Pattern Analysis and Applications (2024) 27:38
https://doi.org/10.1007/s10044-024-01254-3

SHORT PAPER

Tiny polyp detection from endoscopic video frames using vision transformers

Entong Liu[1] · Bishi He[1] · Darong Zhu[2] · Yuanjiao Chen[1] · Zhe Xu[1]

Received: 14 April 2023 / Accepted: 18 February 2024

Abstract

Deep learning techniques can be effective in helping doctors diagnose gastrointestinal polyps. Currently, processing video frame sequences containing a large amount of spurious noise in polyp detection suffers from elevated recall and mean average precision. Moreover, the mean average precision is also low when the polyp target in the video frame has large-scale variability. Therefore, we propose a tiny polyp detection from endoscopic video frames using Vision Transformers, named TPolyp. The proposed method uses a cross-stage Swin Transformer as a multi-scale feature extractor to extract deep feature representations of data samples, improves the bidirectional sampling feature pyramid, and integrates the prediction heads of multiple channel self-attention mechanisms. This approach focuses more on the feature information of the tiny object detection task than convolutional neural networks and retains relatively deeper semantic information. It additionally improves feature expression and discriminability without increasing the computational complexity. Experimental results show that TPolyp improves detection accuracy by 7%, recall by 7.3%, and average accuracy by 7.5% compared to the YOLOv5 model, and has better tiny object detection in scenarios with blurry artifacts.

Keywords Polyp detection · Endoscopic video analysis · Tiny object detection · Vision transformers · Gastrointestinal diseases

1 Introduction

Gastrointestinal endoscopy is essential for early diagnosis of colorectal and gastric cancer. During the examination, the doctor inserts a flexible tube with a miniature camera and guides it through the digestive tract to detect early precancerous lesions [1].Typically, the miss rate of endoscopy is over 15% [2], and the quality of the examination during surgery usually depends on the doctor's ability to avoid misdiagnosis, which requires a high level of professional knowledge and experience. In areas with poor medical conditions, due to the shortage of endoscopy doctors and different levels of operation, the missed diagnosis rate during fatigue can even reach 27% [1, 3], which not only delays patient treatment but also increases medical costs. Therefore, computer-aided examination is needed to improve the diagnostic capacity and technical level of doctors so as to better serve the majority of patients and reduce further development and deterioration of the disease.

With the continuous development and application of deep learning techniques, object detection has received a lot of attention as an influential research direction in computer vision. Different from traditional methods that manually extract features, deep learning object detectors can automatically learn image features to achieve more accurate and faster object detection. Currently, two different types of detectors are mainly used in deep learning detection tasks: one-stage detectors and two-stage detectors. They have their own strengths and weaknesses and are different in their applicability in different application scenarios. The two-stage detector first generates candidate regions that may contain the target through the region proposal network, and then classifies and regresses these candidate regions to achieve object detection. Common two-stage detectors include R-CNN, Fast R-CNN, and Faster R-CNN. Among them, Faster R-CNN is the most popular and classic one, which achieves end-to-end object detection by introducing Region

✉ Bishi He
hebs@hdu.edu.cn

[1] School of Automation (School of Artificial Intelligence), Hangzhou Dianzi University, Hangzhou, China

[2] Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China

Published online: 04 April 2024

Proposal Network (RPN), with higher detection accuracy and faster detection speed. Ruilin Wang et al. improved the Faster R-CNN [4] network, using ROI alignment instead of ROI pooling, and improving the operation mechanism of non-maximum suppression, achieving favorable results in detection [5]. The one-stage detector will not generate candidate boxes, and directly transform the problem of target border positioning into a regression problem, which has higher detection efficiency than the two-stage detector. Dai et al. proposed Faster R-CNN-based single-stage detector R-FCN, which merges detection and localization into a unified network with shared convolutional feature maps and position sensitive RoI pooling layers for faster and more accurate object detection. In some small target tasks, the detection accuracy of one-stage detectors has reached or even exceeded that of two-stage detectors [6]. Common one-stage detectors include YOLO, SSD, and RetinaNet. For example, Al-Masni et al. compared the performance of YOLOv2 and SSD algorithms in polyp detection in their paper, and found that YOLOv2 had better performance than the SSD algorithm [7]. Redmon et al. introduced a lightweight version of YOLOv3 called Tiny-YOLOv3. They evaluated the performance of Tiny-YOLOv3 and RetinaNet on the COCO dataset and found that Tiny-YOLOv3 was better than RetinaNet in tiny object detection [8]· As the YOLO algorithm is continuously updated, the detection accuracy of YOLOv5 increases by about 10% compared to YOLOv4. Moreover, compared with traditional deep learning algorithms, YOLOv5 has faster processing speed and lighter model size [9], but its ability to obtain global information is limited.

The Transformer model was originally proposed by Google and applied to natural language processing tasks. As the advantages of Transformer in processing lengthy sequence data continue to be discovered, some researchers have begun to explore its application in object detection tasks [10]. The vision Transformer (ViT) model proposed by Dosovitskiy et al. [11] splits image data into a text-like sequence similar to natural language processing. For example, the image is divided into a series of non-overlapping regions, and the feature representation of each region is used as an input sequence, which is then learned and fused using the Transformer model, thus demonstrating the feasibility of the Transformer model for object detection tasks. The Swin Transformer proposed by Liu et al. reduces the computational complexity and memory usage by introducing a Shifted Window mechanism, dividing the input image into tiny blocks, and independently calculating self-attention based on the blocks, while still maintaining excellent object detection performance [12, 13].

Although the above object detection methods commonly have higher accuracy than traditional algorithms, the model has a steep rate of missed detections when processing large sequences of video frames containing pseudo-noise, and the detection accuracy is also low when the scale of the polyp target varies largely in the video frames. Therefore, we propose a tiny polyp detection from endoscopic video frames using Vision Transformers, named TPolyp, which takes into account both global and local feature information and can adapt to extreme scale variations. The main contributions of this paper are as follows:

1. A sliding window-based local self-attention mechanism module is proposed as the backbone to capture local dependencies in input sequences.Compared with the traditional attention mechanism, our innovation is to apply the attention mechanism to the local window, effectively extract the correlation of local features, and thus enhance the modeling ability of the model on the input sequence.
2. The bidirectional feature pyramid network is improved as a Neck part to predict different objects by dynamically selecting different network depths. This innovative design enables the network to make more accurate predictions of objects at different scales at different levels, improving the performance of the model in multi-scale object detection tasks.
3. In the downsampling process, we introduce techniques such as channel attention module, cross-stage connection and channel rearrangement. The application of these innovative techniques helps to further extract the feature information between channels, enhance the receptive field and capture more context information. The design of the downsampling process using a variety of techniques can effectively improve the model's ability to represent image features.
4. The output of the model is divided into four scales, including minimal, small, medium and large. This innovative design enables the model to adapt to different detection difficulties and better handle detection tasks of different scale targets. By introducing multi-scale target processing, our model has stronger adaptability and generalization ability in target detection task.

The paper is structured as follows: Sect. 2 reviews related work. Section 3 describes the main steps of the research methodology. In Sect. 4, the dataset is described and ablation experiments are performed to evaluate performance metrics, demonstrate the working principle of TPolyp method, and provide directions for future research. Section 5 concludes the paper.

2 Related work

Polyps are an essential sign of early colon cancer, so the main purpose of examination is to detect them as early as possible to improve patient survival rates [14]. Automatic detection and localization of polyps in video frames of gastrointestinal endoscopy can help reduce missed and false detections in manual manipulation, improve detection quality and efficiency, and have positive implications for early detection of pre-cancerous lesions.

In recent years, machine learning and deep learning have been widely applied in medical imaging. Machine learning algorithms for polyp image detection mainly use traditional image processing techniques, requiring the manual design of feature extraction procedures and classifiers. Wang et al. [15] proposed a polyp detection framework including image preprocessing, feature extraction, feature selection, and classification steps. Zheng et al. [16] proposed a computer-aided diagnosis method based on image feature extraction and Fisher vector technology. Zhang et al. [17] proposed a polyp detection method based on SIFT features. These approaches rely on expertise and experience and have elevated data quality requirements. In contrast, deep learning algorithms can automatically learn features and have relatively low data quality requirements. Zacharaki et al. [18] used support vector machines and neural networks to detect polyps in computed tomography colonoscopy, employing multiple texture feature extraction methods and comparing their effect. The final results showed that neural networks performed better in detecting polyps with regular textures and had better performance [19]. Early machine learning detection algorithms relied on the color and texture features of polyps, but the large color changes between polyps and the limited visibility of surface textures hindered the applicability of the algorithms [20].

Wang et al. [21] used a region-based CNN model to detect and classify polyps in colonoscopy images, which can detect potential polyp regions in images and classify them. Fang et al. [22] used a method based on convolutional neural networks (CNN) and region-based CNN (R-CNN), where CNN was used to generate candidate regions and R-CNN was used for classification and localization of these candidate regions, achieving polyp region localization in colonoscopy images. These polyp detection methods require separate classification of each region in the image [23], resulting in moderate speed. Urban et al. [24] used an improved RCNN model, incorporating the techniques of Fast R-CNN and Faster R-CNN on the basis of the original RCNN model. This method uses the Region Proposal Network (RPN) to generate candidate regions and uses the RoI pooling layer for object classification and localization, demonstrating good performance.

Due to the diverse appearance of polyps in GI endoscopy images, two-stage detection methods suffer from problems such as mismatch between candidate box scales and targets, and high computational complexity for classification and regression. To better adapt to this characteristic, Xu et al. [25] and Li et al. [26] each made improvements and optimizations to the EfficientDet model from two perspectives. The former uses an attention mechanism based approach to extract image features, while the latter uses a feature pyramid network based approach to extract features at different scales. However, both use compound scaling methods to further improve detection efficiency and performance. This method can dynamically adjust the size and resolution of the input image according to the size and shape differences of polyps in gastrointestinal endoscopy, providing ideas for subsequent multi-scale research [27].

Bychkov et al. [28] first proposed to use ResNet and FPN networks to construct feature pyramids to extract features at different scales, and applied them to object detection tasks, providing ideas for the fusion of deep learning and feature pyramids. Later, Wang et al. [29] proposed a polyp detection method that combines the YOLO algorithm and FPN. Bertrand et al. [30] proposed a polyp detection method based on SSD and FPN. In addition, the network structure based on Focal Loss and FPN has also been proven to be feasible [31]. These methods all use FPN to construct feature pyramids and have achieved excellent results. While FPN performs better in the above scenarios, the most common obstacles in endoscopy detection include artifacts caused by motion, specular reflections, low contrast, bubbles, debris, body fluids, and blood, which are often confused with lesions. Each organ has specific limitations on the use of endoscopes, and the appearance, size, and shape of polyps in gastrointestinal endoscopy are also different, making accurate detection of polyps much more difficult [32–34]. In complex scenarios, the FPN can only construct a feature pyramid between the top and bottom layers of the backbone network and cannot fully integrate features from the middle layers, which may cause the detector to have difficulty in recognizing some medium-sized targets. In addition, the resolution of the features in each layer of the FPN varies significantly, which may cause the object detector to lose accuracy when working with small targets and fail to cover the entire target area when working with large targets.

There are a number of open questions about the FPN structure. Bogusz et al. [35] introduced a method based on PANet (Panoramic Attention Network). This approach starts by using ResNet as the backbone network to generate feature maps at different scales. Then, PANet is used to integrate these feature maps, adaptively adjusting the feature weights through a global attention mechanism, avoiding the problem of insufficient feature integration in FPN and adapting to feature resolution, so that it does not lose accuracy when processing small targets and can cover the entire image when processing large targets, avoiding the problem of feature resolution mismatch in FPN [36]. However, due to the complexity of the PANet network structure and the high requirements for training data and computing resources, Smith et al. [37] proposed an object detection method based on Simplified PANet. This approach first preprocesses colonoscopy images with a single threshold segmentation algorithm, and then uses Simplified PANet to achieve multi-scale feature fusion and object detection. Compared to PANet, Simplified PANet has advantages such as lower computational and storage overhead, simpler and more user-friendly network structure, and higher detection performance. However, it also has certain limitations compared to PANet, such as the inability to adaptively adjust the feature weights. In our approach, instead of the traditional convolutional neural network, we use a sliding window-based local self-attention mechanism and integrate a bidirectional adaptive feature selection mechanism to better integrate local low-level and high-level features. During the sampling process, techniques such as channel attention module, cross-stage connections, and channel reordering are introduced to enhance its global feature integration capability in complex scenes and automatically select useful features for detection tasks, thereby improving adaptability.

3 Method

Although many traditional object detection methods have excellent accuracy and speed in discontinuous image object detection tasks, in polyp detection of gastrointestinal endoscopy videos, due to the pseudo-images caused by camera shaking and the large variation in polyp image scales, there is a high demand for the extraction of global and local features, so there is a problem of low recall when dealing with complex scenes containing a large amount of pseudo-noise video frame sequences, as well as low detection accuracy in cases where the polyp target scale changes greatly in video frames. To overcome these issues in existing techniques, in this paper, we propose a tiny object detection algorithm for gastrointestinal endoscopy video frames. Figure 1 shows the overall framework of TPolyp.

The proposed TPolyp framework mainly consists of four parts. The first part is the data preprocessing and data augmentation module, which is used to obtain continuous video frames of gastrointestinal endoscopy images and preprocess them. Among them, there are eight alternative data augmentation methods that can increase data diversity and improve the generalization ability and robustness of the model. The second part is the backbone of this framework, which uses two multi-head self-attention mechanisms for feature extraction, allowing it to focus on different scales and levels of information simultaneously and thus better adapt to multi-level tasks. The third component is the feature pyramid, which connects the backbone and bidirectional adaptive feature selection. By using cross-stage connections and channel reordering, it solves the problem of inconsistent input image sizes, improves the effect of multiple feature fusion, and bidirectional sampling can better integrate features from different levels. The fourth part is the prediction head, which adds multiple self-attention mechanisms, channel attention modules, cross-stage connections, and channel reordering techniques based on multiple convolutional layers and fully connected layers to predict target class, location, confidence, and additional information.

3.1 Backbone network based on Swin transformer feature extraction

Different from previous polyp target detection methods, Swin Transformer is selected as a multi-scale feature extractor in this paper. The main network structure is shown in Fig. 2a, and swin block is shown in Fig. 2b. In this paper, the input image is divided into a series of overlapping small blocks of the same size. For each small block, its feature representation is extracted, and these feature representations make up the feature map. The feature map is then divided into chunks of the same size, each made up of four adjacent smaller chunks. These large block feature representations are composed of four small block feature representations. These chunks are gradually merged into larger chunks through several patch merging-like operations, eventually producing a global feature representation. Each merge operation merges four adjacent blocks into a larger block and applies a self-attention mechanism on the new block to fuse the feature representations of these blocks, as shown in Fig. 3. This self-attention mechanism is only computed within a local window, so when the window size is fixed, the computational complexity is also fixed. In this way, it can capture feature information at different scales and focus only on local prior knowledge when calculating self-attention mechanisms, thus reducing sequence length and computational complexity.

After data preprocessing, gastroenterology three-channel images containing polyps were input into the feature extractor. The height H and width W of the image are set at 640×640. After cross-stage connection and channel rearrangement operation, CSTR structure as shown in Fig. 1 is used to fuse CSPnet before swin block. The swin block contains the W-MSA module (window long head self-attention) and the SW-MSA module (sliding window long head self-attention). The W-MSA module can reduce the calculation of self-attention operation and feature mapping. SW-MSA module uses the offset window to realize the information exchange between different Windows, and further improves the capability of feature representation. These two self-attention structures are connected in series to form a block. At this time, the dimensions of the image do not change. Under the condition that the size of the output feature matrix is 20×20, the window size corresponding to the fourth stage is only 20×20. First, the image input patch segmentation module is divided into blocks, and every $4 \times 4 = 16$ adjacent

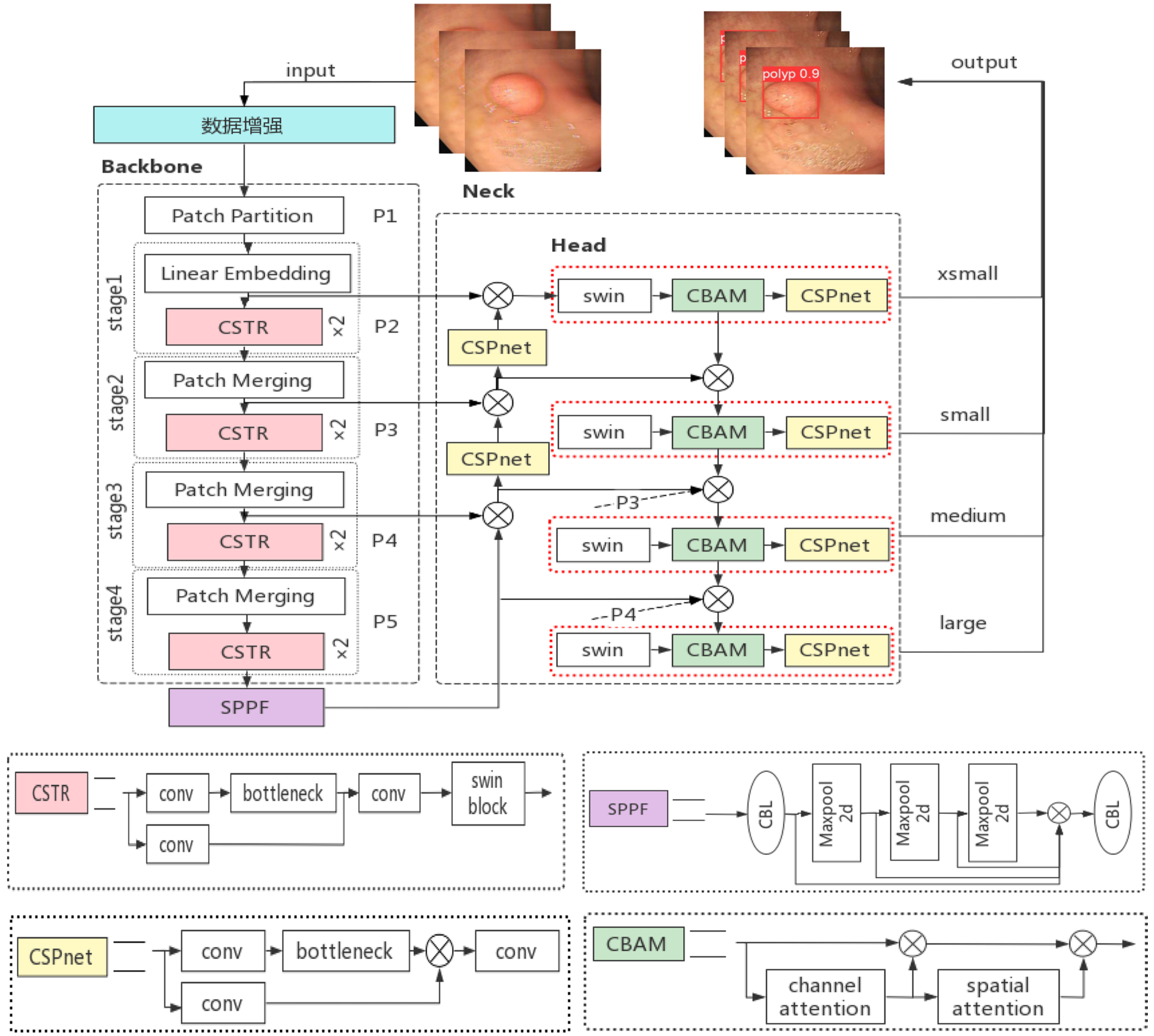


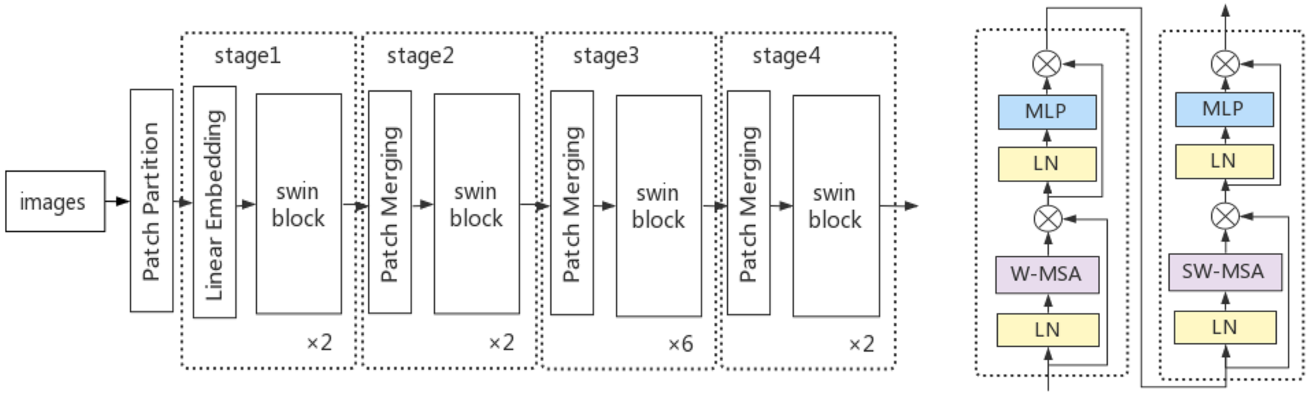
Fig. 1 The overall framework diagram of TPolyp

pixels are set as patches, and each pixel has three values of R, G, and B. After the channel direction is expanded, the channel data of each pixel is transformed linearly by linear embedding layer. Finally, the CSTR shown in Fig. 1 is repeatedly stacked in four stages. The image changes in the four stages of the trunk part model are (640, 640, 3) → (160, 160, 48) → (160, 160, 128) → (80, 80, 256) → (40, 40, 512) → (20, 20, 1024). Table 1 shows the structure configuration information of the four phases.

For polyp detection tasks in gastroenterology video frames, after the Swin Transformer network architecture, the SPPF structure as shown in Fig. 1 is adopted to divide the feature maps into blocks, and the features within each block are maximized to obtain fixed-size feature vectors. Then, the feature vectors of all blocks are spliced together to get the final feature representation. These processing methods only need to operate on the feature graph, so the detection accuracy of the model can be improved without increasing the amount of computation, and the final output can be obtained. At the same time, it can also aggregate different scale receptive fields without changing the size of the feature map, so as to enhance the ability of feature expression.

3.2 Feature pyramid with cross-stage connections and bi-directional sampling

Traditional object detection networks typically start feature fusion from the third layer of features. To improve



(a) main network structure **(b)** swin block

Fig. 2 Swin Transformer network structure

Fig. 3 Diagram of the patch merging process

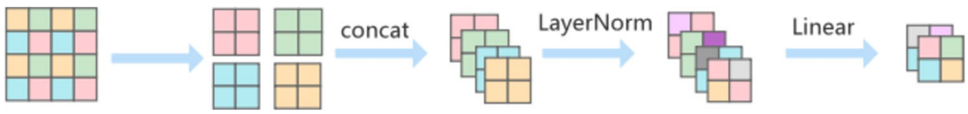
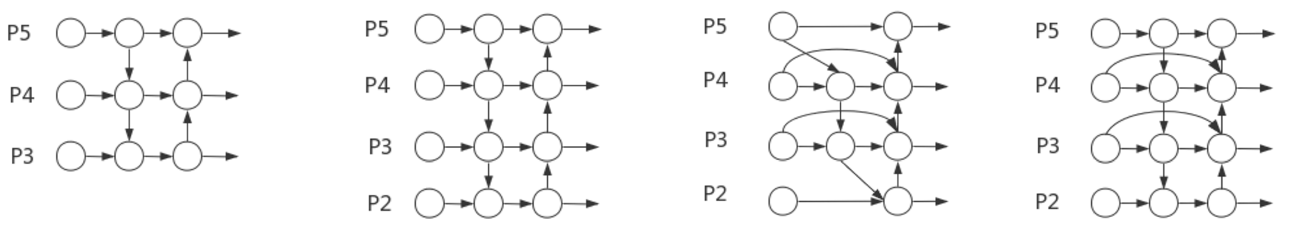


Table 1 Structural configuration information for the four stages

	Stage 1	Stage 2	Stage 3	Stage 4
Channel depth of feature map	128	256	512	1024
Number of heads of self-attentive modules	4	8	16	32
Number of blocks stacked	2	2	2	2

the network's ability to detect small objects, a small object detection layer is introduced in this work and a second feature layer is added to the feature fusion network to preserve shallow semantic information. A 160 × 160 feature map, which was not fused in the feature extraction network, was added to the detection layer, and an upsampling and downsampling operation was added to the feature fusion network to increase the number of detection layers to four. After the addition of the detection layer, the number of output prediction boxes is increased from 9 to 12, which are all prediction boxes with different aspect ratios for small object detection.

Conventional FPN structures have only one-way information flow from top to bottom. As shown in Fig. 4a, the PANet network adds an extra bottom-up path to enhance the information flow, effectively preserving more shallow features. BiFPN is a Google team's improved network structure based on PANet, as shown in Fig. 4c. The original BiFPN network fused features from the third to the seventh of the seven feature layers, and believed that a node with only one input edge would contribute less to the network. Therefore, the feature fusion nodes in the third and seventh layers are removed in this paper to reduce the computational cost. At the same time, proposed a cross-scale concatenation method by adding an additional edge to directly fuse features in the feature extraction network with features of similar size in the bottom-up path, which preserves more shallow semantic



(a) PANet **(b)** PANet+xsmall **(c)** BiFPN **(d)** Improved FPN

Fig. 4 Pyramid structure of the four types of bidirectional fusion features

information without losing too much relatively deep semantic information. The feature fusion network of YOLOv5 uses the PANet structure. As shown in Fig. 4b, since the network added a small object detection layer and added the second layer of features, which was not involved in feature fusion, to the feature fusion network, too much shallow semantic information was retained, leading to a serious loss of deep semantic information in the network and making the features relatively complex for the network. Therefore, it is important to retain more relatively deep semantic information. This study uses an improved bidirectional fusion network, as shown in Fig. 4d, which adds cross-scale connections to fuse more features without increasing too much computational cost.

3.3 Prediction head with multiple channel self-attention mechanisms

The prediction heads in this study include multiple channel self-attention mechanisms, which can aggregate and fuse features from different scales to improve the model's cross-scale detection capability. The multi-head self-attention mechanism can model contextual information in the feature maps, enable dynamic adjustment and fusion of the feature maps, and improve the detection accuracy and robustness of the model. By combining channel and spatial attention mechanisms and cross-stage partial networks, the feature expression and discriminability can be further improved without increasing the computational complexity, thus improving the accuracy and robustness of the model. Swin Block enables cross-scale feature fusion and solves the problem of detecting targets at different scales. By combining techniques such as CBAM and CSPNet, the cross-scale fusion ability of features can be further improved to improve the detection accuracy and robustness of the model. Meanwhile, the prediction head is shown in Fig. 3. Efficient feature fusion and prediction can be achieved with a small number of parameters, which optimizes the computational efficiency of the model and better adapts to the problem of detecting polyps with blurred features.

4 Experiments and results

4.1 Dataset

Two public datasets for polyp detection, LDPolypVideo [38] and Hyper-Kvasir [39], were used to evaluate different approaches. Among them, LDPolypVideo (training set: 22,310 frames, test set: 2479 frames). The video frame images in the training set have frame-level labels, that is, each frame image contains information about the size and location of the polyp. Hyper-Kvasir (training set: 889 frames, test set: 111 frames). The two datasets contain polyps of different sizes and shapes, as shown in Figs. 5 and 6. We kept the same data set Settings for TPolyp and all other methods for the sake of fairness of the experiment.

In Figs. 5 and 6a represents the annotated anchor boxes for all samples, (b) represents the distribution of anchor box positions in all samples, and (c) represents the distribution of anchor box dimensions in all samples.

4.2 Evaluation metrics

In object detection tasks, Precision, Recall, Precision-Recall curve (PR curve), mean Average Precision (mAP) are commonly used to evaluate the performance of the model. Here, mAP is the average of all class APs and is computed for the entire dataset. In this experiment, only one class needs to be detected and evaluated – with or without polyps. In

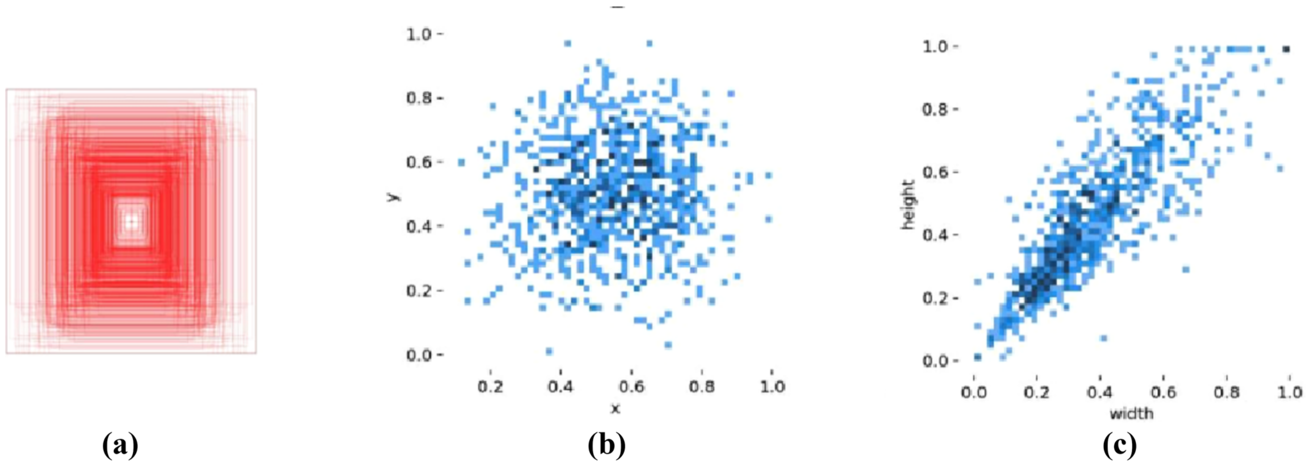


Fig. 5 Hyper-Kvasir dataset

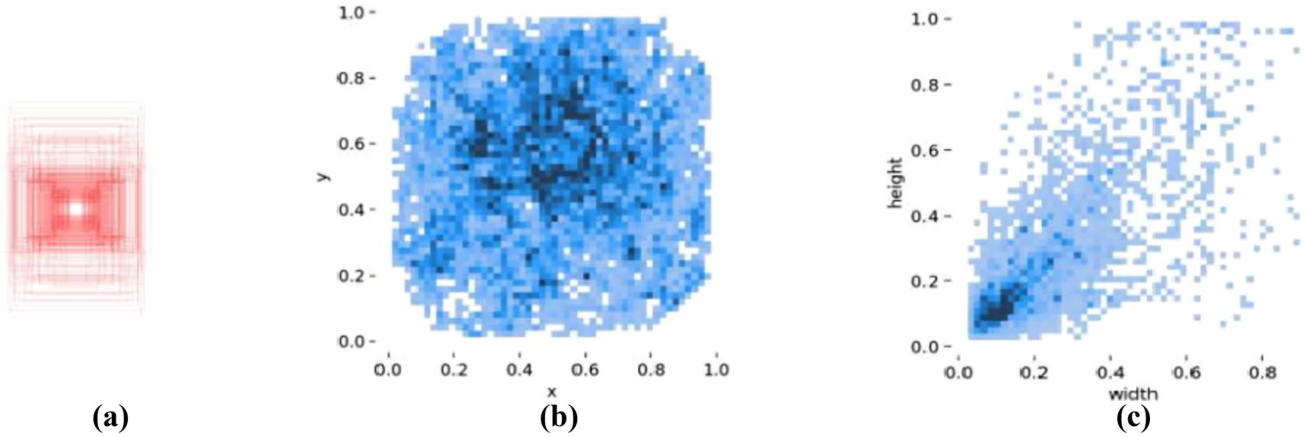


Fig. 6 LDPolypVideo dataset

the actual calculation, the number of positive and negative samples for the true and predicted bins needs to be counted, and TP, FP, FN, and TN are used to do this:

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/TP + FN \tag{2}$$

IoU is used as the basis for marking the boundary boxes as TP, FP, FN, and TN. The calculation method is as follows: Generally, the prediction box whose IoU is greater than 0.5 is marked as TP. The calculation of mAP0.5 refers to the average accuracy when the IoU threshold is 0.5. Therefore, the above target detection and estimation indicators can be calculated and defined as follows:

Mean Average Precision (mAP)—a comprehensive index of precision and recall, whose value is the area under the PR curve drawn with Recall as the horizontal axis and Precision as the vertical axis. Its discrete form is used in practice:

$$AP = \sum_{K=1}^{n} P(k)\Delta R(k) \tag{3}$$

In Eq. (3), $\Delta R(k)$ is an interval that is evenly divided into n segments between 0 and 1 on the x-axis

$$mAP = \frac{1}{|c|}\sum_{c\in C} AP(c) \tag{4}$$

In Eq. (4), c represents the class.

The loss function consists of three main parts: classification loss, object loss, and localization loss. In this study, both the classification loss and objectness loss are calculated using binary cross-entropy (BCE) loss. In practical applications, the model automatically calculates the predicted values, and by continuously adjusting the parameters of the loss function, the model can be continuously improved, ultimately achieving optimal performance [40]. The CIoU loss is used to compute the localization loss.The CIoU loss is based on the DIoU loss and adds a factor that considers the consistency of aspect ratios between the predicted and ground truth boxes [41, 42]. The formula is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{5}$$

Here, b and b^{gt} correspond to the center points of B and B^{gt}, respectively. ρ(•) denotes the Euclidean distance, and c represents the diagonal distance between the minimum enclosing rectangles of B and B^{gt}. α is a parameter used for balancing the ratio.

$$\alpha = \upsilon/(1 - IoU + \upsilon) \tag{6}$$

υ Here, is used to measure the consistency of aspect ratios, which is defined as follows:

$$\upsilon = \frac{4}{\Pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{7}$$

The CIoU loss function combines the advantages of GIoU and DIoU loss functions and can solve the problem of zero IoU loss value when the prediction box and ground truth box do not intersect. At the same time, this function takes into account the overlap area and distance between the central points of the predicted boxes, and adds a parameter that measures the aspect ratio agreement between the predicted and ground truth boxes. These improvements further improve the convergence rate and regression accuracy of the model.

4.3 Data enhancement

In this study, the following methods were used to carry out data enhancement, such as scaling and cropping at 50%

ratio, horizontal inversion, adjustment of HSV hue, saturation and brightness of images, etc., and Mixup algorithm [43] was used to mix images of different types through simple linear transformation of input image data. M During the training process, these enhancements will generate more training data, thereby improving the accuracy and robustness of the model. The TPolyp network model was trained by using enhanced images and unenhanced images of LDPolypVideo dataset, respectively, and two different training models were obtained to verify the influence of image enhancement on detection results. The test results are shown in Table 2. The mAP of the model on the unenhanced data set is 97.7%, while the model mAP on the enhanced data set is 99.5%, which confirms the necessity and effectiveness of data enhancement.

4.4 Experimental design and performance analysis

This study was implemented on top of Ubuntu 16.04 LTS 64-bit operating system, using Python programming language and PyTorch deep learning framework to train object detection models on NVIDIA GTX3080. The initial learning rate was set to 0.01, and the training loss was optimized using the momentum-based stochastic gradient descent (Momentum SGD) method. The number of iterations is set to 100, the batch size is set to 16, and the weight decay is set to 0.0005. Precision, Recall, and mean average precision (mAP) were used as evaluation metrics to comprehensively evaluate the performance of the deep learning algorithm.

Our model consists of three parts: backbone, feature pyramid, and prediction head. In its backbone, two multi-head self-attention mechanisms are used for feature extraction. The feature pyramid part, connecting the trunk and bidirectional sampling adaptive feature selection before using cross-stage connection and channel rearrangement, and then bidirectional sampling can better integrate features of different levels. In the last part of prediction, multi-head self-attention mechanism, channel attention module, cross-stage connection and channel rearrangement are added on the basis of multiple convolution layers and fully connected layers. In order to further verify the effectiveness of this research method, the model proposed in this study was compared with some mainstream target detection models, such as Faster-RCNN, CenterNet [44] and TransVOD [45], and the test results are shown in Table 3. As can be seen from Table 3, the detection results of the proposed model are better than those of Faster-RCNN, CenterNet, TransVOD and other detection models.

Figure 7 shows the visualization of different models. It can be seen that our model has a lower miss rate when body fluid reflection, lens blur and prediction target are small, which improves the prediction score and accuracy, and further proves the effectiveness and accuracy of this model.

In Fig. 7 The first row represents the image containing GT frame of polyps, and the second and third rows are the prediction results of YOLOv5 and Ours models, respectively. The first three columns of images are from the LDPolypVideo dataset and the last three columns of images are from the Hyper-Kvasir dataset. The bounding box is green for correct position and red for incorrect prediction.

Table 2 The effect of data enhancement on detection results

Data processing	Precision	Recall	mAP0.5
Un-Enhanced	0.978	0.980	0.977
Enhanced	**0.990**	**0.989**	**0.995**

Bold values represent the best results of various indicators generated after data augmentation

4.5 Ablation experiments

To verify the effectiveness of the research approach described in this paper, we perform ablation experiments on the modified network model. In this study, YOLOv5 is used as the baseline network and seven different architectures are trained, including CSTR as the backbone, improved cross-stage bidirectional sampling as Neck, and multiple channel self-attention mechanisms fused in the prediction head, to validate their advantages and show the effect parameter of their combination.The specific performance of each module after improvement is shown in Table 4.

The backbone consists of a stack of convolutional and CSPNet structures, and Neck uses a PANet module with

Table 3 Experimental results of different models

Model	LDPolypVideo			Hyper-Kvasir		
	Precision	Recall	mAP0.5	Precision	Recall	mAP0.5
Faster-RCNN	0.772	0.696	0.732	0.688	0.467	0.556
CenterNet	0.746	0.654	0.697	0.706	0.438	0.540
TransVOD	0.793	0.696	0.741	0.919	0.920	0.920
YOLOv5	0.922	0.914	0.920	0.906	0.845	0.875
Ours	**0.990**	**0.989**	**0.995**	**0.956**	**0.973**	**0.970**

Bold values represent the best performance results of the five models on two different datasets

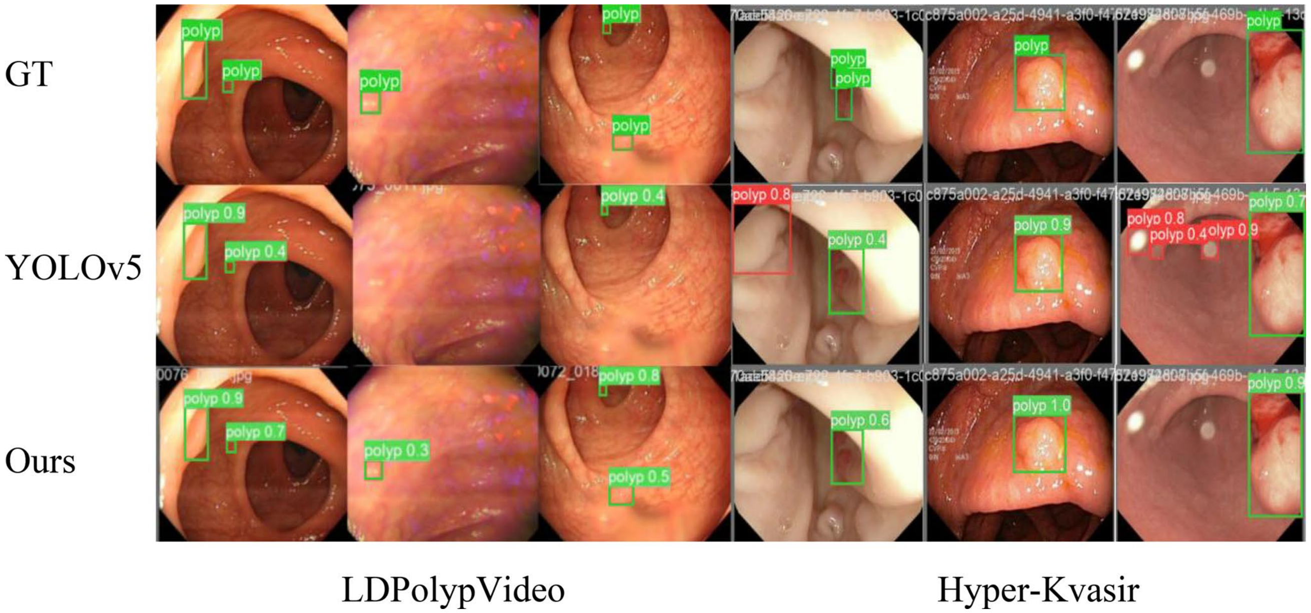


Fig. 7 Comparison of detection performance between YOLOv5 and TPolyp on LDPolypVideo and Hyper Kvasir datasets

Table 4 Results of ablation experiments

CSTR-backbone	Improved-FPN	SCC-head	Precision	Recall	mAP0.5
			0.922	0.914	0.920
√			0.969	0.925	0.954
	√		0.963	0.937	0.950
		√	0.983	0.920	0.925
√	√		0.984	0.971	0.988
√		√	0.956	0.973	0.985
√	√	√	**0.990**	**0.989**	**0.995**

Bold values represent the best results from the ablation experiment

three upsampling layers to ensure that the output of the prediction head includes four scales for tiny object detection. In the second set, the backbone part is used as a variable and a Swin Transformer based feature extraction network is embedded, resulting in an average accuracy increase of 3.4 percentage points after aggregating the receptive fields at different scales. In the third set, the Neck part is used as a variable and replaced with improved cross-level bidirectional sampling to better preserve features at various levels. Experimental results show an average accuracy increase of 3 percentage points. In the fourth set, only the prediction head was changed to a more sophisticated attention mechanism, with only a slight improvement in detection metrics. The first four experiments show that all three changes contribute to optimizing the detection performance, but the effect of the last two improvements is not as significant as the second improvement in the backbone. This may be because adding additional feature fusion steps does not achieve the best results without accurately capturing rich local and global information. Therefore, we perform the fifth and sixth experiments, keeping the improved backbone and adding cross-stage bidirectional sampling and multi-channel prediction heads, respectively. The results show that combining the changes in the backbone and neck, as well as the backbone and prediction head, both increase the average accuracy by about six percentage points. Finally, the model proposed in this study integrates all three improvements and achieves the best performance compared to the baseline, with a 7% increase in accuracy, a 7.3% increase in recall, and a 7.5% increase in average precision, outperforming alternative detection models.

5 Research limitations and future directions

Currently, the number of gastrointestinal endoscopy datasets is still limited and most of them contain noise and other interference factors. Compared to natural image datasets, the quality of these datasets still needs to be improved. Therefore, future research can focus on designing video denoising algorithms and expanding endoscopy datasets, and analyzing the robustness of the algorithms.

Global information on individual discontinuous images is relatively limited in the video detection process of gastrointestinal endoscopy. When significant noise and artifacts are present, the use of consecutive video frames can raise recall index. However, due to the limited information interaction between video frames, the performance under blurry detection conditions still lags behind that of sharp video frames. Therefore, future research can consider detecting global features of video segments. Video clips cover temporal and spatial information interactions and contain additional latent feature information. Moreover, in the presence of significant artifacts and noise, video segments can complement each other and share information, thereby improving the detection capability and raising the recall index. At the same time, improving the accuracy and generalization of the detection to be closer to real-world scenarios is a critical area for future research.

6 Conclusion

In this paper, we propose a tiny polyp detection from endoscopic video frames using Vision Transformers, named TPolyp, which addresses the characteristics of polyps in video frames, such as varying sizes, significant artifacts and noise, and complex feature information. The algorithm mainly consists of a backbone network based on Swin Transformer feature extraction, a feature pyramid with inter-stage connections and bidirectional sampling, and a prediction head that integrates multiple channel self-attention mechanisms. By further extracting inter-channel features, increasing the receptive field, and capturing additional information, the model has the ability to obtain local dependencies in the input sequence and adapt to four types of target variations from ultra-narrow to large scales, with the ability to detect, localize, and provide prediction scores. Compared to the YOLOv5 model, the proposed model improves mAP by 7%, recall by 7.3%, and mean accuracy by 7.5%, demonstrating better overall performance. In addition, this research could assist less experienced imaging physicians in medical diagnosis by helping them detect lesions that are difficult to identify with the naked eye, reducing false negative rates, improving diagnostic accuracy and detection efficiency, and promoting early detection and treatment of diseases, thereby improving patient survival.

Acknowledgements This work was supported by the Science and Technology Plan Project of Hangzhou China (Grant No. 2021WJCY258).

Author contributions Guarantors of integrity of entire study, all authors; study concepts/study design, Bishi He, Entong Liu, Zhe Xu; data acquisition, Darong Zhu,Yuanjiao Chen; data analysis and interpretation, Bishi He,Entong Liu, Darong Zhu; manuscript drafting or manuscript revision for important intellectual content, Bishi He, Entong Liu, Zhe Xu; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; experimental studies, all authors.

Declarations

Conflict of interest The authors declare that they have no conflict of competing interests.

Ethical and informed consent The datasets analyzed during this study are obtained from public datasets. LDPolypVideo: https://github.com/dashishi/LDPolypVideo-Benchmark. Hyper-Kvasir: https://datasets.simula.no/hyper-kvasir/

References

1. Ahn SB, Han DS, Bae JH, Byun TJ et al (2012) The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut Liver 6(1):64
2. Lee J, Park SW, Kim YS et al (2017) Risk factors of missed colorectal lesions after colonoscopy. Medicine 96(27):e7468
3. Pu LZCT et al (2020) Computer-aided diagnosis for characterisation of colorectal lesions: a comprehensive software including serrated lesions. Gastrointest Endosc 92:891–899
4. Ren S et al (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
5. Wang R, Zhang W, Nie W, Yu Y (2020) Gastric polyps detection by improved faster R-CNN. In: Proceedings of the 2019 8th international conference on computing and pattern recognition (ICCPR '19). Association for Computing Machinery, New York, NY, USA, pp 128–133. https://doi.org/10.1145/3373509.3373524
6. Ren S et al (2017) Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
7. Al-Fedaghi S, Bayoumi M (2019) Authentication modeling with five generic processes. Int J Adv Comput Sci Appl (IJACSA). https://doi.org/10.14569/IJACSA.2019.0100947
8. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767
9. Bochkovskiy A et al (2020) YOLOv5: improved performance, and on-device training. arXiv preprint arXiv:2006.05597
10. Vaswani A et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T et al (2021). An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
12. Su J, Zhou B, Jie Z, Zhu J, Ding C, Zhuang Y, Liu S, Li G, Wang Y, Li Z, Xiao B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10257–10266

13. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z et al. (2021). Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030
14. Siegel R, DeSantis C, Jemal A (2014) Colorectal cancer statistics, 2014. CA A Cancer J Clin 64(2):104–117
15. Wang Y, Dorner S, Ecker R (2010) A framework for automatic polyp detection in colonoscopy images. Med Image Anal 14(4):616–629
16. Zheng Y, Wang X, Song Y, Wang H (2018) Computer-aided diagnosis for colonoscopy by using bag-of-visual-words and Fisher vector techniques. J Med Syst 42(2):31
17. Zhang X, Chen Y, Song Y (2016) A novel approach for automated polyp detection in colonoscopy images via SIFT features. J Med Syst 40(6):136
18. Zhou SK et al (2021) A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 109(5):820–838. https://doi.org/10.1109/JPROC.2021.3054390
19. Zacharaki et al (2009) A comparative study of texture features for the detection of colonic polyps in computed tomography colonography
20. Tajbakhsh N et al (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312. https://doi.org/10.1109/tmi.2016.2535302
21. Wang P, Xiao X, Glissen Brown JR, Berzin TM (2018) Automatic detection of colonic polyps in endoscopic images using region-based convolutional neural networks. IEEE J Biomed Health Inform 22(5):1495–1505
22. Fang Y, Zhang J, Zhang Y, Gao Y (2016) Polyp detection using convolutional neural networks and region-based fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention, vol 9902, pp 62–70
23. Wang Y, Li L, Wang H, Gao X, Xia Y (2016) Polyp detection in colonoscopy videos using region-based convolutional neural networks. In: International conference on medical image computing and computer-assisted intervention, vol 9901, pp 473–481
24. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W et al (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155(4):1069–1078
25. Xu Y, Chen W, Zhang X, Wang J (2021) EfficientDet-based colonic polyp detection in colonoscopy images. IEEE Trans Med Imaging 40(1):73–83
26. Li H, Li X, Liang J, Li F (2020) EfficientDet-based automatic polyp detection for colonoscopy images. IEEE J Biomed Health Inform 24(2):566–574
27. Tan M, Le QV (2020) EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
28. Bychkov D, Linder N, Annus P, Kõks S (2018) Detecting lesions in colorectal cancer with deep learning. Med Image Anal 49:88–97. https://doi.org/10.1016/j.media.2018.04.002
29. Wang Z, Dong D, Wu L, Chen S, Liu F (2018) Towards accurate polyp detection with YOLO. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 1576–1580. https://doi.org/10.1109/BIBM.2018.8621135
30. Bertrand R, Marion R, Boudiaf M, Chambon S (2019) Towards real-time lesion detection in colonoscopy using single shot detectors. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), pp 1003–1007. https://doi.org/10.1109/ISBI.2019.8759374
31. Wang S, Wang R, Zhang X, Wang L, Zhang J (2020) Polyp detection in colonoscopy using focal loss convolutional neural networks. J Healthcare Eng 2020:8895832. https://doi.org/10.1155/2020/8895832
32. Pu LZCT, Maicas G, Tian Y, Yamamura T, Nakamura M, Suzuki H, Singh G, Rana K, Hirooka Y, Burt AD et al (2020) Computer-aided diagnosis for characterisation of colorectal lesions: a comprehen-sive software including serrated lesions. Gastrointest Endosc 92:891–899
33. Liu Y, Tian Y, Maicas G, Pu LZCT, Singh R, Verjans JW, Carneiro G (2020) Photoshopping colonoscopy video frames. In: 2020 IEEE 17th international symposium on biomedical imaging (ISBI). IEEE, pp 1–5
34. Tajbakhsh N et al (2015) Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In: 2015 IEEE 12th international symposium on biomedical imaging (ISBI). https://doi.org/10.1109/isbi.2015.7163821.
35. Bogusz A, Moscicki J, Skomorowski M et al (2020) Polyp detection in colonoscopy images using panoramic attention network. IEEE J Biomed Health Inform 24(10):2926–2935. https://doi.org/10.1109/JBHI.2020.3003653
36. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
37. Smith J (2020) Simplified PANet for polyp detection in colonoscopic images. IEEE Trans Med Imaging 39(8):2560–2569. https://doi.org/10.1109/TMI.2020.2975962
38. Ma Y, Chen X, Cheng K, Li Y, Sun B (2021) LDPolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: International conference on medical image computing and computer-assisted intervention. Springer, Berlin, pp 387–396
39. Borgli H et al (2020) Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data 7(1):1–14
40. MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge
41. Rezatofighi H, Tsoi N, Gwak JY et al (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 658–666
42. Zheng Z, Wang P, Liu W et al (2020) Distance-IoU loss: faster and better learning for bounding box regression. In: AAAI, pp 12993–13000
43. Zhang H et al (2017) mixup: Beyond empirical risk minimization
44. Zhou X, Wang D, Philipp K (2019) Objects as points
45. Zhou Q et al (2022) TransVOD: end-to-end video object detection with spatial-temporal transformers

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.