ISSN 1007-9327 (print)
ISSN 2219-2840 (online)

World Journal of *Gastroenterology*

World J Gastroenterol 2024 December 28; 30(48): 5104-5224

Published by Baishideng Publishing Group Inc

World Journal of Gastroenterology

Contents

Weekly Volume 30 Number 48 December 28, 2024

EDITORIAL

ORIGINAL ARTICLE

Retrospective Study

Clinical Trials Study

Observational Study

Basic Study

LETTER TO THE EDITOR

ABOUT COVER

Editorial Board Member of *World Journal of Gastroenterology*, Angela Peltec, PhD, Associate Professor, Department of Internal Medicine, Discipline of Gastroenterology, State University of Medicine and Pharmacy "Nicolae Testemitanu", Chishinev 2019, Moldova. apeltec@yahoo.com

AIMS AND SCOPE

The primary aim of *World Journal of Gastroenterology* (*WJG*, *World J Gastroenterol*) is to provide scholars and readers from various fields of gastroenterology and hepatology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online. *WJG* mainly publishes articles reporting research results and findings obtained in the field of gastroenterology and hepatology and covering a wide range of topics including gastroenterology, hepatology, gastrointestinal endoscopy, gastrointestinal surgery, gastrointestinal oncology, and pediatric gastroenterology.

INDEXING/ABSTRACTING

The *WJG* is now abstracted and indexed in Science Citation Index Expanded (SCIE), MEDLINE, PubMed, PubMed Central, Scopus, Reference Citation Analysis, China Science and Technology Journal Database, and Superstar Journals Database. The 2024 edition of Journal Citation Reports® cites the 2023 journal impact factor (JIF) for *WJG* as 4.3; Quartile: Q1. The *WJG*'s CiteScore for 2023 is 7.8.

RESPONSIBLE EDITORS FOR THIS ISSUE

NAME OF JOURNAL
World Journal of Gastroenterology

ISSN
ISSN 1007-9327 (print) ISSN 2219-2840 (online)

LAUNCH DATE
October 1, 1995

FREQUENCY
Weekly

EDITORS-IN-CHIEF
Andrzej S Tarnawski

EXECUTIVE ASSOCIATE EDITORS-IN-CHIEF
Jian-Gao Fan (Chronic Liver Disease)

EDITORIAL BOARD MEMBERS
http://www.wjgnet.com/1007-9327/editorialboard.htm

PUBLICATION DATE
December 28, 2024

PUBLISHING PARTNER
Shanghai Pancreatic Cancer Institute and Pancreatic Cancer Institute, Fudan University
Biliary Tract Disease Institute, Fudan University

INSTRUCTIONS TO AUTHORS
https://www.wjgnet.com/bpg/gerinfo/204

GUIDELINES FOR ETHICS DOCUMENTS
https://www.wjgnet.com/bpg/GerInfo/287

GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH
https://www.wjgnet.com/bpg/gerinfo/240

PUBLICATION ETHICS
https://www.wjgnet.com/bpg/GerInfo/288

PUBLICATION MISCONDUCT
https://www.wjgnet.com/bpg/gerinfo/208

POLICY OF CO-AUTHORS
https://www.wjgnet.com/bpg/GerInfo/310

ARTICLE PROCESSING CHARGE
https://www.wjgnet.com/bpg/gerinfo/242

STEPS FOR SUBMITTING MANUSCRIPTS
https://www.wjgnet.com/bpg/GerInfo/239

ONLINE SUBMISSION
https://www.f6publishing.com

PUBLISHING PARTNER's OFFICIAL WEBSITE
https://www.shca.org.cn
https://www.zs-hospital.sh.cn

World Journal of Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

World J Gastroenterol 2024 December 28; 30(48): 5111-5129

DOI: 10.3748/wjg.v30.i48.5111

ISSN 1007-9327 (print) ISSN 2219-2840 (online)

ORIGINAL ARTICLE

Retrospective Study

Image detection method for multi-category lesions in wireless capsule endoscopy based on deep learning models

Zhi-Guo Xiao, Xian-Qing Chen, Dong Zhang, Xin-Yuan Li, Wen-Xin Dai, Wen-Hui Liang

Specialty type: Gastroenterology and hepatology

Provenance and peer review: Unsolicited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's classification

Scientific Quality: Grade B, Grade C, Grade C, Grade C

Novelty: Grade B, Grade B, Grade B, Grade C

Creativity or Innovation: Grade B, Grade B, Grade B, Grade C

Scientific Significance: Grade A, Grade B, Grade B, Grade B

P-Reviewer: Gao J; Long Z; Pan Y

Received: June 2, 2024

Revised: September 8, 2024

Accepted: October 8, 2024

Published online: December 28, 2024

Processing time: 179 Days and 19.7 Hours

Zhi-Guo Xiao, Xian-Qing Chen, Dong Zhang, Xin-Yuan Li, Wen-Xin Dai, Wen-Hui Liang, School of Computer Science Technology, Changchun University, Changchun 130022, Jilin Province, China

Zhi-Guo Xiao, School of Computer Science Technology, Beijing Institute of Technology, Beijing 100811, China

Corresponding author: Zhi-Guo Xiao, PhD, Additional Professor, School of Computer Science Technology, Changchun University, No. 6543 Weixing Road, Chaoyang District, Changchun 130022, Jilin Province, China. 3220215169@bit.edu.cn

Abstract

BACKGROUND

Wireless capsule endoscopy (WCE) has become an important noninvasive and portable tool for diagnosing digestive tract diseases and has been propelled by advancements in medical imaging technology. However, the complexity of the digestive tract structure, and the diversity of lesion types, results in different sites and types of lesions distinctly appearing in the images, posing a challenge for the accurate identification of digestive tract diseases.

AIM

To propose a deep learning-based lesion detection model to automatically identify and accurately label digestive tract lesions, thereby improving the diagnostic efficiency of doctors, and creating significant clinical application value.

METHODS

In this paper, we propose a neural network model, WCE_Detection, for the accurate detection and classification of 23 classes of digestive tract lesion images. First, since multicategory lesion images exhibit various shapes and scales, a multidetection head strategy is adopted in the object detection network to increase the model's robustness for multiscale lesion detection. Moreover, a bidirectional feature pyramid network (BiFPN) is introduced, which effectively fuses shallow semantic features by adding skip connections, significantly reducing the detection error rate. On the basis of the above, we utilize the Swin Transformer with its unique self-attention mechanism and hierarchical structure in conjunction with the BiFPN feature fusion technique to enhance the feature representation of multicategory lesion images.

RESULTS
The model constructed in this study achieved an mAP50 of 91.5% for detecting 23 lesions. More than eleven single-category lesions achieved an mAP50 of over 99.4%, and more than twenty lesions had an mAP50 value of over 80%. These results indicate that the model outperforms other state-of-the-art models in the end-to-end integrated detection of human digestive tract lesion images.

CONCLUSION
The deep learning-based object detection network detects multiple digestive tract lesions in WCE images with high accuracy, improving the diagnostic efficiency of doctors, and demonstrating significant clinical application value.

Key Words: Human digestive tract; Artificial intelligence; Deep learning; Wireless capsule endoscopy; Object detection

Core Tip: In clinical practice, wireless capsule endoscopy is commonly used to detect lesions in the digestive tract and search for their causes. Here, we propose a multilesion classification and detection model to automatically identify 23 types of lesions in the digestive tract, and accurately mark the lesions. The model can improve the diagnostic efficiency of doctors and their ability to identify the categories of digestive tract lesions.

Citation: Xiao ZG, Chen XQ, Zhang D, Li XY, Dai WX, Liang WH. Image detection method for multi-category lesions in wireless capsule endoscopy based on deep learning models. *World J Gastroenterol* 2024; 30(48): 5111-5129
URL: https://www.wjgnet.com/1007-9327/full/v30/i48/5111.htm
DOI: https://dx.doi.org/10.3748/wjg.v30.i48.5111

INTRODUCTION

Diseases of the human digestive tract involve the esophagus, stomach, small intestine, and large intestine, encompassing a wide variety of types and symptoms, with common diseases including esophagitis, gastritis, hemorrhage, and duodenal ulcers. In 2019, 7.32 billion cases of digestive diseases were reported, with 2.86 billion prevalent cases, resulting in 8 million deaths[1]. The global burden of digestive tract diseases is enormous, especially in less developed regions, where inadequate early screening and diagnosis result in high mortality rates. Early diagnosis can significantly improve treatment outcomes, improve patient prognosis, and reduce treatment costs. Therefore, improving the early diagnosis rate of digestive tract diseases has become an important challenge in the medical field[2].

Wireless capsule endoscopy (WCE) technology is an important medical technology for the examination of the digestive tract and other organs[3]. This has led to the widespread use of WCE technology in clinical examinations of the digestive tract since its introduction in 2001. WCE enables noninvasive examination of the entire digestive tract through a microcapsule that is swallowed by the patient, covering areas of the small intestine that are difficult to reach with conventional endoscopes. This technological advantage makes WCE important in the early screening of digestive diseases. WCE is simple, noninvasive, requires no anesthesia for the patient, and the examination is easy to perform in daily life, which significantly improves patient acceptance. However, WCE has several significant shortcomings. First, WCE takes continuous pictures inside the patient's body for 6 to 8 hours, generating approximately 50000 to 80000 images[4-6]. The processing and analysis of these images requires many human resources. Even for experienced endoscopists, meticulous analysis of images takes at least 2 to 3 hours. This high-intensity manual processing is not only time-consuming but also prone to fatigue, increasing the risk of a missed diagnosis and misdiagnosis. Second, owing to the complex anatomy of the human digestive tract, lesions may present different morphologies and characteristics in different parts of the body, which further increases the difficulty of image analysis. In particular, certain lesions that are small, irregular in shape, or obstructed by digestive tract contents may not be accurately identified in the image, increasing the possibility of misdiagnosis.

Deep learning methods enable automatic feature learning, reducing the dependence on manually crafted features. Moreover, technologies in computer vision offer tools and methods for medical image processing, empowering doctors to analyze and comprehend images for easier diagnosis. These include object detection, image segmentation, feature extraction, and shape analysis, among others. Among them, object detection is the most critical, and choosing a good object detection method can efficiently and accurately detect the pathology of WCE images. Moreover, the emergence of outstanding classification networks such as AlexNet[7], VGGNet[8], and ResNet[9], which are based on deep learning methods, along with the R-CNN[10-12] series, YOLO[13-22], SSD[23], and other classical object detection models, has contributed to the detection of WCE lesion images. Therefore, rapid advancements have been made in lesion detection techniques based on WCE images.

Unfortunately, the detection of digestive tract lesions is a challenging task because of poor image quality. First, WCE images frequently suffer from noise and low visual quality owing to the complex environment and poor lighting conditions in the digestive tract. Second, numerous contents in the digestive tract, such as food, stool, bile, and air

bubbles, can affect lesion detection. Moreover, the uncontrolled peristaltic motion of the WCE as it traverses the digestive tract leads to highly variable image quality. Moreover, different parts of the digestive tract, such as the esophagus, stomach, and small intestine, present a variety of colors and textures. Finally, different types of disease lesions (*e.g.*, polyps, tumors, ulcers, *etc.*) vary significantly in morphology, size, and color, making it difficult for traditional image analysis algorithms to detect all types of lesions accurately in complex and diverse lesion images. These challenges greatly complicate the detection of digestive tract lesions in WCE images.

In response to the aforementioned challenges associated with multicategory WCE lesion images of the digestive tract, drawing inspiration from the YOLOv8 model structure, we propose the design of the WCE_Detection model based on the YOLOv8 backbone network to address the challenging task of multicategory lesion image detection. The main contributions of this paper are as follows: (1) To the best of our knowledge, this paper presents the first deep learning framework designed specifically for the detection of 23 classes of digestive tract lesions in WCE images; (2) The neck part is redesigned by introducing BiFPN, which significantly enhances feature extraction and detail information capture. Moreover, the Swin Transformer module is incorporated to improve the network's ability to localize regions of interest accurately in large-area images through its self-attention mechanism, thereby enhancing the network capacity to handle complex scenes and object diversity, leading to improved detection accuracy and efficiency; (3) A new detection head is added to the original three detection heads in the head part. This allows the network to extract feature information more profoundly, facilitating better capture and learning of the multilevel features of the object, thereby improving detection accuracy; and (4) Our proposed WCE_Detection detector achieves the detection of 23 types of digestive tract lesions through an end-to-end approach. The detection results of the model on the WCE lesion image dataset outperform those of advanced detection frameworks, indicating that the model is clinically valuable.

MATERIALS AND METHODS

Dataset construction

This study utilized a self-constructed WCE lesion image dataset, as shown in Figure 1, derived from PillCamSB series capsule images. The dataset comprises 1374 WCE images saved in JPG format, encompassing 23 lesion categories, including gastric ulcers, colon polyps, and others. To ensure the uniformity of the dataset, we preprocessed the collected images of different sizes and uniformly resized them to 640 × 640 × 3 pixels in RGB channel format. This resolution choice was made to consider the efficient use of computational resources while ensuring image detail. Each image was manually annotated by a professional physician *via* LabelImg open source data annotation software (version v1.8.5) developed by HeartEx Labs, which strictly adhered to the following rules. First, optimal labeling involves expanding the original lesion's length and width by approximately 1.33 times. Second, if the lesion site appeared slender and irregular, the expansion area of the label was appropriately reduced or not expanded. Third, the labeling cannot exceed the effective area of the image region[24]. Once labeling was completed, the coordinates and categories of the lesion positions were saved as .txt files. The number of labeled lesions corresponded to the number of labeled bounding boxes in each image. To safeguard patient privacy, we anonymized the data by removing any personal information that could identify the patient. We adhered to relevant regulations and ethical norms, especially medical information privacy regulations, during dataset construction to ensure adequate protection of patients' personal privacy.

Dataset splitting and augmentation

To evaluate the effectiveness of our detection model, we partitioned the dataset into a training set, a validation set, and a testing set, with percentages of 70%, 20%, and 10%, respectively. To optimize the detection model, we combined the training and validation sets to construct the training dataset. During training, we observed that without employing data augmentation, the effectiveness of training was compromised because of the limited total number of WCE lesion images and the variability in image quality. Therefore, the model faced challenges in capturing WCE image patterns. We employed an online data augmentation strategy to enhance the training data in real time, aiming to improve the model's ability to handle complex scenes. We used image enhancement techniques such as mosaic, mixup, and HSV in the training process. Mosaic was used to stitch four different images to expand the field of view and improve the detection of multiple categories of lesions. Mixup was used to enrich the distribution of the training samples by linearly combining the two images and enhancing the robustness to different lesion features. HSV was used to simulate different lighting conditions and improve the adaptability to image quality by adjusting the hue, saturation, and luminance. By adjusting hue, saturation, and brightness to simulate different lighting conditions, the adaptability to image quality can be enhanced. These data enhancement strategies significantly increased the diversity of the training data and alleviate the challenges associated with the limited number of images and large quality variations. With the increase of the enhancement data, the performance of the model on the validation and test sets significantly improved, and the occurrence rates of leakage and misdetection significantly decreased.

Experimental setup

We implemented the WCE_Detection model *via* PyTorch 2.1.0. All our models were trained and tested *via* a Tesla V100-SXM2-16GB GPU from the Alibaba Cloud. During the training phase, we utilized a pretrained weight file from YOLOv8m, significantly reducing the training time. We trained for 300 epochs on the training set *via* the AdamW optimizer with an initial learning rate of 0.01. The batch size was set to 16, and the weight_decay was set to 0.0005. The specific parameter settings are shown in Table 1.

Table 1 Experimental parameter settings

Parameter name	Parameter value
Initial learning rate	0.01
Learning rate float	0.01
Epochs	300
Batch size	16
Optimizer	AdamW
Weight_decay	0.0005
Momentum	0.937

YOLOv8 model comparison and limitations

YOLOv8 (you only look once version 8) is a prominent algorithm in the field of object detection, and consists of five models: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x[20]. When training the model with our own images of 23 types of WCE lesions, we discovered that the results of YOLOv8m significantly outperformed those of YOLOv8n, YOLOv8s, YOLOv8l, and YOLOv8x, with a difference of > 3% between the mAP50 and mAP50: 90 values. Moreover, the computational cost of the YOLOv8m model is not the highest among the five models, and its inference speed is also greater than that of YOLOv8l and YOLOv8x. In general, among the training results of YOLOv8m, the mAP50 and mAP50: 90 values are both the highest, and the inference speed and computing cost have certain advantages among the five models. The details are shown in Table 2. However, upon further analysis of the YOLOv8m training results, we discovered that the model is less effective for extreme-sized objects, fuzzy objects, and complex background detections. This paper discusses the deficiencies in detection and proposes solutions.

We propose the WCE_Detection network model for processing multicategory WCE lesion images of the digestive tract, as shown in Figure 2. On the basis of the YOLOv8 backbone network, we redesigned the model and introduced the BiFPN network structure[25] in the neck part. This structure not only optimizes the original network architecture but also helps reduce the loss of feature information, and enhances the feature extraction capability. In terms of head design, we added one new detection head to the original three, ensuring comprehensive coverage of objects at different scales, and accurate detection of objects of various sizes. Moreover, to explore the potential of the self-attention mechanism in prediction, we replaced the original c2f module in the neck with the Swin Transformer module, further improving the prediction performance of the model, and enhancing the network's ability to handle complex scenes and object diversity.

WCE_Detection model

Adding a lesion detection predictor header: After an in-depth study of the WCE dataset, we observed significant shape and scale differences among the lesions. To address this challenge, we drew inspiration from the literature[26], where the performance was significantly improved, particularly in detecting objects at different scales, demonstrating the effectiveness of the module. Therefore, we added a lesion detection prediction head in particular. Combined with the other three prediction heads, they collectively form a four-detection-head structure for the model, as shown in Figure 3. In contrast to the original YOLOv8, which only upsamples twice in the neck structure, the improved network in this paper upsamples three times. The third upsampling layer is fused with the second layer of the backbone network, followed by the downsampling operation. This results in four detection layers: 160 × 160, 80 × 80, 40 × 40, and 20 × 20. The input images are enriched with features from the second layer of the backbone network, thereby deepening the network's depth, and enabling the extraction of feature information into a deeper network. This enhances the multiscale learning capability, improving the model's detection performance in complex scenes.

With this design, we have successfully alleviated the detrimental effects of the drastic variance in the object scale, thereby improving the stability and accuracy of the model. As a result, WCE_Detection now features four detection heads, each dedicated to recognizing micro, small, medium, and large lesion images. While the addition of detection heads incurs a certain level of computation and memory consumption, it is undoubtedly beneficial for enhancing the performance of multiscale lesion object detection. With this enhancement, WCE_Detection can better adapt to lesions of different scales and achieve more precise and comprehensive object detection.

Introduction of the Swin transformer module: WCE lesion images affect image quality due to the lack of light in the digestive tract and the poor angle of shooting the lesions, and it is difficult to extract features effectively if only a CNN is used for feature extraction. Moreover, the same lesions have certain contextual information in the shooting process. As shown in Figure 4, CNNs face difficulty in obtaining contextual information when extracting features. The Transformer [27] employs a self-attention mechanism, which can highlight the features of the detected object and suppress the background features. Initially, the Transformer model achieved significant success in natural language processing. The Vision Transformer (ViT) network[28] applies the Transformer to the field of computer vision. However, a fundamental disparity exists between natural language and images, and the application of the Transformer in the image domain encounters two challenges. First, when applied to natural language, it divides the input into fixed-size tokens. Similarly, computer vision divides an image into fixed-size tokens. However, the features within the image can vary significantly, and the performance of the ViT may not be satisfactory in different scenarios, as shown in Figure 5. Second, when

Table 2 Training results of the YOLOv8 versions using our dataset

Model	Parameter (M)	mAP50	mAP50:95	GFLOPS	FPS
YOLOv8n	2.87	86.0	66.2	8.2	400.00
YOLOv8s	10.62	86.1	67.6	28.7	283.03
YOLOv8m	24.67	88.6	68.3	79.1	189.47
YOLOv8 L	41.62	85.3	67.1	165.5	149.25
YOLOv8x	65.01	83.2	64.1	258.2	114.94

applying the Transformer in natural language processing, the computational complexity is related to the square of the token. Similarly, in computer vision, if the input feature map is 56 × 56, it involves over 3000 matrix operations of length and width, which is a substantial computational load, especially for large images, rendering it unacceptable. This is the reason why ViT is difficult to use widely.

The Swin Transformer[29] model combines the self-attention mechanism of the Transformer and the local perception property of the CNN at the same time, and the shifted window idea is proposed to establish information communication between different windows. This makes the computational complexity linearly related to the size of the input image, which overcomes the limitations of the ViT model in processing large-scale images and the ability to extract features at different scales. This makes the model widely used in all computer vision fields.

In contrast to VIT, which generates a feature map as an indivisible whole, the Swin Transformer adopts a hierarchical Transformer architecture. This architecture reduces the spatial size of the feature map layer-by-layer to extract high-level features, capturing different scales of information at different levels. The Swin Transformer introduces the window self-attention mechanism, which divides the image into multiple windows, and computes the self-attention independently in each window. This significantly reduces the computational cost and preserves the relevance of the local information. However, this segmentation of the image reduces the integration of the global information. Therefore, employing the concept of shifted windows, the model adjusts the window positions in two consecutive Transformer layers. This enhances the model's capability for global information interaction, allowing the Swin Transformer to capture richer global contextual information while maintaining efficient computation, leading to higher performances and lower computational costs.

The Swin Transformer structure comprises four stages for obtaining feature mappings, as shown in Figure 6. Each stage consists of two parts, except for the first stage, which includes linear embedding and a Swin Transformer block, the. The remaining three stages consist of patch merging and a Swin Transformer block. Patch merging, among these components, operates similarly to a pooling operation but without causing information loss. Following each stage of processing, the resolution is halved from the original resolution, while the number of channels doubles compared with that in the previous stage. The Swin Transformer block, which serves as the algorithm's core, comprises window multihead self-attention (W-MSA) and shifted-W-MSA (SW-MSA), as shown in Figure 7. Thus, the number of layers in the Swin Transformer should be an integer multiple of 2, one for W-MSA and one for SW-MSA. This module is succeeded by a 2-layer multilayer perceptron (MLP) with a rectified linear unit (ReLU) interspersed. Both before and after each MSA module and MLP layer, a LayerNorm layer and residual connections are applied. The window-based W-MSA module and the shift-window-based SW-MSA module are sequentially applied to two consecutive Swin Transformer blocks, and the process of computing the feature maps in consecutive Swin Transformer blocks is shown below:

$$\hat{Z}^l = W - MSA(LN(Z^{l-1})) + Z^{l-1} \quad (1)$$

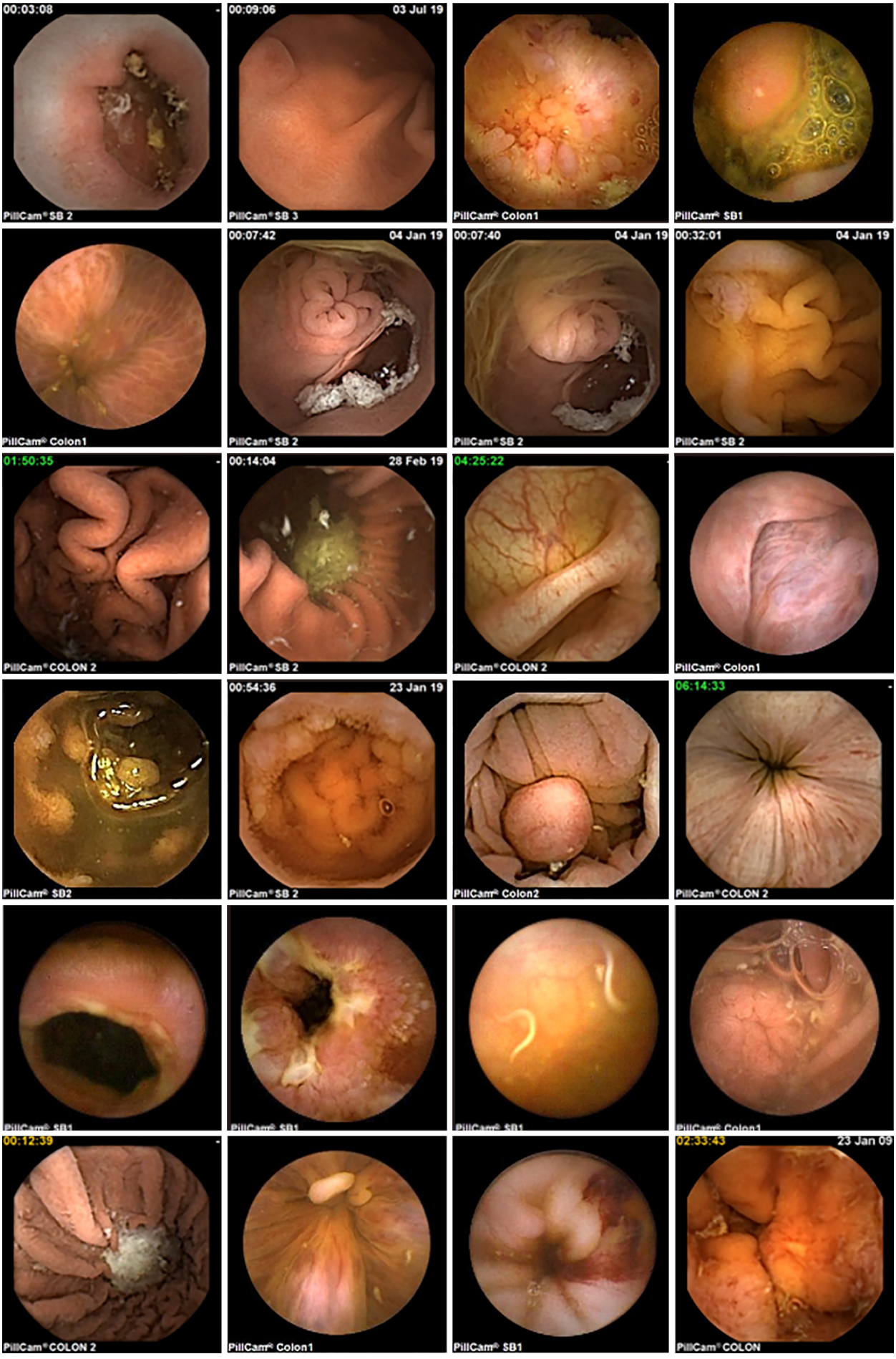
$$Z^l = MLP(LN(\hat{Z}^l)) + \hat{Z}^l \quad (2)$$

$$\hat{Z}^{l+1} = SW - MSA(LN(Z^l)) + Z^l \quad (3)$$

$$Z^{l+1} = MLP(LN(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \quad (4)$$

Where $\hat{z}^l$ represents the output from the encoder of W-MSA. Equation (1) performs multihead calculations and captures the relationships between different parts of the sequence. z^l represents the output after processing by the MLP module. The main function of this layer is to transform the input data into a more discriminative feature representation, thus improving the performance of subsequent tasks.

BiFPN-based feature fusion: Features in WCE lesion images have different shapes and sizes, and the same lesion may have different shapes and sizes during the shooting process. If traditional feature fusion is used, important features may be ignored and lost. The feature pyramid network (FPN)[30] and path aggregation network (PAN)[31] structures are mainly used in YOLOv8 for the fusion of features of different scales. The structures are shown in Figure 8A and B. FPN feature fusion usually combines the semantic information of high-resolution feature maps with the detailed information of low-resolution feature maps through top-down paths and lateral connections. However, the FPN suffers from two problems in feature fusion. First, the FPN fuses features only through upsampling and lateral connectivity, which does not fully utilize the complementary nature of the features at each level, and may lead to important information being


00:03:08
PillCam® SB 2
00:09:06
03 Jul 19
PillCam® SB 3
PillCam® Colon1
PillCam® SB1
PillCam® Colon1
00:07:42
04 Jan 19
PillCam® SB 2
00:07:40
04 Jan 19
PillCam® SB 2
00:32:01
04 Jan 19
PillCam® SB 2
01:50:35
PillCam® COLON 2
00:14:04
28 Feb 19
PillCam® SB 2
04:25:22
PillCam® COLON 2
PillCam® Colon1
PillCam® SB2
00:54:36
23 Jan 19
PillCam® SB 2
PillCam® Colon2
06:14:33
PillCam® COLON 2
PillCam® SB1
PillCam® SB1
PillCam® SB1
PillCam® Colon1
00:12:39
PillCam® COLON 2
PillCam® Colon1
PillCam® SB1
02:33:43
23 Jan 09
PillCam® COLON

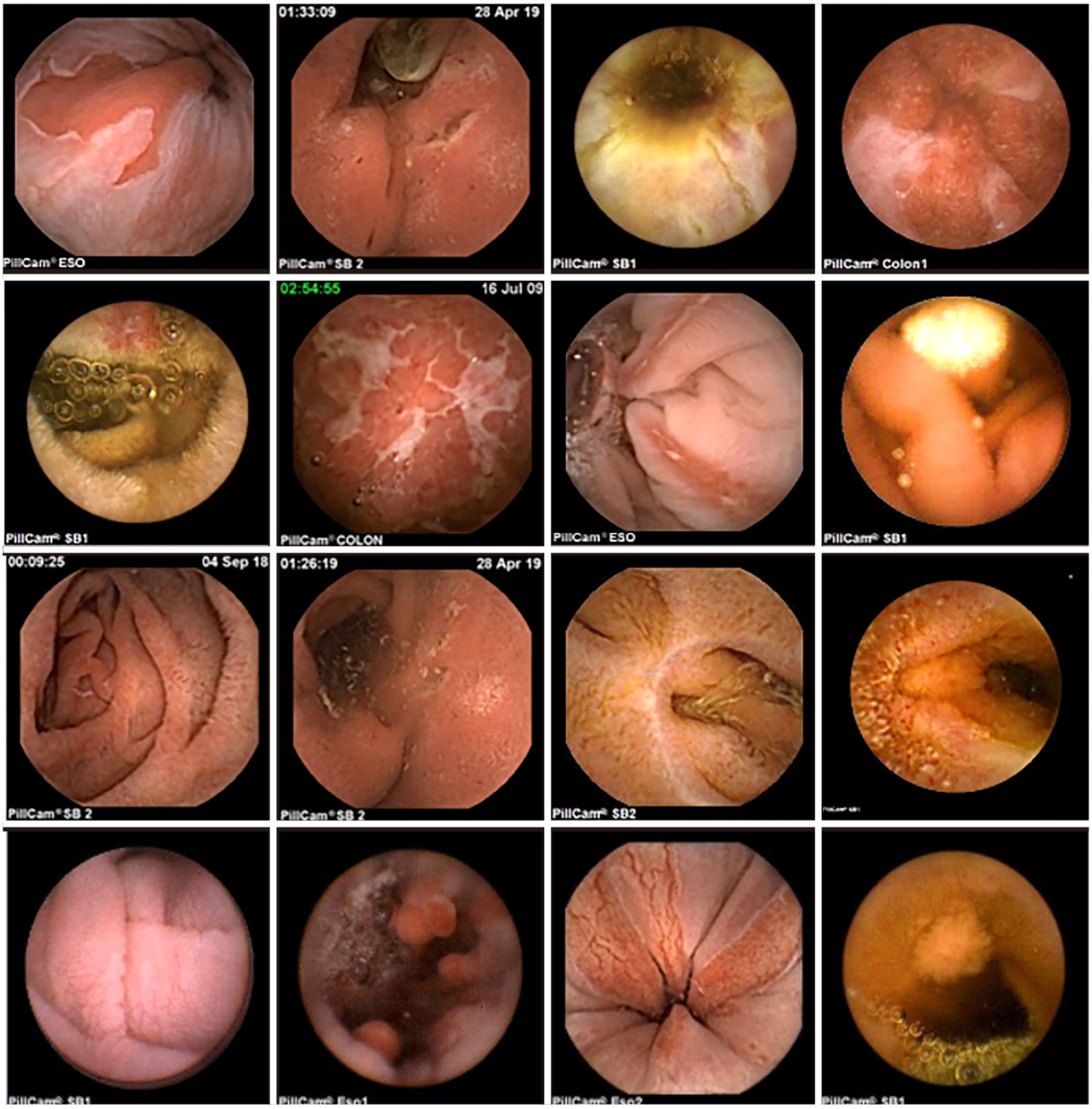


Figure 1 Dataset example.

overlooked. Second, in the top-down path, the spatial information transfer path of the feature map is lengthy, which may lead to the loss of information from the poor fusion of the higher level with the lower level. The PAN is an improvement of the FPN, which optimizes the feature transfer path by introducing lateral connections from the lower to the upper level. It improves the network accuracy, but a larger network size leads to an increase in the number of parameters, making the computational efficiency relatively low. With increasing network layers, the feature information gradually transitions from low-dimensional to high-dimensional, and each layer of the network may produce a certain degree of feature information loss in this process.

BiFPN adopts a bidirectional feature fusion approach, which makes better use of the contextual information of different layers of the features, as shown in Figure 8C. The BiFPN module consists of two phases: Top-down feature fusion and bottom-up feature fusion. First, in the top-down feature fusion stage, the high-level feature maps are refined by upsampling, and are fused with the low-level feature maps so that the high-level feature maps can obtain richer contextual information. Second, in the bottom-up feature fusion stage, the low-level feature map is coarsened by downsampling and fused with the high-level feature map. In this way, the low-level feature maps can obtain more detailed information. Each layer in the BiFPN structure is connected to the upper and lower neighboring layers, and this bidirectional connection design can effectively fuse and interact the feature information of different layers, improve the feature expression ability, and reduce the loss of the feature information. The BiFPN structure performs feature fusion at different scales through the upper and lower branches. Higher-level semantic information can be introduced while maintaining high-resolution features to improve the detection model's ability to perceive objects at different scales. The

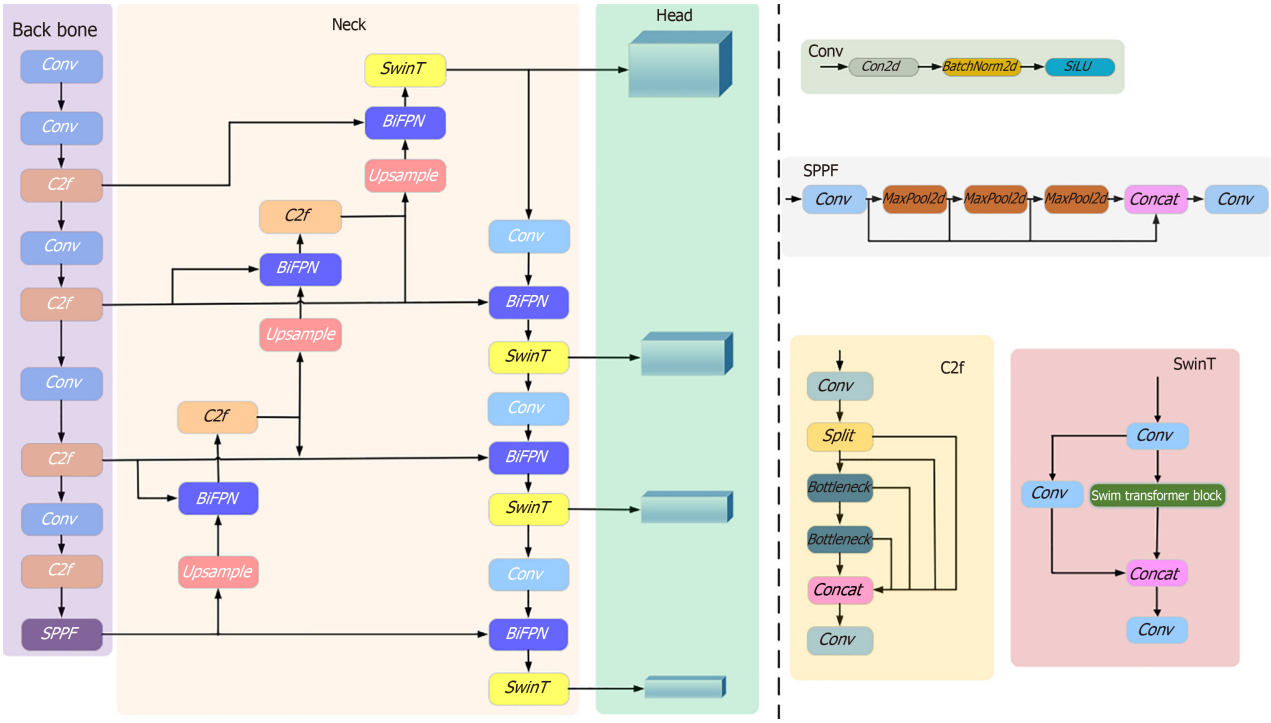


Figure 2 Wireless capsule endoscopy_Detection model structure. Conv: Convolution; SPPF: Spatial pyramid pooling fast; SwinT: Swin transformer; SiLU: Sigmoid linear unit.

feature fusion process of the BiFPN structure adopts a dynamic weight allocation method, which adaptively adjusts the weight of each feature channel according to the quality and importance of the features, ensuring the effective use of the important features. As shown in Equation (5):

$$O = \sum_i \frac{\omega_j}{\varepsilon + \sum_j \omega_j} I_i \quad (5)$$

Where ω represents the weights learned by the features, I_i represents the input feature mapping, the stable value coefficient ε = 0.0001, the weight range is reduced to 0-1, and the training speed is fast and efficient. Moreover, the BiFPN structure can be stacked as many times as needed, and each repetition improves the depth scalability of the network through further feature fusion and optimization, which can be adapted to object detection tasks with different complexity and accuracy requirements.

The BiFPN uses separable convolutional fusion features and adds batch normalization and activation after each convolution. Let us take layer 5 as an example, as shown in Figure 8C. P_5td is the intermediate feature of layer 5, P_5out is the output feature of layer 3, and P_5out is obtained *via* weighted fusion of the inputs P_5td and P_5in from layer 5 and P_4out from layer 4. The expressions for P_5td and P_5out are shown in Equation (6) and (7), respectively.

$$P_5td = Conv\left(\frac{\omega_1 P_5in + \omega_2 Resize(P_6in)}{\omega_1 + \omega_2 + \varepsilon}\right) \quad (6)$$

$$P_5out = Conv\left(\frac{\omega'_1 P_5in + \omega'_2 P_5td + \omega'_3 Resize(P_4out)}{\omega'_1 + \omega'_2 + \omega'_3 + \varepsilon}\right) \quad (7)$$

Where Conv is a depth-separable convolution operation, Resize is an upsampling or down sampling operation, ω_1, ω_2, ω'_1, ω'_2 and ω'_3 are weight parameters, and ε is a parameter used to avoid instability of values due to too small weight parameters.

RESULTS

Model performance evaluation metrics

To comprehensively evaluate the performance of the WCE_Detection model, we utilize the mean average precision (mAP), precision, and recall as the primary indicators for performance evaluation, paying special attention to the mAP50 and mAP50: 95 metrics. These two metrics are positively correlated with model performance, with higher values indicating a better performance. Moreover, we employ the precision-recall (P-R) curve and confusion matrix to provide a more intuitive evaluation of the model's performance. True positive (TP) indicates samples correctly identified as positive, false positive (FP) represents negative samples incorrectly classified as positive, and false negative (FN) indicates

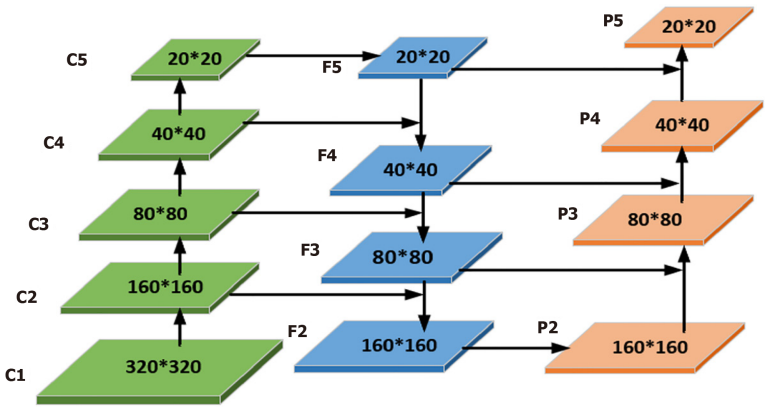


Figure 3 Wireless capsule endoscopy_Detection network structure diagram. C1, C2, C3, C4, C5: Layers 1, 2, 3, 4, and 5 of the backbone network; F2, F3, F4, F5: Layers 2, 3, 4, and 5 of the neck network; P2, P3, P4, P5: 2nd, 3rd, 4th, 5th detection head.

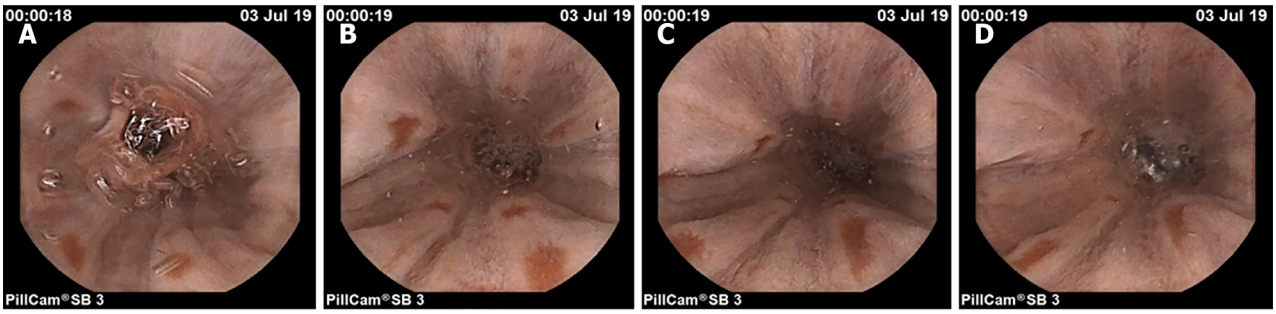


Figure 4 Context information of the reflux esophagitis lesion. A: Captured at 00:00:18, representing the earliest result; B: Captured at 00:00:19, representing an earlier frame at this time point; C: Captured at 00:00:19, representing a continued frame subsequent frame, showing further details of the lesion; D: Captured at 00:00:19, representing a later frame at this time point, where the lesion area might reveal new angles or further details due to the capsule's movement.

positive samples incorrectly classified as negative.

Precision indicates the proportion of the number of samples predicted as positive to the number of all samples predicted as positive. The higher the precision is, the better the performance of the classifier. The specific expression is shown below:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

Recall is the ratio of the number of instances correctly identified as positive categories to the number of instances of all the true categories, as shown in the expression below:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

The average precision represents the average detection precision of the single-category model, which is a coordinate system established with the recall rate as the horizontal coordinate, and the precision rate as the vertical coordinate, where P denotes precision and r denotes recall. r is the horizontal axis, P is the vertical axis, and the area enclosed under the P-R curve is formed on the basis of a certain threshold value. The expression is shown below:

$$AP = \int_0^1 P(r)\,dr \quad (10)$$

The average accuracy mean value represents the average accuracy of all categories N. A larger value indicates a higher accuracy of the object detection model. The specific expression is shown below:

$$mAP = \frac{\sum_{k=1}^{N} AP(k)}{N(class)} \quad (11)$$

The AP value is the area enclosed under the P-R curve, reflecting the model's accuracy performance under different recall rates. ap is the sum of the AP values for all the categories, whereas N (class) denotes the total number of categories. mAP50 is obtained by calculating and averaging the AP values for each category with the intersection and intersection over union (IoU) of the predicted bounding box to the ground truth bounding box set to 0.5. In contrast, mAP50:95 was obtained by calculating and averaging the AP values for each category separately when the IoU was increased from 0.5 to

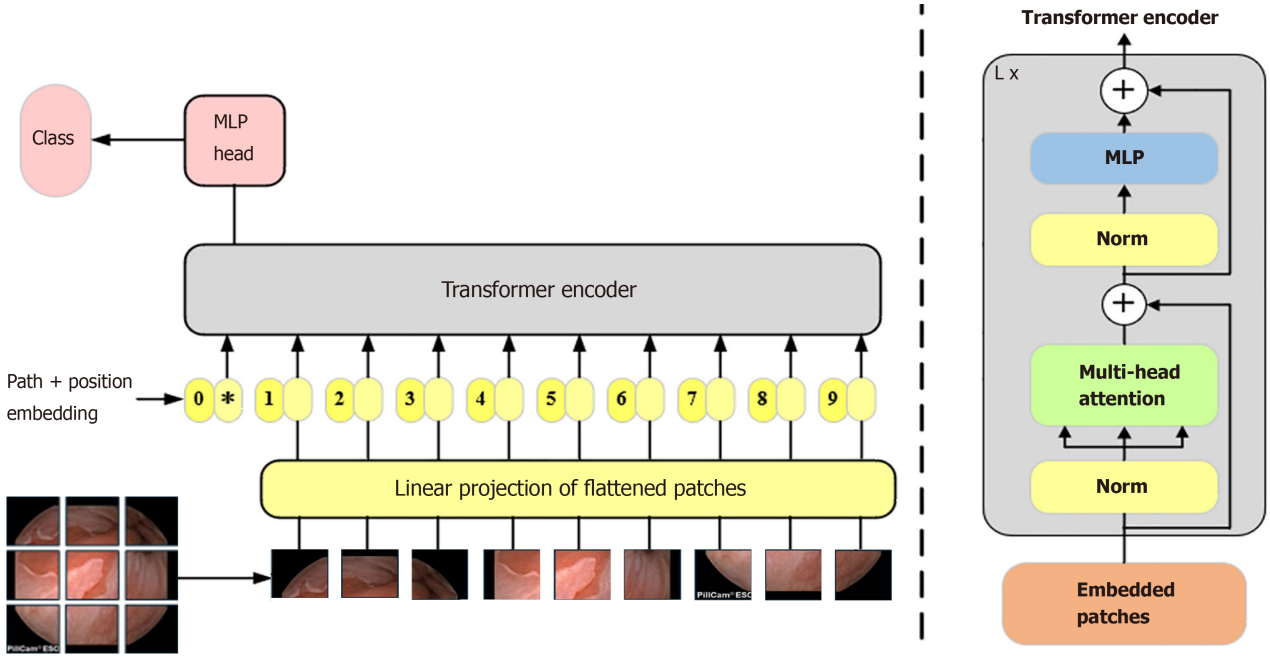


Figure 5 Vision transformer. MLP: Multilayer perceptron; L: Layer.

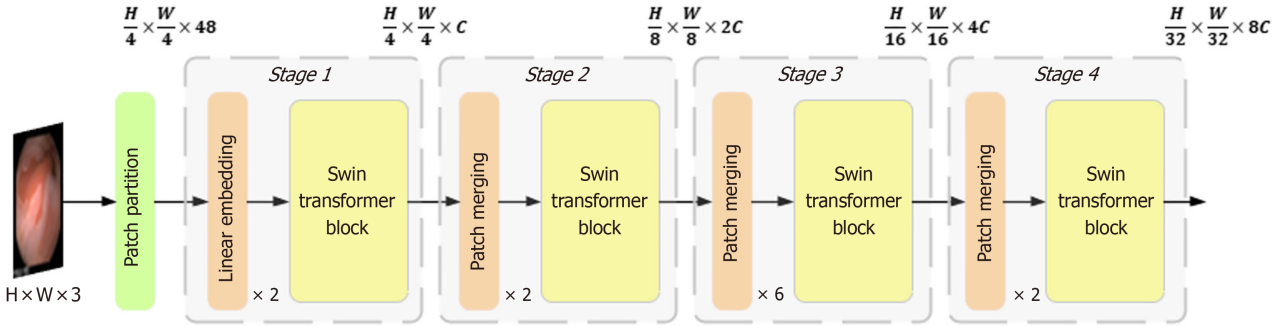


Figure 6 Swin transformer model structure. H: Height; W: Width; C: Channels.

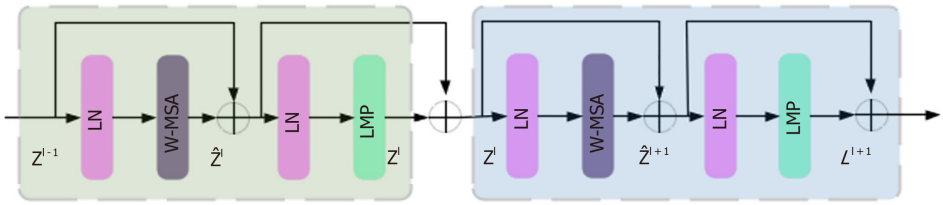


Figure 7 Two consecutive Swin transformer blocks. LN: Layer normalization; W-MSA: Window multihead self-attention; MLP: Multilayer perceptron; SW-MSA: Shifted window multihead self-attention.

0.95 by 0.05 each time. Notably, both precision and recall are considered at an IoU threshold of 0.5 for both the mAP50 and mAP50:95 calculations. These metrics provide us with a comprehensive and detailed view of model performance evaluation.

Experimental results and analysis of the WCE_Detection model

We compare the performance of the proposed method with that of state-of-the-art (SOTA) image detection methods on WCE lesion images. This includes one-stage, two-stage algorithms as well as anchor-based, anchor-free algorithms, and to guarantee the fairness of the experiments, and the reliability of the results, the training process was conducted *via* a Tesla V100-SXM2-16GB GPU on AliCloud, employing the standard AdamW optimizer. The initial learning rate was initialized at 0.01 and dynamically adjusted throughout the training process. Additionally, the batch size was set to 16, and training was performed for 300 epochs. For data preprocessing, we uniformly adjusted the 640 × 640 pixel size for all the input

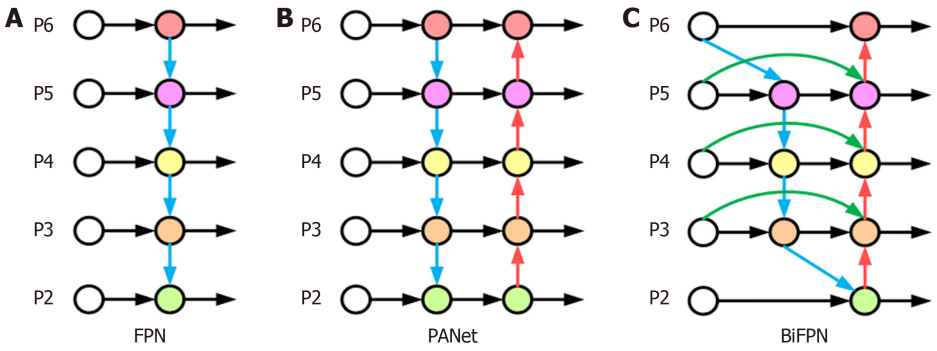


Figure 8 Different feature fusion structures. A: Feature pyramid network (FPN); B: Path aggregation network; C: Bidirectional FPN. P2: Represents layer 2 feature maps; P3: Represents layer 3 feature maps; P4: Represents layer 4 feature maps; P5: Represents layer 5 feature maps; P6: Represents layer 6 feature maps. FPN: Feature pyramid network; PANet: Path aggregation network; BiFPN: Bidirectional feature pyramid network.

WCE lesion images. Data enhancement techniques such as mosaic, mixup, and HSV were introduced, and their performance in complex scenes was improved. The results of the comparison experiments are shown in Table 3[32,33].

Table 3 shows the comparative performance of the proposed WCE_Detection model against several mainstream image detection methods on WCE lesion images. The results indicate that WCE_Detection significantly outperforms the other models in terms of both the mAP50 and mAP50:90 metrics, especially in the multicategory lesion detection task. Specifically, one-stage algorithms such as the YOLOv5, YOLOv6, and SSD algorithms perform differently in terms of FPS, with SSD being slightly worse. The performance is close in terms of the mAP metric, with an mAP50 of approximately 80%. The detection effect for most lesions is good, but the model's ability to capture global information is limited, which makes the detection effect for some lesions, such as duodenal bulbar ulcers and gastritis, poor. YOLOv7 performs poorly for some esophageal protruding lesion tumor masses and colocystosis hemorrhoids, which may be related to its inability to capture small-scale lesions effectively in the feature extraction stage. In contrast, YOLOv9 and RT-DETR perform well, with mAP50 values exceeding 80%, and the model has better detection results for foci such as colocystosis hemorrhoids, gastritis, *etc.*, indicating that the model is able to better cope with complex digestive tract images. However, the FPS performance does not surpass that of the other algorithms. For the two-stage algorithm, Faster R-CNN fails to perform as expected on the WCE dataset, despite its theoretically finer candidate region generation and classification capabilities. This is likely because they generate excessive candidate regions in complex backgrounds, leading to an increased false detection rate, which significantly affects lesions such as gastritis and colonic mucosal melanosis. Moreover, the long computational process limits its efficiency in practical applications. Our proposed WCE_Detection model performs excellently, with a certain advantage in FPS over other algorithms, while the mAP50 reaches 91.5%, and there is also a significant improvement in the mAP50:90 metric. This performance improvement is due mainly to the innovative aspects of our model design. First, the multidetection head strategy substantially enhances the model’s capacity to detect multiscale lesions, ensuring effective recognition regardless of lesion size. Second, BiFPN provides richer fusion of semantic information during feature extraction, especially when dealing with complex backgrounds and multicategory lesions, and BiFPN can significantly reduce the false detection rate. In addition, the introduction of the Swin Transformer module further improves the model's ability to capture global information, and enhances the feature representation of lesions, resulting in a more stable performance in complex WCE image environments. Although the WCE_Detection model performs well on most categories, there is still room for improvement in the model's detection performance for some rare or morphologically complex lesion types. This is largely attributed to the small number of these lesions in the training dataset, and the model's limitations in learning features for these categories. Future work will aim to expand the diversity of the dataset, and will explore more effective feature enhancement and data generation techniques to further improve the model's detection performance on all categories.

We first plotted the P-R curve for each category, as shown in Figure 9. The P-R plot illustrates the relationship between the precision and recall of the model. Precision measures the accuracy of the positive predictions, and recall describes the ability of the model to capture all the positive samples. Ideally, curves near the top right corner indicate higher precision and recall, which means better model performance. As shown in Figure 9, with the threshold of the IoU set to 0.5, sixteen lesions have AP values above 90%, and more than twenty lesions have AP values above 80%, of which only three lesions, stomach, luminal stenosis, and esophageal protruding lesion tumor masses, have AP values below 80%. This is because the training data for these three categories are significantly lower than those of the other categories, and the data are not rich enough. However, it can also be seen from the experimental results, that more than eleven single-category lesions were detected with an accuracy of more than 99.4%, which has good clinical application value.

We simultaneously plotted a confusion matrix for each category, as shown in Figure 10. We need to consider the multiple categories of lesions used in this study, which may make the information presented in the images unclear. Therefore, we should analyze it from different perspectives. The confusion matrix is used to show the classification effect of the model on different categories. The rows and columns of the confusion matrix represent the true and predicted categories, respectively, and the values in the diagonal region represent the proportion of correctly predicted categories, whereas the values in the other regions represent the proportion of incorrectly predicted categories. For lesions with relatively low accuracy in the P-R plot, the same phenomenon of high detection error is present in the confusion matrix; the main reason for this is that, at the data level, such lesions have a small number of training samples and a single

Table 3 Training results of different detection models on the dataset

Model	mAP50	mAP50:90	FPS
YOLOv5	78.3	58.9	176.25
YOLOv6	79.6	58.1	316.63
YOLOv7	75.8	56.4	206.37
YOLOv9[32]	82.0	64.7	94.30
SSD	78.6	57.9	84.62
Faster R-CNN	79.0	60.9	61.98
RT-DETR[33]	81.9	63.9	143.46
WCE_detection	91.5	68.6	129.70

WCE: Wireless capsule endoscopy.

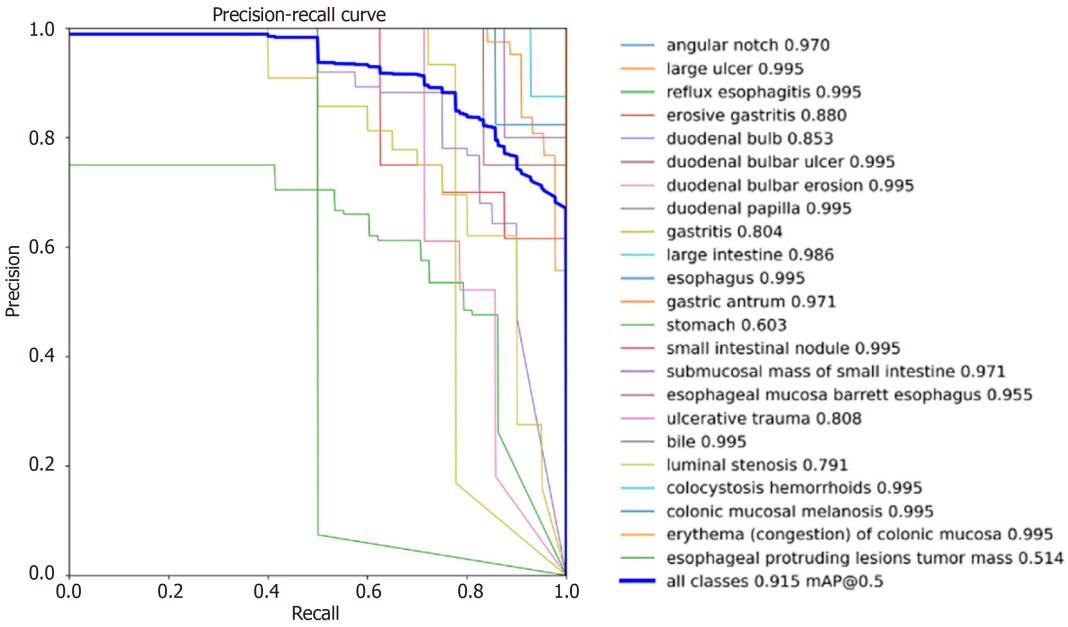


Figure 9 Precision-recall curves of the wireless capsule endoscopy_detection model for the dataset.

source. Second, such lesions are interfered with by food or air bubbles, which increases the difficulty of lesion detection. However, in the confusion matrix, a relatively clear diagonal line can still be seen in the middle, which indicates that the proposed model makes relatively accurate predictions for all categories, demonstrating the effectiveness of the model. Figure 11 shows the detection results for images of different lesions in the digestive tract.

As seen from the images of digestive tract lesion detection presented in the figure, the internal environment of the digestive tract is complex, not only containing multiple types of lesions but also being interfered with by insufficient lighting conditions and other substances such as air bubbles in the digestive tract. These interfering factors, together with the variable morphology and color of the lesions themselves, as well as the multiscale features they present, undoubtedly pose a great challenge to the detection of digestive tract lesions. However, the WCE_Detection model shows an excellent coping ability in the face of these difficulties, successfully overcoming them and achieving excellent detection results. This performance fully demonstrates the accuracy and reliability of WCE_Detection in detecting digestive tract lesions in complex environments. However, the dataset for certain individual lesions is significantly smaller than those for other categories, resulting in limited accuracy during detection. As shown in Figure 12, images A and B illustrate instances of incorrect lesion detection. In image A, the esophageal protruding lesion is misclassified as a gastritis lesion, whereas in image B, the luminal stenosis lesion is misclassified as a gastric antrum lesion. Images C and D depict cases of reduced detection accuracy for individual lesions. This is primarily attributed to the limited number of samples for certain rare lesion types, which hinders the model's ability to fully learn these features. To address this issue, we propose generating additional scarce lesion images to expand the dataset by leveraging the generator and discriminator structure within generative adversarial networks (GANs)][34] to alleviate the data imbalance problem and improve the detection performance.

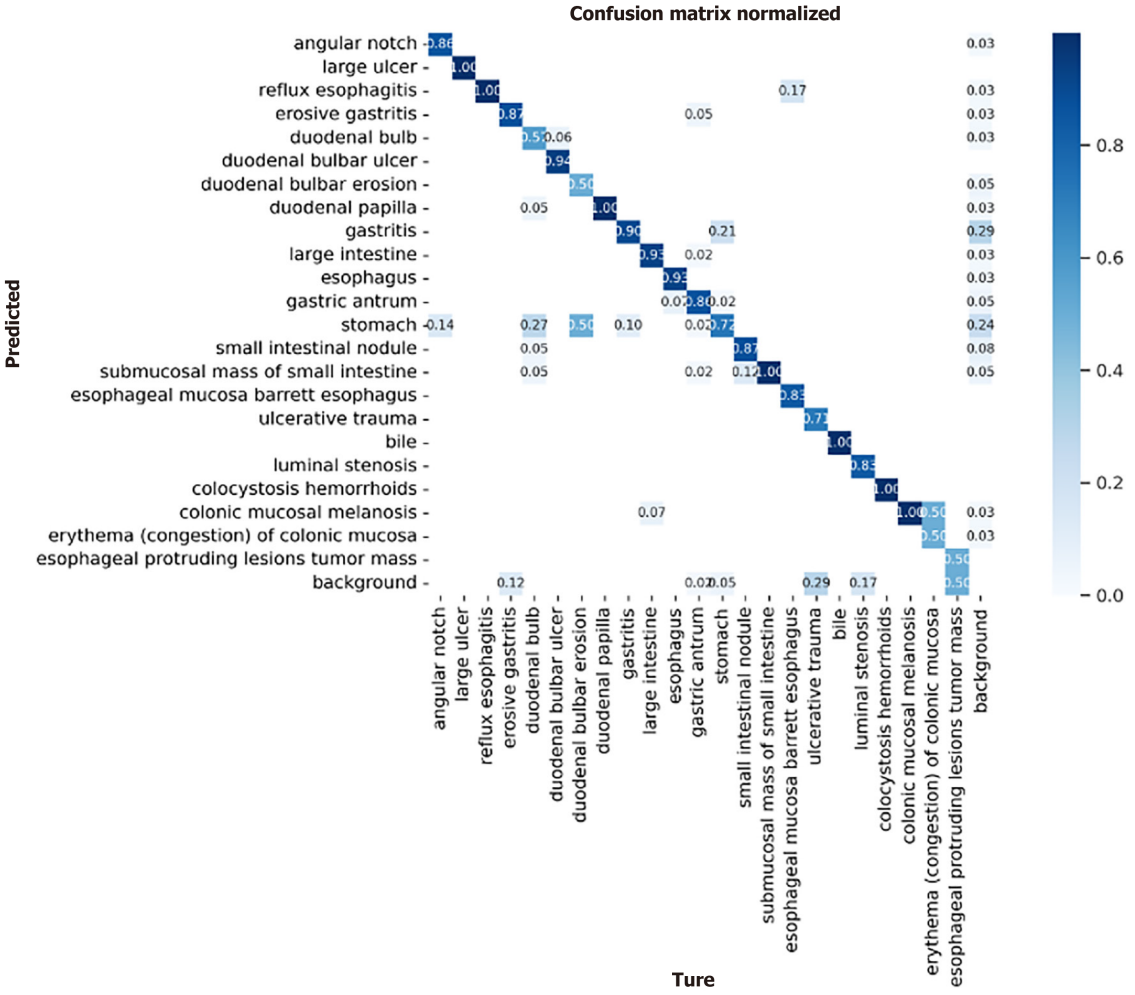


Figure 10 Confusion matrix of the wireless capsule endoscopy_detection model.

DISCUSSION

In recent years, WCE image lesion detection *via* traditional machine learning methods and deep learning-based methods has greatly improved the efficiency and convenience of physicians. Recent research has provided a comprehensive survey and systematic summary of the trends in WCE image lesion detection techniques[35-37], with a special focus on the accurate identification of pathological abnormalities, such as polyps, hemorrhages, tumors, ulcers, and other lesions.

Conventional machine learning-based lesion detection is divided into two main steps. First, feature extraction, where various features, such as color, texture, and shape, are extracted from the WCE image by manually designing or selecting a series of feature descriptions to represent the image as vectors or feature sets that can be processed by machine learning algorithms. These feature collections are then used to train machine learning models, such as support vector machines (SVMs)[38,39], decision trees[40], or K-means clustering algorithms[41], during the model training phase for the classification of lesions. For example, the methods[38,41] both focus on the polyp detection task for WCE images, but they are unique in their feature extraction and methods. Specifically, Hwang and Celebi[41] used Gabor texture features for image analysis, whereas Yuan *et al*[38] used the scale invariant feature transform algorithm to extract texture features from the neighborhood points of the key points. Moreover, there are differences in classification methods. Hwang and Celebi[41] used the K-means clustering algorithm for image classification, whereas Yuan *et al*[38] chose SVM as the classification method. As another example, Pogorelov *et al*[39] proposed a fast bleed detection method for WCE videos, which takes the color and texture of the bleed region as features, and then uses an SVM for classification. Yeh *et al*[40] extracted color vector and grayscale covariance matrix features from WCE images, which was followed by a series of statistics such as the mean and variance of the obtained matrix to form features for classification. The performances of decision trees and support vector machines in terms of ulcers, bleeding, and normal images were further analyzed for classification.

Deep learning-based lesion detection in WCE images is based mainly on CNNs for lesion detection, multiclassification of lesions, and so on. For example, Aoki *et al*[42] proposed a deep learning-based method for the detection of erosions and ulcerations in WCE images, which used SSD convolutional neural network learning to train cropped WCE images. Li *et al*[43] used the CNN models LeNet, AlexNet, GoogLeNet, and VGGNet convolutional neural networks to detect intestinal bleeding, and used the strategy of image augmentation during the training process, which provides good robustness in the face of complex and variable intestinal images. Shin *et al*[44] proposed the detection of colon polyps in WCE images on the basis of an improved Faster R-CNN model, which Shin *et al*[44] adopts the network structure of Inception ResNet

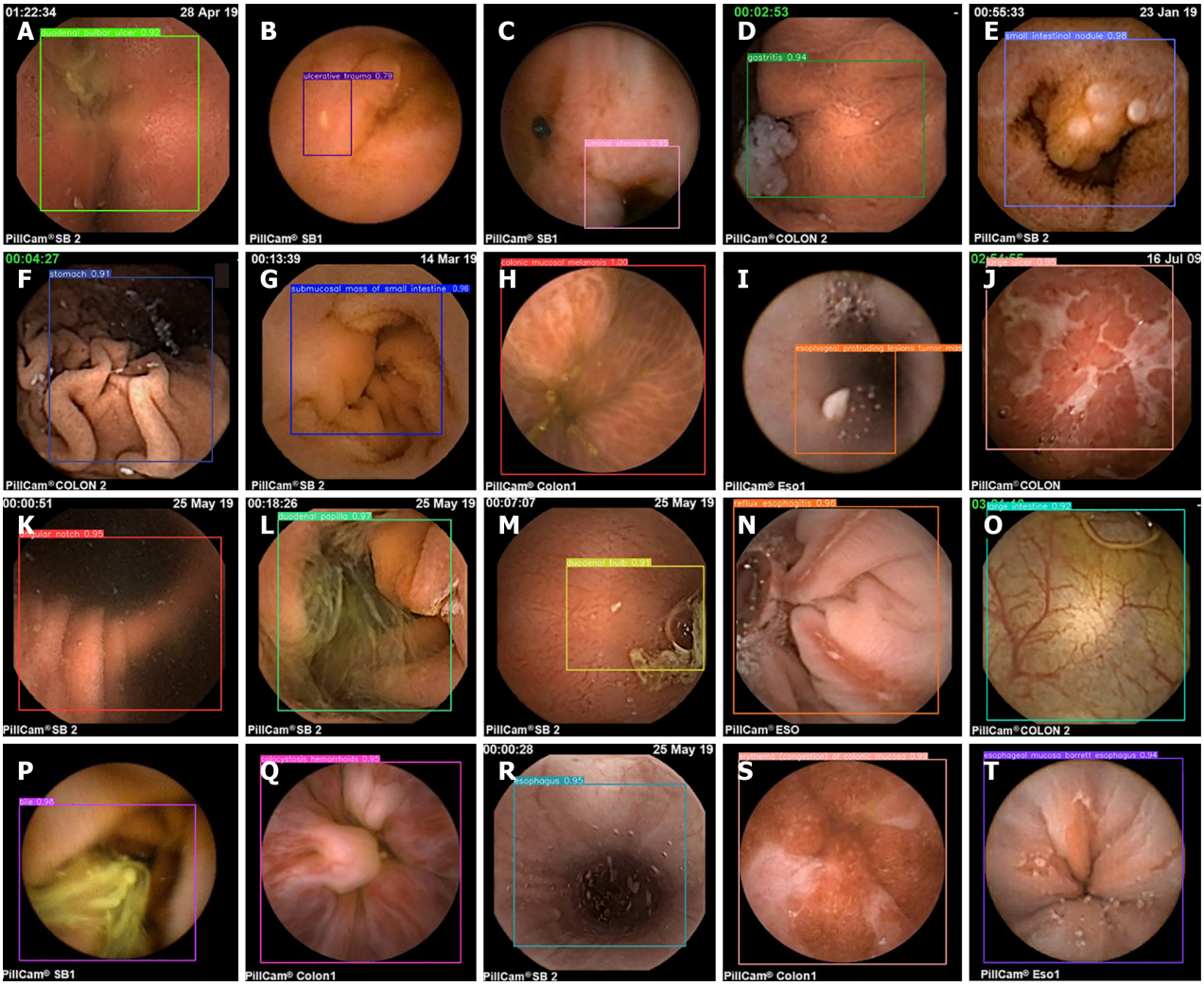


Figure 11 Wireless capsule endoscopy_detection model detection visualization results. Different letters in the image represent different types of lesions. A: Duodenal bulbar ulcer; B: Ulcerative trauma; C: Luminal stenosis; D: Gastritis; E: Small intestinal nodule; F: Stomach; G: Submucosal mass of the small intestine; H: Colonic mucosal melanosis; I: Esophageal protruding lesion tumor mass; J: Large ulcer; K: Angular notch; L: Duodenal papilla; M: Duodenal bulb; N: Reflux esophagitis; O: Large intestine; P: Bile; Q: Colocystosis hemorrhoids; R: Esophagus; S: Gastric antrum; T: Duodenal bulbar erosion.

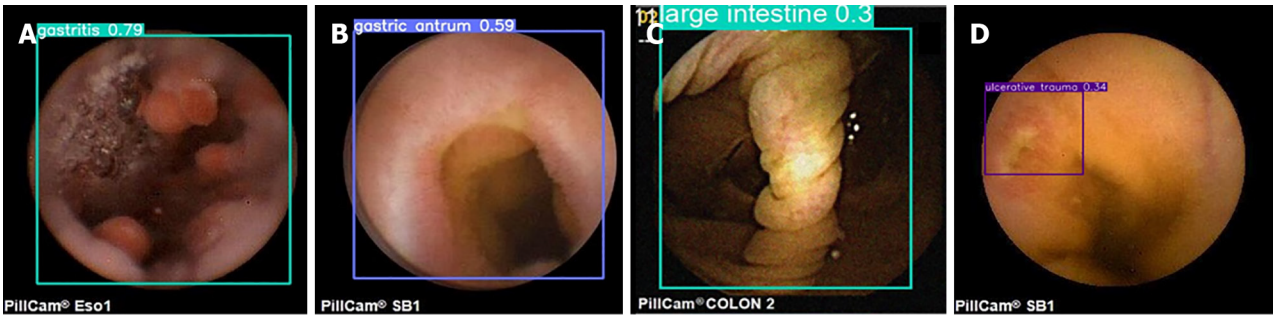


Figure 12 Example of a wrongly detected lesion. A and B: Examples of misdetected lesions; C and D: Lesions with low detection accuracy.

in the backbone network to optimize the performance of the model. Moreover, to further improve the detection accuracy, Shin *et al*[44] also adopted a new strategy to replace the ROI pooling layer in the original model to improve the accuracy of detecting polyps in the reflective region. Thuan *et al*[45] introduces an encoder-decoder model, RaBiT, which integrates a lightweight Transformer-based architecture in the encoder to model multilevel global semantic relationships. The decoder consists of multiple BiFPNs and reverse attention modules to enhance the fusion of multilevel feature mappings, gradually refining polyp boundaries to improve the model's generalization ability. Park and Lee[46], which is based on the U-Net framework, proposes a novel deep learning model, SwinE-Net, for polyp segmentation, which effectively combines CNN-based EfficientNet with the Swin Transformer. The Swin Transformer primarily facilitates accurate and robust medical segmentation while preserving global semantics without sacrificing the low-level features provided by the

CNNs. These studies demonstrate that existing deep learning models have made significant advancements in addressing the detection of specific lesions in WCE images.

Although a great deal of current research has focused on the detection of pathological abnormalities in WCE images and videos, most studies have focused on the detection of a single or a few types of lesions. However, in real medical practice, many different types of lesions may exist in the digestive tract at the same time, which requires the detection algorithms to be able to identify and process multiple types of pathological abnormalities at the same time. Therefore, to achieve the detection of multicategory lesion images in the digestive tract, we design an end-to-end convolutional neural network, the WCE_Detection model. First, a multidetection head strategy is adopted in the target detection network to enhance the model's ability to capture lesions of various sizes and morphologies by detecting them at different feature scales, which significantly reduces the missed detection rate of small-scale or irregular lesions. Second, BiFPN is used to fuse deep and shallow features through jump connections, so that shallow high-resolution features are combined with deep high-semantic information, which enhances the model's ability to detect fine-grained lesions, and thus reduces the detection error rate in complex scenes. In addition, we incorporate the Swin Transformer to further enhance the model's capability for multiscale feature representation through its self-attention mechanism and hierarchical structure. The combination of the Swin Transformer with BiFPN enables the model to better handle long-distance dependencies and global features, thus improving the feature representation and detection of different types of lesions. These innovations significantly improve the detection efficiency and accuracy of our proposed method, which is used to detect 23 types of lesions in the digestive tract, and provide doctors with more comprehensive and accurate diagnostic information.

To further verify the effectiveness of the improved algorithm, this paper redesigned the YOLOv8 backbone network as the basis, introduced improved modules one-by-one, and conducted multiple experiments. The experimental results are shown in Table 4. In the table, "Y" means that the module is designed on the original YOLOv8 algorithm for the experiment.

Table 4 presents the experimental results obtained with varying numbers of detection heads, highlighting the impact of increasing the number of detection heads on the model's ability to identify targets across different scales. While using two detection heads allows for the detection of targets at multiple scales, the performance remains inadequate for handling more complex, multiscale scenarios. In YOLOv8, there are three detection heads. Compared with two detectors, three detectors can provide better detection results in complex scenes. By further increasing the number of detection heads to four, more pronounced scale variations can be effectively managed, leading to improved detection of small, large, and medium targets. However, expanding the number of detection heads to five yields only marginal improvements in mAP50, while significantly increasing both the model complexity and inference time.

Table 5 shows the results of different neck experiments. The neck structure is the key component connecting the backbone network and the detection head, and is responsible for fusing different scale features to improve the detection of multiscale targets. The structure of the FPN is relatively simple, with lower computational overhead and faster inference speed, but the FPN adopts a unidirectional top-down feature fusion strategy, which may lead to information loss or failure to make full use of the underlying detailed features. In YOLOv8, the PANet structure is used, which makes the multiscale features more fully fused, and improves the detection of targets at different scales. However, redundant information may also be introduced, and the network may produce a certain degree of feature information loss. The BiFPN structure provides a better feature weighting mechanism on the basis of the effect of feature fusion and computational efficiency, which improves the detection ability for multiscale targets, especially small target detection, and is suitable for the pursuit of high-precision and complex scenarios.

Table 6 shows that the CBAM module enhances feature extraction through channel attention and spatial attention mechanisms. However, its attention mechanism relies primarily on convolutional operations, which limits its ability to model global features. The SE module applies channel weighting through global average pooling, with minimal computational overhead that does not significantly increase network complexity. However, it only enhances features in the channel dimension, limiting its ability to model spatial features. This restriction can lead to the neglect of complex spatial relationships, resulting in suboptimal detection performances. ViT effectively handles long-range dependencies and complex global contextual information, offering significant advantages in scene understanding and object relevance modeling. However, at high resolutions, ViT demands substantial computational resources and results in longer inference times, making it less suitable for real-time applications. The Swin Transformer combines the global feature modeling advantages of ViT with the efficiency of convolutional networks, achieving multiscale feature extraction through a hierarchical design that enhances computational efficiency (Table 7).

Adding extra detection heads: We add additional detection heads to perform the detection task more effectively with multiscale detection heads, and reduce the semantic gap by fusing different levels of feature information. Although the number of layers of WCE_Detection changes from 295 to 361, and the FPS changes from 189.47 to 179.19 after we add an extra detection head, the detection speed decreases, but the mAP50 increases from 88.6 to 89.2, which also shows the effectiveness of the extra detection head.

Replacement of the neck *via* the BiFPN structure: Although the introduction of additional detection heads helps to extract more layers of feature information, effectively fusing these features is still a challenge. An inappropriate fusion strategy may lead to redundancy or loss of information, thus affecting the final detection results. We choose to use the BiFPN feature pyramid network, which enables better fusion and utilization of features at different scales. When we used the BiFPN structure, the total number of layers of the model did not change, the FPS changed from the original 179.17 to 183.35, the detection speed increased, and the mAP50 increased from 89.2 to 90.6, so it was effective when the neck of the model was replaced with the BiFPN structure.

Table 4 Experimental results of different numbers of detection heads

Method	P2	P3	P4	P5	mAP50	GFLOPS	FPS
Method 1	Y	N	N	N	87.1	47.9	233.65
Method 2	N	Y	N	N	88.6	79.1	189.47
Method 3	N	N	Y	N	89.2	126.7	179.19
Method 4	N	N	N	Y	89.5	166.3	153.95

Y: Means that the module is designed on the original YOLOv8 algorithm for the experiment; N: Means that the module was not used in the model design.

Table 5 Experimental results of different necks

Method	FPN	PANet	BiFPN	mAP50	GFLOPS	FPS
Method 1	Y	N	N	86.3	58.6	226.63
Method 2	N	Y	N	88.6	79.1	189.47
Method 3	N	N	Y	90.1	98.4	178.36

Y: Means that the module is designed on the original YOLOv8 algorithm for the experiment; N: Means that the module was not used in the model design.

Table 6 Experimental results of different attention modules

Method	CBAM	SE	VIT	Swin Transformer	mAP50	GFLOPS	FPS
Method 1	Y	N	N	N	88.9	133.6	163.96
Method 2	N	Y	N	N	87.2	121.5	173.73
Method 3	N	N	Y	N	89.2	219.9	103.76
Method 4	N	N	N	Y	90.7	159.7	149.58

Y: Means that the module is designed on the original YOLOv8 algorithm for the experiment; N: Means that the module was not used in the model design.

Table 7 Ablation experiment results

Method	p4	BiFPN	Swin Transformer	mAP50	GFLOPS	FPS
Method 1	N	N	N	88.6	79.1	189.47
Method 2	Y	N	N	89.2	126.7	179.19
Method 3	Y	Y	N	90.6	138.4	183.35
Method 4	Y	Y	Y	91.5	203.6	129.70

Y: Means that the module is designed on the original YOLOv8 algorithm for the experiment; N: Means that the module was not used in the model design.

Introduction of the Swin Transformer module: Based on the successful introduction of the BiFPN structure to enhance the performance of multiscale feature fusion and object detection, we further explored the optimization of the model structure. To this end, we decided to add the Swin Transformer after BiFPN to take advantage of its powerful self-attention mechanism and efficient computational properties, which enhances the network's ability to handle complex scenes and object diversity, and further improves the accuracy and robustness of the model. Our idea is well proven; when we add the Swin Transformer, the mAP50 increases from 90.6 to 91.5; of course, adding the Swin Transformer will increase the computations, the number of layers of the model increases from 361 to 441, and the FPS changes from 183.35 to 129.70, although the detection speed becomes slower, but it is far greater than the effect of real-time detection, and the change in mAP50 is also very obvious, which fully proves the effectiveness of the module.

In summary, through ablation experiments, we have successfully verified the effectiveness of each module of the WCE_Detection model proposed in this paper, which not only improves the mAP50 by 2.9 percentage points to 91.5 but

also has a detection speed that far exceeds that of real-time detection, reaching 129.70, making our method superior to the existing techniques.

The WCE_Detection model holds significant clinical relevance for early diagnosis and treatment. By automating digestive tract lesion detection, the model significantly reduces physicians' workload in capsule endoscopy image analysis, facilitates rapid and accurate lesion identification, shortens the diagnosis time, and improves the detection rate of early-stage lesions. This advancement enables patients to receive timely treatment, reduces the incidence of misdiagnosis and underdiagnosis, and ultimately enhances overall healthcare quality. In clinical applications, real-time performance, deployability, and seamless integration with existing clinical workflows are critical considerations. We will continue optimizing the computational efficiency and resource consumption of our models to ensure their efficient operation across various devices and their seamless integration into clinical workflows; they will become a valuable aid for diagnosis and treatment.

While our WCE_Detection model demonstrates substantial advantages in multicategory lesion detection, it is essential to acknowledge certain limitations in dataset size and lesion type coverage. Despite the current dataset encompassing 23 types of digestive tract lesions, real clinical scenarios may involve additional lesion types, and particularly rare or complex lesions. Moreover, although data augmentation was employed to enhance the model's generalization capabilities, the limited dataset size may still impact its performance across a broader range of scenarios. Therefore, future research should emphasize the following aspects. First, addressing complex and rare lesions will be a primary focus of upcoming studies. We aim to expand the dataset and optimize the model specifically for rare lesions. Second, we will further optimize the model's resource consumption to ensure an efficient performance even on resource-limited devices. Finally, we intend to extend the WCE_Detection model's techniques to other medical imaging domains, such as endoscopy, computed tomography, or magnetic resonance imaging. By applying these techniques across other medical imaging technologies, we anticipate enhancing both the breadth and depth of disease detection, thereby supporting clinical diagnosis in additional fields.

CONCLUSION

In this paper, we propose a multicategory lesion detection model, WCE_Detection, for WCE images, which enhances the accuracy and efficiency of detecting digestive tract lesions by incorporating multidetection heads, BiFPN, and Swin Transformer modules. The model is designed on the basis of the YOLOv8 backbone network to enable effective end-to-end detection of 23 digestive tract lesions. The experimental results demonstrate that WCE_Detection attains SOTA performance in terms of lesion detection accuracy, efficiency, and multicategory detection capability, offering robust technical support for the early diagnosis and treatment of digestive tract diseases. In practical applications, the model assists physicians in analyzing WCE images more rapidly and accurately, thereby reducing diagnostic time and enhancing diagnostic quality. Future work will focus on expanding the diversity of the dataset, particularly by leveraging GANs to generate additional images of rare lesions, thereby addressing the issue of data imbalance. Additionally, we will explore deploying the model on resource-constrained devices to further optimize computational efficiency. We aim to extend the model's application to other medical imaging domains to contribute to the advancement of clinical diagnostic techniques, and provide further innovations in clinical diagnostic tools.

FOOTNOTES

Author contributions: Xiao ZG and Chen XQ designed the research and wrote the manuscript; Zhang D and Dai WX collected and analyzed the data; Li XY and Liang WH performed data processing; All authors revised the manuscript and approved the final manuscript.

Supported by The Science and Technology Development Center of The Ministry of Education, No. 2022BC004.

Institutional review board statement: This study was approved by the Ethics Committee of the Affiliated Hospital of Changchun University, No. CCU2023043105.

Informed consent statement: The need for informed consent was waived owing to the retrospective nature of the study.

Conflict-of-interest statement: All the authors report no relevant conflicts of interest for this article.

Data sharing statement: The datasets during the current study are not publicly available due to patient privacy and copyright issues but are available from the corresponding author upon reasonable request at 3220215169@bit.edu.cn.

Country of origin: China

ORCID number: Zhi-Guo Xiao 0000-0002-6145-2807; Xian-Qing Chen 0009-0001-8911-3524.

S-Editor: Li L
L-Editor: A
P-Editor: Yu HG

REFERENCES

1 **Wang Y**, Huang Y, Chase RC, Li T, Ramai D, Li S, Huang X, Antwi SO, Keaveny AP, Pang M. Global Burden of Digestive Diseases: A Systematic Analysis of the Global Burden of Diseases Study, 1990 to 2019. *Gastroenterology* 2023; **165**: 773-783.e15 [PMID: 37302558 DOI: 10.1053/j.gastro.2023.05.050]
2 **Fu Y**, Zhang W, Mandal M, Meng MQ. Computer-aided bleeding detection in WCE video. *IEEE J Biomed Health Inform* 2014; **18**: 636-642 [PMID: 24608063 DOI: 10.1109/JBHI.2013.2257819]
3 **ASGE Technology Committee**, Wang A, Banerjee S, Barth BA, Bhat YM, Chauhan S, Gottlieb KT, Konda V, Maple JT, Murad F, Pfau PR, Pleskow DK, Siddiqui UD, Tokar JL, Rodriguez SA. Wireless capsule endoscopy. *Gastrointest Endosc* 2013; **78**: 805-815 [PMID: 24119509 DOI: 10.1016/j.gie.2013.06.026]
4 **Souaidi M**, Ansari ME. A New Automated Polyp Detection Network MP-FSSD in WCE and Colonoscopy Images Based Fusion Single Shot Multibox Detector and Transfer Learning. *IEEE Access* 2022; **10**: 47124-47140 [DOI: 10.1109/access.2022.3171238]
5 **Soffer S**, Klang E, Shimon O, Nachmias N, Eliakim R, Ben-Horin S, Kopylov U, Barash Y. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc* 2020; **92**: 831-839.e8 [PMID: 32334015 DOI: 10.1016/j.gie.2020.04.039]
6 **Chen H**, Chen J, Peng Q, Sun G, Gan T. Automatic hookworm image detection for wireless capsule endoscopy using hybrid color gradient and contourlet transform. 2013 6th International Conference on Biomedical Engineering and Informatics; 2013; Hangzhou, China: IEEE, 2013: 116-120 [DOI: 10.1109/bmei.2013.6746918]
7 **Krizhevsky A**, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017; **60**: 84-90 [DOI: 10.1145/3065386]
8 **Simonyan K**, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 Preprint. Available from: arXiv:1409.1556 [DOI: 10.48550/arXiv.1409.1556]
9 **He K**, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, United States. IEEE, 2016: 770-778 [DOI: 10.1109/cvpr.2016.90]
10 **Girshick R**, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014; Columbus, OH, United States. IEEE, 2014: 580-587 [DOI: 10.1109/cvpr.2014.81]
11 **Girshick R**. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV); 2015; Santiago, Chile: IEEE, 2015: 1440-1448 [DOI: 10.1109/iccv.2015.169]
12 **Ren S**, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* 2017; **39**: 1137-1149 [PMID: 27295650 DOI: 10.1109/TPAMI.2016.2577031]
13 **Redmon J**, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, United States. IEEE, 2016: 779-788 [DOI: 10.1109/cvpr.2016.91]
14 **Redmon J**, Farhadi A. YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, United States. IEEE, 2017: 6517-6525 [DOI: 10.1109/cvpr.2017.690]
15 **Redmon J**, Farhadi A. Yolov3: An incremental improvement. 2018 Preprint. Available from: arXiv:1804.02767 [DOI: 10.48550/arXiv.1804.02767]
16 **Bochkovskiy A**, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. 2020 Preprint. Available from: arXiv:2004.10934 [DOI: 10.48550/arXiv.2004.10934]
17 **Jocher G**, Stoken A, Chaurasia A, Borovec J, NanoCode012, TaoXie, Kwon Y, Michael K, Changyu L, Fang J, V A, Laughing, tkianai, yxNONG, Skalski P, Hogan A, Nadar J, imyhxy, Mammana L, AlexWang1900, Fati C, Montes D, Hajek J, Diaconu L, Minh MT, Marc albinxavi, fatih, oleg, wanghaoyang0106. Ultralytics yolov5. [cited 3 October 2024]. Available from: https://github.com/ultralytics/yolov5/tree/v7.0/
18 **Li C**, Li L, Geng Y, Jiang H, Cheng M, Zhang B, Ke Z, Xu X, Chu X. YOLOv6 v3.0: A full-scale reloading. 2023 Preprint. Available from: arXiv:2301.05586 [DOI: 10.48550/arXiv.2301.05586]
19 **Wang C**, Bochkovskiy A, Liao HM. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Vancouver, BC, Canada: IEEE, 2023: 7464-7475 [DOI: 10.1109/cvpr52729.2023.00721]
20 **Jocher G**, Chaurasia A, Qiu J. Ultralytics YOLOv8. [cited 3 October 2024]. Available from: https://github.com/ultralytics/ultralytics/
21 **Long X**, Deng K, Wang G, Zhang Y, Dang Q, Gao Y, Shen H, Ren J, Han S, Ding E, Wen S. PP-YOLO: An effective and efficient implementation of object detector. 2020 Preprint. Available from: arXiv:2007.12099 [DOI: 10.48550/arXiv.2007.12099]
22 **Huang X**, Wang X, Lv W, Bai X, Long X, Deng K, Dang Q, Han S, Liu Q, Hu X, Yu D, Ma Y, Yoshie O. PP-YOLOv2: A practical object detector. 2021 Preprint. Available from: arXiv:2104.10419 [DOI: 10.48550/arXiv.2104.10419]
23 **Liu W**, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg AC. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision - ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016 [DOI: 10.1007/978-3-319-46448-0_2]
24 **Xiao Z**, Feng LN. A Study on Wireless Capsule Endoscopy for Small Intestinal Lesions Detection Based on Deep Learning Target Detection. *IEEE Access* 2020; **8**: 159017-159026 [DOI: 10.1109/access.2020.3019888]
25 **Tan M**, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, United States. IEEE, 2020: 10778-10787 [DOI: 10.1109/cvpr42600.2020.01079]

26 **Zhu X**, Lyu S, Wang X, Zhao Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021; Montreal, BC, Canada. IEEE 2021: 2778-2788 [DOI: 10.1109/iccvw54120.2021.00312]

27 **Vaswani A**, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS); 2017 Dec 4-9; Long Beach, CA, United States. NIPS, 2017; 30

28 **Dosovitskiy A**, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020 Preprint. Available from: arXiv:2010.11929 [DOI: 10.48550/arXiv.2010.11929]

29 **Liu Z**, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021; Montreal, QC, Canada. IEEE, 2021: 9992-10002 [DOI: 10.1109/iccv48922.2021.00986]

30 **Lin T**, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature Pyramid Networks for Object Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, United States. IEEE, 2017: 936-944 [DOI: 10.1109/cvpr.2017.106]

31 **Liu S**, Qi L, Qin H, Shi J, Jia J. Path Aggregation Network for Instance Segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, UT, United States. IEEE, 2018: 8759-8768 [DOI: 10.1109/cvpr.2018.00913]

32 **Wang CY**, Yeh IH, Liao HYM. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. 2024 Preprint. Available from: arXiv:2402.13616 [DOI: 10.48550/arXiv.2402.13616]

33 **Zhao Y**, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, Chen J. Detrs beat yolos on real-time object detection. 2023 Preprint. Available from: arXiv:2304.08069 [DOI: 10.48550/arXiv.2304.08069]

34 **Goodfellow I**, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS); 2014 Dec 8-13; Montreal, Canada. NIPS, 2014: 2672-2680

35 **Rahim T**, Usman MA, Shin SY. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Comput Med Imaging Graph* 2020; **85**: 101767 [PMID: 32966967 DOI: 10.1016/j.compmedimag.2020.101767]

36 **Muhammad K**, Khan S, Kumar N, Del Ser J, Mirjalili S. Vision-based personalized Wireless Capsule Endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges. *Future Gener Comput Syst* 2020; **113**: 266-280 [DOI: 10.1016/j.future.2020.06.048]

37 **Xiao J**, Meng MQ. A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images. *Annu Int Conf IEEE Eng Med Biol Soc* 2016; **2016**: 639-642 [PMID: 28268409 DOI: 10.1109/EMBC.2016.7590783]

38 **Yuan Y**, Li B, Meng MQ. Improved Bag of Feature for Automatic Polyp Detection in Wireless Capsule Endoscopy Images. *IEEE Trans Automat Sci Eng* 2016; **13**: 529-535 [DOI: 10.1109/tase.2015.2395429]

39 **Pogorelov K**, Suman S, Azmadi Hussin F, Saeed Malik A, Ostroukhova O, Riegler M, Halvorsen P, Hooi Ho S, Goh KL. Bleeding detection in wireless capsule endoscopy videos - Color versus texture features. *J Appl Clin Med Phys* 2019; **20**: 141-154 [PMID: 31251460 DOI: 10.1002/acm2.12662]

40 **Yeh J**, Wu T, Tsai W. Bleeding and Ulcer Detection Using Wireless Capsule Endoscopy Images. *J Softw Eng Application* 2014; **07**: 422-432 [DOI: 10.4236/jsea.2014.75039]

41 **Hwang S**, Celebi ME. Polyp detection in Wireless Capsule Endoscopy videos based on image segmentation and geometric feature. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing; 2010; Dallas, TX, United States. IEEE, 2010: 678-681 [DOI: 10.1109/icassp.2010.5495103]

42 **Aoki T**, Yamada A, Aoyama K, Saito H, Tsuboi A, Nakada A, Niikura R, Fujishiro M, Oka S, Ishihara S, Matsuda T, Tanaka S, Koike K, Tada T. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc* 2019; **89**: 357-363.e2 [PMID: 30670179 DOI: 10.1016/j.gie.2018.10.027]

43 **Li P**, Li Z, Gao F, Wan L, Yu J. Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images. 2017 IEEE International Conference on Multimedia and Expo (ICME); 2017; Hong Kong, China. IEEE, 2017: 1518-1523 [DOI: 10.1109/icme.2017.8019415]

44 **Shin Y**, Qadir HA, Aabakken L, Bergsland J, Balasingham I. Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches. *IEEE Access* 2018; **6**: 40950-40962 [DOI: 10.1109/access.2018.2856402]

45 **Thuan NH**, Oanh NT, Thuy NT, Perry S, Sang DV. Rabit: An efficient transformer using bidirectional feature pyramid network with reverse attention for colon polyp segmentation. 2023 Preprint. Available from: arXiv:2307.06420 [DOI: 10.48550/arXiv.2307.06420]

46 **Park K**, Lee JY. SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J Comput Des Eng* 2022; **9**: 616-632 [DOI: 10.1093/jcde/qwac018]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: office@baishideng.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com