DFANet: Dual-Branch Feature Fusion with Multi-Attention Decoding for Colonoscopy Polyp Segmentation

Kokkanti Srinija
Department of AIML
CBIT, Gandipet
Hyderabad, India
srinijakr@gmail.com

Mamidi Harshini
Department of AIML
CBIT, Gandipet
Hyderabad, India
mamidiharshini09@gmail.com

Panigrahi Srikanth
Department of AIML
CBIT, Gandipet
Hyderabad, India
srikanth.panigrahi@gmail.com

***Abstract*—Accurate colonoscopic polyp segmentation is essential for early colorectal cancer detection, yet remains challenging due to low contrast, specular highlights, blurred boundaries, and large variations in polyp appearance. We propose DFANet, a dual-branch feature aggregation framework that combines transformer-driven global semantic modeling MiT-B5 with CNN-based fine-grained representation learning EfficientNet-B5. A channel–spatial fusion module adaptively integrates multi-scale encoder outputs through element-wise aggregation, channel attention, and spatial attention, followed by a DSC Transformer that enhances contextual reasoning before decoding. The multi-stage decoder progressively restores resolution while refining structural details to produce accurate segmentation masks. Extensive experiments on five benchmark datasets demonstrate consistent performance gains, with DFANet achieving improvements of up to 3–7% in Dice and mIoU over state-of-the-art methods on Kvasir-SEG, CVC-ClinicDB, BKAI-IGH, ETIS-LaribPolypDB, and CVC-ColonDB.**

***Index Terms*—Medical image segmentation, Dual Attention,Depthwise Convolution Transformer.**

I. Introduction

Colorectal cancer (CRC) is a leading cause of cancer-related deaths, with over 1.9 million cases and 930,000 deaths reported in 2020 [1]. Early polyp detection is crucial, yet colonoscopy remains operator-dependent and prone to missing small or low-contrast lesions [2], motivating deep learning–based CAD systems for reliable segmentation. However, CNN-based models lack long-range context, transformers are computationally expensive, and hybrid methods often remain complex. To address this, we propose **DFANet**, a dual-branch framework combining EfficientNet-B5 for spatial features and MiT-B5 for global semantics, with cross-branch channel–spatial fusion and a multi-attention DualDecoder (AG, ASPP, STA) for accurate, boundary-aware segmentation. DFANet achieves strong performance across multiple colonoscopy datasets with good generalizability and efficiency.

The main contributions of this work are summarized as follows:

- We present DFANet, a dual-branch polyp segmentation framework that integrates CNN-based local feature extraction with transformer-based global semantic reasoning, enabling robust handling of low contrast, irregular polyp structures, and complex colonoscopic backgrounds.
- We introduce a cross-branch channel–spatial fusion mechanism that aligns and integrates multi-scale EfficientNet-B5 and MiT-B5 features, together with a multi-attention DualDecoder that incorporates Attention Gates (AG), Atrous Spatial Pyramid Pooling (ASPP), and Spatial–Temporal Attention (STA) to refine skip connections, expand multi-scale receptive fields, and improve boundary localization.
- Extensive experiments on five benchmark datasets—Kvasir-SEG, CVC-ClinicDB, BKAI-IGH, ETIS-LaribPolypDB, and CVC-ColonDB—demonstrate that DFANet achieves consistent state-of-the-art performance, surpassing recent CNN-, transformer-, and hybrid-based models across Dice, mIoU, recall, and precision.

II. Literature Survey

Polyp segmentation has seen substantial progress with deep learning, evolving from CNN-based architectures to transformer-driven and hybrid designs. Early approaches such as U-Net [3], UNet++ [4], ResUNet++ [5], HarDNet-MSEG [6], and DoubleU-Net [7] effectively captured local spatial details and multi-scale features through encoder–decoder structures, residual connections, and attention mechanisms. However, purely convolutional models often struggle to capture long-range dependencies and global contextual information in complex colonoscopic scenes. To address these limitations, transformer-based models including TransResU-Net [8], Swin-UNet [9], HiFiSeg [10], and vision-language guided approaches introduced global receptive fields and enhanced semantic reasoning. Despite improving boundary delineation and contextual modeling, transformer-only frameworks frequently suffer from high computational overhead and limited generalization across diverse datasets.

Hybrid CNN-transformer architectures have improved polyp segmentation through complementary feature fusion and multi-scale representation learning [11]–[14].

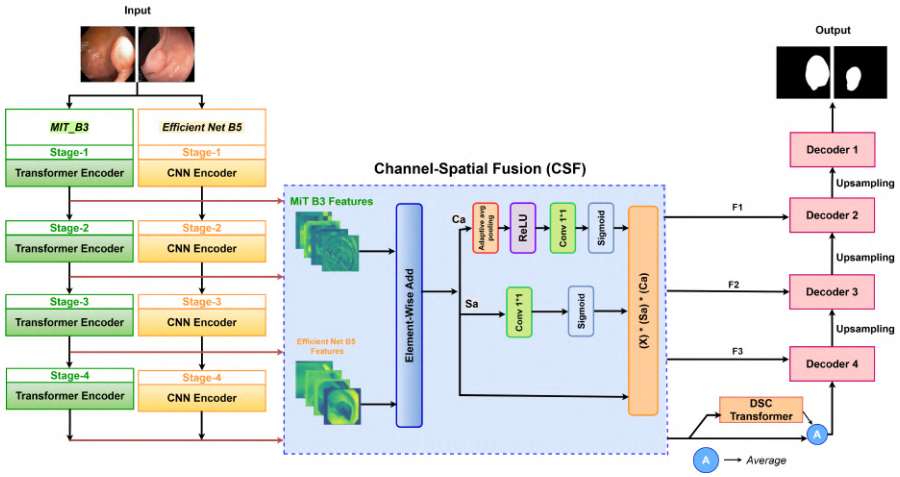

Fig. 1. Overall Architecture of **DFANet**

Research Gap: Achieving accurate segmentation while maintaining computational efficiency remains challenging. DFANet addresses this via efficient dual-branch fusion and lightweight attention-guided decoding.

III. Proposed Method

This section presents the proposed **DFANet** (*Dual-Branch Feature Fusion with Multi-Attention Decoding*) framework, a hybrid segmentation architecture that integrates a convolutional encoder (EfficientNet-B5) and a transformer encoder (MiT-B5). The model follows a dual-branch encoder, cross-branch fusion, and multi-attention decoding strategy. The overall pipeline is shown in Fig. 1.

A. Architecture Overview

DFANet adopts an encoder–fusion–decoder design where two heterogeneous encoders extract complementary feature representations. The CNN branch specializes in capturing fine-grained texture and boundary information, while the transformer branch models global dependencies and semantic structure.

Given an input image $X \in \mathbb{R}^{1\times H\times W}$, the two encoders generate four hierarchical feature levels:

$$\{E_0, E_1, E_2, E_3\} \quad \text{(EfficientNet-B5 branch)},$$

$$\{M_0, M_1, M_2, M_3\} \quad \text{(MiT-B5 branch)}.$$

These multi-scale representations are fused and passed to the multi-attention decoder for final mask prediction.

B. Dual Branch Feature Extraction

1) Convolutional Branch (EfficientNet-B5): The convolutional encoder extracts local spatial cues and fine texture patterns. EfficientNet-B5 offers strong representational capacity through compound scaling of depth, width, and resolution.

2) Transformer Branch (MiT-B5): The MiT-B5 transformer encoder from SegFormer captures global semantic dependencies and long-range spatial structure. Its hierarchical outputs (M_0–M_3) provide large-receptive-field representations complementary to the CNN features.

C. DSC Transformer Refinement Module

A DSC Transformer is employed after feature fusion to enhance contextual reasoning. It combines depthwise separable convolutions for local feature extraction with lightweight self-attention for global dependency modeling. The refined features are subsequently passed to the multi-attention decoder for segmentation.

D. Multi-Attention DualDecoder

The decoder consists of three stacked DualDecoder blocks, each containing the following:

1) Attention Gate (AG): Given decoder input g and skip feature x, AG generates an attention coefficient:

$$\alpha = \sigma\big(\psi(\mathrm{ReLU}(W_g g + W_x x))\big), \qquad \hat{x} = x \odot \alpha.$$

This suppresses background noise and preserves lesion-relevant activations.

E. Multi-Attention DualDecoder

The decoder consists of three stacked DualDecoder blocks, each containing:

- an Attention Gate (AG) for skip refinement,
- an Atrous Spatial Pyramid Pooling (ASPP) module for multi-scale context,
- a Spatial–Temporal Attention (STA) unit for boundary enhancement.

1) Attention Gate (AG): This suppresses background noise and preserves lesion-relevant activations.

2) Atrous Spatial Pyramid Pooling (ASPP): ASPP aggregates multi-scale context through dilated convolutions. This broadens receptive fields and improves robustness to varying polyp size and shape.

TABLE I
PERFORMANCE COMPARISON OF CNN AND TRANSFORMER BACKBONES

Model	P.	R.	F1	IoU	D.	Acc.
ResNet34 [15]	0.81	0.88	0.84	0.73	0.84	0.94
ResNet50 [15]	0.89	0.89	0.89	0.80	0.89	0.96
DenseNet121 [16]	0.89	0.87	0.88	0.80	0.88	0.96
EffNet-B0 [17]	0.88	0.88	0.88	0.78	0.88	0.96
EffNet-B1 [17]	0.88	0.88	0.88	0.79	0.88	0.97
EffNet-B2 [17]	0.88	0.89	0.89	0.80	0.89	0.96
EffNet-B3 [17]	0.87	0.91	0.89	0.80	0.89	0.96
EffNet-B4 [17]	0.88	0.90	0.90	0.81	0.90	0.96
EffNet-B5 [17]	0.89	0.91	0.87	0.78	0.87	0.96
VGG19-BN [18]	0.88	0.88	0.88	0.79	0.88	0.96
MobileNetV2 [19]	0.85	0.90	0.87	0.78	0.87	0.96
MiT-B0 [20]	0.87	0.88	0.87	0.78	0.87	0.96
MiT-B1 [20]	0.90	0.86	0.88	0.80	0.88	0.96
MiT-B2 [20]	0.89	0.91	0.89	0.81	0.89	0.96
MiT-B3 [20]	0.90	0.91	0.90	0.83	0.90	0.97
MiT-B4 [20]	0.84	0.94	0.88	0.80	0.88	0.96
MiT-B5 [20]	0.89	0.92	0.91	0.82	0.91	0.97
IncResNetV2 [21]	0.85	0.87	0.86	0.76	0.86	0.95
InceptionV4 [21]	0.82	0.90	0.85	0.75	0.85	0.95
Swin-T [22]	0.88	0.88	0.88	0.79	0.88	0.97
Swin-B [23]	0.87	0.87	0.87	0.78	0.87	0.97
PVTv2-B0 [24]	0.88	0.88	0.88	0.78	0.88	0.97
PVTv2-B2 [24]	0.89	0.87	0.87	0.78	0.87	0.97
PVTv2-B3 [24]	0.87	0.90	0.88	0.79	0.88	0.97
ViT-Base [25]	0.83	0.90	0.87	0.77	0.87	0.97
ViT-Small [25]	0.88	0.89	0.89	0.80	0.89	0.97

3) Spatial–Temporal Attention (STA): STA enhances salient spatial regions . This improves boundary accuracy and reduces artifacts.

F. Training Setup

DFANet is trained end-to-end using a composite hybrid loss:

$$L = 0.3L_{BCE} + 0.3L_{Dice} + 0.4L_{Tversky}, \tag{1}$$

where Tversky Loss is employed to address class imbalance between foreground and background regions.

IV. EXPERIMENTAL RESULTS

All experiments were implemented in PyTorch and conducted on Google Colab using an NVIDIA Tesla T4 GPU. The model was trained for 100 epochs. Data augmentation included random flipping, rotation, and scaling. AdamW was employed with an initial learning rate of 1×10^{-4} and a ReduceLROnPlateau scheduler. The proposed DFANet has been evaluated on three widely used colonoscopic polyp segmentation datasets: Kvasir-SEG, CVC-ClinicDB, and BKAI-IGH NeoPolyp-Small. Each dataset was treated as an independent benchmark to assess the model's generalization capability across variations in polyp size, morphology, imaging quality, and acquisition modalities. Kvasir-SEG [38] contains 1,000 annotated frames, CVC-ClinicDB [39] provides 612 high-quality endoscopic images, and BKAI-IGH [40] includes 1,200 high-resolution images acquired under both White Light Imaging (WLI) and Flexible Imaging Color Enhancement (FICE). DFANet was optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a `ReduceLROnPlateau` scheduler that reduces the learning rate by a factor of 0.5 after three epochs without validation improvement. To improve both pixel-wise accuracy and boundary consistency, we employed a composite loss function comprising Binary Cross-Entropy, Dice Loss, and Twerky Loss with respective weights of 0.3, 0.3, and 0.4, enabling robust performance across diverse imaging conditions and polyp morphologies.

Table I presents a detailed comparison of CNN and transformer backbones integrated into a U-Net segmentation framework. Among CNN-based encoders, the EfficientNet family consistently outperforms traditional convolutional networks such as ResNet34, ResNet50, DenseNet121, and MobileNetV2, with EfficientNet-B3 and B4 delivering the strongest results (Dice = 0.89 and 0.90; IoU = 0.80 and 0.81). In contrast, EfficientNet-B5 shows slightly lower performance (Dice = 0.87, IoU = 0.78), indicating that larger variants do not necessarily translate to better segmentation accuracy. Among Transformer-based backbones , MiT-B3 (SegFormer) achieves the most balanced and robust performance, attaining a Dice score of 0.90 and the highest IoU of 0.83, while MiT-B5 performs competitively with a Dice of 0.91 but exhibits a slightly lower IoU (0.82) and higher computational cost. Despite this, DFANet adopts MiT-B5 as its transformer branch because its richer hierarchical feature representation and lightweight multi-scale attention align more effectively with the proposed dual-branch fusion strategy, enabling superior cross-branch complementarity and maintaining strong transformer-level accuracy.

A. Performance Analysis of DFANet Across Multiple Colonoscopic Imaging Datasets

Table II presents a comprehensive quantitative comparison of DFANet against state-of-the-art CNN- and transformer-based segmentation models across five benchmark datasets. DFANet demonstrates consistently strong performance, achieving an mIoU of 0.8191 and DSC of 0.9006 on Kvasir-SEG, outperforming classical methods such as U-Net, HarDNet-MSEG, and TransNetR. On CVC-ClinicDB, DFANet attains 0.8731 mIoU and 0.9322 DSC, surpassing leading architectures including UACANet and DeepLab variants. The model further achieves 0.8510 mIoU and 0.9195 DSC on the challenging BKAI-IGH dataset, exceeding competitors like PraNet, CaraNet, and TGANet. DFANet also shows robust performance on ETIS-LaribPolypDB with 0.8104 mIoU and 0.8953 DSC, improving over SANet and G-Cascade. Finally, on CVC-ColonDB, DFANet obtains 0.7917 mIoU and 0.8837 DSC, consistently outperforming U-Net, PraNet, LDNet, and LSSNet. These results collectively demonstrate DFANet's strong generalization, boundary precision, and stability across diverse imaging conditions and dataset complexities.

The qualitative results in Fig. 2 show that DFANet produces accurate and visually consistent polyp segmentations across diverse imaging conditions.

The model preserves fine boundaries even in low-contrast or cluttered regions, benefiting from the complementary strengths of its CNN and transformer branches. Grad-CAM visualizations further indicate that DFANet attends to clinically relevant

TABLE II
QUANTITATIVE COMPARISON OF DFANET WITH STATE-OF-THE-ART METHODS ACROSS FIVE BENCHMARK DATASETS

(a) Kvasir-SEG

Method	mIoU	DSC	Recall	Precision	F_2
U-Net [3]	0.7472	0.8264	0.8504	0.8703	0.8353
U-Net++ [4]	0.7420	0.8228	0.8437	0.8607	0.8295
ResU-Net++ [5]	0.5341	0.6453	0.6964	0.7080	0.6576
HarDNet-MSEG [6]	0.7459	0.8260	0.8485	0.8652	0.8358
ColonSegNet [26]	0.6980	0.7920	0.8193	0.8432	0.7999
UACANet [27]	0.7692	0.8502	0.8799	0.8706	0.8626
UNeXt [28]	0.6284	0.7318	0.7840	0.7656	0.7507
TransNetR [29]	0.8016	0.8706	0.8843	0.9073	0.8744
Proposed (Ours)	**0.8191**	**0.9006**	**0.8745**	**0.9282**	**0.8847**

(b) CVC-ClinicDB

Method	mIoU	DSC	Recall	Precision	F_2
U-Net [3]	0.5433	0.6336	0.6982	0.7891	0.6563
U-Net++ [4]	0.5475	0.6350	0.6933	0.7967	0.6556
ResU-Net++ [5]	0.3585	0.4642	0.5880	0.5770	0.5084
HarDNet-MSEG [6]	0.6059	0.6960	0.7173	0.8528	0.7010
ColonSegNet [26]	0.5090	0.6126	0.6564	0.7521	0.6246
UACANet [27]	0.6808	0.7659	0.7639	0.8820	0.7599
TransNetR [29]	0.8628	0.9107	0.9044	0.9380	0.9065
Proposed (Ours)	**0.8731**	**0.9322**	**0.9235**	**0.9411**	**0.9270**

(c) BKAI-IGH

Method	mIoU	DSC	Recall	Precision	F_2
U-Net [3]	0.7599	0.8286	0.8295	0.8999	0.8264
DeepLabV3 [30]	0.8314	0.8938	0.8870	0.9333	0.8882
PraNet [31]	0.8264	0.8904	0.8901	0.9247	0.8885
UACANet [27]	0.8275	0.8945	0.8870	0.9297	0.8882
CaraNet [32]	0.8329	0.8962	0.8939	0.9273	0.8937
LDNet [33]	0.8254	0.8927	0.8867	0.9153	0.8874
TGANet [34]	0.8409	0.9023	0.9025	0.9208	0.9002
Proposed (Ours)	**0.8510**	**0.9195**	**0.8796**	**0.9631**	**0.8951**

(d) ETIS-LaribPolypDB

Method	mIoU	DSC	Recall	Precision	F_2
U-Net [3]	0.5482	0.6365	0.6000	0.5100	0.5510
PraNet [31]	0.6838	0.7654	0.7500	0.6520	0.6990
SANet [35]	0.6954	0.7746	0.7700	0.6620	0.7130
LDNet [33]	0.6659	0.7413	0.7400	0.6510	0.6970
G-Cascade [36]	0.7490	0.8199	0.8200	0.7460	0.7820
LSSNet [37]	0.7754	0.8464	0.8500	0.7840	0.8160
Proposed (Ours)	**0.8104**	**0.8953**	**0.8767**	**0.9146**	**0.8841**

(e) CVC-ColonDB

Method	mIoU	DSC	Recall	Precision	F_2
U-Net [3]	0.6129	0.7016	0.7000	0.6280	0.6610
PraNet [31]	0.6907	0.7616	0.7600	0.7230	0.7410
SANet [35]	0.6876	0.7675	0.7600	0.7190	0.7400
LDNet [33]	0.7177	0.7968	0.7800	0.7570	0.7680
G-Cascade [36]	0.7319	0.8095	0.8000	0.7770	0.7900
LSSNet [37]	0.7413	0.8203	0.8200	0.7880	0.8030
Proposed (Ours)	**0.7917**	**0.8837**	**0.8887**	**0.8788**	**0.8867**

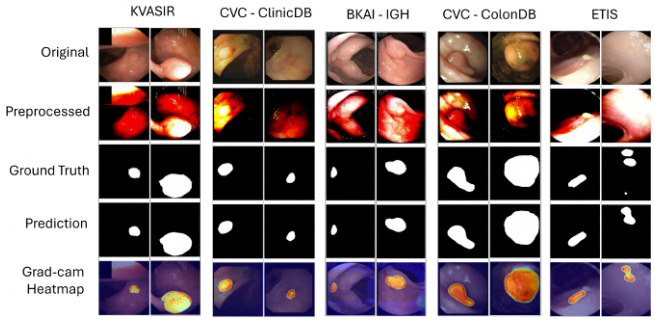

Fig. 2. Qualitative Segmentation Results and Grad-CAM [41] Visualization of DFANet

areas, confirming its reliable feature localization and strong generalization capability.

V. CONCLUSION

This work presented DFANet, a dual-branch feature aggregation network for colonoscopic polyp segmentation that combines global transformer semantics with fine-grained CNN representations. Through a channel–spatial fusion mechanism and DSC Transformer refinement, it effectively integrates multi-scale features for hierarchical decoding, enabling accurate boundary localization and robust mask generation under diverse imaging conditions. Experiments on five benchmark datasets—Kvasir-SEG, CVC-ClinicDB, BKAI-IGH, ETIS-LaribPolypDB, and CVC-ColonDB—show that DFANet consistently outperforms state-of-the-art models across Dice, mIoU, precision, and recall, demonstrating its effectiveness in handling low contrast, variable morphology, and domain shifts. Overall, DFANet is a reliable and computationally efficient framework for real-time polyp segmentation with strong clinical potential, and future work will extend it to video-based segmentation, improve temporal consistency, and develop lighter variants for edge and embedded deployment.

REFERENCES

[1] M. S. Hossain, H. Karuniawati, A. A. Jairoun, Z. Urbi, J. Ooi, A. John, Y. C. Lim, K. M. K. Kibria, A. K. M. Mohiuddin, L. C. Ming, K. W. Goh, and M. A. Hadi, "Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies," *Cancers (Basel)*, vol. 14, no. 7, p. 1732, Mar 29 2022.

[2] S. P. Tholoor, O. Tsagkournis, P. Basford, and P. Bhandari, "Managing difficult polyps: techniques and pitfalls," *Annals of Gastroenterology*, vol. 26, no. 2, pp. 114–121, 2013.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis (DLMIA)*. Springer, 2018, pp. 3–11.

[5] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[6] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," 2021.

[7] D. Jha, M. Riegler, D. Johansen, P. Halvorsen, and H. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 558–564.

[8] N. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha, "Transresu-net: A transformer based resu-net for real-time colon polyp segmentation," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023.

[9] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision – ECCV 2022 Workshops*, ser. Lecture Notes in Computer Science, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Springer, Cham, 2023, vol. 13803, pp. 205–218.

[10] J. Ren, X. Zhang, and L. Zhang, "Hifiseg: High-frequency information enhanced polyp segmentation with global-local vision transformer," *IEEE Access*, vol. 13, pp. 38 704–38 713, 2025.

[11] S. Yi, Y. Zhu, and G. Luo, "Paan: Pyramid attention augmented network for polyp segmentation," in *Proceedings of the Medical Imaging with Deep Learning (MIDL)*, Dec. 2024.

[12] K. Chandrudu, P. Srikanth, K. Sambanna, S. J. Ganji, and S. Shivakeshav, "Mit-daunet: Mit enhanced dual attention unet for medical segmentation," in *2026 National Conference on Communications (NCC)*. IEEE, 2026, pp. 566–571.

[13] P. Srikanth, K. Sambanna, S. J. Ganji, and R. S. Rao, "Fusepolyp: Complementary dual encoder design and fusion strategy analysis for medical polyp segmentation," in *International Conference on Distributed Computing and Intelligent Technology*. Cham: Springer Nature Switzerland, 2026, pp. 436–451.

[14] S. J. Ganji, P. Srikanth, K. Sambanna, and R. S. Rao, "Biencoder-resmambaunet: Dual encoder framework leveraging residual mamba blocks and multi-level semantic convolutional features for polyp segmentation," in *International Conference on Pattern Recognition and Machine Intelligence*. Cham: Springer Nature Switzerland, 2025, pp. 286–295.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 12 077–12 090.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence*, 2017.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[23] Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] W. Wang, E. Yu, J. Xie, X. Li, A. Zhou, X. Sun, J. Xu, L. Zhang, Y. Dong, Z. Wei, T. Lu, C. Shen, and J. Wang, "Pvt v2: Improved baselines with pyramid vision transformer," in *Computational Visual Media*, vol. 8, no. 3. Springer, 2022, pp. 415–424.

[25] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[26] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, vol. 9, pp. 40 496–40 510, 2021.

[27] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*. Association for Computing Machinery, 2021, pp. 2167–2175.

[28] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer Nature Switzerland, 2022, pp. 23–33.

[29] D. Jha, N. K. Tomar, V. Sharma, and U. Bagci, "Transnetr: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing," in *Medical Imaging with Deep Learning (MIDL)*. PMLR, 2024, pp. 1372–1384.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[31] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 12263. Springer, 2020, pp. 263–273.

[32] A. Lou, S. Guan, H. C. Ko, and M. H. Loew, "Caranet: Context axial reverse attention network for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing (SPIE)*, vol. 12032. SPIE, 2022, pp. 81–92.

[33] R. Zhang, P. Lai, X. Wan, D. Fan, F. Gao, X. Wu, and G. Li, "Lesion-aware dynamic kernel for polyp segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 13431. Springer, 2022, pp. 99–109.

[34] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 13431. Springer, 2022, pp. 151–160.

[35] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 12901.

[36] M. M. Rahman and R. Marculescu, "G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE Computer Society, 2024, pp. 7713–7722.

[37] W. Wang, H. Sun, and X. Wang, "Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, ser. LNCS, vol. 15007. Springer Nature Switzerland, 2024, pp. 446–456.

[38] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling. MMM 2020*, ser. Lecture Notes in Computer Science, Y. M. Ro, S. Hwang, T. Kim, W.-T. Chu, and M.-C. Hu, Eds., vol. 11962. Springer, Cham, 2020, pp. 451–462.

[39] J. Bernal, F. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[40] P. N. Lan, N. S. An, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and D. V. Sang, "Neounet: Towards accurate colon polyp segmentation and neoplasm detection," in *Proceedings of the 16th International Symposium on Visual Computing*, 2021.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.