DMP-WCEBleedNet: A YOLOv11-Based Model for Gastrointestinal Bleeding Detection in Wireless Capsule Endoscopy

Yunyun Chai, Yujun Zhang*

Abstract—**Wireless Capsule Endoscopy (WCE) is a crucial non-invasive modality for diagnosing gastrointestinal (GI) disorders. However, automated detection of bleeding lesions remains challenging due to background complexity, variability in lesion scales, and subtle visual features. To address these challenges, we propose DMP-WCEBleedNet, an advanced bleeding detection framework built upon the YOLOv11 architecture. The model incorporates a Poly Kernel Inception Block (PKIBlock), which enhances lesion detection across varying morphologies and sizes using non-dilated multi-scale convolutions. It also integrates a Self-Modulated Feature Aggregation (SMFA) module that combines global contextual information with local texture details through efficient self-attention approximation and fine-detail enhancement mechanisms. Additionally, a Dynamic Detection Head leverages scale-aware, spatial-aware, and task-aware attention to improve robustness under complex visual conditions. Experimental results on the Auto-WCEBleedGen V2 dataset demonstrate that DMP-WCEBleedNet surpasses existing state-of-the-art methods in AP, mAP@0.5, and mAP@0.5:0.95, while maintaining high inference efficiency. These findings highlight the model's strong potential for real-time clinical application in WCE-based bleeding detection.**

Index Terms—**Wireless Capsule Endoscopy, YOLOv11, PKIBlock, SMFA, Dynamic Detection Head**

I. Introduction

SINCE its introduction in the early 21st century, Wireless Capsule Endoscopy (WCE) technology has emerged as a critical tool for diagnosing digestive tract diseases, particularly in the examination of small intestine disorders where it occupies an irreplaceable role [1]. This technology facilitates non-invasive visualization of the entire digestive tract via a miniature camera capsule, addressing the limitation of traditional endoscopy, which leaves approximately 30 % of the small intestine unexamined. It has substantially enhanced the detection rates of conditions such as gastrointestinal bleeding and Crohn's disease [2]. Nevertheless, a single WCE examination can produce between 60,000 and 100,000 images, requiring physicians to spend 2 to 3 hours manually reviewing them. This process is not only inefficient but also increases the risk of misdiagnosis or missed diagnosis due to visual fatigue [3].

Gastrointestinal bleeding, one of the most common indications for WCE, exhibits a broad spectrum of manifestations, including spot bleeding and diffuse oozing. The interpretation of these images is further complicated by factors such as camera angles, intestinal contents, and mucosal movements. Moreover, differentiating true bleeding from non-hemorrhagic red lesions (e.g., angiodysplasia or inflammatory hyperemia) presents significant diagnostic challenges. Consequently, there is an urgent need to develop efficient and accurate computer-aided diagnostic (CAD) systems to automate and enhance WCE image analysis.

In recent years, the rapid progress of deep learning has offered innovative solutions for intelligent WCE image analysis. AI-based systems can efficiently process large datasets, automatically detect abnormal lesions, and support clinicians in decision-making. However, current methods still encounter challenges in terms of robustness under complex conditions, sensitivity to small lesions, and accuracy in multi-class abnormality classification. The Auto-WCEBleedGen international challenge has significantly contributed to advancing WCE intelligence research by providing a large-scale annotated dataset covering 10 types of gastrointestinal (GI) abnormalities, such as bleeding, ulceration, and vascular lesions [4]. The competition prioritizes Balanced Accuracy as the primary evaluation metric and focuses on developing CAD systems with strong generalization capabilities and clinical applicability, thereby promoting the intelligent evolution of WCE technology.

Looking forward, as artificial intelligence continues to integrate more deeply with medical imaging, WCE-assisted diagnostic systems are anticipated to enhance diagnostic efficiency, alleviate clinician workload, and increase early detection rates of GI diseases, ultimately contributing to the realization of precision medicine.

II. Related Work

Driven by the Auto-WCEBleedGen challenge, the bleeding detection research system for WCE images has gradually matured, forming a clear technological evolution trajectory.

In the V1 phase of 2024, researchers primarily focused on adapting and optimizing basic models in an end-to-end manner. Alawode et al. [5] innovatively integrated the DETR framework into WCE image analysis, constructing an integrated detection and classification model with ResNet-50 as the backbone and Transformer encoder-decoder as the core. By training end-to-end, they eliminated reliance on

Manuscript received April 24, 2025; revised June 28, 2025.

This work was supported by the Key Laboratory of Internet of Things Application Technology on Intelligent Construction, Liaoning Province (2021JH13/10200051).

Yunyun Chai is a graduate student of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 1633006219@qq.com).

Yujun Zhang is a Professor of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (Corresponding author, e-mail: 1997zyj@163.com).

non-maximum suppression (NMS) and enhanced model performance through the Hungarian matching strategy. Shekar et al. [6] designed a unified detection architecture based on YOLOv8-X, combining high-quality annotations with a cross-dataset fusion strategy to expand the training set to 6,345 images. This effectively improved the model's generalization ability and established a data construction paradigm for subsequent research. Entering the V2 phase of 2024, research methods became more refined and task-specific. Alavala et al. [7] proposed a cascaded processing framework based on Swin Transformer and RT-DETR, where the front end performed image classification in the Lab color space using Swin Transformer, and the back end conducted multi-scale object detection with RT-DETR. CLAHE preprocessing and Gaussian blur postprocessing were also applied, significantly enhancing robustness in complex scenarios. Lin et al. [8] constructed a two-stage detection process: the first stage used a region proposal network to locate suspicious bleeding areas, while the second stage introduced an attention mechanism for fine classification, improving the detection accuracy of small targets. In the latest 2025 research, Agossou et al. [9] combined ResNet-50 as the feature extractor and YOLOv5 as the detector to build an integrated classification and detection model. They also adopted k-fold cross-validation and multiple image enhancement techniques (e.g., random rotation and color perturbation) to jointly improve the model's adaptability to diverse data, further pushing the performance boundaries of the WCE bleeding detection task [10].

These innovations not only established strong baselines on the Auto-WCEBleedGen dataset but also provided transferable frameworks and methodologies for broader medical image analysis tasks. Despite significant progress, bleeding detection in WCE images still faces the following challenges: (1) Multi-scale detection difficulty—GI bleeding lesions vary widely in size, and conventional CNNs with fixed receptive fields struggle to capture such variation effectively; (2) Feature fusion limitations—Clinical diagnosis requires attention to both local lesion details and global pathological context, which existing models fail to integrate efficiently; (3) Complex background interference—Artifacts such as bubbles, mucus, and food residues often resemble real bleeding in color and texture, leading to high false positive rates; (4) Computational bottlenecks—Although Transformer-based architectures have improved accuracy, their quadratic complexity (e.g., ViT requires up to 184G FLOPs per frame) limits their applicability in real-time clinical settings.

In the development of object detection algorithms, deep learning-based methods are generally categorized into two-stage and single-stage models based on their architectural principles. Two-stage models, exemplified by the R-CNN series, began with the work of Girshick et al. [11] in 2014, who introduced a pioneering pipeline involving region proposal, classification, and localization, significantly improving detection accuracy. Later, in 2015, He et al. [12] proposed Faster R-CNN, which further optimized feature extraction efficiency by incorporating the Region of Interest (RoI) pooling layer. However, such models are typically associated with high parameter complexity and heavy computational costs, making them unsuitable for real-time clinical deployment in WCE scenarios.

In contrast, single-stage models—such as the YOLO series, SSD, and RetinaNet—formulate object detection as a regression problem and enable end-to-end prediction, offering advantages in terms of reduced parameter size and high inference speed. Notably, the YOLO series has undergone multiple iterations of refinement, achieving detection accuracy that rivals or surpasses two-stage models while maintaining real-time performance [13]. These characteristics make YOLO particularly well-suited for deployment on resource-constrained devices, offering an ideal solution for real-time analysis of WCE images.

Based on these advantages, this study adopts YOLOv11 as the baseline framework and proposes an improved model—DMP-WCEBleedNet—specifically tailored for bleeding detection in WCE images. The model introduces three key innovations: (1) The Poly Kernel Inception Block—By introducing multi-scale depthwise separable convolutions and fusing features from different scales, the model effectively enhances the detection ability for multi-scale bleeding regions. (2) Self-Modulated Feature Aggregation (SMFA) module—By combining EASA, LDE, and PCFN mechanisms, the module achieves efficient fusion of local and global features, enhancing the representation of complex bleeding patterns. (3) Dynamic convolution detection head—Designed with a triple attention mechanism to guide the adaptive adjustment of feature responses, improving the model's detection accuracy and robustness in complex backgrounds.

III. PROPAEDEUTICS

YOLOv11 is a universal multi-task visual model that has undergone systematic improvements based on the YOLOv8 architecture. Its core optimizations mainly lie in the design of the backbone module, the enhancement of the attention mechanism, the upgrade of the detection head structure, and the refinement of the training strategy [14], [15]. Compared with YOLOv10, YOLOv11 further improves the efficiency of feature modeling and detection accuracy. Under the premise of maintaining high real-time performance, it achieves better parameter efficiency and generalization ability [16].

In the backbone, YOLOv11 replaces the C2f modules in YOLOv10 with configurable C3K2 modules. Depending on the value of the c3k parameter, the module dynamically switches its structure: when c3k=False, it behaves like C2f with lightweight cross-layer connections; when c3k=True, the internal Bottleneck is replaced by a more powerful C3 structure to enhance multi-scale feature representation [17]. Additionally, a new C2PSA module is introduced after the SPPF layer (as shown in Fig. 1). This module extends C2f by integrating Pointwise Spatial Attention (PSA) with depthwise separable convolutions, aiming to strengthen spatial attention modeling while maintaining computational efficiency. Optional residual connections and feed-forward networks (FFN) are also included to capture nonlinear relationships between local and global features [18].

In the neck, YOLOv11 retains the PAN path aggregation structure from YOLOv8 but replaces all intermediate modules with the C3K2 version to enhance the flexibility and expressiveness of multi-scale feature fusion [19].

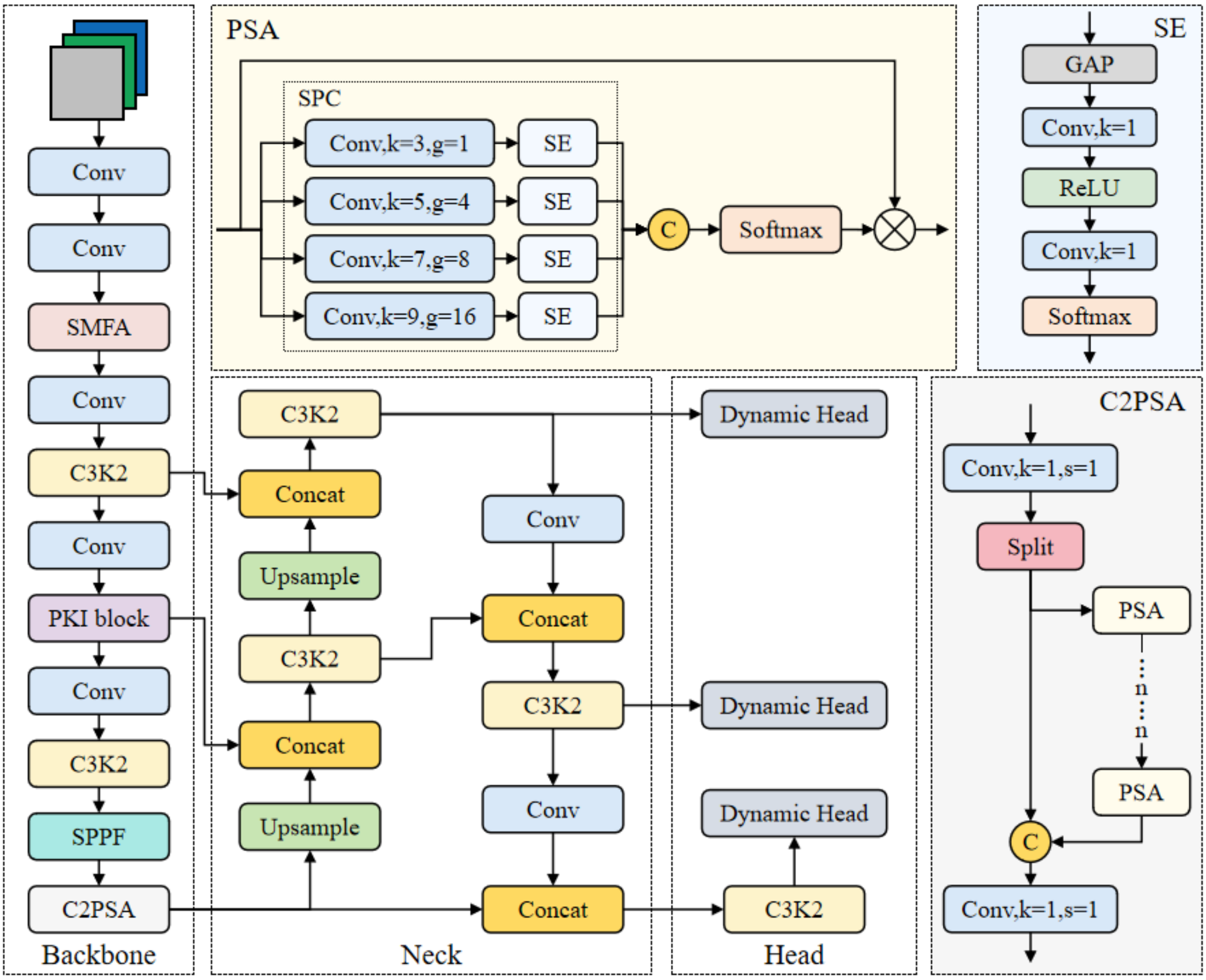

Fig. 1: Overall architecture of the proposed DMP-WCEBleedNet model.

The detection head is also redesigned: the classification branch adopts a dual-depthwise separable convolution structure (DWConv), which significantly reduces redundant computation (up to 75 % compression) while maintaining a wide receptive field [20]. The regression branch combines standard convolution with a distribution-based integral mechanism, outputting a $4 \times regma \times 4$ vector for more precise bounding box modeling. Moreover, the obj branch from YOLOv10 is removed to avoid inconsistencies between training and inference.

In terms of training strategy, YOLOv11 adopts Task-Aligned Assigner for dynamic positive/negative sample assignment, guided by the score formula $t = s^{\alpha} \times u^{\beta}$, where s is the classification score and u is the IoU [21]. The classification branch still uses BCE loss, while the bounding box regression combines Distribution Focal Loss (DFL) and CIoU Loss. DFL models the regression coordinates as a 16-dimensional discrete probability distribution, enhancing robustness to boundary ambiguity. CIoU further incorporates center distance and aspect ratio into the optimization, improving localization accuracy for small objects.

YOLOv11 natively supports five major vision tasks: object detection, instance segmentation, image classification, pose estimation, and oriented bounding box (OBB) detection. Its multi-task adaptability is enabled by: (1) the structural consistency of the universal feature enhancement module C2PSA, (2) a dynamic sample assignment strategy adaptable to various tasks, and (3) the strong generalization ability of DFL loss for handling fuzzy boundaries and scale inconsistencies. Compared to YOLOv8m, YOLOv11m achieves a 1.2 % mAP improvement on the COCO dataset while reducing parameters by 22 % with similar inference speed, demonstrating an optimal balance between accuracy and computational efficiency.

Moreover, YOLOv11 offers high deployment flexibility and can be seamlessly adapted to edge devices, cloud platforms, and NVIDIA GPU-supported environments. This general, lightweight, and accurate model design lays a solid foundation for the high-performance WCE bleeding detection model proposed in this work.

IV. Methods

This paper proposes a WCE bleeding detection model based on the improved YOLOv11 architecture, named DMP-WCEBleedNet, whose overall structure is shown in Figure 1. To address the challenges of the bleeding area in terms of scale, texture and background complexity, the model has carried out three key structural optimizations at the backbone network and detection head levels: introducing PKIBlock (Poly Kernel Inception Block) in the middle of the backbone to jointly model with multi-scale convolution kernels and context anchor attention mechanism (CAA) to enhance the perception ability of the model for bleeding areas in complex scenes; designing a lightweight SMFA module at the front of the backbone to fuse efficient self-attention approximation (EASA) and local detail estimation (LDE) to improve the model's ability to recognize subtle bleeding areas; the detection head part adopts a unified-scale Dynamic Head structure, combining multiple dynamic attention mechanisms that are scale, space and task-aware, to achieve adaptive modulation of features for classification and regression branches. These structures jointly enhance the model's expressive ability and robustness in multi-scale bleeding detection tasks, while maintaining good inference efficiency and deployment flexibility, and possess strong potential for clinical application.

A. Poly Kernel Inception Block (PKI Block)

In the images of gastrointestinal bleeding, the lesion areas exhibit significant differences in morphology and scale, which may manifest as tiny punctate capillary leakage or large-scale diffuse bleeding regions. This heterogeneity in scale span often leads to the model ignoring key structural details during the feature extraction stage, thereby affecting the detection performance. To address this challenge, this paper introduces a multi-scale convolution fusion module - PKIBlock (Poly Kernel Inception Block) into the backbone network of YOLOv11 to replace the original C3k2 module, thereby enhancing the model's ability to model and feature expressiveness for multi-scale bleeding regions.

The PKIBlock is composed of two core sub-modules, as shown in Figure 2, namely the multi-core perception module (PKI Module) and the context anchor attention module (CAA Module). The former focuses on extracting local texture and spatial structure features at different scales, while the latter guides the model to pay attention to representative context regions in the image, thereby achieving the collaborative enhancement of structural breadth and semantic depth.

In the multi-core perception module, multiple groups of depthwise separable convolutions of different sizes (such as 3×3, 5×5, 7×7, 9×9) are used in parallel to model multi-scale responses, thereby expanding the receptive field range while controlling the parameters and computational cost. The outputs of each branch are fused along the channel dimension and then integrated semantically and compressed through a 1×1 convolution to form a multi-scale semantic expression at a unified scale. This structural design takes into account the extraction capabilities of both local texture and contextual information, and is particularly suitable for capturing the spatial variation characteristics of lesion areas in gastrointestinal images due to differences in location and morphology.

Based on this, the CAA module further enhances the model's attention expression on key regions. This module first performs global average pooling on the input features to obtain context summary information; then, it uses a set of orthogonally arranged deep strip convolutions (acting respectively on the horizontal and vertical directions) to approximately simulate large kernel convolution operations, effectively capturing the spatial dependency between distant pixels. Ultimately, the generated attention map is used to perform element-wise modulation on the main branch features, guiding the model to focus more on the ”semantic anchor” regions in the image, such as the center of hemorrhage or clearly defined hemorrhagic lesions, thereby improving the discriminative ability of feature representation.

Finally, the PKIBlock maps the enhanced features to the required number of channels for the backbone network through a 1×11 convolution, while retaining the structural information of the original input features, achieving a cross-module residual connection. This design improves gradient flow, avoids information loss, and maintains compatibility with the original structure of YOLOv11.

B. Self-Modulated Feature Aggregation (SMFA) Module

In the task of identifying gastrointestinal bleeding in WCE images, lesion areas exhibit distinct multi-scale distribution characteristics, ranging from micron-level capillary leakage to centimeter-level ulcer surface bleeding. This scale imbalance poses significant challenges for model detection. Although Transformer-based architectures excel in capturing global context, they suffer from high computational costs and poor real-time performance, making them difficult to directly apply in clinical settings. To address these issues, we introduce a lightweight self-modulated feature aggregation module (SMFA), which aims to balance non-local global modeling capabilities with local detail enhancement capabilities, and improve the model's robustness in identifying multi-scale lesion areas at a lower computational cost.

This module consists of two parallel branches, as shown in Figure 3: an efficient self-attention approximation (EASA) branch for non-local feature extraction, and a local detail estimation (LDE) branch for texture detail enhancement. The two branches are ultimately fused to form a more expressive intermediate representation, providing high-quality feature input for subsequent detection or reconstruction modules.

Feature separation and structural overview: Given the input feature map $F_{in} \in R^{C\times H\times W}$ it is first elevated in dimension to $2C$ through a 1×1 convolution, and then split into two sub-features along the channel dimension:

$$[F_x, F_y] = Split(Conv_{1\times1}(F_{in})), \quad F_x, F_y \in \mathbb{R}^{C\times H\times W} \tag{1}$$

Among them, F_x is input into the EASA branch to extract non-local context, and F_y is sent to the LDE branch to enhance local detail expression.

For the efficient self-attention approximation (EASA) branch, to capture cross-regional structural dependencies at low computational complexity, we design a lightweight attention approximation path. This path builds non-local structure awareness through spatial compression, variance modulation, and resampling mechanisms. First, adaptive max pooling is applied to F_x (with a scaling factor s=8) to reduce the spatial dimension and enhance computational efficiency:

$$F_s = DWConv_{3\times3}(Pool^{(s)}_{\max}(F_x)) \in \mathbb{R}^{C\times\frac{H}{s}\times\frac{W}{s}} \tag{2}$$

Then, spatial variance is introduced as modulation information to perform channel-scale bias enhancement on

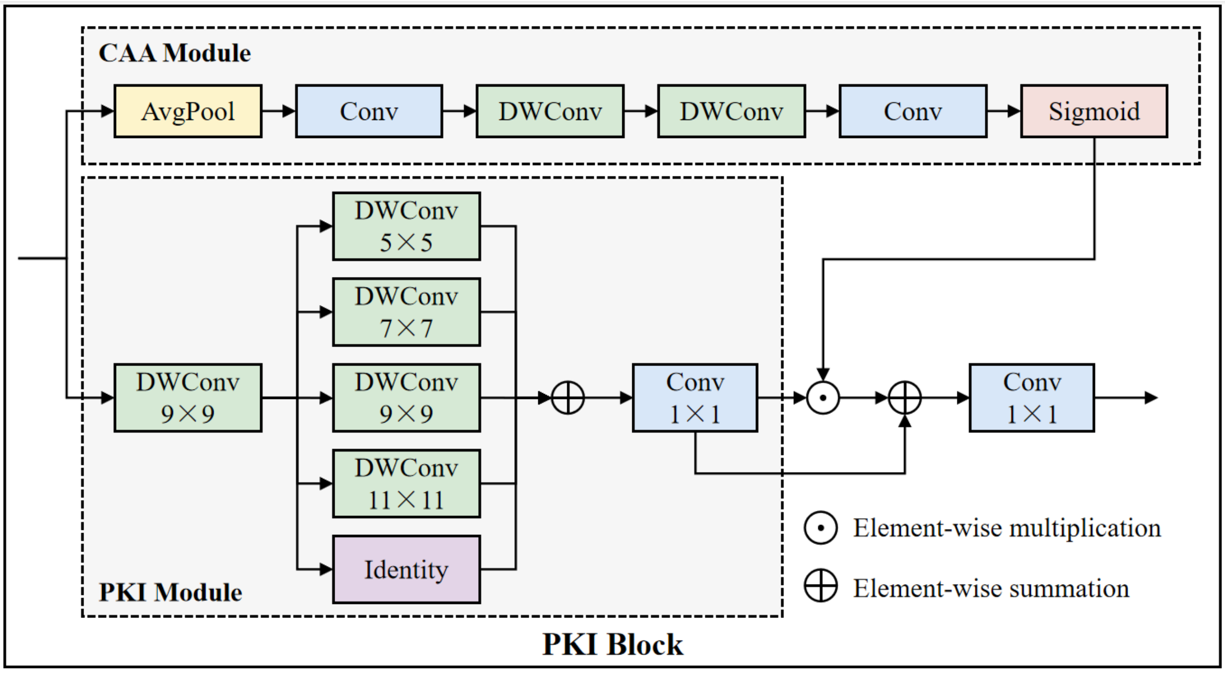

Fig. 2: Structural overview of the PKI Block.

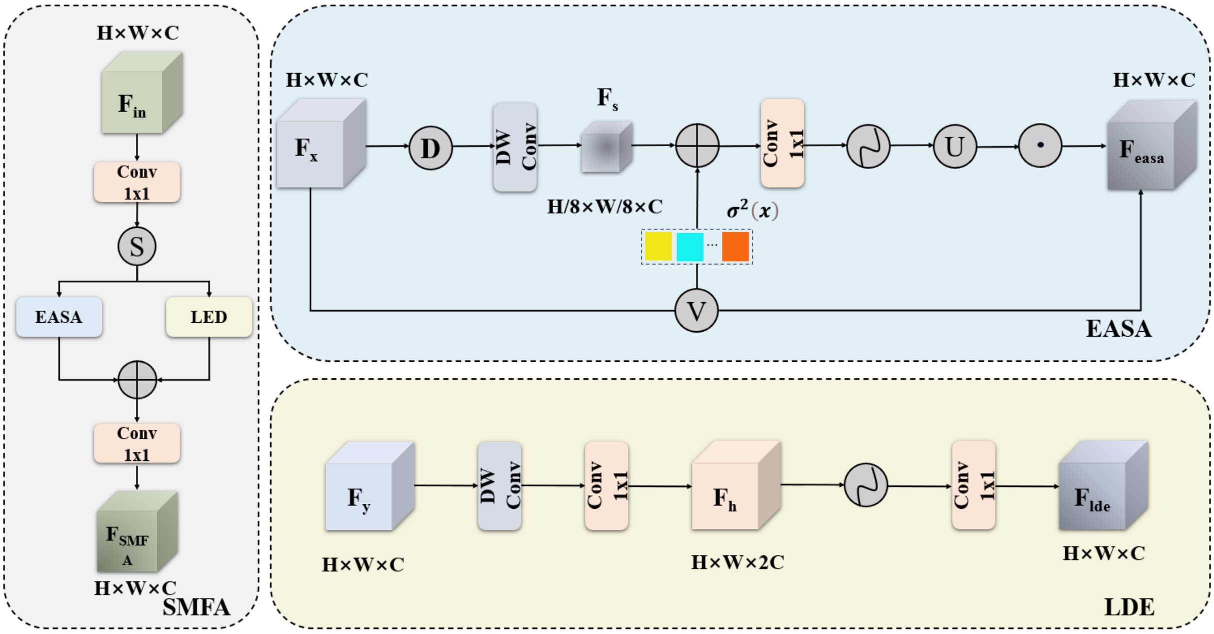

Fig. 3: Architecture of the SMFA Module.

the features:

$$\sigma^2 = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(F_x^{ij}-\mu)^2, \quad \mu = Mean(F_x)$$
$$F_m = Conv_{1\times1}(\alpha \cdot F_s + \beta \cdot \sigma^2) \tag{3}$$

where, $\alpha, \beta \in \mathbb{R}^{C\times1\times1}$are learnable parameters.

Finally, the modulated features are restored to the original size through nearest neighbor upsampling and an activation function, and then fused with the input features:

$$F_{easa} = F_x \odot Upsample(GELU(F_m)) \tag{4}$$

Local Detail Estimation (LDE) branch: To enhance the model's discrimination ability for tiny lesions and marginal hemorrhage, the LDE branch extracts local texture responses through a small receptive field convolution structure. The input Fy first undergoes a depthwise separable convolution to generate detail response features:

$$F_h = DWConv_{3\times3}(F_y) \tag{5}$$

Then, fine-grained enhanced features are constructed through two levels of 1×1 convolution and nonlinear activation functions:

$$F_{lde} = Conv_{1\times1}(GELU(Conv_{1\times1}(F_h))) \tag{6}$$

Feature fusion and module output: The SMFA module fuses the outputs of the two branches through element-wise addition and further compresses the information redundancy with a 1×1 convolution to form the final output feature.

$$F_{SMFA} = Conv_{1\times1}(F_{easa} + F_{lde}) \tag{7}$$

This output has the dual expression ability of global structure and local texture, which can effectively enhance the robustness and detection sensitivity of the model in complex image scenarios.

The SMFA module strikes a balance between high efficiency and strong expressive power in its structure, making it suitable for medical image scenarios where computational resources are a concern. Compared with traditional Transformer modules, SMFA significantly reduces FLOPs consumption. Meanwhile, by introducing variance modulation and local estimation mechanisms, it effectively enhances the modeling quality of bleeding regions at different scales. In WCE image analysis, its enhanced features demonstrate higher recognition accuracy for tiny bleeding points and blurred lesion boundaries, providing more reliable auxiliary support for practical clinical applications.

C. DynamicHead

In the task of gastrointestinal bleeding detection, images often contain complex background elements such as food residues, bubbles, and mucus, which are highly similar in texture and color to real lesions, making false detections very likely. Additionally, bleeding areas may vary significantly in scale among different patients and from different shooting angles, ranging from point-like bleeding to large ulcers. This requires the detection model to not only have good scale adaptability but also possess precise context understanding capabilities. To further enhance the model's detection performance for lesions of different scales and improve its robustness in complex backgrounds, this paper introduces a Dynamic Head in the detection head part of YOLOv11. This module adaptively adjusts the response mode of the convolution kernel at each position by jointly modeling the feature attention in the scale, spatial, and task dimensions, thereby achieving more detailed feature construction and target localization capabilities. The core idea of the Dynamic Head is to uniformly perceive the discriminative information from multiple scale levels, spatial positions, and channel tasks, and enhance the response weight in key areas through a concatenated attention mechanism, thereby suppressing background redundancy and highlighting effective structures.

Specifically, suppose the detection head receives feature maps from different scales (such as P3, P4, P5) of the backbone network, we uniformly reconstruct them into a three-dimensional tensor $F \in \mathbb{R}^{L\times S\times C}$, as shown in Figure 4, where L is the number of feature layers, $S = H \times W$ is the total number of spatial positions, and C is the number of channels. The Dynamic Head uses a cascaded attention mechanism to model this tensor layer by layer. Firstly, the scale-aware attention mechanism dynamically fuses the semantic expression differences among different levels through lightweight convolution to generate weights for each layer and complete weighted summation, thereby enhancing the model's synchronous perception ability for small and large area targets. Secondly, the spatial-aware attention mechanism, based on the deformable sampling strategy, guides the network to focus on discriminative spatial regions (such as lesion edges and central bleeding points) by self-learning position offsets and attention weights, significantly improving the model's accuracy and robustness in spatial target localization. Finally, the task-aware attention mechanism acts on the channel dimension, adaptively activating channels between the classification and regression sub-tasks, and by controlling the activation patterns between channels, achieving task-specific semantic extraction, thereby effectively enhancing the collaborative performance of multi-task detection. Mathematically, this structure can be formalized as:

$$F^{'} = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \tag{8}$$

Among them, π_S, π_L, and π_C respectively represent the scale, spatial, and task attention functions, and $\cdot$ denotes element-wise multiplication. Unlike the traditional multi-head detection branch structure, Dynamic Head integrates all attention operations into a single detection path, greatly simplifying the parameter structure and enhancing the deployability and inference efficiency of the overall model.

V. Dataset and Evaluation Metrics

A. Dataset

The Auto-WCEBleedGen dataset, developed by the MISAHUB team, is specifically designed for bleeding detection and classification in Wireless Capsule Endoscopy (WCE) images, aiming to advance research in this field. The dataset is available in two versions: V1 and V2. Version V1 was first released in collaboration with the 8th International Conference on Computer Vision and Image Processing (CVIP 2023) from August 15 to November 11, 2023, and contains 2,618 WCE frames of both bleeding and non-bleeding conditions. These data are sourced from multiple internet resources and a dataset rich in various types and categories of GI bleeding, covering a wide range of changes throughout the GI tract. Version V2 builds upon V1 by re-labeling multiple bleeding frames and adding new XML and YOLO-TXT formats. All data have been medically verified by professional gastroenterologists to ensure accuracy and reliability. In this study, we utilized the V2 version of the dataset to take full advantage of its improved annotations and more comprehensive data content, thereby better supporting our research objectives.

B. Evaluation Metrics

To comprehensively assess the performance of the proposed DMP-WCEBleedNet model in the WCE image bleeding detection task, this paper adopts the metric system officially recommended by the Auto-WCEBleedGen Challenge. This evaluation system focuses on detection performance, particularly emphasizing the model's ability to precisely locate lesion areas. The main evaluation metrics include precision, recall, average precision (AP), mean average precision (mAP), and its multi-threshold versions (mAP@0.5, mAP@0.5:0.95). In the object detection task, precision measures the proportion of the detection boxes identified as bleeding by the model that are truly lesion areas; recall, on the other hand, reflects the proportion of all true

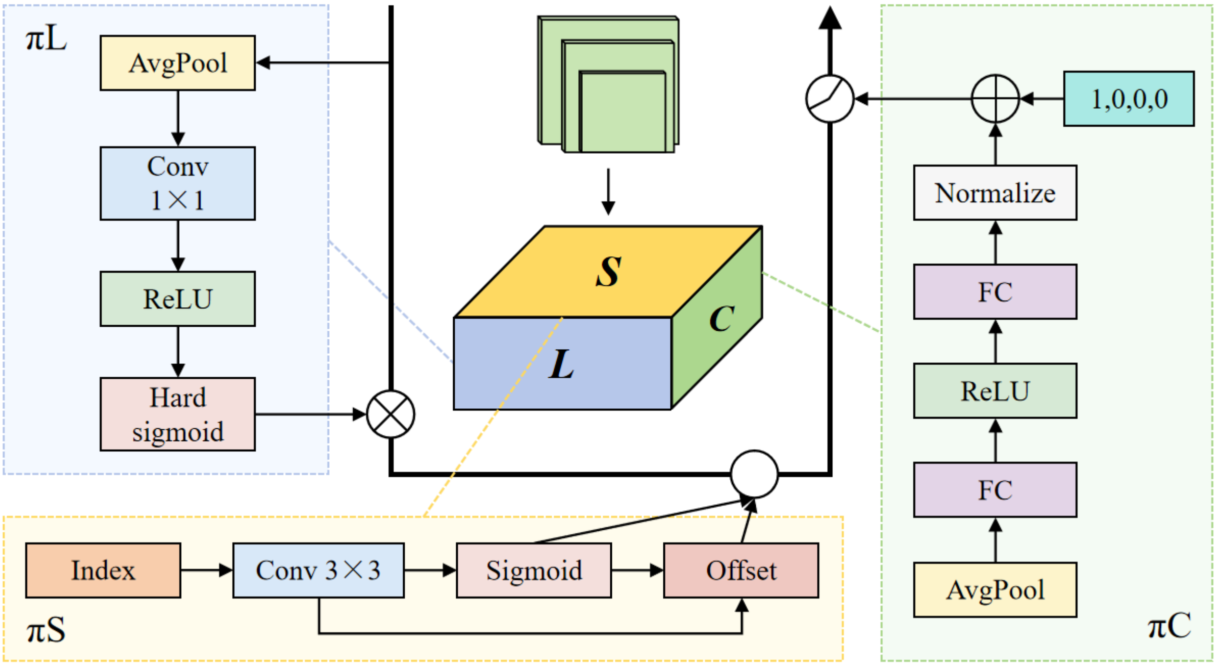

Fig. 4: Workflow of the Dynamic Detection Head.

lesion areas that are successfully detected. Their definitions are as follows:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

Among them, TP represents the number of true bleeding regions detected, FP represents the number of non-bleeding regions falsely detected, and FN represents the number of true bleeding regions missed. Average Precision (AP) is used to measure the detection performance of a single class of targets at different recall rates and is defined as the area under the PR curve (Precision-Recall):

$$AP = \int_0^1 P(R)\,dR \tag{10}$$

Here, P(R) represents the precision at different recall rates. Further, mAP@0.5 indicates the average of AP for all categories when the IoU threshold is 0.5. In single-class tasks, its value is equivalent to AP@0.5, that is:

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

Furthermore, mAP@0.5:0.95 is the average of AP under 10 standards with IoU thresholds ranging from 0.5 to 0.95 (with a step of 0.05), which measures the stability of the model under different matching criteria. Its definition is:

$$\text{mAP@0.5:0.95} = \frac{1}{10}\sum_{i=0}^{9} AP@\left(\text{IoU} = 0.5 + 0.05 \times i\right) \tag{12}$$

This evaluation method strictly adheres to the assessment system of the Auto-WCEBleedGen competition and is applicable to the evaluation requirements of bleeding areas in different scales and complex backgrounds in detection tasks. Particularly, considering that bleeding targets in WCE images are usually small in size, have fine textures and are sparsely distributed, the above indicators not only measure the detection capability but also reflect the reliability and practicality of the model in actual clinical applications.

VI. Experiments and Results Analysis

To comprehensively evaluate the practical performance of the proposed DMP-WCEBleedNet model in the task of bleeding detection in WCE images, this paper designs and conducts two types of experiments: the first is a comparative experiment, where multiple mainstream object detection models are selected as reference objects on the Auto-WCEBleedGen dataset to compare their performance with the model proposed in this paper in various metrics, thereby verifying the effectiveness and advancement of the proposed method as a whole; the second is an ablation experiment, using YOLOv11 as the baseline model, gradually introducing the three improved modules of PKIBlock, SMFA, and Dynamic Head, to analyze the specific contributions of each component to the model's performance, thereby validating the rationality and independent value of the proposed structural design. All related experiments are conducted based on a unified training strategy and evaluation standard to ensure the comparability and objectivity of the results.

TABLE I: Experimental Setup

Parameter	Value
Operating System	Ubuntu 20.04
Deep Learning Framework	PyTorch 2.0.0
Programming Language	Python 3.8
GPU	NVIDIA RTX 4090D (24GB VRAM)
CPU	Intel Xeon Platinum 8481C (16 cores)
RAM	80GB
CUDA Version	11.8

TABLE II: Comparison of Detection Performance on the Auto-WCEBleedGen Dataset. (**Bold** indicates the best result.)

Model Name	AP	mAP@0.5	mAP@0.5:0.95	Recall
YOLOv8-x [6]	0.768	0.768	-	-
Two-Stage WCE [10]	0.7464	-	0.6021	-
YOLOv5-L [4]	-	0.632	0.756	-
DETR-DC5-R101 [7]	-	0.612	0.723	-
WCE Classification & Detection [9]	0.700	0.682	-	-
Transformer-Based [5]	0.7447	0.7328	-	0.7706
Classify ViStA [22]	0.7715	0.726	0.483	-
YOLOv8x [4]	0.650	0.590	0.300	0.59
Divide and Conquer [8]	0.565	0.723	0.434	0.525
YOLOv5nu [4]	0.680	0.630	0.350	0.640
YOLOv8n [4]	0.690	0.630	0.360	0.560
DMP-WCEBleedNet (Ours)	**0.824**	**0.7898**	**0.4831**	**0.6915**

TABLE III: Ablation Study Results on the Auto-WCEBleedGen Dataset. Modules A, B, and C are incrementally added to the baseline YOLOv11. (**Bold** indicates the best result.)

Model	AP	mAP@0.5	mAP@0.5:0.95	Recall
YOLOv11	0.758	0.756	0.46200	0.67100
YOLOv11+A	0.799	0.762	0.46600	0.68100
YOLOv11+A+B	0.81487	0.77062	0.48143	0.66062
YOLOv11+A+B+C	**0.82403**	**0.78979**	**0.48310**	**0.69149**

A. *Experimental Configuration*

This study was conducted in a Linux-based environment with Ubuntu 20.04 as the operating system. Model development and training were carried out using the PyTorch 2.0.0 deep learning framework and Python 3.8. A detailed overview of the hardware and software configurations is provided in Table I.

During the training process, DMP-WCEBleedNet was constructed upon the YOLOv11 architecture, with input images uniformly resized to 640×640 pixels. The model was trained for 300 epochs with a batch size of 12. Stochastic Gradient Descent (SGD) was adopted as the optimizer, configured with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay factor of 0.0005. A linear learning rate decay strategy was utilized, while the cosine annealing scheduler was disabled. To enhance training efficiency and model performance, Automatic Mixed Precision (AMP) was enabled. In terms of data augmentation, techniques such as Mosaic, AutoAugment, and random erasing (with an erasing probability of 0.4) were incorporated, while redundant methods like MixUp and CopyPaste were excluded. The detection task was defined as object detection, with the Intersection over Union (IoU) threshold set to 0.7 and a maximum of 300 detection targets allowed per image. For the sake of reproducibility, the random seed was fixed to 0, and deterministic training behavior was enforced to ensure consistent experimental outcomes.

B. *Comparative Experiments*

To comprehensively validate the effectiveness of the proposed DMP-WCEBleedNet model in the task of bleeding detection using Wireless Capsule Endoscopy (WCE), we conducted a systematic comparative experiment on the Auto-WCEBleedGen dataset with multiple mainstream detection models. These comparison methods are representative models in this task or related medical image analysis fields in recent years, and some of them are also excellent solutions in the Auto-WCEBleedGen competition, covering various design paradigms from classic convolutional architectures to those integrating Transformer mechanisms.

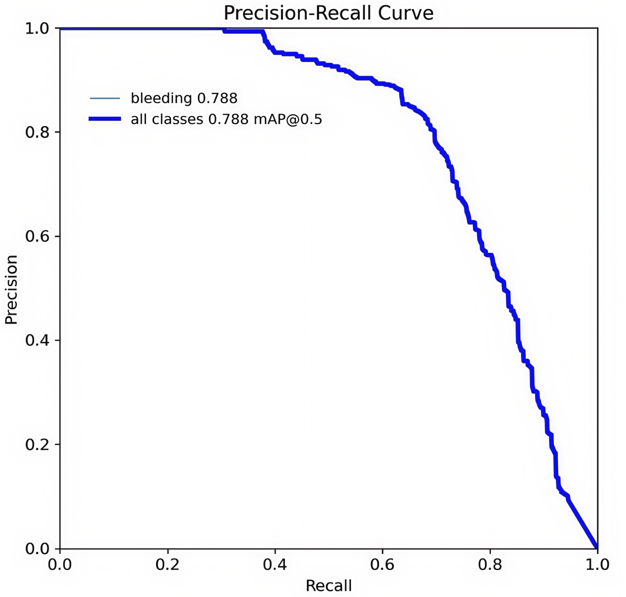

Fig. 5: Precision-Recall Curve of DMP-WCEBleedNet on the Test Set

All models were trained and tested under the same data partitioning and training strategies, using the official release or publicly available and reproducible implementation versions provided by the authors. The comparison models include YOLOv5-L and YOLOv5nu based on lightweight convolutional networks, DETR-DC5-R101 integrating the end-to-end detection paradigm, YOLOv8-x based on the latest structural optimization, Divide and Conquer and ViStA methods adopting multi-stage strategies, as well as Transformer-Based architectures and WCE hemorrhage detection-related models proposed in recent years. All comparison methods were experimented with the publicly available code provided by the authors or their reproducible versions, under unified data partitioning, training strategies,

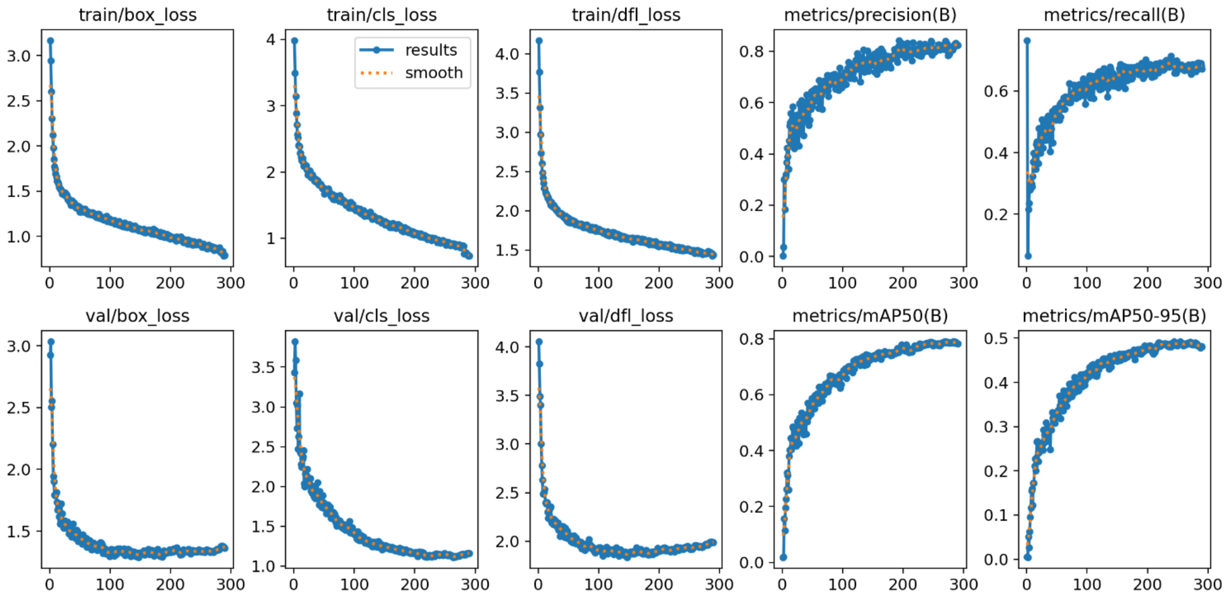

Fig. 6: Loss curves and performance metrics of DMP-WCEBleedNet during training and validation.

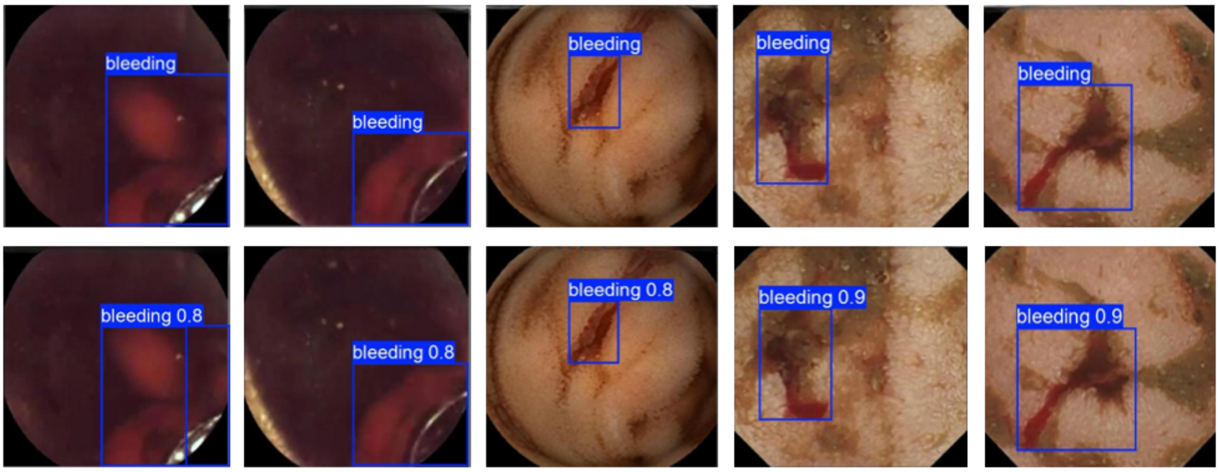

Fig. 7: Visualization of Detection Results by DMP-WCEBleedNet on the Auto-WCEBleedGen Test Set.

and evaluation metrics, to ensure the fairness of the results and the effectiveness of the comparison.

Table 1 presents the comparison results of detection performance on the Auto-WCEBleedGen dataset. It can be seen that DMP-WCEBleedNet performs outstandingly in all evaluation metrics, with AP, mAP@0.5, and mAP@0.5:0.95 reaching 0.824, 0.7898, and 0.4831 respectively, significantly outperforming current mainstream methods. Particularly in the more comprehensive mAP@0.5:0.95 metric, this model still achieves the best result, further verifying its high-precision capability in modeling and identifying multi-scale bleeding regions. Additionally, DMP-WCEBleedNet also demonstrates excellent performance in recall rate, reaching 0.6915, which reflects its robustness in detecting bleeding regions in complex backgrounds.

To further analyze the trade-off between the detection accuracy and recall capability of the model, this paper plots the Precision-Recall (PR) curve of DMP-WCEBleedNet on the test set, as shown in Figure 5. It can be seen from the figure that the model can maintain a high precision rate at different recall rates, and the PR curve shows a generally stable downward trend, reflecting that DMP-WCEBleedNet still has a stable recognition ability in dealing with complex scenes and bleeding areas with ambiguous boundaries. Among them, mAP@0.5 reaches 0.788, further verifying the excellent performance of this model in the detection task.

Based on this, this paper further conducts a visualization analysis of the model's optimization behavior from the perspective of the training process. Figure 6 shows the changes in various loss functions and performance metrics of DMP-WCEBleedNet during the training and validation stages with the number of iterations. The upper row shows the box loss, cls loss, and dfl loss during the training stage, as well as the corresponding precision and recall; the lower row shows the corresponding loss terms and the change trends of mAP@0.5 and mAP@0.5:0.95 during the validation stage. It can be seen that all the losses of the model continuously decrease during the training process, with a clear convergence trend, and the performance metrics steadily improve, demonstrating the good stability and generalization ability of DMP-WCEBleedNet in the feature learning and optimization process, which provides a solid support for its outstanding performance in the bleeding detection task.

C. Ablation Experiments

To evaluate the effectiveness of each proposed module in DMP-WCEBleedNet, we conducted ablation studies by successively introducing Modules A, B, and C into the YOLOv11 baseline. The results of each configuration are shown in Table 2.

Module A introduces a non-dilated multi-scale convolutional kernel design to enhance the model's ability to extract features from bleeding regions of varying sizes. Module B adds the Self-Modulated Feature Aggregation (SMFA) module to improve multi-scale semantic feature fusion. Module C replaces the original detection head with a dynamic convolutional head to adapt to the diversity of lesion regions in both scale and morphology.

Experimental results show that after introducing Module A, the model's AP increased from 0.758 to 0.799 and mAP@0.5 improved by 0.6%, indicating that the multi-scale receptive field design effectively enhanced detection performance. With the addition of Module B, mAP@0.5 increased to 0.7706 and mAP@0.5:0.95 reached 0.4814, demonstrating the effectiveness of the SMFA module in feature fusion. Finally, the introduction of Module C led to the highest detection performance, with mAP@0.5 rising to 0.7898 and recall increasing to 0.6915, further validating the enhanced modeling capability of the dynamic detection head for bleeding targets.

VII. Conclusion

This paper focuses on the automatic detection of bleeding lesions in Wireless Capsule Endoscopy (WCE) images and proposes an efficient detection model, DMP-WCEBleedNet, based on the improved YOLOv11 framework. By integrating a multi-scale convolution fusion structure (PKIBlock), a self-modulated feature aggregation module (SMFA), and a unified-scale Dynamic Head detection head, the model achieves significant improvements in both feature extraction and region discrimination. Experimental results demonstrate that the proposed model substantially outperforms various mainstream detection methods on the Auto-WCEBleedGen dataset, particularly in critical metrics such as mAP@0.5 and Recall. Ablation studies further validate the independent contributions of each module to accuracy enhancement. Moreover, the visualization of detection results presented in Figure 7 highlights the model's excellent practical interpretability, with predicted bleeding areas closely aligning with real annotations and most detection box confidence scores exceeding 0.8. This indicates that the model can accurately locate multi-scale lesions while maintaining high discrimination confidence and stability. In summary, DMP-WCEBleedNet not only ensures superior detection accuracy but also emphasizes model lightweight design and inference efficiency, showcasing substantial potential for clinical deployment.

References

[1] G. Pan and L. Wang, "Swallowable wireless capsule endoscopy: Progress and technical challenges," *Gastroenterology research and practice*, vol. 2012, no. 1, p. 841691, 2012.

[2] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 639–642, IEEE, 2016.

[3] R. Hu, H. Wang, S. Zhang, W. Zhang, and P. Xu, "Improved u-net segmentation model for thyroid nodules," *IAENG International Journal of Computer Science*, vol. 52, no. 5, pp. 1407–1416, 2025.

[4] M. Hub, P. Handa, D. Nautiyal, D. Chhabra, M. Dhir, A. Saini, S. Jha, H. Mangotra, N. Pandey, A. Thakur, *et al.*, "Auto-wcebleedgen version v1 and v2: Challenge, datasets and evaluation," *Authorea Preprints*, 2024.

[5] B. Alawode, S. Hamza, A. Ghimire, and D. Velayudhan, "Transformer-based wireless capsule endoscopy bleeding tissue detection and classification," *arXiv preprint arXiv:2412.19218*, 2024.

[6] P. C. Shekar, V. Kanhangad, S. Maheshwari, and T. S. Kumar, "Automated bleeding detection and classification in wireless capsule endoscopy with yolov8-x," *arXiv preprint arXiv:2412.16624*, 2024.

[7] S. Alavala, A. K. Vadde, A. Kancheti, and S. Gorthi, "A robust pipeline for classification and detection of bleeding frames in wireless capsule endoscopy using swin transformer and rt-detr," *arXiv preprint arXiv:2406.08046*, 2024.

[8] Y.-F. Lin, B.-C. Qiu, C.-M. Lee, and C.-C. Hsu, "Divide and conquer: Grounding a bleeding areas in gastrointestinal image with two-stage model," *arXiv preprint arXiv:2412.16723*, 2024.

[9] B. E. Agossou, M. Pedersen, K. Raja, and A. Vats, "Classification and detection of bleeding in wireless capsule endoscopy (wce)," *Authorea Preprints*, 2025.

[10] S. Neogy, S. Mazumder, N. Chowdhury, T. Sur, and S. Das, "Towards automated screening via two-stage deep learning: A pipeline for classification and localization of bleeding from wireless capsule endoscopy visuals," in *International Conference on Advanced Computing and Applications*, pp. 439–453, Springer, 2024.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "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, 2014.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[13] X. Li and Y. Zhang, "A lightweight method for road damage detection based on improved yolov8n," *Engineering Letters*, vol. 33, no. 1, pp. 114–123, 2025.

[14] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[15] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using yolo: Challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

[16] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, 2024.

[17] C.-Y. Wang, H.-Y. M. Liao, *et al.*, "Yolov1 to yolov10: The fastest and most accurate real-time object detection systems," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.

[18] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022.

[19] C. Wang, Q. Zhang, and J. Huang, "An improved multi-target detection algorithm in uav aerial images based on yolov8s framework," *Engineering Letters*, vol. 33, no. 4, pp. 998–1007, 2025.

[20] D. Hou, Y. Zhang, and J. Ren, "A lightweight object detection algorithm for remote sensing images," *Engineering Letters*, vol. 33, no. 3, pp. 704–711, 2025.

[21] D. Wang, J. Tan, H. Wang, L. Kong, C. Zhang, D. Pan, T. Li, and J. Liu, "Sds-yolo: An improved vibratory position detection algorithm based on yolov11," *Measurement*, vol. 244, p. 116518, 2025.

[22] S. Balasubramanian, A. Abhishek, Y. Krishna, and D. Gera, "Classifyvista: Wce classification with visual understanding through segmentation and attention," *arXiv preprint arXiv:2412.18591*, 2024.