

PolySAGN: Hierarchical Multi-Scale Representation Learning with Scale-Specific Attention for Accurate Polyp Segmentation

Wenqi Zhang, Yue Zhang, Muhammad Fayaz, L. Minh Dang,  
Tan N. Nguyen, Hyeonjoon Moon

PII: S0925-2312(25)03010-3  
DOI: <https://doi.org/10.1016/j.neucom.2025.132338>  
Reference: NEUCOM 132338

To appear in: *Neurocomputing*

Received Date: 1 September 2025  
Revised Date: 24 November 2025  
Accepted Date: 3 December 2025

Please cite this article as: Zhang W, Zhang Y, Fayaz M, Dang LM, Nguyen TN, Moon H, PolySAGN: Hierarchical Multi-Scale Representation Learning with Scale-Specific Attention for Accurate Polyp Segmentation, *Neurocomputing* (2025), doi: <https://doi.org/10.1016/j.neucom.2025.132338>.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: <https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article>. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

1. Proposed a novel three-stream hierarchical network with DASPP and MHSA integration.
2. Designed dual-modal attention modules using diverse pooling and adaptive channel refinement.
3. Developed a scale-specific attention-guided fusion for precise polyp boundary delineation.
4. Validated on five benchmarks, achieving superior mDice and mIoU over state-of-the-art.
5. Ablation and cross-domain studies confirmed effectiveness and strong generalization.

# PolySAGN: Hierarchical Multi-Scale Representation Learning with Scale-Specific Attention for Accurate Polyp Segmentation

Wenqi Zhang<sup>a</sup>, Yue Zhang<sup>a</sup>, Muhammad Fayaz<sup>a</sup>, L. Minh Dang<sup>b,c,d</sup>, Tan N. Nguyen<sup>e</sup> and Hyeonjoon Moon<sup>a</sup>

<sup>a</sup>Department of Computer Science and Engineering, Sejong University, Seoul, 05006, Republic of Korea

<sup>b</sup>The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam

<sup>c</sup>Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam

<sup>d</sup>Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, 05006, Republic of Korea

<sup>e</sup>Department of Architectural Engineering, Sejong University, Seoul, 05006, Republic of Korea

## ARTICLE INFO

### Keywords:

Polyp segmentation  
Colorectal Cancer  
Multi-Scale Attention  
Hierarchical Learning  
Colonoscopy Analysis  
Multi-Stream Network.

## ABSTRACT

Accurate polyp segmentation in endoscopic imagery remains critical for early colorectal cancer detection, requiring robust methods that handle diverse morphological variations and challenging imaging conditions. In this paper, we present a novel hierarchical multi-scale representation learning framework incorporating scale-specific attention mechanisms for precise polyp segmentation called PolySAGN. Our approach addresses fundamental limitations through four key innovations. Initially, we propose a hierarchical multi-stream architecture combining EfficientNet-B7 with parallel Dilated Atrous Spatial Pyramid Pooling (DASPP) and Multi-Head Self-Attention (MHSA) pathways. The integration of DASPPs and MHSA enables simultaneous capture of multi-scale contextual information and global spatial dependencies for enhanced polyp representation. Furthermore, we introduce four complementary pooling operations in the spatial attention and adaptive channel refinement. Additionally, we utilize the scale-specific attention-guided fusion utilizing Convolutional Block Attention Modules (CBAM). Finally, we provide comprehensive experimental validation across five benchmark datasets. Extensive ablation studies validate each architectural component's effectiveness, with backbone analysis and attention mechanism studies. Extensive experiments on ClinicDB, ETIS, ColonDB, Kvasir-SEG, and Endoscene demonstrate state-of-the-art performance with mDice scores of 0.939, 0.809, 0.812, 0.927, and 0.902, respectively, consistently outperforming existing methods.

## 1. Introduction

Colorectal cancer remains one of the leading causes of cancer-related deaths worldwide, and early detection of polyps during colonoscopy plays a crucial role in reducing its incidence [1]. Despite unprecedented advances in medical imaging (MI) technologies, many patients are still diagnosed with advanced metastatic disease due to socioeconomic barriers, healthcare accessibility issues, and insufficient screening programs. Advanced MI has evolved into an indispensable tool for diagnostic assessment and treatment planning, incorporating various imaging modalities that provide critical clinical information [2, 3, 4]. Among the diverse analytical approaches in MI, image segmentation constitutes a foundational technique for localizing and delineating anatomical structures and pathological conditions, particularly in medical imaging applications and endoscopic examinations. This segmentation process involves intricate challenges in achieving accurate anatomical boundary determination necessary for extracting crucial morphological features. Therefore, polyp segmentation (PS) has emerged as a vital application area, focusing on precise detection and accurate boundary extraction of polypoid formations in colonoscopic images. PS remains a challenging domain within computational medical image analysis, where successful implementation offers significant potential for improving early detection capabilities and enabling timely therapeutic measures in colorectal cancer management.

Traditional PS approaches primarily rely on basic feature extraction techniques, incorporating geometric attributes [5], textural characteristics [6], and simple linear iterative clustering superpixel methodologies [7]. However, these conventional methods often produce suboptimal segmentation results and exhibit limited generalization capabilities across varied clinical settings. The advent of deep learning frameworks has triggered remarkable advancements in PS applications, enabling computational systems to learn semantically rich representations from medical imagery, enhance diagnostic accuracy, and demonstrate improved robustness across heterogeneous datasets and clinical applications.

Particularly, encoder-decoder architectural designs [8] have gained significant popularity due to their ability to leverage hierarchical feature hierarchies, facilitating the generation of high-resolution segmentation predictions. More recently, the integration of Transformer-based architectures has enabled the development of multiple high-performing segmentation frameworks [9, 10], incorporating advanced global context modeling mechanisms into medical image segmentation pipelines [11, 12].

Leading approaches in PS include PraNet, Polyp-PVT, UACANet, and CASCADE, each achieving outstanding performance on standard benchmark evaluations [13, 14, 15, 16]. The PraNet architecture [13] presents a parallel reverse attention network design optimized for accurate PS in colonoscopic imagery. This methodology utilizes a parallel partial decoder strategy for hierarchical feature fusion combined with a reverse attention mechanism for improved boundary delineation. Furthermore, a transformer-enhanced PS approach [14] was proposed, featuring three specialized components to effectively combine multi-scale representations, reduce noise artifacts, and enhance robustness under challenging imaging scenarios, outperforming traditional CNN-based methodologies. UACANet [15] introduces an uncertainty-guided PS framework that enhances feature representation through uncertainty-augmented saliency computation integrated within a modified U-Net structure, demonstrating superior performance across five benchmark datasets. CASCADE [16] proposes an attention-guided decoder for transformer-based medical image segmentation that enhances both global dependencies and local contextual modeling through attention gating and convolutional attention components.

Despite these significant contributions, existing PS methodologies face persistent challenges in effectively balancing global context understanding with fine-grained local feature preservation. Current approaches often struggle to capture multi-scale polyp variations while maintaining computational efficiency for real-time clinical deployment. Additionally, the integration of complementary feature representations from different architectural paradigms remains underexplored, limiting the potential for enhanced segmentation accuracy. To address these limitations, this study proposes PolySAGN, a novel Poly-scale Spatial Attention-Guided Network that introduces a hierarchical multi-stream architecture integrated with advanced dual attention and scale-specific fusion mechanisms. Unlike previous approaches, PolySAGN uniquely combines DASPP-based multi-scale contextual encoding with dual-modal attention modules to achieve both fine-grained boundary preservation and global semantic consistency. Furthermore, the proposed attention-guided fusion strategy allows interpretable feature aggregation across scales, providing not only quantitative performance gains but also enhanced clinical interpretability [17].

## 1.1. Research Gaps and Limitations

Despite remarkable progress in polyp segmentation, several fundamental limitations in current methodologies create opportunities for significant improvements that directly motivate our proposed PolySAGN:

- Inadequate Hierarchical Multi-Stream Feature Integration:** Existing approaches [13, 15] predominantly utilize single-pathway architectures that fail to exploit the full potential of parallel hierarchical feature extraction. Current methods lack an effective mechanism to simultaneously process features through multiple streams with different receptive field characteristics, limiting their ability to capture polyps across diverse scale variations and morphological complexities.
- Insufficient Dual-Modal Attention Mechanisms:** Contemporary frameworks [18, 16, 19] employ elementary attention strategies that inadequately address the complementary benefits of spatial and channel attention integration. Existing methods often lack optimal attention modules that can simultaneously perform multi-pooling spatial attention and adaptive channel attention refinement, limiting their discriminative capability in complex endoscopic environments.
- Suboptimal Attention-Guided Feature Fusion:** Current state-of-the-art (SOTA) approaches [13, 14, 15, 16, 20] demonstrate limited effectiveness in combining multi-scale hierarchical features through attention-guided fusion strategies. The absence of systematic integration mechanisms that leverage both convolutional block attention modules (CBAM) and scale-specific attention refinement restricts the models' ability to optimally balance local detail preservation with global context understanding, particularly crucial for accurate polyp boundary delineation.

This work addresses the identified limitations through the following novel contributions:

- Hierarchical Multi-Stream Architecture with DASPP-MHSA Integration:** We propose a novel three-stream hierarchical feature extraction network that effectively combines the EfficientB7 backbone with parallel

processing pathways. Each stream integrates Dilated Atrous Spatial Pyramid Pooling (DASPP) modules with Multi-Head Self-Attention (MHSA) mechanisms, enabling simultaneous capture of multi-scale contextual information and global spatial dependencies for enhanced polyp representation across diverse morphological variations.

2. **Innovative Dual-Modal Attention Mechanisms:** We introduce optimized spatial and channel attention modules that significantly enhance feature discriminability. The spatial attention employs four complementary pooling operations (average, maximum, median, and variance pooling) to capture diverse statistical properties. Moreover, the channel attention mechanism adaptively refines inter-channel relationships, resulting in superior performance with varying illumination and texture characteristics.
3. **Scale-Specific Attention-Guided Feature Fusion Strategy:** We develop an advanced fusion network that systematically integrates multi-scale hierarchical features through CBAM and scale-specific attention refinement. This approach optimally balances local detail preservation with global context understanding, enabling precise polyp boundary delineation.
4. **Comprehensive Evaluation and Superior Performance:** We conduct extensive experimental validation across five benchmark datasets (Endoscene, ClinicDB, ColonDB, ETIS, and Kvasir-SEG), demonstrating SOTA performance. Our PolySAGN achieves superior mDice and mIoU scores compared to existing methods. Comprehensive ablation studies validate the effectiveness of each proposed component and cross-domain generalization capabilities for diverse clinical scenarios.

The organization of this manuscript proceeds as follows: Section 2 provides a thorough examination of existing literature and foundational approaches. Section 3 presents the architectural design and technical implementation of the PolySAGN. Section 4 demonstrates comprehensive experimental validation through rigorous benchmarking and comparative analysis against state-of-the-art methodologies. Section 5 summarizes the research contributions and discusses future research directions in polyp segmentation.

## 2. Literature review

The broader computer vision community has recently introduced universal segmentation frameworks that aim to handle diverse segmentation tasks and datasets within a single promptable model. Segment Anything itself is formulated as a general-purpose segmentation foundation model trained on over one billion masks, exhibiting strong zero-shot generalization across a wide range of natural-image benchmarks [21]. Building on this paradigm, SEEM [22] and X-Decoder [23] unify interactive, generic, referring, and vision–language-aware segmentation within a single decoding space by taking multi-modal prompts (e.g., points, boxes, scribbles, text) as inputs and predicting pixel-level masks and language tokens in a shared representation space. OneFormer [24] further extends this line of work by employing a single architecture conditioned on task tokens to jointly perform semantic, instance, and panoptic segmentation, demonstrating that a single universal model can rival or surpass task-specific networks on several benchmarks. In medical imaging, similar ideas have led to universal or cross-domain segmentation models such as MedSAM [25], SAM-Med2D [26], and UniverSeg [27], which adapt large foundation models to multiple imaging modalities and anatomical structures, enabling few-shot or zero-shot transfer across new tasks. A series of studies dedicated to adapting its powerful prior knowledge to polyp segmentation tasks. Zhao et al. [28] proposed a novel SAM-driven weakly-supervised framework that fosters a collaborative learning process between the segmentation network and SAM to enhance performance, introducing a Cross-aware Feature Aggregation Network (CFANet) to effectively integrate cross-level features and global cues. Sun et al. [29] innovatively introduced visual reference prompting to SAM, leveraging annotated reference images to comprehend specific objects and achieve accurate segmentation in target images. The SAM-EG [30] framework incorporated an Edge Guidance module to capture edge information from inputs and merge it with learned segmentation features and SAM embeddings for refined boundary delineation. While these universal frameworks are highly flexible, they typically incur substantial computational overhead and, in practice, still struggle to match carefully designed task-specific architectures on small, low-contrast, or subtle polyp regions motivating our focus on a polyp-specialized yet computationally efficient architecture in this work.

The incorporation of transformer architectures into medical image segmentation has fundamentally transformed dense prediction tasks through effective self-attention mechanisms that effectively capture global contextual relationships across spatial dimensions. The pioneering Vision Transformer (ViT) [31] established foundational transformer principles for visual understanding by treating images as sequences of patches, demonstrating that attention-based

architectures could rival convolutional approaches in image classification tasks. Building upon this foundation, hierarchical variants such as Hierarchical ViT (HVT) [32] and pooling-based transformers [33] significantly enhanced computational efficiency through progressive token reduction strategies and spatial downsampling mechanisms, making transformer architectures more practical for high-resolution medical imaging applications. Advanced architectures have emerged to address ViT's inherent limitations in capturing fine-grained local features essential for medical segmentation tasks. The Tokens-to-Token ViT (T2T-ViT) [34] addresses these shortcomings through progressive token aggregation mechanisms that preserve local structural information while reducing computational overhead. Similarly, the TNT framework [35] implements dual-level attention mechanisms that simultaneously process global context and local patch details, proving particularly effective for applications requiring fine-grained visual analysis. Pyramid Vision Transformer (PVT) variants [36, 37] have demonstrated that pure transformer architectures can achieve superior performance in segmentation tasks through hierarchical multi-scale feature extraction, establishing new benchmarks for medical image analysis applications. Compared to conventional convolutional architectures, these transformer-based variants provide stronger global contextual reasoning; However, their reliance on patch-based processing and limited integration with convolutional priors may restrict their ability to capture detailed local boundaries that are critical in medical segmentation tasks [38].

Deep learning methodologies for polyp segmentation have undergone remarkable evolution, transitioning from traditional convolutional approaches to sophisticated hybrid architectures that leverage both local and global feature representations. Early hybrid CNN-Transformer frameworks, exemplified by the work of Cai et al. [39], pioneered the combination of local feature extraction capabilities with global modeling mechanisms, demonstrating the synergistic potential of integrating complementary architectural paradigms. These approaches typically employ shallow convolutional layers for detailed texture analysis while utilizing deeper transformer components for spatial relationship modeling and long-range dependency capture. Multi-scale processing strategies have been extensively investigated to address the inherent scale variation challenges in polyp segmentation. Sun et al. [40] incorporated atrous convolutions with varying dilation rates to achieve multi-scale semantic understanding while preserving spatial resolution, effectively addressing the fundamental trade-off between receptive field expansion and detail preservation. Complementary approaches like Psi-Net [41] introduced sophisticated multi-task learning frameworks that simultaneously generate segmentation masks, boundary predictions, and distance transforms through parallel decoder branches, resulting in more coherent and spatially consistent predictions. These multi-objective optimization strategies have proven particularly effective in handling complex polyp morphologies and challenging imaging conditions. Notably, this hybrid paradigm has been effectively extended to other medical imaging domains, such as breast tumor segmentation, where multi-scale fusion strategies directly align with the goals of robust lesion delineation. For instance, HAU-Net [42] embeds an L-G transformer block in skip connections to capture long-range context while preserving local details, and employs a cross-attention block (CAB) in the decoder for inter-layer feature interaction. Similarly, HCTNet [43] integrates Transformer Encoder Blocks (TEBlocks) in the encoder and a spatial-wise cross-attention (SCA) module in the decoder to align semantics across encoding and decoding stages, enhanced by residual connections for multi-scale feature aggregation. While these methods demonstrate improved multi-scale processing, their fusion mechanisms often lack sophistication in effectively combining features across different scales and architectural paradigms, limiting their potential for optimal segmentation performance [44].

Attention-driven architectures have demonstrated remarkable improvements in boundary accuracy and overall segmentation performance. PraNet [13] pioneered the implementation of reverse attention mechanisms that leverage global contextual information derived from parallel partial decoders to progressively refine segmentation boundaries through iterative attention focusing. This approach enables the model to gradually eliminate background interference while enhancing polyp-specific features. Building on similar principles, PolypNet [45] integrated dual-tree wavelet-based pooling with local gradient-weighted embedding techniques, effectively reducing false positive detections in high-intensity regions and improving robustness against noise artifacts commonly encountered in endoscopic imagery. Recent innovations have increasingly focused on clinical deployment optimization and real-time performance requirements. ColonSegNet [46] achieved remarkable real-time performance metrics with 0.8206 Dice coefficient at 182.38 frames per second on the Kvasir-SEG dataset, demonstrating the feasibility of clinical integration without compromising segmentation accuracy. This work established important benchmarks for the accuracy-speed trade-off crucial for practical deployment. MSEG [47] pursued architectural simplification strategies, replacing complex backbone networks with more efficient alternatives while maintaining competitive accuracy levels, thus reducing computational requirements for resource-constrained clinical environments. Although attention-driven architectures

such as PraNet and PolypNet enhance boundary refinement, their attention operations are often confined to single-scale or decoder-specific levels, which limits their capacity to integrate features across resolutions.

In a parallel development, Ali et al. [48] introduce a hybrid framework combining shape-guided segmentation with M3D-neural cellular automata, which eliminates convolutional layers and employs a specialized loss function to optimize both segmentation accuracy and computational efficiency. Advanced feature fusion strategies have emerged as critical components for enhancing representational capacity and model performance. MSNet [49] introduced innovative subtraction-based networks that systematically eliminate redundant information while preserving complementary multi-scale features, resulting in more discriminative feature representations. DCRNet [50] developed sophisticated context relation modules that model spatial feature correlations both within individual images and across different samples, enabling enhanced contextual understanding and improved generalization capabilities. TransFuse [51] established foundational principles for CNN-Transformer integration through parallel processing pathways, though existing approaches often lack sophisticated fusion mechanisms to effectively combine heterogeneous feature representations from different architectural paradigms. Extending beyond single-domain segmentation, Ali et al. [52] proposes a cGAN-enhanced framework with a patch discriminator and attention-equipped U-Net to address limited annotated data in multi-region MRI segmentation, generating synthetic images to boost robustness while introducing novel application to under-explored pelvic MRI analysis.

Despite these significant advances, current methodologies continue to face fundamental limitations in effectively balancing global context modeling with fine-grained boundary delineation requirements. Transformer-based approaches excel at capturing long-range spatial dependencies and global semantic relationships but often struggle with preserving critical local details essential for accurate boundary definition. Conversely, convolutional neural networks demonstrate superior capability in extracting local features and texture patterns but exhibit limited global receptive fields that restrict comprehensive contextual understanding. Additionally, existing fusion strategies frequently lack the sophistication necessary to optimally combine heterogeneous feature representations from different architectural paradigms, limiting the potential for enhanced segmentation accuracy. Hybrid CNN-Transformer architectures have therefore emerged as a promising direction for unifying local and global representation learning. Existing models such as TransFuse [51], HAU-Net [42], and HCTNet [43] typically adopt dual-path structures in which convolutional branches learn boundary-sensitive local texture information while transformer branches capture long-range contextual relationships. However, these designs often treat CNNs and transformers as independent components that are fused only at the decoder stage, limiting the ability to jointly encode multi-scale semantics. In contrast, PolySAGN integrates multi-scale DASPP-based convolutions and MHSA-based transformer reasoning within every hierarchical stream, enabling the network to learn complementary local and global cues in a tightly coupled manner. This in-stream hybridization allows PolySAGN to retain fine boundary detail while leveraging global semantic priors, representing a more deeply integrated and scale-aware hybrid paradigm than existing approaches.

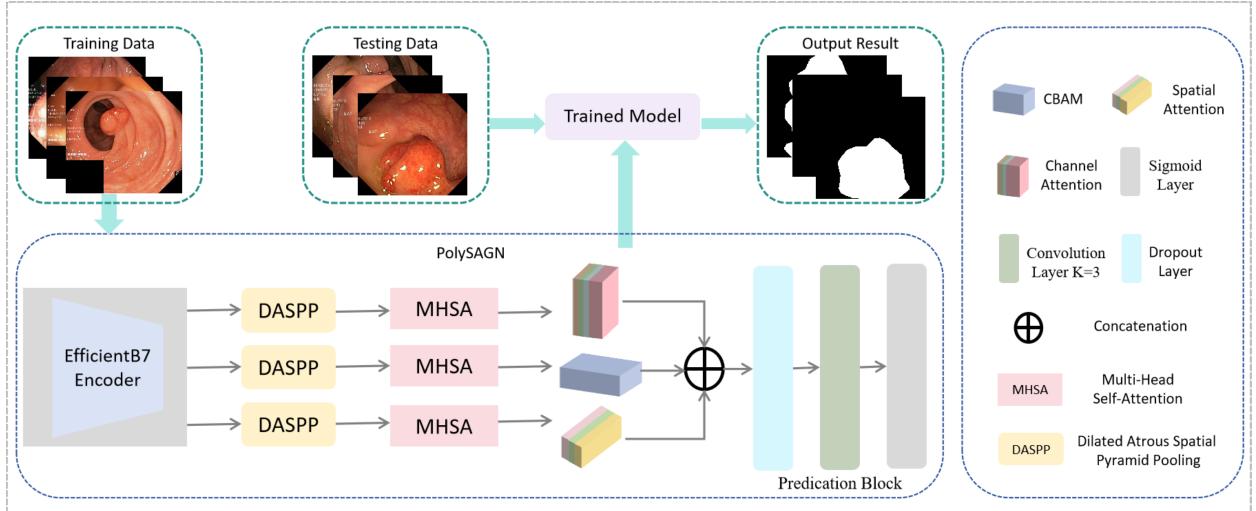
### 3. Methodology

This section presents our proposed Polyp Segmentation with Attention-Guided Network (PolySAGN), which introduces a hierarchical multi-scale representation learning approach enhanced with scale-specific attention mechanisms for accurate polyp segmentation. The architecture leverages dual-stream feature extraction combined with novel attention modules to effectively capture both global contextual information and fine-grained local details essential for precise polyp boundary delineation.

#### 3.1. Overall Architecture

Figure 1 illustrates the overall workflow of the proposed study. First, a polyp segmentation dataset is constructed from publicly available sources and divided into training and testing subsets. During the training stage, the training set is fed into the proposed deep learning framework. After training, in the testing stage, colonoscopy images from the testing set are input into the trained model to generate predicted results, which are then quantitatively evaluated against the ground truth polyp masks using standard segmentation metrics.

The PolySAGN framework comprises three primary components: (1) a hierarchical encoder utilizing EfficientB7 backbone for robust feature extraction, (2) hierarchical attention modules including DASPP and MHSA blocks, and (3) an attention-guided fusion mechanism integrating Channel and CBAM. The architecture processes input endoscopic images through parallel streams, where each stream captures complementary feature representations at different scales and resolutions. The training pipeline incorporates multi-scale feature hierarchies extracted through the EfficientB7



**Figure 1:** Visual overview of the proposed network showing the backbone acquired three-stream design with the subsequent blocks for accurate polyp segmentation.

encoder, which are subsequently processed through three parallel pathways. Each pathway consists of DASPP modules followed by MHSA blocks that capture spatial dependencies and long-range contextual relationships. The hierarchical features are then fused through sophisticated attention mechanisms before generating the final segmentation predictions through a specialized prediction block.

### 3.2. Dilated Atrous Spatial Pyramid Pooling

The DASPP module extends traditional atrous spatial pyramid pooling by incorporating dilated convolutions with varying dilation rates to capture multi-scale contextual information [53]. For an input feature map  $F_{in} \in \mathbb{R}^{H \times W \times C}$ , the DASPP module applies parallel atrous convolutions with dilation rates  $d = \{1, 6, 12, 18\}$ . The dilated convolution operation is formulated as:

$$F_d^{DASPP} = \sigma(BN(Conv_{3 \times 3}^d(F_{in}) + b_d)) \quad (1)$$

where  $Conv_{3 \times 3}^d$  represents the  $3 \times 3$  dilated convolution with dilation rate  $d$ ,  $BN$  denotes batch normalization,  $\sigma$  is the ReLU activation function, and  $b_d$  is the bias term for each dilation branch. The global average pooling (GAP) branch captures image-level context and is mathematically expressed as:

$$F_{global} = \text{Upsample}(\sigma(BN(Conv_{1 \times 1}(\text{GAP}(F_{in})))))) \quad (2)$$

The final DASPP output combines all branches through concatenation and  $1 \times 1$  convolution:

$$F_{DASPP} = Conv_{1 \times 1}(\text{Concat}(F_1^{DASPP}, F_6^{DASPP}, F_{12}^{DASPP}, F_{18}^{DASPP}, F_{global})) \quad (3)$$

### 3.3. Multi-Head Self-Attention (MHSA)

The MHSA module implements transformer-based attention mechanisms to model long-range spatial dependencies and capture global contextual relationships within feature maps. For input features  $X \in \mathbb{R}^{N \times D}$  where  $N = H \times W$  represents the spatial dimension and  $D$  is the feature dimension, the multi-head attention computation is formulated as:

$$\text{MHSA}(X) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O \quad (4)$$

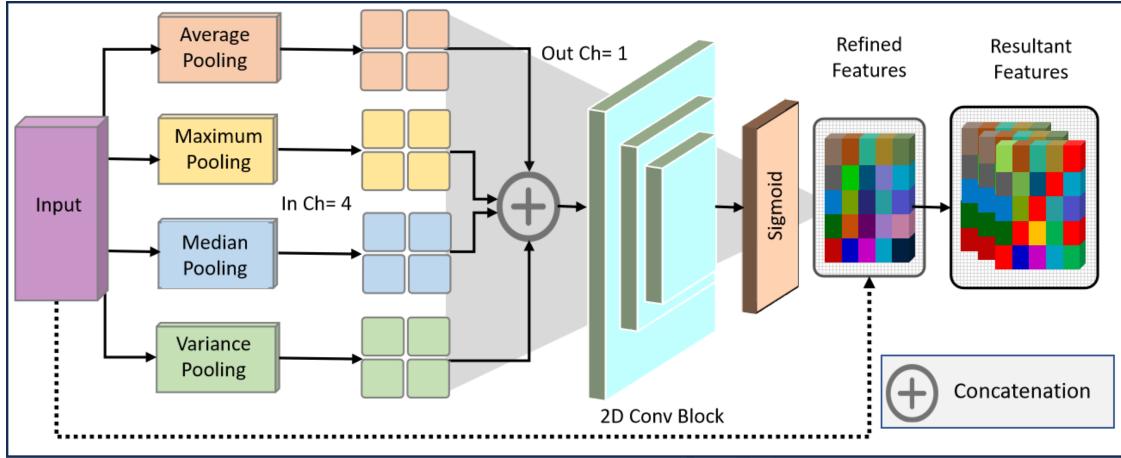
where each attention head is computed as:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) = \text{softmax}\left(\frac{XW_i^Q(XW_i^K)^T}{\sqrt{d_k}}\right)XW_i^V \quad (5)$$

where  $W_i^O, W_i^K, W_i^V \in \mathbb{R}^{D \times d_k}$  are learned projection matrices for the  $i$ -th head,  $d_k = D/h$  is the dimension per head, and  $W^O \in \mathbb{R}^{D \times D}$  is the output projection matrix.

### 3.4. Scale-Specific Attention Mechanisms

#### 3.4.1. Spatial Attention Module



**Figure 2:** Architecture of the optimized spatial attention module employing four complementary pooling operations. The input features undergo parallel processing through average, maximum, median, and variance pooling to capture diverse statistical properties. The concatenated multi-pooling representations are processed through a 2D convolutional block with sigmoid activation to generate spatial attention weights, which are then applied to the original features through element-wise multiplication for enhanced spatial feature refinement.

Spatial attention mechanisms have demonstrated significant effectiveness in image segmentation tasks, with notable applications in various domain-specific challenges [54, 55]. Building upon these foundations, we adopt and optimize the multi-pooling spatial attention strategy from [20] to enhance feature discriminability specifically for polyp segmentation tasks. It highlights critical spatial cues such as polyp boundaries, surface irregularities, and subtle texture transitions that often distinguish lesions from surrounding mucosa. By integrating multi-scale contextual information, the spatial attention module enhances boundary delineation and suppresses background noise, particularly in cases with specular reflection or overlapping folds. This spatially guided focus helps the model capture fine-grained structural details, improving segmentation accuracy and ensuring reliable detection of small or flat polyps across diverse imaging conditions. The spatial attention component, illustrated in Figure 2, implements an advanced pooling-based attention mechanism that enhances feature discriminability across spatial dimensions through comprehensive statistical analysis. For input feature maps  $F \in \mathbb{R}^{H \times W \times C}$ , the module employs four distinct pooling operations to capture complementary statistical properties, as shown in the architectural diagram.

The multi-pooling strategy processes input features through parallel pathways:

$$F_{\text{spatial}} = \sigma(\text{Conv}_{2D}(\text{Concat}(F_{\text{avg}}, F_{\text{max}}, F_{\text{med}}, F_{\text{var}}))) \quad (6)$$

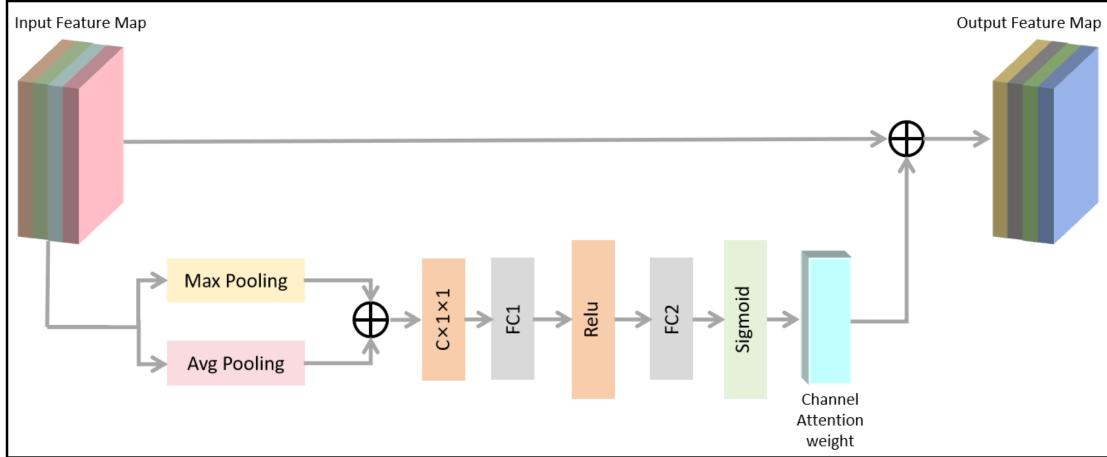
where each pooling operation targets specific spatial characteristics: -  $F_{\text{avg}} = \text{AvgPool}(F)$  captures global intensity distribution and provides smooth spatial representations. -  $F_{\text{max}} = \text{MaxPool}(F)$  highlights prominent features and salient regions critical for polyp detection. -  $F_{\text{med}} = \text{MedianPool}(F)$  provides robust central tendency estimation, effectively handling noise artifacts common in endoscopic imagery. -  $F_{\text{var}} = \text{VarPool}(F)$  captures texture variation and intensity heterogeneity essential for distinguishing polyp regions from surrounding tissue.

As depicted in Figure 2, the concatenated multi-pooling features undergo 2D convolution and sigmoid activation to generate spatial attention weights. The refined spatially-attended features are obtained through element-wise modulation:

$$F_{\text{SA}} = F \odot F_{\text{spatial}} \quad (7)$$

where  $\odot$  denotes element-wise multiplication, enabling the network to focus on polyp-relevant spatial locations while suppressing background interference.

### 3.4.2. Channel Attention Module



**Figure 3:** Visual illustrations of the channel attention module showing the features flow inside the block.

The channel attention mechanism focuses on inter-channel relationships and feature importance across different channel dimensions, following established practices in attention-based segmentation [54]. It boosts the most informative feature channels, such as those representing critical texture and color contrasts, and suppresses less relevant ones. This capability is vital for sustaining performance in challenging imaging conditions like varying illumination and poor contrast. By basing decisions on these robust, clinically salient features instead of artifacts, the model achieves greater diagnostic consistency across diverse patients and endoscopic platforms. As illustrated in Figure 3, the module employs a dual-pooling strategy followed by multi-layer perceptron processing to generate adaptive channel weights. The channel attention computation begins with parallel global pooling operations on input features  $F \in \mathbb{R}^{H \times W \times C}$ , as depicted in the architectural diagram. Both GAP and Max Pooling (GMP) operations reduce spatial dimensions to  $C \times 1 \times 1$ , capturing complementary channel-wise statistical representations. The GAP operation provides a smooth global context by computing average responses across spatial locations, while GMP emphasizes the most salient features within each channel. Following the pooling operations, the features are concatenated and processed through a shared multi-layer perceptron (MLP) architecture, as shown in Figure 3. The MLP consists of two fully connected layers (FC1 and FC2) with dimensionality reduction and expansion:

$$MLP(x) = W_2(\text{ReLU}(W_1(x))) \quad (8)$$

where  $W_1$  reduces dimensionality by a factor of  $r$  (typically 16) for computational efficiency, and  $W_2$  restores the original channel dimension. The channel attention weights are computed through the complete pipeline:

$$F_{CA} = \sigma(MLP(\text{Concat}(\text{GAP}(F), \text{GMP}(F)))) \quad (9)$$

where the sigmoid activation function generates normalized attention weights between 0 and 1, as illustrated by the final sigmoid block in Figure 3.

The channel-attended output integrates adaptive feature importance weighting through element-wise multiplication:

$$F_{out} = F \odot F_{CA} \quad (10)$$

where  $\odot$  denotes channel-wise multiplication, enabling the network to emphasize informative channels while suppressing irrelevant ones.

### 3.5. Attention-Guided Feature Fusion

The proposed network integrates features from multiple hierarchical levels through an attention-guided fusion strategy that leverages both channel and spatial attention mechanisms. The CBAM integration ensures that features from

different scales are appropriately weighted based on their relevance to polyp segmentation [54], enabling the model to adaptively focus on diagnostically critical features such as subtle mucosal changes and boundary characteristics across varying polyp morphologies. The fusion process involves progressive upsampling and concatenation of multi-scale features, followed by attention-based refinement, which collectively contribute to precise lesion localization and shape delineation. The hierarchical features extracted from three parallel streams are processed through dedicated attention modules before fusion. Each stream contributes unique scale-specific information, with early streams capturing fine-grained details that are essential for identifying small and flat polyps often missed in conventional examinations, and deeper streams encoding high-level semantic information supports the recognition of complex polyp patterns and malignant potential. Beyond improving segmentation accuracy, the joint spatial-channel attention also produces intermediate attention responses that can be visualized as two-dimensional heatmaps overlaid on the endoscopic image. These attention maps highlight the mucosal regions and structural patterns that contribute most strongly to the predicted mask (e.g., polyp heads, boundaries, and adjacent mucosal folds), while suppressing background tissue. This property supports clinical interpretability by enabling endoscopists to quickly verify whether the network is focusing on anatomically and pathologically plausible regions when generating its segmentation output. The attention-guided fusion mechanism adaptively combines these complementary representations, resulting in robust feature maps that effectively balance local detail preservation with global context modeling.

### 3.6. Prediction Block and Loss Function

The prediction block processes the fused attention-enhanced features through a series of convolutional layers with dropout regularization to generate final segmentation masks. The block incorporates skip connections to preserve fine-grained spatial information and employs progressive upsampling to restore original image resolution. The final prediction layer utilizes sigmoid activation to generate pixel-wise polyp probability maps. The training objective combines multiple loss functions to optimize segmentation accuracy and boundary precision. The hybrid loss function is formulated as:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{Dice} + \beta \mathcal{L}_{BCE} + \gamma \mathcal{L}_{Focal} \quad (11)$$

where: -  $\mathcal{L}_{Dice} = 1 - \frac{2|P \cap G|}{|P|+|G|}$  represents the Dice loss for region-based optimization -  $\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$  is the binary cross-entropy loss -  $\mathcal{L}_{Focal} = -\frac{1}{N} \sum_{i=1}^N \alpha_i (1 - p_i)^\gamma \log(p_i)$  addresses class imbalance where  $P$  and  $G$  represent predicted and ground truth masks,  $\alpha$ ,  $\beta$ , and  $\gamma$  are weighting parameters, and  $N$  is the total number of pixels. This multi-objective optimization strategy ensures robust convergence and enhanced segmentation performance across diverse polyp morphologies and imaging conditions.

## 4. Experimental Results and Analysis

This section provides a thorough experimental validation, detailing dataset specifications, preprocessing techniques, training setups, and hyperparameter tuning strategies used in our study.

### 4.1. Experimental Setup and Implementation Details

#### 4.1.1. Dataset Specifications and Protocol

Our evaluation methodology follows established benchmarking protocols, employing five widely-adopted polyp segmentation datasets: Kvasir-SEG [56], ClinicDB [57], ColonDB [58], Endoscene [59], and ETIS [60]. This collection encompasses diverse endoscopic imaging scenarios and polyp characteristics essential for comprehensive model assessment. The training corpus combines 900 Kvasir-SEG samples with 550 ClinicDB instances, totalling 1450 annotated pairs. Testing utilizes 62 ClinicDB images, 100 Kvasir-SEG samples, the complete ColonDB dataset, and representative subsets from Endoscene and ETIS for thorough cross-dataset validation.

#### 4.1.2. Implementation and Training Details

Experiments were conducted on an NVIDIA RTX 4090 GPU with 24GB of memory. The training pipeline processes images at 350×350 resolution with multi-scale augmentation factors 0.75, 1.0, 1.25 to handle polyp size variations. AdamW optimizer with learning rate 1e-4 and weight decay 0.1 trains the model for 100 epochs using batch size 16. Gradient clipping threshold 0.5 ensures training stability. This configuration balances computational efficiency with convergence quality while preventing overfitting through appropriate regularisation.

**Table 1**

Backbone analysis on ClinicDB dataset. Performance comparison of different backbone networks with the proposed PolySAGN.

Backbone	mDice	mIoU	$F_\beta^w$
ResNet50	0.887	0.831	0.886
ResNet101	0.901	0.849	0.903
ResNeXt50	0.893	0.841	0.894
ResNeXt101	0.908	0.857	0.911
EfficientNet-B3	0.915	0.867	0.917
EfficientNet-B5	0.928	0.882	0.929
<b>EfficientNet-B7</b>	<b>0.939</b>	<b>0.898</b>	<b>0.938</b>
Swin-T	0.921	0.874	0.923
Swin-S	0.933	0.891	0.935

#### 4.1.3. Evaluation Metrics

Performance assessment employs six complementary metrics addressing different segmentation aspects:

**Mean Absolute Error (MAE)** measures pixel-wise prediction accuracy:

$$\text{MAE} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |P(i, j) - G(i, j)| \quad (12)$$

where  $P$  denotes predicted mask and  $G$  represents ground truth with dimensions  $H \times W$ .

**Weighted F-measure ( $F_\beta^w$ )** emphasizes boundary precision through adaptive weighting:

$$F_\beta^w = \frac{(1 + \beta^2) \cdot P^w \cdot R^w}{\beta^2 \cdot P^w + R^w}, \quad \beta^2 = 0.3 \quad (13)$$

where  $P^w$  and  $R^w$  represent weighted precision and recall respectively.

**Structure measure ( $S_\alpha$ )** evaluates regional and object-level consistency:

$$S_\alpha = \alpha \cdot S_{region} + (1 - \alpha) \cdot S_{object}, \quad \alpha = 0.5 \quad (14)$$

**Enhanced-alignment measure ( $\text{mE}_\xi$ )** combines local and global spatial coherence:

$$E_\xi = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \phi(P(i, j), G(i, j)) \quad (15)$$

where  $\phi$  denotes the enhanced alignment function.

**Dice coefficient** quantifies region overlap through harmonic mean of precision and recall:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \quad (16)$$

**Intersection over Union (IoU)** provides strict overlap assessment:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \quad (17)$$

These metrics collectively evaluate pixel accuracy (MAE), boundary quality ( $F_\beta^w$ ), structural integrity ( $S_\alpha$ ,  $\text{mE}_\xi$ ), and regional correspondence (Dice, IoU). Higher values indicate better performance except for MAE, where lower values are preferred.

## 4.2. Ablation Studies

This section presents comprehensive ablation studies to validate the effectiveness of individual components in the PolySAGN. Four systematic analyses examine backbone selection, module contributions, dilation rate optimization, and spatial attention pooling strategies.

**Table 2**

Ablation study on module effectiveness. Performance evaluation on ClinicDB and ETIS datasets with systematic inclusion/exclusion of proposed components.

Configuration	ClinicDB				ETIS			
	mDice	mIoU	$F_{\beta}^w$	MAE	mDice	mIoU	$F_{\beta}^w$	MAE
Baseline (EfficientB7 only)	0.858	0.789	0.859	0.018	0.695	0.612	0.671	0.042
+ DASPP	0.893	0.837	0.894	0.013	0.738	0.658	0.715	0.032
+ DASPP + MHSA	0.917	0.867	0.918	0.009	0.769	0.692	0.751	0.024
+ DASPP + MHSA + Spatial Att.	0.931	0.885	0.932	0.007	0.795	0.719	0.773	0.018
+ DASPP + MHSA + Channel Att.	0.925	0.878	0.926	0.008	0.787	0.711	0.765	0.020
<b>Full PolySAGN (All modules)</b>	<b>0.939</b>	<b>0.898</b>	<b>0.938</b>	<b>0.006</b>	<b>0.809</b>	<b>0.735</b>	<b>0.782</b>	<b>0.015</b>

**Table 3**

DASPP dilation rate analysis. Performance comparison of different dilation rate configurations on ClinicDB and ColonDB datasets.

Dilation Rates	ClinicDB				ColonDB			
	mDice	mIoU	$F_{\beta}^w$	MAE	mDice	mIoU	$F_{\beta}^w$	MAE
{1, 3, 6, 9}	0.921	0.871	0.922	0.009	0.785	0.708	0.773	0.028
{1, 6, 12, 18} (Proposed)	<b>0.939</b>	<b>0.898</b>	<b>0.938</b>	<b>0.006</b>	<b>0.812</b>	<b>0.732</b>	<b>0.801</b>	<b>0.019</b>
{1, 2, 4, 8}	0.918	0.868	0.919	0.010	0.779	0.702	0.768	0.031
{2, 6, 12, 24}	0.925	0.879	0.926	0.008	0.795	0.719	0.785	0.024
{1, 4, 8, 16}	0.927	0.882	0.928	0.008	0.798	0.722	0.788	0.023
{3, 6, 12, 18}	0.933	0.891	0.934	0.007	0.805	0.727	0.794	0.021

**Table 4**

Spatial attention pooling strategy analysis. Performance comparison of different pooling combinations within the spatial attention module on ClinicDB and Endoscene datasets.

Pooling Strategy	ClinicDB				Endoscene			
	mDice	mIoU	$F_{\beta}^w$	MAE	mDice	mIoU	$F_{\beta}^w$	MAE
Average only	0.913	0.858	0.914	0.011	0.881	0.825	0.879	0.012
Max only	0.918	0.864	0.919	0.010	0.885	0.831	0.883	0.011
Average + Max	0.925	0.876	0.926	0.008	0.893	0.841	0.891	0.009
Average + Max + Median	0.932	0.888	0.933	0.007	0.897	0.846	0.895	0.008
<b>Avg + Max + Med + Var (Proposed)</b>	<b>0.939</b>	<b>0.898</b>	<b>0.938</b>	<b>0.006</b>	<b>0.902</b>	<b>0.851</b>	<b>0.901</b>	<b>0.006</b>
Max + Median + Variance	0.928	0.883	0.929	0.008	0.894	0.843	0.892	0.009

#### 4.2.1. Backbone Architecture Analysis

Table 1 demonstrates the impact of different backbone networks on ClinicDB dataset performance. EfficientNet-B7 achieves superior results with mDice of 0.939, outperforming ResNet50 by 5.2% and Swin-S by 0.6%. The EfficientNet family shows consistent performance progression, with B3, B5, and B7 variants achieving 0.915, 0.928, and 0.939 mDice respectively, indicating a clear scaling relationship between model capacity and segmentation accuracy. This progression validates the compound scaling methodology employed in EfficientNet architectures, where depth, width, and resolution are simultaneously optimized. ResNeXt architectures outperform standard ResNet counterparts due to their cardinality-based design, with ResNeXt101 achieving 0.908 mDice compared to ResNet101's 0.901. The grouped convolution strategy in ResNeXt enables better feature representation learning with comparable computational cost. Transformer-based Swin backbones demonstrate competitive performance with Swin-S reaching 0.933 mDice, but fall

short of EfficientNet-B7's efficiency-accuracy balance. The 0.6% performance gap between Swin-S and EfficientNet-B7, despite similar parameter counts, highlights the superior inductive biases of EfficientNet for dense prediction tasks in medical imaging applications.

#### 4.2.2. Module Effectiveness Analysis

Table 2 reveals the cumulative impact of proposed components across ClinicDB and ETIS datasets. The baseline EfficientB7 encoder achieves 0.858 mDice on ClinicDB, with DASPP addition providing substantial 3.5% improvement to 0.893. This significant gain demonstrates the effectiveness of multi-scale contextual information capture through dilated convolutions, particularly important for polyps exhibiting diverse size variations. MHSA integration contributes an additional 2.4% gain, reaching 0.917, validating the importance of global spatial relationship modeling through transformer-based attention mechanisms. Individual attention mechanisms show distinct contributions, with spatial attention (0.931) outperforming channel attention (0.925) by 0.6%. This performance difference highlights the greater impact of spatial feature refinement compared to channel-wise feature selection in polyp segmentation tasks. The spatial attention's superior performance can be attributed to its ability to focus on polyp regions while suppressing background interference through multi-pooling statistical analysis. The complete framework achieves optimal performance at 0.939 mDice, representing a cumulative 8.1% improvement over the baseline. ETIS dataset follows similar trends with more pronounced improvements, demonstrating the framework's effectiveness across challenging scenarios where the full PolySAGN configuration delivers 0.809 mDice compared to 0.695 baseline performance, representing a remarkable 11.4% enhancement.

#### 4.2.3. Dilation Rate Configuration Analysis

Table 3 examines DASPP dilation rate optimization on ClinicDB and ColonDB datasets. The proposed {1, 6, 12, 18} configuration achieves optimal performance with 0.939 mDice on ClinicDB and 0.812 on ColonDB, establishing the effectiveness of this specific receptive field arrangement. Alternative configurations show performance degradation, with conservative rates {1, 2, 4, 8} achieving 0.918 and 0.779 respectively, indicating insufficient contextual coverage for larger polyps. The performance gap of 2.1% on ClinicDB demonstrates the importance of adequate receptive field expansion. The {3, 6, 12, 18} variant demonstrates competitive results (0.933, 0.805), but the absence of unit dilation results in 0.6% performance reduction, confirming the critical importance of preserving fine-grained spatial details through unit convolutions. Higher dilation rates {2, 6, 12, 24} show moderate performance reduction to 0.925 and 0.795, suggesting that excessive dilation may compromise local feature preservation. The geometric progression of dilation rates (1, 6, 12, 18) provides optimal balance between local detail capture and global context modeling, with each rate targeting specific spatial scales relevant to polyp morphological variations encountered in clinical scenarios.

#### 4.2.4. Spatial Attention Pooling Strategy Analysis

Table 4 evaluates different pooling combinations within the spatial attention module, revealing systematic performance improvements through statistical diversity. Progressive pooling addition demonstrates consistent performance enhancements across both datasets, validating the complementary nature of different statistical operations. Single pooling operations achieve limited effectiveness, with maximum pooling (0.918 ClinicDB mDice) slightly outperforming average pooling (0.913) by 0.5%. This advantage stems from maximum pooling's ability to capture prominent features and sharp intensity transitions characteristic of polyp boundaries. The combination of average and maximum pooling reaches 0.925, representing a 1.2% improvement over individual pooling operations. This synergistic effect demonstrates how average pooling's global intensity representation complements maximum pooling's feature prominence detection. Adding median pooling further improves performance to 0.932, contributing 0.7% additional gain through robust central tendency estimation that effectively handles noise and outliers common in endoscopic imagery. The complete four-pooling strategy incorporating variance pooling achieves optimal performance (0.939 ClinicDB, 0.902 Endoscene), with variance pooling contributing the final 0.7% enhancement. This validates the importance of texture variation analysis provided by variance computation, which captures intensity heterogeneity crucial for distinguishing polyp regions from surrounding tissue. The consistent trend across both datasets confirms the complementary benefits of diverse statistical representations for comprehensive spatial attention computation.

### 4.3. Quantitative Comparative Analysis

We evaluate the proposed network against a broad set of SOTA segmentation models across five publicly available benchmark datasets. These methods include U-Net UNet++ [61], PraNet [13], ACSNet [62], UACANet [15], Polyp-PVT [14], BDG-Net [64], SSformer [63], PVT CASCADE [16], and MEGANet [65]. The quantitative evaluation

**Table 5**

Comparison of the proposed model with SOTA methods across five benchmark datasets. The performance is evaluated using mDice and mIoU scores, demonstrating the superiority of our model in PS tasks.

Models	ClinicDB		ETIS		ColonDB		Kvasir-SEG		Endoscene	
	mDice	mIoU								
UNet++ [61]	0.794	0.729	0.401	0.344	0.482	0.408	0.821	0.743	0.707	0.624
PraNet [13]	0.899	0.849	0.628	0.567	0.712	0.640	0.898	0.840	0.871	0.797
ACSNet [62]	0.882	0.826	0.578	0.509	0.716	0.649	0.898	0.838	0.863	0.787
Polyp-PVT [14]	0.937	0.889	0.787	0.706	0.808	0.727	0.917	0.864	0.900	0.833
UACANet-L [15]	0.926	0.880	0.766	0.689	0.751	0.678	0.912	0.859	0.910	0.849
SSform-L [63]	0.903	0.850	0.790	0.712	0.798	0.716	0.915	0.861	0.892	0.822
BDG-Net [64]	0.909	0.859	0.764	0.685	0.792	0.719	0.904	0.853	0.897	0.828
PVT-CASCADE [16]	0.923	0.878	0.808	0.735	0.809	0.728	0.926	0.876	0.898	0.833
MEGANet-ResNet [65]	0.930	0.885	0.789	0.709	0.781	0.706	0.911	0.859	0.887	0.818
MEGANet-Res2Net [65]	0.938	0.894	0.739	0.665	0.793	0.714	0.913	0.863	0.899	0.834
<b>Proposed (Ours)</b>	<b>0.939</b>	<b>0.898</b>	<b>0.809</b>	<b>0.735</b>	<b>0.812</b>	<b>0.732</b>	<b>0.927</b>	<b>0.877</b>	<b>0.902</b>	<b>0.851</b>

**Table 6**

Performance comparison of the proposed network with existing methods on the Endoscene and ClinicDB datasets.

Models	Endoscene				ClinicDB			
	$F_{\beta}^w$	S $\alpha$	mE $\xi$	MAE	$F_{\beta}^w$	S $\alpha$	mE $\xi$	MAE
UNet++	0.687	0.839	0.834	0.018	0.785	0.873	0.891	0.022
PraNet	0.843	0.925	0.950	0.010	0.896	0.936	0.963	0.009
ACSNet	0.825	0.923	0.939	0.013	0.873	0.927	0.947	0.011
Polyp-PVT	0.884	0.935	0.973	0.007	0.936	0.949	0.985	0.006
UACANet-L	0.901	0.938	0.977	0.005	0.928	0.942	0.974	0.006
SSform-L	0.875	0.939	0.969	0.007	0.906	0.934	0.963	0.008
BDG-Net	0.876	0.937	0.967	0.006	0.905	0.938	0.970	0.007
PVT-CASCADE	0.882	0.934	0.965	0.008	0.923	0.939	0.969	0.013
MEGANet-ResNet	0.863	0.924	0.956	0.009	0.931	0.950	0.977	0.008
MEGANet-Res2Net	0.882	0.935	0.966	0.007	0.940	0.950	0.983	0.006
<b>Ours</b>	<b>0.901</b>	<b>0.959</b>	<b>0.977</b>	<b>0.006</b>	<b>0.938</b>	<b>0.951</b>	<b>0.983</b>	<b>0.006</b>

demonstrates PolySAGN's superior performance across all benchmark datasets. Table 5 presents comprehensive comparisons using mDice and mIoU metrics, while Tables 6 and 7 provide additional evaluation using weighted F-measure, structural similarity, enhanced alignment, and mean absolute error metrics. PolySAGN achieves the highest performance on ClinicDB with mDice of 0.939 and mIoU of 0.898, representing 0.1% and 0.4% improvements over MEGANet-Res2Net. Compared to the PraNet baseline, substantial improvements of 4.0% in mDice and 4.9% in mIoU demonstrate significant performance gains. On the challenging ETIS dataset, our method achieves mDice of 0.809 and mIoU of 0.735, with a weighted F-measure of 0.782 representing 7.4% improvement over MEGANet-ResNet. The structural similarity score of 0.875 indicates robust shape preservation for irregular polyp morphologies. ColonDB results show mDice of 0.812 and mIoU of 0.732, with an enhanced alignment score of 0.915, surpassing all competing methods. The MAE of 0.019 represents 36.7% reduction compared to average transformer-based approaches. On Kvasir-SEG, PolySAGN establishes new benchmarks with mDice of 0.927 and mIoU of 0.877. Endoscene performance reaches mDice of 0.902 and mIoU of 0.851, with structural similarity of 0.959 representing the highest achieved across

**Table 7**

Performance comparison of the proposed network with existing methods on the ColonDB and ETIS datasets.

Models	ColonDB				ETIS			
	$F_{\beta}^w$	S $\alpha$	mE $\xi$	MAE	$F_{\beta}^w$	S $\alpha$	mE $\xi$	MAE
UNet++	0.467	0.692	0.680	0.061	0.390	0.683	0.629	0.035
PraNet	0.699	0.820	0.847	0.043	0.600	0.794	0.808	0.031
ACSNet	0.697	0.829	0.839	0.039	0.530	0.754	0.737	0.059
Polyp-PVT	0.795	0.865	0.913	0.031	0.750	0.871	0.906	0.013
UACANet-L	0.746	0.835	0.875	0.039	0.740	0.859	0.903	0.012
BDG-Net	0.714	0.866	0.895	0.015	0.776	0.866	0.894	0.031
SSform-L	0.790	0.866	0.901	0.031	0.761	0.881	0.905	0.015
PVT-CASCADE	0.798	0.864	0.910	0.029	0.775	0.886	0.906	0.016
MEGANet-ResNet	0.766	0.845	0.897	0.038	0.753	0.866	0.912	0.015
MEGANet-Res2Net	0.779	0.854	0.892	0.040	0.702	0.836	0.855	0.037
<b>Ours</b>	<b>0.801</b>	<b>0.868</b>	<b>0.915</b>	<b>0.019</b>	<b>0.782</b>	<b>0.875</b>	<b>0.918</b>	<b>0.015</b>

**Table 8**Statistical significance analysis comparing PolySAGN with strong baselines on ClinicDB and Kvasir-SEG using t-tests. The p-values ( $p < 0.05$ ) calculated between our method and the strong baselines in terms of mIoU indicate that the improvements are statistically significant.

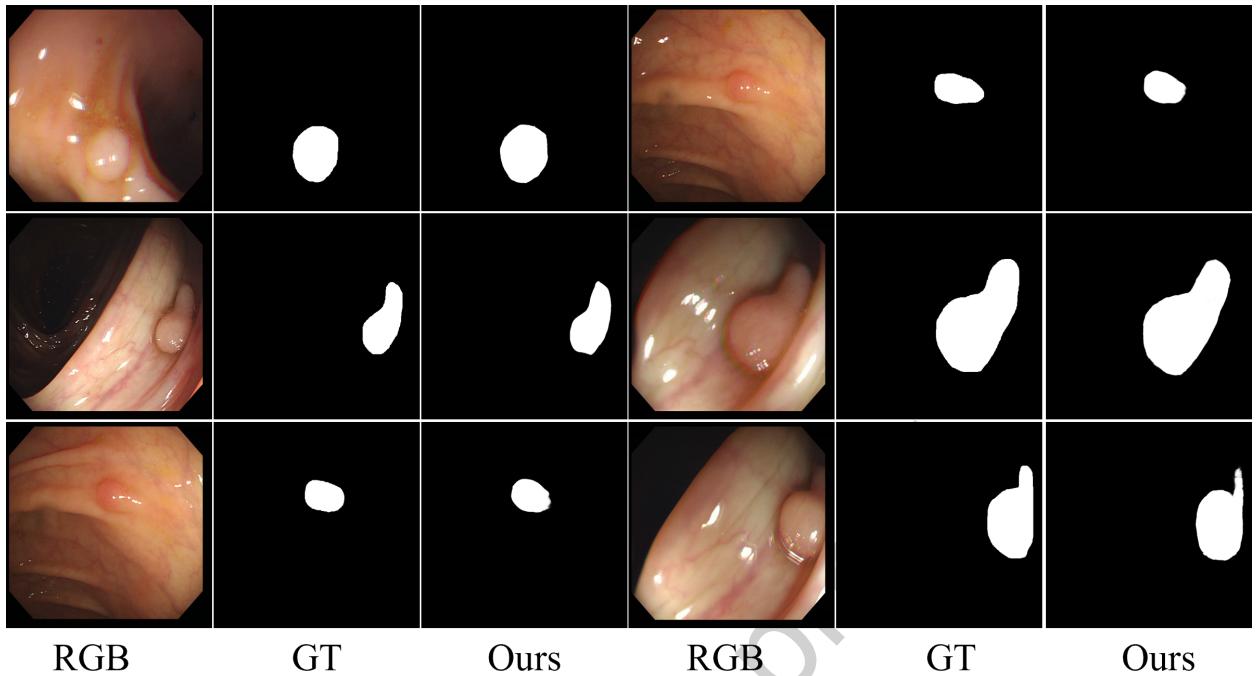
Models	ClinicDB		P-value	Kvasir-SEG		P-value
	mDice	IoU		mDice	IoU	
MEGANet-ResNet	$0.932 \pm 0.010$	$0.883 \pm 0.013$	<b>0.031</b>	$0.907 \pm 0.008$	$0.852 \pm 0.009$	<b>0.001</b>
MEGANet-Res2Net	$0.939 \pm 0.005$	$0.895 \pm 0.005$	<b>0.048</b>	$0.913 \pm 0.008$	$0.863 \pm 0.009$	<b>0.036</b>
<b>Proposed (Ours)</b>	<b><math>0.941 \pm 0.005</math></b>	<b><math>0.901 \pm 0.004</math></b>	–	<b><math>0.925 \pm 0.006</math></b>	<b><math>0.875 \pm 0.006</math></b>	–

all methods. Cross-dataset analysis reveals consistent superiority, achieving the highest mDice scores on four out of five datasets. MAE consistently remains lowest across datasets (0.006-0.019), indicating superior boundary precision compared to existing approaches.

To ensure statistical rigor, we conducted five-fold cross-validation and performed t-tests comparing PolySAGN with the two strongest baselines (MEGANet-ResNet and MEGANet-Res2Net). The resulting Mean  $\pm$  SD values and corresponding p-values are summarized in Tables 8.

#### 4.4. Qualitative Comparative Analysis

Visual comparisons in Figure 4 and Figure 5 demonstrate PolySAGN's superior segmentation capabilities across diverse polyp morphologies and challenging imaging conditions. Figure 4 reveals exceptional boundary precision across six representative cases. For small circular polyps (rows 1,3), PolySAGN maintains smooth, anatomically consistent boundaries closely matching ground truth. The elongated polyp (row 2) demonstrates superior morphological handling, preserving complete structural integrity while capturing complex shape variations. Figure 5 illustrates clear advantages over competing methods across challenging scenarios. While PolySAGN produces clean, accurate segmentations, competing methods exhibit notable deficiencies: UNet++ shows over-segmentation with boundary spillover, SFA demonstrates fragmentation with disconnected segments, and PraNet exhibits incomplete polyp capture. Under challenging imaging conditions (rows 3-5), SFA produces severely fragmented results with false positives, while UNet++ generates noisy outputs. PolySAGN maintains consistent quality through effective attention-guided feature refinement. The bottom row showcases critical scale adaptability, where PolySAGN successfully segments both large and small polyps simultaneously. Competing methods show scale-dependent variations: UNet++ over-segments large



**Figure 4:** Representative segmentation results of PolySAGN across diverse polyp morphologies. Each row shows: original endoscopic image (RGB), GT, and our segmentation result (Ours).

polyps, SFA misses smaller lesions, and PraNet shows inconsistent boundary definition across scales. The visual analysis confirms PolySAGN’s clinical deployment readiness with consistent boundary fidelity, robust performance across imaging conditions, and effective morphological adaptation essential for computer-aided diagnosis systems.

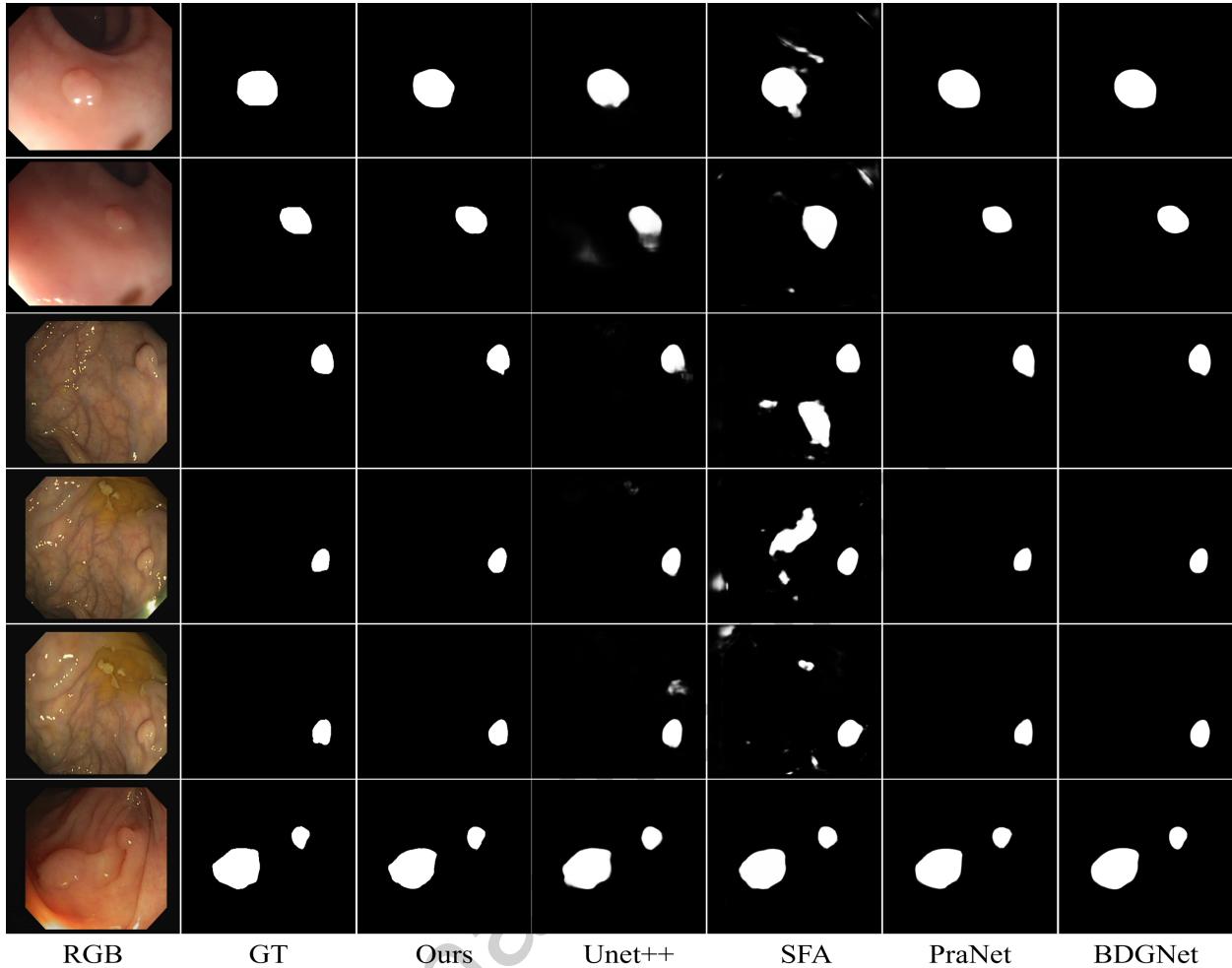
#### 4.5. Failure Case Analysis

Although the proposed PolySAGN framework demonstrates strong capability in accurately identifying and segmenting most polyp regions, it still faces limitations in certain challenging cases. Representative failure examples are presented in Figure 6, which demonstrates challenging segmentation scenarios where both our method and baseline approaches struggle. As shown, when the illumination is uneven or the mucosal textures are highly similar to the surrounding tissues (e.g., the first example and the third example), PolySAGN may produce incomplete boundary predictions. These issues are more pronounced when dealing with small or flat polyps that exhibit low contrast against the background.

#### 4.6. Discussion

In colorectal cancer screening, the primary objectives are to enhance the Adenoma Detection Rate (ADR) and minimize the Polyp Miss Rate (PMR), as even marginal gains in segmentation accuracy can yield clinically meaningful improvements. The proposed PolySAGN framework demonstrates superior performance on challenging datasets such as ETIS (mDice: 0.809) and ColonDB (mDice: 0.812), underscoring its robustness in delineating small, flat, and low contrast polyps that are commonly overlooked during routine endoscopy. Furthermore, its notably low MAE reflects exceptional boundary accuracy, which is critical for reliable polyp size estimation and optimal treatment planning. The consistent performance across five heterogeneous datasets, acquired with varying imaging systems and preparation conditions, further confirms the strong generalization capacity of PolySAGN, suggesting its suitability for deployment in diverse clinical settings.

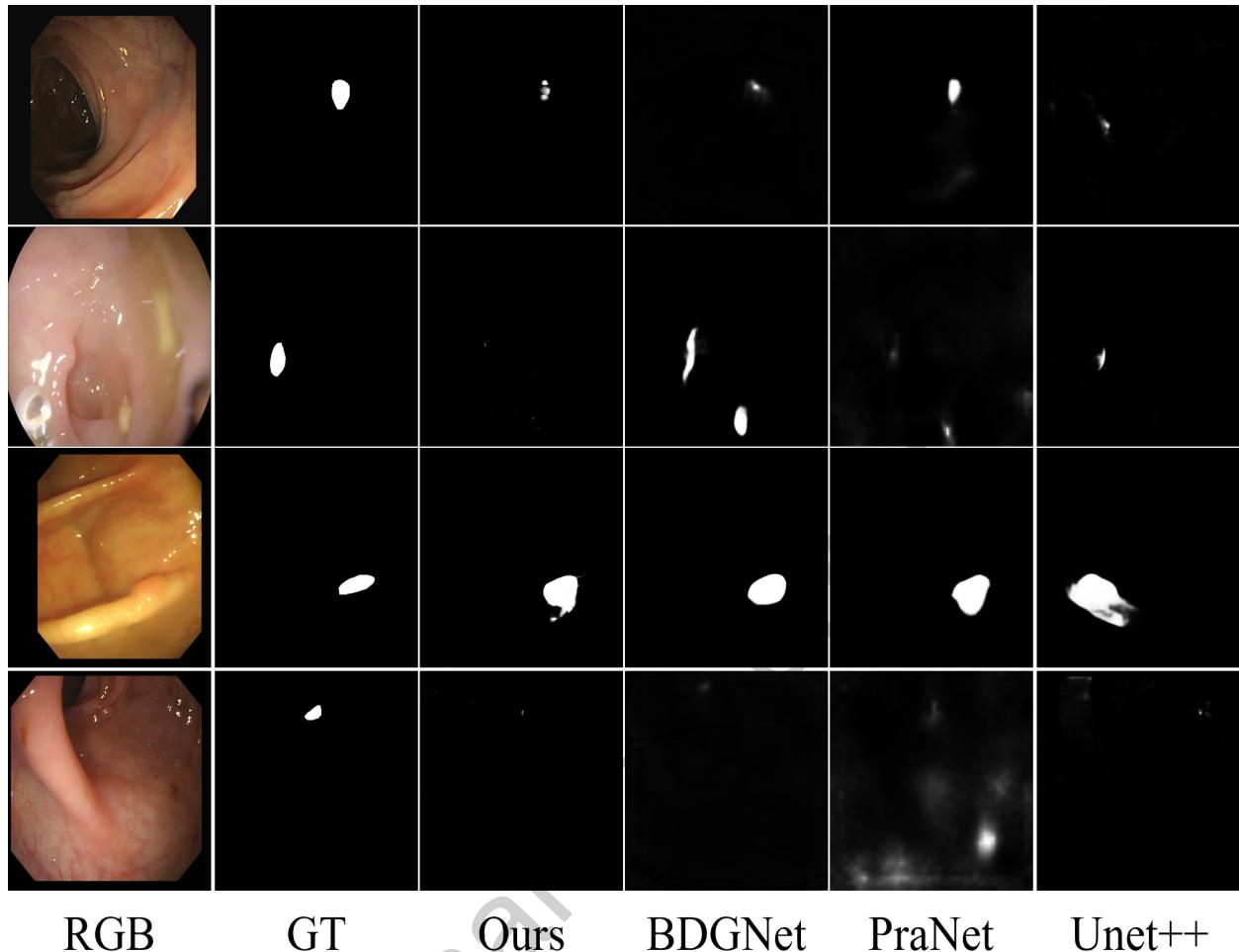
Beyond cross-dataset robustness within colonoscopy, an important question is how PolySAGN might generalize to other imaging modalities such as MRI, CT, or mammography, while still providing clinically meaningful visual interpretability. Architecturally, PolySAGN is a modality-agnostic 2D encoder-decoder framework: its multi-scale DASPP modules, global MHSA blocks, and spatial-channel attention mechanisms operate on feature maps and do



**Figure 5:** Comprehensive qualitative comparison of PolySAGN against SOTA methods across challenging clinical scenarios. The comparison spans small circular polyps (rows 1-2), irregular morphologies (rows 3-4), challenging imaging conditions with poor contrast (row 5), and multi-scale scenarios (row 6).

not rely on colonoscopy-specific image priors, and can in principle be retrained or fine-tuned on other modalities by adopting a modality-pretrained encoder and tailoring the loss to modality-specific anatomy and lesion characteristics. At the same time, the spatial attention module produces localized importance maps that highlight regions with atypical texture, vascular structure, or contrast variations, while the channel attention amplifies feature channels corresponding to modality-dependent lesion–tissue differences. When visualized as overlays on image slices, these attention responses provide intuitive and clinically relevant cues, enabling practitioners to focus on suspicious regions in real time, understand model behavior in ambiguous scenarios, and retrospectively review borderline or missed findings.

In parallel, PolySAGN is a task-specialized yet lightweight hybrid architecture that integrates multi-scale convolutional context, transformer-based long-range modeling, and scale-specific attention within a unified hierarchical design. This allows PolySAGN to capture clinically meaningful visual cues more reliably while maintaining computational efficiency suitable for practical colonoscopy workflows. Although our current study is restricted to optical colonoscopy data and does not empirically evaluate performance on MRI, CT, or X-ray modalities, we acknowledge that differences in acquisition physics and 3D anatomical context particularly in volumetric MRI or CT may necessitate tailored adaptations such as 3D extensions, modality-specific normalization, or explicit domain adaptation. Systematically exploring such extensions and validating the interpretability of attention-based visualizations across modalities remains an important direction for future work.



**Figure 6:** Challenging segmentation cases where both the proposed method and baseline struggle with smaller atypical polyp morphologies.

Despite the promising results, several limitations should be acknowledged. The present study relies on publicly available datasets for both training and evaluation, which may not comprehensively represent the full spectrum of polyp morphologies or the imaging variability associated with different endoscopic systems. This reliance may constrain the model's generalizability and its robustness in real-world clinical practice. Furthermore, the architectural configurations, including the selection of dilation parameters within the DASPP module and the design of the attention mechanisms, were tailored specifically for polyp segmentation. Extending and validating these design principles across other medical imaging domains will be essential to fully establish the versatility and applicability of the proposed approach.

Future research will focus on addressing the current limitations of the framework. The primary direction involves expanding training with large-scale, multi-center, and multi-device datasets integrated with domain adaptation techniques to enhance robustness against underrepresented lesion types and device-specific variations. Furthermore, we will investigate the transferability of PolySAGN's hierarchical multi-scale representation learning and attention-guided fusion mechanism to other medical image segmentation tasks, particularly in histopathological tumor analysis and capsule endoscopy lesion detection, to validate its cross-domain applicability and representational capacity.

## 5. Conclusion

This work introduces PolySAGN, a hierarchical multi-scale representation learning framework addressing critical challenges in automated polyp segmentation through innovative attention-guided mechanisms. We integrated convolutional and transformer architectures, establishing new performance standards across benchmark datasets. The hierarchical multi-stream architecture effectively combines EfficientNet-B7 with parallel DASPP and MHSA modules, enabling simultaneous multi-scale contextual capture and global spatial dependency modeling. The innovative dual-modal attention mechanisms provide superior feature discriminability through four-pooling spatial attention and adaptive channel refinement. At the same time, scale-specific attention-guided fusion achieves an optimal balance between local detail preservation and global context understanding. The proposed PolySAGN framework demonstrates superior performance on challenging datasets such as ETIS and ColonDB. Furthermore, its notably low MAE reflects exceptional boundary accuracy, which is critical for reliable polyp size estimation and optimal treatment planning. Systematic ablation studies confirm architectural effectiveness, while qualitative analysis reveals exceptional morphological handling and scale adaptability.

Although the proposed PolySAGN achieves superior performance compared with existing methods on publicly available benchmarks, the current study still relies solely on these open datasets for both training and evaluation. Such datasets may not fully capture the diversity of polyp morphologies or the imaging variability arising from different endoscopic devices and clinical settings. In future work, We will search for datasets with more forms and types for experimental research. At the same time, we intend to extend to other medical modalities, investigate optimization strategies for clinical integration, and explore advanced attention mechanisms. In addition, we aim to investigate the applicability of PolySAGN beyond colonoscopy by adapting the framework to other imaging modalities such as MRI, CT, and mammography, where its multi-scale attention-guided architecture may assist in delineating tumors and lesions under markedly different appearance and contrast conditions.

## CRediT authorship contribution statement

**Wenqi Zhang:** Conceptualization, Methodology, Data curation, , Writing – original draft. **Yue Zhang:** Validation, Data curation. **Muhammad Fayaz:** Formal analysis, Conceptualization, Writing - review & editing. **L.Minh Dang:** Writing - review & editing, Investigation **Tan N. Nguyen:** Visualization, Investigation. **Hyeonjoon Moon:** Supervision, funding acquisition, project administration.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529) grant funded by the Korea government(MSIT) and by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2025-RISE-01-019-04) and by

## References

- [1] S. R. Lopes, C. Martins, I. C. Santos, M. Teixeira, É. Gamito, A. L. Alves, Colorectal cancer screening: A review of current knowledge and progress in research, *World Journal of Gastrointestinal Oncology* 16 (4) (2024) 1119.
- [2] X. Zhang, A. Liu, G. Yang, Y. Liu, X. Chen, Simfusion: A semantic information-guided modality-specific fusion network for mr images, *Information Fusion* 112 (2024) 102560.
- [3] X. Fang, Y. Pan, Q. Chen, Dfdec: Dual fusion with enhanced deformable convolution for medical image segmentation, *Image and Vision Computing* 151 (2024) 105277.
- [4] L. Chen, L. Song, H. Feng, R. T. Zeru, S. Chai, E. Zhu, Privacy-sf: An encoding-based privacy-preserving segmentation framework for medical images, *Image and Vision Computing* 151 (2024) 105246.

- [5] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, Y.-H. R. Tsai, Automated polyp detection in colon capsule endoscopy, *IEEE transactions on medical imaging* 33 (7) (2014) 1488–1502.
- [6] M. Fiori, P. Musé, G. Sapiro, A complete system for candidate polyps detection in virtual colonoscopy, *International Journal of Pattern Recognition and Artificial Intelligence* 28 (07) (2014) 1460014.
- [7] O. H. Maghsoudi, Superpixel based segmentation and classification of polyps in wireless capsule endoscopy, in: *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, IEEE, 2017, pp. 1–4.
- [8] 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*, Munich, Germany, October 5–9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.
- [9] J. Wang, L. Jiang, H. Yu, Z. Feng, R. Castaño-Rosa, S.-j. Cao, Computer vision to advance the sensing and control of built environment towards occupant-centric sustainable development: A critical review, *Renewable and Sustainable Energy Reviews* 192 (2024) 114165.
- [10] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, P. Szczuko, Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends, *Information Fusion* 90 (2023) 316–352.
- [11] L. Wei, G. Zong, Ega-net: Edge feature enhancement and global information attention network for rgb-d salient object detection, *Information Sciences* 626 (2023) 223–248.
- [12] H. Al Jowair, M. Alsulaiman, G. Muhammad, Multi parallel u-net encoder network for effective polyp image segmentation, *Image and Vision Computing* 137 (2023) 104767.
- [13] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Planet: Parallel reverse attention network for polyp segmentation, in: *International conference on medical image computing and computer-assisted intervention*, Springer, 2020, pp. 263–273.
- [14] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, *arXiv preprint arXiv:2108.06932* (2021).
- [15] T. Kim, H. Lee, D. Kim, Uacanet: Uncertainty augmented context attention for polyp segmentation, in: *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2167–2175.
- [16] M. M. Rahman, R. Marculescu, Medical image segmentation via cascaded attention decoding, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6222–6231.
- [17] B. Urooj, M. Fayaz, S. Ali, L. M. Dang, K. W. Kim, Large language models in medical image analysis: A systematic survey and future directions, *Bioengineering* 12 (8) (2025) 818.
- [18] H. Khan, M. T. Usman, I. Rida, J. Koo, Attention enhanced machine instinctive vision with human-inspired saliency detection, *Image and Vision Computing* 152 (2024) 105308.
- [19] M. T. Usman, H. Khan, I. Rida, J. Koo, Lightweight transformer-driven multi-scale trapezoidal attention network for saliency detection, *Engineering Applications of Artificial Intelligence* 155 (2025) 110917.
- [20] H. Khan, M. T. Usman, J. Koo, Bilateral feature fusion with hexagonal attention for robust saliency detection under uncertain environments, *Information Fusion* 121 (2025) 103165.
- [21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [22] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, Y. J. Lee, Segment everything everywhere all at once, *Advances in neural information processing systems* 36 (2023) 19769–19782.
- [23] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, et al., Generalized decoding for pixel, image, and language, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15116–15127.
- [24] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, H. Shi, Oneformer: One transformer to rule universal image segmentation, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2989–2998.
- [25] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, *Nature Communications* 15 (1) (2024) 654.
- [26] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, K. Li, Sam on medical images: A comprehensive study on three prompt modes, *arXiv preprint arXiv:2305.00035* (2023).
- [27] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, A. V. Dalca, Universeg: Universal medical image segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21438–21451.
- [28] Y. Zhao, T. Zhou, Y. Gu, Y. Zhou, Y. Zhang, Y. Wu, H. Fu, Weakpolyp-sam: Segment anything model-driven weakly-supervised polyp segmentation, *Knowledge-Based Systems* (2025) 113701.
- [29] Y. Sun, J. Chen, S. Zhang, X. Zhang, Q. Chen, G. Zhang, E. Ding, J. Wang, Z. Li, Vrp-sam: Sam with visual reference prompt, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23565–23574.
- [30] Q.-H. Trinh, H.-D. Nguyen, B.-T. N. Ngoc, D. Jha, U. Bagci, M.-T. Tran, Sam-eg: segment anything model with egde guidance framework for efficient polyp segmentation, *arXiv preprint arXiv:2406.14819* (2024).
- [31] D. Alexey, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv: 2010.11929* (2020).
- [32] Z. Pan, B. Zhuang, J. Liu, H. He, J. Cai, Scalable vision transformers with hierarchical pooling, in: *Proceedings of the IEEE/cvf international conference on computer vision*, 2021, pp. 377–386.
- [33] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, S. J. Oh, Rethinking spatial dimensions of vision transformers, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11936–11945.
- [34] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.
- [35] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, *Advances in neural information processing systems* 34 (2021) 15908–15919.
- [36] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

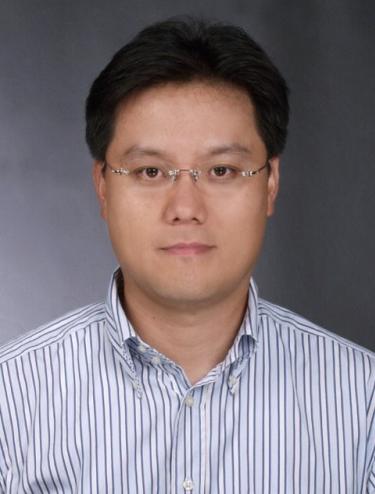
- [37] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, *Computational Visual Media* 8 (3) (2022) 415–424.
- [38] S. U. Amin, S. Taj, A. Hussain, S. Seo, An automated chest x-ray analysis for covid-19, tuberculosis, and pneumonia employing ensemble learning approach, *Biomedical Signal Processing and Control* 87 (2024) 105408.
- [39] L. Cai, M. Wu, L. Chen, W. Bai, M. Yang, S. Lyu, Q. Zhao, Using guided self-attention with local information for polyp segmentation, in: *International conference on medical image computing and computer-assisted intervention*, Springer, 2022, pp. 629–638.
- [40] X. Sun, P. Zhang, D. Wang, Y. Cao, B. Liu, Colorectal polyp segmentation by u-net with dilation convolution, in: *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, 2019, pp. 851–858.
- [41] B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, K. Ram, J. Joseph, M. Sivaprakasam, Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation, in: *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, 2019, pp. 7223–7226.
- [42] H. Zhang, J. Lian, Z. Yi, R. Wu, X. Lu, P. Ma, Y. Ma, Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation, *Biomedical Signal Processing and Control* 87 (2024) 105427.
- [43] Q. He, Q. Yang, M. Xie, Hctnet: A hybrid cnn-transformer network for breast ultrasound image segmentation, *Computers in Biology and Medicine* 155 (2023) 106629.
- [44] S. U. Amin, Y. Jung, M. Fayaz, B. Kim, S. Seo, Enhancing pine wilt disease detection with synthetic data and external attention-based transformers, *Engineering Applications of Artificial Intelligence* 159 (2025) 111655.
- [45] D. Banik, K. Roy, D. Bhattacharjee, M. Nasipuri, O. Krejcar, Polyp-net: A multimodel fusion network for polyp segmentation, *IEEE Transactions on Instrumentation and Measurement* 70 (2020) 1–12.
- [46] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, *Ieee Access* 9 (2021) 40496–40510.
- [47] C.-H. Huang, H.-Y. Wu, Y.-L. Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, *arXiv preprint arXiv:2101.07172* (2021).
- [48] M. Ali, T. Wu, H. Hu, T. Mahmood, Breast tumor segmentation using neural cellular automata and shape guided segmentation in mammography images, *Plos one* 19 (10) (2024) e0309421.
- [49] 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, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24, Springer, 2021, pp. 120–130.
- [50] Z. Yin, K. Liang, Z. Ma, J. Guo, Duplex contextual relation network for polyp segmentation, in: *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*, IEEE, 2022, pp. 1–5.
- [51] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I* 24, Springer, 2021, pp. 14–24.
- [52] M. Ali, H. Hu, T. Wu, M. Mansoor, Q. Luo, W. Zheng, N. Jin, Segmentation of mri tumors and pelvic anatomy via cgan-synthesized data and attention-enhanced u-net, *Pattern Recognition Letters* 187 (2025) 100–106.
- [53] H. Khan, T. Hussain, S. U. Khan, Z. A. Khan, S. W. Baik, Deep multi-scale pyramidal features network for supervised video summarization, *Expert Systems with Applications* 237 (2024) 121288.
- [54] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [55] H. Khan, S. U. Khan, W. Ullah, S. W. Baik, Optimal features driven hybrid attention network for effective video summarization, *Engineering Applications of Artificial Intelligence* 158 (2025) 111211.
- [56] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II* 26, Springer, 2020, pp. 451–462.
- [57] 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, *Computerized medical imaging and graphics* 43 (2015) 99–111.
- [58] N. Tajbakhsh, S. R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, *IEEE transactions on medical imaging* 35 (2) (2015) 630–644.
- [59] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, *Journal of healthcare engineering* 2017 (1) (2017) 4037190.
- [60] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, *International journal of computer assisted radiology and surgery* 9 (2014) 283–293.
- [61] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 2018, pp. 3–11.
- [62] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, Y. Yu, Adaptive context selection for polyp segmentation, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI* 23, Springer, 2020, pp. 253–262.
- [63] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, S. Song, Stepwise feature fusion: Local guides global, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 110–120.
- [64] Z. Qiu, Z. Wang, M. Zhang, Z. Xu, J. Fan, L. Xu, Bdg-net: boundary distribution guided network for accurate polyp segmentation, in: *Medical Imaging 2022: Image Processing*, Vol. 12032, SPIE, 2022, pp. 792–799.

- [65] N.-T. Bui, D.-H. Hoang, Q.-T. Nguyen, M.-T. Tran, N. Le, Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 7985–7994.

Journal Pre-proof

**Declaration of interests**

- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

		Hyeonjoon Moon received the BEng degree in electronics and computer engineering from Korea University, Republic of Korea in 1990, the MEng and the PhD degrees from State University of New York, USA in 1992 and 1999, respectively. He is currently a professor and chairman at Department of Computer Science and Engineering, Sejong University, Republic of Korea. His current research interests include image processing, biometrics, artificial intelligence, and machine learning
		Wenqi Zhang received a Bachelor's degree from the Computer Science and Technology, Shandong University of Technology (SDUT) in 2020, and the Master's degree in Computer Science and Engineering from Sejong University in 2024. Currently, she is pursuing a Ph.D. degree at the Computer Science and Engineering, Sejong University. Her research interests include Deep learning, image processing, and computer vision.
		ZHANG YUE received her B.Eng. degree in 2021 from Yanbian University, China. She is currently pursuing an M.S. degree in Computer Science at Sejong University, South Korea. Her research interests include computer vision, object recognition, object detection, and image segmentation.

	<p>Muhammad Fayaz completed his Bachelor's degree from Islamia College University, Peshawar, and a Master's in Computer Engineering from Cyprus International University, specializing in computer vision, deep learning, and machine learning. He is currently a Research Assistant at the Computer Vision and Pattern Recognition (CVPR) Lab, Sejong University. His research focuses on medical image analysis and land cover mapping, developing deep learning algorithms for improved diagnostic precision and sustainable environmental monitoring.</p>
	<p>L. Minh Dang received the BEng degree in information systems from University of Information Technology, VNU HCMC, Vietnam in 2016, and the PhD degree in computer science from Sejong University, Seoul, Republic of Korea in 2021. Starting from 2017, he joined Sejong University, Republic of Korea. His current research interests include computer vision, natural language processing, and artificial intelligence.</p>
	<p>Tan N. Nguyen received the M.E. degree from Ho Chi Minh City University of Technology (HCMUT), Vietnam, and the Ph.D. degree from Sejong University, South Korea, in 2019. He is currently an Assistant Professor with the Department of Architectural Engineering, Sejong University. His research interests include developing numerical methods, application robust methods to model structure considering modern materials, and new theoretical models. In addition, he also investigates highly nonlinear problems, instability of structures, and applies deep learning to structural analysis.</p>