Journal of Medical Systems (2025) 49:129
https://doi.org/10.1007/s10916-025-02271-3

RESEARCH

Revisiting Challenges in Real-world Video Colonoscopy using End-to-End Two Stream Polyp Detection Transformer (TS-PDTR)

Tianyuan Gan[1] · Chongan Zhang[1] · Peng Wang[1] · Xiao Liang[2] · Xuesong Ye[1]

Received: 28 November 2024 / Accepted: 19 September 2025

Abstract

Accurate polyp detection is essential for the early diagnosis and effective treatment of colorectal cancer (CRC). However, colonoscopy videos in real-world clinical settings present significant challenges, often causing existing algorithms to fail. Compared to single images, videos contain richer temporal and contextual information, making them valuable for developing deep-learning-based detection systems. To address these challenges, we propose an end-to-end Two-Stream Polyp Detection Transformer (TS-PDTR) network. First, our framework uses a two-stream feature extraction network to capture both spatial and temporal features from the RGB frames and optical flow. Then, the proposed Detail-Aware Convolution (DAConv) module enhances fine-grained contextual information in low-level features. Following this, the Detail-Guided Attention (DGA) module generates channel-specific Spatial Attention Maps (SAMs) to refine deep feature maps, improving the model's sensitivity to small and camouflaged polyps. Finally, a Flow Fusion Encoder (FFE) module combines temporal cues from optical flow to increase robustness against poor single-frame image quality. Experiments on three benchmark video colonoscopy datasets show that TS-PDTR consistently outperforms previous state-of-the-art image- and video-based polyp detection methods. Notably, our model achieves a mean Average Precision (mAP) of 33.2 on the most challenging LDPolypVideo dataset. It also improves the mAP to 64.0 and 55.6 on the SUN Colonoscopy Video Database and CVC-VideoClinicDB, respectively. In summary, TS-PDTR is a promising video-based polyp detection method with strong potential for further development and real-world clinical application.

Keywords Video polyp detection · Colonoscopy · Detection transformer · Optical flow

✉ Xuesong Ye
yexuesong@zju.edu.cn

Tianyuan Gan
gantianyuan@zju.edu.cn

Chongan Zhang
kevin_07@zju.edu.cn

Peng Wang
0016902@zju.edu.cn

Xiao Liang
srrshlx@163.com

[1] Biosensor National Special Laboratory, College of Biomedical Engineering and Instrument Science, Zhejiang University, No. 38, Zheda Road, 310027 Hangzhou, Zhejiang, China

[2] Department of General Surgery, Sir Run-Run Shaw Hospital, School of Medicine, Zhejiang University, No. 3, Qingchun East Road, 310016 Hangzhou, Zhejiang, China

List of Abbreviations

AI	Artificial Intelligence
AP	Average Precision
ADC	Angular Difference Convolution
CADe	Computer-Aided Detection
CAD	Computer-Aided Diagnosis
CBAM	Convolutional Block Attention Module
CDC	Central Difference Convolution
CDN	Contrastive DeNoising Training
ConvLSTM	Convolutional Long Short-Term Memory
CRC	colorectal cancer
DAConv	Detail-Aware Convolution
DC	Difference Convolution
DETR	Detection Transformer
DFT	Discrete Fourier Transform
DGA	Detail-Guided Attention
3D-CNN	three-dimensional convolutional neural network
FAM	Feature Attention Mechanism

Published online: 08 October 2025

FFE	Flow Fusion Encoder
FN	false negative
FP	false positive
HDC	Horizontal Difference Convolution
IoU	Intersection over Union
LLMs	Large Language Models
LR	learning rate
LSTM	Long Short-Term Memory
mAP	mean Average Precision
NMS	non-maximum suppression
PEFT	parameter-efficient fine-tuning
RCNN	region-based convolutional neural network
RPN	region proposal network
SAMs	Spatial Attention Maps
SEM	Squeeze-and-Excitation Module
SOTA	state-of-the-art
SSD	Single Shot Detector
TP	true positive
TS-PDTR	Two-Stream Polyp Detection Transformer
VC	vanilla convolution
VDC	Vertical Difference Convolution
ViTs	Vision Transformers
VFMs	Vision Foundation Models
YOLO	You Only Look Once

Introduction

Colorectal cancer (CRC) remains one of the three leading causes of cancer-related deaths worldwide. In China alone, about 521,000 new CRC cases are diagnosed each year, resulting in approximately 248,000 deaths [1], making it the fifth leading cause of cancer mortality in the country. Evidence shows that even a 1% increase in the adenoma detection rate can reduce CRC incidence by 3% to 6% [2]. Therefore, early detection and screening play a critical role in lowering mortality risk. Colonoscopy remains the gold standard for CRC screening and diagnosis. During the procedure, a thin, flexible tube is inserted into the colorectum to allow direct visual inspection by endoscopists. This enables real-time observation and accurate diagnosis through targeted biopsies of suspicious polyps. However, due to long working hours and varying levels of experience, endoscopists may miss polyps or make false detections during colonoscopy. To address this issue, Computer-Aided Diagnosis (CAD) systems have been developed to improve the efficiency and quality of colonoscopy in routine practice, acting as a 'third eye' for endoscopists. In recent years, advances in deep learning have significantly boosted the performance of CAD systems, thanks to breakthroughs in algorithms, improved hardware and computing power, and the growing availability of large annotated endoscopic datasets.

However, most existing methods are designed and validated using well-curated and balanced public colonoscopy datasets of still images. In real clinical scenarios, these methods often fail when facing complex and challenging situations [3], highlighting critical gaps that must be addressed in the development of robust colonoscopy CAD systems. Therefore, we revisit the polyp detection task in real-world scenarios and identify three main challenges in colonoscopy videos: **(1) Poor frame quality.** Unlike videos from ordinary scenes, colonoscopy videos often involve rapid camera movement, instrument manipulation, and a complex intestinal environment. This leads to various disturbances in the visual field, such as motion blur, occlusion, defocus, dispersion, reflections, bubbles, fluid flow, or inadequate bowel preparation, all of which make accurate polyp detection in the single frame difficult. **(2) Camouflaged polyps.** Some polyps are hard to distinguish because they blend in with the surrounding intestinal tissue due to low contrast and high similarity. This camouflage effect can confuse detection models, often resulting in false positives or missed detections. **(3) Small-sized polyps.** Small lesions occupy very few pixels, which can cause fine structural features to be lost in deeper network layers. In addition, their subtle context with limited semantics makes it harder for models to differentiate them reliably.

To tackle these issues, this paper systematically leverages temporal information in colonoscopy videos along with the unique characteristics of colorectal polyps. We propose the Two-Stream Polyp Detection Transformer (TS-PDTR) framework for accurate polyp detection in real-world clinical environments. The main contributions of this study are summarized as follows:

- The Detail-Aware Convolution (DAConv) module is designed to restore fine details of small polyps lost in deep feature maps and to improve the discrimination of camouflaged polyps from the surrounding intestinal walls in the high-dimensional embedding space. By combining parallel vanilla and difference convolutions, DAConv encodes priors that highlight high-frequency texture and structural cues in shallow features.
- The Detail-Guided Attention (DGA) module is introduced to produce channel-specific Spatial Attention Maps (SAMs) in a coarse-to-fine fashion. These SAMs are assigned to high-level features, guiding the model to focus on regions likely to contain small or camouflaged polyps.
- To fully leverage temporal information in colonoscopy videos, we propose a two-stream feature extraction network. The RGB encoder extracts multi-scale spatial features of polyps, while the flow encoder captures temporal representations from optical flow maps. The Flow

Fusion Encoder (FFE) module then fuses these features to compensate for poor single-frame image quality.
- We conduct extensive experiments on three public video polyp detection benchmarks. The TS-PDTR network, integrating DAConv, DGA, and FFE, achieves superior performance compared to existing state-of-the-art (SOTA) image-based and video-based methods, showing strong potential for real clinical use.

Related Work

Polyp Detection

A real-time Computer-Aided Detection (CADe) system plays a crucial role in colonoscopy procedures, assisting clinicians in identifying polyps that may be easily overlooked. In some research contexts, terms such as 'Polyp Detection,' 'Polyp Recognition,' and 'Polyp Classification' have been used interchangeably, leading to confusion. To provide clarity in this paper, 'Polyp Detection' specifically refers to the identification and localization of polyp lesions within a frame, rather than simply determining the presence of polyp lesions at the image level. Numerous studies have investigated computer-aided methods for polyp detection. However, accurately recognizing and localizing polyps remains challenging due to their complex characteristics. The spatial attributes of polyp images, including size, shape, texture, and contrast with their surroundings, vary across frames, making it difficult to consistently detect abnormalities under all conditions [4]. Recent advancements in colonoscopy CAD have emphasized the integration of artificial intelligence (AI) techniques to enhance lesion detection. Various methodologies have been proposed to facilitate polyp detection through machine learning, primarily focusing on handcrafted feature extraction.

In the early stages, Karkanis et al. [5] introduced a descriptor that combines color wavelet features with a sliding window mechanism for polyp detection during colonoscopy. This innovation has spurred subsequent researchers [6–8] to leverage color, texture, and Haar descriptors, enabling the extraction of more intuitive features from colonoscopic images to depict polyp-like anomalies. Recently, advancements in deep learning have facilitated the exploration of deeper neural network layers, automatically extracting more representative features from images [9].

Some object detectors based on deep learning have been proposed that can be transferred to colonoscopy images for polyp detection with a much larger scale dataset. The region-based convolutional neural network (RCNN) algorithm, initially proposed by [10], employs high-capacity convolutional neural networks to identify objects of interest. It autonomously extracts features from specific regions for subsequent analysis. Numerous studies, including [11–13], have validated that RCNN-based models effectively identify polyps of varying shapes and sizes. In 2015, Ren et al. [14] introduced an enhanced version, termed Faster R-CNN. This iteration retains RCNN's essence but incorporates a region proposal network (RPN) that efficiently shares convolutional features with the detection network. Consequently, Faster R-CNN is well-suited for real-time detection applications and has underpinned models like those in [15–18]. Another notable approach emerged in 2016 with the Single Shot Detector (SSD), introduced by [19]. SSD employs a singular deep neural network for object detection and discretizes the bounding box output space into multiple boxes with varied aspect ratios. This approach employs feature maps of varying resolutions to predict object locations and has been incorporated into recent colorectal polyp detection techniques such as those presented in [20, 21]. The You Only Look Once (YOLO) algorithm, initially unveiled in 2016 by [22], stands as a prominent object detection method. YOLOv1 treats detection as a regression task, enabling a singular neural network to predict bounding boxes and class probabilities for an entire image in one pass. Renowned for its real-time processing capability of more than 45 frames per second, the YOLO framework has undergone iterative improvements in both performance and efficiency over the past decade [23–31]. Recent research [18, 32–39] underscores YOLO's prevalent use over alternative detection algorithms.

Detection Transformer

Mainstream detection algorithms had long been dominated by convolutional neural network-based frameworks until recently when Transformer-based detectors made significant strides. End-to-end object detectors are well-known for their streamlined workflows. Carion et al. [40] introduced the first end-to-end object detector based on the Transformer architecture, known as Detection Transformer (DETR). It has gained substantial attention due to its unique characteristics. Notably, DETR eliminates the need for manually designed anchors and non-maximum suppression (NMS) components commonly found in traditional detection pipelines. Instead, it employs bipartite matching and directly predicts one-to-one object relationships. This approach simplifies the object detection process and mitigates performance bottlenecks associated with NMS. Despite its evident advantages, DETR faces two significant challenges: slow convergence and difficulties in optimizing queries. Numerous variants of DETR have been proposed to tackle these issues.

Deformable DETR [41] incorporates an efficient deformable attention mechanism to expedite training convergence

with multi-scale features. Conditional DETR [42] addresses query optimization difficulties by decoupling queries into content and location components. Efficient DETR [43] enhances decoder queries by selecting the top K positions from the encoder's dense predictions. DAB-DETR [44] interprets queries as 4D anchor box coordinates and progressively refines prediction boxes layer by layer within the decoder. More recently, DN-DETR [45] introduces a denoising training approach to address unstable bipartite graph matching so as to speed up DETR training. This method introduces noise-added ground-truth labels and boxes into the decoder, training the model to reconstruct the original ones. Finally, DINO [46] builds several new improvements upon previous works and achieves new SOTA results on COCO [47] detection.

Temporal Information in Colonoscopy Video

To distinguish polyps from healthy tissues in colonoscopy videos, it is natural for the methods mentioned above to initially decompose the input video into a sequence of still images arranged chronologically. Each individual frame includes the visual information of the background and the polyps at the specific moment. Identifying polyps on a static image becomes relatively straightforward when the background is clear, and the polyp is prominent. However, when dealing with colonoscopy videos, mere still image analysis falls short due to the diverse appearances of polyps and the presence of background noise in clinical settings. All these methods have suffered from poor performance. Some researchers have attempted to leverage the relationships among consecutive frames from a temporal perspective to enhance polyp detection in colonoscopy videos. Yu et al. [48] and Puyal et al. [49] employed a three-dimensional convolutional neural network (3D-CNN) to introduce spatial-temporal features by analyzing consecutive polyp frames, thereby improving polyp detection performance. Other studies explored the use of temporal information through tracking-based systems [50, 51]. However, 3D-CNN falls short of meeting real clinical requirements due to its computational complexity, making it challenging to process live streaming data. Additionally, there are uninformative frames that need exclusion during the inference stage, limiting the utility of polyp trackers. Different from these approaches, optical flow, a pattern representing the apparent motion of objects, can be generated in real-time to extract short-term temporal features from video streams. These temporal features capture the displacement of each pixel across frames and visualize the changes in polyp appearance caused by the movement of the colon and camera. Zhang et al. [52] and Zheng et al. [53] employed optical flow to establish connections between feature points across frames, enabling the model to adapt to different views of a polyp as well as its circumstance for robust detection.

Method

This section presents the details of the proposed TS-PDTR framework for accurate polyp detection in colonoscopy videos. An overview of the complete architecture is illustrated in Fig. 1. We first outline DINO, the base detector for our model. We then describe each core component in detail: the Two-Stream Feature Extraction Network, the Detail-Aware Convolution (DAConv) module, the Detail-Guided Attention (DGA) module, and the Flow Fusion Encoder (FFE) module.

DINO Briefing

DINO is an end-to-end DETR-like model that uses a Transformer encoder-decoder architecture and views object detection as a direct set prediction task, thereby eliminating the need for hand-crafted designed components such as NMS or anchor generation. This model enhances its capabilities based on Deformable DETR, DAB-DETR, and DN-DETR. Specifically, DINO incorporates deformable attention modules capable of focusing on a localized set of key sampling points around a reference position within multi-scale feature maps, thereby augmenting computational efficiency and mitigating the slow convergence and limited feature spatial resolution issues in existing DETR variants. Meanwhile, DINO proposes a novel query formulation approach termed mixed query selection that initializes anchor box coordinates as positional queries for the Transformer decoder with the selected top-k features from the Transformer encoder while leaving the content queries as static embeddings and learnable as before. This scheme helps to use better positional information to pool more comprehensive content features from the encoder. Then, the queries are progressively refined layer-by-layer within the decoder for accurate box prediction. To fully exploit the refined box information from subsequent layers for the optimization of parameters of their neighboring early layer, an innovative look-forward-twice technique is introduced for gradient propagation between adjacent layers. Moreover, DINO further presents a unique Contrastive DeNoising Training (CDN) methodology following DN-DETR. The method introduces a query denoising task where ground truth bounding boxes with different intensities of noise are fed into the Transformer decoder to reconstruct the original boxes. This enhances the model's ability to predict 'no-object' for anchors with no nearby objects while reducing the difficulty of bipartite graph matching during training. DINO achieves significant

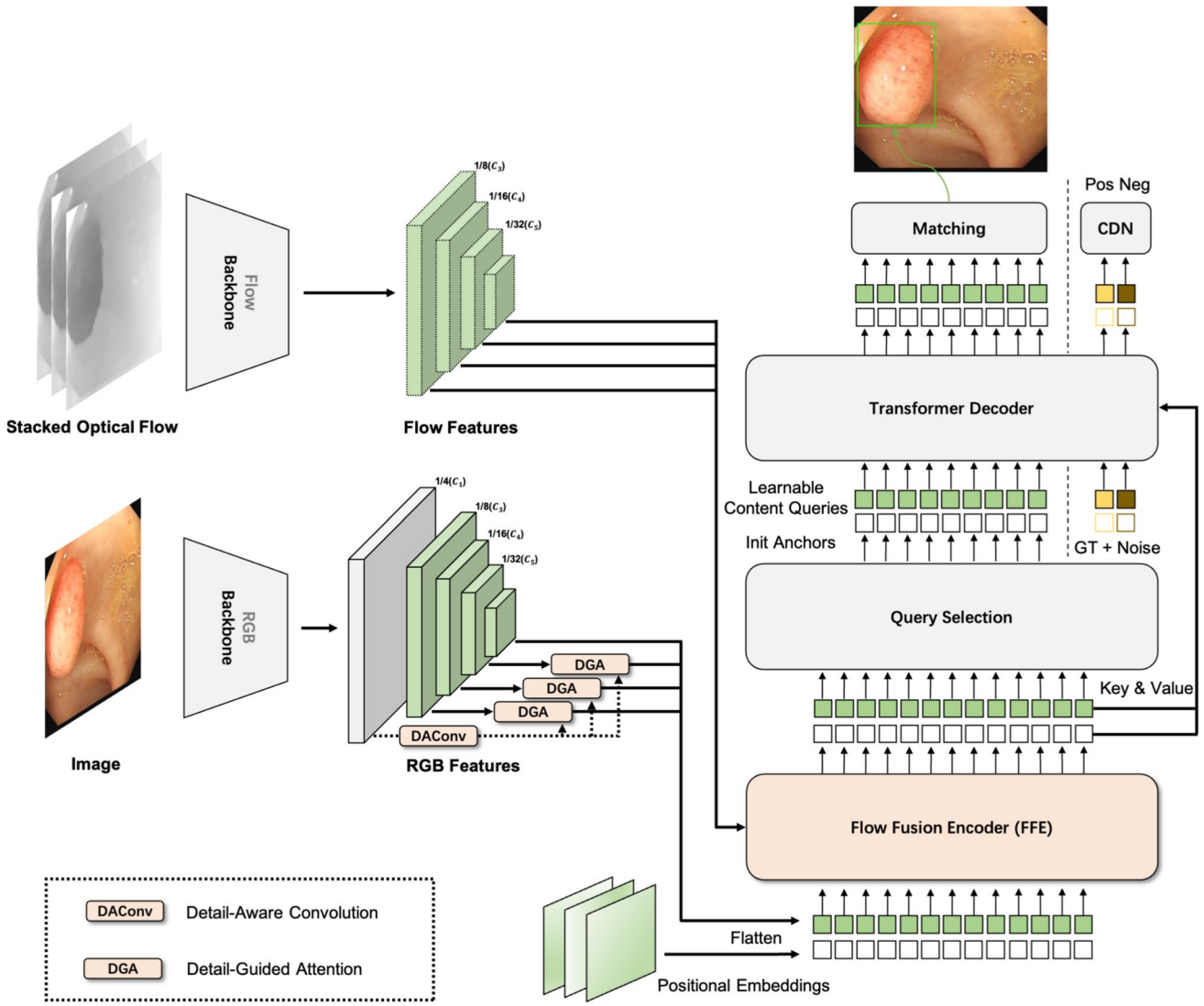

Fig. 1 The overall workflow of the proposed TS-PDTR polyp detection framework for colonoscopy video

advancements in both performance and efficiency compared to previous DETR-like models. Therefore, we leverage the DINO as our foundational detector.

Two-Stream Feature Extraction Network

Colonoscopy videos naturally contain both spatial and temporal information. The spatial component comes from the appearance of each individual frame, which shows the intestinal scene and polyps. The temporal component captures motion between frames, reflecting the movement of the colonoscopy camera and the polyps. To exploit both aspects, our feature extraction network uses a two-stream design (Fig. 1), with separate deep ConvNets or Vision Transformers (ViTs) for the spatial and temporal branches.

Spatial Branch The RGB branch handles the spatial stream, extracting features from individual video frames. Static appearance alone often provides useful clues, as some polyps are clearly visible in the single frame. As shown in Section "Model Components Ablation", detection using only the spatial branch already yields strong results. Since this branch is essentially an image classifier, it can benefit from advances in large-scale image recognition and be pre-trained on datasets like ImageNet [54]. Implementation details are given in Section "Implementation Details".

Temporal Branch To model temporal features in the colonoscopy video, we adopt an optical flow stream, inspired by its proven success in video recognition [55]. The temporal branch uses a dedicated flow backbone to capture motion cues and boost detection accuracy. Unlike the RGB branch,

this stream processes stacked optical flow displacement fields of several consecutive frames, with mean flow subtracted. This explicit motion representation reduces the burden of implicit motion estimation by the detection network.

Specifically, the dense optical flow is treated as a set of displacement vector fields d_t between each pair of consecutive frames t and $t+1$. The displacement vector at a pixel (u, v) in frame t, denoted by $d_t(u, v)$, describes how this point shifts to its new position in the next frame $t+1$. The horizontal and vertical components of these vectors, d_t^x and d_t^y, are used as image channels, making them suitable for feature extraction with deep neural networks. To capture motion across multiple frames, the flow channels $d_t^{x,y}$ from L consecutive frames are concatenated, producing an input with $2L$ channels.

Formally, letting w and h denote the frame width and height, the input volume $I_\tau \in \mathbb{R}^{w\times h\times 2L}$ for any frame τ is defined as follows:

$$I_\tau(u, v, 2k-1) = d^x_{\tau+k-1}(u, v) \tag{1}$$

$$I_\tau(u, v, 2k) = d^y_{\tau+k-1}(u, v) \tag{2}$$

where $u \in [1, w]$, $v \in [1, h]$, and $k \in [1, L]$. For any given point (u, v), the channels $I_\tau(u, v, c)$, with $c \in [1, 2L]$, encode its motion over the sequence of L frames.
In addition, zero-centering the network input is generally beneficial, as it helps the model better exploit the rectified non-linearity. In our case, the displacement vector components can take both positive and negative values, which naturally fits the concept of centering. Across diverse motion scenarios, movement in any direction holds equal probability. However, when computing optical flow for a frame pair, the result can be strongly affected by large displacements caused by rapid camera movement during colonoscopy. Therefore, compensating for camera motion is necessary, which requires estimating the global motion component and subtracting it from the dense flow. In our approach, we implement this by simply subtracting the mean vector from each displacement field d. Despite this adjustment, the backbone architecture for optical flow largely mirrors that of the RGB branch, with the main differences being the input layer's channel count and the network initialization strategy.

Detail-Aware Convolution

In colorectal polyp detection, most existing methods typically use vanilla convolution (VC) layers to extract image features [15, 20, 32]. Standard convolution layers search an unconstrained solution space (often initialized randomly), which can limit their ability to represent fine textural and structural details. For small or camouflaged polyps, their small size and strong similarity to the surrounding intestinal tissue make high-frequency information (e.g., edges and contours) crucial for preserving fine details. However, vanilla convolutions naturally focus more on low-frequency content and often neglect high-frequency details. Some earlier studies [56, 57] introduced edge priors into polyp detection networks to recover sharper boundaries. Building on these ideas, we propose a Detail-Aware Convolution (DAConv) module (see Fig. 2) that explicitly captures both low- and high-frequency components. This is accomplished by integrating carefully designed priors into standard convolution layers. Specifically, by combining Difference Convolution (DC) layers with the vanilla convolution in parallel.

Before describing DAConv in detail, we briefly revisit Difference Convolution. Previous works [58, 59] define DC as a convolution operation applied to pixel differences: first, pixel differences are computed, then convolved with learned kernels to produce feature maps. This process enhances the representation and generalization capacity of standard convolutions. Common DC variants include Central Difference Convolution (CDC) and Angular Difference Convolution (ADC). As shown in Fig. 2, CDC computes pixel differences based on the central pixel and its immediate neighbors, emphasizing local intensity variations, while ADC calculates pixel differences based on angular relationships between the pixels, capturing local directional features. These methods have shown effectiveness in tasks such as edge detection [59], image dehazing [60], and face anti-spoofing [58]. To our knowledge, this is the first time DC has been applied to colorectal polyp detection.

In our design, we deploy five convolution layers in parallel for feature extraction in the DAConv, including four DCs and one vanilla convolution. The DC layers use tailored strategies for computing pixel pair differences to explicitly encode prior information. Different from the prior usage of CDC and ADC, as illustrated in Fig. 2, our DAConv also includes Horizontal Difference Convolution (HDC) and Vertical Difference Convolution (VDC) to embed traditional local descriptors (like Sobel, Prewitt, or Scharr) directly into the convolution process. Both HDC and VDC explicitly embed gradient priors, improving the capturing of subtle edge details and meaningful local gradient variation cues in the horizontal and vertical directions respectively. Taking HDC as an example, the horizontal gradient is first calculated by measuring the difference between specific pixel pairs. Then, the learned kernel weights are equivalently reformulated and directly convolved with the input features during inference. Notably, the resulting equivalent kernel has the same form as classic local operators, where the horizontal weights sum to zero. Traditional horizontal

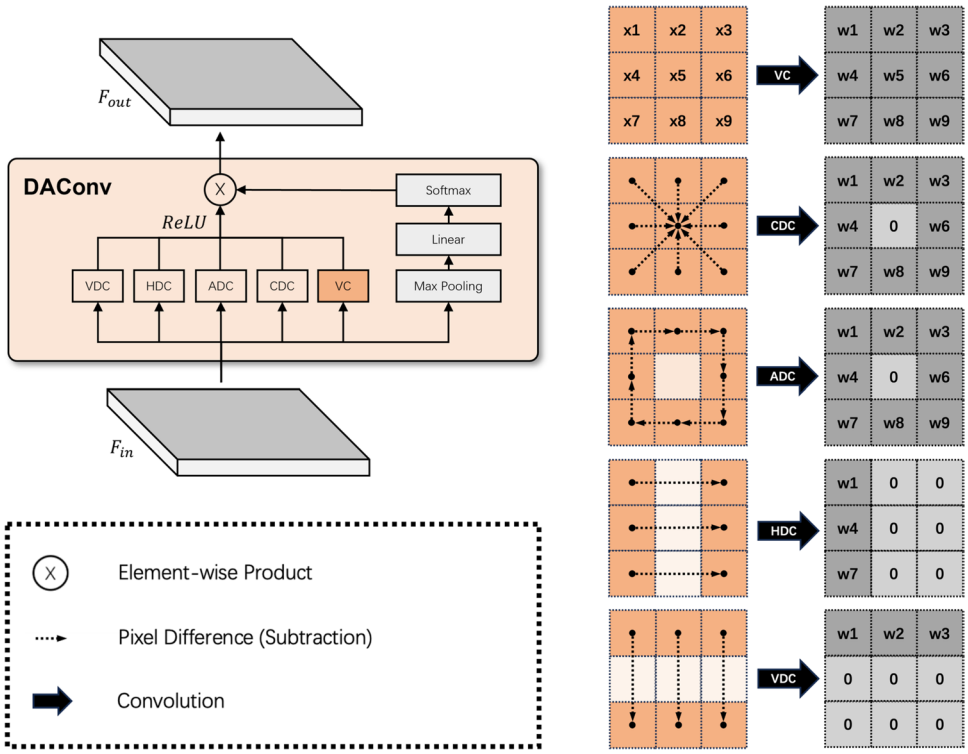

Fig. 2 Illustration of the proposed Detail-Aware Convolution module

kernels like Sobel, Prewitt, and Scharr are special cases of this formulation. The VDC is derived in the same way but uses the vertical gradient instead.

In specific implementation, low-frequency global information is captured by the vanilla convolution, while high-frequency local intensity, direction, and gradient details are strengthened by the DC layers. Unlike the simple summation in previous works, the final output of DAConv is obtained by adaptively combining these features using weights computed via a linear projection and a Softmax function on the input, which allows the model to dynamically emphasize the most relevant features, enhancing detection accuracy, especially for small and camouflaged polyps. Although more advanced strategies for computing pixel differences could further improve polyp detection performance, this is beyond the scope of this study.

Formally, given the input features F_{in}, DAConv produces the enhanced output features F_{out} as follows (bias terms are omitted for clarity):

$$F_{out} = DAConv(F_{in}) = \sum_{i=1}^{5} w_i K_i * F_{in} \tag{3}$$

$$\{w_i\}_{i=1}^{5} = Softmax(Linear(Maxpool(F_{in}))) \tag{4}$$

where $\{K_i\}_{i=1}^{5}$ represents the convolution kernels of VC, CDC, ADC, HDC, and VDC, respectively, * denotes the convolution operation, and $\{w_i\}_{i=1}^{5}$ denotes the adaptive weights to sum up the features learned by different convolution kernels.

Detail-Guided Attention

The Feature Attention Mechanism (FAM) is a widely used approach for guiding the model's focus to the foreground polyp regions. It typically consists of channel attention and spatial attention components, which are applied sequentially to compute attention weights across the channel and spatial dimensions. The channel attention part computes a channel-wise vector $W_c \in \mathbb{R}^{1\times1\times C}$ to re-calibrate the feature maps. Meanwhile, the spatial attention component generates a Spatial Attention Map, denoted as $W_s \in \mathbb{R}^{H\times W}$, to dynamically emphasize informative regions. By assigning different weights to different channels and spatial positions, FAM helps improve polyp detection performance.

However, we argue that there are three main limitations to the current use of FAM. First, there is no interaction between the two attention weights, W_c and W_s, as they are computed sequentially and enhance the features independently. Second, the spatial attention part of FAM only addresses the uneven distribution of polyp regions at the image level through a single-channel W_s, while ignoring variations across different feature channels. In fact, the spatial representation of polyp regions within each feature channel can vary greatly, sometimes even being completely opposite. Therefore, channel-specific SAMs are needed. Third, meaningful representations of polyps are often lost in

deeper layers because small polyps occupy few pixels and often resemble the background. Relying only on deep feature maps to compute attention weights is thus less effective for detecting small or camouflaged polyps.

To overcome these limitations, we propose a Detail-Guided Attention (DGA) module. This module generates channel-specific SAMs for each input feature channel in a progressive manner, ensuring close integration between channel and spatial attention to enable effective information exchange. DGA uses the high-resolution feature map, which captures rich texture and structural details, obtained through DAConv, to guide the generation of SAMs.

The specific steps of the DGA module are shown in Fig. 3. Here, $F_{in} \in \mathbb{R}^{H \times W \times C}$ denotes the input features, while $F_{ref} \in \mathbb{R}^{H_r \times W_r \times C_r}$ represents the high-resolution reference feature map from DAConv. The main goal of DGA is to produce channel-specific SAMs W with the same dimensions as F_{in}. Next, we will explain this process step by step in a formalized manner:

Step 1. Calculation of channel attention W_c and spatial attention W_s

We first compute W_c and W_s using the following equations, according to [61, 62]:

$$W_c = Conv_{1\times1}(ReLU(Conv_{1\times1}(GAP_s(F_{in})))) \tag{5}$$

$$W_s = Conv_{7\times7}([GAP_c(F_{in}), GMP_c(F_{in})]) \tag{6}$$

where $Conv_{k\times k}(\cdot)$ denotes a convolution layer with a $k \times k$ kernel, and $[\cdot,\cdot]$ indicates channel-wise concatenation. $GAP_s(\cdot)$, $GAP_c(\cdot)$, and $GMP_c(\cdot)$ represent global average pooling over spatial dimensions, global average pooling over channels, and global max pooling over channels, respectively.

To reduce parameter count and control model complexity, the first 1×1 convolution within W_c reduces the channel dimension from C to $\frac{C}{r}$ (where r is the reduction ratio, set to 16 in our implementation), while the second 1×1 convolution restores it back to C.

Step 2. Calculation of coarse channel-specific SAMs W_{coa} Next, W_c and W_s are fused using simple addition (following broadcasting rules) to produce the coarse SAMs $W_{coa} \in \mathbb{R}^{H\times W\times C}$:

$$W_{coa} = W_c + W_s \tag{7}$$

Note that W_{coa} and F_{in} are channel-wise aligned because W_c applies at the channel level.

Step 3. Calculation of refined channel-specific SAMs W The final refined channel-specific SAMs W are generated by refining each channel of W_{coa} under the guidance of the corresponding reference detail features F_{ref}. Specifically, F_{ref} is first down-sampled to match the height and width of the coarse SAMs W_{coa}. Subsequently, both W_{coa} and F_{ref} undergo an alternating channel shuffle operation as described in [63]. This step, combined with the subsequent group convolution layers, greatly reduces the parameter count. The detailed formula is as follows:

$$W = \sigma(GConv_{3\times3}(CS([LMP(F_{ref}), W_{coa}]))) \tag{8}$$

where σ is the sigmoid function, $LMP(\cdot)$ is local max pooling for down-sampling, $[\cdot,\cdot]$ is the channel-wise concatenation, $CS(\cdot)$ is the channel shuffle operation, and $GConv_{k\times k}(\cdot)$ is a group convolution with a $k \times k$ kernel. In our implementation, the group number equals C.

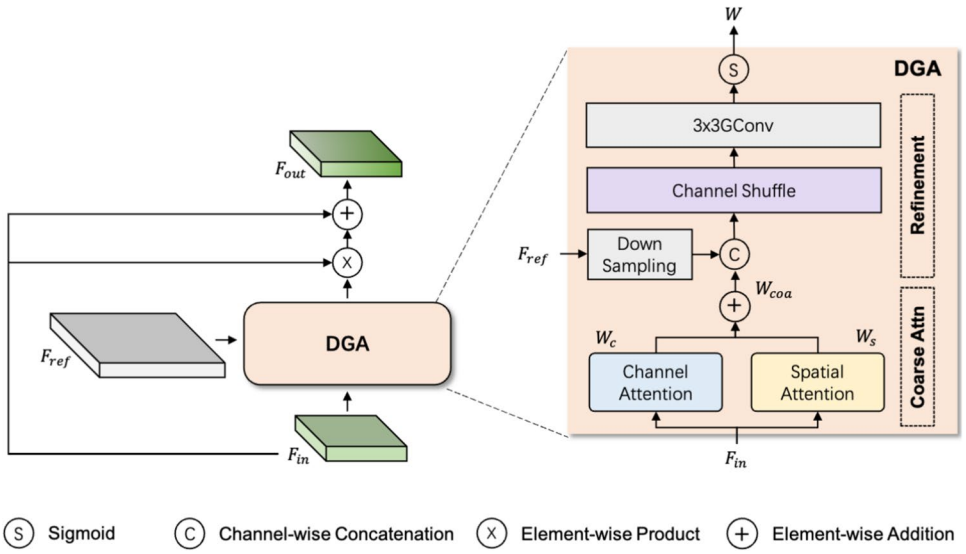

Fig. 3 Illustration of the proposed Detail-Guided Attention module

In summary, the DGA module generates a unique Spatial Attention Map for each channel, directing the model's focus to important regions within every channel. Finally, the output features F_{out} are computed under the guidance of W, with a skip connection included to mitigate gradient vanishing and support stable learning. As a result, the network can better emphasize informative features and improve detection performance.

Flow Fusion Encoder

As illustrated in Fig. 4, the Flow Fusion Encoder (FFE) module comprises N layers, each containing a self-attention block, a flow cross-attention block, and a feed-forward network. Following the standard Transformer architecture, the RGB features initialized with positional embeddings are first processed by the self-attention block as input queries. Next, the flow cross-attention operation integrates the optical flow features into the RGB queries. After passing through the feed-forward network, each encoder layer outputs the updated features, which are then used as queries for the next layer. After N layers of fusion encoding, the final fused features are produced.

Self-Attention To reduce computational costs and efficiently integrate information from multi-scale feature maps, we implement the self-attention operation using multi-scale deformable attention [41]. Each RGB query interacts with only a small, fixed set of corresponding key elements. This process is achieved through key sampling around a reference point, regardless of the spatial resolution of the feature maps, as illustrated below:

$$SA(Q) = MSDeformAttn(z_q, p_q, \{x^l_{RGB}\}^L_{l=1}) \quad (9)$$

where q indexes a query element Q with content feature z_q and the normalized coordinates of the 2-D reference point $p_q \in [0,1]^2$. $\{x^l_{RGB}\}^L_{l=1}$ denotes the input multi-scale RGB feature maps, where l indexes the input feature level and each $x^l_{RGB} \in \mathbb{R}^{W_l \times H_l \times C}$.

Flow Cross-Attention The flow cross-attention operation is also implemented based on the multi-scale deformable attention, but it differs from self-attention in the sampling source of the key elements. The specific flow cross-attention process can be formulated as:

$$FCA(Q) = MSDeformAttn(z_q, p_q, \{x^l_{Flow}\}^L_{l=1}) \quad (10)$$

where $\{x^l_{Flow}\}^L_{l=1}$ represents the multi-scale optical flow feature maps produced by the temporal branch described in Section "Two-Stream Feature Extraction Network".

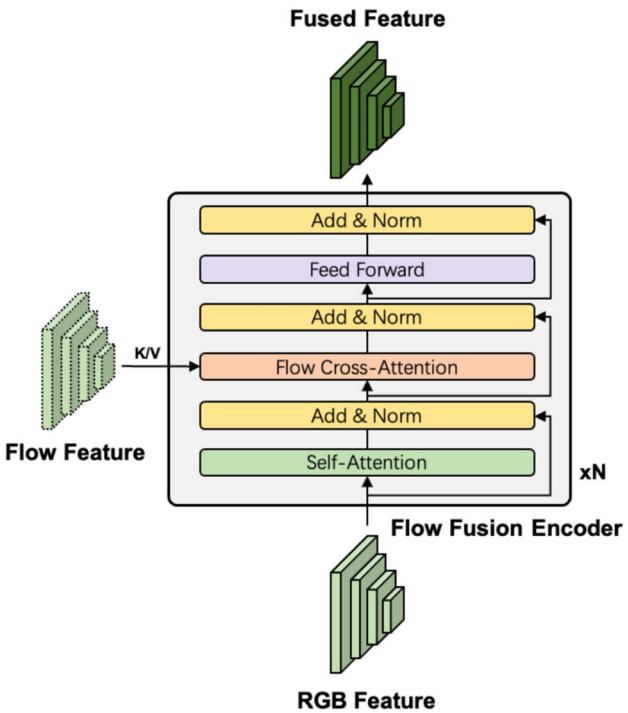

Fig. 4 Illustration of the proposed Flow Fusion Encoder module

Experimental Results

Datasets

This study evaluates the effectiveness of the proposed method using three publicly available video format polyp detection benchmarks, including CVC-VideoClinicDB [64, 65], SUN Colonoscopy Video Database [66], and LDPolyp-Video [67], which are widely adopted by the current mainstream polyp detection methods. For the fairness of the experiments, all methods in our experiments follow the same data division strategy. The detailed information of these datasets is summarized in Table 1.

CVC-VideoClinicDB CVC-VideoClinicDB [64, 65] comprises over 40 short and long video sequences extracted from routine colonoscopy examinations conducted at the Hospital Clinic of Barcelona, Spain. This database covers different scenarios that a computer-aided polyp detection system should face, and is adapted by the GIANA challenge of MICCAI. Notably, the polyp frames are labeled and reviewed by clinical experts. While only the training data consisting of 18 sequences is available with annotations, we have manually partitioned it into two sets: CVC-VideoClinicDB$_{train}$ (14 video sequences; 9470 images) and CVC-VideoClinicDB$_{test}$ (remaining 4 video sequences, numbered #2, 5, 10, and 18; 2484 images) following [53]

Table 1 Details of the three video format datasets used in this work

Dataset	Image size	Total	Training	Testing
CVC-VideoClinicDB	384x288	11954 (18 samples)	9470 (14 samples)	2484 (4 samples)
SUN Colonoscopy Video Database	1158x1008 to 1240x1080	49136 (100 samples)	38416 (79 samples)	10720 (21 samples)
LDPolypVideo	560x480	40186 (160 samples)	24789 (100 samples)	15397 (60 samples)

and [68]. Further details about the dataset can be found in the GIANA website [69].

SUN Colonoscopy Video Database SUN Colonoscopy Video Database [66] comprises the still images of videos, which are collected at the Showa University Northern Yokohama Hospital. Every frame in the database was annotated by the expert endoscopists at Showa University. The database includes 49136 polyp frames taken from 100 polyps with different shapes, sizes, locations and pathological diagnoses. These frames were fully annotated with bounding boxes. Non-polyp scenes of 109554 frames are also included in the database. More detailed statistics and information can be found from the SUN website [70] and its paper [66]. Only the polyp scenes of 100 samples are used in the performance evaluation. We split them into SUN Colonoscopy Video Database$_{test}$ set (totally 10720 frames; case of #1, 3, 7, 16, 19, 24, 26, 35, 36, 44, 47, 56, 59, 65, 68, 72, 79, 86, 89, 90, 96 in polyp samples) and SUN Colonoscopy Video Database$_{train}$ set (totally 38416 frames; the rest 79 cases in polyp samples).

LDPolypVideo LDPolypVideo dataset [67] was publicly released at MICCAI2021, and to the best of our knowledge, it represents the most extensive publicly available colonoscopy video database. Unlike other public colonoscopy datasets that are often highly curated and balanced, LDPolypVideo represents a more realistic clinical scenario. As depicted in [3, 4], LDPolypVideo contains challenging situations that are prone to algorithmic failures. These challenging cases encountered in real-world settings result in a significant drop in detection performance, posing problems that must be addressed in the development of colonoscopy CAD systems. It comprises 40,186 frames of 200 polyps extracted from 160 colonoscopy videos, each with bounding boxes for every polyp. Additionally, it contains 103 videos, comprising 861,400 frames, which have only been simply annotated with video-level annotations that indicate the presence of polyps. The dataset exhibits a broad range of polyp types, sizes, and morphologies, all captured within complex bowel environments, including motion blurs and specular reflections. We divide the labeled frames into LDPolypVideo$_{train}$ set and LDPolypVideo$_{test}$ set according to the original split in the paper. Finally, the LDPolypVideo$_{train}$ set and the LDPolypVideo$_{test}$ set contain a total of 100 videos (24,789 frames) and 60 videos (15,397 frames), respectively. More detailed information about the dataset can be found in the conference paper [67].

Evaluation Metrics

This paper employs the COCO-style mean Average Precision (mAP) as a key evaluation metric to assess the detection performance of the proposed TS-PDTR detector and its counterparts.

To be specific, the true positive (TP), false positive (FP), and false negative (FN) parameters are computed first, where TP stands for the detection results within polyp ground truth bounding boxes with Intersection over Union (IoU) $\geq threshold$, FP stands for the detection results outside polyp ground truth bounding boxes or IoU $< threshold$, and FN stands for no detection found for frame with polyps. Based on the parameters above, Average Precision (AP) value can be calculated for different recall values over 0 to 1 with the settings consistent to the COCO dataset [47]. Following that, mAP metric can be represented using the following formula:

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k \tag{11}$$

where AP_k represents the Average Precision of class k, while n is the number of classes.

In this study, we report the mAP$_{0.5:0.95}$ result, which is the average mAP at different IoU thresholds (ranging from 0.5 to 0.95, with an increment of 0.05). In the subsequent text, we refer to it simply as mAP. We also provide the mAP$_{0.5}$ and mAP$_{0.75}$ results utilizing IoU thresholds of 0.5 and 0.75, respectively.

Implementation Details

We implemented the TS-PDTR based on the *PyTorch* library and the *mmdetection* framework [71], using four NVIDIA TESLA A100 GPUs with 40GB memory for training and testing. For training specifics and hyper-parameters, compared to DINO [46], we use a 6-layer Flow Fusion Encoder (N in Section "Flow Fusion Encoder" equals 6) to replace its original 6-layer Transformer encoder. The AdamW

optimizer with weight decay of 1e-4 was used to train our model with batch-size 16. We set the initial learning rate (LR) as 1e-4 and adopt a simple LR scheduler, which drops LR at the 30-th epoch by multiplying 0.1 for the total 36 epoch training settings (denoted as 3× schedule). We use a ResNet-50 [72] and a SwinL [73] backbone for our main results and SOTA model, respectively. The same multi-scale setting as in DINO to use 4 scales in ResNet-50-based models and 5 scales in SwinL-based models is employed. The input RGB-flow pairs are resized to 480×480 (1.5× larger scale for TS-PDTR with SwinL, 720×720) during both the training and testing phases. For data augmentation, we simply utilize random horizontal flip and rotation with a probability of 0.5. Note that the RGB and flow input pairs should be applied with the same augmentation to maintain alignment. The stack number L of the optical flow input is set to 4 according to the ablation results in Section "Ablation Study". For optical flow generation, dense flow maps are computed using the PWC-Net [74] with default parameters between consecutive frames, and the mean flow subtraction is applied for global camera motion compensation, as described in Section "Two-Stream Feature Extraction Network". For module initialization, the backbone of the RGB branch is pre-trained on ImageNet to achieve better feature representation ability (ImageNet-1K for ResNet-50 and ImageNet-22K for SwinL, respectively). Different from it, the optical flow branch is trained from scratch with all convolutional layers initialized using a Kaiming normal distribution [75]. Likewise, the proposed DAConv, DGA, and FFE modules are all initialized with Kaiming's method. The experimental settings of other competitors follow the best practice given in their papers.

Comparisons with SOTA Methods

In this section, we compare the proposed TS-PDTR quantitatively and qualitatively with previous advanced networks, including image-based models Faster R-CNN [14], RetinaNet [76], SSD [19], FCOS [77], CenterNet [78], YOLOv5 [27], YOLOX [28], Deformable DETR [41], DINO [46] and video-based models FGFA [79], MEGA [80], TransVOD [81], STMN [82], Hybrid 2D/3D CNN [49], STFT [68], and YONA [83]. The related detection experiments are respectively conducted on three different datasets to comprehensively verify the effectiveness of the proposed network in polyp detection. Table 2 shows the final quantitative analysis results on three datasets, and Figs. 5 and 6 illustrate the final qualitative visualization results.

Quantitative Comparison

As shown in Table 2, we select mAP, $mAP_{0.5}$, and $mAP_{0.75}$ to quantitatively analyze the detection ability of each network. To ensure reliability, all results are reported based on the average of five independent training runs with different random seeds. Uncertainty is also provided for reference,

Table 2 Quantitative comparison with other SOTA Image-based and Video-based methods on three different colonoscopy video datasets. Results are reported as mean ± SD across five independent runs with different random seeds. The best results for each dataset are red indexed, and the second-best results are blue indexed.

	Methods	CVC-VideoClinicDB			SUN Colonoscopy Video Database			LDPolypVideo		
		mAP	$mAP_{0.5}$	$mAP_{0.75}$	mAP	$mAP_{0.5}$	$mAP_{0.75}$	mAP	$mAP_{0.5}$	$mAP_{0.75}$
Image-based	Faster R-CNN [14]	46.1±0.4	78.5±0.4	50.4±0.3	54.7±0.3	89.5±0.2	62.2±0.3	22.4±0.3	47.5±0.2	17.8±0.3
	RetinaNet [76]	47.5±0.7	80.0±0.6	54.4±0.6	55.9±0.4	88.8±0.2	62.1±0.3	21.5±0.5	43.6±0.4	17.5±0.3
	SSD [19]	44.8±0.9	79.7±0.7	47.7±0.9	53.4±0.6	88.0±0.5	61.2±0.5	25.1±0.7	53.2±0.5	20.6±0.6
	FCOS [77]	48.0±0.3	84.3±0.5	52.0±0.5	55.1±0.4	93.3±0.4	58.0±0.3	22.9±0.4	51.0±0.3	16.6±0.3
	CenterNet [78]	48.5±0.3	78.1±0.3	54.8±0.2	54.2±0.3	90.4±0.4	60.0±0.2	24.5±0.2	49.8±0.3	20.3±0.3
	YOLOv5 [27]	46.5±0.5	76.9±0.4	52.6±0.3	59.6±0.3	91.4±0.4	69.7±0.3	26.1±0.4	52.2±0.3	22.2±0.3
	YOLOX [28]	49.5±0.6	84.7±0.7	51.5±0.7	58.3±0.4	90.5±0.6	66.5±0.5	29.5±0.4	62.1±0.5	23.8±0.4
	Deformable DETR [41]	48.4±0.3	78.3±0.3	59.3±0.2	56.5±0.2	90.4±0.3	64.8±0.2	25.9±0.3	54.5±0.4	21.4±0.3
	DINO [46]	50.2±0.2	84.4±0.3	60.2±0.3	58.1±0.2	91.2±0.2	67.3±0.2	27.0±0.3	55.3±0.3	23.1±0.2
Video-based	FGFA [79]	50.5±0.2	85.3±0.2	54.6±0.2	57.2±0.1	90.6±0.2	65.3±0.1	28.7±0.2	60.5±0.3	23.1±0.2
	MEGA [80]	48.5±0.3	84.9±0.5	48.0±0.4	57.6±0.2	90.8±0.2	67.2±0.3	26.8±0.4	57.7±0.4	21.6±0.4
	TransVOD [81]	50.4±0.2	85.5±0.4	54.6±0.3	58.6±0.2	93.1±0.3	66.1±0.3	27.2±0.3	59.6±0.4	20.9±0.4
	STMN [82]	45.2±0.7	79.1±0.9	48.5±0.7	54.4±0.5	89.9±0.6	61.8±0.6	22.7±0.6	49.7±0.5	17.2±0.7
	Hybrid 2D/3D CNN [49]	49.3±0.4	82.6±0.5	52.5±0.4	58.1±0.3	90.7±0.3	68.0±0.2	27.5±0.4	58.8±0.5	22.4±0.5
	STFT [68]	51.9±0.2	86.9±0.3	56.7±0.3	59.3±0.2	92.2±0.2	68.0±0.1	28.2±0.3	58.1±0.4	23.4±0.3
	YONA [83]	53.9±0.4	87.3±0.3	61.1±0.4	60.5±0.2	91.6±0.3	69.9±0.3	30.5±0.3	62.6±0.4	25.3±0.4
Ours	TS-PDTR (R50)	53.7±0.2	87.2±0.2	58.9±0.1	61.7±0.1	92.9±0.1	71.5±0.1	31.6±0.2	64.2±0.3	26.5±0.2
	TS-PDTR (SwinL)	55.6±0.3	89.6±0.3	59.8±0.2	64.0±0.2	94.1±0.3	74.2±0.2	33.2±0.3	65.4±0.3	28.2±0.3

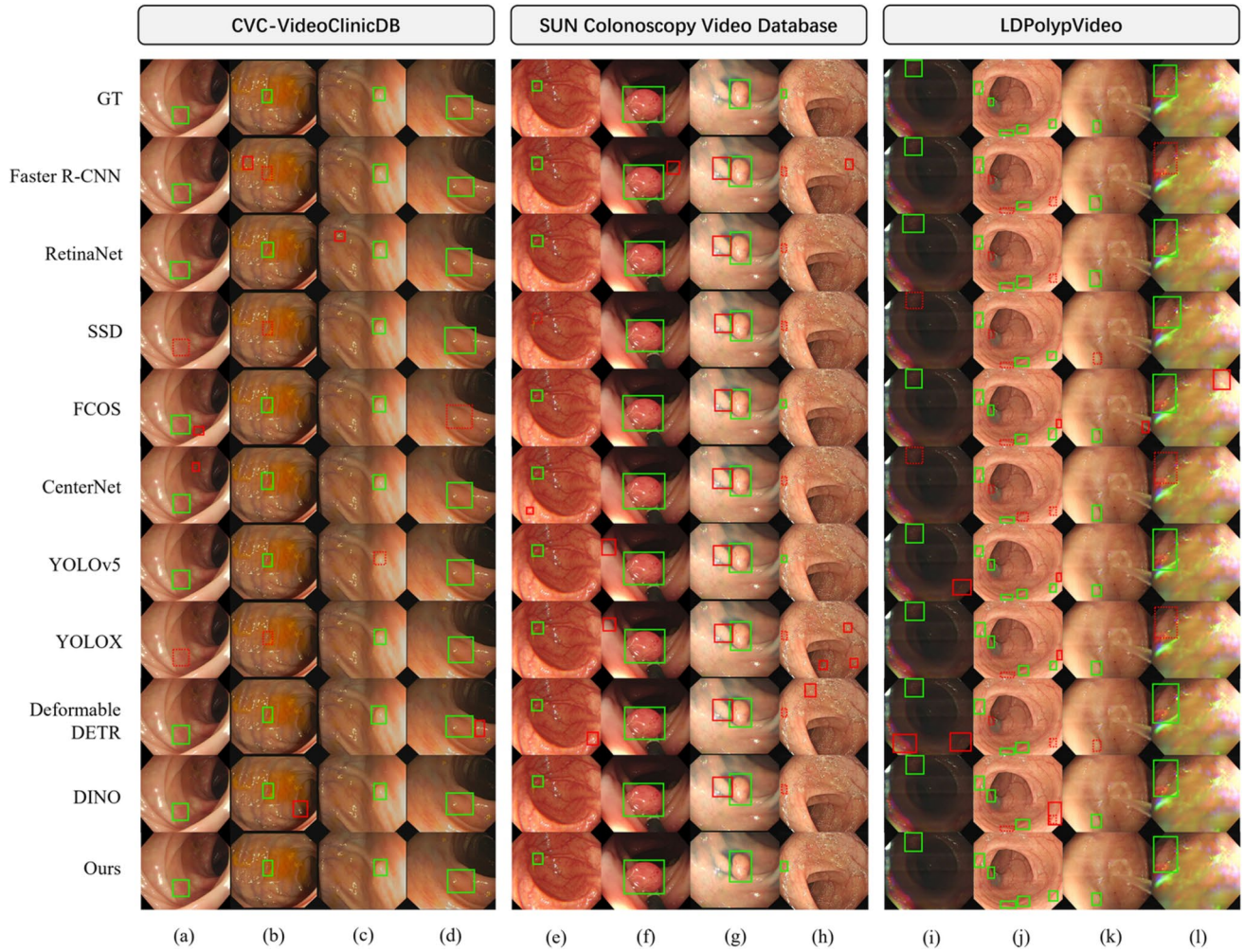

Fig. 5 Visualization of the qualitative comparison results with SOTA image-based polyp detection methods on some still frames from three different colonoscopy video datasets. The green solid line, red solid line, and red dotted line rectangular boxes denote the true positive, false positive, and false negative, respectively

presented in the form of mean ± standard deviation (SD). The red index is the optimal result of different networks on these datasets, and the blue index is the second-best result.

Firstly, compared with the DINO baseline, our TS-PDTR with three novel designs significantly improved the mAP score by 3.5, 3.6, and 4.6 on three benchmarks, demonstrating the effectiveness of the model design. Besides, we found that all of the best indicators and most of the second-best indicators are distributed in the proposed TS-PDTR network. Meanwhile, the exclusive third-best index of TS-PDTR appeared in the evaluation on the CVC-VideoClinicDB, which is a small-sized dataset with relatively simple and homogeneous cases, and the performance gap between our method and the suboptimal YONA model is small (only 0.2 mAP). Furthermore, we noticed that the overall performance of the detection network on the LDPolypVideo dataset is obviously lower than that on the other two datasets. As a large-scale polyp detection benchmark, the polyps sampled in the LDPolypVideo under different conditions reflect multiple challenges that clinicians should face in the colonoscopy practice. In this more realistic setting, we found that video-based SOTAs generally performed better than image-based ones, illustrating the importance of using temporal information when detecting polyps in continuous colonoscopy video streams. Among them, our proposed TS-PDTR network with Swin-L backbone achieves the highest value of 33.2 mAP, which has notable advantages compared with the second-best index except our method. The interval between mAP is as high as 2.7. However, STMN consistently performs poorly across all three video colonoscopy datasets, with results even lower than most image-based methods. This indicates that the Convolutional Long Short-Term Memory (ConvLSTM) architecture is not suitable for colonoscopy polyp detection because its reliance on long-term temporal consistency does not align with the characteristics of endoscopic videos, where polyps are mostly static

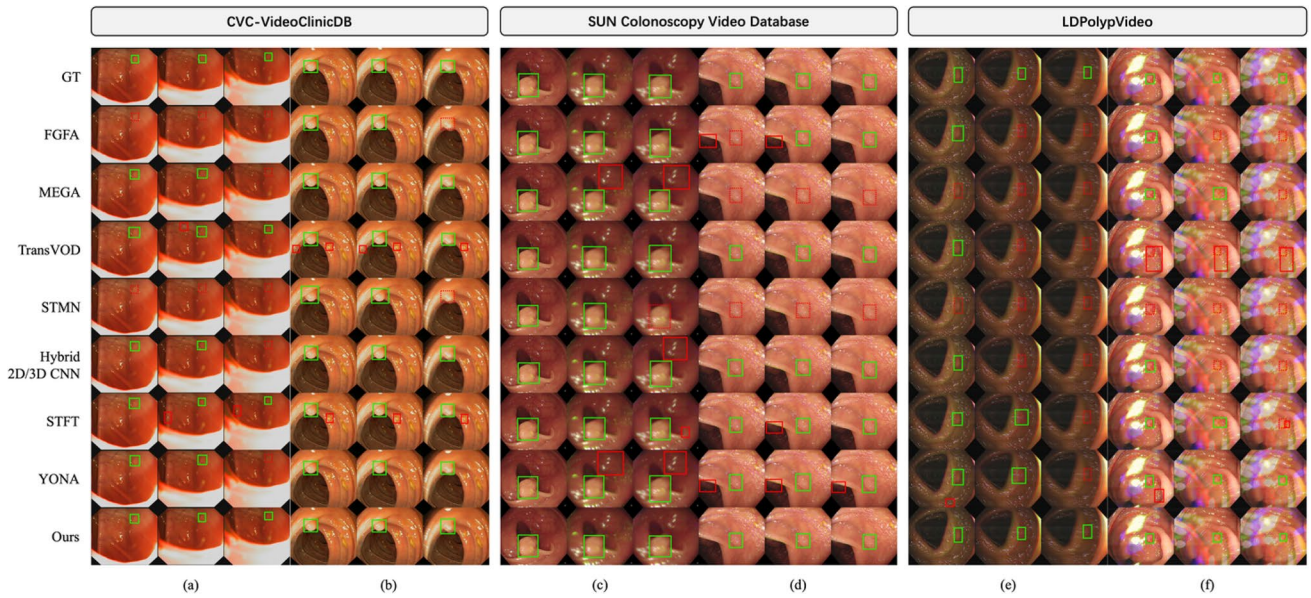

Fig. 6 Visualization of the qualitative comparison results with SOTA video-based polyp detection methods on some video clips from three different colonoscopy video datasets. The green solid line, red solid line, and red dotted line rectangular boxes denote the true positive, false positive, and false negative, respectively

while the camera motion is irregular and dominant. This mismatch causes the Long Short-Term Memory (LSTM) to capture redundant background motion rather than enhancing the fine spatial details needed to detect small or camouflaged polyps.

In summary, Table 2 illustrates that the proposed TS-PDTR in this paper has good quantitative results across these three datasets under different colonoscopy inspection conditions and environments, which indicates that the proposed TS-PDTR has strong model learning ability and robust video polyp detection effect.

Qualitative Comparison

For more intuitive observation of detection performance on the three datasets of various advanced networks, this section selects four still polyp frames and two video polyp clips from each of the three datasets for qualitative visualization experiments, as shown in Figs. 5 and 6 for image-based models and video-based models, respectively. The green solid line, red solid line, and red dotted line rectangular bounding boxes denote the true predictions as well as the ground truth labels, false detections, and missed detections, respectively.

The twelve sample still frames in the Fig. 5 reflect the different hard cases and complex situations in colonoscopy polyp detection. The target areas in column (a), (c), and (d) are flat and have high similarity with their surrounding intestinal wall so that some detectors are in failure. The polyps in column (b), (h), and (k-l) are so small or under the interference of floating excrement, bubbles, and water flow that it is difficult to locate and causes missed diagnosis. The column (e) shows a polyp in the inflammatory environment. The color of its background is relatively red compared to others. In the column (f), a big adenoma is being removed with electrosurgery. Methylthionine was injected around the polyp in column (g) to better determine its boundary. However, all of the models except our TS-PDTR misidentify the background folds as polyps, resulting in over-detection. In the column (i), insufficient brightness affects the judgment of the polyp. The multiple polyps in column (j) are small and have unobvious edges. It confused all other competitors, leading to either false detection or missed detection. Only TS-PDTR generates clear and accurate polyp bounding boxes with robust detection capability. In summary, our model achieves significant improvements in polyp detection performance over previous state-of-the-art image-based detectors, particularly in challenging cases that are prone to algorithm failure, primarily camouflaged polyps (columns a/c/d/j) and small polyps (columns b/e/h/j/k).

Figure 6 demonstrates six sample video clips whose current frames are of poor quality under different challenging scenarios. Specifically, they are: (a) concealed polyp with partial overexposure; (b) motion blur caused by fast moving speed; (c) optical defocus of endoscopy; (d) complex background with high reflection; (e) motion blur and insufficient brightness; (f) occlusion caused by dispersion. From the qualitative comparison results of our method with its counterparts in these representative video clips with data corruptions to some extent, we can find that our TS-PDTR is able

to capture useful information from supporting frames even under severe quality issues of the current input frame and locate polyps more precisely for various difficulties encountered in realistic colonoscopy.

Ablation Study

Model Components Ablation

In order to dig deeply into the role of each network module in the proposed TS-PDTR, a series of ablation experiments are conducted in this section. Metrics are evaluated on the polyp detection task with the most challenging LDPolypVideo dataset. The image-based DINO detector is used as the baseline network.

As shown in Table 3, the mAP score increases to 28.0 after adding our DGA module into the baseline for channel-specific spatial re-weighting. Moreover, attention mechanisms found within DGA are separately employed, namely the Squeeze-and-Excitation Module (SEM) [62] and the Convolutional Block Attention Module (CBAM) [61], to replace our DGA module for comparison. SEM is a kind of mainstream channel attention mechanism, while CBAM contains both channel attention and spatial attention with slightly different implementations. The results indicate that it is critical to pay attention to channel-wise polyp feature distribution difference.

In addition, adding the DAConv module greatly improves the mAP by 2.1 to 30.1. This achieves the largest performance gain among the three designed modules. To further validate the effectiveness of the proposed DAConv module in enhancing fine-grained polyp features, we extracted and visualized the feature maps of several representative small and camouflaged polyps before and after applying DAConv. These feature maps were first averaged and then transformed into the frequency domain by the Discrete Fourier Transform (DFT) to intuitively reveal the differences in spectral characteristics. Figure 7 demonstrates that the DAConv module makes up for the object features and suppresses the background features, which can reflect the modeling ability of the module to mine the shallow texture and structure features (see the green dotted circle regions in the third column, compared with the second column). Besides, as shown in the last two columns, the DFT visualizations illustrate that the frequency spectrum is dominated by low-frequency components before DAConv, indicating that fine details such as edges and textures are underrepresented and partly buried in the dominant background signals. After DAConv, the corresponding DFT results clearly show an increase in the energy of high-frequency components, which directly corresponds to sharper contours and richer texture information of the target polyp regions. Meanwhile, we observe that spurious high-frequency signals from the background are effectively suppressed, which reduces irrelevant noise that may otherwise hinder detection accuracy. This frequency-domain evidence confirms that DAConv simultaneously amplifies useful structural information and suppresses redundant background details, thereby improving the model's sensitivity to subtle boundaries and camouflaged polyps that are easily missed. This demonstrates the unique advantage of DAConv in complementing vanilla convolution by explicitly boosting the representation of high-frequency cues essential for robust detection in complex colonoscopy scenes.

The final row of DGA+DAConv+FFE is equivalent to the full version of our proposed TS-PDTR, which further adds the FFE module. It obtains an mAP of 31.6, 1.5 higher than that of the DGA+DAConv variants. This proves that our FFE module plays a key role in effectively unearthing useful feature representations in the continuous neighborhood from the temporal branch. Overall, all the modules are necessary for precise detection compared with the baseline results. By combining the three network modules, our TS-PDTR model can achieve its best polyp detection performance.

Study on Optical Flow Representations

To provide the most meaningful temporal information for polyp detection in colonoscopy video sequences, we conducted a thorough analysis of optical flow representation on the highly challenging LDPolypVideo dataset. First, we evaluated the effect of different single-frame optical flow generation methods [74, 84–86] on polyp detection performance, as shown in Table 4. The results indicate that the quality of optical flow estimation has a substantial effect on the detector's capability. For example, using the RAFT [86] method for optical flow estimation to assist polyp

Table 3 Ablation experiments of the proposed modules

DGA	SEM	CBAM	DAConv	FFE	mAP	$mAP_{0.5}$	$mAP_{0.75}$
Baseline					27.0	55.3	23.1
✓					28.0	56.3	22.7
	✓				27.3	55.6	21.9
		✓			27.5	58.6	19.0
✓			✓		30.1	60.5	24.2
✓			✓	✓	31.6	64.2	26.5

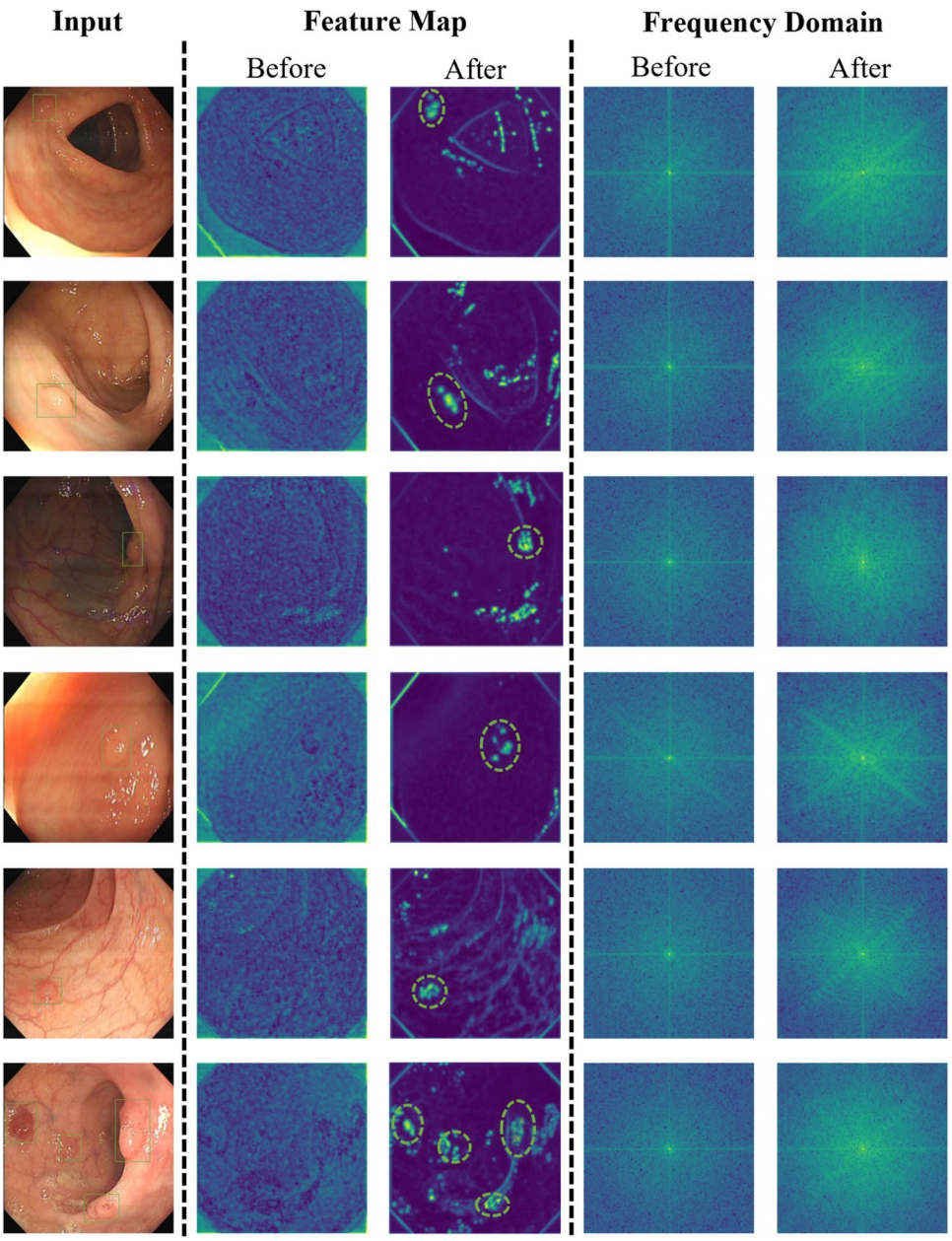

Fig. 7 Visualization of representative feature maps of small and camouflaged polyps before and after applying the DAConv module, along with their corresponding DFT spectra. The DFT images (last two columns) demonstrate that DAConv effectively enhances useful high-frequency signals (highlighting edges and textures) while suppressing background noise, which improves the model's ability to detect subtle and camouflaged lesions

Table 4 Ablation results of using different single-frame optical flow generation methods in the temporal branch

Flow Estimation Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$
FlowNet [84]	30.9	62.7	25.3
LiteFlowNet [85]	30.9	63.0	24.9
PWC-Net (ours) [74]	31.6	64.2	26.5
PWC-Net* [74]	30.5	61.9	24.1
RAFT [86]	29.3	59.1	23.6

* denotes the flow estimation method without global camera motion compensation

Table 5 Ablation results of using different multi-frame optical flow aggregation strategies in the temporal branch

Flow Aggregation Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$
Weighted Addition [87]	30.4	61.5	24.5
Stacked (ours)	31.6	64.2	26.5

localization resulted in even lower performance than the baseline with merely RGB stream (DINO+DGA+DAConv variants in Table 3) by 0.8 mAP. Inaccurate optical flow introduces erroneous temporal information, which interferes with the fused features. Using PWC-Net [74] to estimate the optical flow field achieved the best result of 31.6 mAP. In addition, omitting the mean flow subtraction operation also led to a notable drop in detection result by 1.1 mAP, highlighting the importance of endoscope camera motion compensation.

Furthermore, we assess the influence of multi-frame optical flow aggregation strategies on our method. Table 5 illustrates that our approach significantly outperforms the mainstream weighted-sum aggregation method widely used in natural image processing, achieving an mAP improvement of 1.2. Unlike images of natural scenes, which typically contain salient objects with rich textures and structures, colonoscopy images lack such distinct features. As a result, the shallow layers of the temporal network in our TS-PDTR framework need to track pixel-level motion frame by frame to effectively distinguish between foreground and background. Therefore, feeding multiple frames of raw optical flow vectors as stacked inputs provides richer and more continuous motion information. In contrast, the weighted-addition paradigm, which compresses the flow into only two channels, tends to emphasize the global motion between the first and last frames of the video sequence, rather than capturing detailed local temporal dynamics.

We also investigated the impact of different stacked flow numbers on TS-PDTR in Table 6. Under the detection mAP metric on the LDPolypVideo dataset, using $L = 4$ supporting optical flow frames achieves the best accuracy. We visualize the trends for clearer observation. In Fig. 8, performance of TS-PDTR improves first as more reference frames are utilized and then declines. This is consistent with our conjecture. On the one hand, the fast camera-moving nature of colonoscopy video will introduce large variance in the foreground features of target polyps. Therefore, collaborating too many optical flow frames will increase the misalignment between adjacent frames and lead to poor detection performance. On the other hand, insufficient stacked flow frames also make it difficult to extract the necessary representations to assist detection due to the possible existence of consecutive frames with poor quality. Those abnormalities disrupt the integrity of background feature structures and thus affect the effectiveness of temporal information aggregation.

Table 6 Ablation results of using different stacked flow numbers in the temporal branch

Stacked flow number L	mAP	$mAP_{0.5}$	$mAP_{0.75}$
1	30.4	62.2	25.2
2	30.9	63.0	25.7
3	31.3	63.9	26.2
4	31.6	64.2	26.5
5	31.5	64.1	26.3
6	31.5	64.2	26.1
7	31.1	62.9	25.9
8	30.6	61.9	25.6

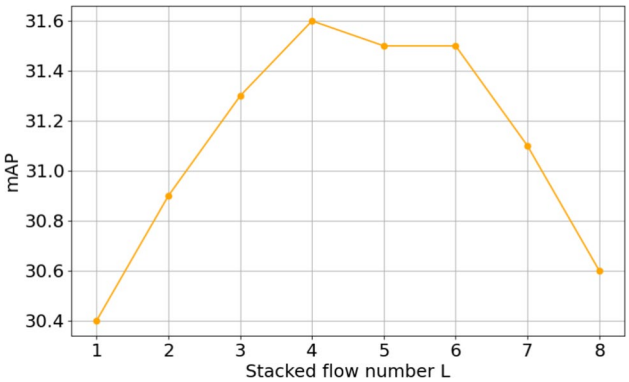

Fig. 8 The performance of TS-PDTR on the LDPolypVideo when using different stacked flow number for reference

Table 7 Ablation results of using different feature fusion methods

Fusion Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$
Addition	29.2	59.1	24.9
Concatenation	30.3	60.5	25.7
Ours	31.6	64.2	26.5

Impact of Temporal-Spatial Feature Fusion Methods

We conducted a comparative analysis of our proposed two stream feature fusion method FFE with other temporal-spatial fusion methods to evaluate its utility. In the manner of feature fusion, element-wise addition and channel-wise concatenation operations are the most fundamental ways, which are adopted by many previous studies in various domains. The experimental results are presented in Table 7. Our FFE achieves the best performance metric of 31.6, 2.4 and 1.3 higher than that of addition and concatenation, respectively. Combined with the visualization results in the last three columns of Fig. 9, the fused feature heatmaps obtained through FFE exhibit a stronger response to the target polyps compared to other competing fusion methods. The simple addition and concatenation approaches fail to address the mismatch issue before fusion, leading to the activation of unrelated background regions.

For better understanding the learned multi-scale deformable flow cross-attention modules, we visualize sampling points and attention weights of the last layer in FFE on the optical flow input, as shown in the Cross Attention column in Fig. 9. For readability, we combine the sampling points and attention weights from feature maps of different resolutions into one picture. We noticed that most of the sampling points with high attention weights are already distributed around the target polyps. The visualization also demonstrates that the proposed multi-scale deformable flow cross-attention module can adapt its sampling points and attention weights according to different scales and shapes of the polyp objects. Our FFE module effectively leverages the complementary temporal information from stacked optical flow data to detect hard-case polyps even in scenarios where current frames with poor quality have sparse useful features.

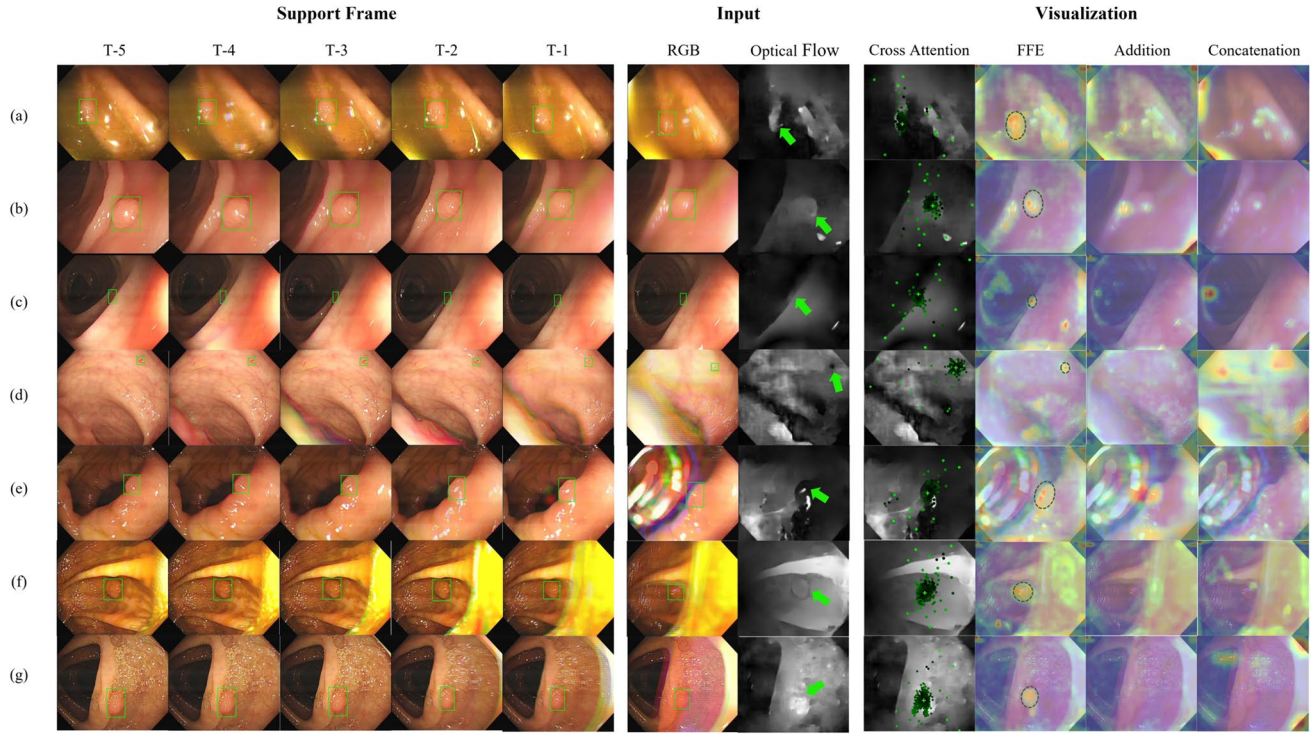

Fig. 9 Visualization of the two stream feature fusion procedure. (a) to (g) are different challenge polyp cases in frames with poor quality (see **RGB** column, the annotated bounding boxes are shown as green rectangles). The **Optical Flow** column illustrates the stacked optical flow results from support frames before current time T. For readability, we calculate their root mean square values along the stacked channel dimension. The positions pointed by the green arrows show obvious differences that can be directly distinguished by the human eye. The **Cross Attention** column visualize the sampling points and multi-scale deformable cross attention weights from feature maps of different resolutions in one picture. Each sampling point is marked as a filled circle whose color depth indicates its corresponding attention weight. The reference point is shown as green plus marker, which is also equivalent to query point in encoder. Last three columns show the fused feature heatmaps of FFE, Addition, and Concatenation methods

Robustness Evaluation

Data Corruption

To simulate visual degradations encountered in real-world colonoscopy, we evaluate the TS-PDTR model on corrupted versions of the LDPolypVideo$_{test}$ set, while keeping the LDPolypVideo$_{train}$ set unchanged. Four common types of image corruption in colonoscopy are considered: Gaussian noise, motion blur, overexposure, and underexposure. Each corruption type is applied at five severity levels using handcrafted algorithms following the protocol in [88], as illustrated in Fig. 10. The results in Table 8 show that the proposed method exhibits strong robustness against mild to moderate corruption (Severity 1–3), with less than a 3% drop in mAP metric. However, performance degradation becomes more pronounced under severe corruption (Severity 5), particularly in the cases of gaussian noise and motion blur. Although such levels of degradation are rare in clinical practice, our experiments help establish the safety margin for deploying the proposed model.

Distribution Shift

Although our TS-PDTR method achieves advanced performance in the video polyp detection task, one critical limitation lies in the assumption that the training and test datasets share an identical distribution. In practice, however, colonoscopy images are typically collected by hospitals using different endoscopic devices and clinical protocols across diverse patient populations, leading to variations in appearance distributions (termed as domain shifts).

To evaluate the model's generalization ability across domains, we conduct cross-dataset experiments using LDPolypVideo, the SUN Colonoscopy Video Database, and CVC-VideoClinicDB. In each experiment, the training subset of one dataset is used for model training, while the test subsets of the other two datasets are used for evaluation. As shown in Table 9, compared with the reference results from the training subset of the same domain, our model achieves consistent or even superior detection performance across most scenarios, demonstrating its robustness to domain shifts in clinical environments to some extent. However, when

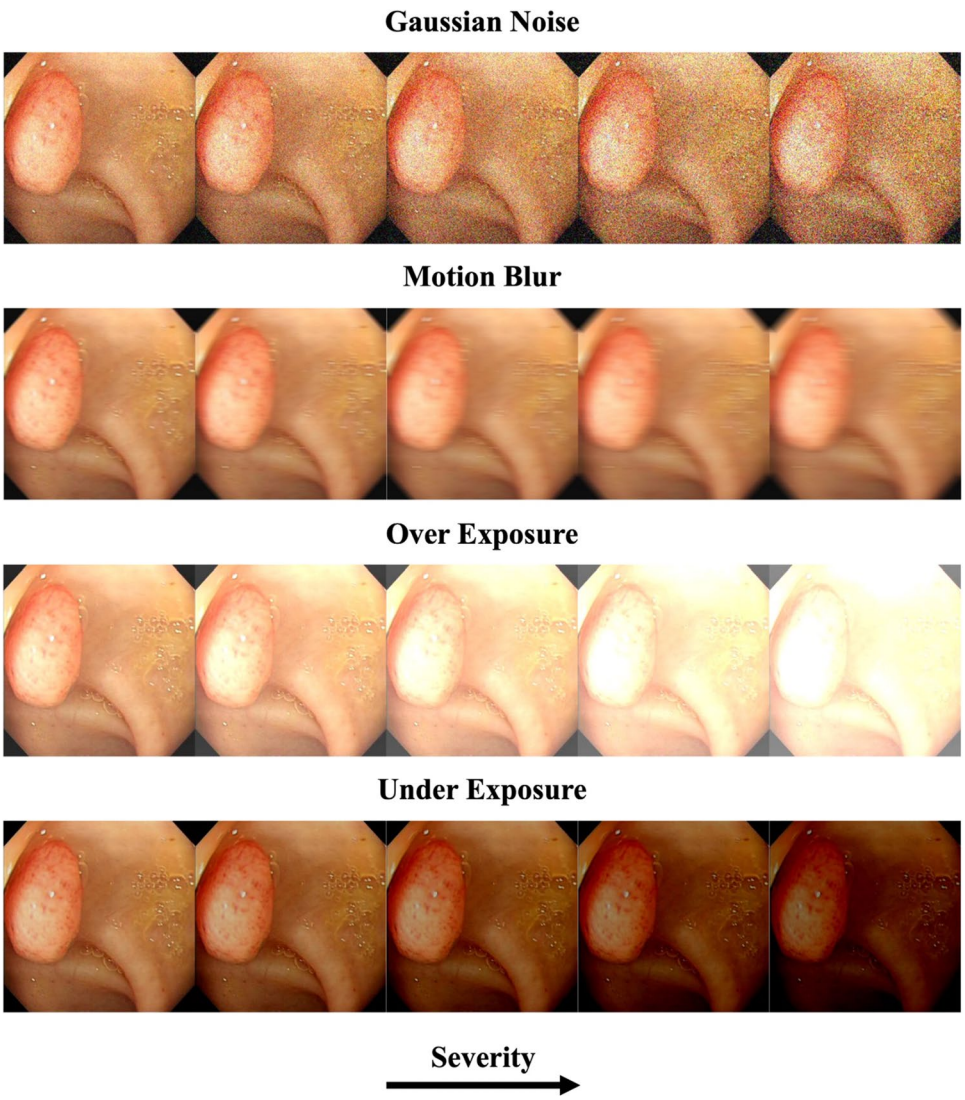

Fig. 10 Examples of four common types of image corruption in colonoscopy, generated using handcrafted algorithms: Gaussian noise, motion blur, overexposure, and underexposure. Each type of corruption has five levels of severity

Table 8 Robustness evaluation results under common types of data corruption at varying severity levels. 'Clean' refers to the original images without degradation while higher numbers following 'Severity' indicate increasing levels of corruption. The numbers in parentheses denote the performance change relative to the 'Clean' images

Corruption Type	mAP					
	Clean	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5
Gaussian Noise	31.6	31.5 (-0.1)	31.2 (-0.4)	30.8 (-0.8)	29.4 (-2.2)	27.5 (-4.1)
Motion Blur		31.4 (-0.2)	31.0 (-0.6)	30.7 (-0.9)	28.9 (-2.7)	27.1 (-4.5)
Overexposure		31.5 (-0.1)	31.3 (-0.3)	31.0 (-0.6)	30.2 (-1.4)	29.5 (-2.1)
Underexposure		31.6 (0.0)	31.3 (-0.3)	31.1 (-0.5)	30.5 (-1.1)	29.9 (-1.7)

Table 9 Robustness evaluation results on distributionally shifted data. Test performances from the same training dataset are underlined as reference baselines, and the values in parentheses indicate the metric changes relative to their respective baselines

Source Domain	mAP on Target Domain		
	CVC-VideoClinicDB$_{test}$	SUN Colonoscopy Video Database$_{test}$	LDPolypVideo$_{test}$
CVC-VideoClinicDB$_{train}$	<u>53.7</u>	60.5 (-1.2)	25.5 (-6.1)
SUN Colonoscopy Video Database$_{train}$	51.8 (-1.9)	<u>61.7</u>	27.2 (-4.4)
LDPolypVideo$_{train}$	52.4 (-1.3)	62.5 (+0.8)	<u>31.6</u>

evaluated on LDPolypVideo$_{test}$, the model's performance drops significantly, by 4.4 mAP when trained on SUN Colonoscopy Video Database$_{train}$, and by 6.1 mAP when trained on CVC-VideoClinicDB$_{train}$. This degradation is likely caused by a substantial domain shift between the test and training sets, leading to relative overfitting to the training data distribution. This outcome reveals a failure case of our model and underscores the need for further research to improve its domain generalization capability for polyp detection.

Computational Cost

For practical implementation, reasonable computational overhead is essential when introducing new modules. In Table 10, we report the per-GPU memory usage and total training time of our two-branch method, TS-PDTR. The evaluation is conducted on the representative LDPolypVideo dataset with a 3× training schedule. Other detailed settings are consistent with those described in Section "Implementation Details". The 'enhanced R50' configuration for the spatial branch refers to a vanilla ResNet-50 backbone [72] augmented with our proposed DAConv and DGA modules, designed to better handle small and camouflaged polyps. Table 10 illustrates that, compared to the original residual RGB image baseline, the augmented version incurs only a minor increase in training time (less than 2 hours) and GPU memory usage (approximately 4%). Moreover, due to the simple design of the optical flow encoder and the efficiency of the deformable cross-attention mechanism, the two-stream variant with the additional temporal branch remains computationally affordable, requiring 26.5 GB of per-GPU memory and 21.8 hours of training time, while obtaining a significant improvement of 4.6 mAP in polyp detection performance (as shown in Table 3).

Discussion

Accurate detection of polyps during colonoscopy is of great importance for the diagnosis and treatment of colorectal cancer. This paper takes three core challenges encountered in real-world video colonoscopy as the entry point and proposes a novel video polyp detection method, called TS-PDTR, which introduces fine-grained texture information as well as temporal representation information to solve the issue of missed detection and false detection. Specifically, two backbone encoders are used to extract spatial features and temporal features from the RGB stream and optical flow stream, respectively. The encoders build multi-level feature maps of different resolutions. Then, a DAConv module is proposed to enhance the detailed texture and structure information from the low-level feature map. Further, a DGA module is utilized to generate the channel-specific SAMs from this reference shallow feature map and assign them to deep feature maps so that the model is sensitive to small and camouflaged polyps. To further improve the robustness of the detector to poor imaging quality in the single frame, a FFE module is proposed to fuse the temporal embeddings from optical flow feature maps as compensation. To evaluate the performance of our proposed TS-PDTR, we conducted comprehensive quantitative and qualitative experiments on three video colonoscopy polyp detection benchmarks. The experimental results demonstrate that TS-PDTR has a more competitive detection capability compared to the most advanced polyp detection algorithms. The best detection results were obtained by our methods on all three datasets, among which the results of mAP reach 55.6, 64.0, and 33.2 on CVC-VideoClinicDB, SUN Colonoscopy Video Database, and LDPolypVideo, respectively.

However, TS-PDTR still has limitations. First, although the network can better solve the problems suffered in video colonoscopy polyp detection, it inevitably increases the complexity of the model when introducing the proposed spatial-feature augmentation modules and the temporal feature extraction and fusion branch to improve the detection ability. Although the experimental results in Table 10 indicate that the increased computational overhead during training is affordable, the model architecture must be further streamlined as the scale of video colonoscopy datasets continues to grow exponentially. The memory bank and memory attention mechanism proposed by SAM2 [89] represent a promising direction to explore in the future, as they can effectively leverage temporal information in video sequences without requiring an additional temporal branch. Simultaneously, model light-weighting and inference acceleration strategies must be considered in the future to satisfy the efficiency requirements of clinical colonoscopy, including real-time capability and low latency, in the inference stage.

Second, as discussed in Section "Robustness Evaluation", although our TS-PDTR method is robust to mild to moderate data corruption, occasional severe image degradation in clinical practice can still lead to algorithm failure. In practice, clinicians should be particularly cautious in situations that prone to mistakes to ensure accurate detection outcomes. To assist clinicians in understanding hard cases, we classify common errors into three categories in Table 11 to

Table 10 Training time and GPU memory cost in four A100 40GB GPUs, implemented in PyTorch

Spatial branch	Temporal branch	Training Time	Memory/ per GPU
vanilla R50	w/o	10.1 hrs	14.3 GB
enhanced R50	w/o	12.0 hrs	14.9 GB
enhanced R50	w/	21.8 hrs	26.5 GB

Table 11 Error taxonomy of potential algorithm failure scenarios during clinical colonoscopy. Each category includes a description of the issues that may affect detection accuracy, along with corresponding recommendations for clinicians

Error Type	Description	Clinical Recommendations
Hard-to-detect lesions	Small, flat, and camouflaged polyps with high similarity to surrounding intestinal tissue or without obvious boundary, making them difficult to detect.	Clinicians should exercise caution when dealing with low-contrast or poorly defined lesions, especially small or flat polyps. Additional diagnostic methods, such as image zoom or manual inspection, may be helpful.
Severe background interference	Poor bowel preparation may lead to fecal matter, fluid, or bubbles obstructing the view, as well as inflammatory bowel walls or excessive bowel folds caused by insufficient insufflation.	Ensure adequate bowel preparation to minimize interference. In cases with poor image view, it is recommended to pause and clean the area to ensure clear visualization.
Poor colonoscopy imaging quality and operation issues	Issues such as underexposure/overexposure leading to images being too dark or too bright, rapid camera movement causing blur and motion artifacts, optical defocus resulting in blurry images, partial high reflections, color dispersion, and obstruction caused by water flow for cleaning or surgical instruments.	Minimize rapid camera moving, and ensure optimal exposure settings to avoid overexposure or underexposure. In case of focus issues, pausing the procedure to readjust the camera or focus may improve detection accuracy. Avoid high reflection areas and obstructions caused by endoscopy tools. Re-examine according to standard operation procedures after the colonoscopy is adjusted to a normal imaging quality and stable status.

draw particular attention during routine colonoscopy. Moreover, the current model exhibits notable failure margins in terms of domain generalization capability. Therefore, further improvements are necessary to enhance its polyp detection performance under significant distribution shifts before deployment in real-world endoscopic applications. Given that our network adopts the DINO as the basis, this DETR-like framework inherits the common characteristic of the Transformer-based models, which feature impressive scalability with increasing training data size, similar to Large Language Models (LLMs). Exploring advanced large-scale unsupervised pre-training strategies or parameter-efficient fine-tuning (PEFT) techniques based on powerful pre-trained Vision Foundation Models (VFMs) presents promising future directions for the research community.

Third, as TS-PDTR serves as the core algorithm of the colonoscopy CADe system, further software optimization is required to facilitate its adaptation to clinical deployment in a live system. A key aspect is the incorporation of inference-stage post-processing functions, including threshold adjustment and temperature scaling, which allow endoscopists to tailor the system according to their preferences and tolerance levels for polyp detection. By adjusting the output confidence threshold, clinicians can strike a balance between precision and recall, ensuring that critical polyps are not missed while minimizing false positives. Precision-recall curves further assist in identifying optimal thresholds under different operational requirements. Temperature scaling provides an additional calibration mechanism by smoothing the model's confidence outputs, aligning them more closely with true probabilities. This mitigates the risk of overconfident predictions in edge cases, thereby enhancing the reliability and clinical credibility of the TS-PDTR model. Beyond thresholding and calibration, practical deployment also requires robust alert management strategies. Alert persistence and de-duplication mechanisms are vital to prevent redundant notifications that may distract clinicians. To address persistence, we suggest a time-window mechanism whereby alerts are issued only upon the first detection of a polyp across consecutive frames, with the alert duration scaled to the model's confidence. This reduces repetitive notifications while maintaining timely feedback. To further handle duplication, IoU-based strategies are useful, in which bounding boxes with overlaps above a set threshold are merged, ensuring that a single polyp does not generate multiple alerts despite variations in camera angle or motion. Together, these software-level designs not only improve the efficiency and robustness of TS-PDTR in real-world scenarios but also ensure adaptability to diverse clinical environments, enabling the CADe system to deliver stable, precise, and clinically meaningful assistance in dynamic colonoscopy settings.

Conclusion

In conclusion, this paper introduces TS-PDTR, an advanced video polyp detection method integrating tailored DAConv, DGA, and FFE modules to address the challenges of polyp detection in real-world colonoscopy. The proposed framework achieves state-of-the-art performance across diverse experimental settings, demonstrating its potential for advancing computer-aided diagnostic systems and facilitating practical deployment in clinical applications.

Author Contributions Tianyuan Gan: Conceptualization of this study, Investigation, Methodology, Software, Formal analysis, Writing original draft, Project administration. Chongan Zhang: Data curation, Investigation, Validation, Visualization, Writing original draft. Peng Wang: Investigation, Writing original draft. Xiao Liang: Supervision, Review &Editing. Xuesong Ye: Methodology, Supervision, Review & Editing, Resources.

Funding This research was supported by the Key Research and Development Plan of Zhejiang Province (Grant Nos. 2022C03086), and the National Key Research and Development Project (Grant Nos. 2022YFB4700803).

Data Availability The public video colonoscopy datasets used in this study are available on their original papers and websites.

Code availability The source code for the proposed TS-PDTR framework is publicly available on the GitHub website: https://github.com/GTYuantt/TS-PDTR.git

Declarations

Competing interests The authors declare no competing interests.

Ethical approval The research conducted for this paper adheres to ethical principles and guidelines concerning the utilization of publicly available datasets. The datasets employed in this study, CVC-VideoClinicDB, SUN Colonoscopy Video Database, and LDPolypVideo are publicly accessible resources without individual identifiers, thus obviating the need for specific consent from individuals.

Clinical trial number Not applicable.

References

1. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, Cercek A, Smith RA, Jemal A (2020) Colorectal cancer statistics, 2020. CA: a cancer journal for clinicians 70(3):145–164
2. Anderson JC, Butterly LF (2015) Colonoscopy: quality indicators. Clinical and translational gastroenterology 6(2):e77
3. Nogueira-Rodríguez A, Reboiro-Jato M, Glez-Peña D, López-Fernández H (2022) Performance of convolutional neural networks for polyp localization on public colonoscopy image datasets. Diagnostics 12(4):898
4. Gan T, Jin Z, Yu L, Liang X, Zhang H, Ye X (2023) Self-supervised representation learning using feature pyramid siamese networks for colorectal polyp detection. Scientific Reports 13(1):21655
5. Karkanis SA, Iakovidis DK, Maroulis DE, Karras DA, Tzivras M (2003) Computer-aided tumor detection in endoscopic video using color wavelet features. IEEE transactions on information technology in biomedicine 7(3):141–152
6. Alexandre LA, Nobre N, Casteleiro J (2008) Color and position versus texture features for endoscopic polyp detection. In: 2008 International Conference on BioMedical Engineering and Informatics, IEEE, pp 38–42
7. Ameling S, Wirth S, Paulus D, Lacey G, Vilarino F (2009) Texture-based polyp detection in colonoscopy. In: Bildverarbeitung für die Medizin 2009: Algorithmen—Systeme—Anwendungen Proceedings des Workshops vom 22. bis 25. März 2009 in Heidelberg, Springer, pp 346–350
8. Tajbakhsh N, Gurudu SR, Liang J (2013) A classification-enhanced vote accumulation scheme for detecting colonic polyps. In: Abdominal Imaging. Computation and Clinical Applications: 5th International Workshop, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013. Proceedings 5, Springer, pp 53–62
9. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. nature 521(7553):436–444
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
11. Tashk A, Nadimi E (2020) An innovative polyp detection method from colon capsule endoscopy images based on a novel combination of rcnn and drlse. In: 2020 IEEE Congress on Evolutionary Computation (CEC), IEEE, pp 1–6
12. Qadir HA, Shin Y, Solhusvik J, Bergsland J, Aabakken L, Balasingham I (2019) Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better? In: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), IEEE, pp 1–6
13. Chen BL, Wan JJ, Chen TY, Yu YT, Ji M (2021) A self-attention based faster r-cnn for polyp detection from colonoscopy images. Biomedical Signal Processing and Control 70:103019
14. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28
15. Qian Z, Lv Y, Lv D, Gu H, Wang K, Zhang W, Gupta MM (2020) A new approach to polyp detection by pre-processing of images and enhanced faster r-cnn. IEEE Sensors Journal 21(10):11374–11381
16. Nadimi ES, Buijs MM, Herp J, Kroijer R, Kobaek-Larsen M, Nielsen E, Pedersen CD, Blanes-Vidal V, Baatrup G (2020) Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy. Computers & Electrical Engineering 81:106531
17. Li J, Zhang J, Chang D, Hu Y (2019) Computer-assisted detection of colonic polyps using improved faster r-cnn. Chinese Journal of Electronics 28(4):718–724
18. Krenzer A, Hekalo A, Puppe F (2020) Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks. In: EndoCV@ ISBI, pp 58–63
19. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, pp 21–37
20. Liu M, Jiang J, Wang Z (2019) Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. IEEE Access 7:75058–75066
21. Tanwar S, Vijayalakshmi S, Sabharwal M, Kaur M, AlZubi AA, Lee HN (2022) Detection and classification of colorectal polyp using deep learning. BioMed Research International 2022
22. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
23. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
24. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767
25. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934
26. Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pp 13029–13038
27. Jocher G (2020) YOLOv5 by Ultralytics. https://doi.org/10.5281/zenodo.3908559 , https://github.com/ultralytics/yolov5
28. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) YOLOX: Exceeding yolo series in 2021. arXiv:2107.08430
29. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, Ke Z, Li Q, Cheng M, Nie W, Li Y, Zhang B, Liang Y, Zhou L, Xu X, Chu X, Wei X, Wei X (2022) Yolov6: A single-stage object detection framework for industrial applications. arXiv:2209.02976

30. Wang CY, Bochkovskiy A, Liao HYM (2022) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696
31. Jocher G, Chaurasia A, Qiu J (2023) YOLO by Ultralytics. https://github.com/ultralytics/ultralytics
32. Eixelberger T, Wolkenstein G, Hackner R, Bruns V, Mühldorfer S, Geissler U, Belle S, Wittenberg T (2022) Yolo networks for polyp detection: A human-in-the-loop training approach. In: Current directions in biomedical engineering, De Gruyter, pp 277–280
33. Doniyorjon M, Madinakhon R, Shakhnoza M, Cho YI (2022) An improved method of polyp detection using custom yolov4-tiny. Applied Sciences 12(21):10856
34. Karaman A, Pacal I, Basturk A, Akay B, Nalbantoglu U, Coskun S, Sahin O, Karaboga D (2023) Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc). Expert Systems with Applications 221:119741
35. Reddy JSC, Venkatesh C, Sinha S, Mazumdar S (2022) Real time automatic polyp detection in white light endoscopy videos using a combination of yolo and deepsort. In: 2022 1st International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS), IEEE, pp 104–106
36. Ou S, Gao Y, Zhang Z, Shi C (2021) Polyp-yolov5-tiny: A lightweight model for real-time polyp detection. In: 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), IEEE, pp 1106–1111
37. Lalinia M, Sahafi A (2023) Colorectal polyp detection in colonoscopy images using yolo-v8 network. Signal, Image and Video Processing pp 1–12
38. Krenzer A, Banck M, Makowski K, Hekalo A, Fitting D, Troya J, Sudarevic B, Zoller WG, Hann A, Puppe F (2023) A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks. Journal of imaging 9(2):26
39. Wan J, Chen B, Yu Y (2021) Polyp detection from colorectum images by using attentive yolov5. Diagnostics 11(12):2264
40. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, Springer, pp 213–229
41. Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159
42. Meng D, Chen X, Fan Z, Zeng G, Li H, Yuan Y, Sun L, Wang J (2021) Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 3651–3660
43. Yao Z, Ai J, Li B, Zhang C (2021) Efficient detr: improving end-to-end object detector with dense prior. arXiv preprint arXiv:2104.01318
44. Liu S, Li F, Zhang H, Yang X, Qi X, Su H, Zhu J, Zhang L (2022) Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv:2201.12329
45. Li F, Zhang H, Liu S, Guo J, Ni LM, Zhang L (2022) Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13619–13627
46. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, Ni LM, Shum HY (2022) Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv:2203.03605
47. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
48. Yu L, Chen H, Dou Q, Qin J, Heng PA (2016) Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE journal of biomedical and health informatics 21(1):65–75
49. Puyal JGB, Brandao P, Ahmad OF, Bhatia KK, Toth D, Kader R, Lovat L, Mountney P, Stoyanov D (2022) Polyp detection on video colonoscopy using a hybrid 2d/3d cnn. Medical Image Analysis 82:102625
50. Zhang R, Zheng Y, Poon CC, Shen D, Lau JY (2018) Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern recognition 83:209–219
51. Yu T, Lin N, Zhang X, Pan Y, Hu H, Zheng W, Liu J, Hu W, Duan H, Si J (2022) An end-to-end tracking method for polyp detectors in colonoscopy videos. Artificial Intelligence in Medicine 131:102363
52. Zhang P, Sun X, Wang D, Wang X, Cao Y, Liu B (2019) An efficient spatial-temporal polyp detection framework for colonoscopy video. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, pp 1252–1259
53. Zheng H, Chen H, Huang J, Li X, Han X, Yao J (2019) Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained cnn. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, pp 79–82
54. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25
55. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems 27
56. Agrahari H, Iwahori Y, Bhuyan MK, Ghorai S, Kohli H, Woodham RJ, Kasugai K (2014) Automatic polyp detection using dsc edge detector and hog features. In: ICPRAM, pp 495–501
57. Tajbakhsh N, Gurudu SR, Liang J (2015) Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging 35(2):630–644
58. Yu Z, Zhao C, Wang Z, Qin Y, Su Z, Li X, Zhou F, Zhao G (2020) Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5295–5305
59. Su Z, Liu W, Yu Z, Hu D, Liao Q, Tian Q, Pietikäinen M, Liu L (2021) Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5117–5127
60. Chen Z, He Z, Lu ZM (2024) Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. IEEE Transactions on Image Processing
61. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
62. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
63. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6848–6856
64. Bernal JJ, Histace A, Masana M, Angermann Q, Sánchez-Montes C, Rodriguez C, Hammami M, Garcia-Rodriguez A, Córdova H, Romain O, Fernández-Esparrach G, Dray X, Sanchez J (2018) Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In: Proceedings of 32nd CARS conference
65. Angermann Q, Bernal J, Sánchez-Montes C, Hammami M, Fernández-Esparrach G, Dray X, Romain O, Sánchez FJ, Histace A (2017) Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In: Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. Springer, p 29–41

66. Misawa M, Kudo Se, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, Baba T, Ishida F, Itoh H, Oda M, Mori K (2021) Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy 93(4):960–967
67. Ma Y, Chen X, Cheng K, Li Y, Sun B (2021) Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, pp 387–396
68. Wu L, Hu Z, Ji Y, Luo P, Zhang S (2021) Multi-frame collaboration for effective endoscopic video polyp detection via spatial-temporal feature transformation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 302–312
69. Bernal J, Histace A (2021) Giana challenge website. https://giana.grand-challenge.org/
70. Itoh H, Misawa M, Mori Y, Oda M, Kudo SE, Mori K (2020) Sun colonoscopy video database. http://amed8k.sundatabase.org/, accessed 2020
71. Chen K, Wang J, Pang J, Cao Y, Xiong Y, Li X, Sun S, Feng W, Liu Z, Xu J, Zhang Z, Cheng D, Zhu C, Cheng T, Zhao Q, Li B, Lu X, Zhu R, Wu Y, Dai J, Wang J, Shi J, Ouyang W, Loy CC, Lin D (2019) MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155
72. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
73. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
74. Sun D, Yang X, Liu MY, Kautz J (2018) Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
75. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
76. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
77. Tian Z, Shen C, Chen H, He T (2019) Fcos: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355
78. Zhou X, Wang D, Krähenbühl P (2019) Objects as points. arXiv:1904.07850
79. Zhu X, Wang Y, Dai J, Yuan L, Wei Y (2017) Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision, pp 408–417
80. Chen Y, Cao Y, Hu H, Wang L (2020) Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10337–10346
81. Zhou Q, Li X, He L, Yang Y, Cheng G, Tong Y, Ma L, Tao D (2022) Transvod: end-to-end video object detection with spatial-temporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence
82. Xiao F, Lee YJ (2018) Video object detection with an aligned spatial-temporal memory. In: Proceedings of the European conference on computer vision (ECCV), pp 485–501
83. Jiang Y, Zhang Z, Zhang R, Li G, Cui S, Li Z (2023) Yona: you only need one adjacent reference-frame for accurate and fast video polyp detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 44–54
84. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2462–2470
85. Hui TW, Tang X, Loy CC (2020) A lightweight optical flow cnn—revisiting data fidelity and regularization. IEEE transactions on pattern analysis and machine intelligence 43(8):2555–2569
86. Teed Z, Deng J (2020) Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision, Springer, pp 402–419
87. Fan L, Zhang T, Du W (2021) Optical-flow-based framework to boost video object detection performance with object enhancement. Expert Systems with Applications 170:114544
88. Hendrycks D, Dietterich T (2019) Benchmarking neural network robustness to common corruptions and perturbations. arXiv:1903.12261
89. Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L, et al (2024) Sam 2: Segment anything in images and videos. arXiv:2408.00714