Polyp Segmentation in Colonoscopy Images

Marcio P. Ferreira(✉), Giulia de A. Freulon, Daniel G. Piorsky, Alexandre C. P. Pessoa, Darlan B. P. Quintanilha, and Aristófanes C. Silva

Applied Computing Group (NCA - UFMA), Federal University of Maranhão, São Luís, MA, Brazil
{marcio.ferreira,alexandre.pessoa,dquintanilha,ari}@nca.ufma.br,
{giulia.freulon,daniel.piorky}@discente.ufma.br

Abstract. Colorectal cancer is a prevalent form of cancer, often detectable through polyps in the gastrointestinal tract. Unfortunately, these polyps typically do not display noticeable symptoms, making early detection challenging. While procedures like colonoscopy and endoscopy can identify polyps, they can miss some, leading to the need for a more automated approach. One innovative solution is capsule endoscopy, which records detailed images of the gastrointestinal tract over an extended period. However, the massive volume of data generated necessitates automation for efficient analysis. Artificial intelligence, particularly convolutional neural networks (CNNs) like TransUNet, can be crucial in quickly and accurately identifying suspicious areas in capsule endoscopy images. This study focuses on automating polyp detection using TransUNet and aims to enhance the early detection of colorectal cancer. The research utilizes the Kvasir-SEG database, containing polyp images and annotated segmentation masks. Various CNN architectures, like UNet, ResUNet, and ResUNet++, are employed, with metrics like Dice Loss and Tversky Loss used for performance evaluation through techniques like cross-validation. Results demonstrate that the TransUNet approach, leveraging transformers in its encoding layers, achieved 66% Dice Score, outperforming other architectures like UNet and ResUNet in this metric, however it did not surpass the ResUNet++ network. In conclusion, the TransUNet model shows potential for automating polyp detection in gastrointestinal images, offering a valuable tool in the fight against colorectal cancer. Integrating advanced technology into medicine promises more accurate and efficient gastrointestinal care.

Keywords: Video Capsule Endoscopy · Transformers · Polyp Segmentation

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 185–194, 2024.
https://doi.org/10.1007/978-3-031-60665-6_14

1 Introduction

Colorectal cancer, a disease of significant impact, ranks second most prevalent cancer among women and the third most common among men [6]. In its early stages, this form of cancer often originates as polyps, abnormal tissue growths on the mucous membrane lining the interior of the digestive system. Sadly, these polyps tend not to exhibit discernible symptoms, complicating their detection. However, screening procedures such as colonoscopy and endoscopy can usually identify these formations. It is concerning to note that studies point to a tendency to overlook the presence of polyps during these examinations, with detection failure rates ranging from 14 to 30%, depending on the size of the polyp [6].

Colonoscopy and endoscopy are highly invasive examinations that can potentially lead to undesirable side effects such as nausea, vomiting, intestinal pain, and even minor bleeding. These effects result from the procedure involving the endoscope. An alternative to these approaches is the adoption of capsule endoscopes, which incorporate cameras inside them. Patients swallow these capsules, and they naturally traverse the entire path that would be explored by the endoscope. During this journey, the capsule continuously records images and videos of the digestive tract, which are transmitted to a receiving device attached to the patient's waist throughout the procedure.

The advantages of using capsules in examinations are that they avoid the side effects commonly associated with colonoscopy and endoscopy procedures, and they are considerably less invasive. However, a significant drawback of using capsules is that the procedure generates a large number of images or a very long video since the capsule naturally travels through the entire digestive tract, which takes around 8 to 12 h [4].

Given the general challenge of detection error rates and the need to analyze a large amount of material from capsule examinations, one solution is the automation of polyp detection, such as the use of image segmentation techniques. Image segmentation is a technique that aims to highlight an area of interest in an image, with the area, in this case, being the polyp itself.

Improving the efficiency of polyp detection implies a decrease in colorectal cancer rates. Thus, finding ways to enhance the identification rate of these structures and automate this process emerges as an achievement of immeasurable value for people's health and quality of life. An approach with potential for this automation lies in the application of computer vision techniques, particularly in the field of image segmentation. Image segmentation, a sophisticated technique, aims to isolate relevant information within an image in this context, explicitly highlighting the area of interest of the polyps. This approach facilitates the interpretation of results and signals a promising path for optimizing the early detection of these anomalies [6]. The objective of this study is to evaluate the effectiveness of computer vision and neural network approaches in automating the process of detecting polyps in images of the gastrointestinal system. Through extensive testing and analysis, this research aims to contribute valuable insights into the performance, strengths and limitations of the proposed methods. The following sections will delve deeper into the experimental setup,

dataset, and evaluation of results, clarifying practical implications and advances achieved through comprehensive testing.

2 Related Work

Research involving technologies such as machine learning and CNN for medical image analysis is widely conducted by researchers from various public and private institutions. Given this large number of researchers, it is possible to find a wide variety of research and different techniques in the field. In Table 1, a selection of related works and their respective techniques are presented.

Table 1. Related works with datasets and techniques used.

Work	Dice	Dataset	Technique
JHA et al. [7]	81.33%	Kvasir-SEG	ResUNet and modifications
TOMAR et a. [10]	85.76%	Kvasir-SEG	DDANet
SRIVASTAVA et al. [8]	92.17%	Kvasir-SEG	MSRF-Net
YEUNG et al. [11]	91%	Kvasir-SEG	Focus U-Net
JHA et al. [5]	92.93%	CVC-ClinicDB	ResUNet++ with CRF and TTA

In a study conducted by [7], ResUNet and its enhanced variations were employed for the segmentation of polyps in the Kvasir-SEG dataset. In this context, the more refined iteration of the network, known as ResUNet++, achieved an impressive Dice score of 81.33%.

DDANet is an architecture based on a dual-attention decoder, and in experiments conducted by [10], it achieved a remarkable Dice score of 85.76% when applied to the Kvasir-SEG dataset.

In the work by [8], the MSRF-Net was introduced as an innovative approach to medical image segmentation in [8], utilizing DSDF (Dual-Scale Dense Fusion) blocks. The results demonstrated a Dice score of 92.17% when applied to the Kvasir-SEG dataset.

The Focus U-Net [11], a network that incorporates the Focus Gate, a model that combines channel-based and spatial-based attention mechanisms. This approach achieved a Dice score of 91% when evaluated on the Kvasir-SEG dataset.

A study conducted by [5] employed the ResUNet++ architecture and improved its performance through the integration of techniques such as Conditional Random Field (CRF) and Test-Time Augmentation (TTA). As a result of these optimizations, a remarkable Dice score of 92.93% was achieved on the CVC-ClinicDB dataset using ResUNet++ in conjunction with the CRF approach.

This work contributes to tests of Kvasir-SEG on TransUNet Network and several tests carried out with this network, and comparison of its performance with UNet and other networks used in medical image segmentation.

3 Materials and Methods

3.1 Dataset

The dataset used was Kvasir-SEG. This dataset consists of 1,000 images of polyps and their respective masks annotated by professionals in the field. The images range from 332×487 to 1920×1072 pixels. Kvasir-SEG was created to be a public dataset that researchers could use in their studies, as obtaining such a dataset for research purposes is challenging.

3.2 Pre-processing

Initially, tests were conducted using the Kvasir-SEG dataset, using the images at their original resolutions. Subsequently, new experiments were carried out after resizing the images to a resolution of 512×512 pixels while preserving their proportions. Based on these initial tests, superior performance was observed with the resized images. Based on these results, it was decided to adopt the resized images as the standard for the subsequent research steps.

3.3 Method

UNet. UNet is a Convolutional Neural Network (CNN) originally developed for medical image segmentation. Its main objective was to achieve highly accurate segmentation while efficiently using computational resources. The term "UNet" derives from its "U"-shaped architecture. This characteristic design of the architecture is responsible for its nomenclature.

The UNet architecture can be divided into two main parts: the downsampling layers and the upsampling layers. In the downsampling phase, the input image goes through a series of operations. First, consecutive convolutions with 3×3 filters are applied, followed by Rectified Linear Activation (ReLU) activation, which solves gradient problems by mapping positive values to themselves and negative values to zero. Next, the downsampling layers incorporate Max Pooling operations with 2×2 filters. Max Pooling is a common technique in CNNs to reduce the dimensionality of feature maps.

In the upsampling layers, the outputs of the downsampling layers go through a sequence of Upsampling operations with 2×2 filters. At each Upsampling step, the output is concatenated with the corresponding output from the downsampling layer. Finally, the last layer performs a convolution with a 1×1 filter, which maps the image's features. This final layer is followed by a sigmoid function, which generates a binary segmentation mask.

TransUNet. TransUNet is a neural network architecture based on transformers, designed to overcome the limitations inherent to CNNs. CNNs often exhibit poor performance on target structures that display wide inter-patient variations in terms of texture, shape, and size due to the nature of multiple convolution

operations. While convolutions are effective at identifying local features, they do not excel at capturing long-range information or global context.

As a solution to this challenge, [1] proposed incorporating self-attention mechanisms, akin to transformers. By adopting this approach, TransUNet aims to address the need to capture long-range relationships between image elements effectively, allowing for a more effective understanding of contextual and global information. This contributes to improving the segmentation and understanding of complex and variable structures among different medical images.

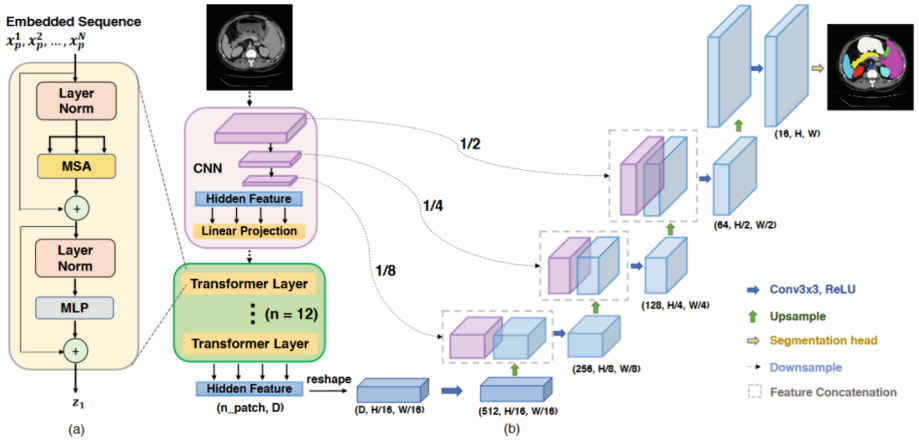

Fig. 1. TransUNet structure [10].

As we can see in Fig. 1, TransUNet adopts a structure similar to U-Net, however, its fundamental distinction lies in the addition of transformer layers, which are not present in the conventional UNet. These transformer layers can establish long-range connections between different regions of the image.

The operation of TransUNet consists of the image initially passing through convolution layers to extract relevant features from the input image. Subsequently, these features are passed to the transformer layers, whose purpose is to capture relationships that extend over a significant distance between different regions of the image.

Within the transformer layers, the following operations take place:

- Layer Normalization: Normalizes input values for each channel of the layer, which helps regularize and normalize activations, making the learning process more stable;
- Multi-Head Self-Attention (MSA): Applies multiple self-attention, where the input is projected into various attention dimensions, allowing the network to focus on relevant information at different positions and scales. This allows the network to capture long-range relationships and broader contexts;

- Layer Normalization: A second layer of normalization is applied after the self-attention operation for the output of that layer;
- Multi-Layer Perceptron (MLP): A layer for learning complex relationships between input features. This helps transform and merge information obtained from the previous operation, allowing the network to capture non-linear relationships.

After obtaining the hidden features from the output of the transformer layers, they are passed through a set of Upsampling layers, which expand the resolution of the feature maps. Then, the features are concatenated with the corresponding outputs from the attention layers in the Encoder. This ultimately leads to image segmentation at the network's output.

Loss Function. Due to the common class imbalance in medical image segmentation, experiments were conducted with loss functions in the UNet and TransUNet architectures, namely, Dice loss [2] and Tversky loss [9]. The Dice loss assigns equal weights to penalize false positives and false negatives evenly. In contrast, the Tversky loss is an extension of the Dice loss that allows for the assignment of different weights to false positives and false negatives, enabling fine-tuning between an emphasis on precision (reduction of false positives) and recall (reduction of false negatives). The formula for the Tversky loss is as follows:

$$T(\alpha, \beta) = 1 - \frac{\sum_{i,j} p_{ij} \times t_{ij}}{\sum_{i,j} p_{ij} \times t_{ij} + \alpha \times \sum_{i,j} p_{ij} \times (1 - t_{ij}) + \beta \times \sum_{i,j} (1 - p_{ij}) \times t_{ij}}$$

where p_{ij} represents the model's prediction probability for the pixel or element at position (i, j), and t_{ij} is the true (label) value of the pixel or element at position (i, j). The parameters α and β adjust the weight assigned to false positives and false negatives, respectively.

In the experiments with both the UNet and TransUNet architectures, the Tversky loss proved to be more effective.

3.4 Evaluation of Results

The evaluation metrics used in the proposed method were Dice score, Precision, Recall, and Mean Intersection over Union (mIOU). The calculations for these metrics were based on the following formulas:

- Dice Score = 2 * (Precision * Recall)/(Precision + Recall)
- Precision = TP/(TP + FP)
- Recall = TP/(TP + FN)
- mIOU = TP/(TP + FP + FN)

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives [3].

4 Results

In this section, the results of the performance evaluation of segmentation on the Kvasir-SEG dataset (resized while maintaining proportions) using the TransUNet network are presented. A comparison is conducted between the ADAM (Adaptive Moment Estimation) and SGD (Stochastic Gradient Descent) optimizers in various configurations.

Initially, hyperparameter optimization tests were conducted with HyperOpt library. These tests consisted of 20 evaluations, each involving training for 100 epochs using the Tversky Loss function. The data split for training was 70% for training, 15% for validation, and 15% for testing.

Based on the results obtained in this optimization phase, a new training was conducted using these values. The initial tests revealed better performance with the SGD optimizer compared to ADAM. Table 2 summarizes the performance metric results for each optimizer.

Table 2. Comparative Test between ADAM and SGD optimizers in the TransUNet network.

Optimizer	Dice	mIOU	Recall	Precision
ADAM	0.4058	**0.6346**	**0.7236**	0.2820
SGD	**0.7020**	0.4358	0.6519	**0.7605**

Based on this preliminary analysis, the choice of the SGD optimizer was made to proceed with further tests. The next step involved conducting a cross-validation with 5 folds, each containing 200 images. In each cross-validation run, 3 folds were used for training, 1 for validation, and 1 for testing, ensuring that each fold was used once for testing and once for validation. The configurations obtained in the hyperparameter optimization stage for the SGD optimizer were used. The results of this process can be seen in Table 3.

Table 3. Cross-validation in TransUNet.

Fold	Dice	mIOU	Recall	Precision
1	**0.7217**	**0.4805**	**0.7212**	0.7222
2	0.6364	0.3521	0.6530	0.6207
3	0.6795	0.3047	0.6397	0.7245
4	0.6844	0.4752	0.6560	0.7153
5	0.6191	0.1189	0.5317	**0.7409**
Average	0.6682	0.3467	0.6403	0.7047
Standard Deviation	0.0412	0.1327	0.0700	0.0469

Additionally, cross-validation was conducted with the UNet network after preliminary hyperparameter tuning to make a comparison between the methods. The results of this evaluation are summarized in Table 4.

Table 4. Cross-validation in UNet.

Fold	Dice	mIOU	Recall	Precision
1	0.5639	0.0571	0.5649	0.5785
2	0.5836	0.0576	0.5424	0.6584
3	**0.6190**	0.0579	0.5784	**0.6797**
4	0.5439	0.0559	**0.6207**	0.5042
5	0.5530	**0.0580**	0.5218	0.5915
Average	0.5726	0.0573	0.5656	0.6026
Standard Deviation	0.0275	0.0007	0.0363	0.0615

When comparing the results, it becomes evident that the utilization of transformers in the encoding (encoder) phase of the TransUNet network demonstrates superiority over the UNet approach.

Furthermore, to have a more robust comparison we performed a cross-validation of the ResUNet and ResUNet++ networks from the work of Jha et al. [7] for comparison purposes only, following preliminary hyperparameter adjustments. The results obtained from this evaluation can be seen in Table 5 and Table 6.

Table 5. Cross-validation in ResUNet.

Fold	Dice	mIOU	Recall	Precision
1	**0.7099**	0.4380	0.6520	0.0469
2	0.6198	0.4407	0.6227	**0.7964**
3	0.6560	0.4376	0.6577	0.7914
4	0.6941	**0.4485**	**0.7095**	0.6963
5	0.6254	0.4394	0.5778	0.7824
Average	0.6610	0.4408	0.6439	0.7768
Standard Deviation	0.0368	0.0040	0.0467	0.0439

Table 7 compares the average cross-validation results of the networks. While TransUNet achieved a superior Dice score compared to ResUNet, it couldn't surpass ResUNet++ in any evaluation metric.

Table 6. Cross-validation in ResUNet++.

Fold	Dice	mIOU	Recall	Precision
1	0.7927	0.7656	0.6165	**0.9456**
2	0.7836	0.7728	0.6806	0.9001
3	**0.8611**	**0.8124**	**0.7241**	0.9347
4	0.7914	0.7771	0.6729	0.8743
5	0.7205	0.7180	0.5653	0.8912
Average	0.7898	0.7692	0.6519	0.9092
Standard Deviation	0.0548	0.0317	0.0531	0.0287

Table 7. Comparison of cross-validations

Network	Dice	mIOU	Recall	Precision
TransUNet	0.6682	0.3467	0.6403	0.7047
UNet	0.5726	0.0573	0.5656	0.6026
ResUNet	0.6610	0.4408	0.6439	0.7768
ResUNet++	**0.7898**	**0.7692**	**0.6519**	**0.9092**

5 Conclusion

Based on experiments conducted with different networks using the Kvasir-SEG dataset, it can be concluded that the TransUNet network's approach, which incorporates transformers into the encoding layers, demonstrated superior performance in terms of the Dice score compared to the UNet and ResUNet architectures, although it did not surpass the results of ResUNet++. The results obtained indicated significant improvements in TransUNet in evaluation metrics such as Dice score, recall, and precision, and, even though it did not surpass ResUNet++, it effectively extracts relevant features from medical images.

Furthermore, the additional use of cross-validation methods on the networks, as applied in the work of Jha et al. [7], contributed to obtaining more consistent and robust results. This approach allowed for a more comprehensive validation of the networks, providing a more reliable view of real performance in different scenarios.

In light of the foregoing, it becomes apparent that the outcomes of this study yield valuable insights pertinent to the choice of neural network architectures in the context of medical image segmentation tasks. These findings underscore the transformative potential inherent in the employment of the transformer-based approach, exemplified by TransUNet. Nevertheless, it is essential to underscore that further research and experimentation remain imperative in the pursuit of enhanced network performance. This includes the exploration of alternative loss functions and the incorporation of post-processing techniques to attain superior results.

Moreover, given the promising trajectory outlined by this study, several opportunities for future research emerge. One possible direction would be the exploration of hybrid strategies that combine the advantages of the TransUNet approach with other cutting-edge architectures, such as Generative Adversarial Networks (GANs), aiming to further improve the accuracy and fidelity of segmentations obtained. Additionally, it would be interesting to explore modifications to the base UNet architecture, such as the incorporation of multi-scale attention mechanisms to enrich the network's contextualization capability, which could yield interesting results. Finally, expanding to more diverse and challenging datasets, as well as applying the TransUNet approach to other areas of medical image analysis, would open up new research frontiers and solidify its potential impact on clinical practice and the advancement of medical science.

References

1. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint: arXiv:2102.04306 (2021)
2. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945). http://www.jstor.org/stable/1932409
3. Diniz, J., et al.: Detecção de covid-19 em imagens de raio-x de tórax através de seleção automática de pré-processamento e de rede neural convolucional. In: Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde. pp. 162–173. SBC, Porto Alegre, RS, Brasil (2023). https://doi.org/10.5753/sbcas.2023.229576, https://sol.sbc.org.br/index.php/sbcas/article/view/25286
4. Gupta, A., Singh, A., Shah, D.: Capsule endoscopy. StatPearls [Internet] (2023). https://doi.org/10.1001/jamanetworkopen.2022.29881, https://www.ncbi.nlm.nih.gov/books/NBK546951/
5. Jha, D., et al.: A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE J. Biomed. Health Inform. **25**(6), 2029–2040 (2021)
6. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y., et al. (eds.) MultiMedia Modeling. Lecture Notes in Computer Science(), vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
7. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255. IEEE (2019)
8. Srivastava, A., et al.: MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J. Biomed. Health Inform. **26**(5), 2252–2263 (2021)
9. Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A.: Loss functions and metrics in deep learning. a review. arXiv preprint: arXiv:2307.02694 (2023)
10. Tomar, N.K., et al.: DDANet: dual decoder attention network for automatic polyp segmentation. In: Del Bimbo, A., et al. (eds.) Pattern Recognition. ICPR International Workshops and Challenges. Lecture Notes in Computer Science(), vol. 12668, pp. 307–314. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68793-9_23
11. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Focus U-Net: a novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput. Biol. Med. **137**, 104815 (2021)