Development of double U-Net Model using depthwise separable convolution and convolutional block attention module (CBAM) for polyp medical image segmentation

Wiharto Wiharto [a,1,*], Maulana Daffa 'Athaullah Yahya [a,2], Umi Salamah [a,3]

[a] Department of Informatics, Universitas Sebelas Maret, Surakarta, Indonesia
[1] wiharto@staff.uns.ac.id; [2] maulanadaffa@student.uns.ac.id; [3] umisalamah@staff.uns.ac.id
* corresponding author

ARTICLE INFO

Article history
Received September 30, 2024
Revised December 15, 2025
Accepted April 10, 2026
Available online August 31, 2026

Keywords
Colorectal cancer
CBAM
Double U-Net
MobileNetV2
Polyp segmentation

ABSTRACT

One of the most common types of cancer is colorectal cancer, which is commonly caused by colon polyps. Early detection via colonoscopy plays a crucial role in preventing the advancement of cancer, even though initial symptoms may not always be noticeable. In this context, computer-assisted methods for image segmentation in colonoscopy significantly aid in polyp detection, reduce the workload on healthcare professionals, and enhance diagnostic precision. A widely used approach in medical image segmentation is the U-Net model, known for its ability to merge information across multiple resolutions through its encoder-decoder structure. Nonetheless, several U-Net-derived architectures struggle to effectively capture deeper features and maintain computational efficiency. This research introduces a modified Double U-Net architecture that integrates MobileNetV2 as its encoder to alleviate computational complexity, while incorporating depthwise separable convolutions and Convolutional Block Attention Modules (CBAM) in the skip connections to enhance both accuracy and efficiency. The proposed model demonstrated notable performance, achieving an F1-score of 95.16% and IoU of 90.63% on the CVC-Clinic DB dataset, and an F1-score of 89.04% with an IoU of 80.47% on the Kvasir-SEG dataset, with a parameter count of 1.96 million. This approach not only increases segmentation precision but also maintains computational efficiency, making it a viable option for detecting colorectal polyps.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Colorectal cancer ranks among the leading types of cancer in Indonesia, with its incidence rising each year [1]. Based on the GLOBOCAN 2022 data, it is the fourth most prevalent cancer, with 35,676 reported cases and a mortality rate of 7.9% [2]. Detecting colorectal polyps early is critical, but it is often hampered by the subtlety or absence of symptoms, making them easy to overlook until the condition worsens [3], [4]. Colonoscopy is widely used for polyp detection, and the resulting images are commonly processed through segmentation techniques to aid in analysis [5]. Therefore, automatic segmentation systems play a pivotal role in improving detection efficiency and facilitating faster and more accurate diagnoses [6].

Techniques like Convolutional Neural Networks (CNN) are vital in automatically isolating the Region of Interest (RoI) for polyp detection in medical images [7], [8]. Nonetheless, key obstacles in

medical image segmentation include the scarcity of high-quality annotated datasets, inconsistencies across patient scans, and poor image clarity [9]. U-Net, a prominent CNN-based architecture, has proven highly effective in addressing these issues. Its encoder-decoder structure extracts critical contextual information while restoring image resolution. The use of skip connections ensures that significant details are preserved, enhancing the precision of segmentation [10].

The U-Net architecture has evolved to become even more powerful for medical image segmentation tasks [11]. An example is DDANet, which incorporates two parallel decoders along with residual and SE blocks. This model achieved a dice coefficient of 0.7874 and an Intersection over Union (IoU) score of 0.7010 on the Kvasir-SEG dataset. However, the increased architectural complexity raises the demand for computational resources and makes optimization more challenging [12]. Another architecture, ResUnet++, developed by Jha et al. [13], combines residual units, SE blocks, ASPP, and attention mechanisms. This resulted in a dice score of 81.33% and an IoU of 79.27% on the Kvasir-SEG dataset, and a dice score of 79.55% with an IoU of 79.62% on the CVC Clinic DB dataset.

Although ResUnet++ demonstrates notable improvements, it still leaves room for enhancing the ability to capture information across different scales and optimizing computational performance. Double U-Net tackles these issues by stacking two U-Net layers sequentially, allowing for deeper feature extraction and a better grasp of contextual information that may be missed by the standard U-Net. The inclusion of a pre-trained VGG-19 encoder, combined with ASPP and SE blocks, further aids in capturing more comprehensive contextual information [9].

Double U-Net has been evaluated using the CVC-Clinic DB dataset, where it obtained a dice coefficient of 87.81% and an IoU of 78.81%. Although these outcomes are encouraging, VGG-19 has certain limitations, including its large model size and less-than-optimal computational efficiency, which may hinder performance and convergence speed [7]. Sultana [14] reported that MobileNetV2 achieved an accuracy of 98.67% in medical image detection and classification, surpassing VGG-19 with 98.20% accuracy and ResNet50 with 97.67%. Bor et al. [15] discovered that combining MobileNetV2 with LSTM yielded the highest F1-score of 98.30%, outperforming other pre-trained models, such as MobileNetV1, DenseNet121, DenseNet169, ResNet50, and ResNet101. MobileNetV2's advantage is its capability to maintain high accuracy by utilizing depthwise separable convolution, which minimizes computational complexity without diminishing performance [16].

Depthwise Separable Convolution (DWSC) is known for being more computationally efficient while capturing more comprehensive features compared to traditional convolutions [17]. In the PSNet architecture, DWSC was used in the encoder to enhance local feature extraction, proving effective in segmenting intricate polyps and resulting in a dice score of 92.8% and an IoU of 87.9% on the CVC-Clinic DB dataset [18]. Similarly, in MDSU-Net, DWSC was substituted for standard convolutions in the encoder, decreasing the parameter count from 34.5 million to 23.5 million, thus accelerating training and improving the dice coefficient by 0.039 in medical images [19]. This underscores that DWSC not only boosts computational efficiency but also enhances accuracy in medical image segmentation.

In the context of medical image segmentation, U-Net frequently faces challenges in capturing essential details, which can lead to reduced accuracy when dealing with more intricate images. Zhao [20] developed SCAU-Net by incorporating spatial and channel attention modules, drawing inspiration from CBAM, into the U-Net framework. This modification enhanced the dice score by 1% and IoU by 1.5%. Khanh et al., [21] showed that the Enhanced U-Net, incorporating a spatial-channel attention mechanism in the skip connection, enhances the encoder's capability to capture contextual details and reduces the disparity between features in the encoder and decoder.The inclusion of scAG raised the dice coefficient to 71.72%, outperforming U-Net and U-Net++ in the task of medical image segmentation, showcasing the notable potential of combining spatial and channel attention mechanisms for optimal results.

Based on the above studies, the research identifies several key gaps and proposes novel approaches to address them.

- Computational efficiency: The large size and demands of VGG-19-based models limit their use. This study replaces the first encoder with MobileNetV2 for higher efficiency without compromising performance.
- Segmentation performance: U-Net's standard convolution layers are computationally heavy. Replacing them with Depthwise Separable Convolutions (DWSC) reduces complexity while maintaining accuracy.
- Feature processing: Standard U-Net models struggle with deep contextual information. This research uses CBAM in skip connections to improve focus on crucial channel and spatial features.

The study aims to create a medical image segmentation model that is both efficient and accurate. It focuses on improving the model's capacity to understand more complex and relevant information, enabling automatic detection and identification of polyps.

2. Method

This study is systematically organized, beginning with the "Dataset" and "Dataset Processing," followed by the "Model Architecture Design." It then covers "Hyperparameter Tuning" to optimize parameters before "Model Training," and concludes with "Performance Evaluation" to assess accuracy and effectiveness. The complete process is illustrated in Fig. 1.

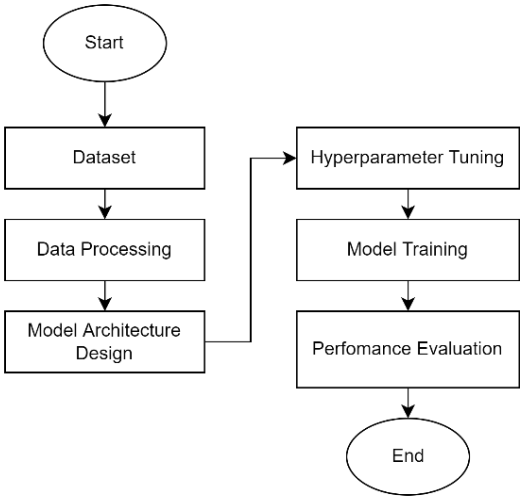


Fig. 1. Research Flow Diagram

2.1. Dataset

This study uses two datasets, namely CVC-Clinic DB and Kvasir-SEG. CVC-Clinic DB was developed by Centro de Investigación Biomédica en Red (CIBER-BBN) and consists of 612 polyp images with a resolution of 288 x 368 pixels taken from colonoscopy procedures. The images have been manually labeled to highlight the position and boundaries of the polyps [22]. Kvasir-SEG is provided by the Multimedia Signal Processing group at Simula Research Laboratory and contains 10,000 polyp images with resolutions varying from 332 x 487 to 1920 x 1072 pixels that are also manually annotated by medical experts [23]. Example images and ground truth from both datasets can be seen in Fig. 2 and Fig. 3.

2.2. Dataset Processing

The dataset is resized to 256 x 256 pixels for consistency. Data augmentation is performed using 25 techniques, including cropping, rotations, flips, distortions, and adjustments to brightness, contrast, and color, as well as blurring, noise, and dropout strategies. These augmentations increase data variety, improve model robustness, and reduce overfitting [9]. The processed dataset is split into 80% training, 10% validation, and 10% testing sets. Table 1 and Fig. 4 display the dataset split and augmentation

outcomes, with Fig. 4 (a) showing Center Crop, Fig 4. (b) Grayscale, Fig 4. (c) Channel Shuffle, and Fig 4. (d) Coarse Dropout.

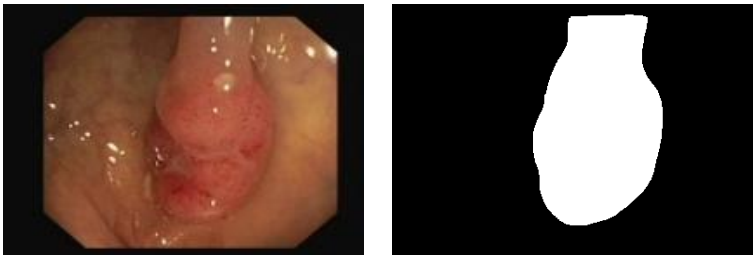


Fig. 2. Polyp image and ground truth from CVC-Clinic DB Dataset

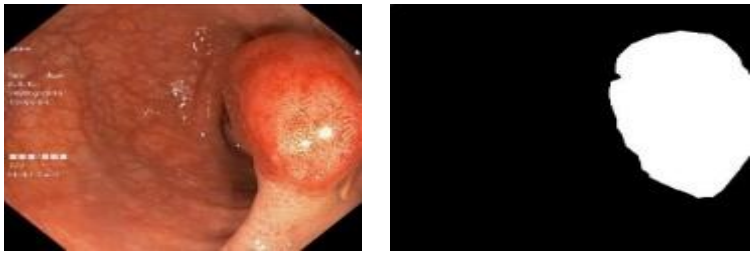
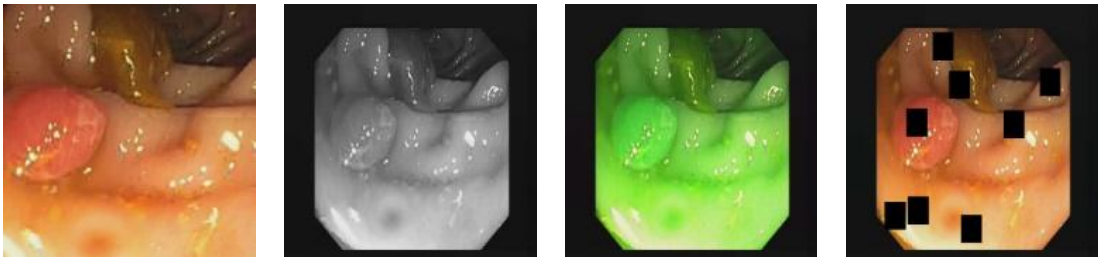


Fig. 3. Polyp image and ground truth from Kvasir SEG Dataset



(a) (b) (c) (d)

Fig. 4. Image augmentation techniques: Center Crop (a), Grayscale (b), Channel Shuffle (c), and Coarse Dropout (d)

Table 1. Dataset split composition

Dataset	Total	Training	Validation	Testing
CVC-ClinicDB	612	490	61	61
Kvasir-SEG	1000	800	100	100

2.3. Model Architecture Design

To improve the accuracy of medical image segmentation, especially in colonoscopy images, the study employed a Double U-Net architecture. This model was selected due to its dual encoder-decoder structure, allowing for more comprehensive multi-scale feature extraction [9]. It helps the model retain fine details while integrating multi-resolution information, which is crucial for detecting small polyps often missed by conventional U-Net models. Details of the architecture are shown in Fig. 5.

This Double U-Net architecture is enhanced with MobileNetV2 as its first encoder, DWSC layer, and CBAM on skip connections. MobileNetV2 was selected due to its efficiency in feature extraction, utilizing inverted residuals and DWSC layers, allowing for a reduction in parameters while maintaining performance [16]. To improve computational performance and allow the model to handle more detailed data with less computational demands, DWSC is also used in the convolution block [17].

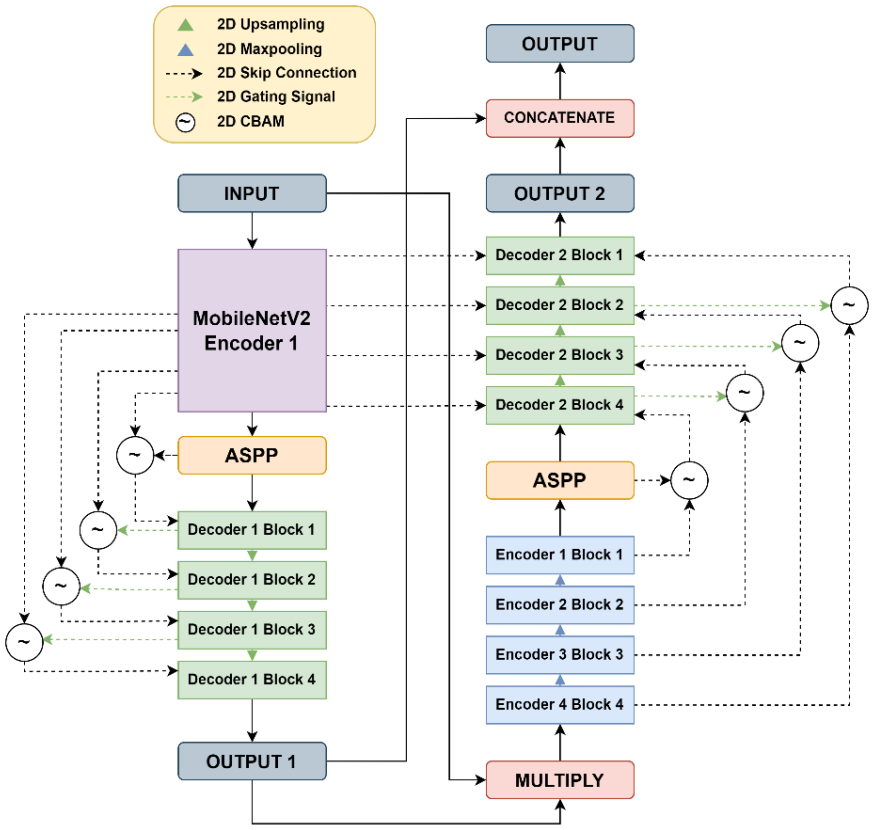


Fig. 5. Composition of proposed Double U-Net with DWSC and CBAM architecture

The convolution block architecture of the encoder and decoder in Fig. 6 processes the input data through several stages to produce a better feature representation. The process starts with depthwise separable convolution using a 3x3 filter, which consists of two stages: depthwise convolution to handle each channel separately and pointwise convolution to combine the results. This 3x3 filter effectively captures local patterns in the image. Subsequently, the features undergo normalization to expedite the process of convergence, and the activation function incorporates non-linearity to enable the model to comprehend intricate patterns [24].

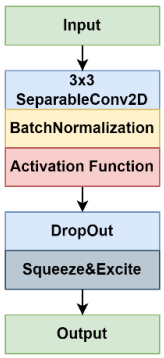


Fig. 6. Encoder and decoder block structure

By randomly deactivating some neurons during training, dropout reduces overfitting and increases the model's noise resistance. Squeeze and excite blocks then amplify important features by adjusting the weights between channels. This approach improves the accuracy in important feature detection and makes the architecture more efficient and accurate for image segmentation and classification tasks.

Meanwhile, CBAM is applied to skip connections to ensure the model focuses on relevant features, both spatially and channel-wise, to improve polyp segmentation in colonoscopy images. Channel

attention and spatial attention are progressively integrated to enhance the feature representation by employing CBAM to skip connections. Channel attention amplifies important features at the channel level, while spatial attention amplifies important locations at the spatial level [21]. Here in Fig. 7 is the architecture of CBAM.

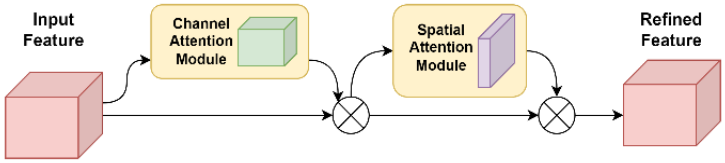


Fig. 7. CBAM architecture mechanism

As shown in Fig. 8, channel attention in CBAM quantifies the weight assigned to each channel in the input feature with dimensions (H,W,C), where H stands for height, W for width, and C for the total number of channels. The process begins with global average pooling and max pooling, both producing a vector with dimensions (1,1,C). These vectors are passed through two dense layers with shared weights: the first reduces the channel dimension using ReLU activation, and the second restores the original channel size. The results are summed and activated using a sigmoid function to generate a channel attention map with dimensions (1,1,C). The appropriate channels are then strengthened by multiplying this map element-by-element with the original input features, providing an output with the same dimensions [20].

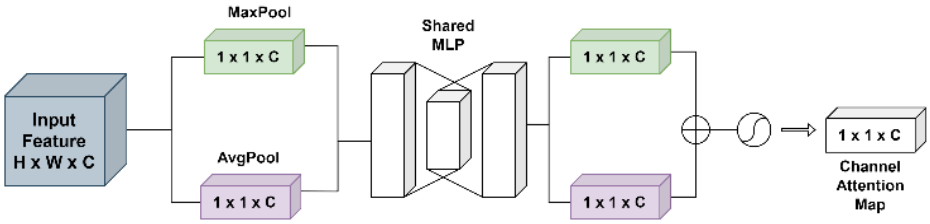


Fig. 8. Layers structure of channel attention in CBAM

Spatial attention in CBAM, as shown in Fig. 9, focuses on the spatial location of the most relevant features. Maps with dimensions (H,W,1) are produced by processing input features with dimensions (H,W,C) using average pooling and max pooling along the channel dimension.

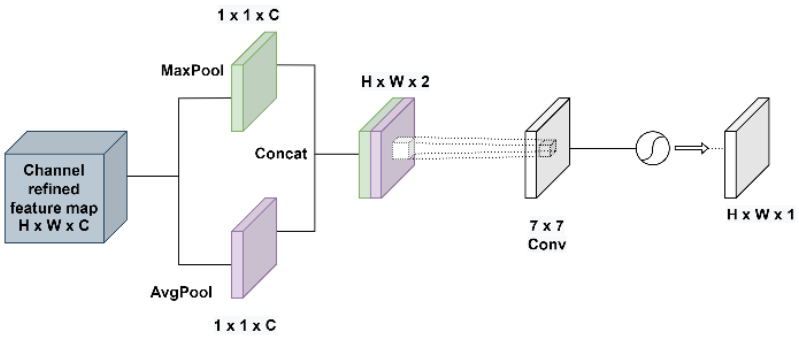


Fig. 9. Layers structure of channel attention in CBAM

These two maps are then combined into a single map with dimensions (H,W,2), which is passed through a convolution layer with a 7x7 kernel and sigmoid activation. Important spatial locations are amplified by multiplying the resulting spatial attention map, which has dimensions (H,W,1), element-wise with the original input characteristics. This process ensures that only the most relevant spatial features are highlighted, supporting better segmentation accuracy [25]. The total segmentation

performance of the model is improved by the channel and spatial attention combined in CBAM, which ensures that the processed features are more informative.

2.4. Hyperparameter Tuning

Hyperparameter tuning in this study is performed to enhance model performance by adjusting key parameters, including the optimizer, learning rate, activation function, and loss function. Experiments were conducted to test various configurations and determine the most optimal combination. This approach helps improve the accuracy and overall performance of the model in image segmentation tasks.

2.5. Model Training

The training process involves compiling the neural network with an appropriate optimizer, loss function, and evaluation metrics, followed by forward and backward passes to update weights based on loss gradients. Validation data is used to monitor performance and prevent overfitting with techniques like model checkpointing and learning rate scheduling [26]. When learning rate scheduling is used, the learning rate is lowered by a factor of 0.1 if the validation loss fails to go lower for five consecutive epochs. After completing the specified epochs or when validation performance plateaus, the model is tested on unseen data to ensure generalization. The study utilizes the Kaggle platform with a P100 GPU, Python, and TensorFlow 2.6.0. The initial experiment configuration can be seen in Table 2.

Table 2. Initial training configuration

Hyperparameter	Value
Batch	8
Activation Function	ReLU
Loss Function	Dice Loss
Epochs	35
Optimizer	Nadam
Learning Rate	0.0001

2.6. Performance Evaluation

The evaluation of model results is conducted by assessing performance on test data using key metrics, specifically F1-Score and Intersection over Union (IoU). These metrics are crucial for assessing the general performance and segmentation abilities of the model. The evaluation results are based on the formulas provided in Equations (1) and (2).

$$F1\ Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \tag{1}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{2}$$

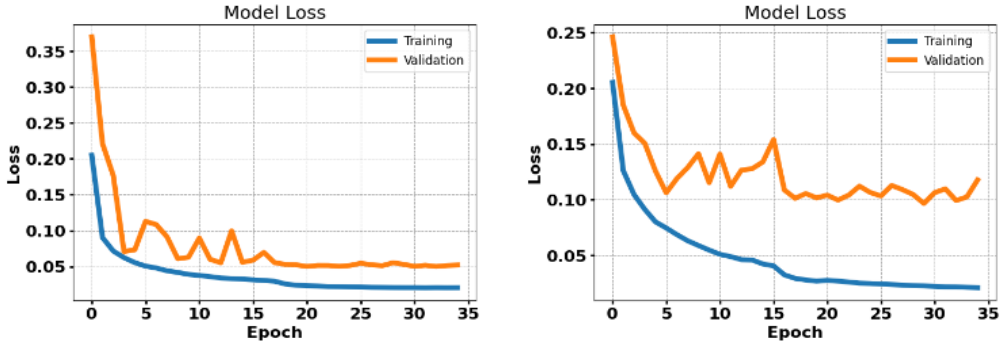
3. Results and Discussion

After performing hyperparameter tuning, the optimizer, learning rate, activation function, and loss function are finalized for the best combination. The selected configuration, shown in Table 3, yielded the lowest loss value.

Table 3. Hyperparameter Tuning Results

Hyperparameter	Value
Batch	8
Activation Function	ReLU
Loss Function	Dice Loss
Epochs	35
Optimizer	RMSprop
Learning Rate	0.001

Fig. 10 (a) presents the loss function graph for the CVC-Clinic DB dataset, while Fig. 10 (b) shows the loss for the Kvasir-SEG dataset. The results demonstrate that the proposed approach generalizes well to the validation data and learns efficiently from the training set.



(a) (b)

Fig. 10. Comparison of training and validation loss of CVC Clinic DB (a) and Kvasir SEG (b) Datasets

Based on the figure above, for the CVC-Clinic DB dataset, the training loss decreases significantly from 0.1864 to 0.0595 between epochs 1 and 4, while the validation loss drops from 0.7524 to 0.1178, indicating that the model quickly learns important features. However, the validation loss begins to fluctuate after epoch 5, with a spike at epoch 5 (0.7098), likely due to the increasing complexity of the validation data. For the Kvasir SEG dataset, the training loss consistently decreases from 0.2038 to 0.0238 up to epoch 34, while the validation loss also fluctuates, with a spike at epoch 28 (0.2731), but stabilizes at 0.1105 by the end of training. After training, the model was tested, and the results are shown in Table 4.

Table 4. Proposed model testing results

CVC Clinic DB		Kvasir SEG		Params (M)
F1-Score	IoU	*F1-Score*	IoU	
95.28%	90.84%	89.08%	80.69%	1.96

The proposed model effectively detects polyps in colonoscopy images, achieving high accuracy and capturing critical details. This balance of precision and sensitivity enhances its reliability for aiding diagnosis. Fig. 11 and Fig. 12 visualize the qualitative results of the proposed models segmentation.

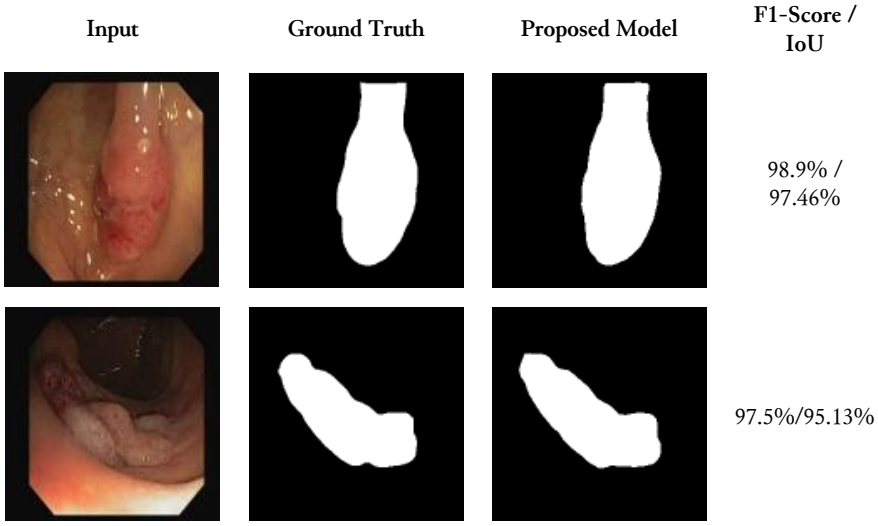


Fig. 11. Qualitative results of proposed model on CVC-Clinic DB Dataset

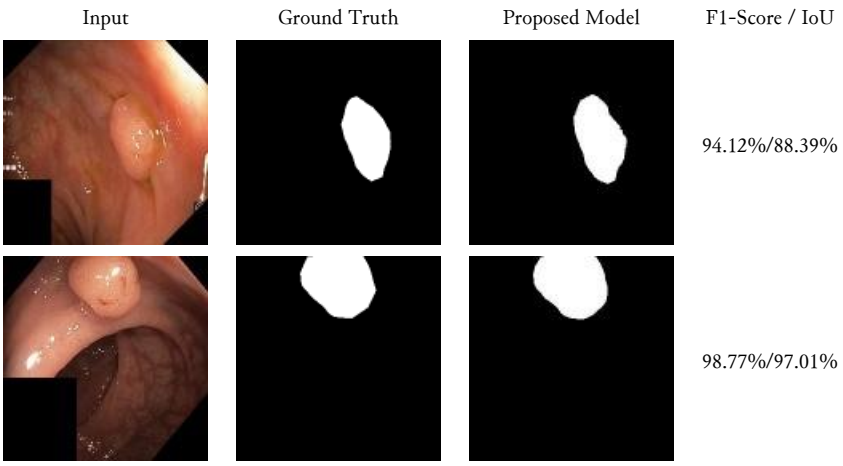


Fig. 12. Qualitative results of the proposed model on the Kvasir SEG Dataset

3.1. Comparison with the Original Double U-Net

A comparison is made between the original Double U-Net architecture and the proposed Double U-Net model. The original architecture uses two encoder-decoders with a VGG-19 encoder, without additional blocks such as DWSC or CBAM at the skip connection. The training loss and validation loss results of the original model are shown in Table 5.

Table 5. Comparison Results of Proposed Model with Double U-Net

Model	CVC Clinic DB		Kvasir SEG		Params (M)
	F1-Score	*IoU*	*F1-Score*	*IoU*	
Double U-Net	92.39%	86.11%	81.29%	73.32%	29.3
Proposed Model	**95.28%**	**90.84%**	**89.08%**	**80.69%**	**1.96**

The comparison results demonstrate that the proposed model offers significant advantages over the original Double U-Net. It achieved a 2.77% F1-Score improvement and a 4.52% IoU improvement on the CVC Clinic DB dataset. The proposed approach increased the IoU by 7.37% and the F1-Score by 7.79% for the Kvasir SEG dataset. Additionally, the proposed model achieves a reduction of 27.34 million parameters, making it much more efficient than the original Double U-Net. Fig. 13 and Fig. 14 visualize the qualitative results of the two segmentation models.

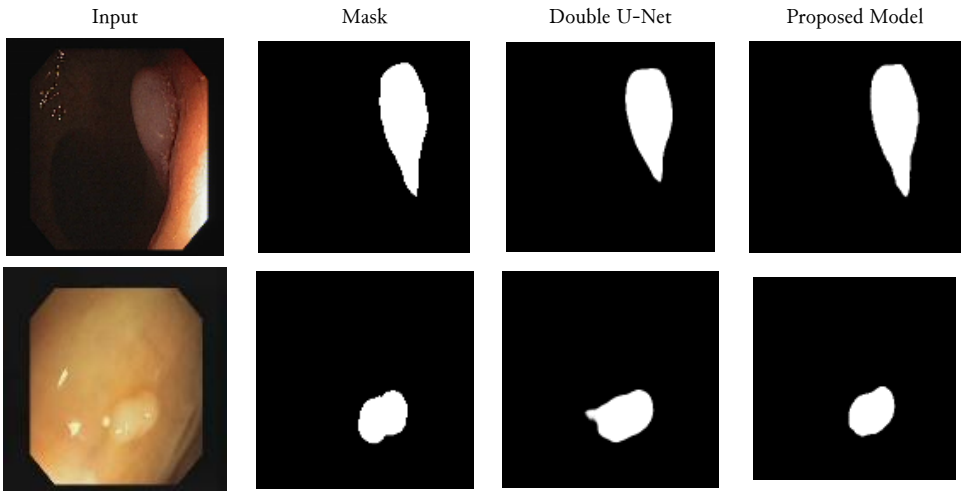


Fig. 13. Qualitative results of original Double U-net with proposed model on CVC-Clinic DB Dataset

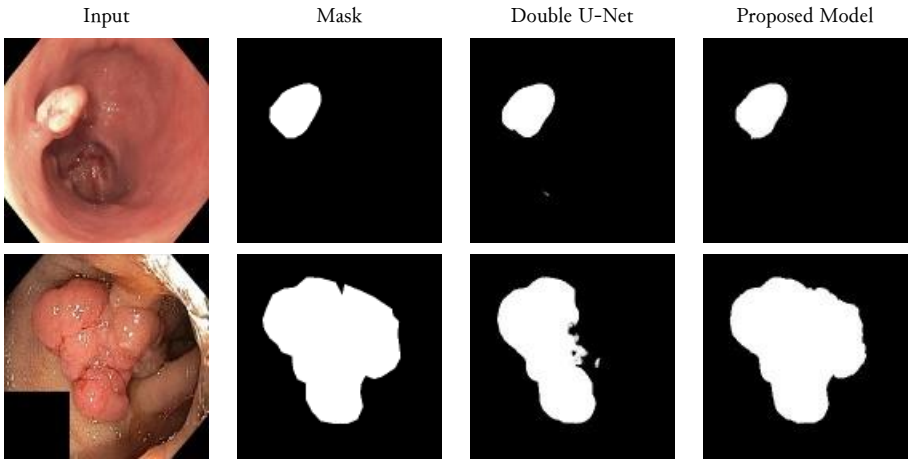


Fig. 14. Qualitative results of original Double U-net with proposed model on Kvasir SEG Dataset

3.2. Impact of pre-trained MobileNetV2 Encoder

This ablation experiment aims to evaluate the impact of replacing the MobileNetV2 encoder with VGG-19 on medical image segmentation performance, while maintaining the DWSC and CBAM layers at skip connections. After training, the model was tested on the test datasets, and the results are shown in Table 6.

Table 6. Evaluation Results of Model Comparison Excluding MobileNetV2

Model	CVC Clinic DB		Kvasir SEG		Params
	F1-Score	*IoU*	*F1-Score*	*IoU*	(M)
Double U-Net	92.39%	86.11%	81.29%	73.32%	29.3
Proposed Model	**95.28%**	**90.84%**	**89.08%**	**80.69%**	**1.96**

The results show that replacing MobileNetV2 with VGG-19 leads to a decrease in F1-Score and IoU values. In addition, there is an increase in parameters, which indicates that the model with MobileNetV2 is more computationally efficient compared to the model using VGG-19. Fig. 15 and Fig. 16 visualize the qualitative results of the two segmentation models.

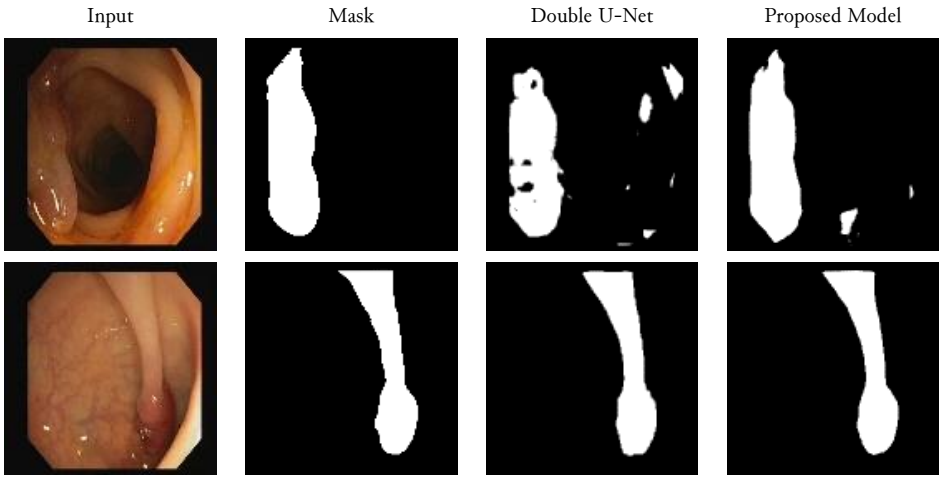


Fig. 15. Qualitative results of encoder replacement with the proposed model on CVC-Clinic DB Dataset.

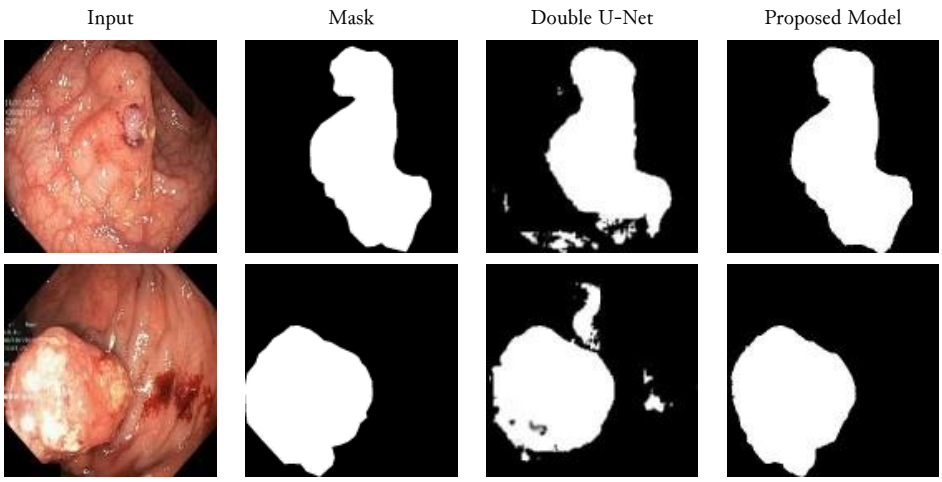


Fig. 16. Qualitative results of encoder replacement with proposed model on Kvasir SEG Dataset

3.3. Impact of DWSC

This ablation experiment aims to evaluate the impact of changing the convolution layer on medical image segmentation performance, keeping the MobileNetV2 and CBAM encoders at the skip connection. After training, the model was tested on unseen datasets, and the results are shown in Table 7.

Table 7. Evaluation Results of Model Comparison Excluding DWSC

Model	CVC Clinic DB		Kvasir SEG		Params (M)
	F1-Score	*IoU*	*F1-Score*	*IoU*	
Model without DWSC	94.74%	89.86%	86.64%	77.16%	2.32
Proposed Model	**95.28%**	**90.84%**	**89.08%**	**80.69%**	**1.96**

The results show that replacing DWSC with traditional block convolution leads to a decrease in F1-Score and IoU values. In addition, an increase in parameters indicates that the model with DWSC is more computationally efficient. Fig. 17 and Fig. 18 visualize the qualitative results of the two segmentation models.

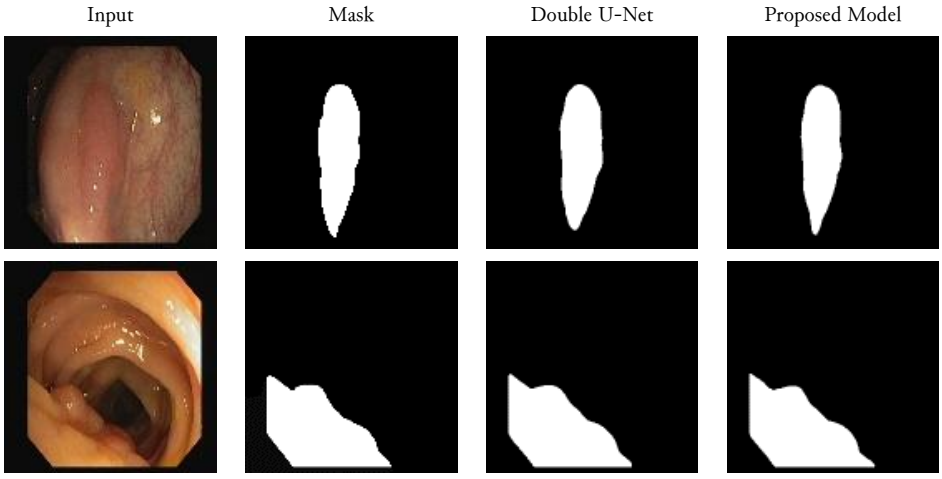


Fig. 17. Qualitative results of DWSC replacement with the proposed model on the CVC-Clinic DB Dataset

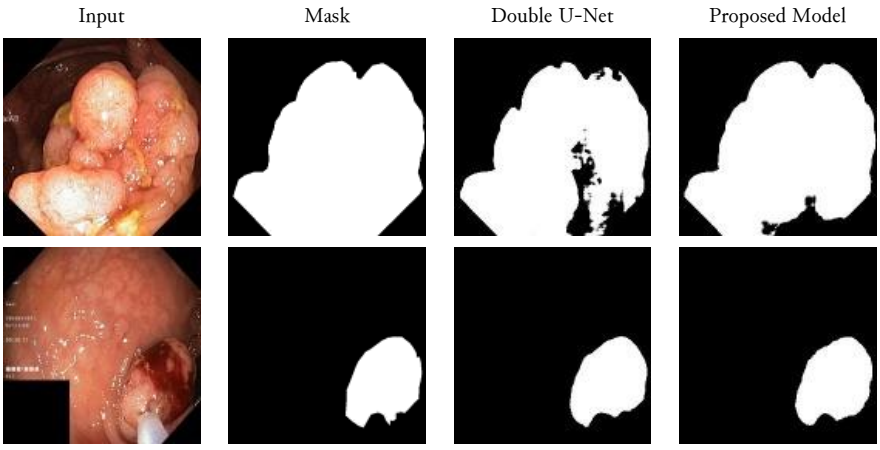


Fig. 18. Qualitative results of DWSC replacement with proposed model on Kvasir SEG Dataset

3.4. Impact of CBAM

This ablation experiment aims to evaluate the impact of removing CBAM at the skip connection on medical image segmentation performance, while maintaining the MobileNetV2 encoder and DWSC layer. After the training process, testing was conducted on unpredicted datasets, and the results are shown in Table 8.

Table 8. Evaluation Results of Model Comparison Excluding DWSC

Model	CVC Clinic DB		Kvasir SEG		Params (M)
	F1-Score	IoU	*F1-Score*	IoU	
Model without CBAM	94.32%	89.12%	87.89%	79.19%	**1.85**
Proposed Model	**95.28%**	**90.84%**	**89.08%**	**80.69%**	**1.96**

The results show that the removal of CBAM results in a decrease in F1-Score and IoU values, indicating a decrease in segmentation performance. CBAM itself plays an important role in improving the accuracy of the model by giving more attention to relevant features in skip connections. Although the addition of CBAM increases the number of parameters slightly, this increase is insignificant when compared to the positive impact on segmentation performance. Fig. 19 and Fig. 20 visualize the qualitative results of the two segmentation models.

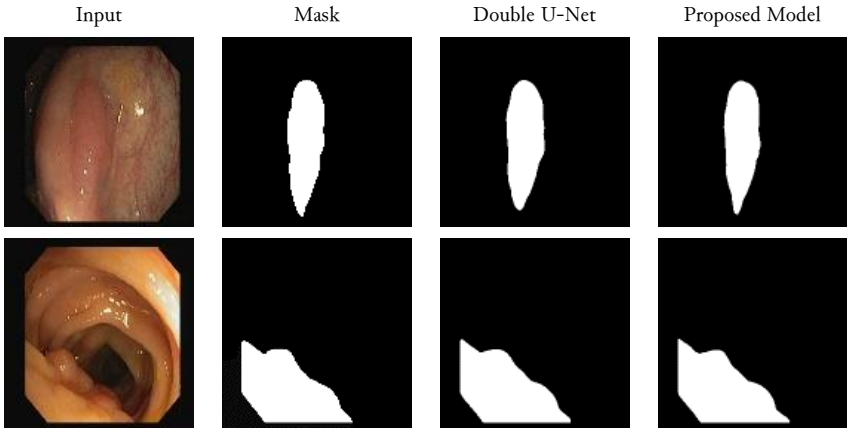


Fig. 19. Qualitative Results of CBAM Replacement with Proposed Model on CVC-Clinic DB

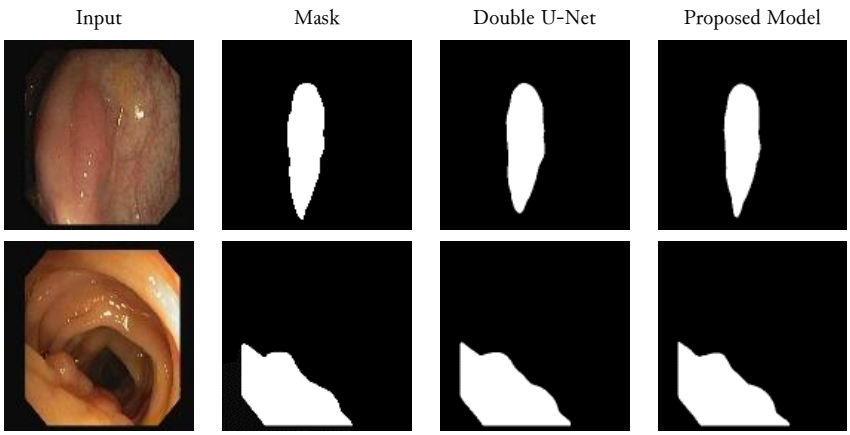


Fig. 20. Qualitative Results of CBAM Replacement with Proposed Model on Kvasir SEG

3.5. Comparison with Prior Work

The performance outcomes of the proposed approach are compared in the last phase of the model evaluation procedure, which follows a series of hyperparameter tuning and ablation experiments. A number of earlier investigations have been compared with the proposed model. Table 9 displays the comparison findings for the F1-Score, IoU, and overall number of model parameters.

The proposed model shows significant performance improvement over previous methods. On the CVC-Clinic DB dataset, F1-score and IoU increased by 2.89% and 4.73%, respectively, compared to the original Double U-Net. For the Kvasir SEG dataset, the F1-score and IoU increase by 3.91% and 6.35% respectively compared to the DCA method. The model has a parameter count of 1.96 million which is the lowest among previous studies. This shows the efficient use of resources while still improving the segmentation accuracy.

Table 9. Comparison of the Results of the Proposed Model with Previous Research

Metode	CVC Clinic DB		Kvasir SEG		Params (M)
	F1-Score	*IoU*	*F1-Score*	*IoU*	
U-Net [10]	82.30%	75.50%	81.80%	74.60%	7.11
Double U-Net [9]	92.39%	86.11%	81.29%	73.32%	29.3
ColonSegNet [27]	88.62%	82.48%	82.06%	72.39%	5.01
U-Net++ [28]	79.40%	72.90%	82.10%	74.30%	9.04
ResU-Net [29]	77.90%	n/a	79.10%	n/a	8.06
ResU-Net++ [13]	79.55%	79.62%	81.33%	79.27%	4.06
DCA [30]	90.86%	83.47%	85.16%	74.34%	30.68
Proposed Model	**95.28%**	**90.84%**	**89.08%**	**80.69%**	**1.96**

4. Conclusion

This research successfully developed a medical image segmentation model by modifying the Double U-Net architecture. These modifications include the use of MobileNetV2 pre-trained encoders to reduce model complexity, the application of DWSC, which aims to improve computational efficiency without sacrificing accuracy, and the addition of CBAM at skip connections to strengthen the focus on important features during segmentation. Experimental results show that the proposed model achieves an F1-score of 95.28% and an IoU of 90.84% on the CVC Clinic DB dataset. On the Kvasir-SEG dataset, the model achieved an F1-score of 89.08% and an IoU of 80.69%, which is superior to previous methods. In addition, the modification successfully reduced the number of parameters to about 1.96 million, which is much more efficient than the original Double U-Net study with 29.3 million parameters. Overall, this study shows that the modified Double U-Net architecture is able to improve medical image

segmentation performance while maintaining optimal computational efficiency, making it suitable for clinical applications that require a combination of high accuracy and efficiency.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] I. Labeda *et al.*, "Colorectal cancer survival rates in Makassar, Eastern Indonesia: A retrospective Cohort Study," *Ann. Med. Surg.*, vol. 74, Feb. 2022, doi: 10.1016/j.amsu.2021.103211.

[2] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/caac.21834.

[3] H. P. Irianto and A. S. Putranto, "Correlation between clinical characteristics of colorectal cancer and nutritional status in patients undergoing surgery," *Bali Med. J.*, vol. 12, no. 1, pp. 1059–1062, Apr. 2023, doi: 10.15562/bmj.v12i1.4057.

[4] E. Rondonotti *et al.*, "Clinical Validation of BASIC Classification for the Resect and Discard Strategy for Diminutive Colorectal Polyps," *Clin. Gastroenterol. Hepatol.*, vol. 18, no. 10, pp. 2357-2365.e4, Sep. 2020, doi: 10.1016/j.cgh.2019.12.028.

[5] B. C. Bag, H. K. Maity, and C. Koley, "U-Net MobileNetV2: Medical Image Segmentation Using Deep Neural Network (DNN)," *J. Mech. Contin. Math. Sci.*, vol. 18, no. 1, Jan. 2023, doi: 10.26782/jmcms.2023.01.00002.

[6] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022, doi: 10.1049/ipr2.12419.

[7] A. M. Khan, A. Ashrafee, F. S. Khan, M. B. Hasan, and M. H. Kabir, "AttResDU-Net: Medical Image Segmentation Using Attention-based Residual Double U-Net," in *2023 International Joint Conference on Neural Networks (IJCNN)*, Jun. 2023, pp. 1–8, doi: 10.1109/IJCNN54540.2023.10191528.

[8] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[9] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Jul. 2020, vol. 2020-July, pp. 558–564, doi: 10.1109/CBMS49503.2020.00111.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[11] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, doi: 10.1109/ACCESS.2021.3086020.

[12] N. K. Tomar *et al.*, "DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation," 2021, pp. 307–314, doi: 10.1007/978-3-030-68793-9_23.

[13] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*, Dec. 2019, pp. 225–2255, doi: 10.1109/ISM46123.2019.00049.

[14] Z. Sultana, M. Foysal, S. Islam, and A. Al Foysal, "Lung Cancer Detection and Classification from Chest

CT Images Using an Ensemble Deep Learning Approach," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, May 2024, pp. 364–369, doi: 10.1109/ICEEICT62016.2024.10534468.

[15] N. Bor, T. T. Sivri, N. P. Akman, A. Berkol, and Y. Ekici, "Breast Cancer Detection Using Various Classification Models Combined with Transfer Learning," in *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*, Dec. 2022, pp. 1–6, doi: 10.1109/ICAIoT57170.2022.10121840.

[16] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation," *CoRR*, vol. abs/1801.0, pp. 1–14, 2018, [Online]. Available at: https://arxiv.org/pdf/1801.04381.

[17] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," no. April 2017, 2017, [Online]. Available at: http://arxiv.org/abs/1704.04861.

[18] J. Lewis, Y.-J. Cha, and J. Kim, "Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images," *Sci. Rep.*, vol. 13, no. 1, p. 1183, Jan. 2023, doi: 10.1038/s41598-023-28530-2.

[19] Y. Zhou, X. Kang, F. Ren, H. Lu, S. Nakagawa, and X. Shan, "A multi-attention and depthwise separable convolution network for medical image segmentation," *Neurocomputing*, vol. 564, p. 126970, Jan. 2024, doi: 10.1016/j.neucom.2023.126970.

[20] P. Zhao, J. Zhang, W. Fang, and S. Deng, "SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation," *Front. Bioeng. Biotechnol.*, vol. 8, Jul. 2020, doi: 10.3389/fbioe.2020.00670.

[21] T. L. B. Khanh *et al.*, "Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging," *Appl. Sci.*, vol. 10, no. 17, p. 5729, Aug. 2020, doi: 10.3390/app10175729.

[22] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/j.compmedimag.2015.02.007.

[23] D. Jha *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," 2020, pp. 451–462. [Online]. Available at: http://link.springer.com/10.1007/978-3-030-37734-2_37.

[24] L. Nanni, S. Brahnam, M. Paci, and S. Ghidoni, "Comparison of Different Convolutional Neural Network Activation Functions and Methods for Building Ensembles for Small to Midsize Medical Data Sets," *Sensors*, vol. 22, no. 16, p. 6129, Aug. 2022, doi: 10.3390/s22166129.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11211 LNCS, Springer Verlag, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[26] S. Jahangard, M. Bonyani, and A. Khosravi, "DoubleU-Net++: Architecture with Exploit Multiscale Features for Vertebrae Segmentation," *Electr. Eng. Syst. Sci.*, pp. 1–23, Jan. 2022, [Online]. Available at: https://arxiv.org/abs/2201.12389.

[27] D. Jha *et al.*, "Real-Time Polyp Detection, Localization and Segmentation in Colonoscopy Using Deep Learning," *IEEE Access*, vol. 9, pp. 40496–40510, 2021, doi: 10.1109/ACCESS.2021.3063716.

[28] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[29] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018, doi: 10.1109/LGRS.2018.2802944.

[30] G. C. Ates, P. Mohan, and E. Celik, "Dual Cross-Attention for medical image segmentation," *Eng. Appl. Artif. Intell.*, vol. 126, p. 107139, Nov. 2023, doi: 10.1016/j.engappai.2023.107139.