RESEARCH ARTICLE

ColoSegNet: Visual Intelligence Driven Triple Attention Feature Fusion Network for Endoscopic Colorectal Cancer Segmentation

Talha Salman[1], Alexandros Gazis[2,3], Aizaz Ali[4], Muhammad Hassan Khan[5], Muhammad Ali[6], Haseeb Khan[7] and Hasnain Ali Shah[8,*]

[1] Erciyes University, Kayseri, Turkey
[2] Department of Electrical and Computer Engineering, School of Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
[3] Heriot-Watt University, Edinburgh, United Kingdom
[4] University of the West of Scotland, London E14 2BE, United Kingdom
[5] Khyber Medical University, Peshawar, Pakistan
[6] University of Haripur, Khyber Pakhtunkhwa, Pakistan
[7] Northwest School of Medicine, Peshawar 25000, Pakistan
[8] School of Computing, University of Eastern Finland, Joensuu 80100, Finland

Abstract

Accurate segmentation of colorectal cancer (CRC) from endoscopic images is crucial for computer-aided diagnosis. Visual intelligence enhances detection precision, supporting clinical decision-making. However, current segmentation methods often struggle with accurately delineating fine-grained lesion boundaries due to limited context comprehension and inadequate attention to optimal features. Additionally, the poor fusion of multi-scale semantic cues hinders performance, especially in complex endoscopic scenarios. To address these issues, we introduce ColoSegNet, a Visual Intelligence-Driven Triple Attention Feature Fusion Network designed for high-precision CRC segmentation. Our approach begins with an analysis of backbone networks to identify optimal intermediate-level feature extraction. The architecture includes a non-local attention module to capture long-range dependencies, enhanced by channel and spatial attention mechanisms for better feature selectivity across semantic hierarchies. Multi-scale fusion strategies are used to preserve both low-level details and high-level context, enabling accurate localization under challenging conditions. Extensive evaluations across four publicly available endoscopic datasets, along with cross-dataset assessments, demonstrate the effectiveness and generalizability of ColoSegNet, outperforming state-of-the-art segmentation methods.

Keywords: colorectal segmentation, visual intelligence,

Academic Editor:
Xue-Bo JIN

Submitted: 04 May 2025
Accepted: 20 May 2025
Published: 25 June 2025

Vol. 2, **No.** 2, 2025.
10.62762/TIS.2025.385365

***Corresponding author:**
Hasnain Ali Shah
alishah@uef.fi

Citation
Salman, T., Gazis, A., Ali, A. Khan, M. H., Ali, M., Khan, H., & Shah, H. A. (2025). ColoSegNet: Visual Intelligence Driven Triple Attention Feature Fusion Network for Endoscopic Colorectal Cancer Segmentation. *ICCK Transactions on Intelligent Systematics*, 2(2), 125–136.

medical imaging, attention mechanisms, feature fusion, semantic segmentation, deep learning.

1 Introduction

Colorectal cancer (CRC) is among the most prevalent and deadly forms of cancer worldwide. According to recent global estimates, CRC is the third most common cancer worldwide [1]. However, early detection and accurate diagnosis play a pivotal role in reducing mortality rates. Endoscopic imaging is a widely adopted modality for the screening of CRC, allowing real-time visualization of lesions, polyps, and mucosal abnormalities. However, manual interpretation of endoscopic images is subject to intra- and inter-observer variability, as well as operator fatigue. In this context, visual intelligence, powered by deep learning, has emerged as a promising solution for automating CRC image analysis, particularly for semantic segmentation, where precise delineation of lesion regions is critical for surgical planning and disease assessment.

Despite significant advances in medical image segmentation, CRC segmentation remains challenging due to several inherent factors. First, the lesion boundaries are often irregular, blurred, or embedded within complex textures, making it difficult for standard convolutional neural networks (CNNs) to capture subtle spatial variations. Second, existing methods often overlook the importance of modeling both global dependencies and fine-grained local structures simultaneously, which are essential for accurate boundary refinement. Third, many architectures fail to effectively fuse multi-level features, resulting in an incomplete semantic representation, especially under varied lighting conditions, specular reflections, or occlusions in real-world endoscopy. These limitations necessitate the development of a more robust and context-aware segmentation framework tailored to the visual complexity of CRC imagery.

Several recent approaches have made notable progress in polyp segmentation (PS), including PraNet by Fan et al. [2], Polyp-PVT by Dong et al. [3], UACANet by Kim et al. [4], and CASCADE by Rahman and Marculescu [5]. The method [2] introduced a parallel reverse attention framework designed to enhance boundary localization in colonoscopy images. Their model integrates a partial decoder for aggregating salient features and employs reverse attention for progressive refinement. Dong et al. [3] developed a transformer-based architecture that incorporates modules for multi-scale fusion, noise suppression, and contextual robustness, resulting in high performance in challenging conditions. Kim et al. [4] proposed UACANet, which enhances feature representation through uncertainty-aware learning, using an augmented saliency map embedded within a modified U-Net to achieve strong performance across five datasets. Finally, CASCADE by Rahman and Marculescu [5] focuses on an attention-guided decoder that integrates attention gates and convolutional modules to effectively capture both long-range dependencies and local context, thereby improving segmentation accuracy.

1.1 Limitations

Recent advances in CRC segmentation have highlighted significant challenges in existing methodologies. Many current approaches demonstrate limited generalization and contextual understanding, primarily due to their dependence on local attention mechanisms and single-scale feature aggregation. Although recent efforts such as dual-encoder multi-scale fusion networks [6] and multi-scale efficient transformer architectures [7] have made progress in addressing scale variability, challenges in long-range dependency modeling and cross-dataset generalization persist across the broader literature. A common limitation among various architectures is their inability to effectively capture long-range dependencies. Additionally, they often struggle with the effective fusion of low-level and high-level features, while also neglecting fine-grained spatial details. This results in suboptimal performance, especially evident in lesions characterized by irregular shapes, indistinct boundaries, or variable textures. Furthermore, there is a paucity of methodologies that ensure robust cross-dataset generalization across varied clinical environments. These identified limitations drive the development of our proposed ColoSegNet, which strategically employs triple attention mechanisms coupled with a multi-scale skip-connected fusion strategy. This design aims to significantly enhance the robustness, accuracy, and clinical applicability of CRC segmentation techniques.

1.2 Contributions

Our work makes the following main contributions:

- **Early Intermediate Feature Analysis:** We performed systematic intermediate experiments to pinpoint the most discriminative and contextually rich feature representations. This investigation shaped the architectural design, directing optimal feature selection to enhance attention alignment and fusion effectiveness.

- **Triple Attention Integration:** We introduce a unified triple attention scheme that sequentially combines non-local, channel, and spatial attention modules. This design enables effective modeling of both long-range dependencies and localized semantic cues, enhancing the discriminative power of features extracted at different network depths.
- **Multi-Scale Skip-Connected Feature Fusion:** We utilized a multi-scale feature fusion strategy to aggregate features on multiple scales and semantic levels. This mechanism ensures the preservation of spatial details while incorporating contextual richness, significantly improving segmentation accuracy for lesions with varied shapes and textures.
- **Extensive Experiments and Cross Evaluation:** To validate the robustness and generalization of our approach, we conduct comprehensive experiments on four publicly available endoscopic CRC datasets. Additionally, our model is trained on a combined KvasirSEG and ClinicDB subset and evaluated on four independent benchmark datasets, demonstrating strong generalization to unseen data distributions across diverse clinical conditions.

2 Related Work

Attention mechanisms have become foundational in medical image analysis and computer vision tasks by selectively emphasizing informative features, with demonstrated applications in colorectal polyp segmentation using parallel reverse convolutional attention [8], adaptive lesion-aware attention for 3D colorectal tumor segmentation [9], and CNN-based attention mechanisms for colorectal tumor segmentation from CT scans [10]. In medical imaging specifically, attention-based architectures have been employed to improve lesion localization, boundary refinement, and contextual reasoning [28]. Recent breakthroughs in machine learning have propelled its application across various medical tasks, providing robust assistance in clinical decision-making [11–13]. The foundation for modern medical image segmentation was established with the U-Net architecture, which introduces an encoder-decoder structure with skip connections that preserve spatial information crucial for delineating irregular boundaries. Fang et al. [14] pioneered the integration of area-boundary constraints through selective feature connection mechanisms, while Hatamizadeh et al. [15] advanced boundary awareness through specialized edge-aware loss functions. In the same year, Jha et al. [16] introduced ResUNet++, incorporating residual connections with multi-level feature fusion across decoder stages to improve segmentation performance on challenging polyp datasets. Murugesan et al. [17] proposed a multi-task learning framework that simultaneously optimizes edge detection and region segmentation, leading to more coherent polyp delineation. Building on these foundations, Hu et al. [18] addressed the challenge of polyp size variability through pyramid pooling-based feature extraction strategies, enabling effective multi-scale representation for accurate polyp segmentation. He et al. [19] further refined this approach by introducing dynamically adaptive mechanisms that adjust feature processing based on input characteristics. During this period, Zhou et al. [20] extended the U-Net concept with U-Net++, implementing nested and dense skip pathways to bridge the semantic gap between encoder and decoder features.

The year 2021 marked significant advancements with transformer architectures revolutionizing the capture of global contextual information. TransUNet by Chen et al. [21] strategically combined transformers with convolutional features to leverage both global relationships and local details. Lin et al. [22] used a lightweight ViT to reduce computational complexity while preserving global context modeling capabilities. Similarly, MedT, proposed by Valanarasu et al. [23], applied self-attention mechanisms specifically tailored for medical image segmentation tasks. Wei et al. [24] developed SANet during this period, employing shallow attention modules to enhance feature discrimination at polyp boundaries. Yue et al. [25] develop uncertainty-aware models to address variations in visual conditions and enhance robustness against illumination changes and complex backgrounds.

More recent innovations have further refined polyp segmentation approaches. In 2022, Cai et al. [30] presented a guided self-attention approach that incorporates local information to enhance context modeling, effectively combining spatial locality with global attention to achieve robust

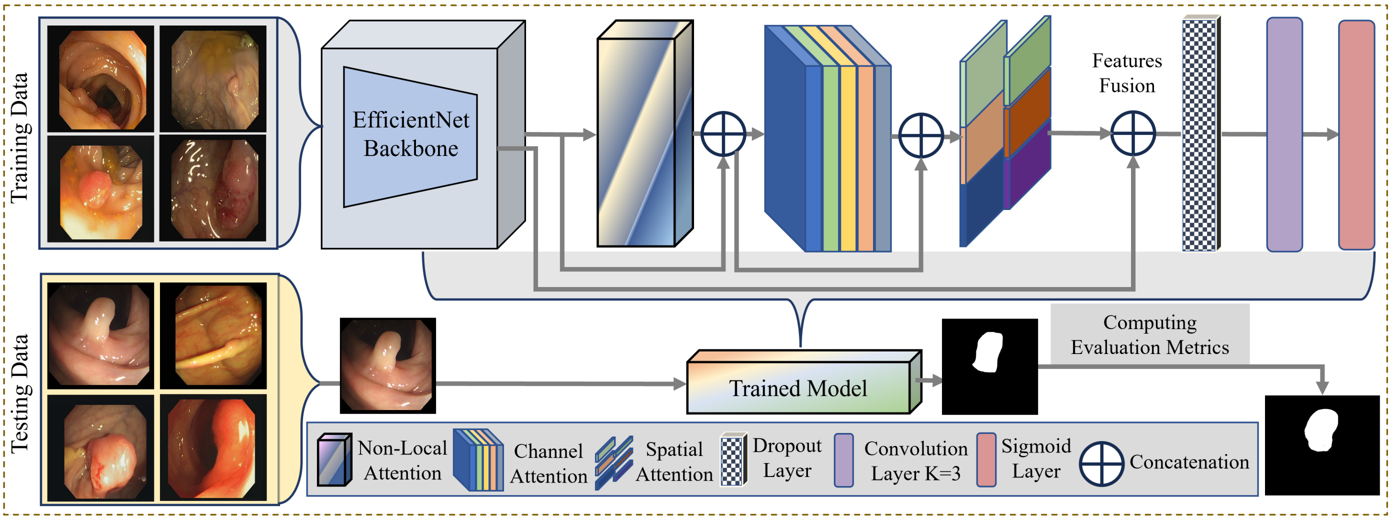

Figure 1. The visual illustration of the overall model architecture.

performance across polyps of various scales. Yang et al. [31] advanced segmentation performance through conditional-synergistic convolution and lesion decoupling strategies, enabling more discriminative feature representations for medical image segmentation. By 2023, Jeong et al. [32] explored lightweight architectures that successfully reduced model complexity, though often at the expense of boundary precision. The most recent contributions in 2024 include the work of Ahamed et al. [27], who demonstrated the efficacy of attention-guided feature selection across multiple scales, and Sinha and Dolz [26], who showcased the benefits of semantic integration across feature hierarchies through multi-level fusion. Although these methodologies have significantly enhanced the domain, they primarily utilize either inflexible feature fusion frameworks with limited adaptability or overly complex attention mechanisms that compromise computational efficiency; for instance, multi-level feature fusion networks that combine multiple attention modules can achieve strong segmentation results but at the cost of increased architectural complexity [29]. Our investigation addresses this constraint by presenting computationally efficient progressive attention modules that adaptively combine features across various scales, thereby harmonizing segmentation precision with real-time processing capabilities, an essential factor for clinical implementation.

3 Proposed Methodology

In this section, we present our proposed **ColoSegNet**, a visual intelligence-driven segmentation framework that integrates triple attention mechanisms and hierarchical feature fusion for precise CRC segmentation from endoscopic images. The architecture comprises four major stages as shown in Figure 1: (1) backbone-based deep feature extraction, (2) global context modeling via non-local attention, (3) selective feature enhancement using channel attention, and (4) spatial detail preservation through optimized spatial attention. The network is optimized using a binary cross-entropy loss tailored for pixel-wise segmentation tasks.

3.1 Deep Feature Extraction

The proposed network leverages a lightweight yet powerful convolutional backbone for initial feature extraction. Among various modern architectures, **EfficientNetB7** has demonstrated exceptional performance across classification and segmentation benchmarks due to its compound scaling strategy, which uniformly scales network width, depth, and resolution. Its design strikes a balance between accuracy and computational efficiency, making it suitable for deployment in real-time clinical settings. In **ColoSegNet**, intermediate features from selected stages of EfficientNetB7 are extracted and processed to preserve both low-level texture and high-level semantic cues. These features serve as input to the subsequent attention modules, ensuring rich contextual information flow while maintaining spatial fidelity, which is crucial for delineating ambiguous polyp boundaries.

3.2 Non-Local Attention for Global Context Modeling

Traditional convolutional operations are inherently local and often fail to capture long-range dependencies.

To address this, we incorporate a non-local attention module inspired by the self-attention mechanism in transformers. The non-local block computes responses at a position as a weighted sum of features at all positions in the input, thereby enabling the network to model long-range spatial dependencies that extend beyond the local receptive field.

Mathematically, given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the non-local operation can be defined as:

$$\mathbf{Y}_i = \frac{1}{\mathcal{C}(\mathbf{X})} \sum_{\forall j} f(\mathbf{X}_i, \mathbf{X}_j) g(\mathbf{X}_j) \tag{1}$$

where $f(\cdot,\cdot)$ computes a pairwise affinity between positions i and j, $g(\cdot)$ transforms the input, and $\mathcal{C}$ is a normalization factor. This design enables the model to focus on relevant semantic regions across the entire field of view, which is particularly beneficial for large or irregularly shaped tumors.

In our network, the features extracted by EfficientNetB7 are first passed through this non-local module to enrich them with global context before entering the subsequent attention layers.

3.3 Channel Attention for Selective Feature Emphasis

After global enhancement, we introduce a channel attention block to emphasize the most informative feature channels as shown in Figure 2. Inspired by the squeeze-and-excitation (SE) mechanism, channel attention computes inter-channel dependencies and assigns adaptive weights to highlight relevant semantic dimensions. This is achieved by applying global average pooling followed by a gating mechanism composed of two fully connected layers with a non-linear activation in between. Formally, given a feature map $\mathbf{F}$, the recalibrated output $\mathbf{F}'$ is:

$$\mathbf{F}' = \sigma(\mathrm{FC}_2(\delta(\mathrm{FC}_1(\mathrm{GAP}(\mathbf{F}))))) \cdot \mathbf{F} \tag{2}$$

where σ denotes the sigmoid function, δ is the ReLU activation, and GAP is global average pooling. This enables the network to emphasize semantically informative channels that contribute most to accurate lesion segmentation.

3.4 Optimized Spatial Attention

To preserve and emphasize spatial structure, we incorporate a spatial attention module following channel refinement. While conventional spatial attention modules typically use a 7×7 convolution kernel, our design reduces the kernel size to 3×3 to lower computational complexity without sacrificing performance. This optimization is particularly effective in endoscopic scenarios where high-resolution details are vital for delineating lesion borders.

The spatial attention mechanism first applies average and max pooling across channels, concatenates the results, and then applies a convolution to generate a spatial attention map. This map modulates the input feature map to strengthen spatially relevant regions:

$$\mathbf{M}_{\text{spatial}} = \sigma(\mathrm{Conv}_{3\times 3}([\mathrm{AvgPool}(\mathbf{F}); \mathrm{MaxPool}(\mathbf{F})])) \tag{3}$$

$$\mathbf{F}'' = \mathbf{M}_{\text{spatial}} \cdot \mathbf{F} \tag{4}$$

This component ensures refined spatial focus on polyp regions while maintaining real-time processing efficiency.

3.5 Loss Function

To train the proposed **ColoSegNet** for binary segmentation, we employ the binary cross-entropy (BCE) loss. BCE measures the discrepancy between predicted probabilities and ground truth labels for each pixel and is well-suited for foreground-background segmentation tasks:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{5}$$

where p_i and y_i denote the predicted probability and ground truth label for the i^{th} pixel, respectively. This formulation encourages accurate pixel-wise classification and helps the network learn both lesion presence and shape.

4 Experimental Analysis

4.1 Implementation Details and Optimization Strategy

Our experiments were conducted on a computing infrastructure featuring an NVIDIA RTX 3090 GPU with 24 GB of memory capacity. To address the heterogeneous nature of polyp morphology, we implemented a scale-adaptive training regime that enhances robustness to variations in size. Images underwent preprocessing to a standardized dimension of 224×224 pixels. The training spanned 100 epochs, with mini-batches of 8 samples, striking an optimal balance between computational demands and convergence quality. After comprehensive hyperparameter tuning experiments, we selected the AdamW optimization algorithm with an initial learning rate of 5×10^{-4} and weight decay parameter

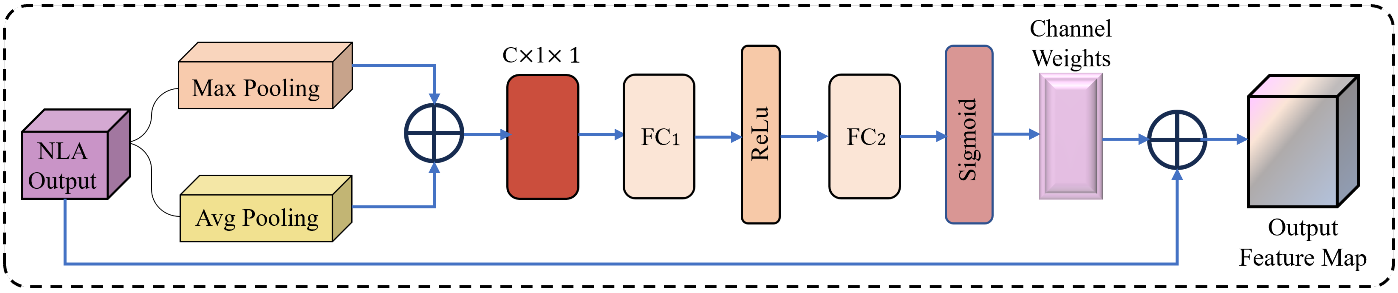

Figure 2. Visual illustration of the feature flow within the Channel Attention mechanism.

Table 1. Performance comparison of different backbone architectures on Endoscene and ClinicDB datasets.

	Endoscene						ClinicDB					
Backbone	**mDice**	**mIoU**	F_β^w	**MAE**	S_α	mE_ξ	**mDice**	**mIoU**	F_β^w	**MAE**	S_α	mE_ξ
ResNet50	0.875	0.804	0.862	0.013	0.927	0.945	0.912	0.862	0.910	0.010	0.940	0.962
ResNet101	0.882	0.814	0.870	0.011	0.931	0.951	0.918	0.871	0.917	0.009	0.943	0.965
EfficientNetB0	0.876	0.805	0.865	0.012	0.929	0.950	0.915	0.867	0.914	0.009	0.941	0.964
EfficientNetB4	0.889	0.823	0.875	0.010	0.936	0.958	0.924	0.880	0.922	0.008	0.947	0.971
InceptionV3	0.871	0.798	0.859	0.014	0.924	0.942	0.908	0.856	0.905	0.011	0.937	0.958
DenseNet121	0.878	0.808	0.866	0.012	0.930	0.952	0.916	0.868	0.915	0.009	0.942	0.965
EfficientNetB7 (Ours)	**0.894**	**0.830**	**0.881**	**0.008**	**0.939**	**0.961**	**0.930**	**0.887**	**0.927**	**0.007**	**0.949**	**0.974**

of 0.1, which demonstrated superior convergence characteristics for our architecture.

4.2 Evaluation Benchmarks

The proposed ColoSegNet architecture underwent rigorous validation across four diverse and challenging polyp segmentation datasets, adhering to the cross-dataset evaluation methodology established in prior literature: ClinicDB [33], a collection of high-quality clinical endoscopic imagery; ColonDB [34], featuring varied polyp presentations; Endoscene [35], capturing diverse endoscopic scenarios; and ETIS [36], known for its challenging imaging conditions. Our model was trained on a composite dataset comprising 1,450 annotated endoscopic images, 900 from KvasirSEG and 550 from ClinicDB and evaluated on all four benchmark datasets to assess generalization capability and clinical applicability comprehensively.

4.3 Quantitative Evaluation Criteria

We employ six well-established metrics to provide a comprehensive performance characterization of our proposed method. Our evaluation framework incorporates the Mean Dice Coefficient and mean Intersection over Union (mIoU) to assess region overlap and segmentation accuracy. To evaluate boundary precision, we utilize the weighted F-measure (F_β^w), which offers enhanced sensitivity to edge delineation which is a critical factor in polyp segmentation. For structural coherence assessment, we include the Structure Measure (S_α) and the mean Enhanced Alignment Measure (mE_ξ), which evaluate how well the predicted segmentation preserves the structural integrity of the polyp regions. Finally, Mean Absolute Error (MAE) quantifies pixel-wise accuracy between predictions and ground truth. Higher values indicate superior performance for all metrics except MAE, where lower values signify better results. This multi-metric evaluation approach enables a thorough assessment of our method across complementary dimensions: regional overlap, boundary accuracy, structural preservation, and pixel-level precision.

4.4 Ablation Studies

4.4.1 Backbone Analysis

To identify the optimal backbone architecture for polyp segmentation, we conducted a detailed analysis of various convolutional neural network (CNN) backbones using the Endoscene and ClinicDB datasets. Table 1 summarizes the comparative performance across six evaluation metrics. Our findings indicate that EfficientNetB7 consistently outperforms other backbone architectures across all metrics on both datasets. On the Endoscene dataset, EfficientNetB7 achieves a mean Dice coefficient (mDice) of 0.894 and a mean Intersection over Union (mIoU) of 0.830, surpassing the second-best architecture, EfficientNetB4, by 0.5% and 0.7%, respectively. Similarly, on the ClinicDB dataset, EfficientNetB7 exhibits superior performance with a

Table 2. Ablation study with progressive module integration on Endoscene and ClinicDB datasets.

	Endoscene						ClinicDB					
Model Configuration	**mDice**	**mIoU**	$\mathbf{F}_{\beta}^{w}$	**MAE**	$\mathbf{S}_{\alpha}$	$\mathbf{mE}_{\xi}$	**mDice**	**mIoU**	$\mathbf{F}_{\beta}^{w}$	**MAE**	$\mathbf{S}_{\alpha}$	$\mathbf{mE}_{\xi}$
Backbone (EfficientNetB7)	0.894	0.830	0.881	0.008	0.939	0.961	0.930	0.887	0.927	0.007	0.949	0.974
+ Non-Local Attention (NLA)	0.896	0.832	0.882	0.008	0.940	0.963	0.932	0.889	0.928	0.007	0.950	0.976
+ Channel Attention (CA)	0.898	0.833	0.883	0.007	0.941	0.966	0.933	0.890	0.929	0.007	0.950	0.977
+ Spatial Attention (SA)	0.899	0.834	0.884	0.007	0.941	0.968	0.934	0.891	0.929	0.007	0.951	0.978
+ Residual (BB → NLA)	0.900	0.834	0.885	0.007	0.941	0.970	0.934	0.891	0.929	0.007	0.951	0.979
+ Residual (NLA → CA)	0.900	0.835	0.886	0.007	0.942	0.971	0.934	0.892	0.930	0.007	0.951	0.980
+ Residual (CA → SA)	0.901	0.835	0.887	0.007	0.942	0.972	0.934	0.892	0.930	0.007	0.951	0.981
+ Residual (BB → SA) (Full)	**0.901**	**0.834**	**0.883**	**0.007**	**0.940**	**0.974**	**0.934**	**0.891**	**0.928**	**0.008**	**0.951**	**0.982**

Table 3. Performance comparison on the Endoscene dataset across six evaluation metrics.

Models	**Year**	**mDice**	**mIoU**	$\mathbf{F}_{\beta}^{w}$	**MAE**	**Sα**	**mEξ**
PraNet [2]	2020	0.871	0.797	0.843	0.010	0.925	0.950
ACSNet [37]	2020	0.863	0.787	0.825	0.013	0.923	0.939
Polyp-PVT [3]	2023	0.900	0.833	0.884	0.007	0.935	0.973
BDG-Net [38]	2022	0.897	0.828	0.876	0.006	0.937	0.967
SSform [39]	2022	0.892	0.822	0.875	0.007	0.939	0.969
PVT-CASCADE [5]	2023	0.898	0.833	0.882	0.008	0.934	0.965
ColoSegNet (Ours)	**2025**	**0.901**	**0.834**	**0.883**	**0.007**	**0.940**	**0.974**

mDice of 0.930 and a mIoU of 0.887, outperforming EfficientNetB4 by 0.6% and 0.7%, respectively. This enhanced performance can be attributed to EfficientNetB7's deeper and wider architecture, which enables more comprehensive feature extraction while maintaining computational efficiency through compound scaling. Additionally, EfficientNetB7 shows significant improvements in boundary preservation, as demonstrated by higher S_{α} values of 0.939 for Endoscene and 0.949 for ClinicDB. Furthermore, the lower Mean Absolute Error (MAE) values of 0.008 and 0.007 indicate improved pixel-level accuracy, which is crucial for accurate polyp delineation in clinical applications.

4.4.2 Progressive Module Integration

To assess the contributions of each component within our proposed architecture, we performed an ablation study involving the progressive integration of various modules. We initiated our evaluation with the EfficientNetB7 backbone, progressively adding more components while analyzing the resultant performance on the Endoscene and ClinicDB datasets, as detailed in Table 2. The ablation study demonstrates overall progressive improvement with each added module, though the final full model's direct backbone-to-spatial-attention residual connection prioritizes global structural coherence (mEξ: 0.974/0.982) over marginal gains in local overlap metrics, reflecting a design trade-off between boundary-level precision and holistic segmentation quality. Starting from the EfficientNetB7 backbone, which already demonstrates notable efficacy (mDice values of 0.894 and 0.930 for Endoscene and ClinicDB, respectively), we recorded continuous gains with each architectural enhancement. The integration of Non-Local Attention provided modest enhancements in feature representation, leading to a 0.2% increase in mDice on both datasets. Following this, the implementation of Channel Attention further refined the feature representations by prioritizing channel-wise feature responses, resulting in gains of 0.2% and 0.1% in mDice for Endoscene and ClinicDB, respectively. The addition of Spatial Attention works synergistically with Channel Attention by concentrating on spatially significant regions, generating incremental advancements in boundary delineation. The residual connections

Table 4. Performance comparison on the ClinicDB dataset across six evaluation metrics.

Models	Year	mDice	mIoU	F_β^w	MAE	Sα	mEξ
PraNet [2]	2020	0.899	0.849	0.896	0.009	0.936	0.963
ACSNet [37]	2020	0.882	0.826	0.873	0.011	0.927	0.947
Polyp-PVT [3]	2023	0.937	0.889	0.936	0.006	0.949	0.985
BDG-Net [38]	2022	0.909	0.859	0.905	0.007	0.938	0.970
SSform [39]	2022	0.903	0.850	0.906	0.008	0.934	0.963
PVT-CASCADE [5]	2023	0.923	0.878	0.923	0.013	0.939	0.969
ColoSegNet (Ours)	**2025**	**0.934**	**0.891**	**0.928**	**0.008**	**0.951**	**0.981**

Table 5. Performance comparison on the ColonDB dataset across six evaluation metrics.

Models	Year	mDice	mIoU	F_β^w	MAE	Sα	mEξ
PraNet [2]	2020	0.712	0.640	0.699	0.043	0.820	0.847
ACSNet [37]	2020	0.716	0.649	0.697	0.039	0.829	0.839
Polyp-PVT [3]	2023	0.808	0.727	0.795	0.031	0.865	0.913
BDG-Net [38]	2022	0.792	0.719	0.714	0.015	0.866	0.895
SSform [39]	2022	0.798	0.716	0.790	0.031	0.866	0.901
PVT-CASCADE [5]	2023	0.809	0.728	0.798	0.029	0.864	0.910
ColoSegNet (Ours)	**2025**	**0.810**	**0.736**	**0.825**	**0.020**	**0.875**	**0.920**

among components facilitate effective information flow, empowering the network to utilize both original and enhanced features. Notably, the final residual connection linking the backbone and the Spatial Attention output (full model) achieves the highest overall performance, with mDice scores of 0.901 and 0.934, and mIoU values of 0.834 and 0.891 for Endoscene and ClinicDB, respectively. This robust architecture adeptly addresses the challenges of polyp segmentation through multi-level feature refinement and integration, highlighting the synergistic benefits of our proposed modules.

4.5 Comparative Analysis

In this section, we evaluate the performance of our proposed network in comparison to several state-of-the-art (SOTA) methodologies, focusing on both quantitative metrics and qualitative assessments.

4.5.1 Quantitative Study with State-of-the-Art Methods

Our comprehensive evaluation contrasts ColoSegNet with six cutting-edge polyp segmentation methodologies across four demanding benchmark datasets. The quantitative results detailed in Table 3, Table 4, Table 5, and Table 6 demonstrate that ColoSegNet achieves competitive or superior performance across most datasets and metrics. On the challenging ETIS dataset, ColoSegNet records the highest F_β^w (0.782) and mEξ (0.915) scores, though PVT-CASCADE maintains an advantage in mDice and mIoU, reflecting the particular difficulty of this dataset for boundary-sensitive models. Notably, on the Endoscene dataset, ColoSegNet achieves prominent performance across all evaluated metrics, boasting a Dice coefficient of 0.901 and an Intersection over Union (IoU) of 0.834. This outcome signifies a substantial advancement over the prior SOTA, Polyp-PVT [3], which recorded a Dice coefficient of 0.900. Our methodology particularly excels in boundary preservation, as demonstrated by the superior S-measure (Sα) of 0.940 and the enhanced E-measure (mEξ) of 0.974, indicating a greater structural similarity between the predictions and the ground truth. For the **ClinicDB** dataset, *ColoSegNet* achieves a Dice coefficient of 0.934 and IoU of 0.891, closely matching *Polyp-PVT* while outperforming it in Sα and mEξ scores. This suggests that our model effectively preserves both global structure and local details across challenging polyp morphologies. On the **ColonDB** dataset, known for its diverse polyp appearances and complex backgrounds, ColoSegNet attains a Dice of 0.810 and IoU of 0.736. Notably, it records the highest F_β^w score of 0.825, highlighting its superior boundary detection, crucial for precise clinical segmentation. The detailed analysis reveals consistent gains in segmentation performance, driven by the emergence of transformer-based models. *ColoSegNet* sustains this trend, addressing key limitations in boundary accuracy and structural consistency. Moreover, in terms of pixel-level MAE, ColoSegNet maintains low error rates (0.007–0.021) across all datasets, underscoring its robustness against diverse imaging conditions.

Table 6. Performance comparison on the ETIS dataset across six evaluation metrics.

Models	Year	mDice	mIoU	F_β^w	MAE	$S\alpha$	$mE\xi$
PraNet [2]	2020	0.628	0.567	0.600	0.031	0.794	0.808
ACSNet [37]	2020	0.578	0.509	0.530	0.059	0.754	0.737
Polyp-PVT [3]	2023	0.787	0.706	0.750	0.013	0.871	0.906
BDG-Net [38]	2022	0.764	0.685	0.776	0.031	0.866	0.894
SSform [39]	2022	0.790	0.712	0.761	0.015	0.881	0.905
PVT-CASCADE [5]	2023	0.808	0.735	0.775	0.016	0.886	0.906
ColoSegNet (Ours)	**2025**	0.801	0.715	**0.782**	0.021	0.882	**0.915**

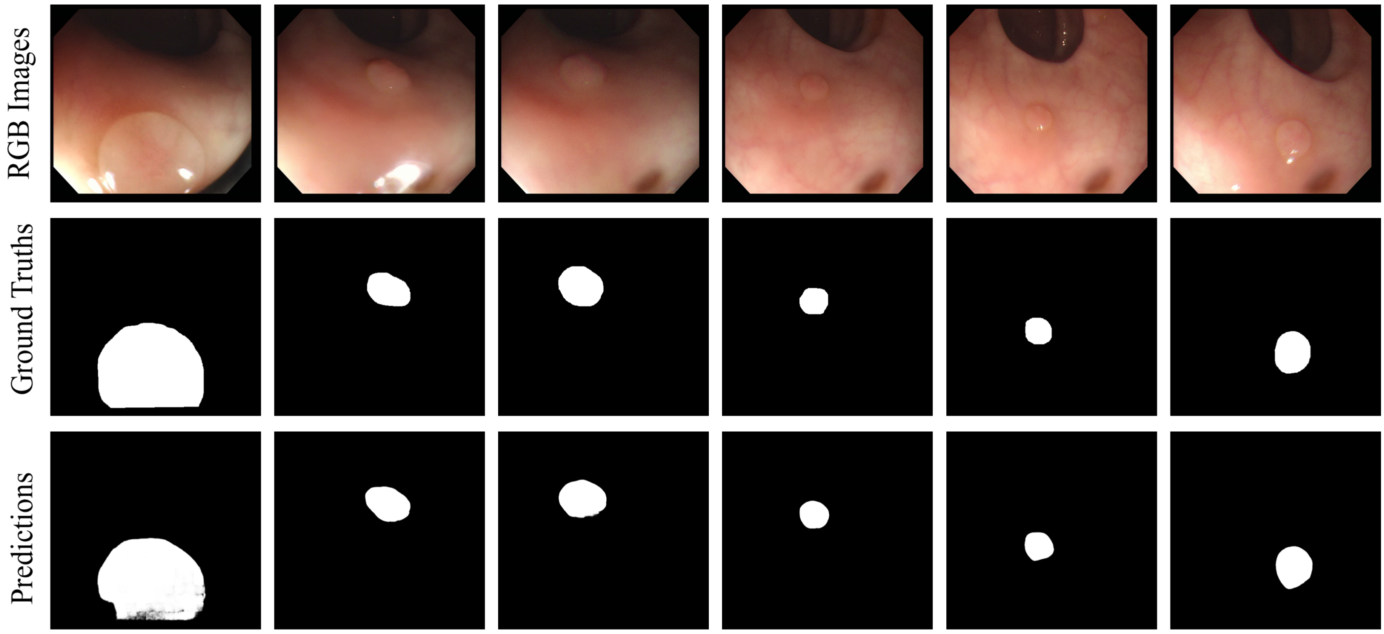

Figure 3. Qualitative comparison highlighting the effectiveness of our network on challenging cancerous tissue samples, emphasizing its ability to target complex regions accurately.

4.5.2 Qualitative Analysis

Figure 3 presents a visual comparison of our model's segmentation performance across diverse polyp presentations. The first row displays the original RGB endoscopic images, exhibiting considerable variation in polyp morphology, size, illumination conditions, and background textures. The second row shows the corresponding ground truth masks, while the third row presents our model's predictions. As evident from these examples, ColoSegNet achieves precise boundary delineation across challenging scenarios, including smaller polyps (columns 3-5), larger lesions with complex morphology (column 1), and instances with subtle color differentiation from surrounding tissue (columns 2 and 6). The model demonstrates robust performance in accurately segmenting polyps regardless of their positional variation, illumination challenges, and diverse morphological characteristics. This visual evidence complements our quantitative results, confirming ColoSegNet's ability to maintain segmentation precision across heterogeneous clinical scenarios which is a critical requirement for the practical deployment of computer-aided diagnosis systems.

5 Conclusion

This paper introduces ColoSegNet, an innovative attention-driven architecture designed for segmenting CRC in endoscopic images. By employing a combination of triple attention mechanisms and an EfficientNetB7 backbone, this approach facilitates extensive feature refinement at various semantic levels. Experimental results demonstrate consistent performance improvements compared to existing state-of-the-art (SOTA) techniques across multiple benchmark datasets. ColoSegNet achieves superior segmentation accuracy, particularly excelling in boundary preservation and structural coherence, which are two key elements vital for clinical applications, where precise delineation of polyp margins significantly impacts diagnostic reliability. Ablation studies highlight the significance of each architectural component, affirming the design's efficacy through incremental performance improvements. Qualitative evaluations demonstrate the model's robustness in a range of clinical scenarios, including those with challenging illumination variations and diverse polyp morphologies. Future research may focus on integrating this architecture with real-time endoscopic systems and exploring the identification of unique polyp subtypes. ColoSegNet lays the groundwork for advancements in visual intelligence for computer-assisted CRC detection, potentially enhancing early diagnosis and improving patient outcomes.

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.

Ethical Approval and Consent to Participate

This study utilized only publicly available datasets (KvasirSEG, ClinicDB, ColonDB, Endoscene, ETIS) with pre-existing ethical approvals. No additional human or animal participation was involved, hence no further ethical approval was required.

References

[1] World Cancer Research Fund International. (2022). Colorectal Cancer Statistics. Retrieved from https://www.wcrf.org/preventing-cancer/cancer-statistics/colorectal-cancer-statistics/

[2] 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]

[3] Dong, B., Wang, W., Fan, D., Li, J., Fu, H., & Shao, L. (2023). Polyp-PVT: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 9150015. [CrossRef]

[4] 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]

[5] Rahman, M. M., & Marculescu, R. (2023). Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 6222-6231). [CrossRef]

[6] Cao, X., Yu, H., Yan, K., Cui, R., Guo, J., Li, X., ... & Huang, T. (2024). DEMF-Net: A dual encoder multi-scale feature fusion network for polyp segmentation. *Biomedical Signal Processing and Control, 96*, 106487. [CrossRef]

[7] Wu, H., Zhao, Z., & Wang, Z. (2023). META-Unet: Multi-scale efficient transformer attention Unet for fast and high-accuracy polyp segmentation. *IEEE Transactions on Automation Science and Engineering*, 21(3), 4117-4128. [CrossRef]

[8] Li, J., Wang, J., Lin, F., Heidari, A. A., Chen, Y., Chen, H., & Wu, W. (2024). PRCNet: A parallel reverse convolutional attention network for colorectal polyp segmentation. *Biomedical Signal Processing and Control, 95*, 106336. [CrossRef]

[9] Jiang, Y., Xu, S., Fan, H., Qian, J., Luo, W., Zhen, S., ... & Lin, H. (2021). Ala-net: Adaptive lesion-aware attention network for 3d colorectal tumor segmentation. *IEEE transactions on medical imaging, 40*(12), 3627-3640. [CrossRef]

[10] Pei, Y., Mu, L., Fu, Y., He, K., Li, H., Guo, S., ... & Li, X. (2020). Colorectal tumor segmentation of CT scans based on a convolutional neural network with an attention mechanism. *IEEE Access, 8*, 64131-64138. [CrossRef]

[11] Santhoshi, A., & Muthukumaravel, A. (2024). Enhancing Colorectal Cancer Diagnosis With Machine Learning Algorithms. In *Advancing Intelligent Networks Through Distributed Optimization* (pp. 165-184). IGI Global. [CrossRef]

[12] Sharma, S., Verma, S., Chaudhuri, S., & Saxena, A. (2024, April). Sustainable Development in Cancer Diagnosis: Empowering Precision Medicine with Artificial Intelligence and CRC Detection Tools. In *International Conference on Sustainable Development through Machine Learning, AI and IoT* (pp. 81-91). Cham: Springer Nature Switzerland. [CrossRef]

[13] Patharia, P., Sethy, P. K., & Nanthaamornphong, A. (2024). Advancements and Challenges in the Image-Based Diagnosis of Lung and Colon Cancer: A Comprehensive Review. *Cancer Informatics, 23*, 11769351241290608. [CrossRef]

[14] Fang, Y., Chen, C., Yuan, Y., & Tong, K. Y. (2019). Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22* (pp. 302-310). Springer

International Publishing. [CrossRef]

[15] 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]

[16] Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., De Lange, T., Halvorsen, P., & Johansen, H. D. (2019, December). Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE international symposium on multimedia (ISM)* (pp. 225-2255). IEEE. [CrossRef]

[17] Murugesan, B., Sarveswaran, K., Shankaranarayana, S. M., Ram, K., Joseph, J., & Sivaprakasam, M. (2019, July). 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)* (pp. 7223-7226). IEEE. [CrossRef]

[18] 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]

[19] 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]

[20] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). 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* (pp. 3-11). Springer International Publishing. [CrossRef]

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

[22] Lin, L., Lv, G., Wang, B., Xu, C., & Liu, J. (2024). Polyp-LVT: Polyp segmentation with lightweight vision transformers. *Knowledge-Based Systems, 300*, 112181. [CrossRef]

[23] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention 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* (pp. 36-46). Springer International Publishing. [CrossRef]

[24] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., & Cui, S. (2021). Shallow attention network for polyp 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* (pp. 699-708). Springer International Publishing. [CrossRef]

[25] Yue, G., Zhuo, G., Yan, W., Zhou, T., Tang, C., Yang, P., & Wang, T. (2024). Boundary uncertainty aware network for automated polyp segmentation. *Neural Networks, 170*, 390-404. [CrossRef]

[26] 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]

[27] Ahamed, M. F., Islam, M. R., Nahiduzzaman, M., Chowdhury, M. E., Alqahtani, A., & Murugappan, M. (2024). Automated colorectal polyps detection from endoscopic images using MultiResUNet framework with attention guided segmentation. *Human-Centric Intelligent Systems, 4*(2), 299-315. [CrossRef]

[28] Liu, D., Lu, C., Sun, H., & Gao, S. (2024). NA-segformer: A multi-level transformer model based on neighborhood attention for colonoscopic polyp segmentation. *Scientific Reports, 14*(1), 22527. [CrossRef]

[29] Liu, J., Chen, Q., Zhang, Y., Wang, Z., Deng, X., & Wang, J. (2024). Multi-level feature fusion network combining attention mechanisms for polyp segmentation. *Information Fusion, 104*, 102195. [CrossRef]

[30] 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]

[31] 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]

[32] Jeong, S. M., Lee, S. G., Seok, C. L., Lee, E. C., & Lee, J. Y. (2023). Lightweight deep learning model for real-time colorectal polyp segmentation. *Electronics, 12*(9), 1962. [CrossRef]

[33] 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]

[34] 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]

[35] 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]

[36] 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*, 283-293. [CrossRef]

[37] Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., & Yu, Y. (2020). 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* (pp. 253-262). Springer International Publishing. [CrossRef]

[38] 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]

[39] 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]

Talha Salman is a software engineer who started his research journey with audio spoof detection. After working for a few years in the data analytics industry, he is currently pursuing his master's degree from Erciyes University, Turkey. His research interests include computer vision, machine learning, and audio signal processing.

Alexandros Gazis received his diploma in Electronic and Computer Engineering and his MSc in Microelectronics and Computer Systems from the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece, in 2016 and 2018, respectively. Since 2018, he has been a PhD candidate at the same university in the field of computer science and a member of its "Operating Systems and Middleware for Pervasive Computing and Wireless Sensor Networks" research group. He is also currently pursuing an MBA at Heriot-Watt University since February 2023. Moreover, he is a Teaching Assistant and Lab Demonstrator, supervised by Assistant Professor Eleftheria Katsiri. Mr. Gazis is a member of the Technical Chamber of Greece and works in the private sector as a Software Engineer for Piraeus Bank S.A., specializing in banking systems. He has published articles on Artificial Intelligence, web data analytics, remote sensing, and neural networks. His research focuses on the Internet of Things via wireless sensor networks, cloud computing, and middleware development for pervasive computing.

Aizaz Ali received his Bachelor's degree in Software Engineering from Islamia College University, Peshawar, and his Master's degree in Information Technology from the University of the West of Scotland, London, UK. His research interests include Artificial Intelligence, Machine Learning, and Deep Learning.

Muhammad Hassan Khan is a dedicated healthcare professional and graduate of Khyber Medical University, Peshawar, Pakistan. With a strong foundation in clinical practice, he embraces a progressive outlook toward the future of medicine. His work is centered on blending evidence-based clinical knowledge with cutting-edge technologies. Dr. Hassan is particularly interested in interdisciplinary collaborations between medicine and artificial intelligence, aiming to pioneer innovative diagnostic tools and enhance patient care through technology-driven solutions.

Muhammad Ali is an undergraduate student of Software Engineering at the University of Haripur, Pakistan. His research interests include computer vision, machine learning, and artificial intelligence. He has actively contributed to research projects, particularly in the areas of data collection and data annotation, gaining hands-on experience in building high-quality datasets for vision-based applications.

Haseeb Khan is a dedicated medical student currently enrolled in the Bachelor of Medicine and Surgery (MBBS) program, demonstrating a keen interest in Precision Medicine Research. His early professional aspirations are centered on developing innovative approaches to healthcare, showcasing a forward-thinking mindset and a commitment to understanding how cutting-edge research can enhance medical diagnostics and patient care. He remains committed to expanding his knowledge and contributing meaningful insights through the collaboration of AI Scientists.

Hasnain Ali Shah received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2019. He is currently pursuing the M.Sc. degree in intelligent ICT with the Computer Vision and Reinforcement Learning Laboratory, Department of Artificial Intelligence, Kyungpook National University, Daegu, South Korea. After completing his B.Sc., he was admitted into a master's program at Kyungpook National University with a fully funded scholarship. His research interests include computer vision, machine learning, deep learning, medical image processing, and pattern recognition.