Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(9): 6658–6687.
DOI: 10.3934/era.2026291
Received: 27 May 2026
Revised: 12 July 2026
Accepted: 16 July 2026
Published: 04 August 2026

Research article

BKMNet: A boundary-enhanced state space model with Kalman filtering for adenomatous polyp segmentation

Jiaoju Wang[1,2,*], Bing Tan[1,2], Jingming Li[3], Yiming Shu[4], Meiling Tang[4], Binghua Tao[4], Muzhou Hou[4,5,*] and Shuijiao Chen[6,*]

[1] School of Mathematics and Statistics, Nanyang Normal University, Nanyang 473061, China

[2] Applied Mathematics Center of Henan Province, Nanyang 473061, China

[3] School of Civil Engineering and Architecture, Nanyang Normal University, Nanyang 473061, China

[4] School of Mathematics and Statistics, Central South University, Changsha 410083, China

[5] School of Computer Science and Engineering, Hunan University of Information Technology, Changsha 410151, China

[6] Department of Gastroenterology, Xiangya Hospital of Central South University, Changsha 410008, China

* **Correspondence:** Email: wjj@nynu.edu.cn, houmuzhou@hnuit.edu.cn, shuijiaoxy@csu.edu.cn.

Abstract: Colorectal cancer is one of the leading causes of cancer-related mortality, and both early detection and the removal of adenomatous polyps are effective treatment measures. However, identifying adenomatous polyps remains challenging due to the diverse polyps, boundaries, and intestinal mucosal background. Therefore, we have proposed a boundary-enhanced adenomatous polyp segmentation network with Kalman-mamba (BKMNet), which integrates visual state space mechanisms and Kalman filtering for precise segmentation and identification of adenomatous polyps. Specifically, we adopted a multi-frequency combined attention block to enhance feature representation by focusing on varying channel frequency components. A fast visual state space module extracts multi-level feature maps, the partial channel transformation aggregates the surrounding contextual information, and the structure-aware local-enhanced residual state space module enhances structure awareness. A Kalman-mamba layer aims for long-range dependency modeling. The decoder has a dual attention module with multi-level features, and a refined differential and boundary-enhanced module performs edge detection and global edge feature learning. Results demonstrated that BKMNet outperforms state-of-the-art methods in terms of segmentation accuracy and clinical applicability. This study provides a reliable auxiliary diagnostic tool for supporting clinicians in improving diagnostic efficiency and reducing misdiagnosis rates.

Keywords: mamba; Kalman filtering; boundary learning; segmentation; adenomatous polyps

1. Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in recent years, which accounts for 10% of global cancer incidence and 9.4% of cancer deaths, just below lung cancer, which comprises 18% of deaths [1]. Clinical studies have confirmed that more than 90% of colorectal cancers evolve from malignant transformation of adenomatous polyps, which are the most critical precancerous lesions of colorectal cancer [2]. Different from nonadenomatous polyps, adenomatous polyps have inherent potential canceration characteristics, and can be classified into villous adenomatous polyps (VAPs) and tubular adenomatous polyps (TAPs) based on their morphological structure and texture. Villous adenomas pose a substantially elevated risk of malignant transformation compared to tubular adenomas [3]. Early accurate screening, localization, and precise segmentation of adenomatous polyps can effectively block the progression of colorectal lesions, significantly reduce the incidence of colorectal cancer, and improve the survival rate of patients [4].

Colonoscopy is the gold standard for clinical screening of colorectal polyps, but manual diagnosis relies heavily on the clinical experience of physicians, accompanied by problems such as low efficiency, easy missed diagnosis, and subjective judgment deviation [5, 6]. An effective and objective automated tool is needed to aid clinicians in diagnosing adenomatous polyps quickly and accurately. However, automatic segmentation of adenomatous polyps is a challenging task due to the inherent complexity of polyp lesions and complex intestinal imaging environments. Adenomatous polyps exhibit high heterogeneity in shape, size, and anatomical location, while their texture and color are highly similar to those of normal colonic mucosa. Moreover, their morphological characteristics overlap considerably with non-adenomatous polyps, and the lesion boundaries between adenomatous polyps and surrounding normal intestinal tissues are blurred and indistinct [7]. In addition, multiple external factors further degrade the performance and robustness of detection models. Complex intestinal background noise, including irregular mucosal folds and residual debris, exhibits polyp-like morphological features that disrupt feature extraction and frequently trigger false-positive segmentation results. Meanwhile, significant inter-patient heterogeneity in polyp contour, volume, attachment site, and surface texture further impairs the stability of model segmentation performance. Notably, small adenomatous polyps occupy a few pixels in endoscopic images, leading segmentation models to frequently miss small lesions.

To address the above clinical challenges, deep learning (DL)-based computer-aided diagnosis (CAD) methods have emerged as promising solutions, thanks to their powerful hierarchical feature learning capabilities [8–13]. Such techniques have been widely applied in multiple clinical scenarios [14–16]. In the field of colorectal polyp analysis, a series of segmentation networks represented by UNet-based segmentation models and transformer-based global modeling networks have achieved favorable results in polyp segmentation tasks [17–19]. Convolutional neural networks (CNNs) rely on local convolution kernels to extract feature information, which has a strong ability to capture local texture and structural features of polyps, and has low computational complexity, meeting the real-time requirements of endoscopic diagnosis. On this basis, multi-scale convolution, dilated convolution, channel attention, and spatial attention mechanisms are widely introduced to solve the problems of multi-scale polyp adaptation and feature interference suppression [20–23]. In recent years, the emerging mamba-based state space model (SSM) has broken the long-distance modeling bottleneck of traditional CNNs and the high computational cost limitation of transformers [24, 25]. With its linear computational complexity and excellent long-range dependency modeling ability, mamba-based algorithms have been gradually

applied to medical image segmentation, realizing efficient global feature modeling of polyp images and further improving the overall segmentation accuracy of lesions [26].

Despite the promising progress of existing DL-based polyp segmentation methods, they still have several limitations. CNN-based models rely purely on local convolution operations and lack effective long-range dependency modeling, making them unable to adapt to the substantial morphological heterogeneity of adenomatous polyps and preserve fine-grained features of polyps [27–29]. Although transformer-based models support global feature modeling, their bulky computational architecture and excessive parameter overhead fail to meet the lightweight and real-time requirements of clinical endoscopic diagnosis [30]. Moreover, most vision mamba models simply stack state space modules to complete global feature modeling, ignoring the unique structural characteristics of adenomatous polyps with directional continuity and local fine-grained differences [30]. In addition, many models lack dedicated boundary optimization modules, and they cannot effectively distinguish blurred polyp edge features from complex intestinal background interference.

Aiming at the above challenges, this paper introduces a Kalman filter mechanism on the basis of the mamba module and constructs a lightweight structural-aware boundary enhanced mamba segmentation network (BKMNet) for accurate segmentation of adenomatous polyps. Specifically, a multi-frequency combined attention (MFCA) block is first used to achieve adaptive weighting of multi-scale and multi-frequency features of polyps. A fast visual state space block (FVSSB) is adopted as the basic feature extraction unit. Combined with partial channel transformation (PCT) and a structure-aware local-enhanced residual state space module (SALE_RSSM), this block acquires a powerful local structural feature representation capability with lighter parameters, effectively capturing fine-grained features of small lesions and suppressing complex background noise. To tackle the morphological heterogeneity of adenomatous polyps and enhance the global feature modeling capability, a Kalman-mamba layer is embedded into the network. The mamba realizes efficient long-range dependency modeling with low computational cost, while Kalman filtering adaptively optimizes features, improving model robustness and generalization. In the decoder stage, a dual attention concatenation (DAC) module is utilized to fuse multi-scale hierarchical features and compensate for shallow detail loss during deep feature extraction. Furthermore, a refined differential and boundary-enhanced (RD-BE) module is designed to solve the boundary ambiguity problem. It decouples polyp body and edge features, and leverages an edge-aware graph convolution (EAGC) module to strengthen edge feature learning, achieving precise segmentation of blurred polyp boundaries. Finally, a composite joint loss function is employed to supervise edge features, body features, and overall segmentation results simultaneously, further improving the model's anti-interference ability and clinical generalization performance. Extensive experiments demonstrate that the proposed BKMNet achieves superior segmentation accuracy and differentiation capability with lightweight parameters, which is suitable for clinical auxiliary diagnosis of adenomatous polyps.

The major contributions in this study consist of the following aspects:

- A FVSS block with a SALE_RSSM module is designed to efficiently extract image features. The SALE_RSSM innovatively integrates directional convolution and a local enhanced residual state space mechanism. The directional convolution realizes multi-directional structural feature extraction of adenomatous polyps and generates structural adaptive weight maps, which enhance the perception ability of polyp edges and continuous structural features. The local enhanced residual SSM realizes fine-grained enhancement of local lesion features and suppresses background interference, effectively mitigating the problem that traditional mamba techniques ignore local

structural details.

- A Kalman filter layer is introduced into the encoder tail. Through the parallel state prediction and update mechanism, the feature map is dynamically smoothed and optimized, which improves the discrimination between lesion features and background features, and makes up for the defect of an insufficient anti-interference ability of existing mamba segmentation models.
- The MFCA block and DAC mechanism are applied. The multi-frequency channel attention realizes adaptive weighting of multi-scale and multi-frequency features of polyps, and the feature aggregation structure completes hierarchical fusion of shallow detail features and deep semantic features, which significantly improves the segmentation accuracy of small polyps and blurred edge regions.
- The network adopts edge-body decoupled supervision and a cascaded feature refinement strategy, which adopts the RD-BE module to separately optimize polyp edge contour and internal body features, and further calibrates lesion boundary information through contour graph convolution refinement, realizing high-precision and robust segmentation of adenomatous polyps in complex endoscopic scenes.

2. Related works

2.1. Boundary learning

To enhance the performance of semantic segmentation, some studies have tried to embed the low-level features that contain boundary and edge information into high-level features [31, 32]. For instance, Lin et al. designed a boundary-guided decoder network that uses a boundary mask obtained from a dedicated boundary detection operator as explicit supervision to enhance the learning capacity on the boundary [33]. Wu et al. proposed a novel two-stage brain MRI lesion segmentation method that introduced a boundary deformation module (BDM) and boundary constraint module (BCM) to handle the tough condition of fuzzy boundaries [34]. Li et al. proposed a boundary-guided network with two-stage transfer learning, which consists of the boundary feature extraction module (BFE), deep feature extraction module (DFE), and multi-scale fusion module (MF) to generate boundary maps that guide the decoder in generating prediction maps [35]. Inspired by these studies, we try to improve the polyp segmentation by decoupling the body and edge parts with different supervision for more accurate boundary detail in the segmentation process.

2.2. Transformer and state space models in medical segmentation

The vision transformer has obtained state-of-the-art performance in segmentation tasks in recent years [36–39]. Chen et al. first explored the potential of the transformer in the context of medical image segmentation and proposed a hybrid CNN-transformer as the encoder in their TransUnet for encoding strong global context, by treating the image features as sequences and utilizing the low-level CNN features [40]. Analogously, Zhang et al. proposed the asymmetric compound branch transformer for medical image segmentation, which effectively extracted local and global representations using an asymmetric asynchronous branch parallel structure while reducing unnecessary computational costs [41]. However, transformers typically have high computational complexity and memory consumption, which limits their application in real-time clinical scenarios.

Recently, state space models (SSMs), such as mamba [25], have emerged as efficient alternatives to transformers. SSMs process sequences by modeling state transitions, enabling efficient long-range dependency modeling with linear computational complexity. Mamba has shown promising results in natural language processing and computer vision tasks. In medical imaging, several studies have attempted to apply SSMs to segmentation [42–44]. For example, Mamba-UNet [45] replaces the convolutional layers in U-Net with mamba layers, improving segmentation accuracy while maintaining computational efficiency. However, existing SSM-based models still have room for improvement in terms of multi-scale feature capture and boundary refinement.

2.3. Attention mechanism

The attention mechanism aims to focus on image regions that are more task-related by learning a binary matrix or a weighted matrix, which has proven to be a potential means to enhance deep segmentation models [46, 47]. For instance, an efficient channel prior convolutional attention (CPCA) was proposed to improve medical segmentation [48]. Liu et al. proposed a cross-attention and feature exploration network (CAFE-Net) for polyp segmentation, which used a cross-attention decoder module (CADM) to effectively preserve features from lower layers and restore fine-grained information [49]. Inspired by this, we incorporate the attention mechanism into our model to alleviate problems arising from scale variation and small object instances.

3. Materials and methods

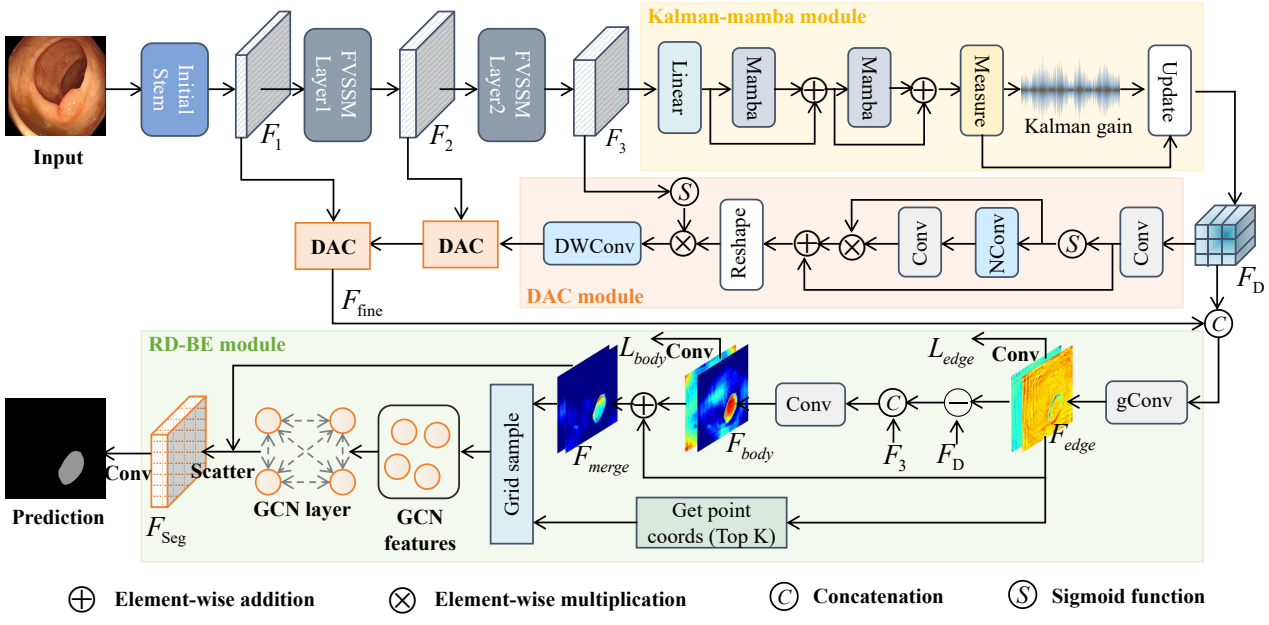

Figure 1. The overall architecture of the proposed BKMNet.

The proposed computer-assisted diagnostic model, designated BKMNet, is designed for the precise localization and identification of adenomatous polyps from endoscopic images. This approach captures

detailed features of the colorectal polyps at various scales and also refines the segmentation boundaries, leading to more reliable diagnostic outcomes. As illustrated in Figure 1, for an input endoscopic image $X \in \mathbb{R}^{H\times W\times 3}$, we aim to predict a pixel-wise label map Y, where each pixel has been assigned a class label, for the differential diagnosis of colorectal polyps. The pipeline begins by feeding X into the efficient visual state space encoder. The initial stem comprises standard convolutions and the multi-frequency combined attention (MFCA) block [50], extracting preliminary features F_1 and enhancing the representation by focusing on varying channel frequency components. Subsequently, the encoder utilizes lightweight FVSSM layers to extract multi-level feature maps F_2 and F_3, which are crucial for subsequent processes, as they provide a rich hierarchy of information that can be utilized for accurate segmentation. After that, a Kalman-mamba layer is designed to enable long-range dependency modeling, resulting in a feature representation F_D. In the decoder, a dual attention module is employed to integrate the different-level features and the refined differential and boundary-enhanced (RD-BE) module executes precise edge detection and global edge feature learning. Leveraging the integrated multi-level features, the RD-BE generates a refined edge map to guide accurate final segmentation. The entire model is supervised using a joint loss function consisting of three terms accounting for the boundary, the polyp body (non-boundary), and the overall segmentation (merge).

3.1. Efficient visual state space encoder

3.1.1. Initial stem

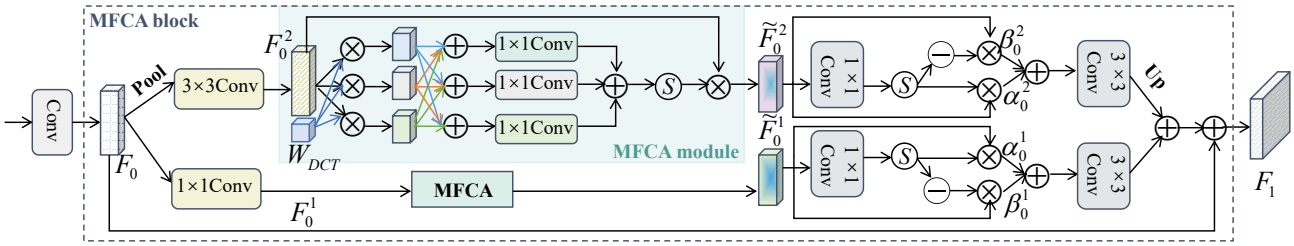

Figure 2. The overall structure of the initial stem layer.

The initial stem layer includes three 3×3 standard convolutions and a multi-frequency combined attention (MFCA) block [50]. This module combines multi-scale branches and triple-pooling multi-frequency aggregation to enhance shallow and subtle polyp textures. As shown in Figure 2, for the input image $X \in \mathbb{R}^{H\times W\times 3}$, the standard convolutions are first used to extract the feature map F_0 for the input of the MFCA block. Once F_0 is obtained, it is then fed into a two-branch architecture with different scales. Features at each scale branch (F_0^1 and F_0^2) are then input into the multi-frequency channel attention (MFCA) module. As shown in Figure 2, the features $F_0^i (i = 1, 2)$ are firstly characterized using discrete cosine transform (DCT) and are compressed into F_{avg}^i, F_{max}^i, and F_{min}^i using global average pooling, global max pooling, and global min pooling, respectively. Subsequently, the output of the MFCA module can be obtained by aggregating each statistic of frequency:

$$\tilde{F}_0^i = F_0^i \cdot f_\sigma(f_{1_2}(F_{avg}^i) + f_{1_2}(F_{max}^i) + f_{1_2}(F_{min}^i)), \tag{3.1}$$

where f_σ is the sigmoid activation function and f_{1_2} denotes two 1×1 convolution layers. The channel recalibrated features $\tilde{F}_0^1$ and $\tilde{F}_0^1$ are then used to determine discriminative boundary cues with various

scales in the spatial domain:

$$\bar{F}_0^i = f_3(\alpha_0^i(\tilde{F}_0^1 \cdot f_\sigma(f_1(\tilde{F}_0^1))) + \beta_0^i(\tilde{F}_0^1 \cdot (1 - f_\sigma(f_1(\tilde{F}_0^1))))), \tag{3.2}$$

where α_0^i and β_0^i are learnable parameters for each scale branch to control the information flow between the foreground and background. f_1 and f_3 denote the 1×1 and 3×3 convolutions, respectively. Finally, the output feature of the initial stem layer (F_1) can be obtained using residual connection:

$$F_1 = F_0 + (\bar{F}_0^1 + f_{up}(\bar{F}_0^2)), \tag{3.3}$$

where f_{up} is the upsample operation.

3.1.2. FVSSM layer

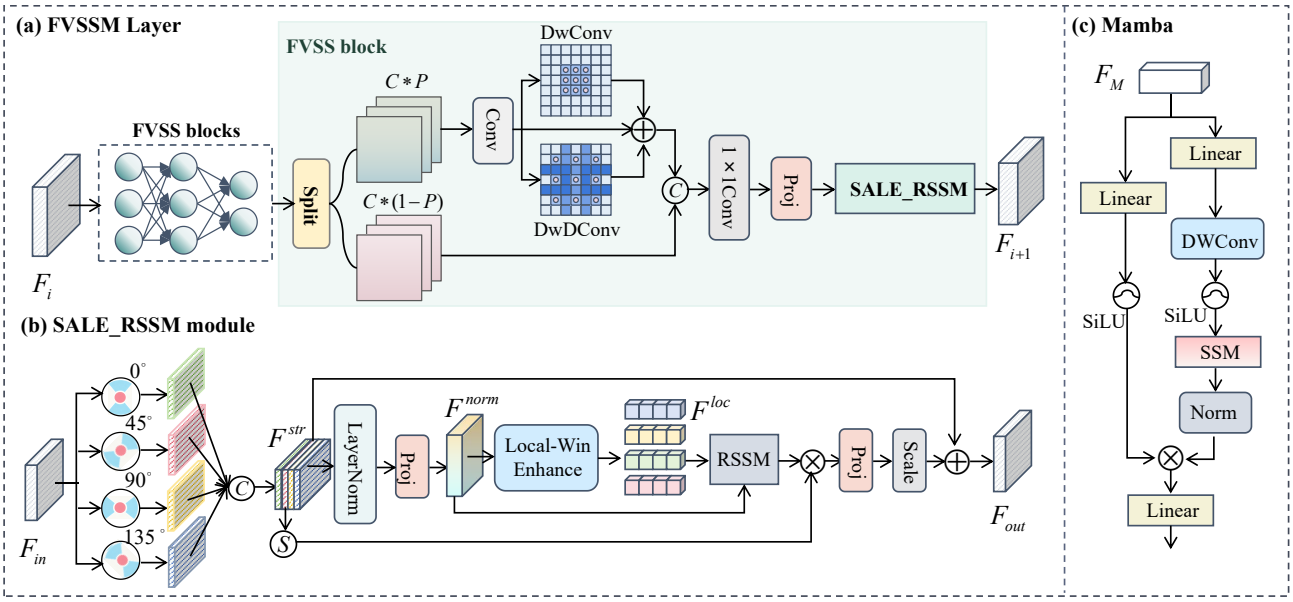

Figure 3. The overall structure of the FVSSM layer and vision mamba block.

The FVSSM layer comprises a series of fast visual state space blocks (FVSSBs), and a FVSSB is the basic feature extraction unit. Combined with partial channel transformation (PCT) and a structure-aware local-enhanced residual state space module (SALE_RSSM), this block acquires a powerful local structural feature representation capability with lighter parameters, effectively capturing the features of small lesions and suppressing background noise. Unlike simple stacked directional convolution and global mamba sequence scanning, SALE_RSSM generates structural gating weights from directional features to constrain the local window-based residual SSM, reducing redundant calculation and strengthening edge structural perception, which constitutes a hybrid structure-stateful modeling paradigm.

As depicted in Figure 3(a), the FVSSB first splits the channels of the input feature map F_i into two parts (F_i^P and F_i^{1-P}) and applies partial channels for identity mapping while the others are used to extract multi-scale context information using the three-branch module. Then, the preliminarily aggregated

output results from the three-branch module concatenate with the shortcut branch to restore the channel number as the F_i, and a 1×1 convolution f_1 is utilized to enable feature interaction among all channels from two branches, as follows:

$$\begin{aligned} F_i^{P1} &= f_{Dw}(f_3(F_i^P)) + f_I(f_3(F_i^P)) + f_{DwD}(f_3(F_i^P)), \\ F_o^C &= Cat(F_i^{1-P}, F_i^{P1}), \end{aligned} \tag{3.4}$$

where $Cat()$ denotes the concat operation, f_{Dw} is a 3×3 depthwise convolution (DwConv) to extract the local contextual information, f_I denotes the identity mapping to retain abundant spatial detail, and f_{DwD} is the dilated convolution to the DwConv with an assigned dilation rate to build the depthwise-dilated convolution (DwDConv) for generating a larger receptive field. Subsequently, a 1×1 convolution and channel projection operation transforms feature F_o^C, generating the input feature F_{in} for the SALE_RSSM.

Figure 3(b) shows the detailed structure of the SALE_RSSM. Four parallel 3×3 depthwise convolutions are designed to focus on four directions: 0°, 45°, 90°, and 135°. Each convolution kernel is initialized to enhance the response of the target direction. The feature maps extracted by the four directional convolutions are denoted as F_{0°, F_{45°, F_{90°, and F_{135°, respectively. The four directional feature maps are concatenated along the channel dimension and then fused into a single structure-aware feature map (F^{str}) through a 1×1 convolution and the Gaussian error linear unit (GELU) activation function. The channel-wise mean of F^{str} is calculated and activated by the sigmoid function to generate a structure weight map (W^{str}), which is used to provide structural priority guidance for subsequent feature modulation:

$$\begin{aligned} F^{str} &= GELU(f_1(Cat(F_{0^\circ}, F_{45^\circ}, F_{90^\circ}, F_{135^\circ}))), \\ W^{str} &= f_\sigma(Mean(F^{str})). \end{aligned} \tag{3.5}$$

The feature map F^{str} is then normalized by LayerNorm and projected through a linear layer to improve the feature expression ability. The subsequent local window enhancement (LWE) module processes the resulting feature F^{norm} by dividing it into non-overlapping windows for local context extraction. This LWE process sequentially involves rearranging F^{norm} for convolution adaptation, applying local window convolution (kernel 4, stride 4) across non-overlapping windows, and utilizing bilinear upsampling to restore the original spatial dimension, thereby yielding the localized feature F^{loc}. After that, the residual state space module (RSSM) replaces the self-attention mechanism with a linear state-space model (SSM) that can capture long-range dependencies in linear time complexity, while incorporating local augmentation features through residual connections. Specifically, a learnable parameter A is used to control the state update, and the local enhanced feature F^{loc} is fused into the state space calculation. Then, two linear layers B and C are used to implement gated activation, and a parameter D is used for direct mapping to retain the original feature information:

$$\begin{aligned} A_\sigma &= f_\sigma(A[None, None, None, :]), \\ F_{state} &= F^{str} \times A_\sigma + F^{loc}, \\ F_B &= Linear(F_{state}, W_B), \\ F_C &= Linear(F_{state}, W_C), \\ F_{ssm} &= F_B \times Tanh(F_C) + D \times F_{state}, \end{aligned} \tag{3.6}$$

where W_B and W_C are the weight parameters of linear layers B and C respectively. The rearranged structure weight map W^{str} is multiplied by F_{ssm} to realize structure-guided local feature modulation. The

resulting feature is projected through the linear layer, and a residual connection is used to superimpose with the input feature F^{norm} to ensure the continuity and integrity of local features (F_{LE}). Finally, the structure-aware feature F^{str} and the local-enhanced feature F_{LE} are concatenated along the channel dimension, and then two 1×1 convolution and GELU activation are used to complete feature fusion and dimension compression, restoring the number of channels to the input channel dimension.

3.1.3. Kalman-mamba layer

The Kalman-mamba layer is designed as the final processing stage of the encoder, which integrates the optimal state estimation capability of the Kalman filter (KF) with the efficient long-sequence modeling of vision mamba. The mamba is responsible for modeling long-range spatial dependencies by selective state compression over the flattened image spatial sequence, generating the features F_{VMB}. The Kalman filtering layer regards the features as noisy observations $\{z_k\}$, solving for the optimal state estimate $\hat{s}_{k|k}$ under the Bayesian minimum mean square error criterion, effectively introducing explicit probabilistic noise modeling and statistical smoothing constraints into the mamba's feature extraction process.

As illustrated in Figure 1, the input feature F_3 is first projected to the target dimension via a 1×1 convolution. It then passes through a sequence of vision mamba blocks (VMBs) with residual connections to capture long-range dependencies. The detailed structure of the VMB is shown in Figure 3(c). It takes the feature F_M as the input and channels it into two parallel branches. In the first branch, the feature F_M is input into a linear layer, a DWConv layer f_{Dw}, and a sigmoid linear unit (SiLU) activation function f_{SiLU}, followed by the state space model (SSM) f_{SSM} and LayerNorm f_{LN}. In another branch, the module expands the feature channels using a linear layer, followed by the SiLU activation function. The above process can be formulated as:

$$\begin{aligned} M_1 &= f_{LN}(f_{SSM}(f_{SiLU}(f_{Dw}(Linear(F_M))))), \\ M_{out} &= Linear(M_1 \cdot f_{SiLU}(Linear(F_M))). \end{aligned} \tag{3.7}$$

The resulting feature $F_{VMB} \in \mathbb{R}^{C' \times H' \times W'}$ from the vision mamba blocks (VMBs) is subsequently refined by the Kalman filter layer (KFL) to suppress noise. Rather than a sequential temporal filter, our KFL is architecturally designed as a parallel Bayesian shrinkage operator that operates across all spatial positions simultaneously. We let the spatial positions be flattened into a sequence indexed by $k = 1, 2, \ldots, L$ ($L = H' \times W'$). For each position k, the feature F_{VMB} provides a context-rich measurement via a learnable linear projection: $z_k = \text{Linear}(F_{VMB,k}) \in \mathbb{R}^D, k = 1, \ldots, L$, where $D = C'/4$.

Distinct from conventional recursive Kalman filters that propagate states over time, our parallel implementation imposes a shared Gaussian prior over all spatial positions for a single inference step. Specifically, the prior state is set to zero ($\hat{s}_{k|0} = 0$) and the prior covariance is initialized as a diagonal matrix $P_{k|0} = p_0 I$ (with a constant $p_0 = 0.01$). The process noise covariance Q_k and measurement noise covariance R_k are derived from learnable log-diagonal parameters to ensure non-negativity. The prediction (prior) step incorporates the learnable scalar transition coefficients $\tau \in \mathbb{R}^D$:

$$\hat{s}_{k|k-1} = \tau \odot \hat{s}_{k|0} = 0, \quad P_{k|k-1} = \tau^2 \odot P_{k|0} + Q_k, \tag{3.8}$$

where $\odot$ denotes element-wise multiplication. Since the initial prior state is zero, the predicted state remains zero, and the prediction step primarily serves to modulate the prior covariance via τ.

Subsequently, the measurement update (correction) is executed in a fully parallelized manner across all L positions. The Kalman gain $K_{Gain} \in \mathbb{R}^D$ is computed element-wise and shared globally across all spatial positions to maintain efficiency:

$$K_{Gain} = P_{k|k-1} \oslash (P_{k|k-1} + R_k), \tag{3.9}$$

where $\oslash$ denotes element-wise division. The posterior state update and covariance update are then derived as:

$$\hat{s}_{k|k} = \hat{s}_{k|k-1} + K_{Gain} \odot (z_k - \hat{s}_{k|k-1}), \tag{3.10}$$

$$P_{k|k} = (\mathbf{1} - K_{Gain}) \odot P_{k|k-1}. \tag{6}$$

This process reveals the synergy between the two modules: the mamba blocks generate the measurement likelihood $p(z_k|s_k)$ by capturing long-range spatial dependencies, while the KFL performs a single-step Bayesian maximum a posteriori estimation to produce the posterior state $\hat{s}_{k|k}$. The Kalman gain K_{Gain} effectively acts as an adaptive, learned shrinkage matrix. It attenuates feature components according to the ratio of process uncertainty to total uncertainty ($P_{k|k-1}/(P_{k|k-1} + R_k)$), thereby suppressing noise while preserving salient contextual cues extracted by the mamba. The globally shared covariance $P_{k|k}$ regularizes the entire feature map, preventing overfitting to local anomalies. Finally, the updated state $\hat{s}_{k|k}$ is projected back to the original channel dimension C' via a linear layer, reshaped to spatial dimensions (H', W'), and passed through a 3×3 convolution with group normalization and GELU activation to enhance local spatial correlations.

3.2. Boundary-enhanced cascade decoder

3.2.1. Dual-attention concatenation module

The dual-attention concatenation (DAC) module aims to integrate the feature information of different levels and refine feature maps. Instead of naive concatenation or additive skip connection, this module adaptively generates positive activation and negative suppression weights to filter invalid background features across scales, achieving lightweight cross-layer feature fusion.

As shown in Figure 1, the DAC module takes the shallow feature F_i and deep feature F_{i+1} as inputs. The deep feature F_{i+1} is compressed by the 1×1 pointwise convolution and the compressed feature map F_i^c is fed into an attention mask implemented by the sigmoid function f_σ. The attention-guided map is further converted to its opposite and a 1×1 convolution is adopted. Then, we apply a linear transformation to amplify non-linearity and learn channel-wise correlations and dependencies within the response scores. The reweighted scores are then added to the F_i^c to get the accurate response F_i^r, which can be written as follows:

$$F_i^r = f_1(-f_\sigma(F_i^c)) \cdot f_\sigma(F_i^c) + F_i^c. \tag{3.11}$$

At the same time, the shallow feature F_i that is rich in texture information is fed into the sigmoid function to convert the feature into probabilities of possible labels for each pixel. Subsequently, we combine the output of the two branches by the element-wise product and apply a 3×3 depthwise convolution f_{Dw} for further integrating local information, as follows:

$$F_i^{DA} = f_{Dw}(f_\sigma(F_i) \cdot F_i^r). \tag{3.12}$$

3.2.2. RD-BE module

The refined differential and boundary-enhanced (RD-BE) module is designed to enhance edge delineation and improve overall segmentation accuracy. We decouple edge and body features in the middle of the network for separate optimization and adopt a contour point graph convolutional network (GCN) for boundary fine-tuning.

As depicted in Figure 1, the deep feature F_D, which captures the rich contextual information of the input image, is concatenated with the high-resolution refined feature F_{fine}. A convolution operation is then applied to this concatenated representation, resulting in a coarse edge feature, denoted as F_{edge}. F_{edge} has rich detailed information and can help outline the boundaries between different objects or regions with improved clarity. Next, we obtain the residual feature F_{res} by subtracting F_{edge} from F_D. To further refine this non-edge information, we adopt another fine-grained feature F_3 generated from the backbone to enhance F_{res} and get body feature F_{body}. This procedure can be formulated as:

$$\begin{aligned} F_{edge} &= f_3(Cat(F_{fine}, F_D)), \\ F_{body} &= f_3(Cat(F_D - F_{edge}, F_3)). \end{aligned} \tag{3.13}$$

The body feature F_{body} and the edge feature are then integrated to compute the merged feature F_{merge}, referred to as $F_{merge} = [F_{body}, F_{edge}]$. This merged representation leverages both detailed structural information from edges and robust contextual information from body features, ultimately leading to improved segmentation performance.

To enforce finer boundary delineation on F_{merge}, we further apply an efficient graph convolution network (GCN) module. As illustrated in Figure 1, given the features F_{edge} and F_{merge}, we first transform F_{edge} to the edge prediction F^e_{seg} and select at most K points with the highest confidence in F^e_{seg}. This process ensures that we focus on the most salient features that are likely to define the boundary accurately. Once these key points are identified, we use the indices of these points to sample the corresponding K point features from F_{merge}, thereby building the graph structure. Assuming a fully connected graph (all K points are connected), the graph convolution operation can be performed as a fully connected layer. The adjacency matrix, which represents the connectivity between the K points, is then computed using a 1×1 convolutional layer. After that, we combine the output point feature and F_{merge}, and obtain the refined segmentation feature F_{Seg}.

3.3. Loss function

The segmentation loss function consists of three parts: edge loss L_{edge}, body loss L_{body}, and merge loss L_{merge} to obtain the predictions of edge, body, and merge parts, respectively. This joint loss to supervise them simultaneously can be formulated as:

$$L = \lambda_1 L_{edge}(F^e_{seg}, G^e_{seg}) + \lambda_2 L_{body}(F^b_{seg}, G^b_{seg}) + \lambda_3 L_{merge}(F^m_{seg}, G^m_{seg}), \tag{3.14}$$

where $\lambda_1 = 1, \lambda_2 = 1$, and $\lambda_3 = 2$ are the balance parameters. F^e_{seg}, F^b_{seg}, and F^m_{seg} denote the predictions of edge part, body part, and merge part, respectively. G^m_{seg} represents the ground truth. G^e_{seg} is the ground truth of the edge part, which is generated from G^m_{seg}. When generating the ground truth of body part G^b_{seg}, the corresponding position of the positive pixel of G^e_{seg} in G^b_{seg} is marked to be ignored and the other pixels of G^b_{seg} are the same as G^m_{seg}. L_{edge} is the Dice loss, and we combine the Dice loss and the

Focal loss to introduce the joint loss function for defining L_{body} and L_{merge} [31], which can be written as follows:

$$L_{merge}(F^m_{seg}, G^m_{seg}) = \mu_1(1 - \frac{1}{T}\sum_{t=0}^{T-1}\frac{y^{tp}(t)}{y^{tp}(t) + y^{fn}(t) + y^{fp}(y)}) - \mu_2\frac{1}{N}\sum_{t=0}^{T-1}\sum_{j=1}^{N} a(t)G^m_{seg}(j,t)(1 - F^m_{seg}(j,t))^{\gamma}F^m_{seg}(j,t)), \tag{3.15}$$

where $\mu_1 = 0.5$ and $\mu_1 = 0.5$ denote balance weights, $y^{tp}(t)$, $y^{fn}(t)$, and $y^{fp}(t)$ denote true positives, false negatives, and false positives for class t and T is the number of categories. N is the total number of pixels in the image. $G^m_{seg}(j,t)$ and $F^m_{seg}(j,t)$ represent the probability for pixel i being of class c in the ground truth and predicted mask, respectively. $a(t) \in [0,1]$ is the weighting factor and $\gamma = 3$ denotes a tunable focusing parameter.

4. Experiments and results

4.1. Dataset

The dataset used in this study includes a total of 2538 endoscopic images collected from Xiangya Hospital in China, consisting of two independent private subsets with separate task applications. As shown in Figure 4, the two datasets include polyps with varying sizes, shapes, and lesion locations, ensuring sufficient data diversity and representation, which can fully verify the comprehensive performance of the segmentation models. As illustrated in Table 1, the first private dataset is exclusively used for the binary classification task (adenomas vs. non-adenomas), containing a total of 1335 images (645 adenomatous polyp images and 690 non-adenomatous polyp images). This dataset only contains binary category labels without fine-grained adenoma subtype information. The second private dataset is used for the polyp subtype classification task (tubular adenomas vs. villous adenomas vs. non-adenomas), consisting of 1203 images, including 352 tubular adenoma images, 187 villous adenoma images, and 664 non-adenomatous polyp images. The two private datasets serve separate experimental tasks and are processed independently without merging. To ensure patient privacy, all personally identifiable information was removed from the endoscopic images. This study was approved by the Ethics Committee of Xiangya Hospital of Central South University, and informed consent was signed by all participants. All the endoscopic images were annotated by a team of three gastroenterologists with over 10 years of experience in the field and reviewed by a gastroenterologist with more than 25 years of experience, ensuring the highest level of accuracy in the annotations. The diagnostic gold standard for adenomatous and non-adenomatous polyps was established through pathological biopsy, a process conducted and interpreted by three pathologists, each with over 20 years of expertise in the field. The diagnostic gold standard for polyp classification (non-adenomatous vs. adenomatous) and adenoma subtyping was determined via pathological biopsy, with all biopsy results interpreted and confirmed by three pathologists with over 20 years of pathological diagnosis experience.

Due to the different resolutions of the collected data, all the endoscopic images are resized to 512×512 as the inputs of the proposed model. The median filter and total variation filter are then used to reduce noise and enhance image quality. The processed endoscopic images are randomly split into three independent subsets: 70% for the training set, 10% for the validation set, and 20% for the testing set. The detailed partitioning statistics of the two independent datasets are summarized in Table 1.

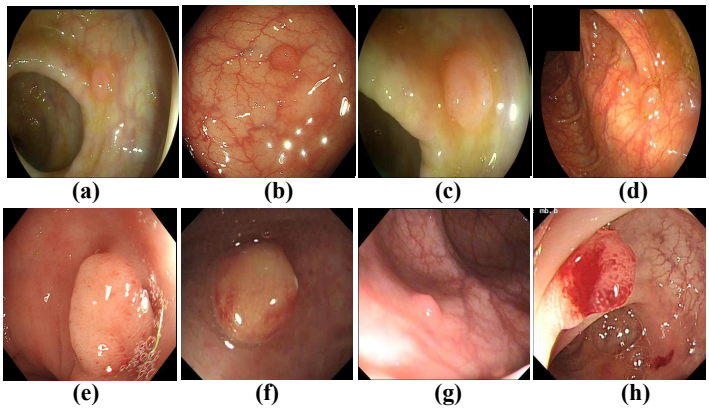

Figure 4. Examples of collected polyp images.

Table 1. Detailed characteristics of the evaluated datasets.

Classes	Total images	Training set		Validation set		Test set	
		Images	Cases	Images	Cases	Images	Cases
Adenomas vs. Non-Adenomas	1335	934	839	134	120	267	240
Adenomas	645	437	392	59	53	149	135
Non-Adenomas	690	497	447	75	67	118	105
Tubular vs. Villous vs. Non-Adenomas	1203	842	747	121	107	240	213
Tubular	352	256	230	40	37	56	51
Villous	187	131	111	15	12	41	38
Non-Adenomas	664	455	406	66	58	143	124

4.2. Evaluation metrics

To fairly compare the polyp segmentation methods, we report the quantitative results by ten evaluation metrics, including intersection over union (IoU), pixel accuracy (Acc), Dice coefficient (Dice), Hausdorff distance (HD), average surface distance (ASD), and sensitivity (Sens), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), and F1-score (F1). These metrics comprehensively evaluate segmentation performance from multiple dimensions: IoU and Dice quantify the overlap between predicted segmentation masks and ground truth labels; pixel accuracy reflects the overall pixel-wise classification correctness; and HD and ASD measure the boundary and surface distance error to characterize the contour-fitting quality of segmented lesions. Meanwhile, sensitivity, specificity, PPV, NPV, and F1-score jointly assess the model's ability to correctly identify polyp regions and background areas. For pairwise comparisons between our proposed method and other models, we conduct a paired Student's t-test with a significance level set to $\alpha = 0.05$. A p-value lower than 0.05 indicates a statistically significant performance gap between the two compared approaches on the corresponding evaluation metric.

4.3. Implementation details

All experiments were implemented in the Pytorch framework and the networks were trained with NVIDIA 3090ti GPUs. All the endoscopic images were resized to $512 \times 512 \times 3$ pixels. The proposed model was trained by the stochastic gradient descent (SGD) optimizer with a batch of 2, a momentum

of 0.9, and a weight decay of 5e-4. The initial learning rate was set to 0.01, and the 'Poly' learning rate policy with factor $(1 - \frac{epoch}{total\_epoch})^{0.9}$ was performed for training our framework. The training was conducted over a total of 180 epochs. In the training phase, Gaussian blur augmentation and bilateral blur augmentation were employed to enrich the image dataset, and data augmentation (such as shifting, tilting, scaling, random clipping, and flipping) was also used to optimize the network's generalization capability.

4.4. Performance of BKMNet

Table 2. Segmentation results for adenomas vs. non-adenomas and specific adenoma subtypes.

Classes	IoU (%)↑	Dice (%)↑	Acc (%)↑	HD↓	ASD↓	Sens (%)↑	Spec (%)↑	PPV (%)↑	NPV (%)↑	F1 (%)↑
Adenomas	89.85 ± 0.85	91.10 ± 0.65	91.78 ± 0.07	6.60 ± 0.03	0.65 ± 0.001	90.94 ± 0.74	99.92 ± 0.02	97.44 ± 0.61	99.72 ± 0.03	94.08 ± 0.65
Non-Adenomas	85.70 ± 0.64	86.85 ± 0.37	89.16 ± 0.02	9.24 ± 0.04	0.92 ± 0.002	88.70 ± 0.35	99.95 ± 0.01	95.08 ± 0.41	99.88 ± 0.01	91.78 ± 0.37
Mean	91.31 ± 0.59	92.51 ± 0.34	93.61 ± 0.05	5.31 ± 0.03	0.53 ± 0.001	93.17 ± 0.33	96.74 ± 0.20	97.37 ± 0.34	98.80 ± 0.19	95.19 ± 0.34
Tubular	87.14 ± 0.03	88.38 ± 0.02	90.79 ± 0.02	8.14 ± 0.04	0.82 ± 0.001	90.78 ± 0.02	99.96 ± 0.01	95.60 ± 0.02	99.90 ± 0.01	93.13 ± 0.02
Villous	88.70 ± 0.83	92.46 ± 0.58	92.67 ± 0.02	2.40 ± 0.02	0.21 ± 0.002	92.33 ± 0.85	99.93 ± 0.01	95.75 ± 0.29	99.88 ± 0.02	94.01 ± 0.58
Non-Adenomas	86.41 ± 0.50	87.12 ± 0.29	90.19 ± 0.02	10.61 ± 0.04	0.91 ± 0.002	90.11 ± 0.16	99.91 ± 0.01	94.25 ± 0.43	99.84 ± 0.01	92.13 ± 0.29
Mean	90.29 ± 0.38	91.83 ± 0.22	93.36 ± 0.02	5.34 ± 0.03	0.51 ± 0.001	93.27 ± 0.25	98.05 ± 0.04	96.32 ± 0.18	99.03 ± 0.01	94.76 ± 0.22

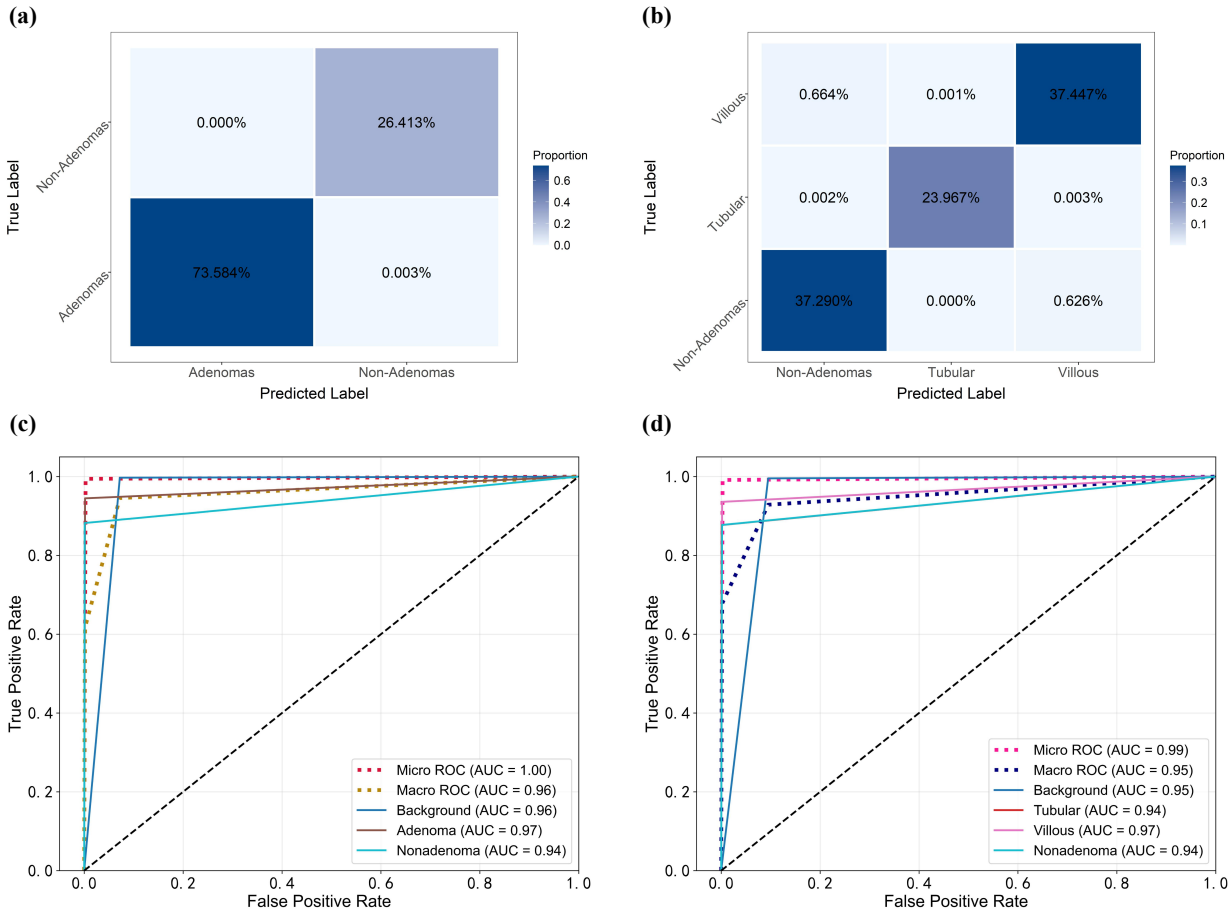

Figure 5. The confusion matrices and ROC curves of the polyp segmentation results. (a) and (c): Adenomas vs. Non-Adenomas; (b) and (d): Tubular vs. Villous vs. Non-Adenomas.

To objectively evaluate the performance of the proposed BKMNet, we conducted extensive testing on the independent dataset of images and five-fold cross-validation was performed. Several widely recognized evaluation metrics to quantify the model's segmentation accuracy and precision were used, and the segmentation results for adenomas versus non-adenomas and specific adenoma subtypes are presented in Table 2.

These results demonstrated that the proposed model can accurately identify and delineate colorectal polyp regions, as evidenced by high mean IoU (91.31%), Dice (92.51%), and accuracy (93.61%) across all classes. Adenomas are segmented with notably higher performance than non-adenomas, yielding an IoU of 89.85% vs. 85.70%, a Dice of 91.10% vs. 86.85%, a lower Hausdorff distance (6.60 vs. 9.24), and an average symmetric surface distance (0.65 vs. 0.92). The sensitivity for adenomas (90.94%) is slightly superior to that for non-adenomas (88.70%), while specificity remains exceptionally high (>99.9%) for both, indicating a very low false-positive rate. The positive predictive value (PPV) and negative predictive value (NPV) are also consistently high, with adenomas showing a PPV of 97.44% and an NPV of 99.72%. These results demonstrate that the proposed model can accurately identify and delineate colorectal polyp regions, showing the remarkable ability to distinguish between adenomatous and nonadenomatous polyps.

When further examining specific adenoma subtypes, villous adenomas achieve the best overall segmentation quality among the three groups, with an IoU of 88.70%, Dice of 92.46%, and a remarkably low HD of 2.40 and ASD of 0.21, suggesting excellent boundary delineation. Tubular adenomas, while still well-segmented, show slightly lower metrics (IoU 87.14%, Dice 88.38%, HD 8.14) compared to villous ones. Non-adenomas in this subtype-specific analysis exhibit the lowest performance (IoU 86.41%, Dice 87.12%, HD 10.61), yet still maintain high specificity and NPV. The mean values across the three subtypes (IoU 90.29%, Dice 91.83%, Acc 93.36%) are comparable to the binary classification results, though the HD is slightly higher (5.34 vs. 5.31) and ASD is marginally lower (0.51 vs. 0.53). Overall, the high IoU and Dice across all classes, coupled with low distance metrics, demonstrate the model's superiority in both tissue classification and spatial segmentation. Its capacity to distinguish adenoma subtypes with varying complexities highlights its potential for clinical applications.

The results of the receiver operating characteristic (ROC) analysis and confusion matrix provide additional quantitative evidence for the efficacy of our segmentation model. As shown in Figure 5(a),(c), in the classification task of Adenomas vs. Non-adenomas, the model achieved an area under the curve (AUC) of 0.97 for adenomas and 0.94 for non-adenomas, demonstrating its robust ability to differentiate between the two classes. The macro-average AUC of 0.96 reflects balanced performance across both categories, indicating that the model does not exhibit significant bias toward either class. The micro-average AUC, which is close to 1.0, suggests an overall high level of precision in predicting class labels. Further quantitative results of the confusion matrix validate the model's reliable segmentation performance by intuitively reflecting pixel-level prediction correctness. For the multi-class classification task involving tubular, villous, and non-adenomas, as shown in Figure 5(b),(d), the model performance was equally impressive. The AUC values for villous and tubular adenomas were 0.97 and 0.94, respectively, highlighting their excellent discriminative power for these subtypes. Even for non-adenomas, the model achieved an AUC of 0.94, consistent with the binary classification results. The macro-average AUC of 0.95 across all three classes further validates the model's consistent performance, regardless of the specific tissue type. The micro-average AUC of 0.99 again underscores the model's exceptional ability to accurately classify samples, effectively capturing the distinct

morphological features of each class. The ROC curves and confusion matrices, in conjunction with the segmentation performance metrics, provide comprehensive and compelling evidence for the reliability and effectiveness of our model in distinguishing between various adenoma subtypes and non-adenomatous tissues.

4.5. Ablation studies

Table 3. The results of the ablation study for the proposed BKMNet.

Methods	Adenomas vs. Non-Adenomas		Tubular vs. Villous vs. Non-Adenomas		Parameter	Flops (G)	P-value
	IoU (%)	Acc (%)	IoU (%)	Acc (%)			
BKMNet w/o MFCA	91.04 ± 0.46	93.16 ± 0.02	90.01 ± 0.31	93.03 ± 0.03	5.24M	17.49	0.1971
BKMNet w/o SALE_RSSM	83.30 ± 0.57	85.76 ± 0.04	82.93 ± 0.48	85.12 ± 0.05	2.13M	13.84	0.0012
BKMNet w/o Kalman-mamba	88.65 ± 0.36	92.30 ± 0.05	88.01 ± 0.26	91.25 ± 0.02	4.77M	17.42	0.0025
BKMNet w/o DAC	90.85 ± 0.27	93.01 ± 0.01	89.90 ± 0.44	93.06 ± 0.02	5.23M	17.88	0.0433
BKMNet w/o RD-BE	89.40 ± 0.33	92.81 ± 0.02	88.68 ± 0.28	91.31 ± 0.03	5.06M	15.35	0.0045
BKMNet	**91.31 ± 0.59**	**93.61 ± 0.05**	**90.29 ± 0.38**	**93.36 ± 0.02**	**5.28M**	**17.95**	–

"w/o" denotes to remove the specific module.

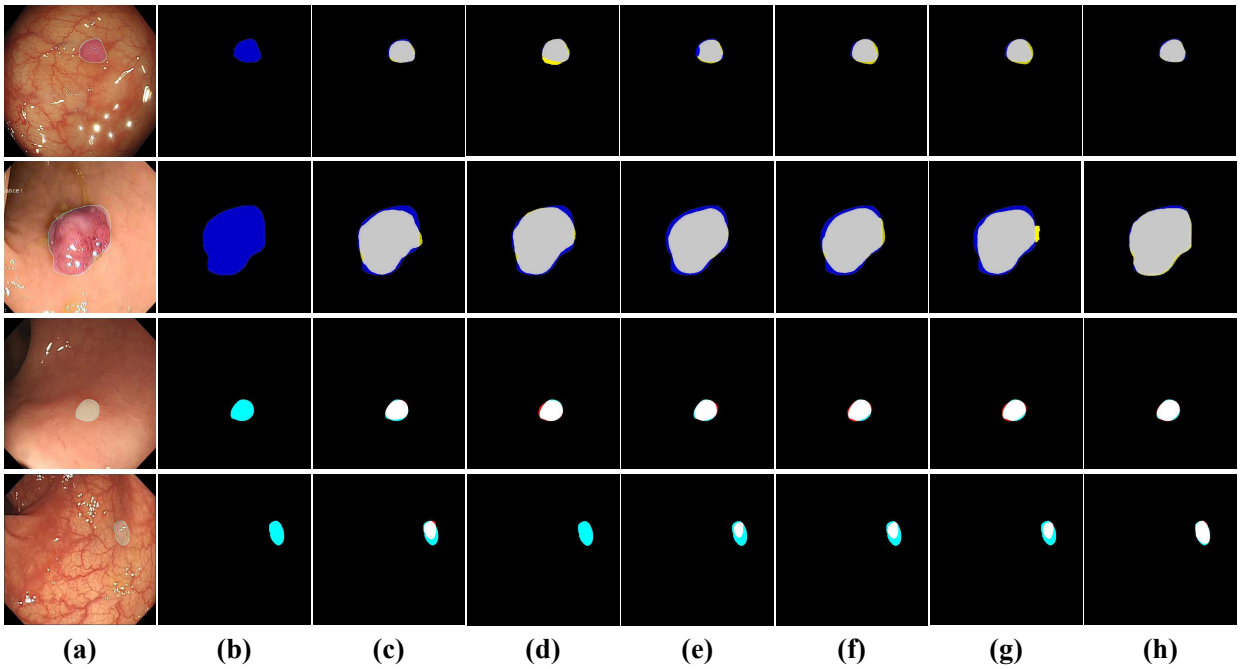

Figure 6. Visual comparison of our ablation studies. (a) Input image; (b) ground truth; (c) BKMNet w/o MFCA; (d) BKMNet w/o SALE_RSSM; (e) BKMNet w/o Kalman-mamba; (f) BKMNet w/o DAC; (g) BKMNet w/o RD-BE; (h) BKMNet. Blue and yellow pixels indicate the ground truth and predictions of adenomatous polyps, respectively. Light gray pixels represent the overlapping regions between the prediction and ground truth. Cyan and red pixels indicate the ground truth and predictions of non-adenomatous polyps, respectively.

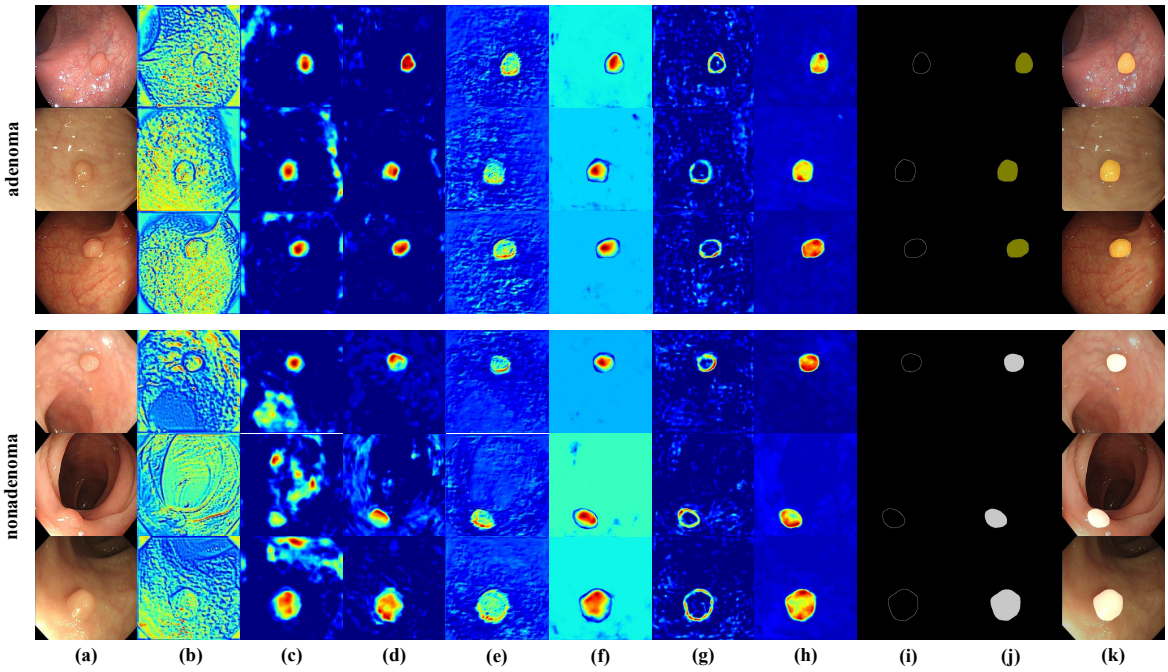

Figure 7. Visualization results of adenomas and non-adenomas. (a) is the input image, (b) is F_1, (c) is F_3, (d) is F_D, (e) is F_{fine}, (f) is F_{body}, (g) is F_{edge}, (h) is F_{Seg}, (i) is the edge prediction, (j) is the prediction and (k) is the overlapping image.

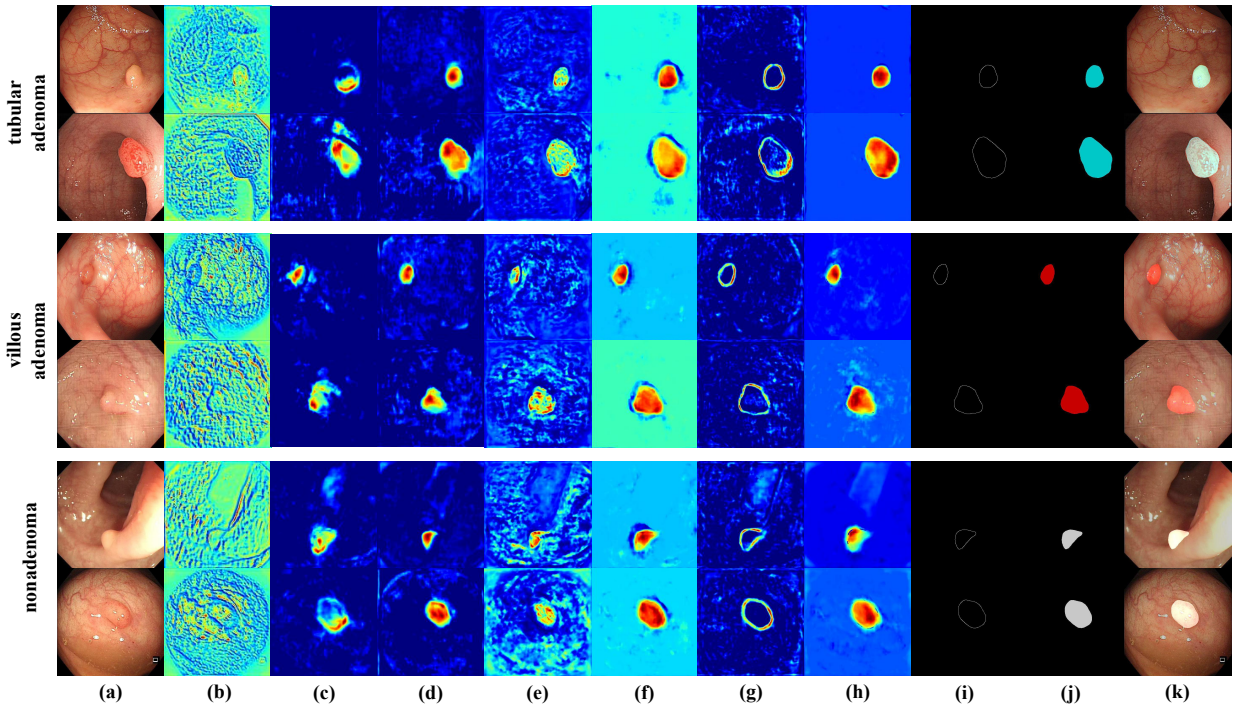

Figure 8. Visualization results of tubular adenomas, villous adenomas, and non-adenomas. (a) is the input image, (b) is F_1, (c) is F_3, (d) is F_D, (e) is F_{fine}, (f) is F_{body}, (g) is F_{edge}, (h) is F_{Seg}, (i) is the edge prediction, (j) is the prediction, and (k) is the overlapping image.

The proposed BKMNet mainly benefits from five key components, including the MFCA block, SALE_RSSM module, Kalman-mamba layer, DAC module, and RD-BE module. Here, we carry out a series of ablation studies to explore the efficacy of each module. The comparison of different configurations was tabulated in Table 3 in terms of trainable parameters, theoretical floating point operations (Flops), and segmentation performance (IoU and Accuracy) for polyps, where "w/o" means to remove the specific module, and values in bold indicate the results of BKMNet. Figure 6 presents the segmentation mask comparison of different model configurations on representative colonoscopic images, including the input image, ground truth (GT), segmentation results of the complete BKMNet model, and results of the five ablation models. To further reveal the internal mechanism of these key modules in optimizing polyp feature representation, Figures 7 and 8 show the feature map visualization results of BKMNet after sequential processing by these modules.

(1) Effectiveness of the MFCA module: This module is a shallow feature enhancement module designed to capture complementary spatial context and multi-frequency channel features at the initial encoding stage, mining texture information of polyp lesions. Quantitative results in Table 3 validate its practical effectiveness. After removing the MFCA module, the complete BKMNet suffers a consistent performance drop on both segmentation subtasks: the IoU declines from 91.31% to 91.04% for adenomas vs. non-adenomas and from 90.29% to 90.01% for tubular vs. villous vs. non-adenomas, accompanied by a slight accuracy reduction. In terms of computational cost, the model without MFCA achieves marginal parameter and GFLOPs reduction (5.24 M vs. 5.28 M, 17.49 vs. 17.95), indicating that MFMSAttentionBlock brings negligible computational overhead but steady performance improvement. Visual results in Figure 6 further support the effectiveness of the MFCA module. The output features F_1 of MFCA are shown in Figures 7(b) and 8(b), which contain a wealth of texture information. By adaptively activating multi-frequency channel responses and aggregating multi-scale spatial features, it enriches shallow feature representation, suppresses low-frequency background noise of intestinal tissues, and highlights high-frequency texture details of polyp surfaces.

(2) Effectiveness of the SALE_RSSM module: This module is verified as the most effective and indispensable core component of BKMNet, dominating the overall segmentation performance. Quantitatively, removing the SALE_RSSM module causes a catastrophic performance drop across all evaluation metrics. Specifically, the IoU drops by 8.01% (91.31% vs. 83.30%) and 7.36% (90.29% vs. 82.93%) on the two polyp segmentation tasks, respectively, and the overall accuracy decreases by nearly 8%. The results indicate that the prominent performance gain of the full model is mainly contributed to by SALE_RSSM. The extremely low P-value proves that the performance improvement brought by SALE_RSSM is statistically significant. Such dramatic quantitative degradation is fully consistent with the visual defects observed in Figure 6, strongly verifying the core effectiveness of SALE_RSSM. Corresponding feature map visualization in Figures 7(c) and 8(c) further reveals its working mechanism: SALE_RSSM significantly enhances the structural contour and local lesion response of polyps, sharply enlarging the feature discrepancy between polyp foreground and complex intestinal background. By integrating direction-aware convolution and a residual state space machine, it captures directional edge continuity of irregular polyps and strengthens local fine lesion features while suppressing background noise.

(3) Effectiveness of the Kalman-mamba layer: This layer is a highly effective high-level semantic enhancement module with excellent performance-cost balance. As illustrated in Table 3, without Kalman-mamba, the IoU drops to 88.65% and 88.01% for the two subtasks, and the accuracy decreases

to 92.30% and 91.25%, with a P-value of 0.0025, demonstrating remarkable statistical differences. Notably, this module only brings tiny increases in parameters (4.77 M vs. 5.28 M) and GFLOPs (17.42 vs. 17.95), realizing significant performance improvement at little computational cost. Qualitative segmentation results in Figure 6 further validate its superior performance. The feature map variations in Figures 7(d) and 8(d) also illustrate the effectiveness of this module. After Kalman-mamba layer processing, the scattered and ambiguous high-level features are optimized into concentrated, accurate, and semantically consistent activation regions.

(4) Effectiveness of the DAC module: This module exhibits prominent effectiveness and ultra-high efficiency in multi-scale feature aggregation. As shown in Table 3, after removing DAC, the model shows performance degradation, with IoU falling to 90.85% and 89.90% for the two segmentation tasks. The P-value less than 0.05 verifies that the performance gain brought by DAC is statistically valid. Moreover, the parameter quantity and GFLOPs remain almost unchanged compared with the full model, indicating that DAC achieves effective feature optimization without introducing redundant computational burden. Visual comparisons in Figure 6 illustrate the practical value of the DAC module. The hierarchical feature maps in Figures 7(e) and 8(e) further demonstrate its fusion superiority: different from traditional fixed skip connections that cause feature aliasing and semantic mismatch, DAC adaptively filters invalid background features and enhances effective lesion responses. It precisely fuses fine-grained shallow edge details and high-level semantic features, effectively compensating for feature loss during downsampling encoding.

(5) Effectiveness of the RD-BE module: To assess the RD-BE module, we replaced it with a full convolution module using F_D and F_{fine} as input to generate the mask. As shown in Table 3, removing this module causes obvious performance degradation, with IoU dropping to 89.40% and 88.68% for the two subtasks. The computational complexity decreases from 17.95 to 15.35 GFLOPs after removal, demonstrating that the moderate computational overhead introduced by edge extraction is fully worthwhile for the prominent performance improvement. Visual results in Figure 6 intuitively reflect the boundary optimization effectiveness of this module. The feature map visualization in Figures 7 and 8 further explains its unique working mechanism. This module decouples the feature maps into body and edge components, resulting in features F_{body} (f) and F_{edge} (g). Each pixel in the body component shares similar feature representations and embodies semantic information, while the edge features encompass more boundary details. The edge-aware graph convolution module was then employed to enhance global feature learning and improve the final prediction, effectively capturing the boundary characteristics of polyps. The segmentation features in (h) exhibit more precise localization and enhanced boundaries, while the prediction masks (j) are generated from these features. The predicted edge prior masks in (i) provide more accurate locations for the objects, offering better priors for identifying the most challenging pixels along the boundary.

4.6. Comparisons with state-of-the-art methods

Table 4. Comparison results with state-of-the-art methods on the two private datasets.

Methods	Parameter	Flops (G)	Mem (MB)	FPS (F/s)	Time (ms)	A vs. N		T vs. V vs. N		P-value
						IoU (%)	Dice (%)	IoU (%)	Dice (%)	
UNet [27]	31.04 M	218.99	118.44	83.55	11.94	85.63 ± 0.27	86.93 ± 0.13	84.42 ± 0.30	85.18 ± 0.15	< 0.005
EBLNet [31]	42.47 M	220.58	164.50	38.50	26.0	86.74 ± 0.59	87.66 ± 0.23	86.05 ± 0.44	87.23 ± 0.19	< 0.005
LETNet [36]	0.95 M	7.31	3.73	16.40	61.01	80.65 ± 0.65	81.96 ± 0.19	80.09 ± 0.42	81.68 ± 0.11	< 0.005
LCNet [28]	0.51 M	3.99	1.97	83.30	12.01	80.68 ± 0.47	81.09 ± 0.24	80.12 ± 0.22	81.09 ± 0.09	< 0.005
Mobile U-ViT [38]	1.39 M	10.03	5.32	51.31	19.51	87.89 ± 0.24	88.97 ± 0.17	86.54 ± 0.37	87.01 ± 0.21	< 0.005
VM-UNet [25]	27.43 M	16.46	104.63	26.60	37.66	90.18 ± 0.21	90.80 ± 0.11	88.11 ± 0.36	88.89 ± 0.12	0.0058
LightM-UNet [42]	5.01 M	2.83	19.12	24.22	41.40	88.37 ± 0.56	89.05 ± 0.22	86.76 ± 0.33	87.91 ± 0.19	< 0.005
AC-MambaSeg [43]	7.99 M	25.10	30.54	12.01	83.41	90.65 ± 0.55	91.02 ± 0.21	88.35 ± 0.37	89.22 ± 0.18	0.0126
BKMNet	**5.28 M**	**17.95**	**20.67**	**12.80**	**78.01**	**91.31 ± 0.59**	**92.51 ± 0.34**	**90.29 ± 0.38**	**91.83 ± 0.22**	–

A vs. N: Adenomas vs. Non-Adenomas; T vs. V vs. N: Tubular vs. Villous vs. Non-Adenomas. FPS: Frames Per Second; Time: Inference Time; Mem: Memory.

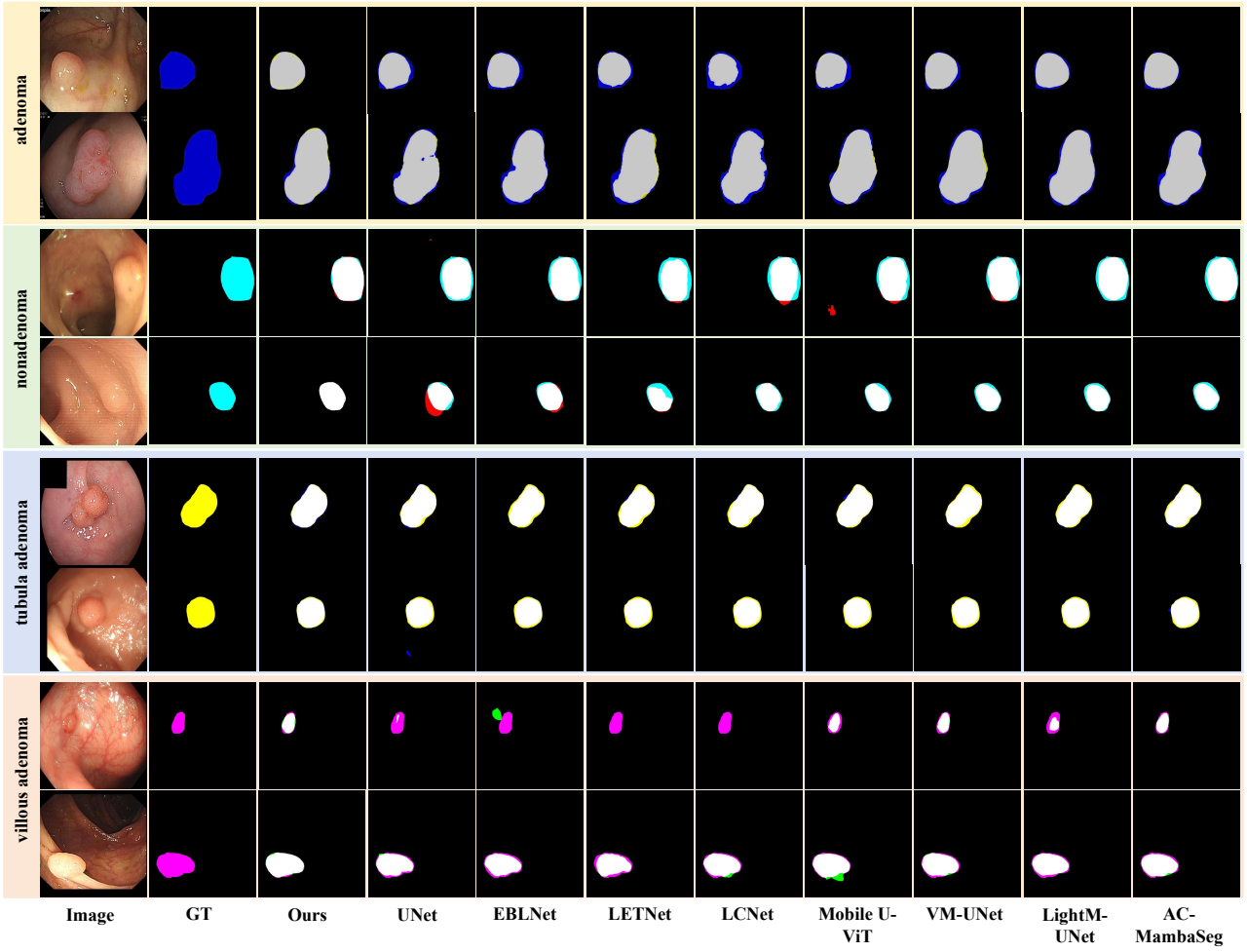

Figure 9. Qualitative comparisons with state-of-the-art methods.

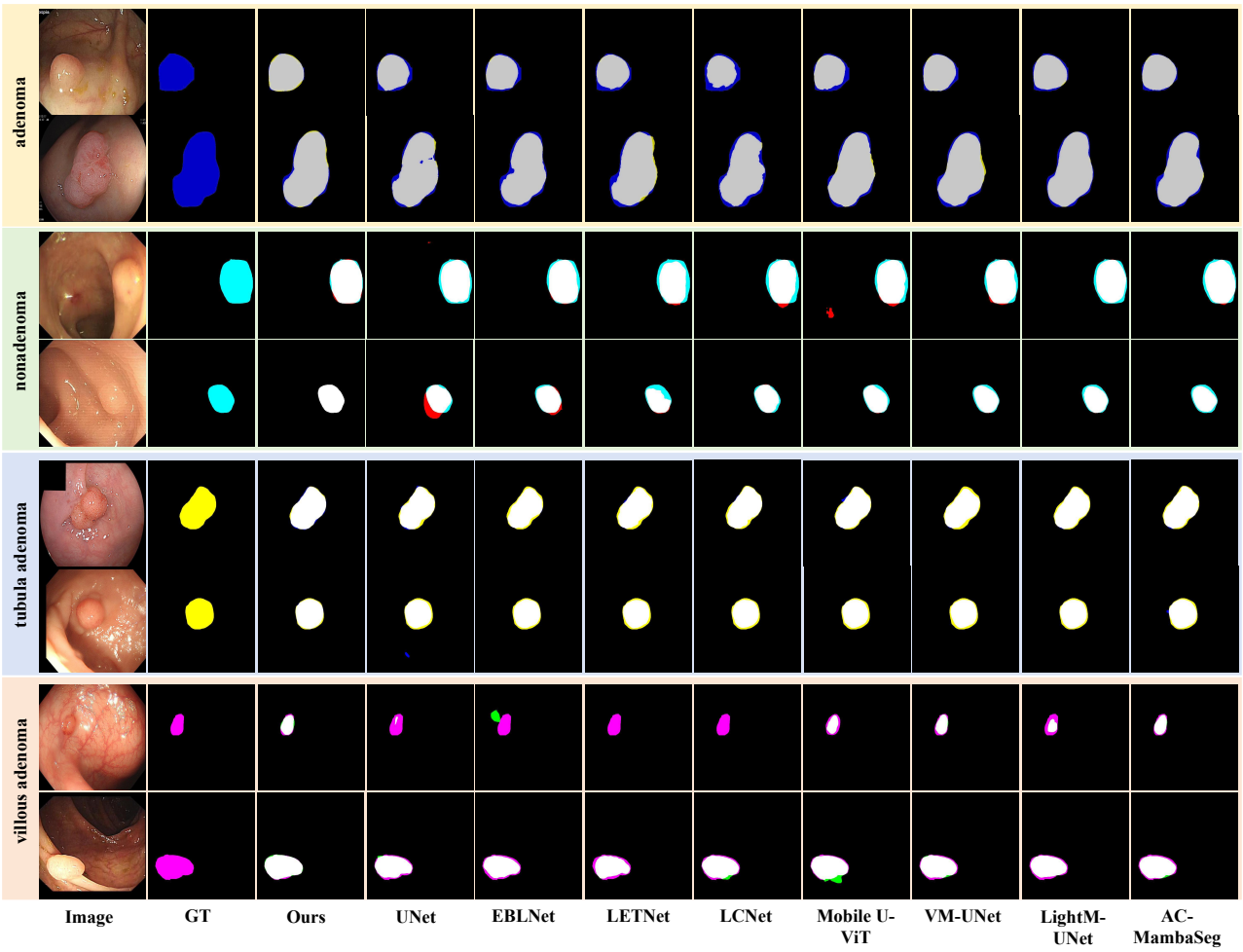

Figure 10. Boundary comparison results with state-of-the-art methods. Light gray pixels represent the overlapping regions between the prediction and ground truth.

To further evaluate the validity of the proposed BKMNet model, we selected some deep learning-based segmentation models such as UNet [27], EBLNet [31], LETNet [36], LCNet [28], Mobile U-ViT [38], VM-UNet [25], LightM-UNet [42], and AC-MambaSeg [43] for this comparison. To ensure a fair comparison, all competing models are trained and tested under identical experimental settings, including unified input image dimensions, consistent training hyperparameters and optimization strategies, identical data augmentation pipelines, and testing on the same private colorectal polyp datasets. Table 4 shows the comparative results on the private dataset, evaluating two critical segmentation tasks: distinguishing adenomas from non-adenomas, and subclassifying tubular, villous, and non-adenomatous lesions. The quantitative results in terms of segmentation performance (IoU, Dice), model complexity (parameter count, GFLOPs, memory occupation), and inference speed (FPS, inference time) are summarized.

As illustrated in Table 4, BKMNet achieves the optimal performance across all evaluation metrics on both polyp segmentation subtasks, significantly outperforming all compared state-of-the-art (SOTA) methods. For the Adenomas vs. Non-Adenomas task, BKMNet obtains the highest IoU of 91.31% and Dice of 92.51%, exceeding the second-best method AC-MambaSeg (90.65% IoU, 91.02% Dice)

by 0.66% and 1.49%, respectively. For the more complex three-category segmentation task (Tubular vs. Villous vs. Non-Adenomas), BKMNet also yields a prominent performance advantage, with IoU and Dice scores reaching 90.29% and 91.83%, which surpass all CNN, transformer, and mamba-based competitors. CNN-based methods (UNet, EBLNet) suffer from a limited feature representation capability, resulting in low segmentation accuracy despite their large model scales. Lightweight CNN models (LETNet, LCNet) greatly reduce parameters and computational overhead but sacrifice segmentation performance severely, failing to meet the precision requirements of clinical polyp segmentation. Mobile U-ViT improves accuracy via transformer-based global modeling, yet its performance is still inferior to mamba-based architectures due to an insufficient local feature refinement capability.

Compared with recent advanced mamba-based segmentation models, BKMNet exhibits more balanced and superior comprehensive performance. VM-UNet achieves competitive accuracy but requires massive computational resources and suffers from low inference efficiency. LightM-UNet realizes lightweight optimization with reduced GFLOPs but lacks elaborate structural and edge feature enhancement, leading to obvious accuracy gaps compared with BKMNet. AC-MambaSeg achieves promising segmentation results, but its parameter scale and memory consumption are higher than BKMNet, and its inference speed is unsatisfactory. In contrast, BKMNet integrates structural awareness, Kalman-optimized global modeling, adaptive feature fusion, and explicit edge refinement modules, which effectively compensate for the local feature deficiency and boundary blur problems, thus achieving better segmentation performance.

In terms of model efficiency, BKMNet achieves an excellent trade-off between segmentation precision and computational cost. The parameter size of BKMNet is only 5.28 M, which is far smaller than heavy-duty models such as UNet (31.04 M), EBLNet (42.47 M), and VM-UNet (27.43 M), demonstrating superior lightweight characteristics. Its GFLOPs and memory consumption are maintained at a low level (17.95 GFLOPs, 20.67 M memory), which is competitive with mainstream lightweight medical segmentation models. In terms of real-time inference performance, BKMNet achieves an FPS of 12.80 and an inference time of 78.01 ms, ensuring stable and efficient inference for clinical colonoscopic video sequences. Different from existing lightweight models that blindly reduce parameters at the cost of accuracy degradation, BKMNet realizes precision improvement through effective module design and feature optimization rather than simple network pruning, which verifies its high-efficiency feature learning mechanism.

We also present some typical scenes of polys and show the visual results generated by our BKMNet and eight other competing methods in Figures 9 and 10. The first and second columns display input images and corresponding ground truths, respectively, with subsequent columns showing the segmentation predictions of these segmentation methods. Polyp boundaries are highlighted by green contours. Light gray pixels represent the overlapping regions between the prediction and ground truth. Notably, when handling the colorectal polyp segmentation task, BKMNet exhibits distinct advantages over competing methods in multiple aspects. First, BKMNet is able to localize the polyp from the complex background more accurately. Our predictions showed accurate edges and were more similar to the ground truths. The kappa coefficient between ground truths and the predictions of our model reaches 0.96, and McNemar's test shows there are no significant differences, indicating a high degree of accuracy and reliability. Second, BKMNet outperforms competing methods in accommodating variations in polyp size and shape, as illustrated in the first to eighth rows. Third, BKMNet is more

suitable for accurately locating polyps with an ambiguous boundary. As shown in the sixth and seventh rows, most segmentation models perform poorly when no clear boundary exists between a polyp and its surrounding tissues. In contrast, BKMNet can produce more accurate predictions, with clearer boundaries. By summarizing the results of quantitative and qualitative comparisons, we can conclude that our proposed BKMNet is more conducive to tackling the challenging polyp-segmentation task than competing methods.

4.7. Comparisons with existing studies

Table 5. Comparison results with existing methods.

Methods	Year	Parameter	Kvasir-SEG		CVC-ClinicDB		CVC-ColonDB	
			IoU	Dice	IoU	Dice	IoU	Dice
UNet [27]	2015	34.51 M	0.746	0.818	0.755	0.823	0.444	0.512
PraNet [29]	2020	32.55 M	0.840	0.898	0.849	0.899	0.640	0.709
MSNet [17]	2021	29.74 M	0.862	0.907	0.879	0.921	0.678	0.755
HSNet [18]	2022	29.26 M	0.877	0.926	0.905	0.948	0.735	0.810
CFANet [47]	2023	25.24 M	0.861	0.915	0.883	0.933	0.665	0.743
Polyp-PVT [37]	2023	44.98 M	0.864	0.917	0.889	0.937	0.727	0.808
EMTS-Net [20]	2023	31.50 M	0.869	0.919	0.885	0.935	0.708	0.788
Meta-UNet [19]	2023	22.21 M	0.864	0.919	0.890	0.939	0.728	0.809
CAFE-Net [49]	2024	35.53 M	0.889	0.933	0.899	0.943	0.740	0.820
RAFPNet [21]	2024	31.83 M	0.869	0.919	0.898	0.940	0.733	0.811
Polyp-Mamba [26]	2025	49.50 M	0.867	0.919	0.896	0.941	0.713	0.791
BMANet [22]	2025	30.59 M	0.879	0.925	0.896	0.942	0.725	0.804
FMCA-Net [23]	2025	28.61 M	0.866	0.919	0.898	0.944	0.723	0.804
BKMNet	-	**5.28 M**	**0.878**	**0.929**	**0.908**	**0.953**	**0.741**	**0.823**

To further verify the effectiveness of BKMNet, we perform a comparative experiment against existing methods using three widely used polyp-segmentation datasets, including Kvasir-SEG [51], CVC-ClinicDB [52], and CVC-ColonDB [53]. This comparison selected some polyp-segmentation methods from the literature, including UNet [27], PraNet [29], MSNet [17], HSNet [18], CFANet [47], Polyp-PVT [37], EMTS-Net [20], Meta-UNet [19], CAFE-Net [49], RAFPNet [21], Polyp-Mamba [26], BMANet [22], and FMCA-Net [23]. As illustrated in Table 5, among all the existing approaches, the proposed BKMNet exhibits remarkable performance, achieving superior or comparable results across all datasets while maintaining a significantly reduced parameter count. On the Kvasir-SEG dataset, BKMNet achieves an IoU of 0.878 and a Dice of 0.929, which is slightly lower than the top-performing CAFE-Net. Notably, on the CVC-ClinicDB dataset, BKMNet outperforms all existing methods with the highest IoU (0.908) and Dice (0.953), highlighting its effectiveness in handling diverse polyp appearances and background complexities. On the more challenging CVC-ColonDB dataset, BKMNet achieves an IoU of 0.741 and a Dice of 0.823, which is comparable to CAFE-Net, demonstrating its robustness in segmenting polyps. A critical advantage of BKMNet lies in its computational efficiency. With only 5.28 M parameters, it represents a substantial reduction in model size compared to other methods. This significant parameter reduction without sacrificing performance underscores the efficiency of BKMNet's architecture, making it more suitable for deployment in resource-constrained environments. The superior performance of BKMNet across all metrics and datasets suggests that its design effectively captures discriminative features relevant to polyp segmentation while maintaining model parsimony.

This balance between accuracy and efficiency is crucial for advancing polyp segmentation from research to practical clinical use, where both diagnostic precision and computational feasibility are paramount.

5. Discussions

As discussed above, this work proposed an automated computer-aided segmentation network (BKMNet) for accurate identification and subtype segmentation of colorectal polyps, and comprehensively demonstrated its superior performance and clinical feasibility for colonoscopic lesion analysis. Since different types of polyps possess distinct malignant transformation potentials and require completely different clinical intervention strategies, accurate, objective, and automated polyp segmentation and classification before clinical treatment is of great significance for early colorectal cancer prevention and individualized therapy formulation. Colorectal endoscopic images commonly contain complex intestinal backgrounds, blurred lesion boundaries, and diverse polyp morphologies, making precise pixel-level segmentation and subtype discrimination extremely challenging. Through the above ablation studies and comprehensive SOTA comparative experiments, we observe that the proposed BKMNet can still efficiently achieve accurate and robust segmentation performance even facing such difficult imaging conditions, relying on the elaborately designed multi-frequency multi-scale attention enhancement, structural local awareness, Kalman-optimized mamba encoding, adaptive decoder fusion, and explicit edge-refinement modules. The synergistic cooperation of these core components enables BKMNet to effectively extract discriminative lesion features, suppress complex background noise, and refine ambiguous polyp boundaries, thereby yielding satisfactory pixel-level segmentation results on both Adenomas vs. Non-Adenomas and Tubular vs. Villous vs. Non-Adenomas tasks.

Benefiting from precise pixel-level segmentation outputs, BKMNet enables objective evaluation of polyp lesion scope, morphological characteristics, and structural abnormalities. Furthermore, this automated segmentation network is capable of capturing subtle morphological differences and boundary changes of polyps that are difficult to distinguish by the naked eye, allowing evaluation of clinical therapeutic effects on lesions and local intestinal regions. It is worth noting that the inference speed and computational efficiency reported in this study are tested. The high efficiency and low parameter redundancy of BKMNet leave sufficient deployment margin for low-computing embedded devices and portable endoscopic equipment. Compared with manual diagnosis, the proposed method achieves stable and high-precision polyp analysis with low computational overhead, greatly improving the efficiency and objectivity of clinical polyp screening, biopsy decision-making, and postoperative assessment. As a time-saving and precise auxiliary diagnosis tool, BKMNet can provide reliable objective references for endoscopists, which effectively reduces physician workload and improves the overall efficiency of endoscopic diagnosis and treatment.

Despite the promising performance and great clinical application potential, the current BKMNet still has several limitations. First, the model still suffers from some extremely challenging cases, where the target polyps are so large that they occupy most of the image with indistinguishable boundaries (see Figure 11(a)), the image is too dark (see Figure 11(b)), or the polyp is too small and has low contrast with the background (see Figure 11(c),(d)). Second, due to a lack of resources, the dataset used in our research was collected from only one hospital. The proposed computer-aided diagnosis method lacks sufficient multi-center cross-device data verification, which limits its external generalization ability in

clinical imaging environments. Third, although BKMNet achieves balanced computational overhead, its multi-module collaborative structure still requires certain memory resources, which can be further optimized for ultra-portable device deployment.

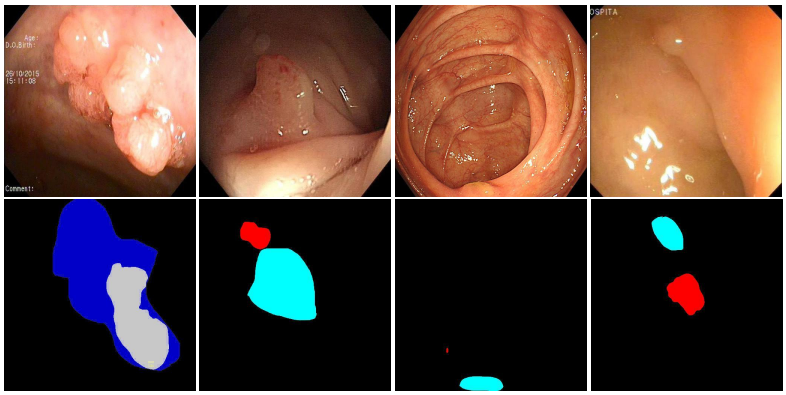

Figure 11. Failure cases. (a) Large polyps, (b) dark image, (c) and (d) small polyps. Blue and yellow pixels indicate the ground truth and predictions of adenomatous polyps, respectively. Light gray pixels represent the overlapping regions between the prediction and ground truth. Cyan and red pixels indicate the ground truth and predictions of non-adenomatous polyps, respectively.

In future work, we will focus on solving the above limitations to further promote the clinical transformation and practical application of the proposed method. First, we will conduct large-scale multi-center clinical validation and external dataset testing to systematically verify the generalization and stability of BKMNet in diverse clinical environments. Second, we will enhance the model's robustness against low-quality and challenging endoscopic images by introducing image preprocessing and anti-interference feature learning strategies. Third, we will further optimize the network structure to construct a more lightweight and efficient version suitable for real-time embedded deployment, providing comprehensive, objective, and efficient auxiliary diagnosis support for standardized clinical colorectal polyp screening and treatment.

6. Conclusions

This work introduces a novel lightweight boundary-enhanced Kalman-mamba segmentation network, called BKMNet, specifically designed for accurate differentiation between adenomatous and non-adenomatous polyps. The encoder's initial stem incorporates a multi-frequency combined attention (MFCA) block, which boosts initial feature extraction by leveraging diverse channel frequency components. Subsequently, lightweight FVSSM layers generate multi-level feature maps, capturing local detailed features and contextual information. The Kalman-mamba layer then reinforces long-range dependency modeling, facilitating a comprehensive spatial understanding of polyps and surrounding tissues. The decoder employs a dual attention module for efficient multi-level feature integration, while the RD-BE module optimizes boundary segmentation, mitigating the ambiguous edge delineation endemic to traditional models. A joint loss function, comprising terms for the boundary,

polyp body, and overall segmentation, ensures robustness across critical diagnostic indicators. Evaluations across private and public datasets confirm BKMNet's superiority over existing state-of-the-art methods in the polyp detection rate, segmentation accuracy, and boundary precision. This robust performance suggests BKMNet can deliver accurate and reliable auxiliary diagnostic information, thereby enhancing the efficiency and quality of endoscopic diagnosis.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Scientific Research Foundation of the Higher Education Institutions of Henan Province (No. 25A520062), National Science Foundation for Young Scientists of Henan Province (No. 262300422566), Doctoral Special Fund Project of Nanyang Normal University (Nos. 2025ZX010, 2025ZX014), and Senior-End Foreign Expert Introduction Project of Henan Province (HNGD2025034).

Conflict of interest

Bing Tan is a special issue editor of the Electronic Research Archive and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

1. D. S. Chan, M. Cariolou, G. Markozannes, K. Balducci, R. Vieira, S. Kiss, et al., Post-diagnosis dietary factors, supplement use and colorectal cancer prognosis: A global cancer update programme (cup global) systematic literature review and meta-analysis, *Int. J. Cancer*, **155** (2024), 445–470. https://doi.org/10.1002/ijc.34906
2. J. Bond, Clinical relevance of the small colorectal polyp, *Endoscopy*, **33** (2001), 454–457. https://doi.org/10.1055/s-2001-14266
3. L. Shi, H. Li, S. Li, S. Lin, Y. Wu, Pseudoinvasion and squamous metaplasia/morules in colorectal adenomatous polyp: A case report and literature review, *Diagn. Pathol.*, **19** (2024), 126. https://doi.org/10.1186/s13000-024-01535-9
4. G. Yue, P. Wei, Y. Liu, Y. Luo, J. Du, T. Wang, Automated endoscopic image classification via deep neural network with class imbalance loss, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–11. https://doi.org/10.1109/tim.2023.3264047
5. L. Lin, G. Lv, B. Wang, C. Xu, J. Liu, Polyp-lvt: Polyp segmentation with lightweight vision transformers, *Knowl. Based Syst.*, **300** (2024), 112181. https://doi.org/10.1016/j.knosys.2024.112181

6. A. Leufkens, M. Van Oijen, F. Vleggaar, P. Siersema, Factors influencing the miss rate of polyps in a back-to-back colonoscopy study, *Endoscopy*, **44** (2012), 470–475. https://doi.org/10.1055/s-0031-1291666
7. Y. Teng, Y. Yang, J. Yang, Q. Lu, J. Shi, J. Xu, et al., Association between triglyceride-glucose index and colorectal polyps: A retrospective cross-sectional study, *World J. Gastroenterol. Endosc.*, **16** (2024), 55. https://doi.org/10.4253/wjge.v16.i2.55
8. D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, et al., Spectralgpt: Spectral remote sensing foundation model, *IEEE Trans. Pattern Anal. Mach. Intell.*, **46** (2024), 5227–5244. https://doi.org/10.1109/tpami.2024.3362475
9. D. Hong, C. Li, N. Yokoya, B. Zhang, X. Jia, A. Plaza, et al., Hyperspectral imaging, *Nat. Rev. Methods Primers*, **6** (2026), 19. https://doi.org/10.1038/s43586-026-00470-x
10. P. Singh, Y. Huang, Akdc: Ambiguous kernel distance clustering algorithm for COVID-19 ct scans analysis, *IEEE Trans. Syst. Man Cybern. Syst.*, **54** (2024), 6218–6229. https://doi.org/10.1109/tsmc.2024.3418411
11. P. Singh, Y. Huang, An ambiguous edge detection method for computed tomography scans of coronavirus disease 2019 cases, *IEEE Trans. Syst. Man Cybern. Syst.*, **54** (2023), 352–364. https://doi.org/10.1109/tsmc.2023.3307393
12. P. Singh, S. S. Bose, A quantum-clustering optimization method for COVID-19 ct scan image segmentation, *Expert Syst. Appl.*, **185** (2021), 115637. https://doi.org/10.1016/j.eswa.2021.115637
13. P. Singh, S. S. Bose, Ambiguous d-means fusion clustering algorithm based on ambiguous set theory: Special application in clustering of ct scan images of COVID-19, *Knowl. Based Syst.*, **231** (2021), 107432. https://doi.org/10.1016/j.eswa.2021.115637
14. Z. Zhu, H. Wang, G. Qi, Y. Li, N. Mazur, Y. Liu, et al., A survey on lightweight technology of neural networks for medical image segmentation, *Pattern Recognit.*, **179** (2026), 113870. https://doi.org/10.1016/j.patcog.2026.113870
15. Y. Ding, S. Li, H. Li, G. Qi, B. Cong, Y. Gong, et al., Physical regularization loss: Integrating physical knowledge to image segmentation, *Int. J. Comput. Vis.*, **134** (2026), 137. https://doi.org/10.1007/s11263-026-02776-5
16. Z. Zhu, Z. Zhang, G. Qi, Y. Li, P. Yang, Y. Liu, Probability map-guided network for 3d volumetric medical image segmentation, *IEEE Trans. Image Process.*, **34** (2025), 7222–7234. https://doi.org/10.1109/tip.2025.3623259
17. X. Zhao, L. Zhang, H. Lu, Automatic polyp segmentation via multi-scale subtraction network, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference*, Springer, (2021), 120–130. https://doi.org/10.1007/978-3-030-87193-2_12
18. W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, C. M. Sham, Hsnet: A hybrid semantic network for polyp segmentation, *Comput. Biol. Med.*, **150** (2022), 106173. https://doi.org/10.1016/j.compbiomed.2022.106173
19. H. Wu, Z. Zhao, Z. Wang, Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation, *IEEE Trans. Autom. Sci. Eng.* **21** (2023), 4117–4128. https://doi.org/10.1109/tase.2023.3292373

20. M. Wang, X. An, Z. Pei, N. Li, L. Zhang, G. Liu, et al., An efficient multi-task synergetic network for polyp segmentation and classification, *IEEE J. Biomed. Health Inform.*, **28** (2023), 1228–1239. https://doi.org/10.1109/jbhi.2023.3273728
21. L. Meng, Y. Li, W. Duan, Three-stage polyp segmentation network based on reverse attention feature purification with pyramid vision transformer, *Comput. Biol. Med.*, **179** (2024), 108930. https://doi.org/10.1016/j.compbiomed.2024.108930
22. Z. Wu, H. Chen, X. Xiong, S. Wu, H. Li, X. Zhou, Bmanet: Boundary-guided multi-level attention network for polyp segmentation in colonoscopy images, *Biomed. Signal Process. Control*, **105** (2025), 107524. https://doi.org/10.1016/j.bspc.2025.107524
23. S. Wang, S. Lin, F. Sun, X. Li, Multi-feature fusion for accurate polyp segmentation using pyramid visual transformers, *Expert Syst. Appl.*, **280** (2025), 127558. https://doi.org/10.1016/j.eswa.2025.127558
24. Y. Ma, Y. Liu, J. Cheng, H. Zhan, G. Qi, X. Zhang, et al., Cihm: Context-insight hybrid mamba for efficient medical image segmentation, *Vis. Intell.*, **4** (2026), 12. https://doi.org/10.1007/s44267-026-00113-5
25. J. Ruan, J. Li, S. Xiang, Vm-unet: Vision mamba unet for medical image segmentation, preprint, arXiv:2402.02491.
26. X. Zhu, W. Wang, C. Zhang, H. Wang, Polyp-mamba: A hybrid multi-frequency perception gated selection network for polyp segmentation, *Inf. Fusion*, **115** (2025), 102759. https://doi.org/10.1016/j.inffus.2024.102759
27. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference*, Springer, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
28. M. Shi, S. Lin, Q. Yi, J. Weng, A. Luo, Y. Zhou, Lightweight context-aware network using partial-channel transformation for real-time semantic segmentation, *IEEE Trans. Intell. Transp. Syst.*, **25** (2024), 7401–7416. https://doi.org/10.1109/tits.2023.3348631
29. D. Fan, G. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, et al., Pranet: Parallel reverse attention network for polyp segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, (2020), 263–273. https://doi.org/10.1007/978-3-030-59725-2_26
30. W. Zhang, B. Su, C. Huangfu, L. Zhang, A review of colorectal polyp segmentation methods, *J. Comput. Electron. Inf. Manag.*, **20** (2026), 111–116. https://doi.org/10.54097/pqj8hf46
31. H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, et al., Enhanced boundary learning for glass-like object segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 15859–15868. https://doi.org/10.1109/iccv48922.2021.01556
32. D. Xie, Y. Zhang, X. Tian, L. Xu, L. Duan, L. Tian, Bgfe-net: A boundary-guided feature enhancement network for segmentation of targets with fuzzy boundaries, *Neurocomputing*, **618** (2025), 129127. https://doi.org/10.1016/j.neucom.2024.129127

33. Y. Lin, D. Zhang, X. Fang, Y. Chen, K. Cheng, H. Chen, Rethinking boundary detection in deep learning models for medical image segmentation, in *International Conference on Information Processing in Medical Imaging*, Springer, (2023), 730–742. https://doi.org/10.1016/j.media.2025.103615
34. Z. Wu, X. Zhang, F. Li, S. Wang, L. Huang, J. Li, W-Net: A boundary-enhanced segmentation network for stroke lesions, *Expert Syst. Appl.*, **230** (2023), 120637. https://doi.org/10.1016/j.eswa.2023.120637
35. S. Li, X. Tang, B. Cao, Y. Peng, X. He, S. Ye, et al., Boundary guided network with two-stage transfer learning for gastrointestinal polyps segmentation, *Expert Syst. Appl.*, **240** (2024), 122503. https://doi.org/10.1016/j.eswa.2023.122503
36. G. Xu, J. Li, G. Gao, H. Lu, J. Yang, D. Yue, Lightweight real-time semantic segmentation network with efficient transformer and cnn, *IEEE Trans. Intell. Transp. Syst.*, **24** (2023), 15897–15906. https://doi.org/10.1109/tits.2023.3248089
37. B. Dong, W. Wang, D. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, *CAAI Artif. Intell. Res.*, **2** (2023), 9150015. https://doi.org/10.26599/air.2023.9150015
38. F. Tang, B. Nian, J. Ding, W. Ma, Q. Quan, C. Dong, et al., Mobile u-vit: Revisiting large kernel and u-shaped vit for efficient medical image segmentation, in *Proceedings of the 33rd ACM International Conference on Multimedia*, (2025), 3408–3417. https://doi.org/10.1145/3746027.3755076
39. H. Sun, Y. Zhang, L. Xu, S. Jin, Y. Chen, Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2025), 7087–7095. https://doi.org/10.1609/aaai.v39i7.32761
40. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., Transunet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306.
41. N. Zhang, L. Yu, D. Zhang, W. Wu, S. Tian, X. Kang, et al., Ct-net: Asymmetric compound branch transformer for medical image segmentation, *Neural Netw.*, **170** (2024), 298–311. https://doi.org/10.2139/ssrn.4331189
42. W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, L. Ma, Lightm-unet: Mamba assists in lightweight unet for medical image segmentation, preprint, arXiv:2403.05246.
43. V. T. Nguyen, V. T. Pham, T. T. Tran, Ac-mambaseg: An adaptive convolution and mamba-based architecture for enhanced skin lesion segmentation, in *International Conference on Green Technology and Sustainable Development*, Springer, (2024), 13–26. https://doi.org/10.1007/978-3-031-76197-3_2
44. Z. Zhang, Q. Ma, T. Zhang, J. Chen, H. Zheng, W. Gao, Switch-umamba: Dynamic scanning vision mamba unet for medical image segmentation, *Med. Image Anal.*, **107** (2025), 103792. https://doi.org/10.1016/j.media.2025.103792
45. Z. Wang, J. Zheng, Y. Zhang, G. Cui, L. Li, Mamba-unet: Unet-like pure visual mamba for medical image segmentation, preprint, arXiv:2402.05079.

46. T. Hussain, H. Shouno, M. A. Mohammed, H. A. Marhoon, T. Alam, Dcssga-unet: Biomedical image segmentation with densenet channel spatial and semantic guidance attention, *Knowl.-Based Syst.*, **314** (2025), 113233. https://doi.org/10.1016/j.knosys.2025.113233

47. T. Zhou, Y. Zhou, K. He, C. Gong, J. Yang, H. Fu, et al., Cross-level feature aggregation network for polyp segmentation, *Pattern Recognit.*, **140** (2023), 109555. https://doi.org/10.1016/j.patcog.2023.109555

48. H. Huang, Z. Chen, Y. Zou, M. Lu, C. Chen, Y. Song, et al., Channel prior convolutional attention for medical image segmentation, *Comput. Biol. Med.*, **178** (2024), 108784. https://doi.org/10.1016/j.compbiomed.2024.108784

49. G. Liu, S. Yao, D. Liu, B. Chang, Z. Chen, J. Wang, et al., Cafe-net: Cross-attention and feature exploration network for polyp segmentation, *Expert Syst. Appl.*, **238** (2024), 121754. https://doi.org/10.1016/j.eswa.2023.121754

50. J. H. Nam, N. S. Syazwany, S. J. Kim, S. C. Lee, Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 11480–11491. https://doi.org/10.1109/cvpr52733.2024.01091

51. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, et al., Kvasir-seg: A segmented polyp dataset, in *International Conference on Multimedia Modeling*, Springer, (2019), 451–462. https://doi.org/10.1007/978-3-030-37734-2_37

52. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, *Comput. Med. Imaging Graph.*, **43** (2015), 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007

53. N. Tajbakhsh, S. R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, *IEEE Trans. Med. Imaging*, **35** (2015), 630–644. https://doi.org/10.1109/tmi.2015.2487997

AIMS Press