Polyp Segmentation in Colonoscopy Images

Tejas Jadhav, Nakul Gupta, Shraddha Gadale
tjadhav@iu.edu, guptanak@iu.edu, sgadale@iu.edu

***Abstract*—This study addresses the critical challenge of polyp segmentation in colonoscopy images, a key diagnostic tool in preventing colon cancer. Polyp segmentation is challenging due to the variability in polyp appearance and the presence of visually similar benign structures. We implemented and assessed enhancements on two state-of-the-art deep learning architectures: DeepLabV3+ with a ResNet50 backbone and ResUNet++ to improve segmentation accuracy and model robustness. We innovated upon these models by integrating randomized activation functions and forming an ensemble from these variants, further enriched by a unique data augmentation strategy. Results demonstrated noticeable improvements in both models' performance. This study illustrates the potential of combining stochastic activation functions with advanced augmentation techniques to significantly enhance the performance of segmentation models in medical imaging.**

***Keywords*—Medical image analysis, semantic segmentation, colonoscopy, polyp segmentation, deep learning**

I. Introduction

Colorectal Cancer (CRC) remains a major global health concern, being one of the leading causes of cancer-related mortality worldwide. Effective early detection and resection of colorectal polyps during colonoscopy examinations are pivotal in preventing CRC. Polyps, particularly adenomatous types, serve as precursors to colorectal cancers. Their detection, however, is complex due to their diverse morphologies, including variations in size, shape, and color, which often result in many polyps remaining undetected during endoscopic procedures.

To enhance polyp detection, Computer-Aided Detection (CAD) systems have been developed to assist clinicians in identifying potential polyps in real-time during endoscopic examinations. These systems act as a secondary observer, drawing the clinician's attention to potential areas of concern on the monitor, thereby reducing the likelihood of oversight. Modern CAD systems not only detect anomalies but also provide pixel-wise segmentation, marking the precise boundaries of polyps for accurate diagnosis and treatment planning.

However, the development of advanced CAD systems is hindered by the substantial costs associated with the collection and annotation of extensive medical datasets required for training. Moreover, the inherent diversity in polyp appearance—from adenomas to hyperplastic and serrated types—and their resemblance to normal mucosal structures or occlusion by stool add significant complexity to the development of effective segmentation algorithms.

In response to these challenges, this project leverages advanced deep learning architectures renowned for their efficacy in semantic segmentation tasks. Specifically, we explore the DeepLabV3+ [1] model equipped with a ResNet50 backbone, which utilizes atrous convolution to capture multiscale information efficiently and incorporates an Atrous Spatial Pyramid Pooling (ASPP) module [2] to improve segmentation across various scales.

Additionally, we examine the U-Net architecture, particularly its advanced variant, ResUNet++ [3]. This model enhances the classic U-Net structure by integrating residual connections to mitigate the vanishing gradient problem and includes innovative features such as residual blocks, squeeze-and-excitation blocks, ASPP, and attention gates. These enhancements facilitate the training of deeper networks, crucial for extracting detailed features in complex medical images.

This study explores various enhancements to these models, motivated by recent advances [4] that suggest the use of stochastic activation functions and innovative augmentation techniques can further improve model performance. Specifically, we employ multiple activation functions randomly across different layers and introduce unique augmentation strategies to expand the diversity of the Kvasir-SEG dataset [5]. We generated five distinct model variations, for both ResUNet++ and DeepLabV3+ architectures, using random activation functions, and trained them independently. Finally, we combined these five models into an ensemble to optimize our results for each architecture. This approach aims to maximize the generalization and robustness of our CAD system, paving the way for more accurate and reliable polyp detection.

II. Dataset

For the critical task of polyp image segmentation, the models must differentiate between two classes

at the pixel level: polyp and non-polyp. This binary classification enables detailed analysis of colonoscopy images, where precise segmentation at the pixel level is crucial for effective medical diagnosis. To evaluate the performance of the model architecture, we utilize the Kvasir-SEG dataset [5], specifically designed for this purpose. The dataset contains 1,000 high-quality annotated polyp images. Each image is paired with a corresponding ground truth mask meticulously annotated by expert endoscopists at Oslo University Hospital, Norway, as shown in Figure 1. This robust dataset offers a diverse range of polyp images, reflecting the variety in appearance that polyps can exhibit, which is essential for training and testing segmentation models to effectively handle real-world clinical scenarios.

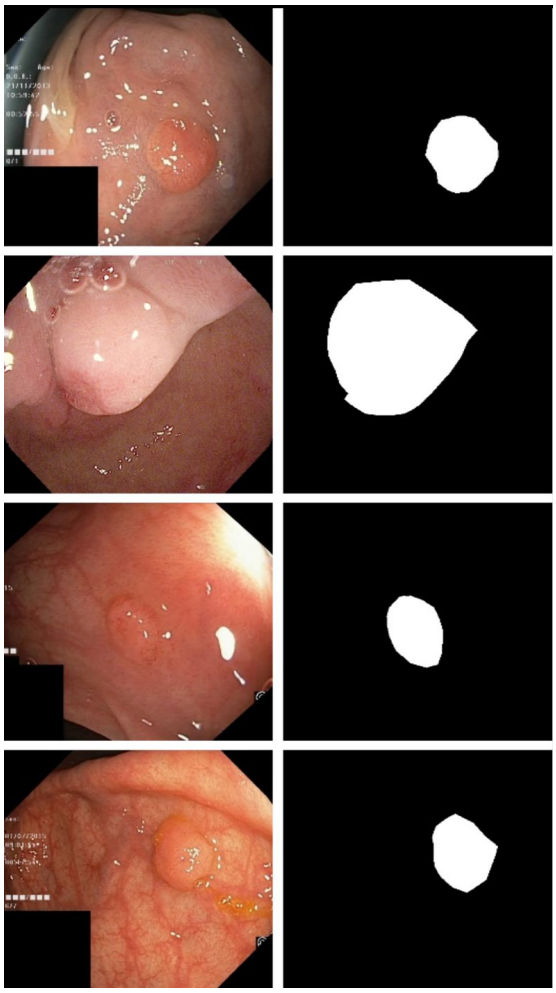

Figure 1: Examples of polyp images and their corresponding masks from the Kvasir-SEG dataset.

A. Augmentation

In medical image analysis, particularly for polyp segmentation in colonoscopy images, data augmentation plays a crucial role in enhancing the robustness and generalization capabilities of deep learning models. The enhancement techniques used in this study not only increase the quantity of training data but also introduce a variety of transformations that simulate different real-world conditions, which a model may encounter in clinical settings.

The image sizes vary between 332 × 487 to 1920 × 1072 pixels, for training purposes, the images are first resized to the input size of 256 × 256 pixels, then we systematically apply transformations- Random Crop, Center Crop, Horizontal Flip, Vertical Flip, Cutout, Scale Augmentation, Random Rotation, Brightness Augmentation, and RGB to Grayscale to each image and corresponding mask.

This diverse set of augmentations is applied iteratively to each image in the Kvasir-SEG dataset, resulting in 31 unique variations per original image expanding the initial dataset of 800 images to a comprehensive set of 24,800 unique training samples, thereby creating a rich and varied dataset. These transformations are crucial for training a robust model capable of high precision and recall, as demonstrated by the enhanced performance metrics of our models after training with this augmented dataset.

III. Methodology

A. DeepLabV3+

DeepLabV3+ with a ResNet50 backbone is a specific configuration of DeepLabV3+ [6] architecture for semantic image segmentation, particularly enhanced by the inclusion of atrous convolutions that facilitate detailed feature extraction across multiple scales without a loss in resolution. The architecture integrates an Atrous Spatial Pyramid Pooling (ASPP) module [2] which efficiently captures context at multiple scales and improves segmentation accuracy significantly. The encoder-decoder structure, with depthwise separable convolutions in the decoder, optimizes the segmentation process by refining the spatial resolution of the output, making it particularly suited for tasks requiring high-detail predictions like polyp segmentation.

For this project, we enhanced the DeepLabV3+ model by integrating Stochastic Activation Selection, a novel approach where various activation functions are randomly assigned to different layers within the network. This method was designed to dynamically investigate non-linear relationships and involved substituting the conventional ReLU activation layers with a diverse array of activation functions, including Leaky ReLU, ELU, PReLU, SReLU, MeLU, GaLU, and Mish. Each network iteration employed a unique combination of these activation functions, randomly selected from our predefined pool for each training cycle. This process

resulted in the creation of multiple distinct networks, each tailored to capture unique aspects of the data.

To leverage the diversity of these networks, we trained five separate instances, each featuring a different set of activation functions. The outputs of these models were then combined by averaging their softmax outputs, forming an ensemble. This ensemble [7] approach helps to mitigate model variance and enhance generalization capabilities, significantly improving the accuracy and robustness of the segmentation results.

The effectiveness of the stochastic activation approach is demonstrated through quantifiable improvements in performance metrics such as the Dice Coefficient (DSC) and Intersection over Union (IoU). During initial testing with a standard DeepLabV3+ model, these metrics showed notable enhancements upon implementing stochastic activation functions, significantly affirming the efficacy of this method in enhancing segmentation accuracy.

B. ResUNet++

ResUNet++ is an advanced iteration of the traditional U-Net and ResUNet architectures, specifically designed to enhance medical image segmentation tasks like polyp detection. This model integrates several sophisticated components: residual blocks to enable deeper network architectures without performance degradation, squeeze-and-excitation (SE) blocks that recalibrate feature channels to emphasize relevant features, Atrous Spatial Pyramid Pooling (ASPP) to capture contextual information at various scales, and attention mechanisms that focus the model's computational power on significant areas of the input. Together, these enhancements significantly improve the network's ability to accurately segment complex medical images by boosting its depth, efficiency, and contextual adaptability.

Similar to the approach taken with DeepLabV3+, ResUNet++ was adapted to include Stochastic Activation Selection to dynamically explore non-linear relationships within the data. By randomly replacing the conventional ReLU activation functions with a variety of alternatives such as Leaky ReLU, ELU, PReLU, and others, each instantiation of the network potentially emphasizes different aspects of the input data. This diverse activation approach fosters a broad exploration of feature space, enhancing the model's ability to generalize across varied and complex image features found in polyp segmentation.

Similarly, five unique versions of ResUNet++ [3] are trained with these different activation functions, and their predictions are then aggregated. This ensemble approach mitigates overfitting and enhances the robustness of the segmentation output by averaging across diverse model interpretations.

The implementation of stochastic activations in ResUNet++ has demonstrated significant improvements in performance metrics. The introduction of diverse activation functions led to a substantial enhancement in key measures such as accuracy and precision. These results underscore the effectiveness of incorporating a variety of activation functions within complex segmentation architectures, confirming the value of this approach in boosting model performance.

C. Implementation details

To effectively address the computational demands, we used Kaggle and Google Colab. On Kaggle, we utilized the NVIDIA Tesla P100 GPU, equipped with 3584 CUDA cores and 16 GB of memory, complemented by an Intel Xeon 2.20 GHz CPU with 4 vCPU cores and 32 GB of memory. Additionally, through Google Colab's Pro plan, we accessed NVIDIA Tesla V100 and L4 GPUs, which offered a "High-RAM" GPU runtime with 53 GB of RAM, and 8 CPU cores.

For the training of the DeepLabV3+ model with stochastic activation functions, we implemented 80 epochs with a learning rate set at 1×10^{-4}.

In contrast, for the ResUNet++ model, we extended the training duration to 100 epochs and employed a lower learning rate of 1×10^{-6} to encourage the development of a more generalized model. This approach is particularly crucial for adapting to the complex variations typically observed in polyp segmentation, as described in [3]. We dynamically adjusted batch sizes, epochs, and learning rates based on specific needs and observed performance trade-offs.

IV. Results

A. Performance Metrics

The primary metrics used to evaluate the segmentation performance in this study are the Dice Coefficient (DSC) and the Intersection over Union (IoU). These metrics are standard for assessing the accuracy of pixel-wise segmentation tasks.

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

where X represents the predicted set of pixels and Y represents the ground truth. The Dice Coefficient measures the overlap between the prediction and the ground truth, with a perfect score of 1 indicating complete overlap and 0 indicating no overlap.

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

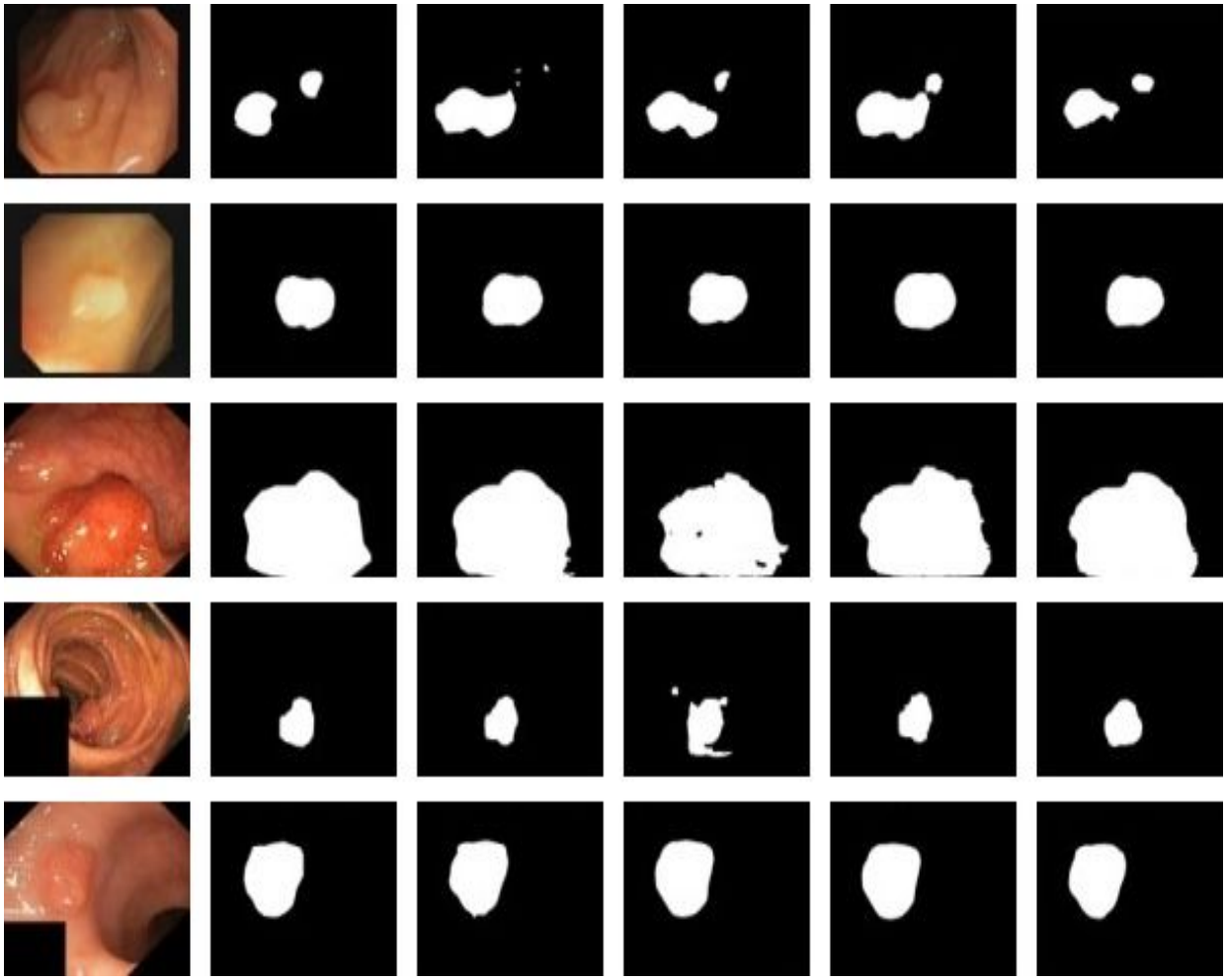

Figure 2: Qualitative results comparison on the Kvasir-SEG dataset. From the left: (1) Image, (2) Ground truth, (3) DeepLabV3+, (4) DeepLabV3+_ran, (5) ResUNet++, and (6) ResUNet++_ran.

IoU also known as the Jaccard index, evaluates the overlap between the predicted segmentation and the ground truth over the union of these two sets. Like the Dice Coefficient, an IoU score of 1 signifies perfect agreement between the prediction and the ground truth, while a score of 0 indicates no overlap.

Model	DSC	mIoU	Precision	Recall
DeepLabV3+	0.7231	0.7063	0.6571	0.7731
DeepLabV3+_ran	0.7852	0.7669	0.7135	0.8395
ResUNet++	0.8132	0.7943	0.7389	0.8695
ResUNet++_ran	0.8347	0.8153	0.7566	0.8924

Table I: The table shows the evaluation results of all the models on the Kvasir-SEG dataset. "ran" denotes a model with random activation.

B. Discussion

The implementation of stochastic activation functions in the DeepLabV3+ and ResUNet++ models led to observable improvements in both the Dice Coefficient and Intersection over Union metrics compared to their standard counterparts, for comparison see I.

DeepLabV3+ [1] Standard Model initially achieved a Dice Coefficient of 0.7231 and mIoU of 0.7063. DeepLabV3+ with Stochastic Activations improved these metrics to a Dice Coefficient of 0.7852 and mIoU of 0.7669, indicating a substantial enhancement in model accuracy and the ability to generalize across diverse segmentation scenarios.

Similarly, for the ResUNet++ model [3], the Standard Model reported a Dice Coefficient of 0.8132 and mIoU of 0.7943. ResUNet++ with Stochastic Activations showed marked improvements, with a Dice Coefficient rising to 0.8347 and mIoU to 0.8153, demonstrating the advantages of integrating diverse activation functions in complex neural network architectures.

V. Conclusion

This project demonstrates the efficacy of integrating stochastic activation functions and advanced data augmentation techniques in the training of semantic segmentation models. The stochastic activation approach introduces variability in the activation layers, effectively mitigating the risk of overfitting by preventing the model from relying too heavily on specific features of the training data. This strategic variability enhances the models' generalization capabilities, enabling them to perform robustly on new, unseen images.

Additionally, the implementation of a comprehensive data augmentation strategy played a pivotal role in the observed improvements. By generating multiple transformations of each training image we significantly enriched the dataset. This not only expanded the quantity of training data but also introduced a greater diversity of scenarios, further training the models to recognize and segment polyps under various conditions and orientations.

References

[1] 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*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3638670

[2] M. Sun, Z. Song, X. Jiang, J. Pan, and Y. Pang, "Learning pooling for convolutional neural network," *Neurocomputing*, vol. 224, pp. 96–104, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231216312905

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

[4] L. Nanni, A. Lumini, S. Ghidoni, and G. Maguolo, "Stochastic selection of activation layers for convolutional neural networks," *Sensors*, vol. 20, no. 6, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/6/1626

[5] K. Pogorelov, K. Randel, C. Griwodz, T. de Lange, S. Eskeland, D. Johansen, C. Spampinato, D. T. Dang Nguyen, M. Lux, P. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," 06 2017. [Online]. Available: https://doi.org/10.1145/3193289

[6] S. O. Atik, M. E. Atik, and C. Ipbuker, "Comparative research on different backbone architectures of deeplabv3+ for building segmentation," *Journal of Applied Remote Sensing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248707204

[7] S. Shrestha, B. Khanal, and S. Ali, "Ensemble u-net model for efficient polyp segmentation," in *MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:235474932

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[9] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:31273727

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

Appendix

Tejas was responsible for data augmentation and the setup of the ResUNet++ model, which involved developing and applying augmentation techniques. He meticulously set up the ResUNet++ architecture, integrating advanced features such as residual blocks and attention mechanisms to optimize its performance for complex image segmentation tasks.

Nakul focused on all initial data-related tasks and the configuration of the DeepLabV3+ model. His duties included acquiring and preprocessing the Kvasir-SEG dataset. He also configured the DeepLabV3+ with ResNet50 backbone and stochastic activation functions.

Shraddha took charge of model evaluation, ensemble strategies, and the comprehensive documentation of the project. She conducted evaluations of both the models and orchestrated the ensemble strategy. Her responsibilities also included the preparation of detailed presentations and reports.