Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.LD30416

CDBT-Unet: A Cross-Attention Transformer-Based Dual-Branch Encoder Framework for Colon Polyp Segmentation

Zihang Li[1*†], Jieqi Li[1†]

[1]*Communication University of China, Beijing, China*
**Corresponding Author. Email: lzh13121107950@cuc.edu.cn*
[†]*These authors contributed equally to this work and should be considered co-first author.*

Abstract. The advancement of colorectal cancer emphasizes how important it is for colonoscopic imaging to accurately segment polyps. Learning-based techniques have made significant progress in the field of polyp medical image segmentation; however, recurring issues such as the identification of small object segments, poorly defined lesion boundaries, and complex backgrounds still exist. In order to overcome these constraints, we introduce CDBT-Unet, a brand-new framework that enhances segmentation performance by integrating two significant innovations. Initially, the transformer layer's convolutional prior speeds up convergence and extracts the fine-grained local texture that is essential for tiny flat polyps. By prioritizing horizontal-vertical background relationships through cross-shaped attention, it improves boundary delineation in complex backgrounds by reducing computation and accelerating convergence. The intricate background and edge blurring issue of polyp segmentation is well-considered in this point. Second, in order to improve accuracy, our dual-path encoder uses the MaxViT block to strategically balance global dependency modeling and local feature preservation. Combining multilevel feature fusion with coordinate space focus mechanisms and channel refinement improves edge response in multiscale fusion. The issue of boundary blurring is the main focus. Under the same experimental setup, our model outperforms the state-of-the-art ConDseg model by 3.72% and the baseline (TransUnet) by 7.32% in terms of Dice scores when tested on the Kvasir-SEG and CVC-ClinicDB datasets. Even in the presence of motion artifacts or low contrast, the framework demonstrates exceptional robustness in segmenting polyps of various sizes. Furthermore, the attention maps that were produced enhanced interpretability and gave physicians practical knowledge about how to make decisions when modeling.

Keywords: Medical Image Segmentation, Cross-Attention Transformer, Dual Path Encoder, TokenMixer, CoordinateSpatial Fusion, TransUnet

1. Introduction

Colorectal cancer, predominantly arising from adenomatous polyps, underscores the critical need for precise polyp segmentation in colonoscopic imaging. Although deep learning approaches have evolved from foundational U-Net [1] architectures to advanced transformer-based [2-5] models like ConDSeg [6], significant challenges persist due to the the identification of small object segments,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30416

poorly defined lesion boundaries and overly complex backgrounds. To address these limitations, we propose CDBT-Unet, featuring two key innovations: (1)a Cross-Shaped Window Transformer block [7] that replaces conventional vision transformers [3-5] through specialized horizontal-vertical attention mechanisms. (2) a dual-branch encoder architecture comprising distinct feature extraction pathways. The Morphology-Aware Branch employs MaxViT blocks [8] to capture global polyp characteristics and spatial hierarchies through their combined channel-wise attention and grid-based operations, effectively modeling long-range dependencies and shape variations that local operators often miss. Conversely, the Texture-Sensitive Branch retains the original TransUnet [9] convolutional backbone, preserving high-frequency edge details and local texture patterns through its inductive bias for spatial locality - features frequently diluted in pure transformer architectures. These complementary pathways are integrated through our novel Coordinate-Spatial Fusion Module, which adaptively combines their outputs using both Coordinate Attention for positionsensitive feature enhancement and SaENet's channel-spatial attention for cross-dimensional context modeling. Coordinate attention encodes spatial position information into channel features and enhances edge responseresponse [10] in multiscale fusion. The channel recalibration [11] further enhances the response to tiny target features, effectively improving the detection and segmentation of small polyps. Experimental results demonstrate CDBT-Unet's superiority over existing approaches, with significant improvements in both mDice score and mIoU metrics compared to baseline and state-ofthe-art methods, while maintaining computational efficiency crucial for clinical deployment. Experimental results demonstrate that CDBT-Unet achieves a 7.32% higher Dice score and 9.17% greater mean IoU than its TransUnet [9] baseline, while surpassing the state-of-the-art ConDSeg model by 3.72% in Dice score and 5.17% in mean IoU.The architecture's particular effectiveness in challenging small-polyp and lowcontrast scenarios, combined with its interpretable attention maps, offers substantial potential for improving early diagnosis and surgical planning in clinical practice.

2. Related work

The evolution of polyp segmentation architectures has been shaped by the need to reconcile conflicting demands of anatomical fidelity and computational practicality. By using symmetric skip connections that preserve spatial details while hierarchically aggregating semantic context, early encoder-decoder frameworks—such as U-Net [1]—established a fundamental paradigm. Nevertheless, the modeling of long-range spatial dependencies is inevitably limited by their dependence on localized convolutional operations, which is a crucial constraint when segmenting polyps with diffuse boundaries or those embedded within intricate mucosal textures. In order to address this, later developments, like TransUNet [9], integrated vision transformers [9] (ViT) to use self-attention mechanisms to capture global contextual relationships. Although these architectures are useful for enhancing coherence for larger lesions, they frequently overlook the directional bias present in endoscopic anatomy, such as the preponderance of vertical luminal structures and horizontal mucosal folds, which results in inefficient attention allocation and unnecessary computational overhead. In order to improve boundary precision by iteratively improving edge predictions, PraNet introduced reverse attention mechanisms. However, its single-scale convolutional processing is unable to adjust to the wide variability in polyp size seen in clinical practice, especially for flat or small lesions (¡5 mm) [12].ConDSeg and other recent state-of-the-art methods use multistage frameworks to enforce prediction consistency and decouple semantic feature learning. Although they show increased robustness, these designs limit their applicability in real-world scenarios with imperfect labels by introducing cascaded computational complexity and

DOI: 10.54254/2755-2721/2025.LD30416

heavily relying on high-quality annotated data for training. Additionally, current approaches usually ignore the synergistic calibration of spatial and channel-wise feature relationships, favoring one over the other. For example, spatial attention modules may not effectively suppress irrelevant background textures, while channel attention mechanisms may not encode positional specificity that is essential for lesion localization. The need for architectures that naturally harmonize multiscale feature representation, adaptive cross-dimensional attention, and directional anatomical priors is highlighted by these unsolved issues.

3. Methodology

3.1. Network architecture

Two key innovations are introduced by CDBT-Unet to improve polyp segmentation through anatomical prior integration, as shown in Fig 1. The traditional skip connection pathway is first redesigned as a Dual-Branch Hierarchical Encoder. The Texture-Preserving Branch uses convolutional operations to preserve high-resolution edge details, while the Morphology-Aware Branch uses MaxViT blocks with grid-based window attention [3] to capture multiscale spatial relationships. Both local boundary characteristics and global polyp morphology can be modeled simultaneously thanks to this dual-path design.A novel Channel-Coordinate Attention Module is used to fuse the branches' multilevel features. It combines two techniques: (1) axial position [13] encoding, which explicitly preserves spatial relationships along endoscopic imaging planes; and (2) entropy-adaptive channel weighting, which suppresses mucosal background interference. Second, the framework includes a Cross-Shaped Window Transformer, which strategically limits self-attention computation to axial stripes that are orthogonal [13]. While removing computationally unnecessary relationships, this directional attention mechanism [14] naturally corresponds with the two most common anatomical orientations in colonoscopy—horizontal mucosal folds and vertical lumen structures. In situations where traditional isotropic attention usually fails, the transformer's anisotropic receptive fields [15] are especially useful for: (1) keeping focus on tiny polyps next to larger lesions, and (2) maintaining boundary continuity across low-contrast regions.

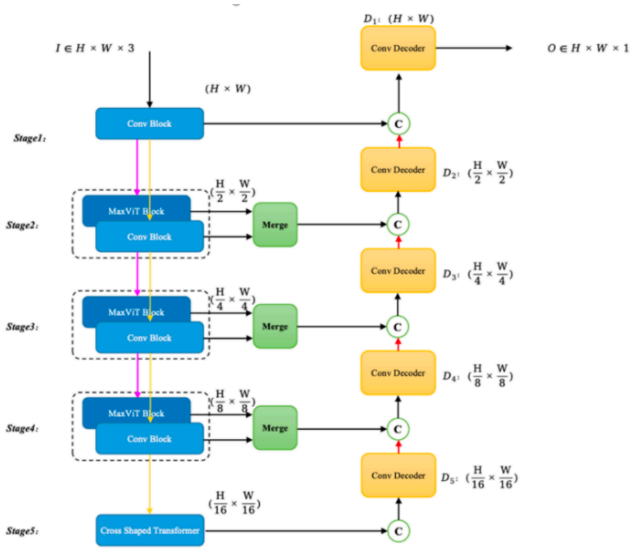

Figure 1. Network architecture

DOI: 10.54254/2755-2721/2025.LD30416

3.2. MaxVit module

The aims to achieve efficient multi-scale feature modelling through the Multi-axis Attention mechanism. It focuses on combining convolution + block attention + grid attention in each module, thus striking a balance between representing local structures and modelling long-range dependencies. Its core design significantly reduces computational complexity while maintaining global modelling capabilities [26-28].

1) Anatomically-Aware Attention Decomposition: Given an input feature map $X \in R^{H \times W \times C}$, MaxViT models long-range dependencies along anatomically salient orientations through axial attention decomposition.

2) Horizontal Attention: Partition the feature map into N_h horizontal strips with dimension $S_h \times W$:

$$Attn_{horizontal}(X) = Softmax\left(\frac{(XW_Q)(XW_K)^T}{\sqrt{d_k}}\right)(XW_V) \tag{1}$$

where W_Q , W_K , W_V are projection matrices. The strip height Sh adapts to anatomical structures, e.g., increasing S_h in regions with prominent mucosal folds.

3) Vertical Attention: Similarly partition into N_v vertical strips $(H \times S_v)$:

$$Attn_{vertical}(X) = Softmax\left(\frac{(XW_Q)(XW_K^T)}{\sqrt{d_k}}\right)(XW_V) \tag{2}$$

4) Multi-Scale Feature Integration: The MaxViT module achieves an optimal balance between global and local feature modeling through its hierarchical design. The multi-axis attention mechanism captures long-range dependencies along anatomically salient orientations [16,28] (vertical lumen structures and horizontal mucosal folds), while the MBConv blocks [17] preserve high-resolution spatial details critical for polyp boundary delineation [26].

3.3. Cross-shaped attention mechanism blocks

Vision Transformer as the last layer of the model downsampling when the attention layer, it makes the downsampled image cut into 9 blocks, and then calculate the relationship between each block and all the other blocks, which makes his convergence slower and too much attention to the global information, the introduction of the cross-shaped attention mechanism can be both concerned about the global information, while not need to calculate the relationship with all the other blocks,. This saves arithmetic and the cross-shaped window also effectively extracts long-distance dependencies.

$$\widehat{X}^l = \mathscr{F}_{CSWin}\left(LN\left(X^{l-1}\right)\right) + X^{l-1} \tag{3}$$

$$X^l = \mathscr{G}_{MLP}\left(LN\left(\widehat{X}^l\right)\right) + \widehat{X}^l \tag{4}$$

The core operation $\mathscr{F}_{CSWin}(\cdot)$ decomposes the standard self-attention into orthogonal directional components:

DOI: 10.54254/2755-2721/2025.LD30416

$$\mathscr{F}_{CSWin}\left(\boldsymbol{Q}\right) = \text{Concat}\left[\underbrace{\text{Attn}(\boldsymbol{Q}_1, \boldsymbol{K}_1, \boldsymbol{V}_1)}_{Horizontal\ scanning}, \underbrace{\text{Attn}(\boldsymbol{Q}_2, \boldsymbol{K}_2, \boldsymbol{V}_2)}_{Vertical\ scanning}\right] \tag{5}$$

For polyp segmentation, this decomposition provides:

·Anisotropic Feature Capture: Horizontal attention sweeps capture longitudinal vessel structures common in gastrointestinal polyps, while vertical attention detects transverse boundary discontinuities.

·Directional Sensitivity: Separate treatment of orthogonal directions improves detection of polyp margins with irregular shapes (e.g., Paris classification Type Is, IIa lesions).

The channel-mixing MLP $\mathscr{G}_{\text{MLP}}(\cdot)$ operates on position-wise features:

$$\mathscr{G}_{MLP}(\boldsymbol{\varkappa}) = W_2\left(GELU\left(W_1\boldsymbol{\varkappa} + b_1\right)\right) + b_2 \tag{6}$$

This can be a good enhancement to the model's ability to model the diversity of polyp morphology.

The combined operations capture clinically relevant patterns:

Table 1. Mathematical-clinical feature correspondence

Mathematical Operation	Clinical Correlation
Horizontal attention	Captures mucosal vascular patterns
Vertical attention	Detects lesion margins against healthy tissue
MLP channel mixing	Fuses RGB/NBI spectral information
Layer normalization	Improving model robustness

3.4. Fusion module

This module combines the difficulties of polyp segmentation, and focuses on the fact that the channel information has the same importance as the coordinate information. In the design of the fusion module, we introduce the channel attention mechanism module (SaE), and at the same time introduce the coordinate attention mechanism.

·Squeeze Aggregated Excitation (SaE) [18] is an improved module based on the channel attention mechanism, which aims to generate more comprehensive channel weights and dynamically adjust the contribution ratio of different feature branches by fusing features of different levels and scales (e.g. shallow details and deep semantics). Enhancing the neural network's ability to perceive key features. Its core design is an extension of the classical SE (Squeeze-and-Excitation) module, which significantly improves the model's representation capability by introducing a multi-branch aggregation strategy and crosslevel information fusion.

· Coordinate Attention [19] extracts horizontal and vertical position-sensitive information by splitting the global twodimensional spatial attention into two one-dimensional directions, horizontal and vertical feature encoding, respectively. The model can accurately capture the positional relationship of the target in space, polyp fuzzy boundaries, and polyp directional texture, which is especially suitable for polyp segmentation, a task that requires precise positioning. The attention mechanism suppresses irrelevant background noise and highlights key features by dynamically adjusting the channel weights, which is especially suitable for small target detection or lowcontrast

DOI: 10.54254/2755-2721/2025.LD30416

regions, such as the complex background of dim light during polyp cutting [30]. The proposed fusion module combines coordinate-aware spatial attention and multibranch channel attention to enhance polyp feature representation [29]. Given an input feature map $X \in \mathbb{R}^{B\times C\times H\times W}$, the fusion process first applies Coordinate Attention (CoordAtt) to model directional dependencies:

$$X_{coord} = X \bigodot (\sigma (W_h \bullet Pool_h (X)) \bullet \sigma (W_w \bullet Pool_w (X))) + X \tag{7}$$

Where

$$Pool_h (X) \in \mathbb{R}^{B\times C\times H\times 1},\ Pool_w (X) \in \mathbb{R}^{B\times C\times 1\times W} \tag{8}$$

$$X_{conv} = ReLU (BN (Conv_{3\times 3} (ReLU (BN (Conv_{3\times 3} (X_{coord})))))) + X_{coord} \tag{9}$$

$$X_{out} = X_{conv} \bigodot \sigma (W_c \bullet [MLP_1 (y) \| MLP_2 (y) \| MLP_3 (y) \| MLP_4 (y)]) \tag{10}$$

where $y = \mathrm{GAP}(X_{\mathrm{conv}}) \in \mathbb{R}^{B\times C}$ and $\|$ denotes concatenation. This hybrid design simultaneously captures position-sensitive patterns, local texture details, and channel-wise interdependencies.

4. Experiment

4.1. Experimental setup

In this study, a multicentre dataset integration strategy was used to organically combine two complementary publicly available datasets, Kvasir-SEG, which contains 1,000 high-quality endoscopic images covering a wide range of polyp morphologies, and CVC-ClinicDB, which provides 612 video frames of colonoscopy with dynamic imaging features. This combination effectively overcomes the limitations of evaluating data from a single source.We evaluate on Kvasir-SEG [20] and CVC-ClinicDB, resizing all images to 256 × 256 pixels. Photometric augmentation is applied during training while preserving anatomical orientation.

Table 2. Datasets specification

Dataset	Images	Polyps	Resolution	Annotation Type
Kvasir-SEG	1,000	1,200	1920 × 1080	Pixel-level
CVC-ClinicDB	612	612	384 × 288	Pixel-level

4.2. Hardware and training protocol

The model was trained on 4× NVIDIA RTX 3090 GPUs using Adam optimizer ($\eta_0 = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with polynomial learning rate decay. The batch size is 16. Each 4 batch in different GPUs to do the training simultaneously:

$$\eta_t = \eta_0 \left(1 - \tfrac{t}{T}\right)^{0.9} \tag{11}$$

where t is the current iteration and T=10,000 is the total training iterations. The learning rate progressively decreased from 1×10^{-3} to 1×10^{-5}, ensuring stable convergence. Evaluation metrics include mDSC, mIoU, Recall, and Precision.

DOI: 10.54254/2755-2721/2025.LD30416

Table 3. Comparison with baseline models

Model	MIoU(%)	mPrecision(%)	mDice(%)
CDBT-UNet	97.27	97.88	98.35
TransUNet	88.10	94.27	91.03
Pra-Ne	91.41	96.33	94.06
ResUnet	84.88	91.48	87.49
ResUnet++	86.05	89.58	88.92
Swin-Unet	79.95	83.35	80.48

4.3. Comparative analysis with baseline models

As demonstrated in Table 3, CDBT-UNet achieves dominant performance across all metrics when compared to classical segmentation architectures. The proposed model attains a 97.27% mIoU, outperforming transformer-based [2] baselines such as TransUNet [9] (88.10%) and SwinUnet [21] (79.95%) by significant margins of 9.17 and 17.32 percentage points respectively, which validates the superiority of our cross-shaped attention over standard global self-attention mechanisms [3,13,22,23] in capturing polyp morphology. Notably, even when compared to specialized polyp segmentation designs like PraNet (91.41% mIoU), CDBT-UNet maintains a 5.86% absolute improvement, highlighting the effectiveness of dual-path feature fusion. The precision-dice gap between our method (97.88% mPrecision / 98.35% mDice) and ResUnet++ [24](89.58% / 88.92%) further confirms enhanced boundary discrimination capability.

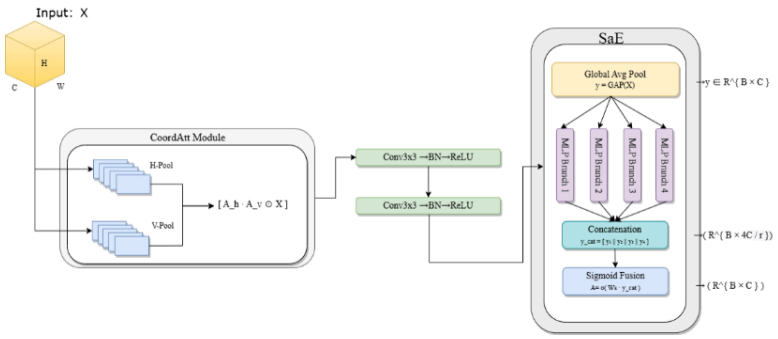

Figure 2. Architecture of the merge block

Table 4. Comparison with state-of-the-art methods

Model	MIoU(%)	mPrecision(%)	mDice(%)
CDBT-UNet	97.27	97.88	98.35
ConDSeg	92.10	96.34	94.63
BMAnet	91.82	95.60	94.40

4.4. Comparison with state-of-the-art methods

Table 4 presents compelling evidence of the advances of CDBT-UNet beyond contemporary SOTA methods. Our model surpasses the cascade architecture of ConDSeg (92.10% mIoU) by 5.17% and

DOI: 10.54254/2755-2721/2025.LD30416

BMAnet's [25]boundaryaware design (91.82%) by 5.45%The 98.35% mDice score not only sets a record, but also demonstrates clinically critical improvements in recall-sensitive scenarios, where even 1% enhancement could reduce diagnostic miss rates substantially.

4.5. Ablation study

To validate the necessity of each core innovation in our framework, we conduct a controlled ablation study under identical training protocols on the dataset. As demonstrated in Table 5, the systematic removal of key components reveals a critical observation:

The experimental results demonstrate the complete model achieves superior performance (98.35% mDice), outperforming the cross-shaped attention-only and dual-branchonly variants by 4.28% and 0.65% respectively in mDice. This indicates non-linear complementarity between the spatial attention mechanism and dual-branch feature fusion.

Table 5. Ablation study results comparison

Model	mPrecision (%)	mIoU (%)	mDice (%)
Only Cross-shaped Attention	94.49	91.42	94.07
Only Dual-branch Encoders	97.15	96.27	97.70
Complete CDBT-UNet	97.88	97.27	98.35

4.6. Segmentation showcase

Figure 3 shows the segmentation of our model, even with complex backgrounds or small target segmentation.



Figure 3. Polyp segmentation results

DOI: 10.54254/2755-2721/2025.LD30416

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention– MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021.

[4] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve J ´ egou. Training data-efficient ´ image transformers & distillation through attention. In International conference on machine learning, pages 10347–10357. PMLR, 2021.

[5] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 22–31, 2021.

[6] Mengqi Lei, Haochen Wu, Xinhua Lv, and Xin Wang. Condseg: A general medical image segmentation framework via contrast riven feature enhancement, 2024.

[7] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with crossshaped windows, 2022.

[8] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer, 2022.

[9] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97: 103280, 2024.

[10] Jingjing Ren, Xiaoyong Zhang, and Lina Zhang. Hifiseg: Highfrequency information enhanced polyp segmentation with globallocal vision transformer. IEEE Access, 2025.

[11] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks. IEEE transactions on medical imaging, 38(2): 540–549, 2018.

[12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation, 2020.

[13] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. arXiv preprint arXiv: 1912.12180, 2019.

[14] Nima Tajbakhsh, Shabana R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging, 35(2): 630–644, 2015.

[15] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24, pages 14–24. Springer, 2021.

[16] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology, 155(4): 1069–1078, 2018.

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4510–4520, 2018.

[18] Mahendran Narayanan. Senetv2: Aggregated dense layer for channelwise and global representations, 2023.

[19] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13708–13717, 2021.

[20] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pal Halvorsen, Thomas de Lange, Dag Johansen, and Havard D. Johansen. Kvasir-seg: A segmented polyp dataset. In Yong Man Ro, WenHuang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve, editors, MultiMedia Modeling, pages 451–462, Cham, 2020. Springer International Publishing.

[21] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, Computer Vision – ECCV 2022 Workshops, pages 205–218, Cham, 2023. Springer Nature Switzerland.

DOI: 10.54254/2755-2721/2025.LD30416

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[23] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In Proceedings of the IEEE/CVF international conference on computer vision, pages 2998–3008, 2021.
[24] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas De Lange, Pal Halvorsen, and H avard D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In 2019 IEEE International Symposium on Multimedia (ISM), pages 225–2255, 2019.
[25] Zihuang Wu, Hua Chen, Xinyu Xiong, Shang Wu, Hongwei Li, and Xinyu Zhou. Bmanet: Boundary-guided multi-level attention network for polyp segmentation in colonoscopy images. Biomedical Signal Processing and Control, 105: 107524, 2025.
[26] Yamagishi Y, Hanaoka S. Ensemble of ConvNeXt V2 and MaxViT for Long-Tailed CXR Classification with View-Based Aggregation [J]. arXiv preprint arXiv: 2410.10710, 2024.
[27] Nguyen-Tat T B, Vo H A, Dang P S. QMaxViT-Unet+: A query-based MaxViT-Unet with edge enhancement for scribble-supervised segmentation of medical images [J]. Computers in Biology and Medicine, 2025, 187: 109762.
[28] Prakash M S, Mallam M. Hybrid Transformers with Multi-scale Feature Extraction for Vision-Language Tasks: CvT, MaxViT, and CoAtNet [J]. IAENG International Journal of Computer Science, 2025, 52(8).
[29] Zhao D, Cai W, Cui L. Adaptive thresholding and coordinate attention-based tree-inspired network for aero-engine bearing health monitoring under strong noise [J]. Advanced Engineering Informatics, 2024, 61: 102559.
[30] Wang Z, Chen Y, Wang F, et al. Improved Unet model for brain tumor image segmentation based on ASPP-coordinate attention mechanism [C]//2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE). IEEE, 2024: 393-397.