CAM-PoEm: Contour-Aware Mamba Bottleneck with Multi-Kernel Positional Embedding-based Encoder for Gastrointestinal Polyp Segmentation

Anonymous submission

Abstract

Accurate segmentation of gastrointestinal polyps is vital for early detection of colorectal cancer, yet remains challenging due to ambiguous boundaries, heterogeneous textures, and high morphological variability. While convolutional neural networks (CNNs) lack effective global context modeling and transformers suffer from quadratic complexity, recent state-space models (SSMs) like Mamba enable efficient linear-time global modeling. However, SSMs remain inherently insensitive to boundary structures that are critical for precise medical segmentation. We propose CAM-PoEm, a novel segmentation framework that introduces a Contour-Aware Mamba (CAM) bottleneck, where learnable contour cues are integrated into Mamba's selective scan mechanism to enable morphology-aware global context modeling with boundary sensitivity. By placing CAM in the bottleneck where fine spatial details are less and boundaries and global features are crucial, we enhance global reasoning while preserving crucial contour details with minimal computational cost. To complement this, we incorporate a Multi-Kernel Positional Embedding (MKPE) encoder that captures multi-scale spatial features, enabling robust local representation alongside global modeling. To our knowledge, this is the first application of contour-sensitive SSMs in medical image analysis. CAM-PoEm achieves state-of-the-art performance across four benchmark datasets—Kvasir-SEG (Dice: 0.91) (Jha et al. 2019a), PolypGen2021 (0.88) (Ali et al. 2021), CVC-ClinicDB (0.91)(Bernal et al. 2015), and CVC-ColonDB (0.88) (Bernal, Sánchez, and Vilariño 2012)—using only 4.96M parameters, demonstrating strong boundary delineation and high computational efficiency.

Introduction

Gastrointestinal (GI) polyps—abnormal growths in the colon, rectum, or stomach—are established precursors to colorectal cancer, a leading cause of global cancer-related deaths (Djinbachian et al. 2020). Timely detection and removal via colonoscopy can reduce incidence by up to 30% (Haggar and Boushey 2009), making accurate polyp segmentation from endoscopic images essential for early diagnosis and improved clinical outcomes. Yet, segmentation remains challenging due to high variability in polyp appearance and indistinct boundaries with surrounding tissues (Pooler et al. 2023).

Deep learning has advanced this task, but existing approaches face key limitations. CNNs like FCNs (Long, Shelhamer, and Darrell 2015), U-Net (Ronneberger, Fischer, and Brox 2015a), and ResUNet++ (Jha et al. 2019b) extract local features effectively but struggle to model long-range dependencies due to limited receptive fields (Li et al. 2024c; Zhao et al. 2024), leading to degraded performance on polyps with irregular boundaries (Xie et al. 2024a). Transformer-based models such as TransUNet (Chen et al. 2021) and Swin-UNet (Cao et al. 2021) capture global context using self-attention (Dosovitskiy et al. 2021; Liu et al. 2024), but their quadratic complexity hinders scalability, and reliance on natural image pretraining introduces domain gaps in medical settings (Vaswani et al. 2017; Kirillov et al. 2023). While attention-based models enhance long-range reasoning, they often yield redundant representations and imprecise segmentation near object boundaries due to pixel-category mixing and lack of explicit contour modeling (You et al. 2025; Li et al. 2024a).

State-space models (SSMs) like Mamba (Gu and Dao 2023) have recently emerged as efficient alternatives, offering linear-time global modeling. Their applications in segmentation—MambaUNet (Wang et al. 2024b), UMamba (Ma, Li, and Wang 2024), VM-UNet (Ruan, Li, and Xiang 2024)—show promise. However, current SSM-based methods remain largely insensitive to spatial boundaries and often operate at a single scale, limiting accuracy in medical segmentation tasks that require fine-grained contour delineation (Xu et al. 2024; Ho et al. 2025a).

To this end, we propose CAM-PoEm, a novel segmentation framework that integrates global context modeling with explicit boundary sensitivity. Central to our approach is the Contour-Aware Mamba (CAM) bottleneck, which introduces learnable contour guidance into Mamba's selective scan for boundary-aware global reasoning. This is complemented by a Multi-Kernel Positional Embedding (MKPE) encoder that captures spatial details across multiple receptive fields. Together, these components enable precise, efficient polyp segmentation in a unified encoder-decoder architecture. We summarize our main contributions as follows:

- **Contour-Aware State-Space Modeling**: We propose Contour-Aware Mamba (CAM)—the first state-space model to explicitly incorporate contour sensitivity into its sequence modeling. By injecting learnable boundary cues via **Learnable Contour Extractor** into Mamba's selective scan mechanism, CAM enables morphology-

aware state transitions, enhancing spatial localization without sacrificing global context modeling. This design allows Mamba to capture long-range dependencies while remaining sensitive to fine boundary structures, a critical aspect in medical image segmentation.

- **Unified Global-Local Architecture Design**: We introduce CAM-PoEm, a novel encoder-decoder framework that strategically combines local and global modeling with boundary precision. MKPE encoders enhance spatial encoding by capturing multi-scale local structures, improving robustness across varying polyp sizes and textures. CAM is positioned in the bottleneck, where semantic abstraction is highest and spatial resolution is lowest—making it the ideal location to inject contour awareness into global reasoning. This placement enables efficient morphology-aware context modeling over compact feature representations, while keeping computational cost low.
- **Comprehensive Empirical Validation across datasets**: We conduct extensive experiments on four benchmark datasets—Kvasir-SEG (Jha et al. 2019a), PolypGen2021 (Ali et al. 2021), CVC-ClinicDB (Bernal et al. 2015), and CVC-ColonDB (Bernal, Sánchez, and Vilariño 2012). CAM-PoEm achieves state-of-the-art performance across all datasets, matching or outperforming exisiting segmentation baselines, particularly in boundary precision and global consistency, while maintaining high computational efficiency.

Related Work

Convolutional and Transformer-Based Architectures

CNN-based models such as U-Net (Ronneberger, Fischer, and Brox 2015b), FCNs (Long, Shelhamer, and Darrell 2014), and ResUNet++ (Jha et al. 2019c) established strong baselines for medical segmentation through local feature extraction and skip connections. Extensions like A-DenseUNet (Safarov and Whangbo 2021) and Dilated U-Net (Karthikha, Jamal, and Rafiammal 2024) improved multi-scale context via atrous and dilated convolutions. However, CNNs inherently struggle with capturing long-range dependencies due to their limited receptive fields and deep-layer inefficiency.

Transformers addressed this via global self-attention, as seen in ViT (Dosovitskiy et al. 2021), TransUNet (Chen et al. 2021), and polyp-specific models like PraNet (Fan et al. 2020a), ColonFormer (Duc et al. 2022), and Polyp-PVT (Dong et al. 2023). Despite improved boundary localization, transformers suffer from high memory cost and require large-scale pretraining, limiting their deployment in medical contexts (Vaswani et al. 2017; Kirillov et al. 2023).

Hybrid models like SwinE-Net (Park and Lee 2022), Focus U-Net (Yeung et al. 2021), and DCATNet (Wang et al. 2025) combine CNNs for local encoding with transformers for global reasoning. These improve performance but introduce significant complexity and often underperform at fine boundary delineation.

Mamba-Based Architectures

Mamba (Gu and Dao 2024) and Vision Mamba (Zhu et al. 2024) introduced selective scanning for linear-time sequence modeling. Vision adaptations like Mamba-UNet (Wang et al. 2024c), VM-UNetV2 (Zhang et al. 2024), and ProMamba (Xie et al. 2024b) showed strong segmentation capabilities via bidirectional and multi-scale extensions. More recent efforts such as RM-UNet (Tang et al. 2024), LKM-UNet (Wang et al. 2024a), and Topo-VM-UNetV2 (Adame et al. 2025) explored recurrence directionality and spatial topology. Nevertheless, these models lack contour guidance mechanisms, limiting their boundary precision in clinical segmentation.

Multi-Kernel Positional Encoding (MKPE)

MKPE enhances spatial awareness through multi-scale receptive fields, as in ConvNeXt-MPE (Mau et al. 2023), PEFNet, and MEP (Gao 2024). MKSA-BAHA (Zhou et al. 2024) extended this to city-scale segmentation via hybrid attention. Yet, many MKPE variants lack adaptability across resolutions and remain computationally expensive for high-resolution medical images.

Contour-Aware Modeling

Boundary-guided models such as BDG-Net (Qiu et al. 2022), Polyper (Shao, Zhang, and Hou 2023), and contour-consistency decoders (Li et al. 2024b) explicitly encode edge priors via auxiliary branches or post-processing. While effective in improving contour alignment, they often rely on external supervision and are not fully integrated into the core modeling pipeline. Recent efforts like LiteMamba-Bound (Ho et al. 2025b) aim for lightweight boundary-aware adaptation but still lack unified, end-to-end design.

Methodology

This section presents the proposed **CAM-PoEm** framework, which introduces Contour-Aware Mamba-based state-space modeling within an encoder-decoder segmentation pipeline. Our motivation stems from the unique challenges in medical image segmentation, especially in gastrointestinal polyp delineation, where ambiguous boundaries, diverse morphologies, and local textural variations require both long-range context modeling and local edge sensitivity. We begin by outlining the overall architecture, followed by an in-depth description of each core component, including theoretical underpinnings and implementation details.

Architectural Overview of CAM-PoEm

CAM-PoEm adopts a U-shaped encoder-decoder framework, augmented with a Mamba-based bottleneck for global context modeling. Let $\mathbf{X} \in R^{B\times 3\times H\times W}$ denote the input medical image, where B is the batch size, and H and W are spatial dimensions. The network outputs a segmentation map $\mathbf{Y} \in R^{B\times C\times H\times W}$, where C is the number of classes (typically $C = 2$ for binary segmentation). Fig. 1 shows the architectural overview.

The encoder path consists of three DoubleConvWithMKPE blocks with channel dimensions 64, 128,

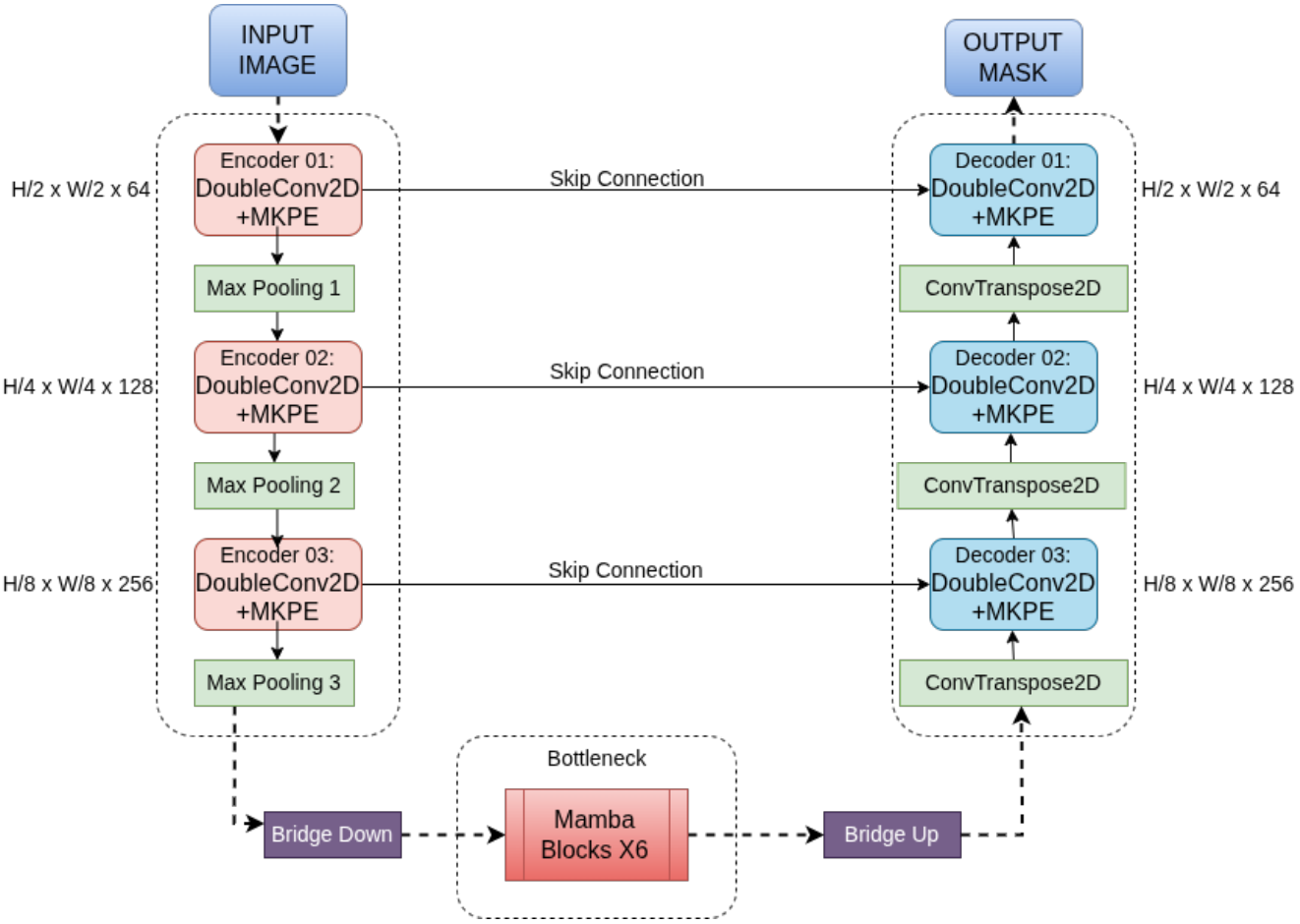

Figure 1: The U-shaped architecture of the proposed model

and 256, each followed by max-pooling to halve spatial dimensions. The bottleneck employs Contour-Aware Mamba Blocks with a feature dimension $d_{model} = 128$, enabling efficient sequence modeling. The decoder path upsamples features via transposed convolutions, concatenates skip connections from the encoder, and applies DoubleConvWithMKPE blocks to refine features. Deep supervision generates auxiliary outputs at intermediate decoder stages, enhancing gradient flow. The final output is processed through a 1x1 convolution and an MKPE module to produce the segmentation map.

The overall architecture can be mathematically formulated as:

$$\mathbf{Y} = f_{MKPE}(f_{\mathcal{D}}(f_{\mathcal{M}}(f_{\mathcal{E}}(\mathbf{X})))) \quad (1)$$

where $f_{\mathcal{E}}(\cdot)$, $f_{\mathcal{M}}(\cdot)$, $f_{\mathcal{D}}(\cdot)$, and $f_{MKPE}(\cdot)$ represent the encoder, Mamba-based bottleneck, decoder, and final MKPE operations, respectively.

Proposed Multi-Kernel Positional Encoding (MKPE) for Local Spatial Awareness

Accurate polyp segmentation requires capturing fine-grained spatial cues and structural variations at multiple scales. Traditional positional encodings, such as sinusoidal embeddings or fixed grid coordinates, often fail to represent such spatial intricacies—especially in medical images with varying morphology and texture.

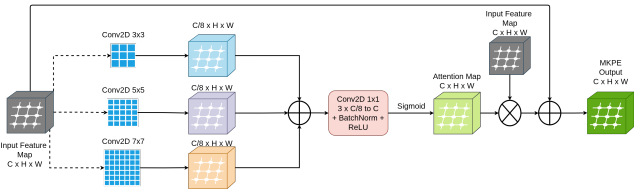

Figure 2: MKPE module architecture. Multiple convolutional branches extract features at different scales, followed by concatenation and attention-based modulation. "⊕" indicates addition; "⊗" denotes element-wise multiplication.

To address this, we propose a **Multi-Kernel Positional Encoding (MKPE)** module, as shown in Figure. 2 which enhances spatial representation through two mechanisms: multi-scale convolutional encoding and attention-based modulation.

Multi-Scale Context Aggregation. Given an input feature map $\mathbf{F} \in R^{B\times C\times H\times W}$, MKPE applies parallel depth-

wise convolutional branches with kernel sizes 3×3, 5×5, and 7×7. These branches capture spatial patterns at varying receptive fields, enabling both local detail sensitivity and broader contextual encoding. The outputs are concatenated along the channel axis to form a unified representation:

$$\mathbf{F}_{multi} = Concat[Conv_{3\times3}(\mathbf{F}), Conv_{5\times5}(\mathbf{F}), Conv_{7\times7}(\mathbf{F})]$$

Attention-Based Spatial Recalibration. To adaptively emphasize spatially informative regions, MKPE employs a lightweight attention block. A series of 1×1 convolutions with batch normalization and ReLU activation compress and recalibrate the aggregated features:

$$\mathbf{A}_{pos} = \sigma\left(f_{conv2}\left(ReLU\left(BN(f_{conv1}(\mathbf{F}_{multi}))\right)\right)\right)$$

where $\mathbf{A}_{pos} \in R^{B\times C\times H\times W}$ is the learned spatial attention map and σ denotes the sigmoid function.

Residual Feature Enhancement. The input is modulated by the attention map via residual fusion:

$$\mathbf{F}_{enhanced} = \mathbf{F} \odot \mathbf{A}_{pos} + \mathbf{F}$$

This operation preserves the semantic richness of the original features while injecting multi-scale spatial priors into the representation.

Integration with Encoder Blocks. Each encoder stage in CAM-PoEm adopts a `DoubleConvWithMKPE` block, where standard convolutional layers are augmented with MKPE. This ensures that early-stage representations encode strong local positional awareness—critical for accurately identifying polyp boundaries and textures.

Overall, MKPE enables the encoder to focus on spatial saliency across multiple scales, providing strong inductive bias for structure-preserving segmentation.

Proposed Learnable Contour Extractor

Segmentation performance heavily relies on precise boundary localization, which becomes especially difficult in medical images with fuzzy or low-contrast edges. We address this by introducing a Learnable Contour Extractor module, which builds on classical edge detection and enhances it with deep feature learning.

We first apply Sobel filters along horizontal and vertical directions to compute intensity gradients:

$$\mathbf{G}_x = \mathbf{S}_x * \mathbf{F}, \quad \mathbf{G}_y = \mathbf{S}_y * \mathbf{F}$$

where $\mathbf{S}_x$ and $\mathbf{S}_y$ are standard 3×3 Sobel kernels, and $*$ denotes convolution. These gradients emphasize abrupt intensity transitions that often align with object boundaries. Next, we compute the magnitude:

$$\mathbf{G}_{mag} = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2 + \epsilon}, \quad \epsilon = 10^{-8} \tag{2}$$

To improve robustness and learn task-relevant boundaries, $\mathbf{G}_{mag}$ is refined through a lightweight enhancement network composed of depthwise separable convolutions (for parameter efficiency) and a 1×1 projection:

$$\mathbf{C}_{feat} = Conv_{1\times1}(\sigma(Conv_{3\times3}(ReLU(BN(\mathbf{G}_{mag})))))) \tag{3}$$

The resulting contour prior $\mathbf{C}_{feat} \in R^{B\times H\times W\times D}$ is aligned with Mamba's latent dimension and used to inject boundary awareness into state-space modeling.

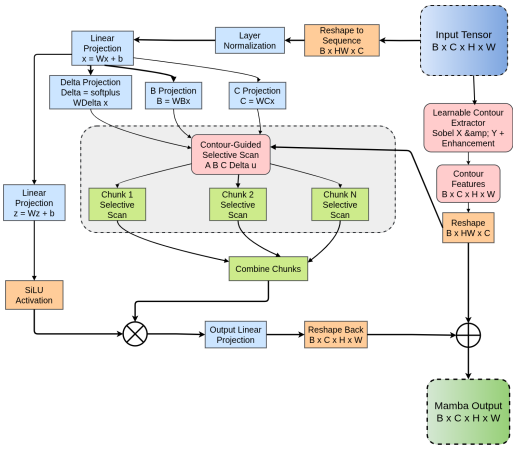

Figure 3: Detailed block diagram of Contour-Aware Mamba Block showing all components and data flow.

Algorithm 1: Contour-Aware Mamba Block (CAM-Block)

Require: Input sequence $x \in R^{B\times L\times D}$

Project input: $[x_{ssm}, x_{res}] \leftarrow Linear_{2D_{inner}}(x)$
Apply depthwise conv: $x_{conv} \leftarrow SiLU(Conv1D_{dw}(x_{ssm}))$
Compute SSM params: $[\delta, B, C] \leftarrow Split(Linear(x_{conv}))$
Smooth decay: $\delta \leftarrow Softplus(Linear(\delta))$
Extract contour prior: $C_{features} \leftarrow LayerNorm(Linear(ContourExtractor(x)))$
Contour-guided scan: $y_{ssm} \leftarrow ContourGuidedScan(x_{conv}, \delta, A, B, C, D, C_{features})$
Channel attention:
    $y_{att} \leftarrow y_{ssm} \odot \sigma(f_{fc2}(ReLU(f_{fc1}(GAP(y_{ssm})))))$
Fuse project: $y \leftarrow Dropout(Linear_D(y_{att} \odot SiLU(x_{res})))$
return y

Proposed Contour-Aware Mamba (CAM) Bottleneck

Mamba with linear scan is a state-space model that captures long-range dependencies in sequences via selective scanning with linear time complexity. However, it lacks spatial inductive bias, treating all tokens uniformly—limiting its utility in segmentation. We propose the Contour-Aware Mamba (CAM) block, which injects anatomical boundary priors to guide selective scanning. Figure 3 shows the detailed diagram.

The standard Mamba recurrence is:

$$\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{u}_t \tag{4}$$

$$\mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t + \mathbf{D}\mathbf{u}_t \tag{5}$$

where $\mathbf{u}_t$ is the input token at time t, and $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t, \mathbf{C}_t, \mathbf{D}$ are learnable matrices governing the state transitions and output projection.

Algorithm 1 outlines the forward pass of the Contour-Aware Mamba Block (CAM-Block), which integrates con-

tour structure into the selective scanning mechanism of a state-space model (SSM).

Local context extraction. Lines 1–2 split the input sequence into two branches—one for SSM computation and one for residual fusion. The SSM branch uses **depthwise separable convolution**, which preserves local structure with far fewer parameters than standard 1D conv, improving efficiency without sacrificing performance. SiLU activation adds non-linearity.

SSM parameter generation. Lines 3–4 extract decay parameters δ and transformation matrices B, C using linear layers. A `Softplus` ensures positivity in decay dynamics, enhancing stability during scanning.

Contour-aware enhancement. Line 5 introduces the core innovation: explicit contour conditioning. A lightweight `ContourExtractor` (Equation 3) module captures boundary information from the original input x. Contour features $\mathbf{C}_{features} \in R^{B\times L\times D}$ are reshaped to match the sequence dimension, and a contour weight is computed:

$$\mathbf{W}_c = \sigma(\mathbf{C}_{feat}) \tag{6}$$

Selective scanning. In Line 6, we perform `ContourGuidedScan`, a modified SSM scan, where $\mathbf{W}_c$ modulates the scan interval Δ_t, effectively increasing dwell time near edges. The scan specially incorporates $\mathbf{W}_c$, into the the output projection matrix $\mathbf{C}_t$ only, leaving the internal state dynamics intact. The modulation is applied as:

$$\mathbf{C}_t^{mod} = \mathbf{C}_t \odot (1 + \alpha \cdot \mathbf{C}_{features}) \tag{7}$$

where α is a learnable scalar controlling the strength of contour influence, and $\odot$ denotes element-wise multiplication. This ensures edge-aware semantics are emphasized in the readout phase while preserving internal recurrence via $\bar{\mathbf{A}}_t, \bar{\mathbf{B}}_t$. We modulate only $\mathbf{C}_t$ to avoid disrupting the stable and efficient dynamics of Mamba's core SSM.

Channel-wise recalibration. Lines 7–8 apply squeeze-and-excitation (SE) attention to enhance informative channels. This is lightweight due to reduced inner dimensionality and focuses on high-salience semantic content.

Residual fusion. Line 9 projects the attended output and fuses it with the residual stream through element-wise multiplication. This adds robustness while maintaining gradient flow and semantic integrity.

Deep Supervision and Final Prediction

To enhance learning and ensure better gradient propagation, we apply deep supervision across multiple decoder stages. Intermediate segmentation maps are upsampled and compared with ground truth using a combined Dice and Binary Cross Entropy (BCE) loss. The overall loss is given by:

$$\mathcal{L}_{total} = \sum_i w_i(\mathcal{L}^i_{Dice} + \mathcal{L}^i_{BCE}) \tag{8}$$

where w_i denotes the weighting coefficient for each stage.

Experiments

Datasets

We evaluate CAM-PoEm on four widely-used polyp segmentation benchmarks. Kvasir-SEG (Jha et al. 2019a)consists of 1,000 high-quality images with corresponding polyp masks. CVC-ClinicDB (Bernal et al. 2015) contains 612 clinical images with expert-annotated polyp masks. CVC-ColonDB (Bernal, Sánchez, and Vilariño 2012) provides 380 low-resolution polyp images and masks, representing challenging real-world scenarios. Finally, PolypGen21 (Ali et al. 2021) includes 807 images with masks covering polyps, normal tissue, and false-positive regions, facilitating evaluation under realistic clinical variability.

Implementation Details

The model was implemented in PyTorch and trained on NVIDIA GPUs using mixed precision for computational efficiency. Training utilized a combined loss function integrating Dice and Focal losses, weighted at 0.7 and 0.3 respectively, to effectively handle class imbalance. Optimization was done using the AdamW optimizer with an initial learning rate of 1.8e-4, alongside a cosine annealing scheduler and a 15-epoch warm-up period. Input images were resized to 352×352, and training was performed with a batch size of 16 for up to 200 epochs with early stopping patience of 30 epochs. The architecture includes 6 sequential Mamba blocks with a model dimension of 128 and a dropout rate of 0.2.

Results and Analysis

Performance of CAM-PoEM Across Datasets. We evaluated our proposed CAM-PoEM model on four widely used polyp segmentation benchmarks: Kvasir, ClinicDB, ColonDB, and PolypGen. Our model achieves consistent and competitive performance across all datasets with a lightweight architecture of only 4.96M parameters and 51.10 GFLOPs. Specifically, CAM-PoEM achieves a mean Dice score of 91.65% and mIoU of 85.35% on Kvasir, and 91.50% Dice and 85.89% IoU on ClinicDB, demonstrating strong generalization on clean and high-resolution datasets. On the more challenging ColonDB and PolypGen datasets—which contain low-quality, diverse, and complex samples—our model maintains high accuracy, with Dice scores of 88.31% and 88.27%, and IoUs of 81.83% and 83.20%, respectively. These results highlight CAM-PoEM's robustness in capturing both global context and fine-grained boundaries across varying polyp morphologies.

Quantitative Comparison with State-of-the-Art Methods. We compare CAM-PoEM with recent state-of-the-art segmentation models in Table 1. Our model achieves top-tier performance while being significantly more efficient in terms of parameter count and computational cost. Notably, CAM-PoEM outperforms larger models such as UACANet-L (69.16M) and SSFormer-L (66.22M) on ColonDB and PolypGen in terms of Dice and IoU, despite having an order of magnitude fewer parameters. On ColonDB, our model

Methods	Params (M)	FLOPs (G)	Kvasir		ClinicDB		ColonDB		PolypGen	
			mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
UNet (Ronneberger, Fischer, and Brox 2015c)	34.53	65.53G	81.80	74.60	82.30	75.50	51.20	44.40	-	-
UNet++ (Zhou et al. 2018)	25.09	84.30	82.10	74.30	79.40	72.90	48.30	41.00	-	-
AttnUNet (Oktay et al. 2018)	34.88	66.64	83.49	76.84	86.46	79.66	52.33	45.10	-	-
DeepLabv3+ (Chen et al. 2018)	39.76	14.92	89.06	83.72	91.24	84.91	65.32	57.85	-	-
PraNet (Fan et al. 2020b)	32.55	221.90	89.80	84.00	89.90	84.90	70.90	64.00	-	-
CaraNet (Lou, Guan, and Loew 2023)	46.64	11.48	89.75	83.25	91.70	85.34	65.55	57.72	-	-
UACANet-L (Kim, Lee, and Kim 2021)	69.16	31.51	90.02	85.25	91.02	84.76	69.81	63.10	-	-
ACSNet (Liu et al. 2021)	46.02	29.45	89.80	83.80	88.20	82.60	71.60	64.90	-	-
SFA (Luo et al. 2019)	-	-	72.30	61.10	70.00	60.70	46.90	34.70	-	-
Yolo-SAM2 (Mansoori et al. 2024)	-	-	86.60	76.40	-	-	-	-	80.80	67.80
LGPS (Tesema et al. 2025)	–	–	-	-	-	-	-	-	72.99	78.67
CAM-PoEM (Ours)	**4.96**	51.10	**91.65**	**85.35**	**91.50**	**85.89**	**88.31**	**81.83**	**88.27**	**83.20**

Table 1: Quantitative comparison of CAM-PoEM with state-of-the-art models on four benchmark datasets: Kvasir, ClinicDB, ColonDB, and PolypGen. The reported results for baseline methods are sourced from their respective papers or public benchmarks and are included solely for reference, as variations in dataset splits, preprocessing pipelines, or image resolutions may exist. A dash ('–') denotes that the corresponding result was not available or not reported in the source.

achieves the highest Dice (88.31%) and IoU (81.83%), and performs on par with leading models on other datasets.

Qualitative and Failure Case Analysis

Qualitative Analysis. We visualize segmentation results in Figure 4 to qualitatively compare CAM-PoEm with state-of-the-art baselines, including UNet, Attention-UNet, and SwinUNet. Evaluation is performed on representative samples from the **Kvasir** and **PolypGen** datasets. CAM-PoEm consistently produces crisp and accurate contours, effectively capturing both small and multiple polyps per frame—especially in Kvasir. In PolypGen, our model adapts robustly to variations in lighting, tissue texture, and polyp morphology, outperforming others in cases with low contrast or irregular shapes. Baseline models tend to under-segment or overlook faint boundaries, particularly in multi-polyp or cluttered backgrounds. These results underscore the strength of our boundary-sensitive global modeling under real-world clinical variations.

Failure Case Analysis. We further examined cases from the **CVC-ClinicDB** dataset to evaluate potential failure modes (Figure 5). Unlike Kvasir or PolypGen, several ground truth masks in ClinicDB were either overly coarse or poorly aligned with visible polyp boundaries. In such scenarios, CAM-PoEm's output—although visually and anatomically plausible—exhibited reduced overlap with the annotated ground truths, leading to lower Dice or mIoU scores. This mismatch highlights a key limitation: high structural sensitivity may conflict with sub-optimal label quality, resulting in underreported performance despite clinically valid predictions. Such observations suggest the need for uncertainty-aware or soft-label evaluation in future polyp segmentation benchmarks.

Ablation Study

We conduct comprehensive ablation studies to evaluate the individual and combined contributions of the two key components of *CAM-PoEm*: (i) the **Contour-Aware Mamba (CAM)** bottleneck, and (ii) the **Multi-Kernel Positional Embedding (MKPE)** encoder. Experiments are conducted on the Kvasir-SEG dataset under consistent training settings, and performance is reported as mean Intersection over Union (mIoU) and Dice coefficient (mDice) across three runs, along with standard deviation (SD).

Ablation of Contour-Aware Mamba Bottleneck To assess the effectiveness of our proposed CAM bottleneck, we compare the segmentation performance of models with the standard Mamba block versus the contour-aware variant. As shown in Table 2, introducing contour cues into the Mamba scan mechanism enhances boundary sensitivity, leading to better segmentation results—especially when combined with MKPE. While CAM slightly performs better than standard Mamba in isolation, it yields notable improvement in the final model, confirming that contour-guided global modeling complements local spatial features.

Ablation of Multi-Kernel Positional Embedding (MKPE) Encoder We evaluate the contribution of MKPE by replacing the standard CNN encoder with our proposed multi-kernel positional embedding module. As presented in Table 3, MKPE substantially improves spatial representation and performance when used in conjunction with both standard Mamba and CAM bottlenecks. This highlights the value of multi-scale context modeling at the encoder stage, which captures heterogeneous polyp structures more effectively.

Ablation on making different mamba parameters Contour Aware

The Mamba architecture includes four learnable parameters: **A, B, C**, and **D**. In our Contour-Aware Mamba (CAM) design, we selectively introduce inductive bias only in the **C** parameter, as it directly influences the input-to-state projection and spatial transformation. To evaluate the impact of modifying different parameters, we conducted an ablation by integrating contour-awareness into various combinations.

Why only C? The **C** parameter projects spatial features from input to state space. By incorporating contour priors into this mapping, the model can better preserve structural boundaries without over-parameterizing. In contrast, modi-

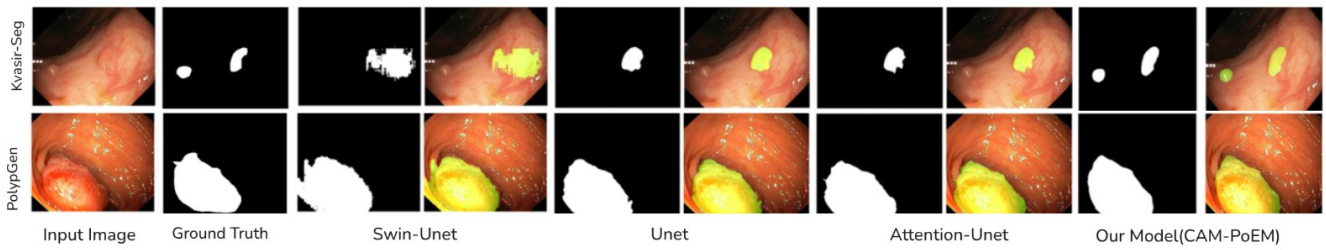

Figure 4: Qualitative comparison of CAM-PoEm against UNet, Attention-UNet, and SwinUNet on challenging samples from Kvasir and PolypGen. CAM-PoEm achieves sharper contours, detects small polyps, and maintains segmentation fidelity under visual ambiguity.

Model Variant	**Dice**	**Gain**
CNN + Mamba	0.8873 ± 0.0044	Reference
CNN + CAM	0.8891 ± 0.0037	+0.2%
MKPE + Mamba	0.8834 ± 0.0039	—
MKPE + CAM (Ours)	**0.9165 ± 0.0035**	**+3.3%**

Table 2: Ablation study highlighting the contribution of the proposed **Contour-Aware Mamba (CAM)**. CAM enhances boundary-sensitive global modeling and consistently improves performance, particularly when combined with the MKPE encoder.

Model Variant	**Dice**	**Gain**
CNN + Mamba	0.8873 ± 0.0044	Reference
MKPE + Mamba	0.8834 ± 0.0039	-0.4%
CNN + CAM	0.8851 ± 0.0037	—
MKPE + CAM (Ours)	**0.9165 ± 0.0035**	**+2.2%**

Table 3: Ablation study highlighting the contribution of the **Multi-Kernel Positional Embedding (MKPE)**. MKPE enhances multi-scale local representation and synergizes effectively with CAM to achieve superior segmentation performance.

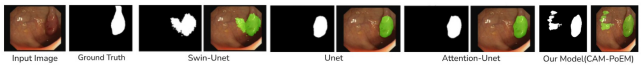

Figure 5: Failure case from CVC-ClinicDB. CAM-PoEm produces clinically reasonable segmentation masks, but diverges from ground truth due to imprecise annotations—highlighting dataset-dependent limitations.

Modified Params	**mIoU**	**mDice**	**Total Epochs**
CAM-A (A only)	0.8257	0.8929	120
CAM-B (B only)	0.8100	0.8819	93
CAM-A-B-D	0.8122	0.8817	101
CAM-C (Ours)	**0.8535**	**0.9161**	–
CAM-C+D	0.8320	0.8900	–

Table 4: Ablation on modifying different Mamba parameters with contour-aware bias. CAM-C (modifying only **C**) yields the best results.

fying other parameters such as **A** (state transition), **B** (input mixing), or **D** (output projection) either introduces noise or results in diminished boundary sensitivity. Combining multiple parameters (CAM-A-B-D or CAM-C+D) adds unnecessary complexity and leads to lower performance. These findings validate our design choice of isolating the inductive bias to only **C**, ensuring both efficiency and accuracy in boundary-sensitive segmentation tasks.

Discussion

Global Modeling with Bottleneck-Positioned CAM. CAM-PoEm demonstrates that injecting *learnable contour priors* into state-space models (SSMs) enables boundary-sensitive sequence modeling without compromising global context. By placing the *Contour-Aware Mamba (CAM)* block in the bottleneck—where spatial resolution is reduced but semantic abstraction is highest—the model leverages low-dimensional, high-level features for global reasoning, optimizing both accuracy and efficiency. This design aligns with encoder-decoder principles where context modeling at the bottleneck yields maximal receptive field coverage at minimal computational cost. The CAM block modulates Mamba's output projection with contour priors while preserving its stable recurrent dynamics, making it suitable for memory-constrained clinical segmentation tasks.

Contour-Aware Mamba via Linear Token Scanning. The introduction of contour-awareness into Mamba's linear-time selective scan addresses a core limitation of existing SSM-based vision models: the uniform treatment of spatial tokens. CAM-PoEm injects edge-weighted priors into the projection matrix C_t, enabling the model to allocate greater emphasis to boundary-adjacent regions during output generation. This preserves Mamba's $\mathcal{O}(N)$ complexity while enabling morphology-aware token processing—a crucial improvement for segmenting ambiguous or irregular lesions in medical imaging.

Conclusion

Combined with the *Multi-Kernel Positional Embedding (MKPE)* encoder, which captures local structure at multiple receptive fields, CAM-PoEm achieves state-of-the-art segmentation performance across four polyp benchmarks using only 4.96M parameters. Ablation studies confirm the *synergistic effect* between CAM and MKPE, delivering both global coherence and fine-grained boundary localization.

References

Adame, D.; Nunez, J. A.; Vazquez, F.; Gurrola, N.; Li, H.; Tang, H.; Fu, B.; and Gu, P. 2025. Topo-VM-UNetV2: Encoding Topology into Vision Mamba UNet for Polyp Segmentation.

Ali, S.; Jha, D.; Ghatwary, N.; Realdon, S.; Cannizzaro, R.; Salem, O.; Lamarque, D.; Daul, C.; Ånonsen, K. V.; Riegler, M.; Halvorsen, P.; Rittscher, J.; de Lange, T.; and East, J. 2021. PolypGen: A Multi-Center Polyp Detection and Segmentation Dataset for Generalisability Assessment. Multi-center endoscopic polyp dataset with 1,000+ images from 6 medical centers, arXiv:2106.04463.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43: 99–111. Official journal of the Computerized Medical Imaging Society.

Bernal, J.; Sánchez, J.; and Vilariño, F. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9): 3166–3182. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2021. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv:2102.04306.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv:1802.02611.

Djinbachian, R.; Iratni, R.; Durand, M.; Marques, P.; and von Renteln, D. 2020. Rates of Incomplete Resection of 1- to 20-mm Colorectal Polyps: A Systematic Review and Meta-Analysis. *Gastroenterology*, 159(3): 904–914.e12.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2023. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *CAAI Artificial Intelligence Research*, 9150015. Article ID: 9150015.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published as a conference paper at ICLR 2021, arXiv:2010.11929.

Duc, N. T.; Oanh, N. T.; Thuy, N. T.; Triet, T. M.; and Dinh, V. S. 2022. ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation. *IEEE Access*, 10: 80575–80586.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020a. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. Published in MICCAI 2020, arXiv:2006.11392.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. arXiv:2006.11392.

Gao, W. 2024. MEP: Multiple Kernel Learning Enhancing Relative Positional Encoding Length Extrapolation. arXiv:2403.17698.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Published at ICLR 2024, arXiv:2312.00752.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

Haggar, F. A.; and Boushey, R. P. 2009. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4): 191–197.

Ho, Q.-H.; Nguyen, T.-N.-Q.; Tran, T.-T.; and Pham, V.-T. 2025a. LiteMamba-Bound: A lightweight Mamba-based model with boundary-aware and normalized active contour loss for skin lesion segmentation. *Methods*, 235: 10–25.

Ho, Q.-H.; Nguyen, T.-N.-Q.; Tran, T.-T.; and Pham, V.-T. 2025b. LiteMamba-Bound: A Lightweight Mamba-Based Model with Boundary-Aware and Normalized Active Contour Loss for Skin Lesion Segmentation. *Methods*, 235: 10–25.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2019a. Kvasir-SEG: A Segmented Polyp Dataset. Medical image segmentation dataset, arXiv:1911.07069.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; de Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019b. ResUNet++: An Advanced Architecture for Medical Image Segmentation. arXiv:1911.07067.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; de Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019c. ResUNet++: An Advanced Architecture for Medical Image Segmentation. Published in IEEE Access 2021, arXiv:1911.07067.

Karthikha, R.; Jamal, D. N.; and Rafiammal, S. S. 2024. An approach of polyp segmentation from colonoscopy images using Dilated-U-Net-Seg – A deep learning network. *Biomedical Signal Processing and Control*, 93: 106197.

Kim, T.; Lee, H.; and Kim, D. 2021. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2167–2175. ACM.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. Meta AI Research, arXiv:2304.02643.

Li, L.; Lian, S.; Luo, Z.; Wang, B.; and Li, S. 2024a. Contour-aware consistency for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 89: 105694.

Li, L.; Lian, S.; Luo, Z.; Wang, B.; and Li, S. 2024b. Contour-Aware Consistency for Semi-Supervised Medical Image Segmentation. *Biomedical Signal Processing and Control*, 89: 105694.

Li, W.; Xiong, X.; Li, S.; and Fan, F. 2024c. HybridVPS: Hybrid-Supervised Video Polyp Segmentation Under Low-Cost Labels. *IEEE Signal Processing Letters*, 31: 111–115.

Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; and He, Z. 2024. A Survey of Visual Transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 7478–7498. Early Access.

Liu, Z.; Wang, L.; Zhang, Q.; Tang, W.; Yuan, J.; Zheng, N.; and Hua, G. 2021. ACSNet: Action-Context Separation Network for Weakly Supervised Temporal Action Localization. arXiv:2103.15088.

Long, J.; Shelhamer, E.; and Darrell, T. 2014. Fully Convolutional Networks for Semantic Segmentation. *arXiv preprint*, abs/1411.4038. Published in CVPR 2015.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038.

Lou, A.; Guan, S.; and Loew, M. 2023. CaraNet: context axial reverse attention network for segmentation of small medical objects. *Journal of Medical Imaging*, 10(01).

Luo, S.; Li, X.; Zhu, R.; and Zhang, X. 2019. SFA: Small Faces Attention Face Detector. *IEEE Access*, 7: 171609–171620.

Ma, J.; Li, F.; and Wang, B. 2024. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv:2401.04722.

Mansoori, M.; Shahabodini, S.; Abouei, J.; Plataniotis, K. N.; and Mohammadi, A. 2024. Self-Prompting Polyp Segmentation in Colonoscopy using Hybrid Yolo-SAM 2 Model. arXiv:2409.09484.

Mau, T.-H. N.; Trinh, Q.-H.; Bui, N.-T.; Tran, M.-T.; and Nguyen, H.-D. 2023. Multi Kernel Positional Embedding ConvNeXt for Polyp Segmentation. arXiv:2301.06673.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; Glocker, B.; and Rueckert, D. 2018. Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999.

Park, K.-B.; and Lee, J. Y. 2022. SwinE-Net: Hybrid Deep Learning Approach to Novel Polyp Segmentation Using Convolutional Neural Network and Swin Transformer. *Journal of Computational Design and Engineering*, 9(2): 616–632.

Pooler, B. D.; Kim, D. H.; Matkowskyj, K. A.; Newton, M. A.; Halberg, R. B.; Grady, W. M.; Hassan, C.; and Pickhardt, P. J. 2023. Growth rates and histopathological outcomes of small (6-9 mm) colorectal polyps based on CT colonography surveillance and endoscopic removal. *Gut*, 72(12): 2321–2328.

Qiu, Z.; Wang, Z.; Zhang, M.; Xu, Z.; Fan, J.; and Xu, L. 2022. BDG-Net: Boundary Distribution Guided Network for Accurate Polyp Segmentation. In Išgum, I.; and Colliot, O., eds., *Medical Imaging 2022: Image Processing*. SPIE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015a. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015b. U-Net: Convolutional Networks for Biomedical Image Segmentation. Published in MICCAI 2015, arXiv:1505.04597.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015c. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Ruan, J.; Li, J.; and Xiang, S. 2024. VM-UNet: Vision Mamba UNet for Medical Image Segmentation. arXiv:2402.02491.

Safarov, S.; and Whangbo, T. K. 2021. A-DenseUNet: Adaptive Densely Connected UNet for Polyp Segmentation in Colonoscopy Images with Atrous Convolution. *Sensors*, 21(4): 1441. Article Number: 1441.

Shao, H.; Zhang, Y.; and Hou, Q. 2023. Polyper: Boundary Sensitive Polyp Segmentation.

Tang, H.; Huang, G.; Cheng, L.; Yuan, X.; Tao, Q.; Chen, X.; Zhong, G.; and Yang, X. 2024. RM-UNet: UNet-like Mamba with Rotational SSM Module for Medical Image Segmentation. *Signal, Image and Video Processing*, 18: 8427–8443.

Tesema, F. B.; Manzanares, A. G.; Cui, T.; Zhang, Q.; Solomon, M.; and He, S. 2025. LGPS: A Lightweight GAN-Based Approach for Polyp Segmentation in Colonoscopy Images. arXiv:2503.18294.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. Published in NeurIPS 2017.

Wang, J.; Chen, J.; Chen, D.; and Wu, J. 2024a. LKM-UNet: Large Kernel Vision Mamba UNet for Medical Image Segmentation. arXiv:2403.07332.

Wang, Z.; Li, T.; Liu, M.; Jiang, J.; and Liu, X. 2025. DCAT-Net: Polyp Segmentation with Deformable Convolution and Contextual-Aware Attention Network. *BMC Medical Imaging*, 25.

Wang, Z.; Zheng, J.-Q.; Zhang, Y.; Cui, G.; and Li, L. 2024b. Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation. arXiv:2402.05079.

Wang, Z.; Zheng, J.-Q.; Zhang, Y.; Cui, G.; and Li, L. 2024c. Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation.

Xie, J.; Liao, R.; Zhang, Z.; Yi, S.; Zhu, Y.; and Luo, G. 2024a. ProMamba: Prompt-Mamba for polyp segmentation. arXiv:2403.13660.

Xie, J.; Liao, R.; Zhang, Z.; Yi, S.; Zhu, Y.; and Luo, G. 2024b. ProMamba: Prompt-Mamba for Polyp Segmentation.

Xu, Z.; Tang, F.; Chen, Z.; Zhou, Z.; Wu, W.; Yang, Y.; Liang, Y.; Jiang, J.; Cai, X.; and Su, J. 2024. Polyp-Mamba: Polyp Segmentation with Visual Mamba. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 510–521. Cham: Springer Nature Switzerland. ISBN 978-3-031-72111-3.

Yeung, M.; Sala, E.; Schönlieb, C.-B.; and Rundo, L. 2021. Focus U-Net: A Novel Dual Attention-Gated CNN for Polyp Segmentation During Colonoscopy.

You, C.; Jiao, L.; Li, L.; Liu, X.; Liu, F.; Ma, W.; and Yang, S. 2025. Contour Knowledge-Aware Perception Learning for Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5): 4560–4575.

Zhang, M.; Yu, Y.; Gu, L.; Lin, T.; and Tao, X. 2024. VM-UNET-V2: Rethinking Vision Mamba UNet for Medical Image Segmentation.

Zhao, X.; Tang, F.; Wang, X.; and Xiao, J. 2024. SFC: Shared Feature Calibration in Weakly Supervised Semantic Segmentation. arXiv:2401.11719.

Zhou, X.; Wu, G.; Sun, X.; Hu, P.; and Liu, Y. 2024. Attention-Based Multi-Kernelized and Boundary-Aware Network for Image Semantic Segmentation. *Neurocomputing*, 597: 127988.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. arXiv:1807.10165.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model.