RESEARCH ARTICLE

Scale-Specific Visual Sensing for Colonoscopy Polyp Segmentation via Hybrid CNN-Transformer Attention

Ikram Majeed Khan[1,*] **and Wisal Khan**[2]

[1] Coventry University, Coventry CV1 5FB, United Kingdom
[2] Northwest School of Medicine, Hayatabad, Peshawar 25000, Pakistan

Abstract

Automated colonoscopy constitutes a critical visual sensing task in clinical diagnostic pipelines, where precise segmentation of colorectal polyps from endoscopic image streams is essential for timely cancer diagnosis and prevention. Nevertheless, current segmentation methods contend with intrinsic variability in polyp appearance, differences in size, shape, and texture, while preserving computational efficiency necessary for clinical implementation. In this paper, we present a novel segmentation architecture that integrates scale-specific attention mechanisms within a hybrid CNN-Transformer backbone to address these challenges. Our model employs Coordinate Attention for high-resolution feature maps to preserve spatial details essential for boundary delineation, and Channel Attention for deep semantic features to enhance discriminative capacity. These representations are progressively integrated through a hierarchical decoder with specialized fusion modules: Semantic Fusion for high-level features, and Detail-Preserving Fusion for low-level features. The proposed architecture achieves state-of-the-art performance across five benchmark datasets, demonstrating that scale-specific visual sensing can deliver superior generalization and robustness in challenging clinical imaging scenarios, with direct applicability to real-time intelligent sensing systems for gastrointestinal diagnostics.

Keywords: visual sensing, medical imaging, hybrid CNN-Transformer, multi-scale attention, semantic fusion, deep learning.

1 Introduction

Colonoscopy represents a paradigmatic visual sensing task in which intelligent systems must interpret complex endoscopic image streams under real-time constraints to support clinical decision-making. Colorectal cancer (CRC) constitutes the third most frequently diagnosed malignancy globally and continues to be the second primary cause of cancer-associated mortality [1]. Most CRC cases arise from precancerous lesions, known as colon polyps, which are abnormal tissue proliferations arising from the mucosal layer of the colon. These polyps, particularly adenomatous variants, carry a substantial risk of malignancy, making early identification and excision essential therapeutic measures. Colonoscopy is widely recognized as the clinical gold standard for polyp identification and removal, and prompt

Submitted: 03 January 2026
Accepted: 02 March 2026
Published: 28 June 2026

Vol. 3, **No.** 2, 2026.
10.62762/TSCC.2026.664028

***Corresponding author:**
✉ Ikram Majeed Khan
Khani72@coventry.ac.uk

Citation
Khan, I. M., & Khan, W. (2026). Scale-Specific Visual Sensing for Colonoscopy Polyp Segmentation via Hybrid CNN-Transformer Attention. *ICCK Transactions on Sensing, Communication, and Control*, 3(2), 109–123.

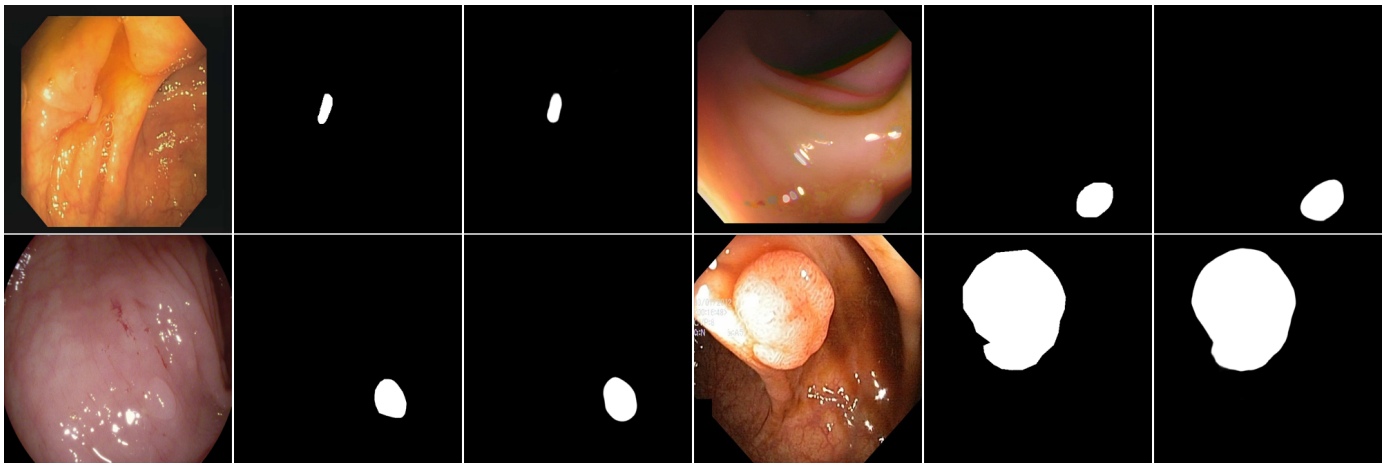

Figure 1. Predicted segmentation outputs from our model under complex polyp conditions.

diagnosis markedly increases the five-year survival rate [2].

However, manual outlining of polyp margins is subjective and prone to error due to variations in clinician skill, fatigue, and the inherent challenges of interpreting visual cues. Consequently, there is a pressing need for automated and accurate polyp segmentation systems that can serve as the perception backbone of intelligent clinical sensing pipelines, augmenting diagnostic decision-making and reducing variability [3]. Automated segmentation of polyps is a challenging task due to their heterogeneous appearance. Polyps can range from tiny sub-millimeter nodules to large, irregular masses and exhibit diverse morphologies, including sessile, pedunculated, and flat types, with textures ranging from smooth to villous [4]. These variations require robust models that capture multi-scale spatial information and contextual dependencies while preserving fine boundary details, as evidenced by pyramid pooling and multi-scale feature extraction strategies in recent polyp segmentation research [5]. Although recent advances in deep learning, particularly encoder-decoder frameworks like U-Net [6], have led to significant performance improvements, several key challenges remain unresolved:

- **Loss of fine details:** Traditional convolutional networks downsample spatial resolutions through pooling and strided convolutions, leading to blurred boundaries and poor localization of small or subtle polyps. This feature misalignment during aggregation and the neglect of boundary information have been identified as primary contributors to segmentation degradation [7].
- **Ineffective multi-scale fusion:** Current feature fusion techniques often assume homogeneous feature relevance across spatial regions and scales, neglecting the diverse characteristics between central polyp regions and complex boundaries.
- **Inadequate boundary modeling:** Existing boundary-aware approaches relying on edge prediction or auxiliary tasks [8, 9] frequently fail to capture transitional features at polyp margins, resulting in coarse or incomplete segmentation.

In addition to architectural constraints, multiple practical issues obstruct the clinical adoption of leading models. Several techniques demonstrate limited generalization across diverse colonoscopy datasets owing to differences in equipment, imaging settings, and patient demographics, as systematic cross-dataset evaluations have revealed significant performance gaps between training and unseen test distributions [10]. Moreover, transformer-based frameworks, while effective at capturing long-range dependencies [11], often incur substantial computational cost and inference latency, reducing their suitability for deployment in real-time visual sensing systems where continuous endoscopic image streams must be processed within clinical time constraints [12, 13]. Driven by these challenges, we introduce a new lightweight architecture that achieves an effective trade-off among spatial accuracy, semantic richness, and computational efficiency. The proposed method combines convolutional inductive priors with transformer-driven global representation learning, while maintaining scale-aware features through an improved attention and fusion mechanism. Furthermore, we strengthen the training process using multi-scale supervision, enabling stable performance across diverse lesion characteristics and varying

imaging environments.

The main contributions of our work are summarized as follows:

- **Hybrid Feature Extraction Framework:** We design a hybrid backbone that combines ConvNeXt-T convolutional blocks with vision transformer modules, enabling effective modeling of both local textures and global contexts crucial for polyp segmentation.
- **Scale-Specific Attention Mechanisms:** We introduce a novel attention strategy that applies distinct mechanisms at different feature levels: Coordinate Attention (COA) for preserving spatial detail in shallow features, and Channel Attention for enhancing semantic selectivity in deeper layers.
- **Hierarchical Feature Fusion Decoder:** We construct a decoder with two specialized fusion modules. Semantic Fusion Module (SFM) and Detail-Preserving Fusion Module (DPFM) are each tailored to the information content and resolution of the respective feature stage.
- **Multi-Scale Supervision Strategy:** We employ a progressive supervision scheme at multiple decoder outputs to enable balanced optimization and accurate segmentation.
- **State-of-the-Art Sensing Performance:** Comprehensive experiments on five challenging benchmark datasets show that our method consistently outperforms existing approaches, validating its potential as a reliable perception module within intelligent clinical sensing systems. In particular, it achieves better detection of small polyps and more accurate delineation of intricate boundaries under varying conditions (see Figure 1).

2 Related Work

Our research builds upon advances in three interconnected areas: segmentation backbone architectures, polyp-specific modeling strategies, and adaptive multi-scale integration mechanisms.

2.1 Advances in Segmentation Architectures

Convolutional neural networks (CNNs) have significantly advanced medical image segmentation. Long et al. [14] laid the foundation with fully convolutional networks (FCNs) for pixel-level segmentation. The introduction of U-Net by Ronneberger et al. [15] was a milestone, leveraging an encoder-decoder architecture with skip connections to preserve spatial information. U-Net++ [16] further refined this architecture by introducing dense skip pathways to reduce the semantic gap between the encoder and decoder features. Boundary-aware segmentation has become increasingly important for clinical precision. Fang et al. [17] proposed integrating area-boundary constraints, while Hatamizadeh et al. [18] introduced edge-aware loss functions. Recent advances have included vision transformer architectures, such as TransUNet [19] and MedT [20], which apply self-attention to capture global context. Despite their effectiveness, transformers often demand significant computational resources, limiting their integration into real-time clinical sensing pipelines where low-latency inference is critical for continuous endoscopic monitoring.

2.2 Polyp Segmentation Strategies

Polyp segmentation relies on common medical imaging techniques but presents distinct challenges due to variation in polyp size, texture, and overall shape [21]. U-Net variants remain dominant: MSNet [22] employs dense multi-scale connections, and PraNet [23] focuses on difficult boundaries via reverse attention mechanisms. Multi-scale representation is crucial in this field. Cai et al. [24] combined CNNs and transformers to enhance scale robustness. Lou et al. [25] proposed CaraNet, a context axial reverse attention network that leverages channel-wise feature pyramids to improve segmentation accuracy for small medical objects including colorectal polyps. For boundary enhancement, methods have progressed from probability map adjustments (e.g., SANet [26]) toward multi-task learning strategies that integrate edge-based and region-level supervision [11, 12]. Although frequency-aware analysis remains underinvestigated in this field, it shows promise for decoupling fine structural details from broader contextual information. Lightweight architectures for polyp segmentation seek to reduce parameter counts while maintaining boundary precision; however, balancing these competing objectives remains challenging, as multi-scale feature fusion strategies that prioritize efficiency often sacrifice fine-grained boundary delineation [10, 27].

2.3 Adaptive Multi-scale Feature Fusion

Considering the diversity in polyp scale, multi-scale fusion is essential for attaining precise segmentation performance. Vision-based models have proposed techniques such as adaptive kernel convolutions [28] to handle multi-scale feature representations. Within medical imaging, parallel multi-scale attention strategies [29] have demonstrated that progressively fusing features at different resolutions improves segmentation performance across polyps of varying sizes and morphologies. He et al. [30] proposed adaptive mechanisms for scale handling, while Tomar et al. [31] and Sinha and Dolz [32] demonstrated the advantages of multi-level semantic integration, incorporating size-aware and region-level feature hierarchies to improve segmentation of variable-scale polyps. Srivastava et al. [33] proposed MSRF-Net, a multi-scale residual fusion network that progressively exchanges multi-scale information across encoder stages, improving biomedical image segmentation. Our approach introduces progressive attention modules that adaptively fuse features across scales, balancing accuracy with real-time feasibility.

3 Proposed Methodology

3.1 Feature Extraction with Vision Transformer and ConvNeXt

To address the dual challenges of computational efficiency and multi-scale feature learning, we propose a hybrid backbone that combines the strengths of convolutional neural networks (CNNs) and vision transformers (ViTs). Specifically, we use ConvNeXt-T blocks [50] for shallow feature extraction, integrated with Swin Transformer modules [49] for deep semantic modeling, as our foundational feature extractor. This design, as shown in Figure 2, leverages the inductive biases and spatial locality of convolutions alongside the global context modeling capabilities of transformers, resulting in a hierarchical representation well-suited for polyp segmentation tasks. The input image, resized to 224×224, is processed through a series of convolutional stages followed by transformer blocks. The architecture incorporates local-to-global attention mechanisms and progressive channel expansion to enhance feature diversity across different layers. As a result, the hybrid backbone generates four distinct feature maps at multiple scales:

- $\mathbf{x}_1 \in \mathbb{R}^{B\times 96\times 56\times 56}$: ConvNeXt-extracted shallow features capturing fine-grained spatial details and texture information;
- $\mathbf{x}_2 \in \mathbb{R}^{B\times 192\times 28\times 28}$: ConvNeXt-derived features with enhanced boundary awareness and local pattern recognition;
- $\mathbf{x}_3 \in \mathbb{R}^{B\times 384\times 14\times 14}$: Swin Transformer-generated mid-level features encoding semantic information through self-attention;
- $\mathbf{x}_4 \in \mathbb{R}^{B\times 768\times 7\times 7}$: deep transformer-enriched representations capturing long-range dependencies and global context.

The progressive channel expansion from 96 to 768 channels ensures increasingly rich feature representations while maintaining the spatial hierarchy critical for accurate segmentation. The ConvNeXt blocks in the early stages employ large-kernel (7×7) convolutions with depth-wise separable operations, efficiently extracting local patterns with minimal computational overhead. The Swin Transformer modules embedded in deeper layers employ window-based self-attention with shifted windows to capture global relationships while reducing computational complexity. Each stage in the backbone reduces spatial resolution via strided operations while preserving the original aspect ratio, an essential property for precise polyp localization. The shallow ConvNeXt features ($\mathbf{x}_1$, $\mathbf{x}_2$) retain detailed texture and edge information, which is crucial for detecting small and subtle polyps. In contrast, the deeper Swin features ($\mathbf{x}_3$, $\mathbf{x}_4$) exploit transformer-based contextual modeling to disambiguate complex backgrounds and accommodate variations in polyp appearance across diverse endoscopic imaging conditions.

3.2 Spatial Context Encoding via Coordinate Attention

To enhance the representational power of high-resolution feature maps while preserving critical spatial information, we incorporate Coordinate Attention (COA) [48] mechanisms into the first two feature maps ($\mathbf{x}_1$ and $\mathbf{x}_2$) extracted from our ConvNeXt-T blocks. These shallow features ($\mathbf{x}_1 \in \mathbb{R}^{B\times 96\times 56\times 56}$, $\mathbf{x}_2 \in \mathbb{R}^{B\times 192\times 28\times 28}$) contain fine-grained spatial details and boundary cues essential for accurate polyp detection and segmentation. The COA module disentangles spatial dependencies along horizontal and vertical directions independently, thereby enabling more precise positional encoding.

For each feature map $\mathbf{x}_k$ where $k \in \{1, 2\}$, the COA process consists of three main steps:

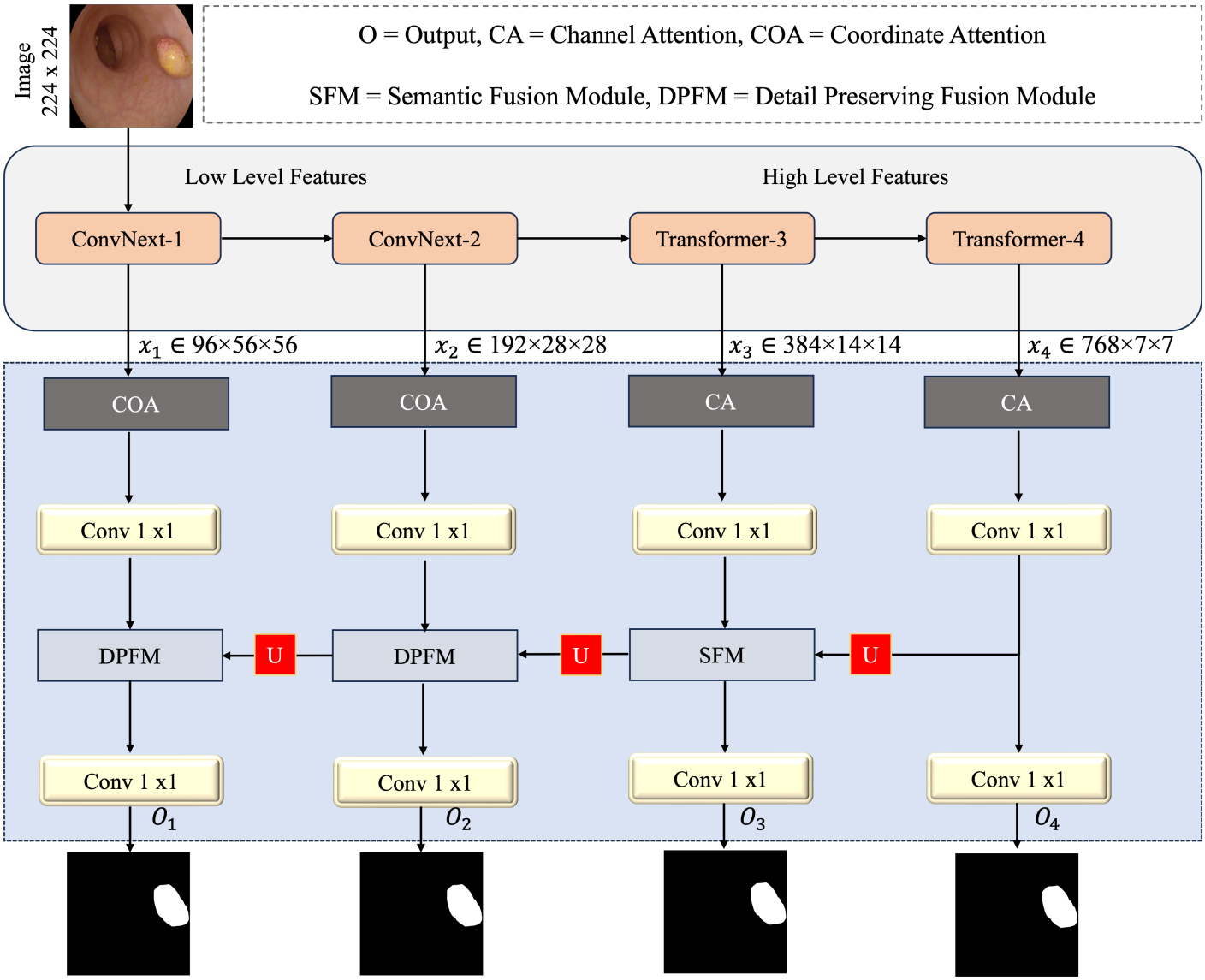

Figure 2. Visual illustration of the overall model architecture.

1. Coordinate-wise Pooling: We generate context descriptors by applying spatial pooling separately along the height (H) and width (W) dimensions:

$$\mathbf{z}_k^h = \frac{1}{W}\sum_{j=1}^{W}\mathbf{x}_k(:,:,:,j), \quad \mathbf{z}_k^w = \frac{1}{H}\sum_{i=1}^{H}\mathbf{x}_k(:,:,i,:) \tag{1}$$

2. Channel Transformation: The pooled features are concatenated and passed through a shared 1D convolution followed by ReLU activation:

$$\mathbf{g}_k = \delta\left(\text{Conv1D}\left([\mathbf{z}_k^h;\mathbf{z}_k^w]\right)\right) \tag{2}$$

where $\delta(\cdot)$ denotes the ReLU activation function and $[\cdot;\cdot]$ represents concatenation along the channel axis.

3. Attention Generation and Modulation: The transformed feature $\mathbf{g}_k$ is split into direction-specific attention maps, which are activated by a sigmoid function and applied to the input features:

$$\hat{\mathbf{x}}_k = \mathbf{x}_k \otimes \left(\sigma(\mathbf{f}_k^h) \otimes \sigma(\mathbf{f}_k^w)\right) \tag{3}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, $\otimes$ is the element-wise multiplication, and $\mathbf{f}_k^h \in \mathbb{R}^{C_k\times H_k\times 1}$, $\mathbf{f}_k^w \in \mathbb{R}^{C_k\times 1\times W_k}$ are the horizontal and vertical attention maps, respectively. Specifically, $C_1 = 96$, $H_1 = W_1 = 56$ for $\mathbf{x}_1$, and $C_2 = 192$, $H_2 = W_2 = 28$ for $\mathbf{x}_2$.

This directional attention mechanism offers several distinct advantages for polyp detection. First, it preserves the full spatial resolution of the feature maps, which is essential for accurate boundary localization. Second, it captures long-range dependencies in a computationally efficient manner, introducing minimal additional overhead. Third, by applying

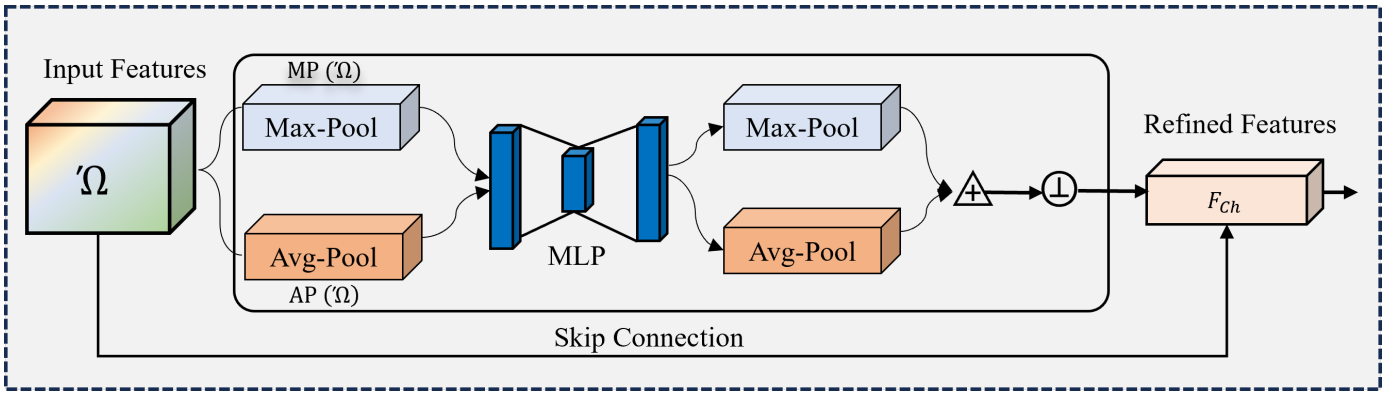

Figure 3. Feature flow inside CA.

direction-specific attention weighting, it enhances sensitivity to polyps with irregular, elongated, or fragmented morphologies. Moreover, it effectively suppresses background noise commonly present in endoscopic imagery while amplifying salient and discriminative regions. The resulting enhanced features, denoted as $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$, retain their original spatial dimensions while embedding enriched boundary and positional context, providing a robust foundation for subsequent multi-scale feature fusion in the decoder.

3.3 Channel Attention for Deep Semantic Feature Refinement

To enhance the discriminative capacity of deep semantic features extracted by our Swin Transformer modules, we integrate Channel Attention (CA, Figure 3) mechanisms into the final two feature maps, following the channel recalibration design of CBAM [47]. Although these features already benefit from the global context modeling inherent in transformers, further refinement is needed to amplify discriminative channels and suppress less relevant ones for more precise polyp localization. The CA module recalibrates channel-wise responses by modeling inter-channel dependencies as empirically validated in the study [47]. For each feature map $\mathbf{x}_k$ where $k \in \{3, 4\}$, the attention computation begins with a global average pooling operation that aggregates spatial information into a channel descriptor:

$$\mathbf{z}_k = \mathcal{F}_{sq}(\mathbf{x}_k) = \frac{1}{H_k \times W_k} \sum_{i=1}^{H_k} \sum_{j=1}^{W_k} \mathbf{x}_k(i, j) \tag{4}$$

where $\mathbf{z}_k \in \mathbb{R}^{B \times C_k \times 1 \times 1}$, with $C_3 = 384$ and $C_4 = 768$.

Next, we pass the pooled descriptor through a bottlenecked excitation block to learn channel-wise attention weights:

$$\mathbf{s}_k = \mathcal{F}_{ex}(\mathbf{z}_k) = \sigma\left(\mathbf{W}_2 \cdot \delta\left(\mathbf{W}_1 \cdot \mathbf{z}_k\right)\right) \tag{5}$$

Here, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C_k}{r} \times C_k}$ and $\mathbf{W}_2 \in \mathbb{R}^{C_k \times \frac{C_k}{r}}$ constitute the learnable parameters of the bottleneck structure, where $r = 16$ is the reduction ratio. The function $\delta(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ denotes the sigmoid activation function, producing the attention map $\mathbf{s}_k \in \mathbb{R}^{B \times C_k \times 1 \times 1}$.

These attention weights are then applied to the original feature map via channel-wise multiplication to recalibrate the feature responses:

$$\hat{\mathbf{x}}_k = \mathcal{F}_{scale}(\mathbf{x}_k, \mathbf{s}_k) = \mathbf{x}_k \otimes \mathbf{s}_k \tag{6}$$

where $\otimes$ indicates element-wise multiplication broadcast across spatial dimensions.

For the deepest feature map $\mathbf{x}_4$, we introduce a progressive attention mechanism involving two consecutive CA blocks separated by a non-linear transformation. The first attention stage is computed as:

$$\tilde{\mathbf{x}}_4 = \mathcal{F}_{scale}\left(\mathbf{x}_4, \mathcal{F}_{ex}(\mathcal{F}_{sq}(\mathbf{x}_4))\right) \tag{7}$$

Then, we apply a depthwise separable convolution followed by a second attention refinement:

$$\hat{\mathbf{x}}_4 = \mathcal{F}_{scale}\left(\text{Conv}_{3\times3}(\tilde{\mathbf{x}}_4), \mathcal{F}_{ex}(\mathcal{F}_{sq}(\text{Conv}_{3\times3}(\tilde{\mathbf{x}}_4)))\right) \tag{8}$$

This progressive design enables deeper recalibration of abstract semantic features while maintaining computational efficiency. The use of a lightweight convolution between attention stages introduces minimal overhead while enabling enhanced nonlinear feature transformation. Channel Attention significantly improves feature discriminability by emphasizing channels that are semantically relevant to polyps and attenuating those associated

with complex background noise. Despite its effectiveness, it adds only marginal computational cost. The refined deep features $\hat{\mathbf{x}}_3$ and $\hat{\mathbf{x}}_4$ thus carry enhanced semantic representations that are crucial for high-level segmentation cues. When these attention-enhanced Swin features are integrated with the spatially-preserved shallow ConvNeXt features $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ from the COA block, they together form a powerful multi-scale representation. This fusion exploits the complementary strengths of the two levels, enabling more accurate and robust polyp segmentation.

3.4 Feature Fusion Decoder for Multi-scale Integration

To effectively leverage the complementary information captured at different scales and through various attention mechanisms, we propose a hierarchical feature fusion decoder. This decoder progressively integrates the refined feature maps $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \hat{\mathbf{x}}_3, \hat{\mathbf{x}}_4$ to generate the final segmentation output. The architecture combines top-down and lateral connections with specialized fusion modules designed to handle the diverse characteristics of features at each level.

3.4.1 Top-Down Feature Pathway

The decoding process begins with the deepest feature map $\hat{\mathbf{x}}_4 \in \mathbb{R}^{B\times 768\times 7\times 7}$ and proceeds through a top-down pathway. The first operation is a 1×1 convolution to reduce the channel dimension:

$$\mathbf{p}_4 = \text{Conv}_{1\times 1}(\hat{\mathbf{x}}_4) \tag{9}$$

where $\mathbf{p}_4 \in \mathbb{R}^{B\times 256\times 7\times 7}$ represents a channel-reduced projection of $\hat{\mathbf{x}}_4$, reducing the channel dimension to 256 across all decoder levels. For each subsequent level $l \in \{3, 2, 1\}$, the following operations are performed:

$$\mathbf{u}_l = \text{Upsample}_{2\times}(\mathbf{p}_{l+1}) \tag{10}$$

$$\mathbf{p}_l = \text{FFM}_l(\mathbf{u}_l, \text{Conv}_{1\times 1}(\hat{\mathbf{x}}_l)) \tag{11}$$

where $\text{Upsample}_{2\times}$ denotes bilinear upsampling by a factor of 2, $\text{Conv}_{1\times 1}$ projects the refined feature map $\hat{\mathbf{x}}_l$ to 256 channels, and FFM_l is the Feature Fusion Module at level l.

3.4.2 Feature Fusion Modules

Each Feature Fusion Module (FFM) is tailored to effectively combine features from adjacent scales while preserving their unique characteristics. For high-level fusion (levels 4 and 3), we use a Semantic Fusion Module (SFM) that prioritizes semantic information from Swin Transformer features:

$$\text{FFM}_3(\mathbf{u}_4, \hat{\mathbf{x}}'_3) = \mathbf{u}_4 + \hat{\mathbf{x}}'_3 + \text{Conv}_{3\times 3}(\text{Concat}[\mathbf{u}_4, \hat{\mathbf{x}}'_3]) \tag{12}$$

where $\hat{\mathbf{x}}'_3 = \text{Conv}_{1\times 1}(\hat{\mathbf{x}}_3)$, and the addition of the concatenation operation through a 3×3 convolution enriches feature aggregation while maintaining semantic consistency.

For low-level fusion (levels 2 and 1), we employ a Detail-Preserving Fusion Module (DPFM) that emphasizes spatial details from ConvNeXt features:

$$\text{FFM}_l(\mathbf{u}_{l+1}, \hat{\mathbf{x}}'_l) = \text{Conv}_{3\times 3}(\text{Concat}[\mathbf{u}_{l+1} \otimes \gamma_l, \hat{\mathbf{x}}'_l]) + \hat{\mathbf{x}}'_l \tag{13}$$

where $l \in \{1, 2\}$, and γ_l is a spatial attention map derived from $\hat{\mathbf{x}}'_l$ that highlights boundary regions:

$$\gamma_l = \sigma(\text{Conv}_{1\times 1}(\text{Conv}_{3\times 3}(\hat{\mathbf{x}}'_l))) \tag{14}$$

This design allows the boundary-aware features from the COA module to guide the fusion process, preserving fine-grained details in the final segmentation.

3.4.3 Multi-scale Supervision and Output Generation

To facilitate effective gradient flow and deep supervision, we generate predictions at each decoder level:

$$\mathbf{o}_l = \text{Conv}_{1\times 1}(\mathbf{p}_l), \quad l \in \{1, 2, 3, 4\} \tag{15}$$

where $\mathbf{o}_l \in \mathbb{R}^{B\times 1\times H_l\times W_l}$ represents the prediction at level l. The final output is derived from the finest-resolution prediction through:

$$\mathbf{Y} = \sigma(\text{Upsample}_{4\times}(\mathbf{o}_1)) \tag{16}$$

where $\mathbf{Y} \in \mathbb{R}^{B\times 1\times 224\times 224}$ is the final segmentation mask, and σ denotes the sigmoid activation function. During training, we employ a weighted multi-scale loss function:

$$\mathcal{L} = \sum_{l=1}^{4} w_l \mathcal{L}_{\text{dice}}(\mathbf{o}_l, \mathbf{Y}_{gt}) + \lambda \mathcal{L}_{\text{bce}}(\mathbf{o}_1, \mathbf{Y}_{gt}) \tag{17}$$

where $\mathcal{L}_{\text{dice}}$ is the Dice loss, $\mathcal{L}_{\text{bce}}$ is the binary cross-entropy loss, $\mathbf{Y}_{gt}$ is the ground truth mask, w_l are level-specific weights (with $w_1 > w_2 > w_3 > w_4$), and λ balances the contribution of the BCE loss.

The proposed decoder architecture effectively integrates features from all scales, leveraging the fine spatial details captured by COA in ConvNeXt

Table 1. Progressive integration of architectural modules and their impact on performance across five benchmark datasets.

Modules	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Backbone	0.821	0.743	0.894	0.839	0.733	0.658	0.712	0.625	0.873	0.811
Backbone + COA	0.847	0.768	0.901	0.856	0.745	0.676	0.725	0.643	0.885	0.828
Backbone + COA + CA	0.879	0.812	0.931	0.887	0.779	0.713	0.758	0.684	0.918	0.865
Ours	**0.903**	**0.837**	**0.939**	**0.896**	**0.812**	**0.736**	**0.801**	**0.719**	**0.925**	**0.878**

Table 2. Comparison of the performance of different backbone architectures across five benchmark datasets.

Modules	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Pure ConvNeXt-T	0.788	0.709	0.872	0.815	0.701	0.623	0.683	0.591	0.851	0.784
Pure Swin	0.803	0.726	0.883	0.826	0.714	0.637	0.695	0.609	0.862	0.798
Hybrid ConvNeXt-T/Swin	0.821	0.743	0.894	0.839	0.733	0.658	0.712	0.625	0.873	0.811
Ours	**0.903**	**0.837**	**0.939**	**0.896**	**0.812**	**0.736**	**0.801**	**0.719**	**0.925**	**0.878**

features $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$, alongside the semantic information emphasized by Channel Attention in Swin Transformer features $\hat{\mathbf{x}}_3$ and $\hat{\mathbf{x}}_4$. The specialized fusion modules at each level ensure that the unique characteristics of each feature scale are preserved during integration, resulting in segmentation masks that exhibit both precise boundary delineation and accurate polyp region identification.

4 Experimental Evaluation Protocol

This section outlines the comprehensive evaluation methodology employed in our study, covering dataset curation, preprocessing techniques, computational infrastructure, and hyperparameter optimization. To thoroughly assess the efficacy of our model, we incorporate five established benchmark datasets. Performance evaluation is carried out using a diverse set of quantitative metrics: Mean Absolute Error (MAE), weighted F-measure (F_β^w), Structure-measure ($S\alpha$), Mean Enhanced-alignment Measure ($mE\xi$), mean Dice coefficient, and mean Intersection over Union (IoU). Additionally, detailed ablation analyses are performed to quantify the contribution of individual architectural components. This experimental framework facilitates both quantitative and qualitative comparisons with contemporary state-of-the-art methods. The experimental results consistently demonstrate that our methodology outperforms existing approaches, positioning it as a robust solution for polyp segmentation tasks.

4.1 Technical Configuration and Training Procedure

All experiments were carried out on a high-end computing setup equipped with an NVIDIA GeForce RTX 4090 GPU with 24GB of VRAM, enabling efficient handling of intensive computations. To address the wide range of polyp scales observed in the dataset, we adopted a multi-scale training strategy to strengthen the model's generalization across varying lesion sizes. All input images were resized to a fixed resolution of 224×224 pixels. The model converged optimally after 100 epochs with a batch size of 16, offering a balanced compromise between accuracy and computational cost. Guided by extensive empirical analysis and established practices in related studies, we used the AdamW optimizer with a learning rate of 0.0005 and a weight decay of 0.1.

4.2 Datasets and Comparative Analysis

We assess our approach using five publicly accessible and challenging benchmark datasets, adhering to the standard evaluation protocols of PraNet [23]: KvasirSEG [34], ClinicDB [35], ColonDB [36], Endoscene [37], and ETIS [38]. The KvasirSEG dataset comprises high-resolution polyp images acquired during diverse endoscopic examinations, whereas ClinicDB contains samples collected from real clinical endoscopy sessions. The ColonDB, Endoscene, and ETIS datasets provide additional complementary benchmarks for cross-dataset evaluation. Our training setup employed a merged dataset of 1,450 samples, including 900 images from KvasirSEG and 550 from ClinicDB. For validation, we used 162 images (100 from KvasirSEG and 62 from ClinicDB), along

Table 3. Impact of the progressive addition of supervision levels on model performance.

Modules	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
L1	0.879	0.812	0.931	0.887	0.779	0.713	0.758	0.684	0.918	0.865
L1 + L2	0.888	0.822	0.926	0.880	0.796	0.720	0.782	0.701	0.917	0.866
L1 + L2 + L3	0.892	0.826	0.930	0.884	0.801	0.726	0.789	0.708	0.920	0.869
L1 + L2 + L3 + L4	0.898	0.832	0.935	0.890	0.807	0.731	0.795	0.714	0.922	0.872
Ours	**0.903**	**0.837**	**0.939**	**0.896**	**0.812**	**0.736**	**0.801**	**0.719**	**0.925**	**0.878**

with the complete ColonDB dataset and selected portions of EndoScene and ETIS. For comparison, we benchmarked our model against several leading polyp segmentation approaches, including U-Net [15], UNet++ [16], PraNet [23], ACSNet [39], UACANet [40], Polyp-PVT [41], BDG-Net [42], SSform [43], PVT-CASCADE [44], MEGANet [45], and EMCAD [46].

4.3 Evaluation Criteria

We evaluate segmentation performance using six standard quantitative metrics commonly adopted by reference methods in polyp segmentation literature. Specifically, we report Mean Absolute Error (MAE) to assess pixel-level accuracy, Weighted F-measure (F_β^w) to emphasize boundary precision, Structure-measure ($S\alpha$) and Mean Enhanced-alignment Measure ($mE\xi$) to evaluate structural consistency and alignment quality, and Mean Dice Coefficient together with Mean Intersection over Union (IoU) to quantify region overlap between predicted masks and ground truth annotations. All metrics are computed at the dataset level, where higher values indicate better performance for all measures except MAE, for which lower values are preferred. This evaluation protocol ensures fair and comprehensive comparison with existing state-of-the-art approaches.

4.4 Ablation Studies

To rigorously evaluate the contribution of each architectural component, we conducted comprehensive ablation experiments by progressively integrating modules into our full model and measuring performance across five benchmark datasets: *Endoscene*, *ClinicDB*, *ColonDB*, *ETIS*, and *Kvasir-SEG*. Table 1 summarizes the performance impact of each module. The baseline model, comprising only the backbone, achieved solid mDice scores ranging from 0.712 (ETIS) to 0.894 (ClinicDB). Integrating the **COA** module led to consistent improvements, notably **+2.6%** on Endoscene and **+1.2%** on Kvasir-SEG, highlighting COA's ability to refine spatially-aware features.

The incorporation of the **Channel Attention (CA)** module contributed additional gains across all datasets, with the most notable improvements on Endoscene (**+1.3%** mDice) and ColonDB (**+1.6%**), indicating CA's strength in modeling inter-channel dependencies. Compared to the Backbone+COA+CA configuration, the full model showed further improvements ranging from **+2.4%** (Endoscene) to **+4.3%** (ETIS), emphasizing the synergistic effect of the integrated components. These results validate the importance of each module and its collective impact on segmentation performance.

4.4.1 Backbone Architecture Comparison

To validate the efficacy of our proposed hybrid backbone, we compared three variants: (1) a pure ConvNeXt-T architecture, (2) a pure Swin transformer, and (3) our hybrid ConvNeXt-T/Swin backbone. Table 2 reports the comparative results. The ConvNeXt-T backbone provided a strong CNN-based baseline, with mDice scores ranging from 0.683 (ETIS) to 0.872 (ClinicDB), but it struggled on datasets that require long-range context modeling, such as ETIS and ColonDB. The pure Swin backbone improved over ConvNeXt-T, particularly on Endoscene (+1.5%) and ColonDB (+1.3%), benefiting from its global attention mechanism. However, it showed limitations in capturing localized, fine-grained details critical for accurate segmentation.

Our proposed hybrid backbone outperformed both individual architectures across all datasets. Compared to ConvNeXt-T, the hybrid model improved by +3.3% (Endoscene) and +3.2% (ColonDB). Relative to Swin, gains were also consistent: +1.7% (ETIS) and +1.1% (Kvasir-SEG). These findings confirm the effectiveness of our hybrid design, which combines the spatial precision of convolutions with the global contextual understanding of transformers.

Moreover, when integrated with our attention modules and multi-scale supervision, the complete model significantly outperformed the hybrid backbone alone, achieving improvements of +8.2% (Endoscene) and +8.9% (ETIS). The hybrid backbone lays a strong foundation, while attention modules enhance multi-scale feature representations.

4.4.2 *Impact of Multi-Loss Supervision*

To investigate the effect of hierarchical supervision, we conducted experiments by incrementally adding supervision levels to the loss function. Table 3 presents the corresponding results across all datasets. Starting with supervision at only the highest-resolution level (L1), the model already achieved competitive

Table 4. Comparison of the proposed model with SOTA methods. The performance is evaluated using mean Dice (mDice) and mean Intersection over Union (mIoU) scores.

Models	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
UNet (2015)	0.710	0.627	0.823	0.755	0.504	0.436	0.398	0.335	0.818	0.746
UNet++ (2018)	0.707	0.624	0.794	0.729	0.482	0.408	0.401	0.344	0.821	0.743
PraNet (2020)	0.871	0.797	0.899	0.849	0.712	0.640	0.628	0.567	0.898	0.840
ACSNet (2020)	0.863	0.787	0.882	0.826	0.716	0.649	0.578	0.509	0.898	0.838
UACANet-S (2021)	0.902	0.837	0.916	0.870	0.783	0.704	0.694	0.615	0.905	0.852
Polyp-PVT (2021)	0.900	0.833	0.937	0.889	0.808	0.727	0.787	0.706	0.917	0.864
BDG-Net (2022)	0.897	0.828	0.909	0.859	0.792	0.719	0.764	0.685	0.904	0.853
SSform-L (2022)	0.892	0.822	0.903	0.850	0.798	0.716	0.790	0.712	0.915	0.861
PVT-CASCADE (2023)	0.898	0.833	0.923	0.878	0.809	0.728	0.808	0.735	0.926	0.876
MEGANet-ResNet (2024)	0.887	0.818	0.930	0.885	0.781	0.706	0.789	0.709	0.911	0.859
PVT-EMCAD-B2 (2024)	0.885	0.812	0.929	0.881	0.819	0.736	0.794	0.717	0.924	0.878
Ours	**0.903**	**0.837**	**0.939**	**0.896**	**0.812**	**0.736**	**0.801**	**0.719**	**0.925**	**0.878**

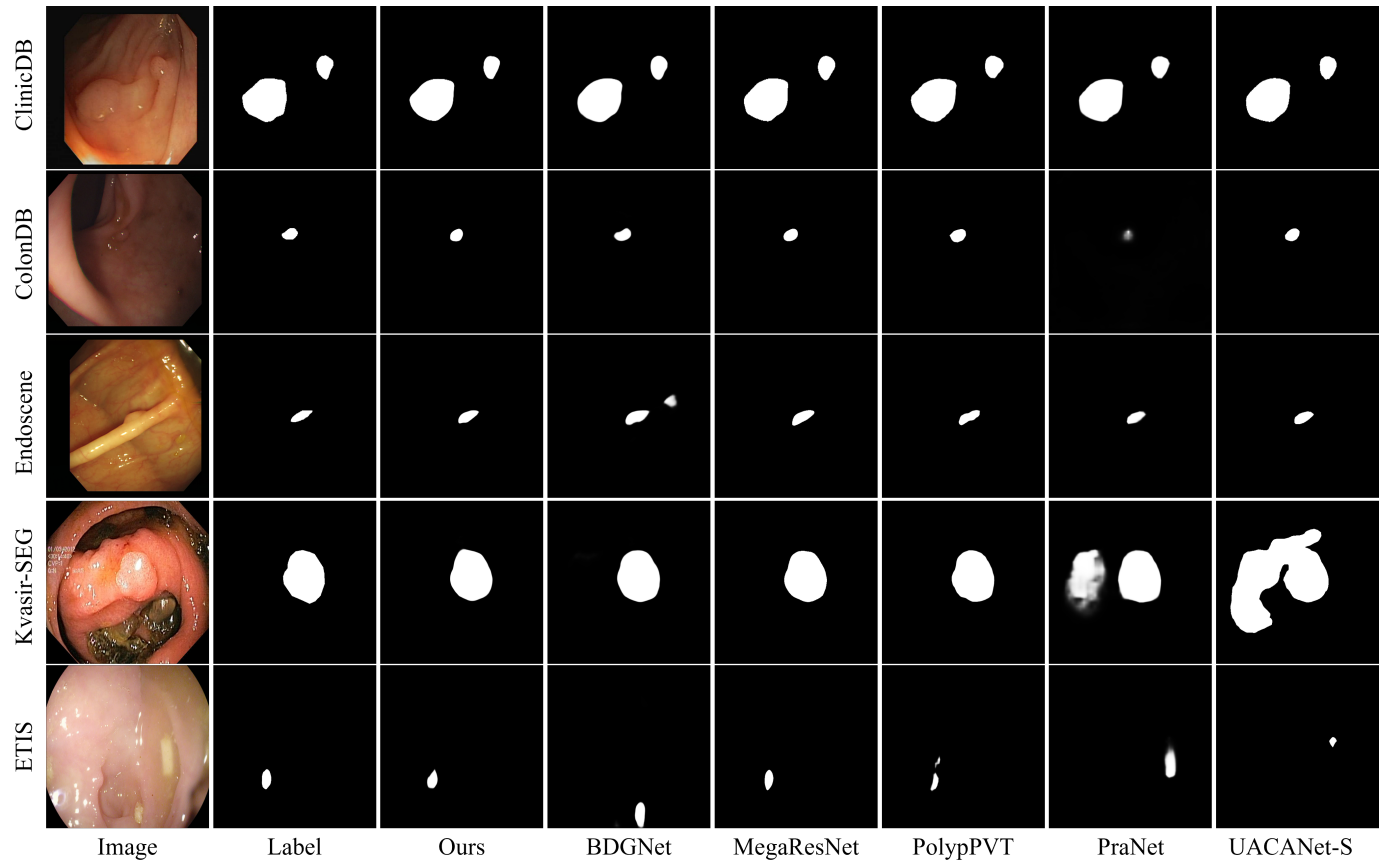

Figure 4. Qualitative comparison of our network with leading approaches including BDG-Net, MEGANet-ResNet, Polyp-PVT, PraNet, and UACANet-S.

Table 5. F-measure (F_β^w) and MAE performance comparison across five benchmark datasets.

Models	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	F_β^w	MAE	F_β^w	MAE	F_β^w	MAE	F_β^w	MAE	F_β^w	MAE
UNet (2015)	0.684	0.022	0.811	0.019	0.491	0.059	0.366	0.036	0.794	0.055
UNet++ (2018)	0.687	0.018	0.785	0.022	0.467	0.061	0.390	0.035	0.808	0.048
PraNet (2020)	0.843	0.010	0.896	0.009	0.699	0.043	0.600	0.031	0.885	0.030
ACSNet (2020)	0.825	0.013	0.873	0.011	0.697	0.039	0.530	0.059	0.882	0.032
Polyp-PVT (2021)	0.884	0.007	0.936	0.006	0.795	0.031	0.750	0.013	0.911	0.023
BDG-Net (2022)	0.876	0.006	0.905	0.007	0.714	0.015	0.776	0.031	0.896	0.028
SSform-L (2022)	0.875	0.007	0.906	0.008	0.790	0.031	0.761	0.015	0.911	0.023
PVT-CASCADE (2023)	0.882	0.008	0.923	0.013	0.798	0.029	0.775	0.016	0.918	0.020
MEGANet-ResNet (2024)	0.863	0.009	0.931	0.008	0.766	0.038	0.753	0.015	0.904	0.026
PVT-EMCAD-B2 (2024)	0.869	0.007	0.927	0.010	0.804	0.028	0.759	0.016	0.920	0.021
Ours	**0.889**	**0.007**	**0.922**	**0.006**	**0.825**	**0.020**	**0.784**	**0.020**	**0.923**	**0.019**

Table 6. S-measure ($S\alpha$) and Mean E-measure ($mE\xi$) performance comparison across five benchmark datasets.

Models	Endoscene		ClinicDB		ColonDB		ETIS		Kvasir-SEG	
	$S\alpha$	$mE\xi$	$S\alpha$	$mE\xi$	$S\alpha$	$mE\xi$	$S\alpha$	$mE\xi$	$S\alpha$	$mE\xi$
UNet (2015)	0.843	0.848	0.889	0.913	0.710	0.692	0.684	0.643	0.858	0.881
UNet++ (2018)	0.839	0.834	0.873	0.891	0.692	0.680	0.683	0.629	0.862	0.886
PraNet (2020)	0.925	0.950	0.936	0.963	0.820	0.847	0.794	0.808	0.915	0.944
ACSNet (2020)	0.923	0.939	0.927	0.947	0.829	0.839	0.754	0.737	0.920	0.941
Polyp-PVT (2021)	0.935	0.973	0.949	0.985	0.865	0.913	0.871	0.906	0.925	0.956
BDG-Net (2022)	0.937	0.967	0.938	0.970	0.866	0.895	0.866	0.894	0.918	0.952
SSform-L (2022)	0.939	0.969	0.934	0.963	0.866	0.901	0.881	0.905	0.923	0.957
PVT-CASCADE (2023)	0.934	0.965	0.939	0.969	0.864	0.910	0.886	0.906	0.928	0.964
MEGANet-ResNet (2024)	0.924	0.956	0.950	0.977	0.845	0.897	0.866	0.912	0.916	0.952
PVT-EMCAD-B2 (2024)	0.921	0.965	0.943	0.973	0.869	0.919	0.877	0.902	0.929	0.966
Ours	**0.942**	**0.976**	**0.947**	**0.982**	**0.875**	**0.920**	**0.884**	**0.919**	**0.932**	**0.954**

results, with mDice ranging from 0.758 (ETIS) to 0.931 (ClinicDB), indicating the significance of fine-grained, high-resolution features. Introducing supervision at level L2 (L1+L2) resulted in noticeable improvements, especially on ColonDB (+1.7%) and ETIS (+2.4%). This suggests that mid-level features contribute valuable boundary-aware information for complex morphologies. The addition of level L3 supervision further improved performance by +0.4% to +0.7% mDice, leveraging the contextual representation capabilities of enriched features. Including supervision at level L4 continued this trend, yielding notable gains on ClinicDB and ETIS, attributed to semantically rich features refined by the CA module. Finally, full supervision at all four levels (L1–L4) delivered the best results, with additional performance boosts of +0.3% to +0.6% mDice over L1–L4. These results demonstrate the effectiveness of multi-scale supervision in enhancing feature refinement and gradient flow.

4.5 Comparison with State-of-the-Art Methods

In this section, we compare the performance of our proposed network against various SOTA approaches, highlighting both quantitative and qualitative analyses.

4.5.1 Quantitative Analysis

Tables 4, 5, and 6 provide a comprehensive comparison between our proposed model and several state-of-the-art (SOTA) methods across

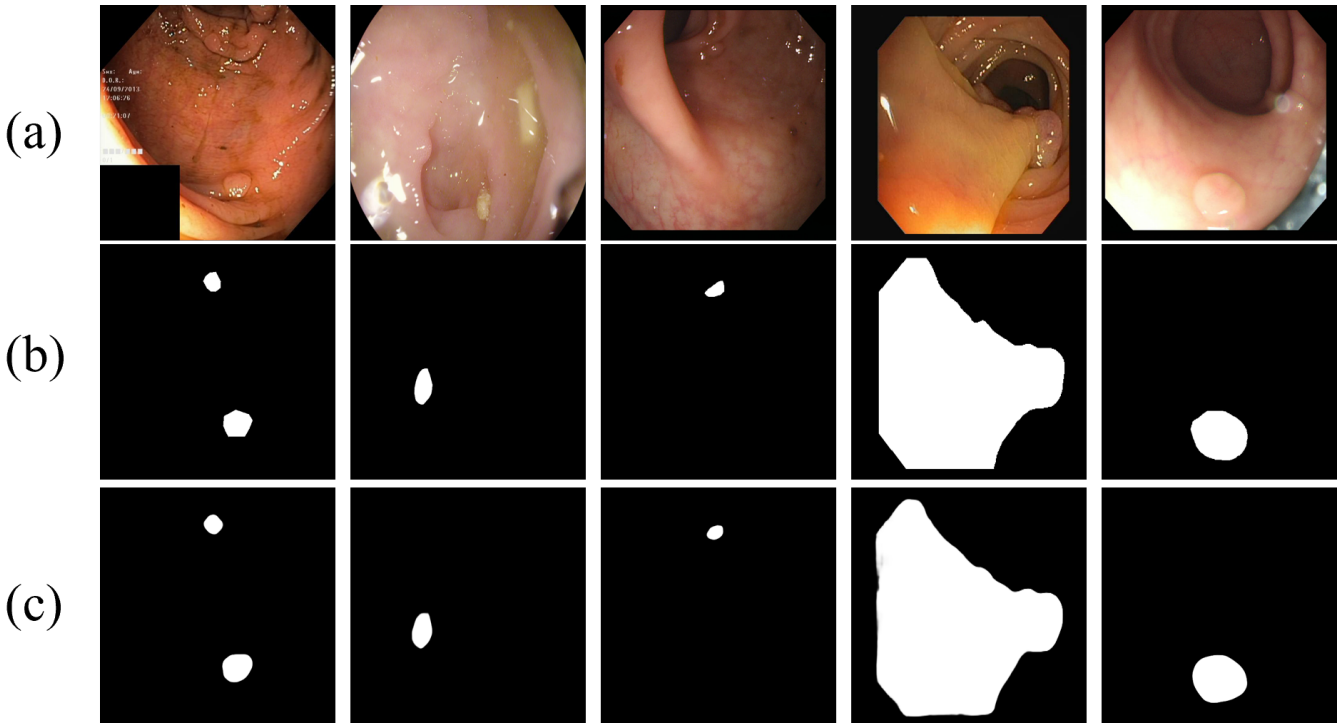

Figure 5. Qualitative comparison of our network on the benchmark dataset. From top to bottom: (a) input image, (b) ground truth, and (c) model predictions.

five benchmark datasets. The evaluation metrics include mean Dice (mDice), mean Intersection over Union (mIoU), weighted F-measure (F_β^w), Mean Absolute Error (MAE), S-measure ($S\alpha$), and Mean E-measure ($mE\xi$). Our model achieves top or near-top performance across all datasets and metrics. Specifically, it attains the highest mDice scores on three out of five datasets (Endoscene, ClinicDB, and Kvasir-SEG), and competitive performance on ColonDB and ETIS, where it closely approaches the leading methods, demonstrating its robustness and generalization ability. For instance, on the challenging ETIS dataset, our model achieves an mDice of 0.801 and an mIoU of 0.719, outperforming most recent methods such as PVT-EMCAD-B2 (mDice 0.794) and MEGANet-ResNet (mDice 0.789), while remaining competitive with PVT-CASCADE (mDice 0.808). Furthermore, in terms of F_β^w and MAE (Table 5), our approach maintains superior precision-recall balance while minimizing pixel-wise prediction errors. This is especially evident in datasets such as ColonDB and ETIS, where accurate segmentation is more challenging due to low contrast and complex backgrounds. Lastly, Table 6 presents S-measure and E-measure scores, highlighting our model's ability to preserve object structure and capture spatial alignment effectively. With the highest $S\alpha$ and $mE\xi$ on Kvasir-SEG and strong results across all datasets, our model confirms its comprehensive strength in both boundary-aware and holistic saliency metrics. These results collectively affirm that our proposed method not only surpasses traditional architectures like UNet and UNet++ but also outperforms advanced Transformer-based and hybrid frameworks.

4.5.2 Qualitative Evaluation

A visual comparative analysis between our architecture and leading approaches is presented in Figures 4 and 5. Figure 4 shows segmentation outputs from our network alongside six contemporary methodologies (BDGNet, MegaRes2Net, MegaResNet, PolyPVT, PraNet, and UACANet-S) across all five datasets. The rows correspond to ClinicDB, ColonDB, Endoscene, Kvasir-SEG, and ETIS, respectively, and each row highlights a case of particular complexity. The comparison particularly emphasizes the effectiveness of our approach in delineating small polyps against heterogeneous backgrounds. Figure 5 extends this analysis by examining our network's adaptability across diverse clinical scenarios. These visualizations substantiate the model's consistent performance across varying imaging conditions, polyp morphologies, and complexity levels. Key findings include remarkable consistency across datasets, indicating strong generalizability, and robust performance in visually challenging environments with confounding elements. The visual evidence from both figures corroborates our quantitative findings, confirming

that our architecture achieves state-of-the-art visual segmentation quality in polyp segmentation. The presented visual comparisons provide qualitative confirmation of our network's improved boundary precision and region coherence compared to existing approaches.

5 Conclusion

This paper introduced a hybrid CNN-Transformer architecture for colonoscopy-based visual polyp sensing that integrates scale-specific attention mechanisms and specialized fusion modules, addressing the core challenges of intelligent medical image sensing systems. By leveraging Coordinate Attention and Channel Attention at different feature scales, our model enhances boundary precision and region coherence across diverse clinical scenarios. The hierarchical decoder, equipped with tailored fusion modules, enables effective multi-scale integration while maintaining computational efficiency through lightweight operations and window-based attention. Extensive evaluations across five benchmark datasets confirm the model's superior performance, especially in challenging cases involving small polyps and complex backgrounds. Future work includes extending this approach to video-based segmentation, other medical imaging tasks, and integrating uncertainty estimation for clinical decision support. Overall, our method advances automated polyp segmentation as a visual sensing capability, balancing perceptual accuracy and computational efficiency for integration into real-time clinical sensing and diagnostic support systems.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Hossain, M. S., Karuniawati, H., Jairoun, A. A., Urbi, Z., Ooi, D. J., John, A., ... & Hadi, M. A. (2022). Colorectal cancer: a review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers, 14*(7), 1732. [CrossRef]

[2] Kim, N. H., Jung, Y. S., Jeong, W. S., Yang, H. J., Park, S. K., Choi, K., & Park, D. I. (2017). Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal research, 15*(3), 411. [CrossRef]

[3] Sanchez-Peralta, L. F., Bote-Curiel, L., Picon, A., Sanchez-Margallo, F. M., & Pagador, J. B. (2020). Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. *Artificial intelligence in medicine, 108*, 101923. [CrossRef]

[4] Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., ... & Lu, H. (2023). M^2 SNet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv preprint arXiv:2303.10894*. [CrossRef]

[5] Hu, K., Chen, W., Sun, Y., Hu, X., Zhou, Q., & Zheng, Z. (2023). PPNet: Pyramid pooling based network for polyp segmentation. *Computers in biology and medicine, 160*, 107028. [CrossRef]

[6] Tomar, N. K., Jha, D., Riegler, M. A., Johansen, H. D., Johansen, D., Rittscher, J., ... & Ali, S. (2022). Fanet: A feedback attention network for improved biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems, 34*(11), 9375-9388. [CrossRef]

[7] Su, Y., Cheng, J., Zhong, C., Jiang, C., Ye, J., & He, J. (2023). Accurate polyp segmentation through enhancing feature fusion and boosting boundary performance. *Neurocomputing, 545*, 126233. [CrossRef]

[8] Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., & Shen, D. (2023). Cross-level feature aggregation network for polyp segmentation. *Pattern Recognition, 140*, 109555. [CrossRef]

[9] Yue, G., Han, W., Jiang, B., Zhou, T., Cong, R., & Wang, T. (2022). Boundary constraint network with cross layer feature integration for polyp segmentation. *IEEE Journal of Biomedical and Health Informatics, 26*(8), 4090-4099. [CrossRef]

[10] Tomar, N. K., Jha, D., & Bagci, U. (2023, January). Dilatedsegnet: A deep dilated segmentation network for polyp segmentation. In *International conference on multimedia modeling* (pp. 334-344). Cham: Springer International Publishing. [CrossRef]

[11] Yang, H., Chen, Q., Fu, K., Zhu, L., Jin, L., Qiu, B., ... & Lu, Y. (2022). Boosting medical image segmentation via conditional-synergistic convolution and lesion decoupling. *Computerized Medical Imaging and Graphics, 101*, 102110. [CrossRef]

[12] Xiao, H., Li, L., Liu, Q., Zhu, X., & Zhang, Q. (2023). Transformers in medical image segmentation: A

review. *Biomedical Signal Processing and Control, 84*, 104791. [CrossRef]

[13] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical image analysis, 88*, 102802. [CrossRef]

[14] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440). [CrossRef]

[15] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. [CrossRef]

[16] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018, September). Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis* (pp. 3-11). Cham: Springer International Publishing. [CrossRef]

[17] Fang, Y., Chen, C., Yuan, Y., & Tong, K. Y. (2019, October). Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 302-310). Cham: Springer International Publishing. [CrossRef]

[18] Hatamizadeh, A., Terzopoulos, D., & Myronenko, A. (2019, October). End-to-end boundary aware networks for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging* (pp. 187-194). Cham: Springer International Publishing. [CrossRef]

[19] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. [CrossRef]

[20] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021, September). Medical transformer: Gated axial-attention for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 36-46). Cham: Springer International Publishing. [CrossRef]

[21] Usman, M. T., Khan, H., Khan, H., Rida, I., Zhu, X., & Koo, J. (2025). HMPFormer: Hierarchical vision transformer with multi-perspective feature learning for precise polyp segmentation. *Image and Vision Computing*, 105777. [CrossRef]

[22] Zhao, X., Zhang, L., & Lu, H. (2021, September). Automatic polyp segmentation via multi-scale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 120-130). Cham: Springer International Publishing. [CrossRef]

[23] Fan, D. P., Ji, G. P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020, September). Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 263-273). Cham: Springer International Publishing. [CrossRef]

[24] Cai, L., Wu, M., Chen, L., Bai, W., Yang, M., Lyu, S., & Zhao, Q. (2022, September). Using guided self-attention with local information for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 629-638). Cham: Springer Nature Switzerland. [CrossRef]

[25] Lou, A., Guan, S., Ko, H., & Loew, M. H. (2022, April). CaraNet: context axial reverse attention network for segmentation of small medical objects. In *Medical Imaging 2022: Image Processing* (Vol. 12032, pp. 81-92). SPIE. [CrossRef]

[26] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., & Cui, S. (2021, September). Shallow attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 699-708). Cham: Springer International Publishing. [CrossRef]

[27] Liu, F., Hua, Z., Li, J., & Fan, L. (2022). Dbmf: Dual branch multiscale feature fusion network for polyp segmentation. *Computers in Biology and Medicine, 151*, 106304. [CrossRef]

[28] Li, X., Wang, W., Hu, X., & Yang, J. (2019, June). Selective Kernel Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 510-519). IEEE. [CrossRef]

[29] Song, P., Li, J., & Fan, H. (2022). Attention based multi-scale parallel network for polyp segmentation. *Computers in Biology and Medicine, 146*, 105476. [CrossRef]

[30] He, J., Deng, Z., & Qiao, Y. (2019). Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3562-3572). [CrossRef]

[31] Tomar, N. K., Jha, D., Bagci, U., & Ali, S. (2022). TGANet: Text-guided attention for improved polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* (pp. 151–160). Springer [CrossRef]

[32] Sinha, A., & Dolz, J. (2020). Multi-scale self-guided attention for medical image segmentation. *IEEE journal of biomedical and health informatics, 25*(1), 121-130. [CrossRef]

[33] Srivastava, A., Jha, D., Chanda, S., Pal, U., Johansen, H. D., Johansen, D., ... & Halvorsen, P. (2021). MSRF-Net: A multi-scale residual fusion network for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics, 26*(5), 2252-2263. [CrossRef]

[34] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., De Lange, T., Johansen, D., & Johansen, H. D. (2019, December). Kvasir-seg: A segmented polyp dataset.

In *International conference on multimedia modeling* (pp. 451-462). Cham: Springer International Publishing. [CrossRef]

[35] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & 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. [CrossRef]

[36] Tajbakhsh, N., Gurudu, S. R., & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging, 35*(2), 630-644. [CrossRef]

[37] Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., ... & Courville, A. (2017). A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering, 2017*(1), 4037190. [CrossRef]

[38] Silva, J., Histace, A., Romain, O., Dray, X., & Granado, B. (2014). Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery, 9*(2), 283-293. [CrossRef]

[39] Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., & Yu, Y. (2020, September). Adaptive context selection for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 253-262). Cham: Springer International Publishing. [CrossRef]

[40] Kim, T., Lee, H., & Kim, D. (2021, October). Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 2167-2175). [CrossRef]

[41] Dong, B., Wang, W., Fan, D. P., Li, J., Fu, H., & Shao, L. (2021). Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*. [CrossRef]

[42] Qiu, Z., Wang, Z., Zhang, M., Xu, Z., Fan, J., & Xu, L. (2022, April). BDG-Net: boundary distribution guided network for accurate polyp segmentation. In *Medical Imaging 2022: Image Processing* (Vol. 12032, pp. 792-799). SPIE. [CrossRef]

[43] Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., & Song, S. (2022, September). Stepwise feature fusion: Local guides global. In *International conference on medical image computing and computer-assisted intervention* (pp. 110-120). Cham: Springer Nature Switzerland. [CrossRef]

[44] Rahman, M. M., & Marculescu, R. (2023, January). Medical Image Segmentation via Cascaded Attention Decoding. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 6211-6220). IEEE. [CrossRef]

[45] Bui, N. T., Hoang, D. H., Nguyen, Q. T., Tran, M. T., & Le, N. (2024, January). MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 7970-7979). IEEE. [CrossRef]

[46] Rahman, M. M., Munir, M., & Marculescu, R. (2024, June). EMCAD: Efficient Multi-Scale Convolutional Attention Decoding for Medical Image Segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11769-11779). IEEE. [CrossRef]

[47] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision* (pp. 3-19). Cham: Springer International Publishing. [CrossRef]

[48] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13713–13722). IEEE. [CrossRef]

[49] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). IEEE. [CrossRef]

[50] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11976–11986). IEEE. [CrossRef]

Ikram Majeed Khan earned his Bachelor's degree in Software Engineering from Islamia College University, Peshawar, and a Master's degree in Computer Science from Coventry University, England, UK. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, and Visual Intelligence. (Email: Khani72@coventry.ac.uk)

Wisal Khan is a committed MBBS student currently pursuing a medical degree at the Northwest School of Medicine. He is passionate about combining clinical knowledge with evidence-based research. He actively participates in academic activities, clinical rotations, and volunteer work, with a strong interest in medical writing, case discussions, and promoting community health awareness. He remains dedicated to expanding his understanding and making meaningful contributions through collaboration with AI scientists. (Email: wisal7377@gmail.com)