Multimedia Tools and Applications (2024) 83:71773–71788
https://doi.org/10.1007/s11042-024-18497-1

Small gastric polyp detection based on the improved YOLOv5

Linfei Wu[1] · Jin Liu[1] · Haima Yang[2] · Bo Huang[1] · Haishan Liu[1] · Shaowei Cheng[1]

Received: 3 June 2023 / Revised: 7 September 2023 / Accepted: 29 January 2024 /
Published online: 6 February 2024

Abstract
Small target polyps are prone to missed detection due to their small coverage area and little information. To address this issue, a modified PATM-YOLO polyp detection model based on YOLOv5 is proposed. The model first addresses the issue of missed detection of small polyps by constructing a detection head for identifying small polyps and using an improved Phase-Aware Token Mixing Module(PATM) attention module to increase the network's attention to small polyps and suppress the model's focus on non-polyp regions. Secondly, an improved Adaptively Spatial Feature Fusion(ASFF) module is proposed to fully utilize multi-scale information, enhancing the network's feature richness. Finally, by introducing the Swin Transformer into the network and determining its optimal placement through experiments, the detection accuracy is maximized without affecting the network's performance. After experimental comparison on the constructed dataset and the public dataset SUN, the proposed PATM-YOLO network model alleviated missed detection in dense and small polyp images, and achieved an precision of 91.3%, which is 8.5% higher than the baseline YOLOv5 network model. This indicates that the detection performance of this model outperforms other classical target detection networks and the original network.

✉ Jin Liu
liujin@sues.edu.cn

Linfei Wu
adrian19970116@163.com

Haima Yang
snowyhm@sina.com

Bo Huang
huangbosues@sues.edu.cn

Haishan Liu
hithsh@163.com

Shaowei Cheng
luochen211203@163.com

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, People's Republic of China

[2] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, 516 Jungong Road, Shanghai 200093, People's Republic of China

Keywords Polyp · Deep learning · YOLOv5 · Small target · Detection

1 Introduction

Gastric cancer is a prevalent disease that poses a serious threat to human health [1]. According to the estimates by the International Agency for Research on Cancer (WHO), around 1.09 million individuals across the world suffered from gastric cancer in 2020, with 769,000 deaths attributed to the disease. This ranks gastric cancer as the sixth most prevalent malignancy and fourth most deadly worldwide. China accounts for 36.42% of new cases and 37.58% of deaths from gastric cancer worldwide, indicating the urgency of prevention and treatment. However, gastric cancer is frequently difficult to detect at an early stage due to the lack of specific symptoms in most gastric cancer patients, which is one of the reasons for the current low diagnosis and treatment rate of early-stage gastric cancer. In addition, gastric polyps can cause surface bleeding due to the corrosive effects of gastric acid, leading to symptoms such as anemia at the later stages. If polyps can be detected at an early stage and undergo regular follow-up or direct surgical treatment, the survival rate of patients can be significantly improved. Therefore, effective prevention and treatment measures should be taken to reduce the harm caused by gastric cancer to humans.

The main method for detecting polyps is gastroscopy screening, which is divided into traditional manual screening and assisted screening methods based on deep learning technology. However, traditional manual screening methods have limitations. Due to the irregular variations in texture features [2, 3], shape [4, 5], size [5], color [6, 7], and other characteristics of polyps, manual screening not only time-consuming to identify polyps but also requires doctors to have relevant knowledge. Moreover, even if the doctor has the appropriate knowledge, factors such as fatigue level [8] and the characteristics of the polyp itself can lead to experienced experts making misdetection or missed detection of some polyps [9]. Therefore, the development of computer-aided diagnosis technology to assist doctors in polyp detection is of great significance.

Due to the significant breakthroughs in GPU computing power, recently, researchers have devoted a lot of effort to computer vision. Compared to conventional manual screening methods, deep learning-based approaches can help doctors focus their attention on identifying suspected polyps, rather than wasting their time on a massive amount of normal images. Therefore, deep learning technology is now utilized in the field of medical imaging [10, 11], aiming to address the above problems while improving the precision of polyp detection and reducing the risk of missed diagnosis and misdiagnosis, providing doctors with more comprehensive, reliable, and efficient diagnostic tools.

In 2020, Deeba et al. [12] introduced a computer-assisted algorithm for the detection of polyps in both colonoscopy and wireless capsule endoscopy (WCE). This algorithm involved several key components, including image enhancement, the generation of saliency maps, and the extraction of histogram of oriented gradients (HOG) features, all of which played a crucial role in the final classification process. By effectively amplifying clinically significant features and reducing the number of search windows through saliency-based selection, the algorithm ultimately improved detection efficiency. In 2021, Qadir et al. [13] created a colon polyp detection system based on F-CNN, which used a two-dimensional Gaussian mask instead of a binary mask, allowing the proposed system to successfully detect the flat and small target polyps with blurred boundaries between the background and the polyp, reducing the rate of missed detection of colon polyps. In 2021, Taş et al. [14] suggested a preprocessing approach that used a super-resolution method based on convolutional neural networks

(SRCNN) to enhance the resolution of colonoscopy images prior to polyp localization. This method improved both recall rate and accuracy of the model compared to the low-resolution case. In 2021, Chen et al. [15] improved the saliency of the polyp area by enhancing the contrast of the input image through the differentiation of foreground and background images. The enhanced data were input into an improved deep residual convolutional neural network and integrated learning method for automatic colon polyp detection. By adding attention modules, the network can focus on useful feature channels and suppress invalid feature channels, greatly improving the precision of the detection network. In 2021, Cao et al. [16] proposed a network for detecting gastric polyps, which incorporated a module that extracted and merged features and was based on the YOLOv3 network. The network utilized both high-level and low-level feature maps' semantic information, resulting in improved detection of small target polyps, with a recall rate of 86.2%. In 2022, Nisha et al. [17] proposed a dual-path convolutional neural network (DP-CNN) that used image enhancement techniques, DP-RNN structure, and the S-shaped classifier to detect polyps, successfully classifying polyps and non-polyp patches in colonoscopy images and reducing complexity with fewer learnable parameters. In 2022, Hu et al. [18] proposed a novel approach, NeutSS-PLP, aimed at extracting polyp regions within colonoscopy images. The method combines neutral uncertainty theory and saliency detection strategies to enhance the identification accuracy of specular reflections in colonoscopy images and to perform suppression. In addition, a two-level short connection to the saliency detection network was introduced to extract multi-level and multi-scale features for better polyp region extraction.

Several effective strategies have been proposed for conventional polyp detection problem, which have performed well in terms of accuracy, recall, and feasibility. However, due to the irregularity, low resolution, and insufficient feature information of polyp targets, conventional detection models often encounter issues of missed detection or false detection when facing such small targets. Therefore, this study optimized the YOLOv5 model for small polyp target detection, including the following aspects: firstly, to address the issue of information loss in small polyp targets, a new network was developed by adding a small target detection head and utilizing Swin Transformer to enhance the network's sensitivity to small targets, thereby improving small polyp detection. Secondly, to fully utilize the information between different scales, the new network integrated the ASFF module, which can be applied to four detection heads. Additionally, to weaken the impact of non-detection object areas in the image on the model results, a more outstanding plug-and-play Res-PATM attention mechanism module was proposed based on the PATM module. The proposed PATM-YOLO algorithm achieved 91.3% precision and 86.6% recall in the constructed dataset and 95.6% precision and 90.8% recall in the public polyp dataset SUN, outperforming other comparison algorithms in both datasets. The presented PATM-YOLO algorithm demonstrates its effectiveness in detecting small polyps, as indicated by these results.

2 Material and methods

2.1 Dataset

In this study, the parts of the collected datasets [23, 24] related to polyps were extracted and combined into a new dataset, in order to test the detection capability of the model with a richer dataset. The dataset constructed in this study consists of 1759 images, most of which are small polyps and can support related detection tasks for small polyps. Table 1 displays the distribution of images used in this study. A total of 1,127 images were used for training the

Table 1 Details regarding the public datasets employed in this study

Dataset	Total	Train set	Val set	Test set	Dataset download link
Synthetic Dataset	1759	1127	281	351	https://github.com/jiquan/Dataset-acess-for-PLOS-ONE https://datasets.simula.no/hyper-kvasir/
SUN	49136	31448	7862	9826	http://sundatabase.org

network model, 281 for validation, and 351 for testing the model's performance. To validate the model's effectiveness, experiments were also conducted on the publicly accessible polyp dataset SUN [25], using the corresponding techniques. SUN is an public dataset for polyp detection, which contains up to 49,136 photos with polyp information collected from 100 patients and divided into 100 parts according to different patients. To ensure experimental objectivity, these photos were randomly partitioned into training, validation, and test sets. Within the training set, 31,448 photos containing polyp information were used for training. Furthermore, the validation set consisted of 7,862 images, while the remaining 9,826 polyp photos were allocated to the test set.

2.2 The improved ultra-small target detection head

Due to the high presence of small polyps in the dataset and the significant irregularity in their shape, texture, and size, YOLOv5 does not perform optimally in detecting these small polyps. To address this issue, this study constructed a detection head for small targets in the model's head to counteract the missed detection that occurs with YOLOv5 in small polyp detection [19]. This approach enhances the model's detection accuracy for small polyps without compromising its ability to detect polyps of other sizes.

2.3 Improved PATM attention module

Currently, attention mechanisms are being widely used in the field of vision. Inspired by this, this study introduces the PATM attention module, which combines the advantages of smaller inductive bias and simpler architecture in MLP, to enhance the network model's attention to effective targets and suppress attention to non-target areas [20], based on the following principle:

The PATM attention module characterizes a token as a wave function that possesses phase and amplitude, defined as follows:

$$\tilde{Z}_p = |Z_p| \odot e^{i\theta_p}, p = 1, 2, \ldots, m, \tag{1}$$

Where i represents the imaginary unit that satisfies $i^2 = -1$, $|\cdot|$ represents the absolute value operator, and $\odot$ represents the element-wise dot product operator. The amplitude $|Z_p|$ represents the real-valued feature for each token, $e^{i\theta_p}$ is a periodic function, and θ_p represents the phase, corresponding to the current position of the token within the wave period. The phase term θ_p affects the summing result of different tokens during aggregation.

Calculate the corresponding amplitude information Z_p and phase information θ_p based on the given input features using Formula 2 and Formula 3, respectively.

$$Z_p = Channel - FC(X_p, W^c) = W^c X_p, p = 1, 2, \ldots, m, \tag{2}$$

$$\theta_p = \Theta(X_p, W^\theta), \tag{3}$$

Where W^c represents a weight that possesses learnable parameters, and W^θ represents learnable parameters.

As Formula 1 is represented in the complex domain, Formula 4 is needed to expand it and represent it in terms of real and imaginary parts.

$$\tilde{Z}_p = |Z_p| \odot cos\theta_p + i|Z_p| \odot sin\theta_p, p = 1, 2, \ldots, m, \tag{4}$$

In the above formula, the real and imaginary parts of complex-valued tokens are signified by two vectors, correspondingly. Then, different tokens $\tilde{Z}_p$ are merged using the $token - FC$ operation, i.e.:

$$\tilde{O}_p = Token - FC(\tilde{Z}, W^t)_p = \sum_q W^t_{pq} \odot \tilde{Z}_q, p = 1, 2, \ldots, m, \tag{5}$$

Where $\tilde{Z} = [\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_m]$ represents all the wave-like tokens in one layer. In Formula 5, the interaction between tokens takes into account both their amplitude and phase information. The resulting output, $\tilde{O}_p$, is represented by complex values that combine the features. Following the common quantum measurement approach that involves projecting a quantum state, characterized by a complex-valued representation, onto an observable real value. The real-valued output O_p is obtained by weighting and summing the real and imaginary parts of $\tilde{O}_p$ with parameters. Combined Formula 5, the output O_p can be obtained:

$$O_p = \sum_q W^t_{pq} Z_q \odot \cos\theta_q + W^i_{pq} Z_q \odot \sin\theta_q, p = 1, 2, \ldots, m, \tag{6}$$

Where W^t and W^i each represent weights with learnable parameters. In the above formula, the phase θ_q dynamically adjusts itself based on the semantic content of the input data. In addition to the unchanging weights, the phase also modulates the aggregating process of different tokens.

As shown in Fig. 1, the PATM attention module generates amplitude and phase information using Formula 2 and the phase estimation function Formula 3, respectively, given the input

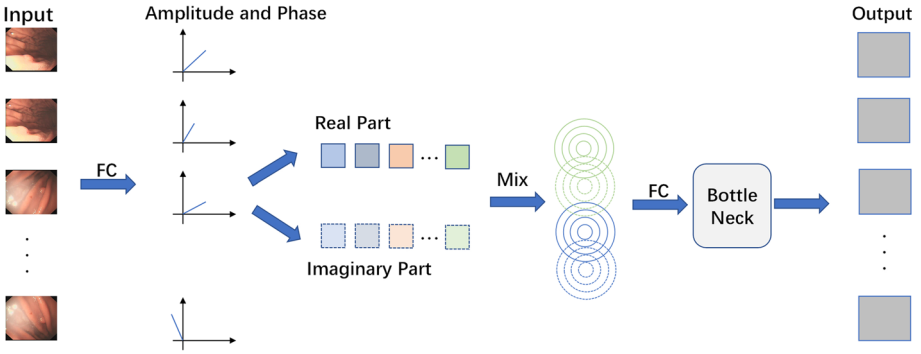


Fig. 1 Schematic diagram of the Res-PATM principle

features. The complex-value representation is obtained by expanding the output as wave-like tokens using Formula 4 and aggregating them with Formula 6. The output features are enhanced by transforming them with another Channel-FC to increase the representational power and fed into the bottleneck residual module to deepen the network, thus improving the detection capability for small targets. This study refers to the improved module as Res-PATM.

To investigate how the improved PATM attention module can be applied with the bottleneck residual module, this study conducted experiments comparing four adding methods as shown in Fig. 2. These experiments aimed to maximize the detection results of the network and determine the optimal number of bottleneck residual modules to be used in the PATM attention module.

The experimental findings presented in Fig. 2 indicate that the optimal performance of the network is achieved when two bottleneck residual modules are incorporated into the PATM attention module, allowing it to concentrate on more relevant information.

2.4 Determining the location of swin transformer

The Swin Transformer-v2 architecture benefits from the shift-window operation, which restricts the attention operation to a window and reduces the computational cost. Additionally, the Patch Merging operation can increase the receptive field and obtain multi-scale features [21]. Swin Transformer-v2 architecture makes the amplitude controllable by applying layer normalization afterwards. Inspired by these methods, this study replaced some of the original Cross Stage Partial(CSP) modules in YOLOv5 with CSP modules based on the Swin Transformer-v2 architecture(Swin-CSP).The schematic diagram related to the Swin-CSP module is shown in Fig. 3.

To further verify the optimal placement of Swin Transformer modules in the network, this study conducted experimental comparisons of the optimal placement positions, as shown in Table 2.

The results presented in Table 2 demonstrate that the detection model performs best when replacing one CSP module in the backbone network and all in the neck, and is able to extract

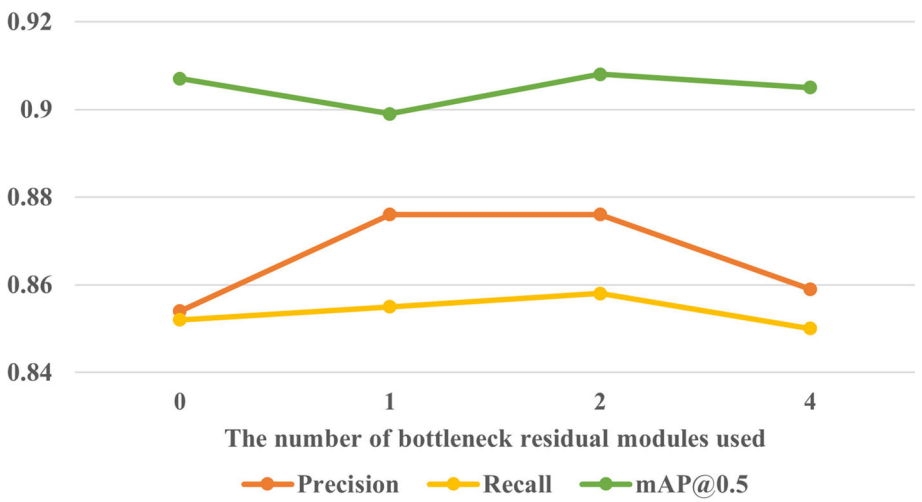


Fig. 2 Comparison of Experimental Results in Four Different methods

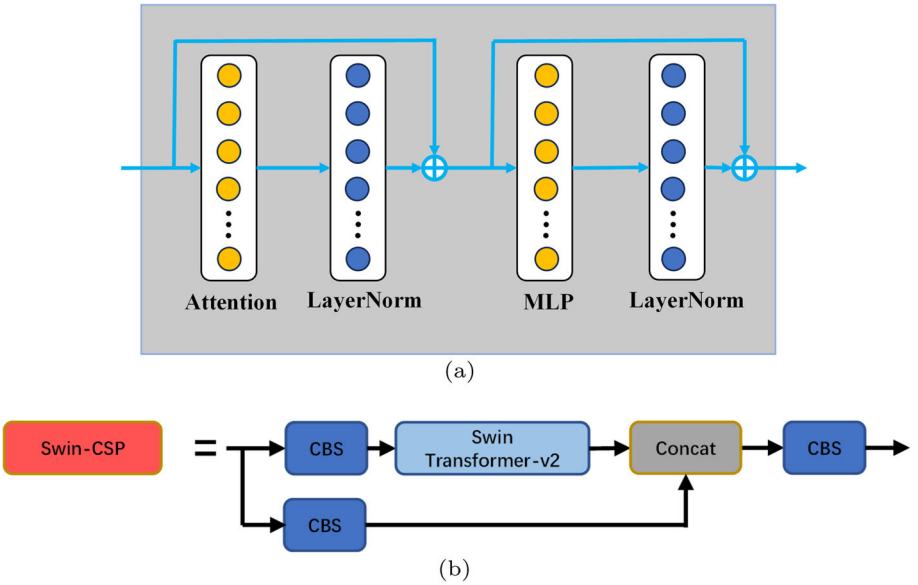


Fig. 3 Illustrations related to Swin Transformer-v2: (a) The schematic diagram of the Swin Transformer-v2 block; (b) Swin-CSP module

more effective features compared to other network structures. Finally, we extended this to the PATM-YOLO model. As depicted in Fig. 5, only one CSP module in the head was replaced, whereas all CSP modules in the neck were replaced.

2.5 The improved ASFF module

The initial structure of YOLOv5 is affected by the irregular variations in size and shape of polyps, resulting in differences in detection difficulty due to polyps of different sizes. To reduce the detection difficulty fluctuations caused by polyps of different sizes, this study introduced the ASFF module [22]. The ASFF module's fundamental concept revolves around empowering the network to dynamically acquire spatial feature weights across various scales during fusion, and its implementation can be divided into two parts: feature size normalization and scale fusion, as follows:

Table 2 Experimental results of optimal placement positions

Number of replacements		P	R	mAP@0.5
Head	Neck			
1	0	0.839	0.844	0.895
2	0	0.842	0.846	0.897
4	0	0.809	0.769	0.845
2	2	0.814	0.860	0.883
1	4	0.846	0.857	0.884
2	4	0.841	0.833	0.880

Unified Feature Size: Due to the various resolutions and channel numbers in the network header, the feature layers will eventually need to perform the summation operation as depicted in Formula 7. Therefore, it is crucial to ensure that each layer has uniform channel numbers and feature map size. This can be achieved by initially utilizing a regular convolution operation to obtain equal channel numbers, and then adjusting the sampling strategy for different levels of upsampling and downsampling to ensure uniform feature map size. As shown in Fig. 4, the blue line indicates the downsampling operation and the red line represents the upsampling operation used to enhance resolution.

As shown in Formula 7, the scale fusion operation is performed on the l-th level as an example. By dividing the feature layers of different resolutions into levels, the same resolution feature maps obtained from the other three levels after the feature size unification operation (i.e., downsampling operation) are weighted and summed to obtain the final features.

$$y_{ij}^{l} = \alpha_{ij}^{l} * x_{ij}^{1\to l} + \beta_{ij}^{l} * x_{ij}^{2\to l} + \gamma_{ij}^{l} * x_{ij}^{3\to l} + \delta_{ij}^{l} * x_{ij}^{4\to l} \tag{7}$$

Where α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l} and δ_{ij}^{l} are the spatial feature fusion weights of their corresponding feature maps relative to the feature map of level l, and these weights are shared across channels. It is worth noting that α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l} and δ_{ij}^{l} are subject to the constraints of $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}+\delta_{ij}^{l} = 1$ and α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l}, $\delta_{ij}^{l} \in [0, 1]$ and are defined as follows:

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}} + e^{\lambda_{\beta_{ij}}^{l}} + e^{\lambda_{\gamma_{ij}}^{l}} + e^{\lambda_{\delta_{ij}}^{l}}} \tag{8}$$

Where α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l} and δ_{ij}^{l} are defined by the softmax function with $\lambda_{\alpha_{ij}}^{l}$,$\lambda_{\beta_{ij}}^{l}$,$\lambda_{\gamma_{ij}}^{l}$, and $\lambda_{\delta_{ij}}^{l}$ as the control parameters.

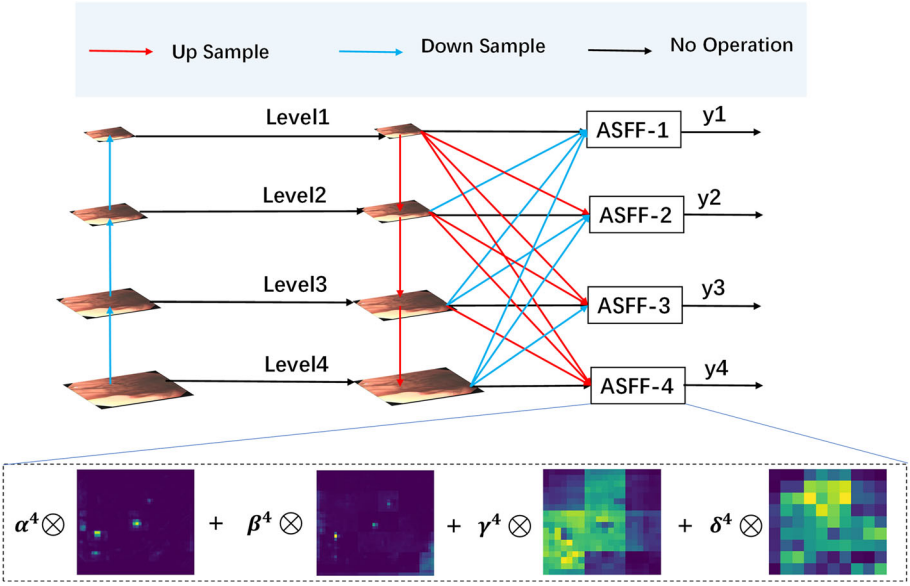


Fig. 4 Framework of ASFF module

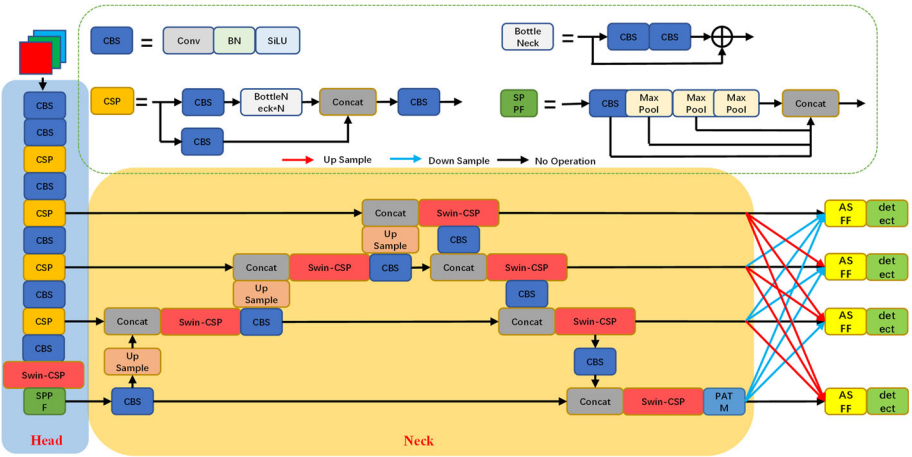


Fig. 5 Structural diagram of the PATM-YOLO algorithm

2.6 The PATM-YOLO algorithm

Based on the above enhancements, the present study suggests an improved algorithm Phase-Aware token Module based YOLOv5(PATM-YOLO) based on YOLOv5, which is dedicated to enhancing the missed detection of dense, small polyps due to the loss of information on small polyps, unevenness in polyp texture and polyp size, and the complexity of the detection background. Figure 5 depicts the structural diagram of the PATM-YOLO algorithm.

3 Results

3.1 Implementation details

3.1.1 Training setting

In this paper, the experimental setup utilizes the Ubuntu 20.04 operating system. The central processing unit (CPU) employed is a 24 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz, while the graphics processing unit (GPU) chosen is the RTX 3090 with 24GB memory capacity. The experimentation environment is configured with Python 3.8.1 scripting language, PyTorch 1.10.0 deep learning framework, and CUDA 11.3 GPU acceleration library.

The key training parameters for the experiments were configured as follows: input image dimensions were set to 640 by 640 pixels, the initial learning rate was established at 0.01, learning rate momentum was assigned a value of 0.937, and the weight decay coefficient was set to 0.0005. The training process spanned 100 epochs with a batch size of 64.

3.1.2 Evaluation metrics

In this research, the effectiveness of pre- and post-improved network models in detecting both dense and small polyp images using a constructed dataset was assessed under similar

experimental conditions. The differences in experimental outcomes were compared to assess the network performance, i.e., the status of missed and false detections. The three main metrics chosen for this study included precision, recall, and mean average precision (mAP), which were calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$mAP@0.5 = \frac{\sum_{i=1}^{N} AP_i}{N}, \tag{11}$$

The formulas presented above use TP to indicate true accurate predictions, FP to indicate false predictions, and FN to indicate false negative i.e., false predictions,and is frequently employed for assessing the overall target detection network model's detection performance.

3.2 Ablation experiment

To further verify the impact of the proposed modules and optimizations on the detection algorithm in polyp detection tasks, this study conducted a set of ablation experiments. Based on the YOLOv5s network, this study added the Swin Transformer network to create YOLOv5s-a, added the ASFF module to create YOLOv5s-b, added the PATM attention module to create YOLOv5s-c, and added the small target detection head to create YOLOv5s-d. The network with all modules added to the YOLOv5s baseline is referred to as the proposed PATM-YOLO network. The results of the ablation experiments are shown in Table 3. In comparison to the YOLOv5 network, the separate addition of each module to the network not only led to a minimum improvement of 1.8% in precision but also yielded a performance increase of at least 1.1% in both recall and mAP@0.5. This underscores the feasibility of enhancing the model.

3.3 Experimental comparison between YOLOv5 algorithm and improved algorithms

In this section, the PATM-YOLO algorithm is compared with the YOLOv5 series algorithms. In order to objectively demonstrate the performance on the dataset, considering all existing YOLOv5 models, including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5x, and YOLOv5l.

Table 3 Ablation experiments of PATM-YOLO

ID	PATM	Small target detection head	Swin Transformer	ASFF	P	R	mAP@0.5
YOLOv5s	×	×	×	×	0.828	0.833	0.873
YOLOv5s-a	×	×	✓	×	0.846	0.857	0.884
YOLOv5s-b	×	×	×	✓	0.872	0.844	0.886
YOLOv5s-c	✓	×	×	×	0.876	0.858	0.908
YOLOv5s-d	×	✓	×	×	0.847	0.861	0.886
PATM-YOLO	✓	✓	✓	✓	0.913	0.866	0.920

Table 4 Comparison of experimental results between the PATM-YOLO and YOLOv5 series networks

Model	Parameters (Million)	Weights (MB)	P	R	mAP@0.5
YOLOv5n	1.8	3.9	0.746	0.792	0.825
YOLOv5s	7.0	14.5	0.828	0.833	0.873
YOLOv5m	20.9	42.2	0.814	0.763	0.828
YOLOv5l	46.1	92.9	0.784	0.798	0.835
YOLOv5x	86.2	173.1	0.809	0.798	0.846
PATM-YOLO	40.9	82.7	0.913	0.866	0.920

Table 4 shows the different performance of PATM-YOLO algorithm and YOLOv5 series algorithms on the constructed dataset.

Figure 6 displays a comparison of performance parameters between the PATM-YOLO algorithm and the YOLO series algorithms, revealing that the PATM-YOLO algorithm exhibits superior precision, recall rate, and mAP@0.5 compared to other models. With reference to Table 4, it can be inferred that the PATM-YOLO algorithm attains an precision rate of 91.3%, a recall rate of 86.6%, and an mAP@0.5 of 92% in the detection experiment of polyp targets when contrasted with the original YOLOv5 series network. This constitutes an improvement of 8.5%, 3.3%, and 4.7%, respectively, over the YOLOv5s baseline network model. It can be observed that the performance of the PATM-YOLO algorithm on the constructed dataset exhibits an advantage.

For specific detection tasks involving dense and small targets, the detection performance of the YOLOv5 and PATM-YOLO network models is shown in Figs. 7 and 8. Figure 7 corresponds to dense polyp images with three targets in the original image, of which the original network detected two targets but missed one, while the improved PATM-YOLO

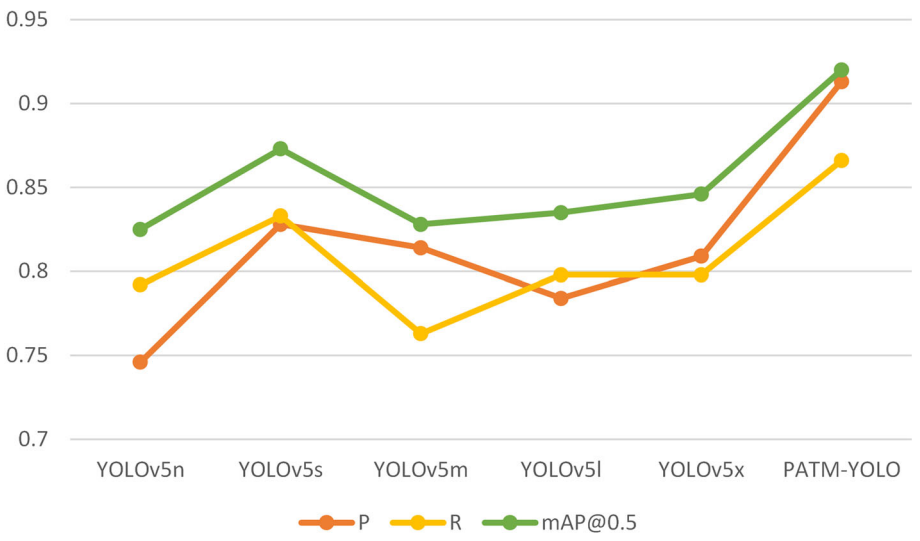
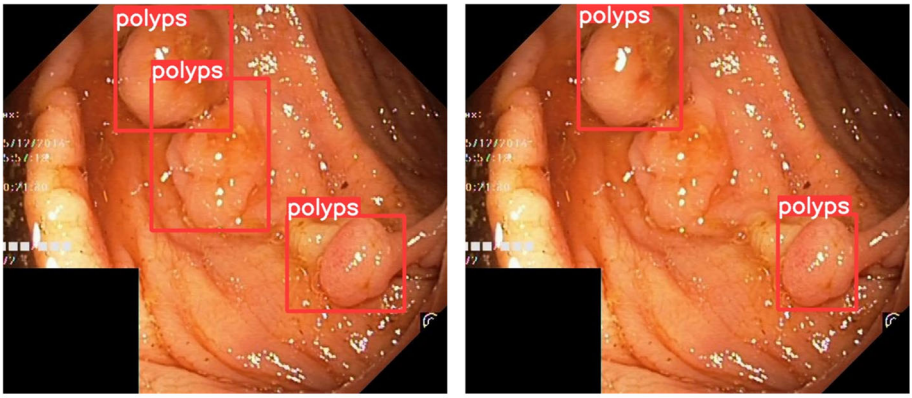


Fig. 6 Comparison of different performance parameters between the PATM-YOLO and YOLOv5 series algorithms



(a) (b)

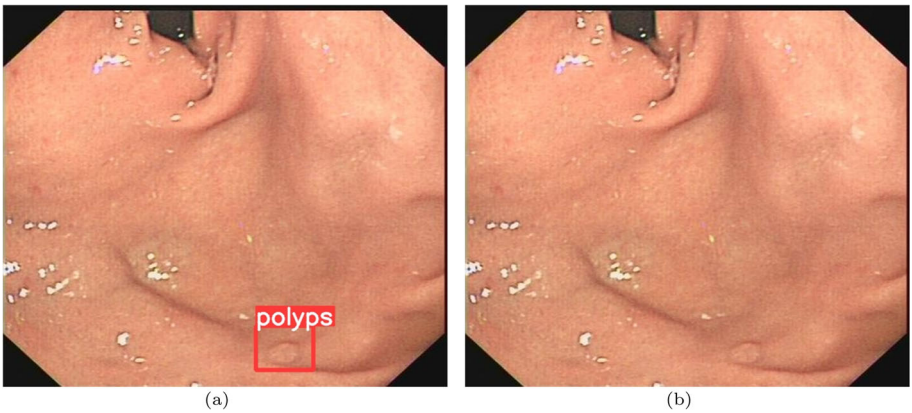
Fig. 7 Comparison of detection results for dense polyp images: (a) PATM-YOLO; (b) YOLOv5

network detected all polyp targets. For small polyp images in Fig. 8, the original network exhibited missed detection, while the improved network was able to detect all targets.

3.4 Comparison of the PATM-YOLO algorithm with other algorithms

To further validate the effectiveness of the PATM-YOLO model, comparative experiments were conducted with other algorithms under the condition of maintaining consistent configuration environments and initial hyperparameters as much as possible.

Table 5 shows the experimental results of the PATM-YOLO algorithm and other algorithms on the dataset constructed in this paper. As shown in the table, under the input size of 640*640, the detection performance of the PATM-YOLO algorithm and other algorithms are outstanding, and are able to achieve better performance that surpasses other algorithms in polyp detection tasks.



(a) (b)

Fig. 8 Comparison of detection results for small polyp images: (a) PATM-YOLO; (b) YOLOv5

Table 5 Comparison between the PATM-YOLO algorithm and other algorithms

Model	P	R	mAP@0.5
YOLOv5s	0.828	0.833	0.873
SSD	0.843	0.867	0.879
YOLOv7	0.904	0.863	0.919
YOLOv8	0.875	0.865	0.865
PATM-YOLO	0.913	0.866	0.920

3.5 Testing of the PATM-YOLO algorithm on the SUN dataset

This section aims to introduce the experiments of the PATM-YOLO algorithm proposed in this study on the public polyp dataset SUN. Similarly, to ensure the fairness of the polyp detection experiments, the experiment is conducted using similar parameter settings as the previous experiments.

Figure 9 shows the training process and validation results of YOLOv5, YOLOv7, YOLOv8, and PATM-YOLO. As shown in the figure, the red line representing YOLOv8 has lower recall, precision, and mAP@0.5 compared to the other three detection algorithms. The proposed PATM-YOLO algorithm can achieve higher precision and recall in a shorter time compared to the baseline YOLOv5 network. Furthermore, although the training process

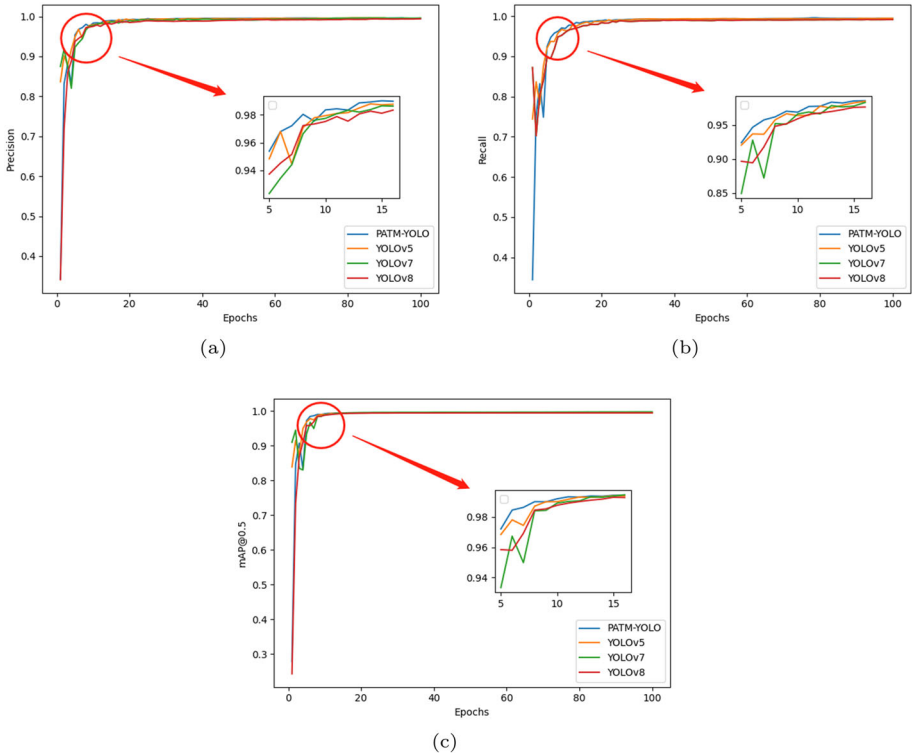


Fig. 9 The validation results of the model training: (a) mAP@0.5; (b) Recall; (C) Precision

Table 6 The comparison results between PATM-YOLO and other algorithms on the SUN dataset

Model	P	R	mAP@0.5
YOLOv3-CSP [26]	0.953	0.750	0.844
YOLOv4-CSP [26]	0.928	0.799	0.974
YOLOv3 [26]	0.958	0.808	0.988
YOLOv4 [26]	0.930	0.802	0.979
YOLOv5-ABS [27]	0.934	0.814	0.922
YOLOv5	0.941	0.888	0.943
YOLOv7	0.952	0.900	0.946
YOLOv8	0.914	0.833	0.924
PATM-YOLO	0.956	0.908	0.961

shows that PATM-YOLO and YOLOv7 have similar precision and recall, the PATM-YOLO algorithm actually outperforms YOLOv7 in the test set. When Table 6 is integrated into the analysis, it can be observed that the PATM-YOLO algorithm on the SUN dataset shows an increase of 0.4% in precision and 0.8% in recall.

It can be observed from Table 6 that the PATM-YOLO algorithm outperformed the detection performance in the public dataset SUN by a significant margin in terms of recall and was more suitable for the polyp detection task compared to other detection networks [26, 27].

4 Conclusions

The study proposes a new method for detecting small polyps in images, called PATM-YOLO. The proposed method addresses the issue of missed detection of small polyps. In terms of network architecture, a detection head is firstly constructed for detecting small targets, followed by an attention mechanism to obtain richer information and limit the influence of background areas in the image on the target. Secondly, the Swin Transformer structure is employed to augment the network's feature extraction capacity. Finally, the ASFF module is incorporated into the network to enhance the integration of multi-scale features and enrich the network's feature diversity. The PATM-YOLO algorithm achieved better performance than other YOLOv5 series algorithms, with an precision of 91.3%, a recall rate of 86.6%, and an mAP@0.5 of 92% on the constructed dataset. In addition, the algorithm also achieved an precision of 95.6% and a recall rate of 90.8% on the public SUN dataset, making it more suitable for polyp detection tasks. The study shows that PATM-YOLO algorithm can improve the detection performance of polyps. In addressing the computational requirements, there is a need for further improvement in the PATM-YOLO algorithm. Enhancing the algorithm to reduce computational costs while maintaining detection accuracy and improving its detection performance to facilitate deployment on resource-constrained devices will be a focal point of our future work.

Author Contributions LW, JL, and HY contributed to conception and design of the study. LW, HY, HL, and SC organized the database. LW, JL and SC performed the statistical analysis. LW wrote the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding This work was supported by the Shanghai Science and Technology Innovation Action Plan(22S31903700 & 21S31904200).

Availability of data and materials Three publicly available polyp datasets of endoscopy images (the PLoS One-Zhang dataset, the Hyper-Kvasir-Segmentation dataset, and the SUN dataset) were used in the experiments of this study. The PLoS One-Zhang dataset can be found at: https://github.com/jiquan/Dataset-acess-for-PLOS-ONE. The Hyper-Kvasir-Segmentation dataset can be found at: https://datasets.simula.no/hyper-kvasir/. The SUN dataset can be found at: http://sundatabase.org.

Code Availibility Not applicable

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval Not applicable

Consent to participate Not applicable

Consent for publication Not applicable

References

1. Brosens LA, Wood LD, Offerhaus GJ, Arnold CA, Lam-Himlin D, Giardiello FM, Montgomery EA (2016) Pathology and genetics of syndromic gastric polyps. International journal of surgical pathology 24(3):185–199. https://doi.org/10.1177/1066896915620013
2. Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F (2009) Texture-based polyp detection in colonoscopy. In: Bildverarbeitung Für die Medizin 2009: Algorithmen—Systeme—Anwendungen Proceedings des Workshops Vom 22. Bis 25. März 2009 in Heidelberg, Springer, pp 346–350. https://doi.org/10.1007/978-3-540-93860-6_70
3. Iakovidis DK, Maroulis DE, Karkanis SA, Brokos A (2005) A comparative study of texture features for the discrimination of gastric polyps in endoscopic video. In: 18th IEEE Symposium on computer-based medical systems (CBMS'05), IEEE, pp 575–580. https://doi.org/10.1109/CBMS.2005.6
4. Hwang S, Oh J, Tavanapong W, Wong J, De Groen PC (2007) Polyp detection in colonoscopy video using elliptical shape feature. In: 2007 IEEE International conference on image processing, IEEE, 2:465. https://doi.org/10.1109/ICIP.2007.4379193
5. Bernal J, Sánchez J, Vilarino F (2012) Towards automatic polyp detection with a polyp appearance model. Pattern Recognition 45(9):3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002
6. Leung WK, Guo C-G, Ko MK, To EW, Mak LY, Tong TS, Chen L-J, But DY, Wong SY, Liu KS et al (2020) Linked color imaging versus narrow-band imaging for colorectal polyp detection: a prospective randomized tandem colonoscopy study. Gastrointestinal Endoscopy 91(1):104–112. https://doi.org/10.1016/j.gie.2019.06.031
7. Alexandre LA, Nobre N, Casteleiro J (2008) Color and position versus texture features for endoscopic polyp detection. In: 2008 International conference on biomedical engineering and informatics, IEEE, 2:38–42. https://doi.org/10.1109/BMEI.2008.246
8. Freedman JS, Harari DY, Bamji ND, Bodian CA, Kornacki S, Cohen LB, Miller KM, Aisenberg J (2011) The detection of premalignant colon polyps during colonoscopy is stable throughout the workday. Gastrointestinal endoscopy 73(6):1197–1206. https://doi.org/10.1016/j.gie.2011.01.019
9. Simmons DT, Harewood GC, Baron TH, Petersen BT, Wang KK, Boyd-Enders F, Ott BJ (2006) Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time. Alimentary pharmacology & therapeutics 24(6):965–971. https://doi.org/10.1016/j.gie.2006.03.026
10. Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, Van Riel SJ, Wille MMW, Naqibullah M, Sánchez CI, Van Ginneken B (2016) Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging 35(5):1160–1169. https://doi.org/10.1109/TMI.2016.2536809
11. Pang S, Ding T, Qiao S, Meng F, Wang S, Li P, Wang X (2019) A novel yolov3-arch model for identifying cholelithiasis and classifying gallstones on ct images. PloS one 14(6):0217647. https://doi.org/10.1371/journal.pone.0217647

12. Deeba F, Bui FM, Wahid KA (2020) Computer-aided polyp detection based on image enhancement and saliency-based selection. Biomed Signal Process Control 55:101530. https://doi.org/10.1016/j.bspc.2019.04.007
13. Qadir HA, Shin Y, Solhusvik J, Bergsland J, Aabakken L, Balasingham I (2021) Toward real-time polyp detection using fully cnns for 2d gaussian shapes prediction. Med Image Anal 68:101897. https://doi.org/10.1016/j.media.2020.101897
14. Taş M, Yılmaz B (2021) Super resolution convolutional neural network based pre-processing for automatic polyp detection in colonoscopy images. Comput Electrical Eng 90:106959. https://doi.org/10.1016/j.compeleceng.2020.106959
15. Chen B-L, Wan J-J, Chen T-Y, Yu Y-T, Ji M (2021) A self-attention based faster r-cnn for polyp detection from colonoscopy images. Biomed Signal Process Control 70:103019. https://doi.org/10.1016/j.bspc.2021.103019
16. Cao C, Wang R, Yu Y, Zhang H, Yu Y, Sun C (2021) Gastric polyp detection in gastroscopic images using deep neural network. PloS one 16(4):0250632. https://doi.org/10.1371/journal.pone.0250632
17. Nisha J, Gopi VP, Palanisamy P (2022) 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
18. Hu K, Zhao L, Feng S, Zhang S, Zhou Q, Gao X, Guo Y (2022) Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement. Comput Biol Med 147:105760. https://doi.org/10.1016/j.compbiomed.2022.105760
19. Zhu X, Lyu S, Wang X, Zhao Q (2021) 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, pp 2778–2788. https://doi.org/10.48550/arXiv.2108.11539
20. Tang Y, Han K, Guo J, Xu C, Li Y, Xu C, Wang Y (2022) An image patch is a wave: Phase-aware vision mlp. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 10935–10944. https://doi.org/10.48550/arXiv.2111.12294
21. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, Ning J, Cao Y, Zhang Z, Dong L, et al. (2022) Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12009–12019. https://doi.org/10.48550/arXiv.2111.09883
22. Liu S, Huang D, Wang Y (2019) Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516. https://doi.org/10.48550/arXiv.1911.09516
23. Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, Hu W, Wang L, Duan H, Si J (2019) Real-time gastric polyp detection using convolutional neural networks. PloS one 14(3):0214133. https://doi.org/10.1371/journal.pone.0214133
24. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD et al (2020) Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data 7(1):283. https://doi.org/10.1038/s41597-020-00622-y
25. Misawa M, Kudo S-E, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, Baba T, Ishida F et al (2021) Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy 93(4):960–967. https://doi.org/10.1016/j.gie.2020.07.060
26. Pacal I, Karaman A, Karaboga D, Akay B, Basturk A, Nalbantoglu U, Coskun S (2022) An efficient real-time colonic polyp detection with yolo algorithms trained by using negative samples and large datasets. Comput Biol Med 141:105031. https://doi.org/10.1016/j.compbiomed.2021.105031
27. Karaman A, Pacal I, Basturk A, Akay B, Nalbantoglu U, Coskun S, Sahin O, Karaboga D (2023) Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc). Expert Syst Appl 221:119741. https://doi.org/10.1016/j.eswa.2023.119741