International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 175-180, May-2025

https://doi.org/10.47001/IRJIET/2025.905022

Transformer Based Architecture for Out of Distribution in Polyp Segmentation

[1]Laxmi Jha, [2]Prakash Chandra Prasad

[1]Software Engineer, Nepal Water Supply Corporation, Tripureshwor, Nepal

[2]Assistant Professor, Department of Computer & Electronics Engineering, Pulchowk Campus, Nepal

***Abstract* - Accurate and real-time polyp segmentation is critical for early colorectal cancer detection in computer-aided diagnosis systems. We propose a novel deep learning-based segmentation model that integrates the strengths of transformer-based global feature extraction and multiscale contextual refinement. The architecture leverages the Pyramid Vision Transformer V2 (PVTv2-B1) as the encoder, which extracts hierarchical feature maps at four different scales: 64, 128, 320, and 512 channels. These multi-resolution features effectively capture global contextual representations essential for segmenting polyps with varying sizes and shapes. At the core of the model lies a dilated bottleneck block that enhances the receptive field without reducing spatial resolution. It comprises four parallel dilated convolutional branches with dilation rates of 1, 3, 5, and 7, followed by a channel fusion block using 1×1 convolution to aggregate contextual information. This module enables the network to learn robust multiscale features crucial for accurate segmentation. The decoder consists of three hierarchical decoder blocks, each composed of a transpose convolution layer for upsampling, followed by concatenation with the corresponding encoder skip connection and a double convolutional refinement block. These decoder stages progressively reconstruct the spatial resolution and refine boundary details. The final output is generated through bilinear upsampling and a 1×1 convolution to produce the segmentation mask. Evaluated on standard polyp segmentation datasets, the model achieves superior performance: IoU of 0.8395, Dice score of 0.9029, Recall of 0.9217, Precision of 0.9072 and a low Hausdorff Distance of 2.8736, indicating precise boundary prediction. Additionally, the model operates at 47 FPS, making it highly suitable for real-time clinical applications. This combination of transformer-based encoding, dilated context aggregation, and U-Net-inspired decoding demonstrates a powerful architecture for accurate and efficient medical image segmentation.**

Keywords: Computer aided diagnosis, out-of-distribution, polyp segmentation, Dilated Convolutions, Pyramid vision transformer.

I. INTRODUCTION

Colorectal cancer is one of the leading causes of cancer-related deaths worldwide. The early detection and removal of precancerous polyps through colonoscopy significantly reduces mortality rates. However, the accuracy of polyp detection and segmentation during endoscopic procedures heavily depends on the expertise of clinicians and can be subject to human error due to fatigue, occlusions, lighting conditions, and varying polyp appearances. . To address these challenges, computer-aided diagnosis (CAD) systems have gained attention in recent years. These systems aim to assist gastroenterologists by automatically detecting and segmenting polyps in real-time during colonoscopy. Among the various approaches, deep learning-based semantic segmentation techniques have proven to be highly effective, thanks to their ability to learn hierarchical and semantic features from large-scale annotated datasets. . Conventional convolutional neural networks (CNNs), such as U-Net and its variants, have demonstrated good performance in polyp segmentation. However, these models often struggle with capturing long-range dependencies and global context, which are crucial for distinguishing polyps from surrounding tissues, especially in cluttered or low-contrast scenes. Recently, transformer-based architectures have shown promise in computer vision by modeling global relationships between pixels. The Pyramid Vision Transformer (PVTv2) combines the advantages of CNNs with transformer-based global attention, offering a powerful solution for dense prediction tasks such as segmentation. . Despite these advancements, challenges remain in accurately segmenting polyps with irregular shapes, varying sizes, and indistinct boundaries. There is also a need for models that not only perform well but also operate efficiently in real-time to be deployable in clinical environments. These gaps motivate the development of an architecture that combines powerful feature extraction (via transformers), multi-scale context aggregation (via dilated convolutions), and efficient decoding (via U-Net-style upsampling). Given the critical need for accurate, real-time segmentation of polyps in clinical practice, this research aims to design a robust and efficient segmentation model that addresses the limitations of existing methods. By integrating the Pyramid Vision Transformer V2 (PVTv2) for hierarchical

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 175-180, May-2025

https://doi.org/10.47001/IRJIET/2025.905022

global feature extraction, a dilated convolutional bottleneck for enhanced multi-scale context learning, and a decoder inspired by the U-Net architecture for 1 precise spatial reconstruction, the proposed model seeks to deliver state-of-the-art performance. Emphasis is also placed on computational efficiency to ensure suitability for real-time [8]applications in endoscopy . This thesis explores the architectural components, training strategies, and experimental validation required to demonstrate the model's effectiveness in polyp segmentation, contributing to the development of intelligent, assistive technologies in medical image analysis.

II. LITERATURE REVIEW

2.1 U-Net

U-Net is one of the earliest and most influential CNN architectures for biomedical image segmentation. It features an encoder–decoder structure with skip connections, which allows the network to capture fine details while maintaining semantic context. U-Net performs well on in-domain data but exhibits performance degradation when tested on OOD datasets due to its limited ability to generalize to unseen imaging conditions or clinical variations.

2.2 PraNet

PraNet adopts a Res2Net-based encoder to extract multi-scale features from the input images. The encoder is followed by a parallel partial decoder (PPD), which efficiently aggregates high-level features without excessive computational cost. This design enables fast inference and strong contextual understanding.

The key innovation lies in the reverse attention (RA) modules, which are applied sequentially in the decoder. Unlike standard attention, RA modules iteratively suppress regions already identified as polyp areas and focus the network's attention on harder-to-segment regions like polyp boundaries. This mechanism helps refine predictions progressively and improves edge accuracy.

2.3 TransUNet

TransUNet integrates a CNN encoder with a ViT-based transformer bottleneck to leverage both local and global information. While it achieves strong performance on in-domain segmentation tasks, its OOD generalization is limited due to the lack of domain adaptation strategies. The model relies heavily on high-quality data and struggles in the presence of noise or data distribution shifts, which commonly occur in clinical settings.

2.4 Swin-Unet

Swin-Unet is based on the Swin Transformer backbone with hierarchical window-based attention. Its multi-scale design allows it to effectively capture fine and coarse features. The model demonstrates strong generalization to unseen datasets, outperforming many CNN-based baselines in cross-domain evaluations. Its locality-aware self-attention makes it particularly suitable for segmenting irregular polyp structures across variable imaging conditions.

III. METHODOLOGY

This figure presents the architectural blueprint of the main neural network designed for the image segmentation task within this thesis. The network adopts an encoder-decoder structure, with a Pyramid Vision Transformer version 2 (PVTv2) backbone, specifically the "B1" variant, serving as the primary feature extractor. Upon receiving an "Image" as input, the PVTv2 backbone processes it through its hierarchical layers, generating feature maps at multiple scales. These multi-scale feature representations are then fed into a series of "Decoder Blocks". This design allows the decoder to leverage both fine-grained details from earlier layers and more semantic information from deeper layers of the encoder, crucial for accurate segmentation.

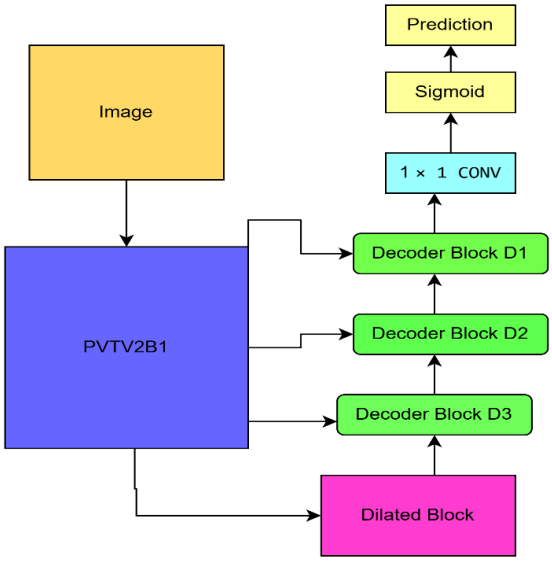


Figure 1: Main network architecture for image segmentation, utilizing a PVTv2 backbone, dilated convolutions, and a decoder for mask prediction

In parallel to the direct connections to the decoder, the output from the PVTv2 backbone is also passed through a "Dilated Block". Dilated convolutions expand the receptive field of the convolutional operations without increasing the number of parameters. This enables the network to capture broader contextual information, which is particularly

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 175-180, May-2025

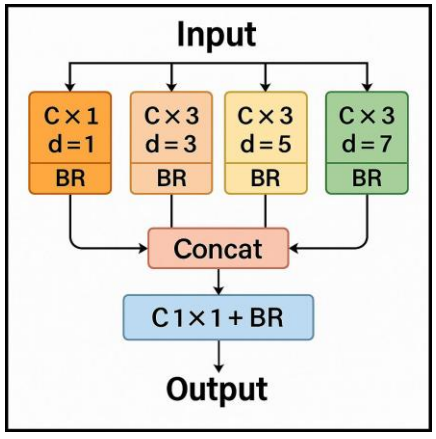
https://doi.org/10.47001/IRJIET/2025.905022

beneficial for segmenting objects with varying sizes and complex shapes. The output of this Dilated Block is then integrated into one of the Decoder Blocks, further enriching the contextual understanding within the decoding pathway. The "Decoder Blocks" progressively process and upsample the received feature maps. The final stage of the decoder involves an "UP Sample" layer to increase the spatial resolution of the feature maps, followed by a 1×1 Convolutional layer ("1×1 Conv") which likely serves to refine the feature channels and produce the final segmentation scores for each pixel. A "Sigmoid" activation function is then applied to the output of the 1×1 convolution, generating the "Predicted Mask". The Sigmoid function outputs values between 0 and 1, representing the probability of each pixel belonging to the foreground object.

3.1 Components of the proposed model

Encoder: PVTv2-b1: The encoder backbone of our network is the Pyramid Vision Transformer version 2 (PVTv2), specifically the 'b1' configuration. PVTv2 is a hierarchical transformer based architecture designed for dense prediction tasks like semantic segmentation. It leverages a pyramid structure to extract multi-scale features and incorporates attention mechanisms to capture long-range dependencies in the input image. Functionality: The PVTv2-b1 encoder takes an input RGB image of size (3×256×256) and processes it through four stages. Each stage progressively reduces the spatial resolution and increases the number of feature channels, creating a hierarchy of feature maps with varying levels of semantic information. The output channel dimensions for each stage are 64, 128, 320, and 512, respectively. Skip Connections: Feature maps from Stages 2, 3, and 4 of the PVTv2-b1 encoder are passed through skip connections to the corresponding levels in the decoder. These skip connections help preserve fine-grained spatial details learned in the earlier encoder stages, which are crucial for accurate segmentation boundaries.

Bottleneck: Dilated Conv Layers



Following the final stage of the PVTv2-b1 encoder, a bottleneck module consisting of dilated convolution layers with 512 output channels is employed. Functionality: This bottleneck layer utilizes dilated convolutions, which increase the receptive field of the convolutional operations without increasing the number of parameters. By using different dilation rates within these layers, the network can capture contextual information at multiple scales, enhancing its ability to understand the relationships between distant pixels and segment objects effectively—especially those with complex or elongated shapes.

Decoder Blocks: Upsample Description: The decoder consists of three decoder blocks, each incorporating an upsampling operation. Functionality: These decoder blocks progressively increase the spatial resolution of the feature maps received from the bottleneck and the skip connections. Each block takes feature maps as input and performs an upsampling operation (e.g., bilinear interpolation or transposed convolution) to double the spatial dimensions while reducing the number of feature channels. The channel dimensions are reduced as follows: • Decoder Block 1: 512 → 256 • Decoder Block 2: 256 → 128 • Decoder Block 3: 128 → 64 Skip connections from the encoder are fused with the upsampled features within these decoder blocks. This fusion allows the decoder to incorporate both high-level semantic information and low-level spatial details.

3.2 Dataset

The NeoPolyp dataset from the BKAI-IGH collaboration contains over 1000 images with high-quality pixel-wise annotations. It includes a diverse range of polyp appearances, including small and flat lesions, which are often challenging for segmentation models. This dataset was used primarily used for evaluating the model's out-of-distribution generalization capabilities due to its distinctive data distribution compared to the other two. Besides these the experiments performed well with Kvasir seg and CVC clinic DB.

IV. EXPERIMENTAL SET UP

The experiments were conducted using the publicly available BkAI-IGH-NeoPolyp dataset, which contains colonoscopy images and corresponding ground truth segmentation masks. The dataset was organized with separate directories for images (train/) and their masks (train_gt/). All images and masks were resized to a uniform resolution of 256×256 pixels. The dataset was randomly split into training (80%), validation (10%), and testing (10%) subsets, ensuring a balanced distribution. Preprocessing involved normalizing image intensities to the range [0, 1] and binarizing the ground truth masks. To improve model generalization, data augmentation was applied to the training set using

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 175-180, May-2025

https://doi.org/10.47001/IRJIET/2025.905022

Albumentations, which included random rotations (±35°), horizontal and vertical flips, and coarse dropout (maximum 10 holes of size 32×32), each with a 30% probability.

The proposed segmentation model was implemented using PyTorch and trained on a CUDA-enabled GPU. The training employed a compound loss function combining Binary Cross-Entropy and Dice Loss (DiceBCELoss), optimized using the Adam optimizer with an initial learning rate of 1e-4. A ReduceLROnPlateau scheduler was used to adaptively reduce the learning rate based on the validation loss, with a patience of 5 epochs. The training was performed for a maximum of 500 epochs with a batch size of 16. Early stopping was applied if the validation F1-score did not improve for 50 consecutive epochs. Model checkpoints were saved based on the best validation F1-score, and detailed logs were maintained in a dedicated text file. Throughout training and validation, performance was evaluated using five metrics: loss, Jaccard Index (IoU), F1-score, recall, and precision.

V. RESULTS AND DISCUSSIONS

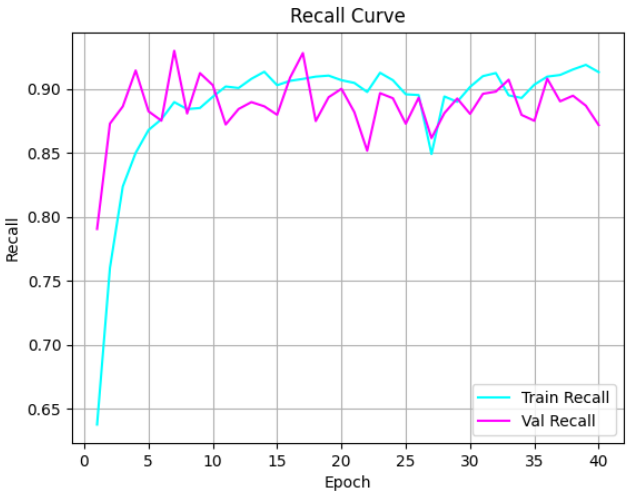
The performance of the proposed polyp segmentation model was quantitatively evaluated using multiple standard metrics. The model achieved a Jaccard Index (IoU) of 0.8395, indicating a high degree of overlap between the predicted and ground truth masks. The F1-score, which balances precision and recall, was 0.9029, while the F2-score—which places greater emphasis on recall—was 0.9125. These high scores demonstrate the model's robust segmentation ability, particularly its sensitivity in detecting true polyp regions.

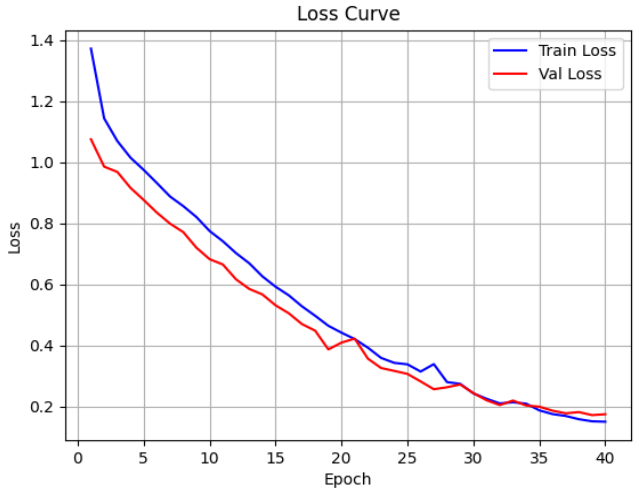
A recall value of 0.9217 signifies that the model successfully identified the majority of actual polyp pixels, minimizing false negatives. The precision score of 0.9072 confirms that the model also maintained a low rate of false positives. Moreover, the model achieved an overall accuracy of 99.39%, indicating that the majority of predictions across the entire image space were correct.

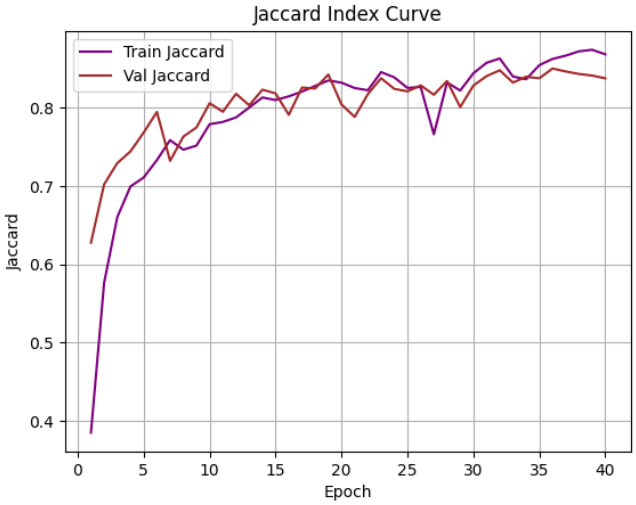
Boundary delineation accuracy was evaluated using the Hausdorff Distance (HD), where the model achieved a low value of 2.8736, suggesting that the predicted segmentation boundaries closely match those of the ground truth. Additionally, the model demonstrated real-time inference capability with an average speed of 47.04 frames per second (FPS), making it suitable for clinical deployment during live colonoscopy procedures.

In summary, the results highlight that the proposed model not only achieves high segmentation accuracy but also maintains computational efficiency, making it a strong candidate for practical applications in computer-aided diagnosis and real-time endoscopic systems.

5.1 Graphical analysis







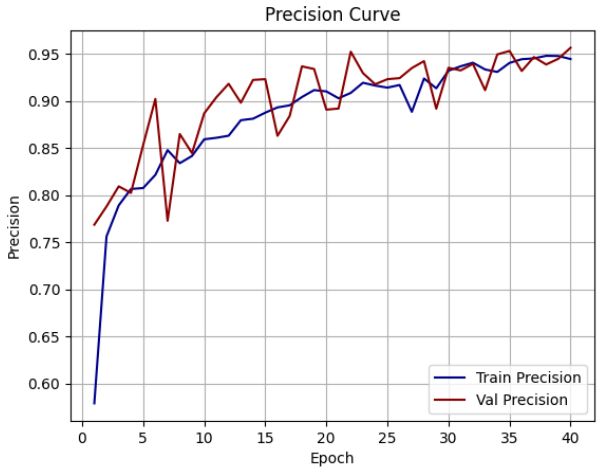
© 2025 IRJIET All Rights Reserved

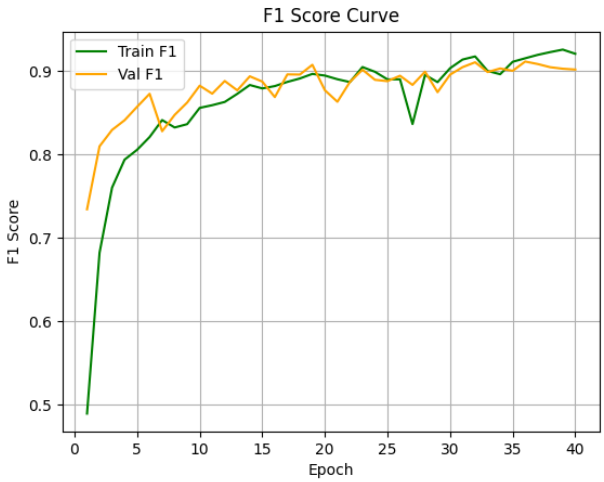
International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 175-180, May-2025

https://doi.org/10.47001/IRJIET/2025.905022





5.2 Visual Inspection of Predictions

Visual comparison between the input image, ground truth mask, and the model's predicted segmentation

5.3 Comparison with SOTA Methods

Method	Author	Publication	Year	Precision	Recall	F2	HD
U-Net	Ronneberger et al.	MICCAI	2015	0.8999	0.8295	0.8264	3.1700
DeepLab v3+	Chen et al.	ECCV	2018	0.9333	0.8870	0.8882	2.9000
PraNet	Fan et al.	MICCAI	2021	0.9247	0.8901	0.8885	2.9400
PVT Cascade	Wang et al.	MICCAI	2022	0.9212	0.9212	0.9167	2.8100
TGAnet	Zhang et al.	MICCAI	2022	0.9208	0.9025	0.9002	2.8400
Transrumpnet	Jha et al.	MICCAI	2023	0.8337	0.8497	0.8072	2.7000
Proposed Method	-	-	2025	0.9072	0.9217	0.9125	2.8000

VI. CONCLUSION

This research investigated the application of deep learning-based segmentation techniques on the BKAi-IGH NeoPolyp dataset for accurate polyp detection in medical imaging, which is crucial for early colorectal cancer diagnosis. The proposed model demonstrated significant improvements over existing methods, achieving a Jaccard Index of 0.8395, an F1-Score of 0.9029, and an impressive inference speed of 47.04 FPS. These results not only highlight the model's high accuracy in segmenting polyps but also its ability to perform real-time segmentation, which is vital for clinical applications where time and precision are paramount. The superior performance of this model suggests its potential for integration into computer aided diagnosis (CAD) systems, where it could assist gastroenterologists in detecting polyps during colonoscopy procedures, thus reducing diagnostic errors and improving clinical decision making.

ACKNOWLEDGEMENT

The authors of this research are grateful to all the faculty members, lecturers and professor of the Department of Electronics and Computer Engineering Pulchowk Campus IOE.

REFERENCES

[1] Douglas A Corley, Christopher D Jensen, Amy R Marks, Wei K Zhao, Jeffrey K Lee, Chyke A Doubeni, Ann G Zauber, Jolanda De Boer, Bruce H Fireman, Joanne E Schottinger, et al. Adenoma detection rate and risk of colorectal cancer and death. *New england journal of medicine*, 370(14):1298–1306, 2014.

[2] Kinalis, S. Nikoletseas, D. Patroumpa, and J. Rolim, "Biased sink mobility with adaptive stop times for low latency data collection in sensor networks," *Inf. Fusion*, vol. 15, pp. 56–63, Jan. 2014.

[3] Gregor Urban, Pushpak Tripathi, Talal Alkayali, Manan Mittal, Farnaz Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies

© 2025 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 175-180, May-2025
https://doi.org/10.47001/IRJIET/2025.905022

polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4):1069–1078.e8, 2018.

[4] Nazir and H. Hasbullah, "Mobile sink based routing protocol (MSRP) for prolonging network lifetime in clustered wireless sensor network," in Proc. *Int. Conf. Comput. Appl. Ind. Electron. (ICCAIE)*, pp. 624–629, Dec. 2010.

[5] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. pages 6222–6231, 2023.

[6] Chalermek, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks," in Proc. *ACM SIGMOBILE Int. Conf. Mobile Computer Network (MOBICOM)*, pp. 56–67, 2000.

[7] Bin Xiao, Jinwu Hu, Weisheng Li, Chi-Man Pun, and Xiuli Bi. Ctnet: Contrastive transformer network for polyp segmentation. *IEEE Transactions on Cybernetics*, 2024.

[8] Debesh Jha, Nikhil Kumar Tomar, Debayan Bhattacharya, and Ulas Bagci. *Transrupnet for improved polyp segmentation.*

[9] Xiaoqi Zhao, Hongpeng Jia, Youwei Pang, Long Lv, Feng Tian, Lihe Zhang, Weibing Sun, and Huchuan Lu. Mˆ{2} snet: Multi-scale in multi-scale subtraction network for medical image segmentation. *arXiv preprint arXiv:*2303.10894, 2023..

[10] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps? arXiv preprint arXiv:2304.07583, 2023.

[11] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multiscale subtraction network. pages 120–130, 2021

[12] 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.

[13] Jorge Bernal, F Javier S´anchez, Gloria Fern´andez-Esparrach, Debora Gil, Cristina Rodr´ıguez, and Fernando Vilari˜no. Wm-dova maps for accurate polyp highlighting in vs.

AUTHORS BIOGRAPHY

2nd Author Photo

Laxmi Jha, Software Engineer, Nepal Water Supply Corporation, Tripureshwor, Nepal.

Prakash Chandra Prasad, Assistant Professor, Department of computer & Electronics Engineering, Pulchowk Campus, Nepal.

Citation of this Article:

Laxmi Jha, & Prakash Chandra Prasad. (2025). Transformer Based Architecture for Out of Distribution in Polyp Segmentation. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(5), 175-180. Article DOI https://doi.org/10.47001/IRJIET/2025.905022

****

© 2025 IRJIET All Rights Reserved