P-ISSN 1991-8941
E-ISSN 2706-6703
Vol. 19, No. 2 (2025)

Journal of University of Anbar for Pure Science
(JUAPS)

journal homepage: juaps.uoanbar.edu.iq

Open Access

Original Paper

Developing an efficient polyp segmentation using residual learning in asymmetric non-local neural networks

Mahmoud H. Farhan [1], **Akeel Abdulraheem Thulnoon** [1*], **Farah Maath Jasem** [1]

[1]Department of Information Technology, College of CS & IT, University of Anbar, Ramadi, Iraq.

ARTICLE INFO

Received: 05/06/2025
Accepted: 03/08/2025
Available online: 19/11/2025
December Issue
10.37652/juaps.2025.161190.1421

Cite @ JUAPS

Corresponding author
Akeel Abdulraheem Thulnoon
akeelalhadithy@uoanbar.edu.iq

ABSTRACT

Colorectal cancer is among the leading causes of death by cancer worldwide. Thus, colorectal cancer needs early diagnosis to help in the management of the disease. Colonoscopy remains the primary tool for polyp detection; however, the procedure sometimes fails due to the challenge of identifying vastly different structures and very small polyps, which can lead to potential overlooks. To overcome the limitation, this work presents a new image segmentation framework based on an Asymmetric Non-local Neural Network (ANN). Having considerably enhanced the accuracy of polyp detection in colonoscopy images, the ANN presents the best features of the non-local operation, which considers both subtle local features and long-range dependencies of an image for its segmentation, thereby facilitating better polyp identifications. Upon evaluating the framework over the Kvasir-SEG dataset, higher results were obtained, with a mean Dice score of 0.947 and a mean IoU (Intersection over Union) of 0.901. This breakthrough is hypothesized to provide a major support for colorectal cancer screening by reducing misdiagnoses and improving overall patient care, aiding in new possibilities for the early detection and effective treatment of this devastating disease.

Keywords: *Asymmetric non-local neural network (ANN), Colorectal cancer, Medical image, Non-local attention, Polyp segmentation*

1 INTRODUCTION

Image segmentation, the technique of dividing an image into portions to identify an image's objects and boundaries, has increased dramatically in various fields, especially in medical imaging [1]. To achieve good object detection performance, most methods are based on hand-crafted features combined with machine learning classifiers such as Random Forests and Support Vector Machines [2]. Additionally, various methods utilize richer contextual information, and these algorithms apply structured prediction methods to enhance efficiency. In most cases, the hand-crafted features are insufficient for real-world applications [3]. Deep learning helps computer vision researchers achieve higher accuracy in several areas, such as semantic segmentation, object identification, and image classification, compared to conventional computer vision. Deep learning methods are also more flexible and domain-specific. The deep learning model autonomously identified specific parts based on a variety of deep neural network architectures, including convolutional neural networks (CNN), auto-encoder networks, and deep belief networks [4].

Among all malignancies [5], colorectal cancer is one of the most significant threats to world health and ranks second in terms of morbidity and mortality. Early detection and treatment depend critically on the accurate localization and segmentation of polyps, which are vital precursors to the disease [6]. The most often used screening method nowadays is colonoscopy; it is a practical approach to detect and remove polyps. Although misdiagnosis rates show 3.5–6% for colorectal cancer and 16.8% for polyps [7], other factors, such as the quality of bowel preparation and the doctor's experience, influence the accuracy of detection. These statistics [8] underline

the pressing need to add computer-aided diagnostic tools to endoscopic operations to lower misdiagnosis rates.

Polyps are a very common condition in the gastrointestinal (GI) system. They can be benign or malignant. Among all, colorectal polyps are abnormal growths in the mucosal membrane lining of the GI system and are considered precursors of colorectal cancer. Anatomical studies have shown that it differs from normal mucosa in size, colour, and surface type. According to morphological deformity in the GI tract, they can occur in a flat, raised, or pedunculated manner. Manual examination has multiple flaws: endoscopists may fail to detect some polyps, especially those with subtle presentations, and the precise molecular etiology underlying the transformation of benign polyps into malignant ones remains unclear [9]. Polyps increase with age, and according to various statistics, polyps are expected in approximately half of the entire population by the time they reach 50 years of age [10]. Figure 1 illustrates sample images from the Kvasir dataset, showcasing polyps with their distinctive characteristics outlined in green, which helps visualize the varied presentations of polyps in colonoscopy images.

Considering all these issues, the requirement for automated and accurate polyp segmentation in clinical practice is of great importance. Polyp segmentation is a complex task due to the variation in sizes and shapes, which further increases the significance of this step. Therefore, the exact segmentation of polyps would indeed be paramount in value toward increasing the level of diagnosis and further enhancing patient outcomes. For instance, advanced segmentation algorithms could assist endoscopists in pinpointing areas of suspicion during colonoscopies, thereby reducing the likelihood of missed polyps and misdiagnosis. In this regard, the development and adoption of automatic segmentation systems are paramount. The systems increase diagnostic accuracy and decrease the burden on health workers with rather complex cases. This technological improvement, therefore, enables timely intervention and treatment, ultimately saving lives.

In this paper, the advancements in polyp segmentation techniques are surveyed with a special focus on the transition from earlier handcrafted feature-based methods to modern DNN-based approaches. Reviewing the current state-of-the-art methods and their applications, it develops a systematic survey of the challenges and possible solutions for this critical area of medical imaging. The key contributions of this paper are as follows:

- *Present the Asymmetric Non-local Neural Network (ANN) Model:* This paper proposes a novel image segmentation framework for segmenting lesions within colonoscopy images, which is based on an asymmetric Non-local Neural Network (ANN). The suggested model enhances segmentation performance by learning from both long-range dependencies and local characteristics.
- *Enhanced Performance in Polyp Segmentation:* The proposed ANN model surpasses state-of-the-art polyp segmentation techniques. With a mean Dice score of 0.947 and mIoU of 0.901 on the Kvasir-SEG dataset, the model's efficacy in detecting polyps with high accuracy was evident.
- *Integration of ResNet-50 with an ANN Decoder:* The model utilizes an ANN-based decoder for feature extraction, in conjunction with a ResNet-50 encoder. This mix allows more precise segmentation by capturing complex interactions between items and combining elements at several levels.
- *Comprehensive Evaluation:* The model was evaluated using a variety of commonly employed metrics, including accuracy, IoU, and Dice coefficient. The robustness of the model was improved by outperforming other state-of-the-art methods.
- *Potential for Broader Application:* The framework's structure and methodology go beyond polyp segmentation, with potential applications in other medical imaging tasks that require precise object identification.

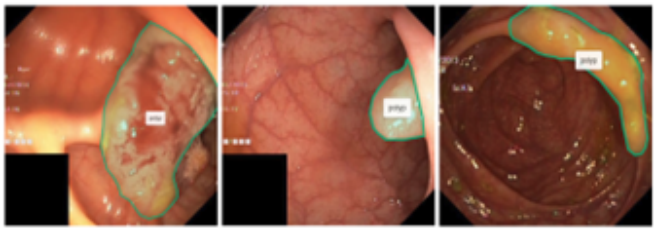

Fig. 1 Sample images from the Kvasir-seg dataset, with the polyp mass outlined in green [11]

2 RELATED WORK

Medical images are now essential to healthcare diagnosis and medical research due to the fast development of medical imaging technology. Medical image

segmentation is critical for automated medical image comprehension and analysis because it typically includes determining the medical image's region of interest. An enhanced understanding and assessment of a medical image result from accurate segmentation [12]. Today, there is a lot of evidence and progress in the use of Machine Learning and Deep Learning approaches in the segmentation of medical images. There are several kinds and formats of medical images that are performed; among the most frequently utilized in the current state of the art are magnetic resonance imaging (MRI), x-ray, and ultrasound images of various body organs [13, 14]. As stated in the preceding section, this research will employ images from colonoscopy recordings, specifically exam frames, to perform polyp segmentation using a supervised learning technique [15].

Image segmentation is the process of categorizing each pixel into predefined labels [16]. Several works have been done in the past few decades on the segmentation of images with various content. In the first methods of polyp segmentation, represented in references [17] and [18], classifiers were used to separate the presence of a polyp from the rest of the image, but such models had very high error rates. Currently, the most widely used methods for polyp identification and segmentation are the convolutional neural networks (CNN) or Fully Convolutional Networks (FCNs) pre-trained methods due to their higher performance level [19]. Another state-of-the-art method, "Pranet", is presented in reference [20], which uses a deep neural network for segmentation and focuses on the area and boundary of the polyp's analysis. The most important aspect of consideration about polyp is that it exists in different sizes and shapes, and considering these two features lead to more accurate detection results.

Safarov et al. [21] Developed a DenseUNet, which is a fully automated model for pixel-wise polyp segmentation. DenseUNet applies a large number of dilated convolutions with many different atrous rates in order to see a large scale without increased computational costs and compromise in spatial information. It also applies an attention mechanism to filter out non-essential information and noise, which may destroy contextual features. DenseUNet attained Dice coefficient scores of 90% and 91% on the Kvasir-SEG and CVC-612 datasets, respectively, proving its effectiveness at identifying polyps in colonoscopy images. Following the work done in segmentation models, authors in [22] came up with Multi-Scale Subtraction Network (MSNet) for automatic polyp segmentation in colonoscopy images. The subtraction unit (SU) was one of the most important components of MSNet, which extracts feature differentials between adjacent levels in the encoder to enable the capture of rich multi-scale feature information. Besides, it is further combined with a LossNet to supervise the learning processes at all feature levels for capturing comprehensive, detailed, and structural information. However, computation complexity and the variety of clinical applications are still clinical challenges for MSNet. Along this line, the authors in [23] proposed the TransFuse architecture, which combines state-of-the-art Transformer models with Convolutional Neural Networks for enhanced medical image segmentation. The fusion module of BiFusion presented the success of feature fusion from both branches and showed state-of-the-art results on various 2D and 3D segmentation tasks: polyp and skin lesion segmentation with a mean Dice score of 0.918 and a mIoU of 0.868. This architecture increases model efficiency and inference speed, which can be deployed on either the cloud or edge. However, the BiFusion module poses some potential limitations due to its complexity and scalability to different medical imaging applications. Furthermore, the Context Fusion Module (CFM) of these three modules in [24] is integrated into Polyp-PVT, including semantic and locational information; the Context Interaction Module (CIM) to locate buried cues on the polyp; and the Spatial Attention Module (SAM) for feature fusion. The framework achieved a mean Dice score of over 0.8 on the Kvasir dataset but falls short of providing effective distinction of boundaries under some lighting conditions. Furthermore, these images can be converted to grayscale. Recently, T. Kim et al. [25] Proposed the Uncertainty Augmented Context Attention Network (UACANet) for polyp segmentation, in which feature representation was done by integrating saliency maps and further combining foreground, background, and uncertain area maps. The Uncertainty Augmented Context Attention Network (UACANet) has achieved a mean Dice score of 76.6% on the ETIS dataset, with an improvement of 13.8% compared to the results obtained using previously available methods. Though efficient in the use of uncertain regions and Parallel Axial Attention for feature extraction, scaling it to other medical imaging tasks remains open for further research on computational efficiency. To enhance segmentation performance, especially of small-sized medical objects, in [26], the Context Axial Reverse Attention Network (CaraNet) was proposed. The CaraNet model integrates an Axial Reverse Attention mechanism with Channel-wise Feature Pyramid blocks, yielding state-of-the-art

segmentation performance. Experiments on different datasets, including brain tumor (BraTS 2018) and polyp datasets (Kvasir-SEG, CVC-ColonDB, CVC-ClinicDB, CVC-300, and ETIS-LaribPolypDB), showed that on the Kvasir-SEG dataset, CaraNet scored a mean Dice score of 0.918 and a mean IoU of 0.865. However, weaknesses exist in bilinear interpolation when used to up-sample feature maps, which may cause a loss of information and produce coarse boundaries, which then could be mitigated. Besides, its backbone has been pre-trained on ImageNet; it does not necessarily conform to the nature of the medical images. In addition, using sliced brain MRI data may not have the right spatial accuracy.

In another approach, Text-Guided Attention Network (TGANet) was introduced by [27] to make polyp segmentation better with the use of text-guided attention to size and number features. The model, using byte-pair encoding and feature enhancement modules, attains a mean IoU of 0.8330 and a mean DSC of 0.8982 on datasets that include Kvasir-SEG. Although TGANet can effectively boost the segmentation accuracy of clinically relevant polyps, an increased level of model complexity and thus computational demands, along with potential challenges to be generalized to diverse polyp types or sizes, might thus adversely affect its overall efficiency and applicability. Presumably Semantic Segmentation Transformer (SFFormer) Model [28] for segmentation of medical images is proposed with a hierarchical Transformer-based encoder and Progressive Locality Decoder (PLD) to focus on local features and reduce attention dispersal. The most important contributions are the close integration of pyramid Transformer architecture aimed to boost generalization, PLD for high-detail information processing, and excellent performance on benchmarks such as ETIS, CVC-ClinicDB, Kvasir. SSFormer outperforms in the segmentation task and potentially brings up the accuracy of medical and general image analysis to a higher level. It is strong for advanced feature integration and high accuracy, while complexity might need huge datasets with very heavy computational resources. Besides, CASCaded Attention DEcoder (CASCADE) was proposed in [29] for medical image segmentation using hierarchical vision transformers for fusion of features and context refinement with the help of attention gates and convolutional modules. Up to 5.07% and 6.16% improvements were achieved on Dice and mIoU scores, respectively, showing the effectiveness of the method for medical and general computer vision tasks. CASCADE demonstrates performance improvement in segmentation accuracy through an advanced hierarchical transformer architecture and its means to integrate features effectively. The complexity might pose a problem, needing high computational resources or not generalizing with equal efficacy in all contexts. PrototypeLab, introduced in [30], is a method to robustly segment polyps by using different imaging modalities (WLI, BLI, LCI, FICE) to create prototypes, representing object characteristics. PrototypeLab improves DICE of ≥90% and mIoU of ≥85% on Kvasir-SEG datasets. Although PrototypeLab provides high segmentation accuracy over many different datasets with various image variabilities, its implementation is complex, and the requirement for large quantities of diverse data to optimize its adaptability may be prohibitive. In [31], the Polyp Segmentation Network (PSNet) has a dual encoder–decoder structure for better polyp segmentation. PSNet was built to incorporate deep learning modules with PS encoders, transformer encoders, PS decoders, enhanced dilated transformer decoders, and merge modules. It combines some state-of-the-art performances on five polyp datasets with mDice and mIoU scores of 0.863 and 0.797, and 0.941 and 0.897, respectively, improved results with a modified dataset. However, this model may become complex, which leads to high computational costs, generalizing difficulties, and insensitivity to other polyp conditions. Finally, CAFE-Net [32] is composed of three major modules, including FSEM (Feature Supplement and Exploration Module), CADM (Cross-Attention Decoder Module), and MFA (Multi-Scale Feature Aggregation Module), which, in principle, can achieve higher accuracy in polyp segmentation with small-sized polyps compared to the state of the art. But generalization over unseen datasets or conditions not properly presented could be an issue, and it still remains difficult to plug into modules a model good at the preserved details.

While analyzing the latest state of the art in polyp segmentation, a number of key trends can be identified. It is possible to trace the progression from classical machine learning approaches through sophisticated architectures of deep learning. Models such as DenseUNet, MSNet, TransFuse, and SSFormer showcase potential when keeping together several techniques combined in an effort toward raising results on segmentation by a margin. However, shortcomings persist concerning the small handling of polyps, providing boundary demarcation, and robustness across different imaging modalities and lighting variations. In addition, most of the techniques in the literature show limitations in terms of their im-

plementation overhead and the time required to achieve the desired results. Our proposed Asymmetric Non-local Neural Network (ANN) aims to address these challenges by leveraging both local features and long-range dependencies, potentially offering a more efficient and accurate approach to polyp segmentation that can be readily applied in real-world clinical scenarios.

3 THE PROPOSED MODEL

An encoder-decoder architecture is the model put out in this work. The main idea of the architecture lies in the ability to encode the input data into a compact, high-level representation by utilizing strong features extracted by the backbone. Then, using the decoder heads, generate the desired output, in this case, an accurate segmentation map. Due to this, very wide flexibility of output format is applicable and the capability of catching both local and global features; therefore, it has been widely applied in a wide spectrum of quite different tasks in computer vision, including semantic segmentation, object detection, and image segmentation. The suggested model uses a ResNet-50 [33] encoder to extract visual features from the images after a data preprocessor has prepared the input data. The decoder, which uses an asymmetric non-local neural network (ANN) model [34], is the vital part. Using the encoded features from the ResNet-50 encoder, the ANN-based decoder produces the segmentation mask as the output. In essence, the ANN model decodes the data by mapping the encoded feature representations to the desired predictions (as illustrated in Figure 2).

3.1 Data preprocessor

The preprocessor basically gets the mask data and image ready in a way that can be used to train a deep learning model for precisely segmenting images. The preprocessor confirms that the model obtains high-quality, consistent, and normalized input data by performing these essential tasks, which are required to achieve optimal performance and accurate outcomes for segmentation. The preprocessor in this work is concerned with:

- **Color Space Conversion:** converts images from the BGR format to RGB format, which is the usual format for almost all deep learning models.
- **Normalization:** Normalization is the process of scaling the pixel data to have a unit variance and centering it around zero. To do this, take each color channel in the training data, subtract its mean, and divide the result by the standard deviation. This stage allows more efficient model learning.
- **Resizing and Padding:** To ensure consistent processing during training, make sure all images and the corresponding masks used for segmentation have the same dimensions (512 x 1024 pixels in this instance). Depending on whether the image and its mask, padding is applied using different values.

3.2 Encoder (backbone)

The encoder is a convolutional neural network (CNN) in charge of obtaining high-level features from the source image. The encoder utilized in this work is a ResNet50 architecture that was pre-trained using an enormous image dataset. The encoder can learn significant features from an input image. This provides an accurate foundation for extracting features. The three primary components of the ResNet50 model are as follows:

A- Stem (Stage 1): Initial Processing The stem processes the original input image through a series of pooling and convolutional layers. These layers hold out operations including:

- **Low-level feature extraction:** The first convolutional layers take out of the image features such as edges, corners, and textures. The basis for additional processing is provided by these features.
- **Down-sampling:** By decreasing the spatial resolution of the feature maps, max pooling layers improve processing efficiency. By decreasing the computational load, this down-sampling enables the network to focus on the most important information.
- **Setting ready for deeper processing:** The stem builds up the feature maps that will be extracted by the later residual layers, which will extract complex and abstract features.

B- Residual Layers for Feature Extraction (from Layer 1 to Layer 4): Four residual layers from Layer 1 to Layer 4 (i.e., from stage 2 to stage 5), each consisting of several bottleneck blocks, constitute the encoder. Using a technique called residual learning, these layers focus on incrementally more advanced characteristics from the input image:

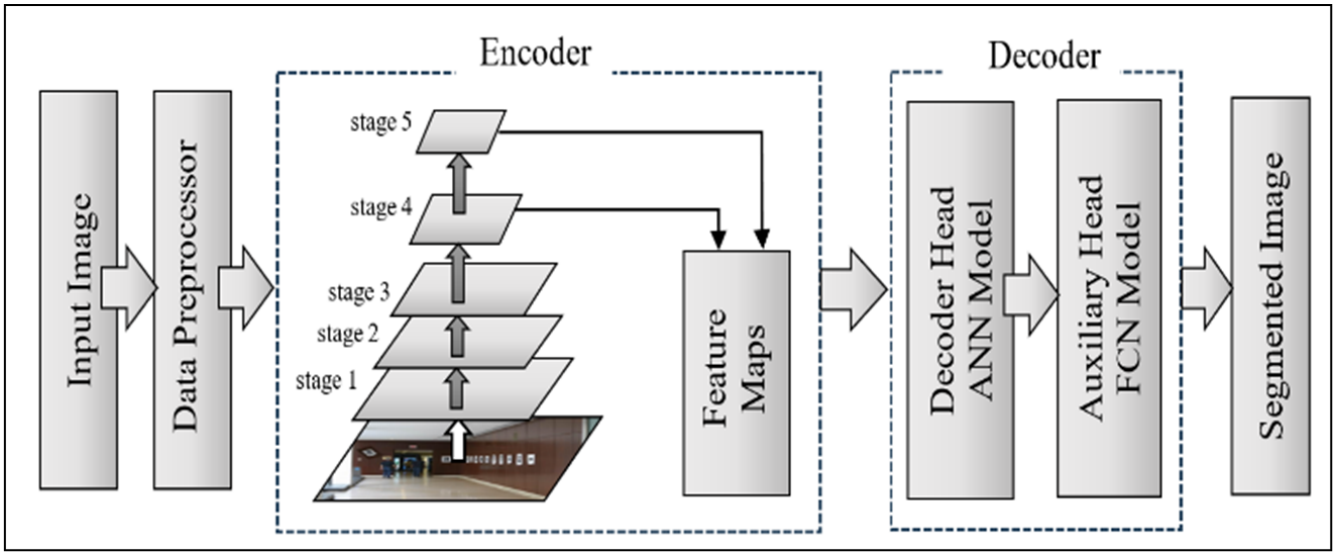

Fig. 2 Overview of the proposed Model

- **Bottleneck blocks:** A "bottleneck" is located in the middle of each bottleneck block, which consists of three convolutional layers. By decreasing the total number of parameters as well as improving information flow, this structure assists in improving the network's precision and efficacy.
- **Residual connections:** The network is able to learn residual functions that improve the features extracted by previous layers because each bottleneck block adds its input to its output. The remaining learning improves the ability of the network to learn complex characteristics and helps in solving the vanishing gradient issue.
- **Increasing complexity:** The network may extract increasingly detailed and complex features as it passes across the layers, while the convolutional layers have more filters as they go. As a consequence, the network may store ever-more-detailed details regarding the image, including the textures, forms, and relationships between items.

The encoder can extract useful representations from the source image by using the pre-trained ResNet50 model, giving it a strong basis for feature extraction. By employing the knowledge obtained from an extensive image dataset, this pre-training improves the capacity of the network for adapting to new tasks and images.

3.3 Decoder head

Understanding complex instances and object recognition is necessary for semantic segmentation. Nevertheless, objects often possess cross-scale interactions, which typical networks fail to capture. An asymmetric non-local neural network is employed to prevent this. The Asymmetric Non-local Neural Network (ANN) is composed of two primary components: the Asymmetric Pyramid Non-local Block (APNB) and the Asymmetric Fusion Non-local Block (AFNB). The APNB employs a pyramid sampling module within the non-local block, which reduces calculations and memory usage while maintaining performance. The AFNB is an adaptation of the APNB that attempts to merge features from multiple stages while taking long-term dependencies into account in order to improve the ability to learn of non-local blocks and the effectiveness of segmentation. Overall, the ANN and all of its constituent parts are shown in Figure 3. The network backbone utilizes ResNet-50 to extract feature maps at various stages. They lack cross-scale context, but they do capture scale-specific patterns. An AFNB fuses adjacent stages to aggregate multi-scale information. Between the last two stages (i.e., Stages 4 and 5), embedding and non-local operations are used. Contextual cues are combined and fused multi-scale features are the result. Because the fused features include complementary details from both stages, their semantic richness is higher.

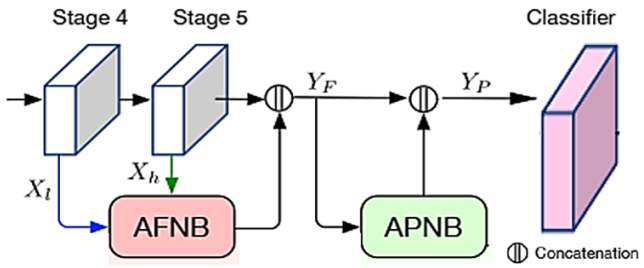

Fig. 3 Asymmetric Non-local Neural Network Model [34]

They take the place of the network's initial higher-level functionalities. This makes it possible for downstream modules to function on a combined multi-scale view, such as the APNB. The fused features are then modelled using the APNB. To gather data from various regions, it first samples representative points using pyramid pooling. Compared to using all points, this reduces computation. All places, regardless of distance, are related by the calculation of similarity between sampled points.

The contextual interactions through the fused map are highlighted by the similarity matrix. An integrated attention map that summarizes affinities is produced by normalizing it. By combining contextual responses, this matrix improves the depiction of each position in relation to its multi-scale surroundings. This allows APNB to record complex object associations that are not limited by distance. By describing correlations among locations from different stages of the network effectively, the APNB output predicts and classifies semantic labels. Through APNB and the combination of information from adjacent scales, difficult scenarios involving items spanning many scales can be identified more easily.

A- AFNB (Long-Range Dependencies) Obtaining long-range relationships among features derived from different encoder levels is the main objective of the AFNB module, Figure 4. It mixes features to create a more accurate representation by using self-attention to determine which are most relevant to one another. The following steps are involved in this process:

- **Key and Query Projection:** This is the next layer, known as convolutional, where features from other layers are projected onto key/query vectors whose main purpose is to extract only the valuable information those features possess or enclose.

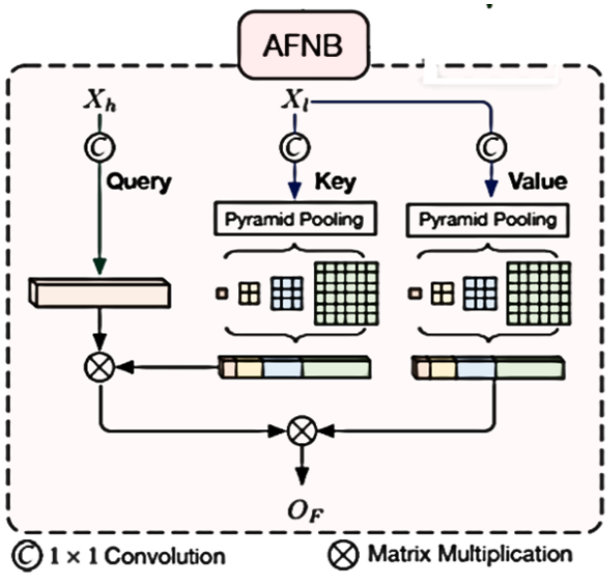

Fig. 4 Asymmetric Fusion Non-local Block (AFNB) [34]

- **Self-Attention:** Query vectors attend to the key vectors, generating a weighted sum of the key vectors, based on their relevance to the query. This feature allows the module to pull out the most relevant features of any level, no matter how far apart in space those features might be.

- **Value Projection:** The attended key vectors are the input to this layer, which changes them into value vectors using a convolutional layer. Vectors are refinements and mixtures of information from relevant features; they capture only important aspects and respect interactions.

- **Output Projection:** The value vectors are concatenated so that more and more contextual information or awareness, via representation, is brought in and then passed to one final feature map. As it captures long-range dependencies between the features using data from many levels, the output feature map better informs the image being processed.

The AFNB module enhances the capability of the decoder in producing an accurate segmentation map, capturing long-range dependencies. It enables the decoder to take into account the connections between features from different image regions and then generate consistent, reliable segmentation results.

B-APNB (Local Context Refinement) The AFNB component enhances the capacity of the decoder. The APNB module aims to enhance its features by employing contextual data from neighboring pixels. In order to increase the accuracy of segmentation, it utilizes self-attention to learn the connections among nearby pixels, Figure 5. The following steps are involved in this process:

- **Key and Query Projection:** Using convolutional layers, the features are projected into key and query vectors. These vectors, which record the important information found in every pixel and its surrounding area, act as a representation of the feature content.
- **Self-Attention:** By combining the key vectors corresponding to their weight and importance to the query, the key vectors are attended to using the query vectors. This enables the module to focus on the nearest pixels that are most important, taking advantage of their connections and the context they offer.
- **Value Projection:** A convolutional layer is then used to project the attended key vectors into value vectors. These vectors gather important characteristics of the local context while taking the connections among pixels into account. They represent the contextually aware and improved data from the adjacent pixels.
- **Output Projection:** A single output feature map is generated by estimating and mixing the value vectors, which results in a more advanced and specifically aware representation. By including data from adjacent pixels and gathering the connections between them, this output feature map offers an improved understanding of the local context.

The ability of the decoder to produce precise and comprehensive segmentation maps is improved by the APNB module by integrating local context. In particular, at object boundaries, this allows the decoder to take into account the relationships among pixels adjacent to it, producing segmentation results that are more accurate and trustworthy.

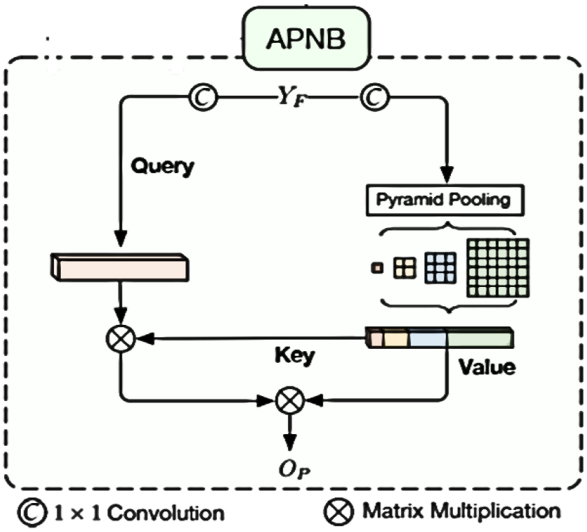

Fig. 5 Asymmetric Pyramid Non-local Block (APNB) [34]

3.4 Auxiliary head

An additional level of supervision is provided during training by the Auxiliary Head (FCN Head). Using features from a previous encoder layer, it produces a segmentation map with low resolution. As a consequence, the model performs more effectively overall, and the training process becomes more regular.

A-Feature Processing (Convolutional Layers): The features from the encoder are processed by a series of convolutional layers that make up the FCN Head. These levels carry out tasks like

- **Extracting relevant information:** Convolutional layers are employed to generate the segmentation map by collecting useful data from the encoder features. They concentrate on the elements of the features that are most essential in establishing each pixel class.
- **Reducing dimensionality:** Convolutional layers decrease the overall size of the features, which enhances their processing efficiency in terms of computation. By doing so, the computational load of the network decreases while the vital information necessary for segmentation remains within it.
- **Preparing for segmentation:** Convolutional layers convert the features into a format that is suitable

for creating the segmentation map, preparing them for the final segmentation layer.

B-Dropout: Preventing Overfitting A dropout layer in the FCN head aids in preventing overfitting during training. During training, dropout randomly eliminates a certain number of neurons, pushing the network to learn adaptive representations that aren't too dependent on any one neuron. This enhances the network's capacity for generalization, helping it to perform well with unknown data [35].

C- Segmentation Layer: Generating the Segmentation Map The segmentation layer (i.e., convolutional layer) creates the segmentation map with low-resolution images. The map has the same dimensions as the input image, and each pixel represents the predicted class based on the features that the preceding convolutional layers processed. An approximation of the object boundaries in the image is provided by this segmentation map.

The model learns more reliable and accurate segmentation task representations with the assistance of the FCN Head, which enhances the model's overall performance during training. Through more regular training, the network becomes less likely to overfit the training set and is better able to adapt to new images. This is made feasible by the additional supervision. With a focus on the distinct contributions made by each component to the general segmentation process, this comprehensive explanation offers an improved understanding of the internal structure of the proposed encoder-decoder model and all of its individual components. It emphasizes how essential each element is to getting highly accurate and thorough segmentation results.

4 EXPERIMENTAL ANALYSIS

The PyTorch framework in Python was used to implement and code the ANN framework independently in this work. The following parameters of a laptop were used to conduct the experiments: AMD Ryzen 7 4800H 64-bit CPU, 16 GB RAM, and 4 GB GPU driver (GTX 1650Ti). Stochastic gradient descent (SGD) is the optimizer utilized for this model, with weight decay set to 0.0005, momentum set to 0.9, and learning rate set to 0.01. In the training phase, the model is trained for 500 epochs with a batch size of 10. Every 100 training epochs, the system is validated.

Kvasir-seg [11] was the dataset used for evaluating the model. This dataset was divided into two sets: the training set, which had 90% of the entire dataset, and the test set, which included the remaining 10%. The result of the model depended on the test set. The applied model was compared using the mean average precision at various IoU thresholds and loss functions in the subsections that followed.

4.1 Evaluation metrics

The well-known segmentation evaluation metrics of recall, precision, intersection over union (IoU), and dice coefficient were employed to evaluate the polyp segmentation. Using well-known parameters like true positive (TP), false positive (FP), and false negative (FN), these metrics were computed as follows:

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{1}$$

$$\text{Recall} = \frac{TP}{(\text{TP} + \text{FN})} \tag{2}$$

- **Intersection over Union (IoU):** is a commonly used metric in the assessment of segmentation methods. The similarity between the genuine mask and the predicted pixels is shown by the equation:

$$\text{IoU} = \frac{TP}{(\text{TP} + \text{FP} + \text{FN})} \tag{3}$$

- **Dice similarity coefficient:** This common measure is used to compare the pixel-by-pixel outcomes of the segmentation that was predicted with the ground truth. The dice coefficient formula can be defined as follows:

$$\text{Dice} = \frac{(2 \times TP)}{(2 \times \text{TP} + \text{FP} + \text{FN})} \tag{4}$$

Figure 6 shows how accurate an ANN head and an auxiliary head were after 500 training epochs. The accuracy of the ANN Head starts out fairly low (around 80 epochs) in the left part of Figure 6, but it quickly improves during the first few epochs of training. Between epochs 100 and 150, a number of about 97% is seen. It stays between 96 and 97% after that, which is a pretty stable range. It appears that the model is working well because its accuracy did not decrease much during the training process. The accuracy of an auxiliary head while it is being trained is shown on the right side of Figure 6. The accuracy goes up and down while training, but most

of the time it goes up when an auxiliary head is added. The pattern shown here indicates the model's auxiliary head is getting better at classifying or segmenting the data.

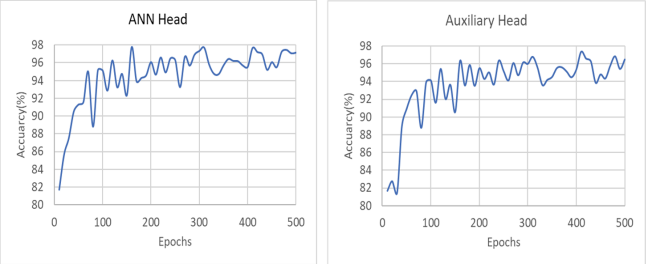

Fig. 6 Accuracy of Segmentation on training phase, ANN Head (Left) and Auxiliary Head(Right)

The results in Figure 7 show that the segmentation model for the polyp was getting better over 500 epochs. There is an obvious rise in metrics such as IoU, accuracy, Dice, and precision. This means that the model learns to segment the polyp better over time. IoU grows from 76.92% after 100 epochs to 83.69% after 500 epochs, for example. This steady improvement indicates a model that learns and improves how it divides things into categories. The high numbers, like 90.67% accuracy after 500 epochs, show that the model has been trained well and can correctly divide this image into its components. This means the model probably works for the task, especially since there aren't any signs of big changes or performance peaks, which means the model continues learning.

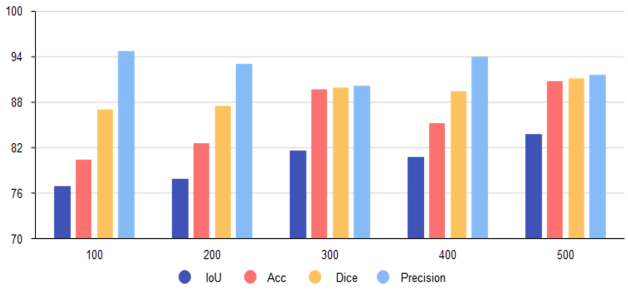

Fig. 7 Evaluation metrics (IoU, Acc, Precision and Dice) for Polyp of the proposed approach

As shown in Figure 8, the consistently high-performance metrics (e.g., IoU close to 96%, accuracy closer to 99%, Dice and precision closer to 97–99%) for background segmentation show that the segmentation is very accurate. As the number of epochs goes up, the model seems to have well-converged and learned to reliably find and separate the backgrounds. These good results point to a model that should be able to achieve high accuracy in real life, especially when reliable background segmentation is important.

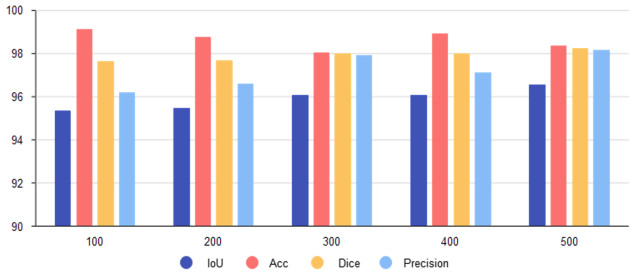

Fig. 8 Evaluation metrics (IoU, Acc, Precision and Dice) for the background of the proposed approach

Figure 9 illustrates that the proposed model performs well across a number of metrics as training runs on. The letter m at the beginning of each metric stands for the mean value across all classes. Precision (mPrecision) goes up from 95.44% to 94.86% over 500 epochs, and recall (mRecall) goes up from 89.75% at 100 epochs to 94.5% at the same 500 epochs. It indicates that the model learns and improves its segmentation strategy over time, as shown by the steady rise in recall and precision and the high values for mDice (e.g., 92.29% at 100 epochs, 94.68% at 500 epochs) and mIoU (e.g., 86.14% at 100 epochs, 90.12% at 500 epochs). The excellent results indicate that the segmentation works well across all time periods, which suggests that the model was well-trained and is probably reliable.

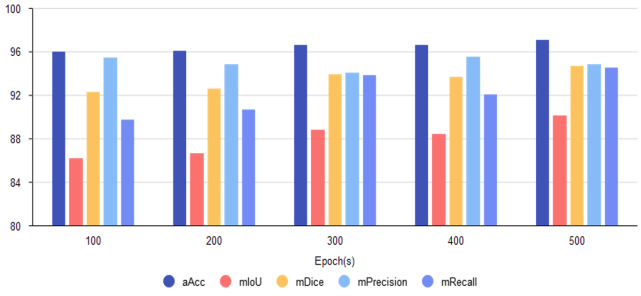

Fig. 9 Evaluation metrics (IoU, Acc, Precision and Dice) for the background of the proposed approach

4.2 Losses functions

Cross-entropy loss is a typical loss function for classification tasks, such as image segmentation, and it is

employed in both heads (ANN and FCN) of the proposed model. It motivates the model to guess each pixel in the input image's right class label. The model is supposed to provide a class label (foreground, background, etc.) for each pixel in the input image when it comes to image segmentation. The difference between the true class probabilities and the predicted class probabilities is measured by the Cross-Entropy Loss. The CrossEntropyLoss is computed as follows, given the true class labels (ground truth) and the predicted class probabilities (p), which are the model's output.

$$L = -\frac{1}{N}\left[\sum_{j=1}^{N}[y\log(p) + (1-y)\log(1-p)]\right] \quad (5)$$

Where the sum is calculated across all pixels and N is the total number of pixels in the input image. The natural logarithm is the log function. When the model predicts incorrectly, the Cross Entropy Loss function penalizes it; this forces the model to learn how to anticipate the proper class labels in order to minimize the loss.

The loss of an ANN head and auxiliary heads over 500 epochs is shown in Figure 10. On the left side of the Figure, the most noteworthy aspect about the loss of ANN Head is how quickly it drops in the first few epochs, roughly from epoch 0 to 50. The loss continues to decrease, but at a much slower and more gradual rate after this initial sharp drop. This means that the model learns quickly at first, but over time, it gets worse at a slower rate. After the first drop, the changes in the loss curve are small and consistent, which means the model is getting closer to a stable state. The loss of an auxiliary head is shown on the right side of Figure 10. This loss is very obvious in the first few epochs, dropping quickly from a value above 0.2 to a value below 0.05. The loss stops decreasing quickly upon the first decrease and stays stable, going up and down in small peaks around an average low value. Overall, the curve is going down, which shows that the model is learning and getting better at what it does. It's possible that more training after 500 epochs might not make a major change or could even cause overfitting. A very successful training process can be seen in Figure 11, where the model's loss rapidly decreases in the first several epochs. The model has essentially converged to a stable performance slow down, as shown by the loss rapidly decreasing to a low value (about 0.1) and then varying around this value.

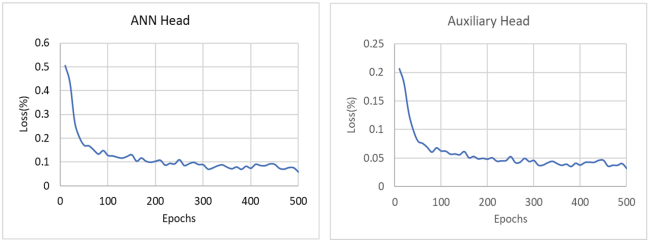

Fig. 10 Losses of Segmentation, ANN Head (Left) and Auxiliary Head (Right)

Consistently low values over the training period indicate that the model has learnt effectively and efficiently and is probably performing well. The reason for successful training is further supported by the overall decrease in loss, which is a good indicator of model progress with few spikes around the stable loss value.

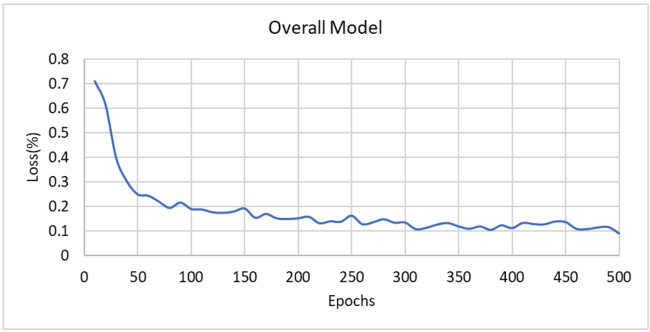

Fig. 11 Loss of Segmentation in Overall Model

The samples of polyp segmentation using the proposed model are shown in Figure 12, demonstrating the accuracy of the model in the majority of cases. Samples have been separated into three categories: small polyps (rows 1-3), medium polyps (rows 4-6), and large polyps (rows 7-9). Each category contains three images, each row of these categories consists of four columns: the original image, the original mask (from the database), the predicted mask (obtained from the proposed model), and the segmented image, where the background is colored green and the polyp is red. When the original mask and the predicted mask are compared, the results demonstrate that the model performs effectively regardless of polyp size. The two masks nearly match in terms of quality. The samples for these categories were presented with care to show how robust the model is against most situations.

4.3 Comparison with previous studies

The model is efficient because it enhances the learning of local and global features equally by employing Asymmetric Pyramid Non-local Block and Asymmetric Fusion Non-local Block components. The first is used for local features, while the second is used for global ones. The model yields better segmentation as these components allow it to process detailed information and wider spatial connections, especially in complicated scenarios where the polyps either blend with the surrounding tissues or where their edges become less distinct. The ResNet-50 encoder can make the model more efficient in recognizing significant features with smooth computing. Charts from Figures 6 to 10 indicate stable training output of precise segmentation for ANN and auxiliary heads. The model significantly reduces false negatives and false positives that are highly required in medical applications because of its accurate metrics and steady alignment of high recall with loss curves. These results make clear the promising potential of the proposed ANN model for polyp segmentation and its potential application in the field of object-delineation-based medical image segmentation tasks.

Table 1 highlights the advantages of the proposed model as compared to the state-of-the-art studies. The model engages with representative performance metrics like Dice coefficient and average Intersection over Union (mIoU). The proposed model shows significant gains over the methods for which this information is produced. For example, the method TransFuse-L [13]yielded a Dice score of 0.918 and mIoU of 0.868, compared to our model's achievement of 0.947 in Dice and 0.901 in mIoU. Similarly, the proposed model yielded far better robustness and segmentation accuracy than SFFormer-L [18] Dice score of 0.936 and mIoU of 0.891. More recent approaches, such as CAFE-Net [22] and PSNet [21] are on the same footing with a proposed solution. Our proposed model somewhat surpassed the commanding performance of CAFE-Net Dice score of 0.943 and a mean Intersection over Union (mIoU) of 0.899.

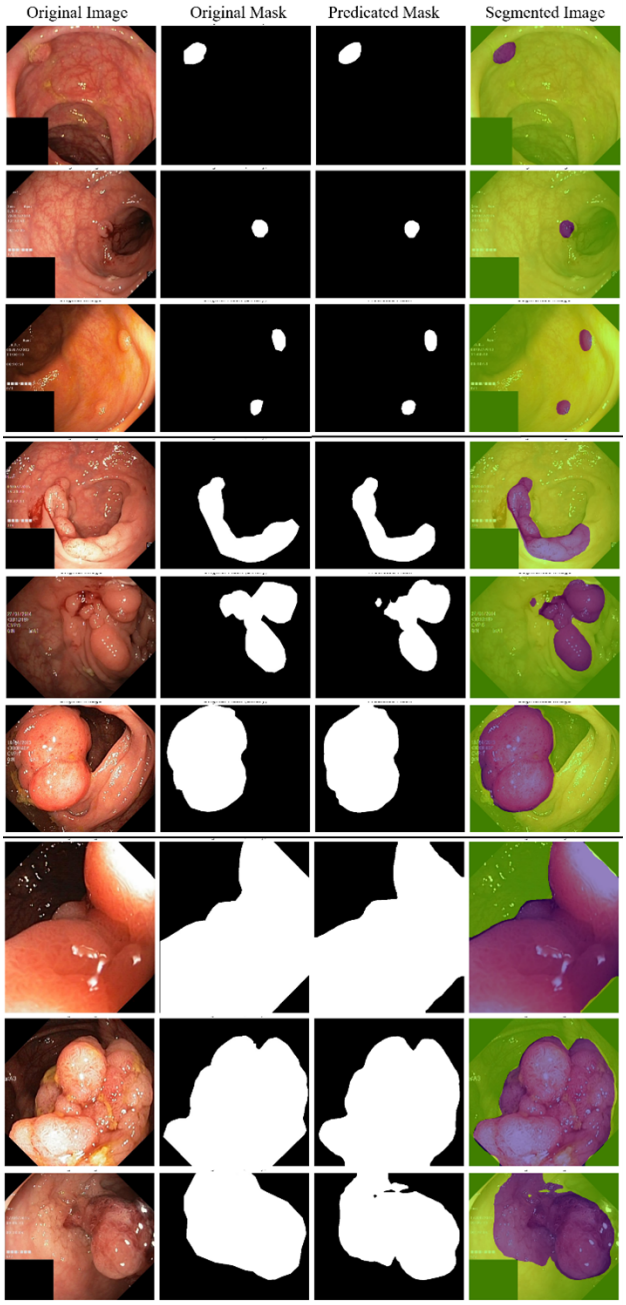

Fig. 12 Polyp image segmentation samples on the Kvasir-SEG using the proposed model

Table 1 Hybrid deep learning methods applied to object recognition systems

Method	Year	Dice	mIoU
MSNet [12]	2021	0.907	0.862
TransFuse-L [13]	2021	0.918	0.868
Polyp-PVT [22]	2021	0.917	0.864
UACANet [15]	2021	0.912	0.859
CaraNet [16]	2022	0.918	0.865
TGANet [17]	2022	0.898	0.833
SFFormer-L [18]	2022	0.936	0.891
PVT-Cascade [19]	2023	0.926	0.878
PrototypeLab [20]	2023	0.909	0.854
PSNet [21]	2023	0.929	0.879
CAFE-Net [22]	2024	0.943	0.899
The Proposed Method	2025	0.947	0.901

The proposed methodology proves highly effective, as evidenced by its strong performance in comparison. The suggested model is effective because of its novel architecture, specifically due to the introduction of an asymmetric pyramid non-local block for the extraction of local features and an asymmetric fusion non-local block for the acquisition of global features. This new integration enables the model to capture specific challenges of polyp segmentation, such as variable size, shape, and subtle borders that might prove challenging for competing models. The ability of the ANN model to learn pixel-level minutiae along with rich spatial connections ensures better accuracy of segmentation. Besides, the use of a pre-trained ResNet-50 encoder will enhance the model's capability to extract all important features while avoiding unnecessary computational complexity. This, therefore, makes the ANN model more accurate and lean in terms of the computer resources involved, especially when compared with models like TransFuse-L and SFFormer-L, which might have scalability and processing challenges. As indicated in Figures 6 to 10, the training process is very consistent, with very good precision and recall, and low false positives and false negatives, proving the reliability of the model for clinical applications. Overall, the robustness of the proposed approach can be drawn from the fact that both the ANN and all the auxiliary heads perform well, as shown in the loss curves. The inclusion of advanced feature extraction and an effective architecture helps our method outperform other competitive models in polyp segmentation and also assists in applying it to other medical image segmentation tasks.

5 CONCLUSION

Our study presents a novel asymmetric non-local neural network for polyp segmentation in colonoscopy images. Based on an integrated ResNet-50 network for encoding and an ANN-based decoder. Our method achieves the best performance, with a mean Dice score of 0.947 and mIoU of 0.901 on the Kvasir-SEG dataset, outperforming recent deep state-of-the-art approaches. The enhanced capability of the model for capturing both local features and long-range dependencies marks a significant advancement in medical image segmentation. These results consequently carry great implications, extending beyond mere polyp segmentation to other areas in medical imaging where accurate object delineation is important. Enhancing the detection of polyps with accuracy would contribute toward earlier identification of colorectal cancer precursors, reduce missed diagnoses, and improve patient outcomes. With all the features and distinctive results that the model provides, it is necessary to point out some limitations. The model increases the complexity of the space resulting from the complexity of the feature extraction model. The complexity of the encoder-interpreter model, knowing that most models operate with the same structure, means they remain within the limitations of the model. Further research directions might consider the performance of the model on diverse datasets, investigate its suitability for other medical imaging tasks, and further modification of the architecture to satisfy real-time clinical usage. This present work represents a leap forward in polyp segmentation, hence introducing an effective method that should improve colorectal cancer screening and diagnosis.

ACKNOWLEDGEMENT

N/A

FUNDING SOURCE

No funds received.

DATA AVAILABILITY

N/A

DECLARATIONS

Conflict of interest

Authors declare no conflict of interest.

Consent to publish

All authors consent to publish in JUAPS.

Ethical approval

N/A

REFERENCES

[1] Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. IET Image Processing. 2022;16(5):1243–1267. 10.1049/ipr2.12419

[2] Jozdani SE, Johnson BA, Chen D. Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. Remote Sensing. 2019;11(14):1713. 10.3390/rs11141713

[3] Dang T, Nguyen TT, McCall J, Elyan E, Moreno-García CF. Two-layer Ensemble of Deep Learning Models for Medical Image Segmentation. Cognitive Computation. 2024;16(3):1141–1160. 10.1007/s12559-024-10257-5

[4] Farhan MH, Shaker K, Al-Janabi S. Efficient Approach for the Localization of Copy-Move Forgeries Using PointRend with RegNetX. Baghdad Science Journal. 2024;21(4):1416. 10.21123/bsj.2023.8304

[5] Mattiuzzi C, Sanchis-Gomar F, Lippi G. Concise update on colorectal cancer epidemiology. Annals of Translational Medicine. 2019;7(21):609–609. 10.21037/atm.2019.07.91

[6] Zhang R, Li G, Li Z, Cui S, Qian D, Yu Y. In: Adaptive Context Selection for Polyp Segmentation. Springer International Publishing; 2020. p. 253–262. 10.1007/978-3-030-59725-2_25

[7] Wang J, Lim CS. Synergistic Multi-Granularity Rough Attention UNet for Polyp Segmentation. Journal of Imaging. 2025;11(4):92. 10.3390/jimaging11040092

[8] Takano D, Omura H, Minamoto T. Detection and Classification Method for Early-Stage Colorectal Cancer Using Dyadic Wavelet Packet Transform. IEEE Access. 2025;13:9276–9289. 10.1109/access.2025.3526786

[9] Blanes-Vidal V, Baatrup G, Nadimi ES. Addressing priority challenges in the detection and assessment of colorectal polyps from capsule endoscopy and colonoscopy in colorectal cancer screening using machine learning. Acta Oncologica. 2019;58(sup1):S29–S36. 10.1080/0284186x.2019.1584404

[10] Afify HM, Mohammed KK, Hassanien AE. An improved framework for polyp image segmentation based on <scp>SegNet</scp> architecture. International Journal of Imaging Systems and Technology. 2021;31(3):1741–1751. 10.1002/ima.22568

[11] Jha D, Smedsrud PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, et al. In: Kvasir-SEG: A Segmented Polyp Dataset. Springer International Publishing; 2019. p. 451–462. 10.1007/978-3-030-37734-2_37

[12] Khalaf M, Dhannoon BN. MSRD-Unet: Multiscale Residual Dilated U-Net for Medical Image Segmentation. Baghdad Science Journal. 2022;19(6(Suppl.)):1603. 10.21123/bsj.2022.7559

[13] Mukhlif AA, Al-Khateeb B, Mohammed MA. An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges. Journal of Intelligent Systems. 2022;31(1):1085–1111. 10.1515/jisys-2022-0198

[14] Al-Waisy AS, Al-Fahdawi S, Mohammed MA, Abdulkareem KH, Mostafa SA, Maashi MS, et al. RETRACTED ARTICLE: COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images. Soft Computing. 2020;27(5):2657–2672. 10.1007/s00500-020-05424-3

[15] Branch MVL, Carvalho AS. Polyp Segmentation in Colonoscopy Images using U-Net-MobileNetV2. arXiv; 2021. 10.48550/ARXIV.2103.15715

[16] Hazzaa F, Udoidiong I, Qashou A, Yousef S. Segment Anything: A Review. Mesopotamian Journal of Computer Science. 2024;2024:150–161. 10.58496/mjcsc/2024/012

[17] Mamonov AV, Figueiredo IN, Figueiredo PN, Richard Tsai YH. Automated Polyp Detection in Colon Capsule Endoscopy. IEEE Transactions on Medical Imaging. 2014;33(7):1488–1502. 10.1109/tmi.2014.2314959

[18] Tajbakhsh N, Gurudu SR, Liang J. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. IEEE Transactions on Medical Imaging. 2016;35(2):630–644. 10.1109/tmi.2015.2487997

[19] Akbari M, Mohrekesh M, Nasr-Esfahani E, Soroushmehr SMR, Karimi N, Samavi S, et al.. Polyp Segmentation in Colonoscopy Images Using Fully Convolutional Network. arXiv; 2018. 10.48550/ARXIV.1802.00368

[20] Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, et al. In: PraNet: Parallel Reverse Attention Network for Polyp Segmentation. Springer International Publishing; 2020. p. 263–273. 10.1007/978-3-030-59725-2_26

[21] Safarov S, Whangbo TK. A-DenseUNet: Adaptive Densely Connected UNet for Polyp Segmentation in Colonoscopy Images with Atrous Convolution. Sensors. 2021;21(4):1441. 10.3390/s21041441

[22] Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, et al. In: PraNet: Parallel Reverse Attention Network for Polyp Segmentation. Springer International Publishing; 2020. p. 263–273. 10.1007/978-3-030-59725-2_26

[23] Zhang Y, Liu H, Hu Q. In: TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. Springer International Publishing; 2021. p. 14–24. 10.1007/978-3-030-87193-2_2

[24] Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. arXiv preprint. 2021. 10.48550/ARXIV.2108.06932

[25] Kim T, Lee H, Kim D. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21. ACM; 2021. p. 2167–2175. 10.1145/3474085.3475375

[26] Lou A, Guan S, Ko H, Loew MH. CaraNet: context axial reverse attention network for segmentation of small medical objects. In: Išgum I, Colliot O, editors. Medical Imaging 2022: Image Processing. SPIE; 2022. p. 11. 10.1117/12.2611802

[27] Tomar NK, Jha D, Bagci U, Ali S. In: TGANet: Text-Guided Attention for Improved Polyp Segmentation. Springer Nature Switzerland; 2022. p. 151–160. 10.1007/978-3-031-16437-8_15

[28] Wang J, Huang Q, Tang F, Meng J, Su J, Song S. In: Stepwise Feature Fusion: Local Guides Global. Springer Nature Switzerland; 2022. p. 110–120. 10.1007/978-3-031-16437-8_11

[29] Rahman MM, Marculescu R. Medical Image Segmentation via Cascaded Attention Decoding. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE; 2023. p. 6211–6220. 10.1109/wacv56688.2023.00616

[30] Tomar NK, Jha D, Bagci U. Prototype Learning for Out-of-Distribution Polyp Segmentation. arXiv; 2023. 10.48550/ARXIV.2308.03709

[31] Lewis J, Cha YJ, Kim J. Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images. Scientific Reports. 2023;13(1). 10.1038/s41598-023-28530-2

[32] Liu G, Yao S, Liu D, Chang B, Chen Z, Wang J, et al. CAFE-Net: Cross-Attention and Feature Exploration Network for polyp segmentation. Expert Systems with Applications. 2024;238:121754. 10.1016/j.eswa.2023.121754

[33] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv; 2015. 10.48550/ARXIV.1512.03385

[34] Zhang Y, Liu H, Hu Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. arXiv; 2021. 10.48550/ARXIV.2102.08005

[35] Farhan MH, Shaker K, Al-Janabi S. Double Dual Convolutional Neural Network (D2CNN) for Copy-Move Forgery Detection. In: 2023 15th International Conference on Developments in eSystems Engineering (DeSE). IEEE; 2023. p. 439–444. 10.1109/dese58274.2023.10100318

How to cite this article

Farhan MH, Thulnoon AA, Jasem FM. Developing an efficient polyp segmentation using residual learning in asymmetric non-local neural networks. Journal of University of Anbar for Pure Science. 2025; 19(2):177-191. doi:10.37652/juaps.2025.161190.1421