

A ResNet34–U-Net Based Framework for Accurate and Efficient Endoscopic Polyp Segmentation

¹ V.Ramya Sri, ²G. Sahitya

¹M.Tech Student (Embedded Systems), ²Assistant Professor

^{1,2}Department of Electronics and Communication Engineering

^{1,2}VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

¹ ramyakolloju1116@gmail.com, ² sahitya_g@vnrvjiet.in

Abstract: Automated polyp segmentation can support computer-aided colonoscopy by providing precise pixel-level delineation of suspicious gastrointestinal regions. This work presents an efficient ResNet34–U-Net framework with a hybrid Tversky–binary cross-entropy (BCE) loss for polyp segmentation on the Kvasir-SEG dataset. A pretrained ResNet34 encoder extracts hierarchical residual features, while a U-Net-style decoder reconstructs the segmentation mask through transposed-convolution upsampling and encoder–decoder skip connections. Input images were resized to 352×352 pixels, normalized using ImageNet statistics, and augmented using flips, rotations, and mild intensity variations. The model was trained with AdamW and evaluated using five-fold image-level cross-validation, with an internal validation subset used for early stopping and threshold selection. Across the five outer test folds, the proposed method achieved a Dice score of 0.8943 ± 0.0092 , IoU of 0.8339 ± 0.0088 , precision of 0.9177 ± 0.0145 , and recall of 0.9030 ± 0.0137 . The model contains 24.45 million trainable parameters and achieved an average inference latency of 2.2818 ± 0.0714 ms per image, corresponding to 438.58 ± 13.27 FPS on the employed GPU. These results demonstrate stable, accurate, and computationally efficient polyp segmentation using a comparatively simple residual encoder–decoder architecture. The framework therefore offers a practical foundation for efficient automated analysis of endoscopic imagery.

IndexTerms - Polyp segmentation, Kvasir-SEG, ResNet34, U-Net, Tversky loss, deep learning, medical image segmentation.

I. INTRODUCTION

Colorectal polyps are abnormal growths of tissue in the gastrointestinal tract that can be connected to the development of colorectal cancer. It is therefore crucial to identify and delineate polyps accurately during colonoscopy for the early clinical assessment. However, the size, shape, texture, illumination and boundary contrast of the polyp vary which makes its manual localization difficult. Automated polyp segmentation can help the clinicians by providing a precise localization of the suspicious regions on a pixel level.

The biomedical image segmentation field has been using deep learning to a great extent. Effective encoder–decoder architecture with skip connections was introduced that preserves spatial information while learning high level semantic features by U-Net [2]. Residual networks further enhanced deep feature extraction by adding shortcut connections of the network which allowed deeper versions of convolutional networks to be trained [3]. Schemes like UNet++ have also been shown to help with better fusion of the features of the encoder–decoder in medical image segmentation [5].

Besides network architecture, another significant factor to take into account about foreground–background imbalance is the selection of a loss function. Tversky loss can penalize the two types of error differently and has been used successfully in the context of medical image segmentation [4]. The Kvasir-SEG dataset includes 1000 polyp images acquired from the endoscopy along with the ground truth segmentation masks and it is widely used to test automated polyp segmentation algorithms [1].

In this study a hybrid Tversky–binary cross entropy loss function is used in conjunction with a ResNet34–U-Net model for gastrointestinal polyp segmentation. A ResNet34 model is used to encode hierarchical features and a U-net like decoder with skip connections to reconstruct the segmentation mask. The proposed configuration is assessed on Kvasir-SEG with fivefold image level cross validation. To measure both segmentation accuracy and computational efficiency, dice coefficient, IoU, precision, recall, inference latency and throughput are used. In this work, we aim to design a simple, accurate and efficient polyp segmentation framework without any unnecessary architectural complexity.

II. LITERATURE SURVEY

The evolution of deep learning-based polyp segmentation has evolved from the traditional encoder–decoder networks to the residual, attention-based, and transformer-based networks. To enhance the representation of features for medical image

segmentation, Jha et al. proposed ResUNet++ that incorporated Residual learning along with the encoding and decoding approach of U-Net. To improve the localization of polyp regions and their boundaries, Fang et al. proposed a selective feature aggregation network with area and boundary constraints [7]. Later, DoubleU-Net architecture is developed by Jha et al. [8] wherein two U-Net are cascaded to improve the segmentation results.

In recent years, attention mechanisms have since been extensively investigated for enhancing the localization of polyps. Fan et al. introduced PraNet that progressively refines uncertain polyp regions by using parallel partial decoding and reverse attention [9]. Jha et al. have also assessed ResUNet++ with conditional random fields and test-time augmentation, as a combination of deep segmentation, post-processing and inference-time augmentation [10]. Kim et al. proposed to add uncertainty-augmented context attention to better model ambiguous polyp areas, called UACANet [11]. Wei et al. [13] proposed to use a shallow attention network to improve the segmentation ability while reducing the depth of the network.

TABLE I. Selected Polyp Segmentation Methods

Ref.	Method	Main Idea
[6]	ResUNet++	Residual U-Net for enhanced feature extraction
[8]	DoubleU-Net	Two-stage U-Net for mask refinement
[9]	PraNet	Reverse attention for polyp localization
[11]	UACANet	Uncertainty-aware context attention
[14]	ColonFormer	Transformer-based global feature modeling
[16]	Polyp-PVT	Pyramid transformer for multi-scale segmentation
Proposed	ResNet34–U-Net + Tversky–BCE	Residual encoder with U-Net decoder and overlap-aware loss

In recent years, more and more studies have been using transformer-based feature learning. To leverage local spatial details and global contextual information, Zhang et al. proposed TransFuse, which is based on convolutional networks and transformers [12]. Duc et al. proposed an efficient transformer-based framework, called ColonFormer, for colorectal polyp segmentation [14]. Lou et al. introduced CaraNet that combines contextual and axial reverse attention, and is specifically designed to segment small medical objects [15]. Dong et al. proposed Polyp-PVT using pyramid vision transformers to represent polyp at multiple scales [16]. More recently, Wang et al. fused the nested U-Net with transformer based multi-scale feature modelling for colorectal polyp segmentation [17].

Despite their good segmentation performance, many of these methods involve the use of attention blocks, multiple network stages, transformer blocks, or complicated feature fusion schemes. These designs can add complexity to the model and the number of calculations needed. Hence, there is still an opportunity for comparatively simple approaches with still efficient feature extraction and inference. Encouraged by this insight, the work presented here uses a pre-trained ResNet34 encoder and a U-Net-like decoder, as well as a hybrid Tversky–BCE loss function, striving to balance segmentation accuracy, simplicity of training, and efficiency of computation.

III. METHODOLOGY

A. Dataset and Preprocessing

The proposed polyp segmentation framework was evaluated using the Kvasir-SEG dataset, which contains 1,000 gastrointestinal endoscopic images with corresponding pixel-level segmentation masks. All image–mask pairs were verified before training to ensure correct correspondence.

The RGB images were scaled to 352×352 pixels using bilinear interpolation, and the segmentation masks were scaled using nearest-neighbour interpolation to maintain the boundaries of the masks. The masks were then transformed into binary format – foreground pixels corresponding to the polyp region and background pixels to the non-polyp regions. The images were normalized using the mean and standard deviation of ImageNet.

To enhance the generalization ability of the model, paired data augmentation was used in the training process to prevent overfitting. Both image and mask were transformed by the same geometric transformations. The augmentations were horizontal flipping, vertical flipping, 90-, 180-and 270-degree rotations and minor changes in brightness and contrast.

B. Proposed ResNet34–U-Net Architecture

The proposed model is composed of a pretrained encoder (ResNet34) and a decoder of U-Net type. The encoder consists of several residual blocks, which are used to learn the hierarchical visual features from the input endoscopic image. Residual learning is adopted to help extracting more features without compromising the stability of optimization.

Transposed convolutions decode the features progressively bringing up the spatial resolution. High-level semantic information is fused with fine spatial details by using skip connections between the similar layer of the encoder and decoder. This can help maintain the boundaries of the polyps and enhance the accuracy of localization.

The number of feature channels is gradually decreased along the decoding pathway and the resolution of the image is incrementally recovered. The final 1×1 convolution generates a one channel segmentation output which shows the probability that every pixel is part of the polyp region.

The overall workflow can be represented as

$$I \rightarrow E_{ResNet34} \rightarrow D_{U-Net} \rightarrow P$$

where I denotes the input image, $E_{ResNet34}$ represents the ResNet34 encoder, D_{U-Net} denotes the U-Net decoder, and P represents the predicted polyp probability map. The overall ResNet34–U-Net segmentation architecture is illustrated in Fig. 1.

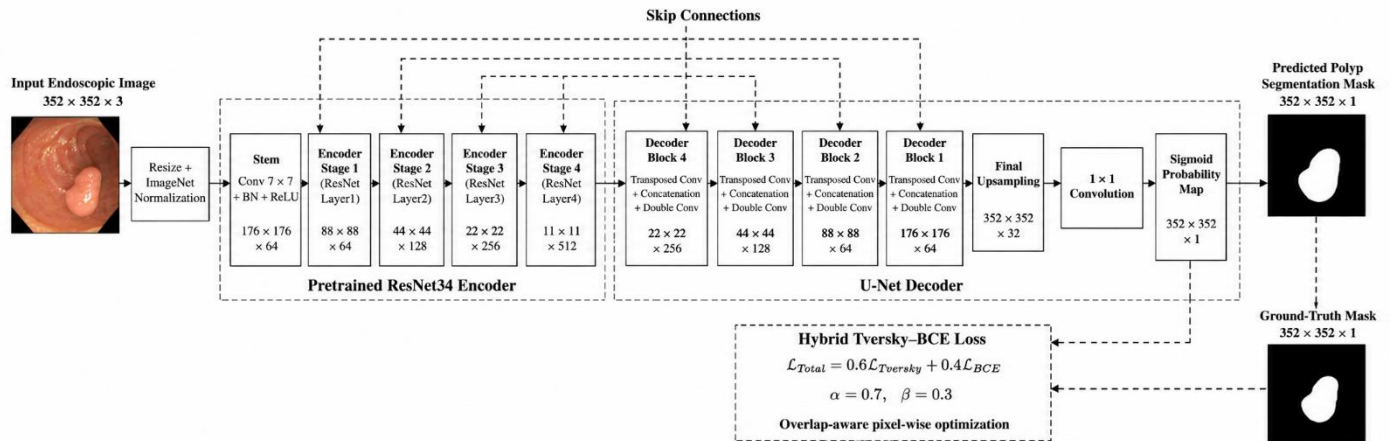


Fig. 1. Proposed ResNet34–U-Net architecture for polyp segmentation.

C. Hybrid Tversky–BCE Loss

To address foreground–background imbalance and improve segmentation overlap, a hybrid loss function combining Tversky loss and binary cross-entropy was used.

The Tversky index is defined as

$$TI = \frac{TP + \epsilon}{TP + \alpha FP + \beta FN + \epsilon}$$

where TP , FP , and FN represent true-positive, false-positive, and false-negative pixels, respectively, and ϵ is a small smoothing constant.

The corresponding Tversky loss is defined as

$$L_T = 1 - TI$$

The final loss function is expressed as

$$L_{Total} = 0.6L_T + 0.4L_{BCE}$$

with

$$\alpha = 0.7, \beta = 0.3$$

The false positive pixels have higher weight which pushes the model to inhibit unwanted foreground predictions. The BCE part offers pixel-level supervision and the Tversky part boosts the regional overlap between the predicted and ground-truth masks.

D. Training Strategy

The model was trained using the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4} . The batch size was set to 16, and the maximum number of training epochs was fixed at 50.

To stabilize training, gradient clipping was done with a maximum norm of 1.0. During the GPU execution, mixed-precision training was applied to reduce the computation burden and memory consumption of the training. A ReduceLROnPlateau scheduler was used to decrease the learning rate when there was no further improvement in the validation Dice score. To avoid unnecessary training and overfitting, early stopping with patience value of 10 epochs was also implemented.

E. Five-Fold Cross-Validation

Five-fold image level cross validation was carried out to get a reliable estimation of the segmentation performance. A total of 200 images were used as the outer test set in each fold. The remaining 800 images were then used for training (720) and internal validation (80).

A model selection, early stopping and segmentation-threshold optimization were performed using the internal validation subset. The outer test set was kept separate during training and validation and was used at the end for evaluation.

The internal validation set was used to evaluate a range of candidate segmentation thresholds that started at 0.30 and ended at 0.70 with a step size of 0.05. The chosen threshold that achieved the maximum Dice validation score was then used to predict on the respective outer test fold.

F. Evaluation Metrics

The performance of the proposed model was done by using Dice coefficient, Intersection over Union, precision, and recall.

The Dice coefficient is equal to

$$Dice = \frac{2TP}{2TP + FP + FN}$$

The Intersection over Union is calculated as

$$IoU = \frac{TP}{TP + FP + FN}$$

Precision is calculated as

$$Precision = \frac{TP}{TP + FP}$$

and recall is calculated as

$$Recall = \frac{TP}{TP + FN}$$

The mean $\pm$ standard deviation of the five outer test folds is reported as the final segmentation performance. Besides segmentation accuracy, the computational efficiency was also assessed, including the number of model parameters and average inference latency per image and frames per second. This allowed to evaluate the proposed framework both in terms of segmentation performance and computational efficiency.

IV. RESULTS AND DISCUSSION

The proposed ResNet34–U-Net framework was evaluated on the Kvasir-SEG dataset using five-fold image-level cross-validation. The model was trained with 720 images, validated with 80 images and independently tested with 200 held-out images for each fold. Only the appropriate inner fold of validation was used to select the segmentation threshold, prior to testing on the outer fold of validation. The quantitative results are shown in Table II in terms of fold.

TABLE II. FIVE-FOLD SEGMENTATION RESULTS

Fold	Dice	IoU	Precision	Recall	Threshold
1	0.9020	0.8413	0.9187	0.9085	0.30
2	0.9021	0.8398	0.9055	0.9245	0.30
3	0.8942	0.8343	0.9346	0.8910	0.40
4	0.8936	0.8347	0.9289	0.8951	0.30
5	0.8795	0.8192	0.9010	0.8958	0.30
Mean $\pm$ SD	0.8943 $\pm$ 0.0092	0.8339 $\pm$ 0.0088	0.9177 $\pm$ 0.0145	0.9030 $\pm$ 0.0137	—

The mean Dice coefficient and IoU of the proposed framework over the five outer test folds is 0.8943 ± 0.0092 and 0.8339 ± 0.0088 , respectively. Precision and recall reached 0.9177 ± 0.0145 and 0.9030 ± 0.0137 , respectively. The results show that the

model is producing a balance segmentation response, that is, high agreement with the ground truth polyp regions and high discrimination between the polyp and background pixels.

The fold-wise Dice performance is illustrated in Fig. 2, where all five outer folds achieved comparable segmentation accuracy, with a mean Dice coefficient of 0.8943. The relatively small variation across folds indicates consistent model performance across different test subsets.

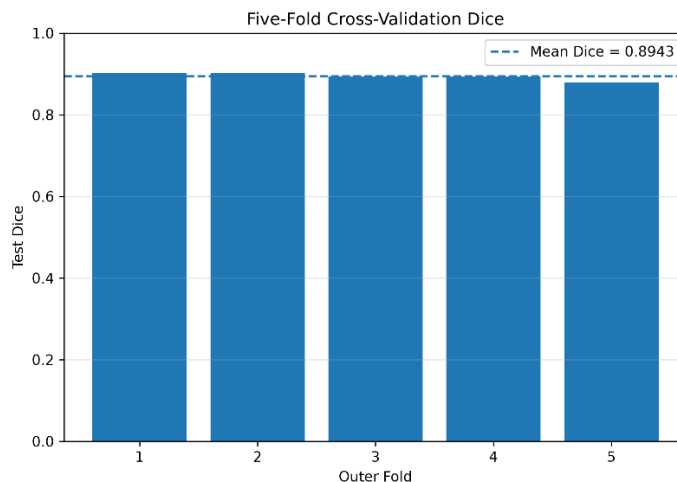


Fig. 2. Dice performance across the five outer cross-validation folds.

The distribution of out-of-fold Dice, IoU, precision, and recall scores across the complete dataset is presented in Fig. 3, providing additional insight into the consistency and variability of the segmentation performance.

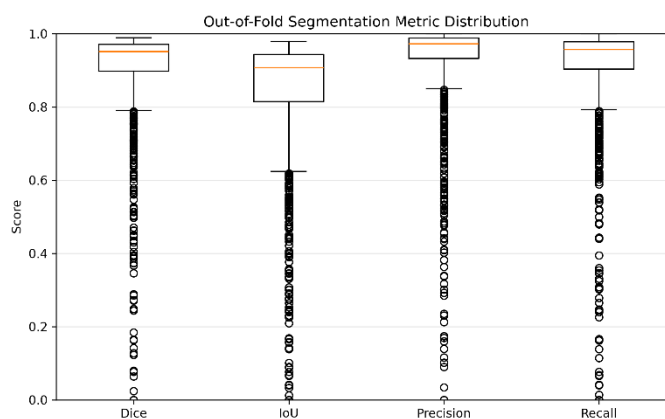


Fig. 3. Distribution of out-of-fold segmentation metrics across Kvasir-SEG.

The fold-wise results also do not exhibit significant variations between different subsets of the data. Fold 2 was the highest with a Dice score of 0.9021, followed by Fold 1 with a Dice score of 0.9020. The highest precision (0.9346) is obtained by Fold 3 while the highest recall (0.9245) is obtained by Fold 2. The held-out subsets were found to have varying levels of difficulty of segmentation, with the smallest Dice value of 0.8795 reported by Fold 5. However, all five folds had relatively high precision and recall, and the low standard deviations for Dice and IoU suggest that the reported performance was not overly sensitive to a good train-test split.

The validation Dice trajectories across the five folds are shown in Fig. 4. All folds exhibited rapid improvement during the initial training epochs followed by stable convergence, indicating consistent learning behaviour across the cross-validation splits.

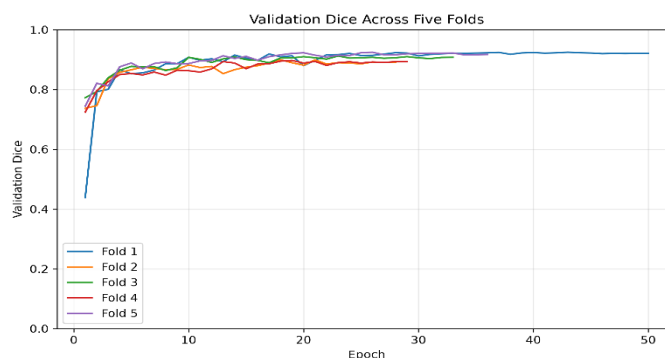


Fig. 4. Validation Dice convergence across the five cross-validation folds.

A significant remark is that the proposed model always achieves a very high accuracy. In this case, the hybrid Tversky–BCE objective emphasizes false positive errors with $\alpha=0.7$, while the latter emphasizes false negative errors with $\beta=0.3$. The precision found of 0.9177 and recall of 0.9030 was found to be consistent with the optimisation behaviour that was intended, where too many background regions of the image surrounding the polyp region are not predicted, but enough background regions are being predicted further away from the polyp region to maintain a high recall. This is especially helpful for endoscopic segmentation, in which a higher number of segments of mucosal structures might compromise the accuracy of the predicted polyp boundaries.

The segmentation threshold had an impact on the resulting binary prediction as well. Four of the five folds chose a threshold of 0.30 and Fold 3 chose 0.40. The proposed evaluation strategy didn't fix the threshold at 0.50 for all the trained models. This enabled the final decision boundary to adjust to the probability distribution generated at each fold, and maintained the outer test images completely separated from the threshold optimisation process. The similarity of the chosen value for the four folds also indicates that a rather low decision value was usually beneficial for this training configuration.

The effect of the segmentation threshold on validation Dice is shown in Fig. 5, where a threshold of 0.30 provided the best mean validation performance across the five folds.

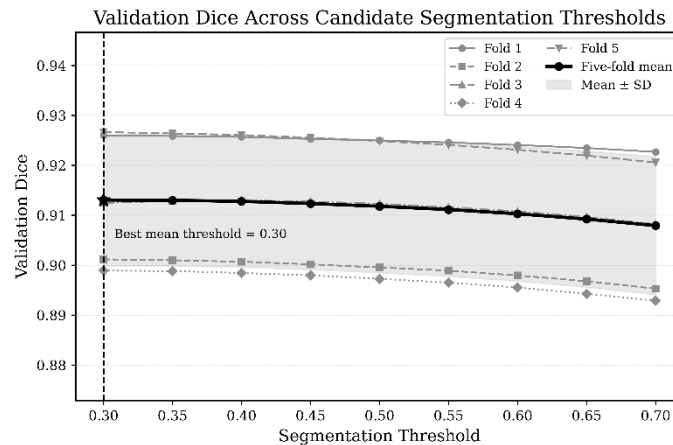


Fig. 5. Validation Dice across candidate segmentation thresholds.

A. Computational Efficiency

In addition to segmentation accuracy, computational efficiency was evaluated to determine whether the proposed architecture could support high-speed inference. There are 24,451,585 trainable parameters in the model. The single-image mean inference time was: 2.2818 ± 0.0714 ms, representing a mean throughput of 438.5753 ± 13.2712 FPS in the used GPU benchmark setting.

TABLE III. OVERALL PERFORMANCE AND COMPUTATIONAL CHARACTERISTICS

Measure	Proposed Method
Dice	0.8943 ± 0.0092
IoU	0.8339 ± 0.0088
Precision	0.9177 ± 0.0145
Recall	0.9030 ± 0.0137
Parameters	24.45 M
Latency	2.2818 ± 0.0714 ms/image
Throughput	438.58 ± 13.27 FPS

This low model-inference latency shows that the residual encoder and U-Net decoder can be used to do fast segmentation on GPU hardware without the need of attention layers and transformer blocks. The reported FPS, however, is of a batch-size-one model inference with the experimental GPU setup and does not necessarily reflect the throughput of a full clinical system, as the timing measurement was performed without considering image acquisition, pre-processing, display, and other end-to-end processing components. It is important to make this difference when looking at the real-time capabilities of the framework.

B. Qualitative Analysis

Quantitative overlap metrics give an overall idea of the accuracy of the segmentation but visual analysis is needed to check how boundaries were localised and how the prediction behaviour was. Best, median and difficult prediction cases were then identified and representative examples of out-of-fold segmentation were selected. Qualitative examples include the original endoscopic image, the ground truth polyp annotation and the predicted polyp mask. The qualitative results depicted in Fig. 3 illustrate the behavior of the model for various difficulties of the segmentation task. The resulting mask in well-segmented examples is able to closely

approximate the shape and size of the annotated polyp area. In median cases there are slight differences in the boundaries but the main outline of the polyp is preserved. For difficult cases, the boundaries of prediction and reference are more different, reflecting the decrease in Dice and IoU for difficult samples. While the quantitative results are presented, the examples give a glimpse of the qualitative results and demonstrate that the errors with which these regions are identified are not simply image-level detection errors, but rather errors in the delineation of the region.

Representative segmentation results for difficult, median, and best-performing test samples are illustrated in Fig. 6, showing the input endoscopic image, corresponding ground-truth mask, and predicted segmentation mask.

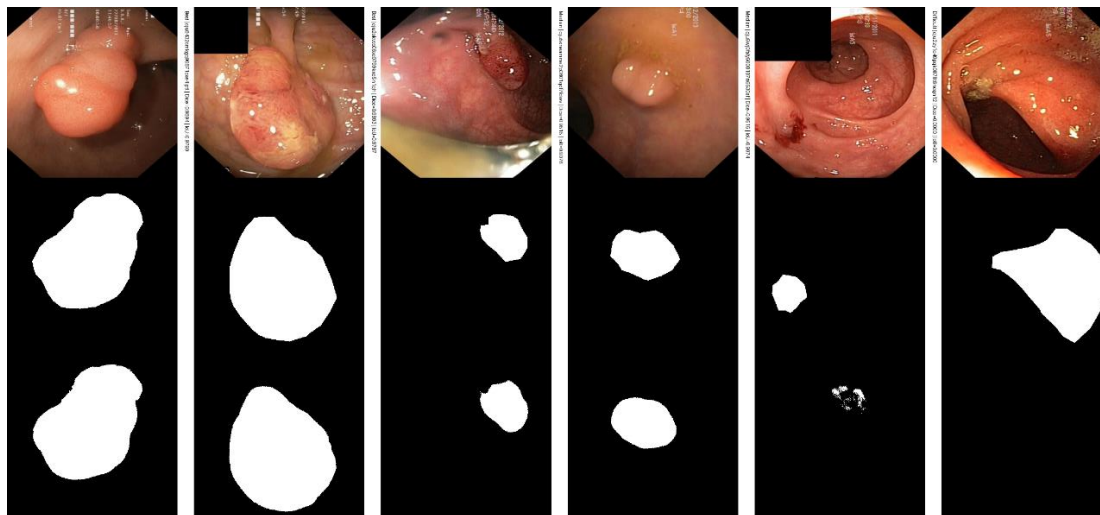


Fig. 6. Qualitative polyp segmentation results for representative best, median, and difficult cases.

C. Discussion

The experimental results demonstrate that the proposed residual encoder–decoder framework is able to achieve accurate polyp segmentation with a relatively simple architecture, even without the addition of transformer and complex attention module. ResNet34 is used for hierarchical feature extraction with the corresponding residual blocks in the network and the U-Net decoder adds back the spatial information through encoder–decoder skip connections. This architecture is further enhanced by the hybrid Tversky–BCE objective, which optimises for overlap among regions and for supervision at the pixel-level.

Recently, the polyp segmentation methods PraNet [9] and UACANet [11] use specialised attention mechanism and the methods ColonFormer [14] and Polyp-PVT [16] use transformer-based global representation. This kind of approaches prove the success of more and more advanced feature modelling. Instead, in this paper, we aim to achieve a feasible tradeoff between segmentation performance and efficiency with the same decoder and backbone, which is a standard Residual CNN. Hence, the added value of the proposed framework needs to be understood more in the light of simplicity, stable performance across cross-validation, overlap-aware optimisation, and high inference efficiency, than just novelty in architecture.

As well, caution needs to be used when comparing with previously published methods of direct numerical comparison. The different training-test splits, image sizes, augmentation techniques, external training sets, and evaluation procedures of published works using Kvasir-SEG introduce variability and differences. For instance, the multi-scale nested U-Net–Transformer study in [17] used a 880/120 split on Kvasir-SEG and achieved a Dice score of 0.942 under this split. Other studies take the trouble to merge Kvasir-SEG and CVC-ClinicDB before testing. The current study, on the other hand, presents the mean and standard deviation for five different outer test folds of the entire Kvasir-SEG data set. Therefore, the values are not considered as absolute one-to-one rankings without repeating of the same experimental protocol.

Other strengths of the evaluation are that no images are reused in any outer test folds and that early stopping and threshold selection are done only on internal validation data and not on the outer test folds. This takes away the reliance on a specific random test set, and is a more meaningful way of estimating within data-set generalisation than reporting one good split. The performance of the segmentation also shows variation in Fold 5, suggesting that the quality of segmentation performance is dependent upon the content of the test data and that the performance of the segmentation on challenging polyp appearances is still an area of potential improvement.

In total, the results show that the ResNet34–U-Net with Tversky–BCE optimisation offers a stable and efficient framework for polyp segmentation with an average measured inference latency of about 2.28 ms per image, a 5-fold Dice coefficient of 0.8943 ± 0.0092 , IoU of 0.8339 ± 0.0088 , precision of 0.9177 ± 0.0145 , and recall of 0.9030 ± 0.0137 . The results suggest that in cases where segmentation accuracy, model simplicity and computational efficiency are desired, the use of residual encoder–decoder networks can be an effective alternative.

VII. CONCLUSION

In this work, an efficient ResNet34–U-Net based segmentation framework with a hybrid Tversky–BCE loss function for segmenting gastrointestinal polyps from the Kvasir-SEG dataset was presented. Using five-fold image-level cross-validation, the model achieved a mean Dice score of 0.8943 ± 0.0092 , IoU of 0.8339 ± 0.0088 , precision of 0.9177 ± 0.0145 , and recall of 0.9030 ± 0.0137 . The framework also showed low inference latency and high throughput, which are indicative of good computational efficiency. In conclusion, we find that using a residual encoder–decoder framework can achieve stable and accurate polyp segmentation despite its relatively simple design, eliminating the need for specific and complex attention or transformer modules. In the future, efforts could be directed on the improvement of the framework's performance in difficult cases and the validation of the framework on other endoscopic datasets.

VIII. FUTURE SCOPE

Future work can focus on deploying the proposed polyp segmentation model on embedded edge devices such as NVIDIA Jetson platforms or Raspberry Pi-based accelerators for real-time endoscopic assistance. Model compression techniques such as quantization and pruning can be explored to reduce memory usage and improve inference speed. The framework can also be validated on additional polyp datasets and real clinical endoscopy videos to evaluate its robustness and practical applicability.

REFERENCES

- [1] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *MultiMedia Modeling*, Lecture Notes in Computer Science, vol. 11962, Springer, Cham, Switzerland, 2020, pp. 451–462, doi: 10.1007/978-3-030-37734-2_37.
- [2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Lecture Notes in Computer Science, vol. 9351, Springer, Cham, Switzerland, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [4] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, vol. 10541, Springer, Cham, Switzerland, 2017, pp. 379–387, doi: 10.1007/978-3-319-67389-9_44.
- [5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science, vol. 11045, Springer, Cham, Switzerland, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
- [6] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, 2019, pp. 225–230, doi: 10.1109/ISM46123.2019.00049.
- [7] Y. Fang, C. Chen, Y. Yuan, and K.-Y. Tong, "Selective feature aggregation network with area-boundary constraints for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*, Lecture Notes in Computer Science, vol. 11764, Springer, Cham, Switzerland, 2019, pp. 302–310, doi: 10.1007/978-3-030-32239-7_34.
- [8] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, 2020, pp. 558–564, doi: 10.1109/CBMS49503.2020.00111.
- [9] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "PraNet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*, Lecture Notes in Computer Science, vol. 12266, Springer, Cham, Switzerland, 2020, pp. 263–273, doi: 10.1007/978-3-030-59725-2_26.
- [10] D. Jha, P. H. Smedsrud, D. Johansen, T. de Lange, H. D. Johansen, P. Halvorsen, and M. A. Riegler, "A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2029–2040, Jun. 2021, doi: 10.1109/JBHI.2021.3049304.
- [11] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty augmented context attention for polyp segmentation," in *Proc. 29th ACM Int. Conf. Multimedia (MM '21)*, 2021, pp. 2167–2175, doi: 10.1145/3474085.3475375.
- [12] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, Lecture Notes in Computer Science, vol. 12901, Springer, Cham, Switzerland, 2021, pp. 14–24, doi: 10.1007/978-3-030-87193-2_2.
- [13] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*, Lecture Notes in Computer Science, vol. 12901, Springer, Cham, Switzerland, 2021, pp. 699–708, doi: 10.1007/978-3-030-87193-2_66.
- [14] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and D. V. Sang, "ColonFormer: An efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80575–80586, 2022, doi: 10.1109/ACCESS.2022.3195241.
- [15] A. Lou, S. Guan, and M. Loew, "CaraNet: Context axial reverse attention network for segmentation of small medical objects," *J. Med. Imag.*, vol. 10, no. 1, Art. no. 014005, 2023, doi: 10.1117/1.JMI.10.1.014005.
- [16] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," *CAAI Artificial Intelligence Research*, vol. 2, Art. no. 9150015, 2023, doi: 10.26599/AIR.2023.9150015.

- [17] Z. Wang, Z. Liu, J. Yu, Y. Gao, and M. Liu, "Multi-scale nested UNet with transformer for colorectal polyp segmentation," *J. Appl. Clin. Med. Phys.*, vol. 25, no. 6, Art. no. e14351, Jun. 2024, doi: 10.1002/acm2.14351.



Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.