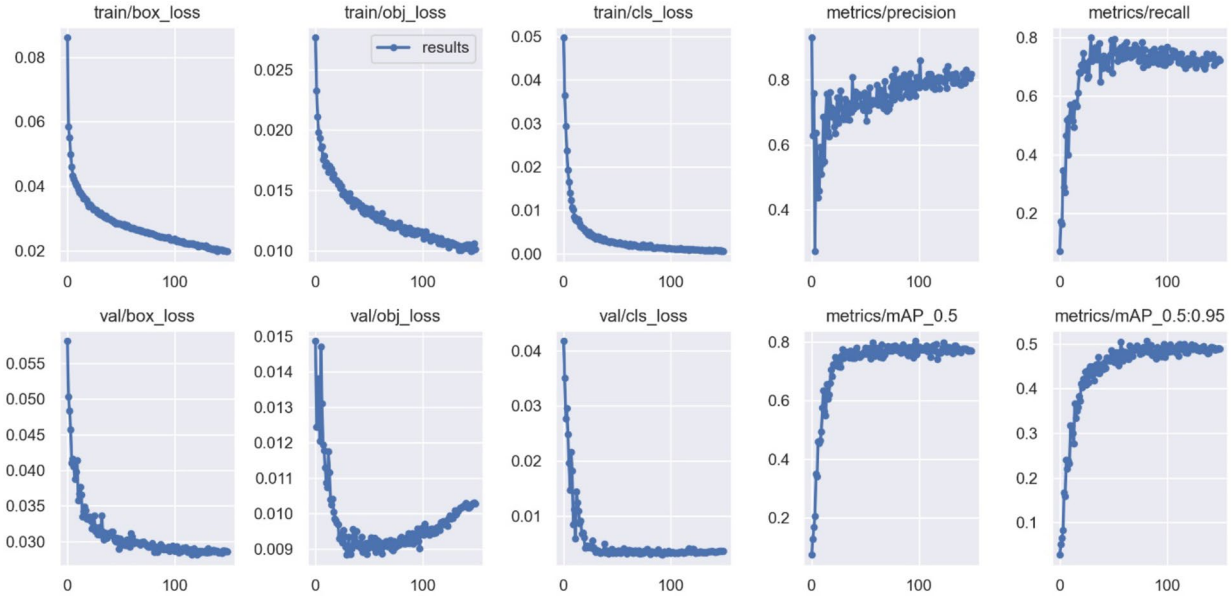

Fig. 17. Training results of YOLOv5s algorithm.

It becomes evident that the enhanced YOLOv5s + SEBiFPN algorithm exhibits greater stability in the decline of the loss function after 200 epochs during training, as shown in Figs. 16 and 17. Additionally, it displays a smoother and less pronounced ascent in the mAP and recall curves. The increased stability of the YOLOv5s + SEBiFPN algorithm is clear.

Conclusion

The primary aim of this study is to investigate a high-performance algorithm for identifying polyp lesions in endoscopic images. A new image dataset containing polyps is gathered for training the proposed algorithm model. The proposed algorithm is an enhanced YOLOv5s, designed to improve the recognition rate of small and medium-sized polyps. Specifically: (1) Implement the attention module in the C3 layer at various positions along the backbone network, thereby directing the network's attention towards the target area for detection via weighting, to acquire more comprehensive information; (2) Enhance the model's pyramid connection method by leveraging bidirectional weighted feature fusion (BiFPN). Empirical evidence based on the dataset supports the efficacy of these improvements through comparative experiments.

In contrast to its predecessor, the enhanced YOLOv5s + SEBiFPN model exhibits superior detection capabilities while maintaining notably low computational complexity. Moreover, it surpasses other exceptional detection algorithms of a similar nature.

However, several limitations persist. The model's performance may degrade in highly challenging scenarios such as extreme lighting conditions, very small or blurred polyps, or when facing certain types of occlusion. Future research should focus on further enhancing the model's robustness against such challenges. Additionally, applying the proposed method to other medical imaging fields, such as lung or skin lesion detection, could explore its broader applicability. For instance, leveraging advancements like the hybrid attention strategy employed in HADCNet for lung infection segmentation[24] or the projective parameter transfer-based sparse learning framework in neuroimaging[25] could offer valuable insights for improving the generalization and performance of models in other medical imaging domains. Addressing these limitations, including improving generalization across diverse datasets, remains a key direction for future work. Furthermore, combining this model with advanced segmentation techniques and incorporating additional contextual information could further improve its accuracy and reliability.

Data availability

The datasets used and analysed during the current study available from the corresponding author on reasonable request.

Received: 9 September 2024; Accepted: 20 February 2025
Published online: 26 February 2025

References

1. Sung, H. et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
2. Ciuti, G. et al. Frontiers of robotic endoscopic capsules: A review. *J. Microbiol. Robot.* **11**, 1–18. https://doi.org/10.1007/s12213-016-0087-x (2016).
3. Choi, I. H. et al. Collagenous gastoduodenitis in the form of a gastric ulcer. *Korean J. Gastroenterol.* **80**, 225–228. https://doi.org/10.4166/kjg.2022.079 (2022).
4. Chao, G., Zhu, Y. & Fang, L. Retrospective study of risk factors for colorectal adenomas and non-adenomatous polyps. *Transl. Cancer Res.* **9**, 1670–1677. https://doi.org/10.21037/tcr.2020.01.69 (2020).
5. Liu, J., Zhang, W., Liu, Y. & Zhang, Q. Polyp segmentation based on implicit edge-guided cross-layer fusion networks. *Sci. Rep.* **14**. https://doi.org/10.1038/s41598-024-62331-5 (2024).
6. Sikkandar, M. Y. et al. Utilizing adaptive deformable Convolution and position embedding for colon polyp segmentation with a visual transformer. *Sci. Rep.* **14**. https://doi.org/10.1038/s41598-024-57993-0 (2024).
7. Xu, C., Fan, K., Mo, W., Cao, X. & Jiao, K. Dual ensemble system for polyp segmentation with submodels adaptive selection ensemble. *Sci. Rep.* **14**, 6152. https://doi.org/10.1038/s41598-024-56264-2 (2024).
8. Angermann, Q., Histace, A. & Romain, O. Active learning for real time detection of polyps in videocolonoscopy. *Procedia Comput. Sci.* **90**, 182–187. https://doi.org/10.1016/j.procs.2016.07.017 (2016).
9. Misawa, M. et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology* **154**, 2027–2029. https://doi.org/10.1053/j.gastro.2018.04.003 (2018).
10. Klare, P. et al. Automated polyp detection in the colorectum: A prospective study (with videos). *Gastrointest. Endosc.* **89**, 576–582. https://doi.org/10.1016/j.gie.2018.09.042 (2019).
11. Liu, M., Jiang, J. & Wang, Z. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access.* **7**, 75058–75066. https://doi.org/10.1109/access.2019.2921027 (2019).
12. Nisha, J. S., Gopi, V. P. & Palanisamy, P. Automated colorectal polyp detection based on image enhancement and dual-path CNN architecture. *Biomed. Signal. Process. Control* **73**, 103465. https://doi.org/10.1016/j.bspc.2021.103465 (2022).
13. Nogueira-Rodríguez, A., Glez-Peña, D., Reboiro-Jato, M. & López-Fernández, H. Negative samples for improving object detection—A case study in AI-assisted colonoscopy for polyp detection. *Diagnostics* **13**, 966. https://doi.org/10.3390/diagnostics13050966 (2023).
14. Gao, J., Xiong, Q., Yu, C. & Qu, G. White-light endoscopic colorectal lesion detection based on improved YOLOv5. *Comput. Math. Methods Med.* **2022**, 9508004. https://doi.org/10.1155/2022/9508004 (2022).
15. Zhu, X., Lyu, S., Wang, X. & Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* 2778–2788. https://doi.org/10.1109/ICCVW54120.2021.00312 (2021).
16. Zhao, J., Zhu, B., Peng, M. & Li, L. Mobile phone screen surface scratch detection based on optimized YOLOv5 model (OYm). *IET Image Proc.* **17**, 1364–1374. https://doi.org/10.1049/ipr2.12718 (2023).
17. Sun, Q., Zhang, X., Li, Y. & Wang, J. YOLOv5-OCDS: an improved garbage detection model based on YOLOv5. *Electronics* **12**, 3403. https://doi.org/10.3390/electronics12163403 (2023).
18. Jiao, L. et al. Brain-inspired remote sensing interpretation: A comprehensive survey. *IEEE J. Sel. Top. Appl. Earth Obs Remote Sens.* **16**, 2992–3033. https://doi.org/10.1109/JSTARS.2023.3247455 (2023).
19. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7132–7141. https://doi.org/10.1109/CVPR.2018.00745 (2018).
20. Sheng, K. & Chen, P. An efficient mixed attention module. *IET Comput. Vis.* **17**, 496–507. https://doi.org/10.1049/cvi2.12184 (2023).
21. Lin, T. Y. et al. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* 2117–2125. https://doi.org/10.1109/CVPR.2017.106 (2017).
22. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8759–8768. https://doi.org/10.1109/CVPR.2018.00913 (2018).
23. Tan, M., Pang, R., Le, Q. V. & EfficientDet Scalable and efficient object detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079 (2020).
24. Chen, Y. et al. HADCNet: automatic segmentation of COVID-19 infection based on a hybrid attention dense connected network with dilated Convolution. *Comput. Biol. Med.* **149**, 105981 (2022).
25. Fei, X. et al. Projective parameter transfer based sparse multiple empirical kernel learning machine for diagnosis of brain disease. *Neurocomputing* **413**, 271–283 (2020).

Acknowledgements
The author thanks and acknowledges the support from the Zhejiang Provincial Key Research and Development program of China under Grant (No. 2020C03074), and the Nanxun scholars program of ZJWEU (No. RC2023010962).

Author contributions
JL Carried out algorithm improvements and wrote the first draft of the paper. JZhao Contributed to the writing of the first draft and performed revisions and editing. YW Collected part of the data and constructed the dataset. JZhu Provided a large amount of data and guided the experiments. YWei, JJ Zhu & XL: Edited and revised the paper, SYan& QZ Polished the paper. All authors reviewed the manuscript.

Funding
This work was supported in part by the Zhejiang Provincial Key Research and Development program of China under Grant (No. 2020C03074) and the Nanxun scholars program of ZJWEU (No. RC2023010962).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.Z.