# Enhancing Polyp Segmentation in Colorectal Cancer using Attention U-Net Model

Matthew Lam
*Department of Electrical and Computer Engineering*
*Toronto Metropolitan University*
Toronto, Ontario
matthew1.lam@torontomu.ca

***Abstract*—The incidence of colorectal cancer among younger adults in Canada is on the rise, highlighting the urgent need for improved methods in polyp segmentation and detection.[1] Accurate segmentation algorithms are crucial for identifying polyps and providing detailed information about their size and shape, which is essential for early diagnosis and treatment. While current segmentation algorithms have shown promising results, there is still significant potential for enhancement to meet clinical needs effectively. This study explores various segmentation models and proposes a project to incorporate attention mechanisms into convolutional neural networks, specifically U-Net, to enhance the accuracy of polyp segmentation. This approach aims to advance the early evaluation and management of polyps, potentially reducing the incidence of colorectal cancer.**

***Index Terms*—U-Net, Attention Mechanism, Colorectal Cancer, Polyp Segmentation**

## I. Introduction

According to the Canadian Cancer Society, colorectal cancer is the second leading cause of cancer death in men and the third in women.[2] The incidence of colorectal cancer in adults under 50 is increasing, particularly among those aged 20-29 and 30-39.[1] Colon polyps, clusters of cells forming on the colon lining, are generally benign but can develop into colorectal cancer.[3] Colonoscopy is the most effective method for capturing medical images to detect colon polyps. Analyzing these images using deep learning techniques for polyp segmentation can significantly improve diagnosis by quickly identifying abnormalities in the colon, allowing for earlier intervention and reducing the risk of colorectal cancer progression. Recent studies have highlighted the potential of various deep-learning architectures in medical image analysis. Among these, attention mechanisms in deep learning models have shown great promise, especially in image segmentation. This research aims to leverage Attention U-Net architecture's capabilities to improve polyps' segmentation accuracy in colonoscopy images.

## II. Literature Survey

Before the advent of deep learning, various techniques, such as edge-based and texture-based methods, were used for image segmentation. However, deep learning models, especially U-Net, have revolutionized segmentation accuracy by automating the feature extraction process. These models utilize large datasets and significant computational resources to learn complex patterns and features directly from raw pixel data, eliminating the need for manual feature extraction.

One notable enhancement in this domain is the use of attention mechanisms, which can significantly improve model performance. The SIA-Unet model by Ye et al. exemplifies this by enhancing gastrointestinal (GI) tract image segmentation from MRI scans. It incorporates short-term sequence information and an attention mechanism that filters out irrelevant information, allowing the model to focus on critical features. This approach resulted in a Dice coefficient of 91.13 percent and a Jaccard coefficient of 88.28 percent. [4]

Similarly, Chong et al. developed the P-Trans U-Net model, which combines transformers and convolutional neural networks (CNNs) to improve the detection and segmentation of lesion edges. This model integrates the strengths of both architectures and has outperformed state-of-the-art models across four distinct medical imaging segmentation tasks. [5]

Another enhancement is modifying the encoder which can be seen in N-Net and Attentional Feature Fusion (AtFF) Module. N-Net enhances U-Net by incorporating two parallel encoder branches, one for fine-grained spatial details and the other for broader contextual information, leading to improved performance on lung and liver CT images, as well as COVID-19 X-ray images.[6] In AtFF, Liu et al. developed a dual-branch approach that merges the strengths of CNNs and Transformers in the UNet's encoder, with the ConvNeXt CNN capturing local details and the Swin Transformer capturing global context, achieving a high Dice score of 91.65 percent in cancer cell segmentation.[7]

The encoder and decoder can both be modified to find the best encoder-decoder combinations based on the specific dataset. The experiment concluded that the best approach for segmenting gastrointestinal tract images involved using EfficientNet B0 as the encoder to downsample the images, the Feature Pyramid Network (FPN) as the decoder, and the Adam optimizer.[8]

Dynamic resolution adjustment in medical image segmentation has been significantly improved with innovative models like Fovea-UNet. Developed by Lui et al., Fovea-UNet employs Fovea Pooling (FP) to dynamically adjust the resolution of different image regions based on their importance, which enhances lymph node detection in colorectal cancer, achieving a high Dice score of 88.51 percent and a sensitivity of 92.82 percent. [9]

In the realm of real-time and resource-constrained segmentation, Jha et al.'s NanoNet stands out. Designed for real-time polyp segmentation during endoscopic procedures, NanoNet utilizes MobileNetV2 and an optimized residual block to enhance responsiveness. However, it sometimes produces overly large segmentation masks for large lesions. [10]

For object detection and segmentation, Murugesan et al. introduced YOLOv3-MSF, a deep learning network aimed at colon cancer staging based on tumor length. Trained on the CVC colonDB database, this model provides bounding boxes for detected objects but lacks detailed segmentation information.[11]

Lastly, Yang et al.'s UNet-D targets artifact detection in endoscopic images. By incorporating a dense layer to maintain low-level spatial information, UNet-D enables real-time artifact segmentation. However, its overall performance is lower compared to other models like UNet and Deeplab, with an F2 score of 0.44. [12]

Together, these models showcase diverse approaches to improving medical image segmentation, whether through dynamic resolution adjustment, real-time responsiveness, object detection, or artifact segmentation.

## III. Problem Statement and Dataset

The increasing incidence of colorectal cancer in Canada, especially among younger individuals, highlights the urgent need for advanced and efficient polyp identification and segmentation techniques to overcome the time-consuming and variable nature of manual colonoscopy image segmentation in clinical settings. This study proposes Attention U-Net to enhance the precision and consistency of polyp segmentation in colonoscopy images.

The Kvasir-SEG dataset, an open-source collection, consists of 1,000 images with resolutions ranging from 332x487 to 1920x1072 pixels. Specifically tailored for polyp segmentation in the gastrointestinal tract, each image in this dataset is paired with a corresponding mask. These masks, originally annotated by a medical doctor, have been reviewed by an expert gastroenterologist to ensure accuracy. The images are provided in RGB format, while the ground truth segmentation masks are binary.

## IV. Preprocessing and Augmentation

The image is first resized to 244 x 244 pixels and then normalized using a mean and standard deviation of 0.5. This preprocessing step ensures that the images are consistent in size and scaled appropriately, making it feasible to load them into memory efficiently.

To enhance the variability of our training dataset, we implemented several augmentation techniques. These techniques include rotating images by 0, 90, 180, or 270 degrees; randomly cropping the images; flipping them horizontally or vertically; and adjusting their brightness, contrast, and hue. These augmentations are designed to improve the model's robustness by exposing it to a diverse array of transformed images, thereby simulating various real-world scenarios.

## V. Model Description

The U-Net model, developed by Ronneberger et al., is renowned for its efficacy in image segmentation tasks, primarily due to its dual-path architecture comprising an encoder and a decoder.[7] The encoder, utilizing convolutional layers, ReLU activations, and max pooling, effectively downsamples the input image to generalize its features. In contrast, the decoder employs up-convolutions and concatenations to recover the spatial resolution, ultimately generating the segmentation mask.

The Attention U-Net, introduced by Oktay et al., builds upon the original U-Net architecture by incorporating self-attention gating mechanisms in the skip connections of the decoder.[9] These attention gates, represented by red circles in the architecture diagrams, introduce an attention coefficient that quantifies the importance of each feature. These coefficients are calculated based on the relevance of features for the specific task, allowing the network to prioritize critical features while down-weighting less important ones. This selective focus improves the network's ability to highlight significant features necessary for accurate segmentation.

The Attention U-Net maintains a similar encoder-decoder structure to the original U-Net. The encoder path includes convolutional blocks with progressively increasing filters: 64, 128, 256, 512, and 1024. Each block consists of two consecutive 3x3 convolutions, each followed by batch normalization and ReLU activation, with max-pooling used to reduce spatial dimensions. In the decoder path, the model upsamples the feature maps back to the original input size using up-convolution blocks with filters: 512, 256, 128, and 64. Each block performs upsampling followed by a 3x3 convolution, batch normalization, and ReLU activation. The attention gates modulate the skip connections by generating attention maps using 1x1 convolutions, ReLU activations, and a sigmoid function to compute the attention coefficients.

## VI. Function Loss

Binary Cross-Entropy (BCE) Loss is commonly used to measure the dissimilarity between the predicted probability map, $S_{\text{pred}}$, and the ground truth map, $S_{\text{gt}}$. The BCE Loss is calculated using the following formula:

$$\text{BCE Loss} = \frac{1}{N}\sum_{n=1}^{N} -w_n \left[S_{\text{gt},n} \cdot \log(S_{\text{pred},n}) + (1 - S_{\text{gt},n}) \cdot \log(1 - S_{\text{pred},n})\right]$$

- $S_{\text{pred},n}$ is the predicted probability of the $n$th pixel,
- $S_{\text{gt},n}$ is the ground truth probability of the $n$th pixel,
- $w_n$ is the weight of the $n$th pixel.

The BCE loss function penalizes incorrect classifications, ensuring the model improves its predictive accuracy over iterations. This function is particularly sensitive to class imbalances, and to mitigate this issue, a weighting factor is applied.

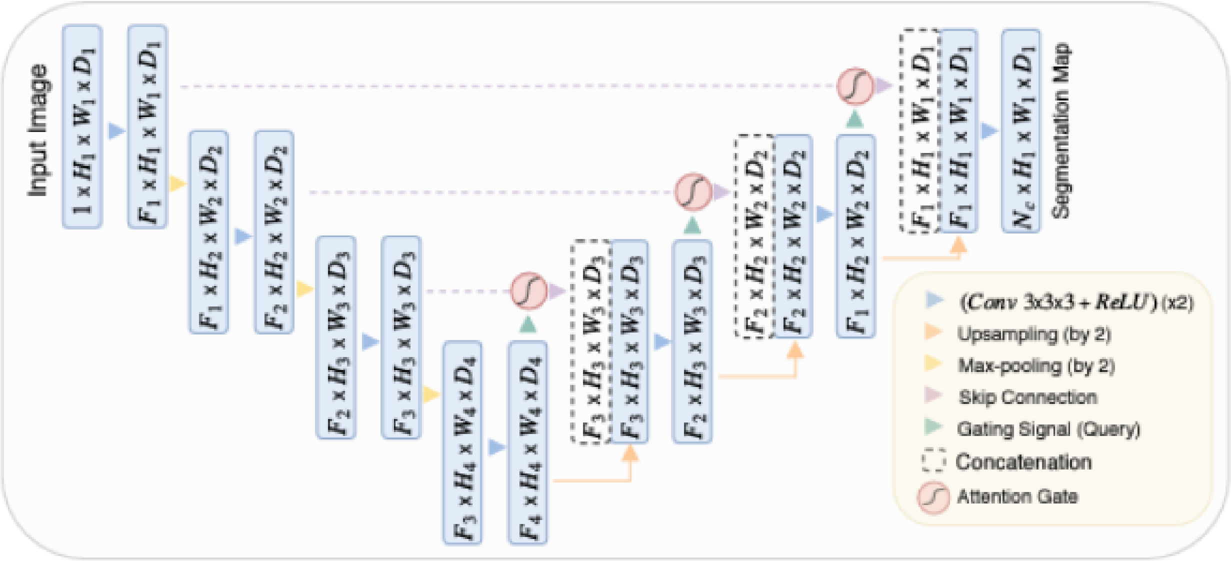

Fig. 1. Shows Attention U-Net illustrated by Oktay et al. [14]

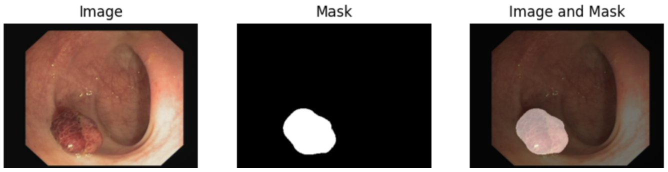

Fig. 2. Show a sample image that includes the following components: a original input image, a binary mask, and the binary mask overlaid on to the original image

## VII. Evaluation

Three methods are used to measure the performance of the model: Accuracy, Jaccard Similarity, and Dice Coefficient.

Accuracy measures the proportion of correctly classified pixels, providing a straightforward metric for overall classification performance.

$$\text{Accuracy} = \frac{\text{Number of Correct Pixels}}{\text{Total Number of Pixels}} \tag{1}$$

Jaccard Similarity (JS) quantifies the overlap between predicted and actual segments by comparing the intersection and union of positive pixels, highlighting the extent of correct coverage.

$$\text{Jaccard Similarity} = \frac{|S_{\text{pred}} \cap S_{\text{gt}}|}{|S_{\text{pred}} \cup S_{\text{gt}}|} \tag{2}$$

Dice Coefficient (DC) assesses the overlap between predicted and actual segments, focusing on the ratio of common pixels to the total positive pixels in both sets.

$$\text{Dice Coefficient} = \frac{2|S_{\text{pred}} \cap S_{\text{gt}}|}{|S_{\text{pred}}| + |S_{\text{gt}}|} \tag{3}$$

## VIII. Results

The results, summarized in Table 1, demonstrate that the Attention U-Net model outperforms the other models in terms of performance metrics. However, it is important to note that accuracy alone is not a sufficient metric for evaluating segmentation models. High accuracy can be misleading, particularly when the segmented object is small, as the large number of background pixels can disproportionately influence the accuracy score. Therefore, it is essential to consider Jaccard Similarity (JS) and Dice Coefficient (DC), to provide a more comprehensive assessment of model performance.

TABLE I
Comparing different models

| **Model** | **Accuracy** | **JS** | **DC** |
|---|---|---|---|
| U-Net [7] | 92.002% | 61.876% | 70.994% |
| R2U-Net [8] | 92.727% | 59.865% | 70.242% |
| Attention U-Net [9] | 93.211% | 62.547% | 71.363% |

The test dataset was processed using the best-performing segmentation model, Attention U-Net. Figure 3 below presents the best and worst two results alongside their ground truth. In the top two cases, the model successfully identifies the core of the segmentation but struggles with accurately segmenting the edges, often displaying uncertain grey transitions in these areas. In contrast, the worst two cases show that the model completely misses the tumor cancer segmentation or even incorrectly classifying them. These errors could be attributed to the small size or irregular shape of the tumors or even larger structures that divert the model's focus.

## IX. Limitation

A significant limitation observed in the model training process is the overfitting, evident from the training Dice score

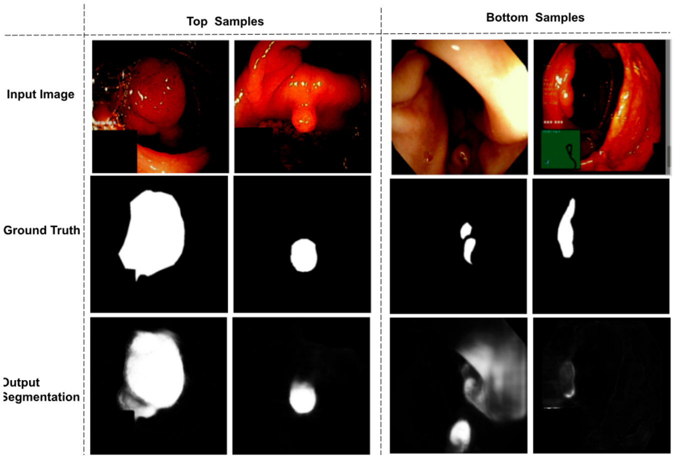

Fig. 3. Top two and bottom two examples based on Dice score, showing input image, ground truth, and output segmentation from the Attention U-Net model.

being higher than the validation Dice score. This overfitting is likely due to the relatively small and homogenous dataset, which consists of only 1,000 images. To address the overfitting issue, we can enhance our dataset by incorporating images from the CVC-ClinicDB collected from real-world clinical settings. [16]

## X. Conclusion

The proposed Attention-Net model demonstrates superior performance in polyp segmentation tasks, significantly enhancing the accuracy and reliability of detecting colorectal abnormalities in colonoscopy images. This model improves feature representation and prioritization by integrating attention mechanisms, leading to more precise segmentation results. These advancements have the potential to facilitate earlier diagnosis and intervention, ultimately contributing to a reduction in colorectal cancer incidence among younger adults in Canada.

Github: https://github.com/MatthewLamBok/Assignment_NN

## References

[1] D. E. O'Sullivan, R. J. Hilsden, Y. Ruan, N. Forbes, S. J. Heitman, and D. R. Brenner, "The incidence of young-onset colorectal cancer in Canada continues to increase," Cancer Epidemiol., vol. 69, p. 101828, Dec. 2020, doi: 10.1016/j.canep.2020.101828.

[2] C. C. S. / S. canadienne du cancer, "Colorectal cancer statistics," Canadian Cancer Society. Accessed: Jun. 23, 2024. [Online]. Available: https://cancer.ca/en/cancer-information/cancer-types/colorectal/statistics

[3] M. Meseeha and M. Attia, "Colon Polyps," in StatPearls, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Jun. 23, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK430761/

[4] R. Ye, R. Wang, Y. Guo, and L. Chen, "SIA-Unet: A Unet with Sequence Information for Gastrointestinal Tract Segmentation," in PRICAI 2022: Trends in Artificial Intelligence, S. Khanna, J. Cao, Q. Bai, and G. Xu, Eds., Cham: Springer Nature Switzerland, 2022, pp. 316–326. doi: 10.1007/978-3-031-20862-1-23.

[5] Y. Chong, N. Xie, X. Liu, and S. Pan, "P-TransUNet: an improved parallel network for medical image segmentation," BMC Bioinformatics, vol. 24, no. 1, p. 285, Jul. 2023, doi: 10.1186/s12859-023-05409-7.

[6] B. Liang, C. Tang, W. Zhang, M. Xu, and T. Wu, "N-Net: an UNet architecture with dual encoder for medical image segmentation," Signal Image Video Process., vol. 17, no. 6, pp. 3073–3081, Sep. 2023, doi: 10.1007/s11760-023-02528-9.

[7] J. Liu, S. Mao, and L. Pan, "Attention-Based Two-Branch Hybrid Fusion Network for Medical Image Segmentation," Appl. Sci., vol. 14, no. 10, Art. no. 10, Jan. 2024, doi: 10.3390/app14104073.

[8] N. Sharma, S. Gupta, M. S. A. Reshan, A. Sulaiman, H. Alshahrani, and A. Shaikh, "EfficientNetB0 cum FPN Based Semantic Segmentation of Gastrointestinal Tract Organs in MRI Scans," Diagnostics, vol. 13, no. 14, p. 2399, Jul. 2023, doi: 10.3390/diagnostics13142399.

[9] Y. Liu, J. Wang, C. Wu, L. Liu, Z. Zhang, and H. Yu, "Fovea-UNet: detection and segmentation of lymph node metastases in colorectal cancer with deep learning," Biomed. Eng. OnLine, vol. 22, no. 1, p. 74, Jul. 2023, doi: 10.1186/s12938-023-01137-4.

[10] D. Jha et al., "NanoNet: Real-Time Polyp Segmentation in Video Capsule Endoscopy and Colonoscopy," in 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2021, pp. 37–43. doi: 10.1109/CBMS52027.2021.00014.

[11] M. Murugesan, R. Madonna Arieth, S. Balraj, and R. Nirmala, "Colon cancer stage detection in colonoscopy images using YOLOv3 MSF deep learning architecture," Biomed. Signal Process. Control, vol. 80, p. 104283, Feb. 2023, doi: 10.1016/j.bspc.2022.104283.

[12] S. Yang and S. Cochran, "Graph-search Based UNet-d For The Analysis Of Endoscopic Images." Accessed: Jun. 23, 2024. [Online]. Available: http://ceur-ws.org/Vol-2366/EAD2019-paper-6.pdf

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 18, 2015, arXiv: arXiv:1505.04597. Accessed: Jun. 21, 2024. [Online]. Available: http://arxiv.org/abs/1505.04597

[14] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," Apr. 2018.

[15] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," May 29, 2018, arXiv: arXiv:1802.06955. doi: 10.48550/arXiv.1802.06955.

[16] D. Jha et al., "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in 2019 IEEE International Symposium on Multimedia (ISM), Dec. 2019, pp. 225–2255. doi: 10.1109/ISM46123.2019.00049.