Viet Nam National University Ho Chi Minh City

University of Science

Faculty of Information Technology

Submission 4
Final

Computer Vision - CS412

21125148 - Mai Tan Dat

21125166 - Do The Phuc

December 21, 2024

1 Introduction

1.1 Introduction to Machine Learning in Image Segmentation for Medical Imaging

Image segmentation is a critical component in the field of medical imaging, enabling the extraction of meaningful information from visual data. By dividing an image into distinct regions or segments, this process facilitates the identification and analysis of anatomical structures, abnormalities, or regions of interest. Traditional methods of image segmentation often relied on manual annotation or rule-based algorithms, which were time-consuming and prone to subjectivity. However, with the advent of machine learning (ML), particularly deep learning techniques, the field has witnessed significant advancements in accuracy, efficiency, and scalability.
Machine learning models, especially convolutional neural networks (CNNs), have revolutionized image segmentation by learning hierarchical features directly from data. These models excel in tasks such as tumor detection, organ segmentation, and lesion localization. By leveraging large datasets and robust computational power, machine learning algorithms can achieve remarkable precision, often surpassing human performance in specific tasks.

1.2 Applications of Polyp Segmentation for Medical Imaging

One prominent application of machine learning in medical image segmentation is in the domain of endoscopic imaging, particularly for polyp detection and segmentation. Polyps are abnormal tissue growths that can develop in the gastrointestinal tract, and their early detection is critical for preventing conditions such as colorectal cancer. Endoscopic images pose unique challenges for segmentation due to factors like varying polyp sizes, shapes, and textures, as well as the presence of occlusions, shadows, and reflections.
Machine learning models, trained on annotated datasets of endoscopic images, can effectively address these challenges. Techniques such as U-Net[1], attention-based mechanisms, and transformers have been employed to enhance the segmentation accuracy. These models not only aid in improving diagnostic efficiency but also assist in the development of computer-aided detection (CAD)[2] systems, reducing the workload of medical professionals and increasing diagnostic consistency.

2 Literature Review

2.1 MetaFormer and CNN Hybrid Model for Polyp Image Segmentation [3]

Motivation The paper aims to solve polyp image segmentation. While CNNs efficiently capture local details and Transformers are effective for global context, each has limitations in handling the other's strengths. The authors propose a hybrid model to overcome these issues by combining both approaches.

Method The proposed model, RAPUNet, utilizes CAFormer (a MetaFormer) as a Transformer backbone, with a custom convolutional block called RAPU (Residual and Atrous convolution in Parallel Unit) to enhance detailed local information. This architecture balances local detail and

global context. The combination of CAFormer and RAPU blocks form an encoder, while the aggregation module acts as a decoder, using the local and high-level features from the encoder to create segmentation mask.

2.2 Using DUCK-Net for Polyp Image Segmentation [4]

Motivation Researchers have developed various deep learning architectures to improve the accuracy and efficiency of polyp segmentation, including U-Net[1] and their variants. While these methods can achieve precise segmentation results, their performance may be less robust when faced with a wide range of polyp characteristics. The authors propose a novel supervised convolutional neural network architecture for image segmentation that uses the encoder-decoder structure of the U-Net[1] architecture with some significant differences.

Method The proposed polyp segmentation solution consists of two novel main components. The first is a novel convolutional block called DUCK that uses six variations of convolutional blocks in parallel to allow the network to train whichever it deems best. The second novel contribution keeps the low-level details by adding a secondary U-Net[1] downscaling layer that does not process the image, so it keeps the low-level details intact.

2.3 EMCAD: Efficient Multi-scale Convolutional Attention Decoding for Medical Image Segmentation[2]

Motivation Traditional U-shaped CNNs (like U-Net[1] and its variants) and attention-based models have achieved notable success but are computationally expensive, limiting their practical application in resource-constrained environments. Vision transformers address some limitations by capturing global dependencies, yet they struggle with local spatial details and are still computationally demanding when combined with convolutions. The authors introduce an efficient multi-scale convolutional attention decoder that enhances feature maps at multiple resolutions, reduces computational cost, and integrates local and global attention through large-kernel grouped convolutions.

Method EMCAD[2] is an efficient multi-scale convolutional attention decoder for medical image segmentation, designed to refine spatial representations while minimizing computational costs. It features a MSCAM(Multi-scale Convolutional Attention Module) that uses depth-wise convolutions to capture multi-scale features and suppress irrelevant regions. Additionally, a Large-kernel Grouped Attention Gate fuses refined features with skip connections using large-kernel convolutions, enhancing local context capture. With minimal parameters and low computational cost, EMCAD[2] can be paired with various hierarchical vision encoders to achieve superior segmentation performance across multiple benchmarks.

2.4 EffiSegNet: Gastrointestinal Polyp Segmentation through a Pre-Trained EfficientNet-based Network with a Simplified Decoder[5]

Motivation The paper aims to improve colon polyp detection for colorectal cancer diagnosis, as current manual methods have significant miss rates. Despite the success of deep learning in medical image analysis, transfer learning approaches for colon polyp segmentation have been underutilized,

with many models still trained from scratch. Existing methods often use symmetric U-shaped networks, which can be inefficient. The authors propose EffiSegNet, a novel architecture that uses EfficientNet as the encoder and a simplified decoder to reduce complexity while improving performance, aiming to surpass current state-of-the-art models.

Method This work introduces EffiSegNet, a novel segmentation framework leveraging transfer learning with a pre-trained CNN classifier as its backbone. Deviating from traditional architectures with a symmetric U-shape, EffiSegNet simplifies the decoder and utilizes full-scale feature fusion to minimize computational cost and the number of parameters.

3 Dataset

Polyp segmentation datasets, including Kvasir-SEG [6], CVC-ClinicDB [7], CVC-ColonDB [8], EndoScene-CVC300 [9], and ETIS-LaribPolypDB [10], are used to evaluate the performance, as they are commonly used in polyp segmentation research.

Dataset	Number of images (train/test)	Resolution
Kvasir-SEG	1000 (900/100)	332x489 to 1920x1072
CVC-ClinicDB	612 (550/62)	384x288
CVC-ColonDB	380 (0/380)	574x500
EndoScene-CVC300	60 (0/60)	574x500
ETIS-LaribPolypDB	196 (0/196)	1255x966

Table 1: Properties of the datasets

4 Method

4.1 RAPUNet

RAPUNet Overall Architecture The model consist of 2 components: the encoder and the aggregation module. In the encoder, the backbone CAFormer is used for extracting image feature (mainly global). Since local detail is the weakness of transformer, custom convolutional block RAPU is created to solve this. The aggregation module consists of 2 parts: High-level feature aggregation and Low-level feature aggregation. An element-wise addition is applied to the output of these 2 parts to combine them, and after that is a sigmoid function to predict the segmentation mask.

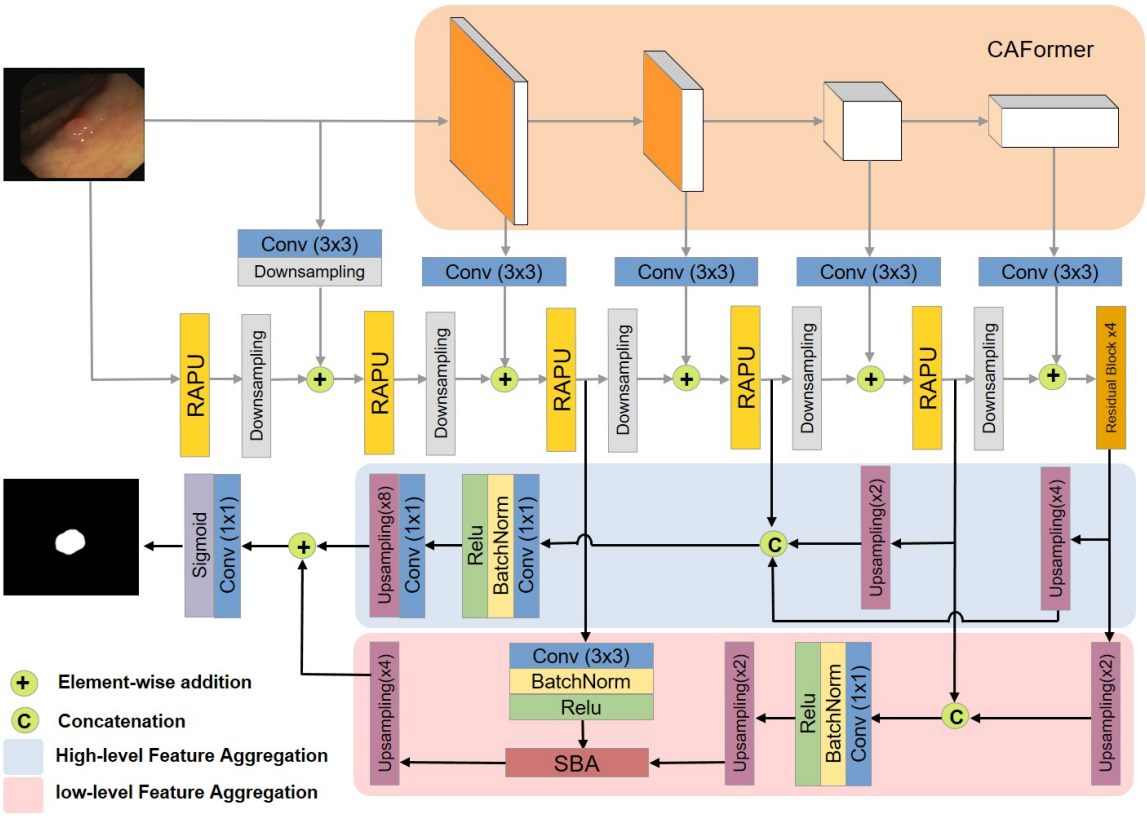

Figure 1: RAPUNet architecture

Encoder CAFormer has a 4-stage framework and employs separable depth-wise convolutions as token mixers in the first two stages and attention modules in the last two stages. The output from each CAFormer layer is merged with the corresponding output from each CNN layer through element-wise addition. The CNN component is constructed using the proposed RAPU units and down-sampling convolutions with stride 2.
RAPU is designed to address the limited low-level detail provided by the CAFormer backbone. The atrous block in RAPU capture the complex surrounding context, while the residual block preserve the small details.

Aggregation Module The proposed aggregation module is a modification of the aggregation module in DuAT [11].
Low-level skip connections are aggregated using SBA to capture detailed information, such as distinct boundaries and small objects. High-level skip connections are combined with the encoder's output to enhance semantic information.
The SBA module ensures a strong combination and precise refinement of features. It receives the output from the first CAFormer layer and integrates it with high-level semantic information from the encoder output and the highest skip connection. The SBA module includes two distinct Re-Calibration Attention Unit (RAU) operations, each addressing missing information in the input data. The RAU filters out irrelevant information from the lower-level input by leveraging contextual details from the high-level input. Conversely, for the high-level input, the RAU incorporates detailed information from the lower-level input, producing fine-grained contours and effectively combining shallow and deep features.

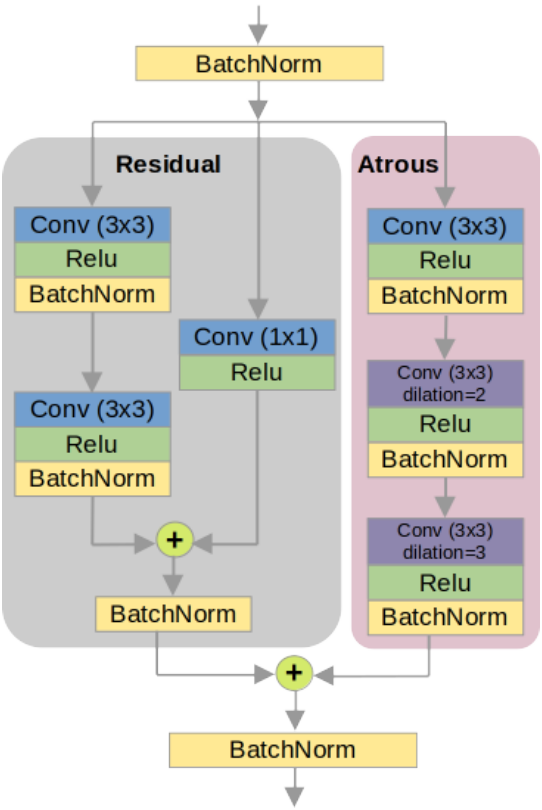

Figure 2: RAPU block

High-level skip connections from the second and third CAFormer layers, along with the encoder output, are concatenated to create a high-level feature map. Concatenation is chosen over element-wise addition to maintain spatial information. Lastly, the output from the SBA is combined with the high-level feature map during the prediction stage.

4.2 U-RAPUNet

In the architecture of RAPUNet (Figure 1), the aggregation module is split into two threads, high-level and low-level features. We suggest that this is unnecessarily complicated and propose a new design for the aggregation module based on the UNet [1] architecture. We call it U-RAPUNet.

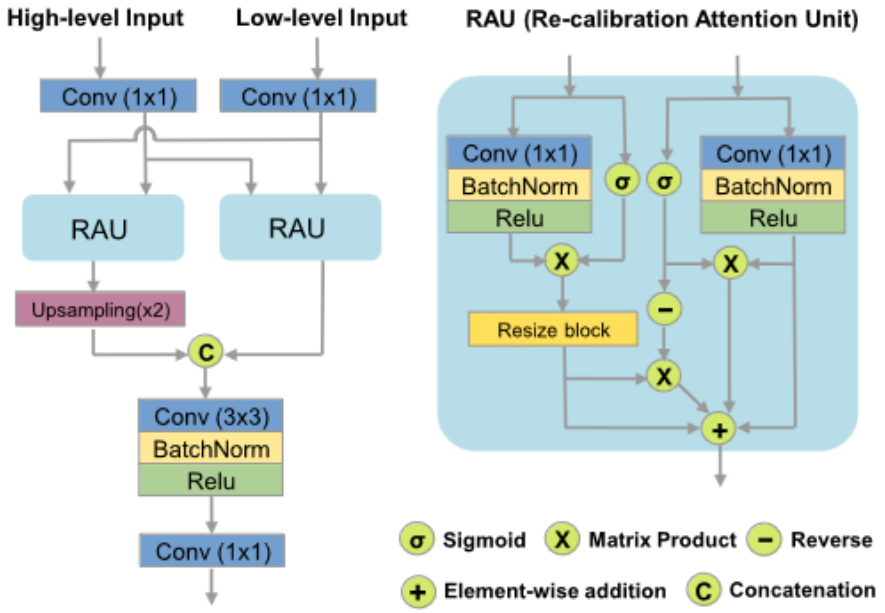

Figure 3: Selective Boundary Aggregation (SBA)

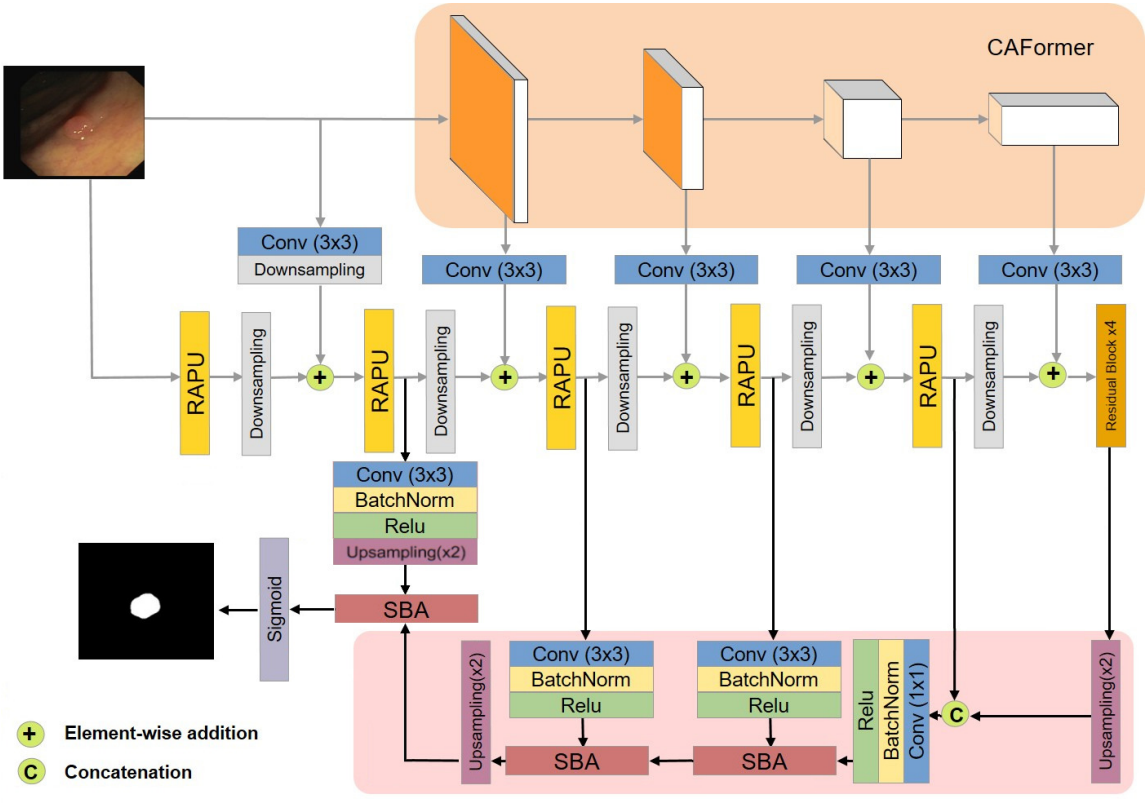

Figure 4: U-RAPUNet

From the encoder, there are 5 outputs to the aggregation module. The first 3 outputs are local

features because the first part of the encoder mainly consists of convolutional layers. The last 2 outputs are global features since the last 2 blocks of the CAFormer [12] backbone are transformers. The global features are concatenated.
There are 3 aggregating steps corresponding to 3 local feature outputs. At each aggregating step, higher-level and lower-level features are merged by a Selective Boundary Aggregation (SBA) module. The output of the intermediate SBAs can also be used for prediction, allowing deep supervision training.
SBA is proposed in the paper DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation [11]. It selectively aggregates the boundary information from low-level features and semantic information from high-level features [11]. SBA consists of 2 Re-calibration attention unit (RAU) blocks that "adaptively picks up mutual representations from two inputs before fusion" [11]. The SBA and RAU in RAPUNet is slightly different from the ones in DuAT, but the idea is still the same.
The implementation for U-RAPUNet can be found at github. The notebook where we run it can be found at kaggle.

5 Experiment and evaluation

We perform deep supervision training with the 3 outputs from 3 SBA modules. Each output goes through a sigmoid function to produce prediction. The loss function is dice loss. For $loss_0$, $loss_1$, $loss_2$ correspond the output in descending order of resolution:

$$loss = 0.6 \times loss_0 + 0.3 \times loss_1 + 0.1 \times loss_2$$

The hyper-parameters for the training process are the same as the ones we use to train RAPUNet. We use the AdamW optimizer with initial learning rate and weight decay both are 10^{-4}, min learning rate and weight decay are 10^{-6}. We employ ReduceLROnPlateau as our lr scheduler with patience is 10 epochs and decay factor is 0.2. Our weight decay scheduler is PolynomialLR with power = 0.2 and total steps is 1000.
The result of U-RAPUNet is better than the RAPUNet that we trained on most datasets except for EndoScene-CVC300.

Dataset	**U-RAPUNet**	Our RAPUNet	The original RAPUNet
CVC-ClinicDB	**0.840**	0.816	0.961
Kvasir-SEG	**0.885**	0.870	0.939
EndoScene-CVC300	**0.842**	0.846	0.906
CVC-ColonDB	**0.749**	0.744	0.776
ETIS-LaribPolypDB	**0.569**	0.557	0.879

Table 2: Test result in mean Dice

6 Conclusion

In this report, we explored the application of machine learning in polyp segmentation. Our proposed U-RAPUNet model, based on RAPUNet, show positive results in various datasets. This highlights

the importance of integrating both global and local feature representations in medical image analysis tasks. Despite the advancements, challenges persist in achieving consistent precision in difficult datasets, such as ETIS-LaribPolypDB.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[2] Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. *EMCAD: Efficient Multi-scale Convolutional Attention Decoding for Medical Image Segmentation*. 2024. arXiv: 2405.06880 [eess.IV]. URL: https://arxiv.org/abs/2405.06880.

[3] Hyunnam Lee and Juhan Yoo. "MetaFormer and CNN Hybrid Model for Polyp Image Segmentation". In: *IEEE Access* 12 (2024), pp. 133694–133702. DOI: 10.1109/ACCESS.2024.3461754.

[4] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. "Using DUCK-Net for polyp image segmentation". In: *Scientific Reports* 13.1 (June 2023). ISSN: 2045-2322. DOI: 10.1038/s41598-023-36940-5. URL: http://dx.doi.org/10.1038/s41598-023-36940-5.

[5] Ioannis A. Vezakis et al. *EffiSegNet: Gastrointestinal Polyp Segmentation through a Pre-Trained EfficientNet-based Network with a Simplified Decoder*. 2024. arXiv: 2407.16298 [eess.IV]. URL: https://arxiv.org/abs/2407.16298.

[6] Debesh Jha et al. "Kvasir-seg: A segmented polyp dataset". In: *International Conference on Multimedia Modeling*. Springer. 2020, pp. 451–462.

[7] Jorge Bernal et al. "WM-DOVA Maps for Accurate Polyp Highlighting in Colonoscopy: Validation vs. Saliency Maps from Physicians". In: *Computerized Medical Imaging and Graphics* 43 (Mar. 2015). DOI: 10.1016/j.compmedimag.2015.02.007.

[8] David Vázquez et al. *A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images*. 2016. arXiv: 1612.00799 [cs.CV]. URL: https://arxiv.org/abs/1612.00799.

[9] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information". In: *IEEE Transactions on Medical Imaging* 35.2 (2016), pp. 630–644. DOI: 10.1109/TMI.2015.2487997.

[10] Jorge Bernal et al. "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge". In: *IEEE Transactions on Medical Imaging* 36.6 (2017), pp. 1231–1249. DOI: 10.1109/TMI.2017.2664042.

[11] Feilong Tang et al. *DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation*. 2022. arXiv: 2212.11677 [cs.CV]. URL: https://arxiv.org/abs/2212.11677.

[12] Weihao Yu et al. *MetaFormer Is Actually What You Need for Vision*. 2022. arXiv: 2111.11418 [cs.CV]. URL: https://arxiv.org/abs/2111.11418.