Computing and Informatics, Vol. 45, 2026, 830–864, doi: 10.31577/cai_2026_4_830

RESCAM-NET: RESIDUAL CONVOLUTION AND ATROUS MULTI-SCALE NETWORK FOR ROBUST MEDICAL IMAGE SEGMENTATION

Fouzia El Abassi*

Computer and Systems Engineering Laboratory (L2IS)
Cadi Ayyad University, UCA, Faculty of Science and Technology (FST)
Av. Abdelkarim Elkhattabi, 40000, Marrakech, Morocco
&
Laboratory LAMIGEP, EMSI Marrakech, Morocco
e-mail: fouzia.elabassi@ced.uca.ma

Aziz Darouichi, Aziz Ouaarab

Computer and Systems Engineering Laboratory (L2IS)
Cadi Ayyad University, UCA, Faculty of Science and Technology (FST)
Av. Abdelkarim Elkhattabi, 40000, Marrakech, Morocco
e-mail: {a.darouichi, a.ouaarab}@uca.ac.ma

Abstract. Digital data and automated solutions have raised the standard of medical image analysis and made it more complex and challenging. Manual examination of medical images is inefficient, subjective, and error-prone due to variation in attributes such as shape, size, and texture. The majority of approaches in the literature, starting from the classical U-Net including its modern variants, usually suffer from overfitting issues, model saturation issues, and inability to capture multi-scale contextual features, reducing generalization across different imaging modalities. To address these challenges, we present ResCAM-Net, an improved encoder–decoder architecture that accurately and efficiently raises segmentation accuracy for different types of medical images. Our model incorporates residual learning, multi-kernel residual convolutions, adaptive feature recalibration using atrous spatial pyramid pooling, and a hybrid triple attention module to enhance feature aggregation and focus on critical regions. ResCAM-Net reduces the number of trainable parameters to 9.91 million, compared with other state-of-the-art architectures, by approximately

* Corresponding author

70 %, which significantly improves computational efficiency and reduces convergence time. Our model performs better on several benchmark datasets: Dice similarity coefficients of 89.84 %, 87.39 %, and 85 % on ISIC-2017 (small), ISIC-2017, and Kvasir-SEG datasets, respectively. Further, the robustness and generalization capabilities of ResCAM-Net were well reflected in these segmentation tasks, where it outperformed existing models both in accuracy and parameter efficiency.

Keywords: Medical image analysis, semantic segmentation, convolutional neural network, colonoscopy, dermoscopy

Mathematics Subject Classification 2010: 68U10, 68T05, 68T45

1 INTRODUCTION

In recent years, the whole world has been moving towards the digitization of data with the help of computer-aided algorithms, resulting in the exposure of a vast amount of information that researchers seek to explore and exploit to help mankind. This digital transformation has extended to the field of medicine as well, where advanced imaging technologies have enabled healthcare professionals to visualize and analyze intricate details of the human body. Medical imaging allows the creation of visual representations of the body's interior for early diagnosis. There are several techniques for obtaining medical images in any plane of space for early detection of cancerous and non-cancerous diseases. Among the best-known techniques are X-ray, Colonoscopy, Dermoscopy, Microscopy, Magnetic Resonance Imaging (MRI), Ultrasound, Computerized Tomography (CT) and Electrocardiogram (ECG).

The manual analysis of medical images presents challenges such as subjectivity, time-consuming, potential errors due to fatigue, the need for specialized expertise and constrained accuracy due to the significant variation in shape, dimensions, location, and texture within medical images. It can be impractical for handling large volumes of images, lacks scalability, and is costly. To mitigate these issues, healthcare institutions increasingly rely on Artificial Intelligence (AI) and computer-assisted tools, which enhance accuracy, speed, and consistency in image analysis. AI contributes to a variety of tasks, including pattern identification [1], categorization [2], and segmentation tasks such as brain and brain tumor segmentation [3], liver and liver tumor segmentation [4], cell segmentation [5], optic disc segmentation [6], as well as lung segmentation, cardiac image segmentation [7] and pulmonary nodule analysis [8], offering a promising solution to improve medical image analysis and diagnosis.

The utilization of Deep Learning techniques, notably Convolutional Neural Networks (CNNs), simplifies extraction and analysis of the image features. This specialization in feature extraction helps to accelerate the processes and significantly improve image segmentation results.

The first appearance of CNN architectures was to solve problems related to the classification of images, such as LeNet [9] and AlexNet [10], after the researchers modified these architectures by increasing the number of layers, namely VGGNet [11]. All of these architectural designs incorporate fully connected layers for classification. However, this approach is unsuitable for object segmentation tasks, where the number of objects of interest within the image segmentation task is variable and not fixed. Consequently, the output layer's dimension cannot remain constant, which prompted the emergence of fully convolutional networks, such as FCN [12], which consists of only convolution layers, allowing it to produce a segmentation map that is the same size as the input image but needs a very large amount of labelled data to produce high-resolution output images. To solve this problem, other encoder-decoder architectures have been developed, the most well-known of which are U-Net [13], SegNet [14] and DoubleU-Net [15].

The classical U-Net and its new state-of-the-art variants still face significant challenges in effectively capturing the complex output feature maps produced by convolutional network units. These are largely due to the dissimilarities in scale of the regions of interest, complexity of the contextual environment, indistinct boundaries, and diversity of textures in medical images. The models, however, usually face several challenges in terms of overfitting issues, inability to handle multi-scale contextual features, and inefficiency in segmenting various medical imaging modality structures. Most of the available models are also generally facing model saturation problems, which affect their generalization performance over various biomedical datasets.

Our contribution (ResCAM-Net) addresses these limitations by enhancing the architecture of ColonSegNet, which was purely designed for colonoscopy images, by integrating several advanced components to improve efficient segmentation in all types of medical images.

ResCAM-Net is an attentional model featuring residual learning, multi-core residual convolutions, adaptive feature recalibration via atrous spatial pyramid pooling, and a hybrid triple attention model integrated into its architecture for improved feature aggregation and precise focusing on critical regions. Furthermore, the proposed model reduces the number of trainable parameters by almost 70 % compared to state-of-the-art architectures, which significantly reduces convergence time. These are the reasons why the model, due to the novelties mentioned, can overcome the limitations of conventional architectures and improve segmentation accuracy, while being more robust and generalizable on different biomedical datasets.

The contribution of this paper is outlined as follows:

- Inclusion of a Residual Module within the convolution layers addresses concerns related to overfitting, model saturation, and notably, enhances the performance of semantic segmentation in the context of medical images.
- The integration of a Multi-Kernel Residual Convolution Module into the bottleneck dramatically widens the perspective for capturing a wide spectrum of semantic global context features across various scales.

- An adapted hybrid Triple Attention Module has been introduced, facilitating the fusion of channel- and spatial-oriented attention in conjunction with a squeeze-and-excitation-based attention mechanism. This refinement results in enhanced interdependencies among channels and spatial relationships within the high-level feature maps.
- We have embedded feature recalibration through the application of Squeeze and Excitation operations within the attention-based Atrous Spatial Pyramid Pooling Mechanism.
- Our methodology's efficacy has been substantiated through rigorous experimentation on three publicly accessible benchmark datasets. Comparative analyses conclusively demonstrate that our approach consistently outperforms state-of-the-art medical segmentation models, particularly in terms of precision and the number of trainable parameters.

This paper is organized as follows: in Section 2, we review state-of-the-art techniques in the field of medical image segmentation. Section 3 gives an overview of the improved approach and describes the proposed model, detailing all its components. Section 4 delineates the datasets and the evaluation criteria adopted for this investigation. Section 5 offers an insight into the experimental outcomes. Lastly, Section 6 presents a discussion and offers concluding remarks.

2 RELATED WORKS

2.1 Convolutional Neural Networks (CNNs)

In the last decade, CNNs have established themselves as the preferred approach in many applications in the field of medical image analysis. The driving forces behind this transformation in deep learning include the accessibility of large datasets, the affordability of high-performance graphics processing units, advances in network architectures, improvements in learning algorithms and the refinement of regularization techniques. The integration of Fully Convolutional Networks (FCNs) has marked a significant advance in the conception of convolutional neural networks (CNNs) for handling semantic image segmentation tasks [12]. It replaces the fully connected layers found in classification networks with a fully convolutional layer. The application of FCNs to medical image segmentation, and in particular to polyp segmentation, has produced remarkable results [16].

FCNs exclusively employ convolutional layers, enabling them to generate segmentation maps of identical dimensions to the input image. By incorporating skip connections, which involve upsampling and fusing feature maps from the model's final layers with those from earlier layers, the model fuses semantic knowledge (extracted from deeper, heavier layers) with appearance details (from the shallower, finer layers) to achieve precise and intricate segmentations. However, this approach often yields lower-resolution output images. Overcoming the necessity for vast

amounts of labelled training data, which poses a significant computational challenge for real-time inference, has led to the emergence of another architectural approach grounded in encoder-decoder networks.

The SimpleCNN-Unet [17] model was proposed to solve the challenge of segmenting the optic disc with high accuracy from medical images for the diagnosis of pathological myopia. It reduces computational cost with no sacrifice of large receptive fields using small-core convolutions. A multi-layer cross-attention gate enhances efficient feature fusion, while large data augmentation handles the limited availability of data. Although the model improves accuracy and speed in segmentation, it still faces some challenges managing long-range dependencies relative to transformer-based methods.

Similarly, the researchers present UCM-Net [18], which is a light and effective network architecture for the segmentation of skin lesions within medical images to provide better diagnosis of melanoma. Considering the high demand for early detection, the deep learning models that impose high computation costs, hence applying these to mobile health will be restricted. UCM-Net managed to overcome this challenge by effectively fusing a multi-layer perceptron MLP and CNN through limited resources. UCM-Net requires much fewer parameters and computational resources. Therefore, it will work well even in resource-constrained environments.

2.2 Encoder-Decoder Architectures

Encoders-decoders are a family of models that learn to map data points from an input domain to an output domain via a two-stage network. The encoder, which executes an encoding function $z = g(x)$, compresses the input x into a latent space representation z, while the decoder $y = f(z)$ predicts the output y from z.

The U-Net model, presented in [13], improves medical image segmentation through a dual-path architecture that includes contraction for feature mapping and dimension reduction, and expansion for feature retrieval and spatial reconstruction. Using skip connections, it efficiently combines low- and high-level information, requiring experimentation to optimize model size for maximum efficiency.

The recently reported CSWin-Unet [19] is a variant of the classical U-Net architecture, which extends a U-shaped network with cross-shaped window self-attention. This approach solves two important challenges associated with medical image segmentation: the limited receptive field of CNNs and the high computational cost of Transformer architectures. It extracts local and global contextual information, while having higher computational efficiency by parallel processing of horizontal and vertical self-attention. The integration of the content-aware reassembly layer (CARAFE) helps the decoder avoid losing edge details during upsampling; hence, it would lead to more accurate segmentation of many modalities. CSWin-UNet has some shortcomings on low-contrast images and small organ segmentation. Due to reliance on pre-trained ImageNet weights, it cannot have full end-to-end applicability in purely medical domains.

Other works involve a UNet-based multi-branch feature fusion network, UMF-Net [20], for colon polyp segmentation. Notice that UMF-Net introduces DCN and a multi-branch feature fusion strategy for capturing both local and global features. It contains three branches: global, residual, and main, which are useful in enhancing feature extraction and refinement. Additionally, the proposed Residual Coarse-to-Fine block improves feature learning for the model. UMF-Net has demonstrated superior performance compared to state-of-the-art models on several tasks related to colon polyp segmentation by achieving higher accuracy with lower computational complexity. Its multi-branch structure may raise some challenges in real-time implementation and additional validation on more diverse datasets for wider clinical use.

Additionally, the Dilated-U-Net-Seg architecture [21] focuses on the integration of dilated convolution into an encoder-decoder structure for segmenting small, flat, and sessile polyps in colonoscopy images. Dilated-U-Net-Seg significantly enhanced the detection of hard-to-detect polyps, providing critical advantages for colorectal cancer screening. The model further suffers from increased computational complexity and relies on publicly available datasets, hence limiting real-time applicability in clinical settings and further validation with clinical data.

Another study proposes a new architecture DI-Unet [22], designed to enhance melanoma segmentation related to challenging color factors, boundary ambiguity, and scale variations within skin lesion images. The authors introduce a dual-branch structure including vertical and horizontal paths in order to improve feature diversity and richness with the help of the Dual Channel Symmetric Convolution Block (DCS-Conv) and Residual Fusion and Selection (RFS) module. The DCS-Conv has enabled the network to capture multi-scale features using convolution kernels of different sizes, while the RFS module uses an auto-attention mechanism that focuses on lesion-specific features and reduces irrelevant artifacts. Better segmentation accuracy achieved with DI-UNet, the authors indicate, may promote very early detection and diagnosis of skin cancer-especially melanoma-by dermatologists, hence potentially assisting clinical decision-making. However, despite its advantages, the complexity of the model may require further optimization to reduce computational costs for practical applications.

2.3 Residual Learning Mechanism

Although a wide range of convolution layers improves the extraction of image characteristics, it can cause problems such as exploding gradients and vanishing gradients. To solve these problems, the concept of residual functions was introduced [23].

ResUNet [24] is an innovative deep neural architecture that combines elements from both U-Net and ResNet [23] models. Unlike traditional convolutional layers, ResUNet employs fundamental building blocks. This architecture consists of three key components: an encoder, a decoder, and a bridge. Instead of employing pooling for feature downsampling, the encoder employs a stride of 2 in the initial convolution block. A transfer layer is positioned before each decoding unit to facilitate feature communication between the encoder and decoder.

An innovative approach named GR-Net in [25] presented a new neural network designed to improve the accuracy of medical image segmentation, in particular the segmentation of polyps in colonoscopic images for colorectal cancer screening. GR-Net employs a global-local learning strategy, using ResNest as a backbone to integrate features from different scales and introducing an axial attention module to deal with positional deviations. In addition, a new local branch is added to extract detailed features from the image, improving segmentation performance over current methods. These advances demonstrate the potential of GR-Net to make a significant contribution to the field of medical image analysis, in particular to improving early detection and diagnosis of colorectal cancer.

Similarly, SMRU-Net [26] proposes a new skin lesion image segmentation approach, which combines a Residual U-Net architecture with Channel-Space Separate Attention and Depthwise Separable Convolutions to handle complex lesion morphology and distracting backgrounds. Multi-scale feature extraction, facilitated by the ConvMixer Block, enables SMRU-Net to outperform state-of-the-art methods in handling blurred regions and complex boundaries, although at the cost of higher computational complexity that will limit its use in resource-constrained environments.

3 METHODOLOGY

3.1 Semantic Segmentation Problem Definition

We consider an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$ defined over a spatial domain $\Omega \subset \mathbb{R}^2$, where H and W are the image height and width, respectively. The total number of pixels in the image is $N = H \times W$. Furthermore, semantic segmentation aims at assigning a category label to every pixel in Ω. For binary segmentation, the goal is to divide Ω into two non-overlapping areas: a region of interest (ROI), denoted by $A \subset \Omega$, and a background region $\Omega \setminus A$.

Let $\hat{\mathbf{Y}} \in [0,1]^{H \times W}$ be the predicted score map, where each element $\hat{\mathbf{Y}}_{i,j}$ represents the probability that pixel $(i,j) \in \Omega$ belongs to the ROI. In the multi-class case, the prediction becomes $\hat{\mathbf{Y}} \in [0,1]^{H \times W \times C}$, where $\hat{\mathbf{Y}}_{i,j,c}$ denotes the probability that pixel (i,j) belongs to class C.

The segmentation model learns a mapping from the image domain to the class probability space:

$$\hat{\mathbf{Y}} = f_\theta(\mathbf{X}), \tag{1}$$

where f_θ denotes the deep neural network parameterized by θ.

3.2 Operational Primitives in Deep Learning

3.2.1 Sigmoid Activation Function

The sigmoid function is a smooth, differentiable activation function that was commonly used in early neural networks due to it being non-linear and having an output

that is bounded. It is utilized for mapping real-valued inputs to the range $(0, 1)$, which it renders extremely easy to probabilistic interpretation. The function gained popularity by virtue of being used with the backpropagation algorithm [27] and can be mathematically expressed as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

where x is a real-valued input.

3.2.2 Rectified Linear Unit (ReLU)

The Rectified Linear Unit (ReLU) is one of the widely used deep learning activation functions, valued for its simplicity and efficiency. ReLU introduces non-linearity with negligible computational cost and has been shown to improve sparse activations as well as aid in overcoming the vanishing gradient problem. ReLU was popularized in deep sparse neural networks [28] and is defined as:

$$\text{ReLU}(x) = \max(0, x), \tag{3}$$

where x is a scalar input.

3.2.3 Convolution Operation

Two-dimensional discrete convolution is among the basic blocks of convolutional neural networks (CNNs) to produce local spatial features from input. It achieves this by convolving a learnable kernel over the input feature map to produce an output feature map, detecting patterns such as edges or textures. Convolutional layers were originally proposed by [29] and later formalized in the document recognition community [30]. The operation with a kernel of size $k \times k$ is defined as:

$$Y_{i,j} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X_{i+m,j+n} \cdot K_{m,n}, \tag{4}$$

where X is the input feature map, K is the convolutional kernel, and Y is the resulting output feature map. The indices i and j refer to spatial positions in the output, and m, n to the positions within the kernel window.

3.2.4 Batch Normalization

Batch normalization is a technique introduced to solve internal covariate shift and enhance training stability and convergence of deep neural networks. By normalizing the input of each layer, it facilitates higher learning rates and reduces sensitivity to initialization. The method was proposed by [31] and operates by normalizing

features within a mini-batch as follows:

$$\begin{aligned} \mu &= \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2, \\ \hat{x}_i &= \frac{x_i-\mu}{\sqrt{\sigma^2+\epsilon}}, \qquad y_i = \gamma\hat{x}_i + \beta, \end{aligned} \tag{5}$$

where x_i is an input value from a mini-batch of size m, μ and σ^2 are the batch mean and variance, and ϵ is a small constant added for numerical stability. The parameters γ and β are learnable, allowing the network to recover the original distribution if necessary.

3.2.5 Pooling Operations

Pooling operations are downsampling techniques commonly used in convolutional neural networks to reduce the spatial dimensions of feature maps while preserving important features. The two most widely used types are max pooling, which selects the maximum value within a local receptive field, and average pooling, which computes the mean over the same region. These operations help control overfitting, reduce computational load, and provide translational invariance. Pooling was first introduced in early convolutional architectures for document recognition tasks by [30].

Max Pooling:

$$Y_{i,j} = \max_{(m,n)\in\mathcal{R}_{i,j}} X_{m,n}. \tag{6}$$

Average Pooling:

$$Y_{i,j} = \frac{1}{|\mathcal{R}_{i,j}|}\sum_{(m,n)\in\mathcal{R}_{i,j}} X_{m,n}. \tag{7}$$

Here, X is the input feature map, Y is the resulting output, and $\mathcal{R}_{i,j}$ denotes the local receptive region centered at position (i, j) in the input.

3.3 Overview of the Proposed Segmentation Network (ResCAM-Net)

The Residual Convolution and Atrous Multi-Scale Network (ResCAM-Net) is a further development of the lightweight ColonSegNet architecture [1], which was originally proposed for colon polyp segmentation tasks. Despite ColonSegNet achieving high performance with fewer trainable parameters, it suffers from limitations in its skip connections, causing loss of information. Besides that, it doesn't take full advantage of high-level feature maps from different receptive fields, which could bring substantial performance improvements. In order to overcome the shortcoming and make the architecture more robust for various medical imaging segmentation

tasks, not limited to colonoscopy segmentation only, several components have been integrated. Among the inspirations of DoubleU-NetPlus architecture [32], these enhancements include Multikernel Residual Convolution, Triple Attention Module, and Squeeze and Excitation-based Atrous Spatial Pyramid Pooling module, all strategically placed between encoder-decoder components.

The full network pipeline can be summarized by the following composite function:

$$\hat{\mathbf{Y}} = \sigma\left(\mathrm{Conv}_{1\times1}\left(f_{\mathrm{Decoder}}\left(f_{\mathrm{MKRC}}\left(f_{\mathrm{ASPP}}(f_{\mathrm{TA}}\left(f_{\mathrm{Encoder}}(\mathbf{X})\right)), \mathrm{Skips}\right)\right)\right)\right) \tag{8}$$

where σ denotes the sigmoid function (Equation (2)).

3.3.1 Encoder Path

As detailed in Figure 1, ResCAM-Net is composed of two parts: the first part is the encoder part, containing convolution and pooling layers to extract the specific characteristics of each image, as well as residual functions to avoid problems of overfitting, vanishing gradients, and model saturation. This part comprises four residual blocks, each composed of two convolution layers with filters of size 3×3, the first followed by a Rectified Linear Unit (ReLU) activation function and Batch Normalization (BN); the second is followed by a BN and also contains an identity path containing a convolution layer with size 1×1 filters followed by a BN and a Squeeze and Excitation (SE) block, as well as two strided convolutions enabling feature maps to be reduced in size without any remarkable loss of spatial information, each comprising a convolution layer with size 3×3 and strides 2 followed by a BN and ReLU activation function. This parameter-efficient design makes for an exceptionally lightweight network, guaranteeing real-time performance.

To more efficiently capture the encoder's high-level multi-scale contextual information and transmit it to the decoder in the bottleneck of the encoder-decoder model, we successively added the Triple Attention (TAM) Module, the Squeeze and Excitation-based Atrous Spatial Pyramid Pooling (SE-ASPP) Module and the Multi-Kernel Residual Convolution (MKRC) Module. All of these components are thoroughly detailed in the subsections below.

The encoder performs hierarchical feature extraction using convolutional layers, residual connections, SE blocks, and attention mechanisms. The input image is first passed through two convolutional layers with a residual shortcut:

$$\mathbf{F}_1 = \delta\left(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{X}))\right), \tag{9}$$

$$\mathbf{F}_2 = \mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{F}_1)), \tag{10}$$

$$\mathbf{F}_3 = \mathrm{SE}\left(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{F}_1))\right), \tag{11}$$

$$\mathbf{F}_{\mathrm{res}_1} = \delta\left(\mathbf{F}_2 + \mathbf{F}_3\right). \tag{12}$$

Here, $\delta(\cdot)$ denotes the ReLU activation function: $\delta(x) = \max(0, x)$.

The subsequent stages of the encoder apply downsampling and deeper residual blocks:

$$\mathbf{F}_{\text{ds}} = \delta\left(\text{BN}(\text{Conv}_{3\times3,\text{stride}=2}(\mathbf{F}_{\text{res}_1}))\right), \tag{13}$$

$$\mathbf{F}_{\text{res}_2} = f^2_{\text{ResBlock}}(\mathbf{F}_{\text{ds}}), \tag{14}$$

$$\mathbf{F}_{\text{pool}} = \text{MaxPool}_{2\times2}(\mathbf{F}_{\text{res}_2}), \tag{15}$$

$$\mathbf{F}_{\text{res}_3} = f^3_{\text{ResBlock}}(\mathbf{F}_{\text{pool}}), \tag{16}$$

$$\mathbf{F}_{\text{stride}} = f_{\text{StrideConv}}(\mathbf{F}_{\text{res}_3}), \tag{17}$$

$$\mathbf{F}_{\text{res}_4} = f^4_{\text{ResBlock}}(\mathbf{F}_{\text{stride}}). \tag{18}$$

Triple Attention is then applied:

$$\mathbf{F}_{\text{TA}} = \text{Concat}(f_{\text{CA}}(\mathbf{F}_{\text{res}_4}), f_{\text{SA}}(\mathbf{F}_{\text{res}_4}), f_{\text{SE}}(\mathbf{F}_{\text{res}_4})). \tag{19}$$

The attention-enhanced feature map is processed using the customized ASPP module:

$$\mathbf{F}_{\text{ASPP}} = f_{\text{ASPP}}(\mathbf{F}_{\text{TA}}). \tag{20}$$

Then, a Multi-Kernel Residual Convolution (MKRC) module refines the features:

$$\mathbf{F}_{\text{MKRC}} = f_{\text{MKRC}}(\mathbf{F}_{\text{ASPP}}). \tag{21}$$

3.3.2 Decoder Part

The second part is the decoder part, which resamples low-resolution feature maps to the size of the original image, recovering the spatial information and fine detail lost during downsampling. It refines and transforms these features into high-resolution per-pixel classification maps, facilitating precise segmentation and localization of objects. It comprises four block residuals and four transposed convolutions.

The decoder stage is mathematically formulated as follows:

$$D_1 = \text{ResBlock}_1\left(\text{Concat}\left(\text{Up}_{\times4}(\text{MKRC}(F_{\text{enc}})), \text{Skip}_3\right)\right), \tag{22}$$

$$D_2 = \text{ResBlock}_2\left(\text{Concat}\left(D_1, \text{Skip}_2\right)\right), \tag{23}$$

$$D_3 = \text{ResBlock}_3\left(\text{Concat}\left(\text{Up}_{\times2}(D_2), \text{Skip}_1\right)\right), \tag{24}$$

$$F_{\text{dec}} = \text{ResBlock}_4(D_3). \tag{25}$$

The final segmentation map is obtained by applying a 1×1 convolution followed by a sigmoid activation:

$$\hat{\mathbf{Y}} = \sigma\left(\text{Conv}_{1\times1}(F_{\text{dec}})\right). \tag{26}$$

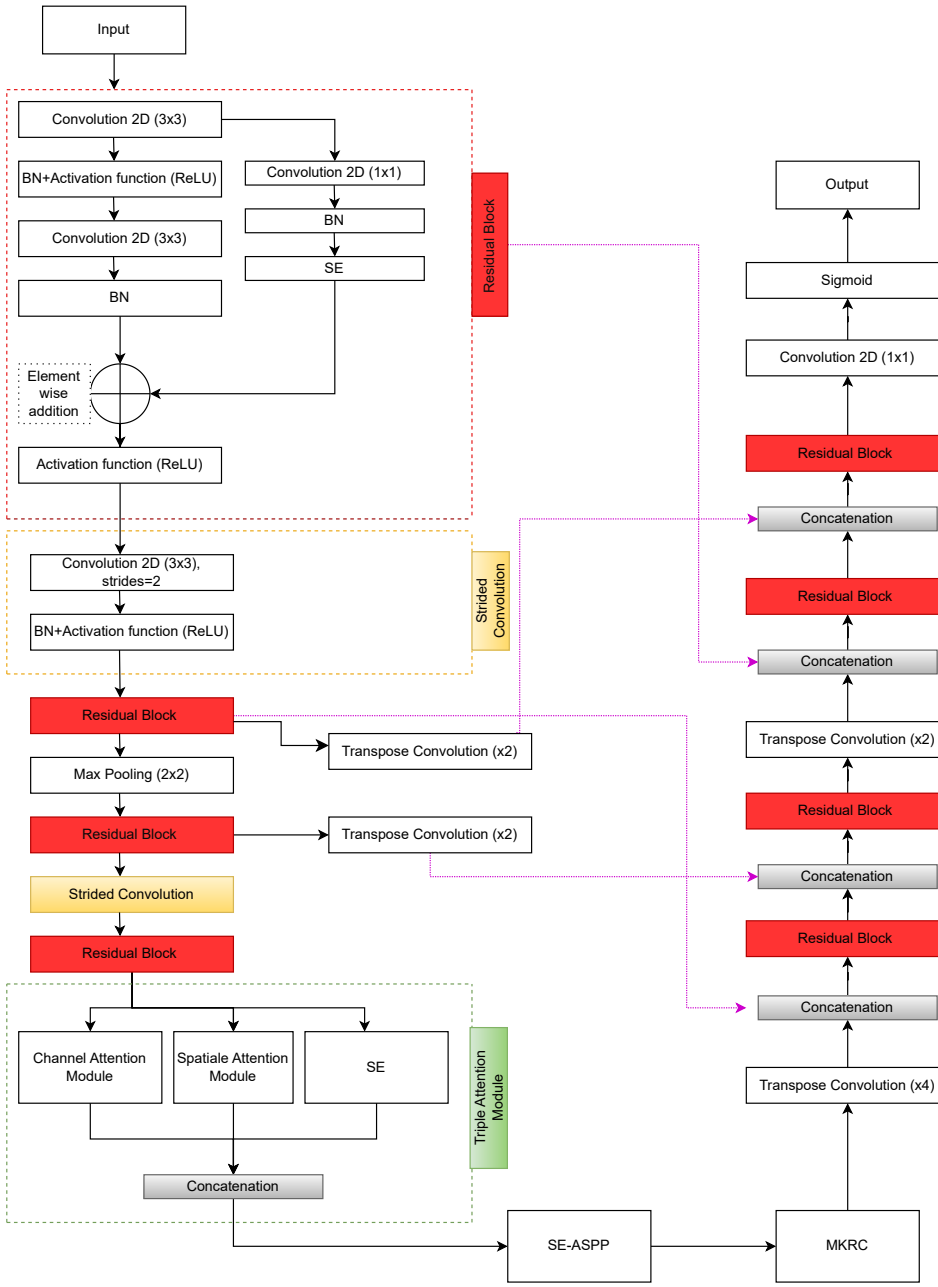


Figure 1. Proposed architecture (ResCAM-Net) components

3.3.3 Atrous Spatial Pyramid Pooling (ASPP)

Every traditional CNN model uses the pooling operation (maximum or average pooling) to reduce the spatial dimensions of the feature maps in order to reduce the number of parameters in the architecture and the computational cost of the subsequent layers; the model can focus on the most relevant information in the feature maps and discard redundant or irrelevant information. But this operation may produce the problem of information loss, especially when the object of interest has a small size. The ASPP [33] surmounts this limitation by using dilated convolutions at different dilation rates to capture multi-scale contextual information and thus considering the object of interest and its surrounding context at multiple scales without losing spatial resolution.

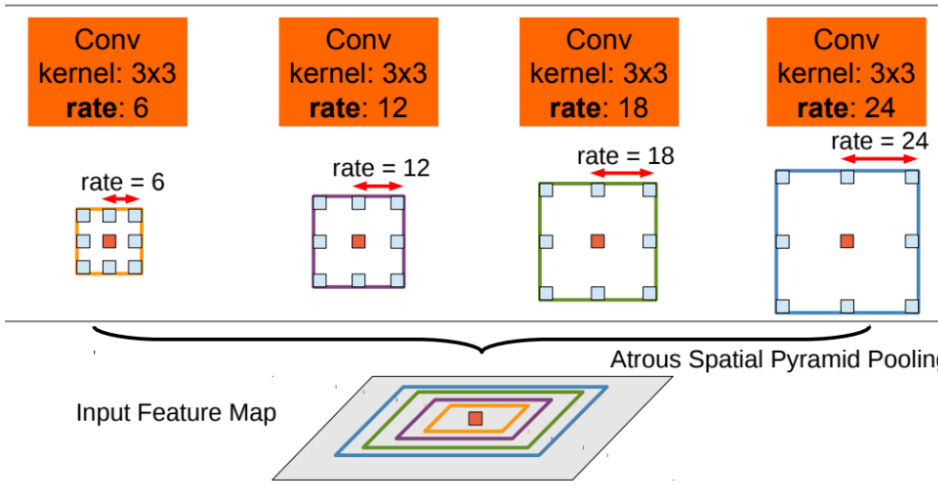


Figure 2. Atrous Spatial Pyramid Pooling (ASPP): For center pixel classification (depicted in orange), ASPP leverages multi-scale features through the use of several concurrent filters with varying rates. Different colors illustrate the effective field-of-views [33].

Atrous Convolution and ASPP

The ASPP based on the Atrous convolution, also referred to as dilated convolution, is a technique designed to enlarge the receptive field of filters without increasing the number of parameters or sacrificing spatial resolution. It introduces a dilation rate r that defines the spacing between kernel elements. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ and a filter $\mathbf{w} \in \mathbb{R}^{K\times K\times C}$, the atrous convolution at location p is computed as:

$$y(p) = \sum_{k\in\mathcal{K}} w(k) \cdot x(p + r \cdot k). \tag{27}$$

Here, $\mathcal{K}$ denotes the kernel grid, and r controls the stride at which input positions are sampled. A larger rate results in a wider receptive field, allowing the network to incorporate broader context information without downsampling the feature map.

As demonstrated in Figure 2, the ASPP module comprises multiple parallel blocks followed by a pooling layer; each block includes a succession of atrous convolutional layer with a specific sampling rate, batch normalization and activation function. The pooled feature map is passed through a 1×1 convolutional layer and then upsampled to match the spatial resolution of the original input image and the output of this operation is concatenated with the output feature maps of each block and then passed through a 1×1 convolutional layer to reduce the number of channels. Given a set of rates $\{r_1, r_2, \ldots, r_n\}$, ASPP computes:

$$y_i(p) = \sum_{k \in \mathcal{K}} w_i(k) \cdot x(p + r_i \cdot k), \quad \text{for each } r_i \tag{28}$$

The resulting feature maps $y_1, y_2, \ldots, y_n$ are then concatenated:

$$y_{\text{ASPP}} = \text{Concat}(y_1, y_2, \ldots, y_n). \tag{29}$$

Customized Atrous Spatial Pyramid Pooling

To further improve the contextual modeling capability of ASPP, we propose a customized version presented in Figure 3 that incorporates seven parallel branches, including one global pooling path and six atrous convolution branches with varying dilation rates: 1, 2, 6, 10, 13, and 16. The global pooling path applies average pooling followed by a 1×1 convolution, batch normalization, and upsampling to match the spatial resolution of other branches. Each of the atrous convolution branches consists of a 3×3 convolution with a specific dilation rate, followed by batch normalization and ReLU activation. The outputs of all branches are concatenated and passed through a 1×1 projection layer with batch normalization and ReLU to fuse the features.

Let $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H \times W \times C}$ be the input feature map. The global context path output is defined as:

$$\mathbf{F}_{\text{global}} = \text{Upsample}\left(\text{BN}\left(\text{ReLU}\left(\text{Conv}_{1\times1}\left(\text{Pool}_{\text{avg}}(\mathbf{F}_{\text{in}})\right)\right)\right)\right). \tag{30}$$

For each atrous convolution branch with dilation rate $r_i \in \{1, 2, 6, 10, 13, 16\}$, the feature map is computed as:

$$\mathbf{F}_{r_i} = \text{BN}\left(\text{ReLU}\left(\text{Conv}_{3\times3}^{r_i}(\mathbf{F}_{\text{in}})\right)\right). \tag{31}$$

The concatenated feature map is then:

$$\mathbf{F}_{\text{concat}} = \text{Concat}(\mathbf{F}_{\text{global}}, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_6, \mathbf{F}_{10}, \mathbf{F}_{13}, \mathbf{F}_{16}). \tag{32}$$

Finally, the features are fused using a 1×1 projection:

$$\mathbf{F}_{\text{out}} = \text{ReLU}\left(\text{BN}\left(\text{Conv}_{1\times1}(\mathbf{F}_{\text{concat}})\right)\right). \tag{33}$$

This design offers several advantages. By incorporating a wide range of dilation rates, the module is able to capture rich, multi-scale information, enhancing the model's ability to recognize objects of varying sizes. The inclusion of a global pooling branch introduces global contextual awareness, which complements the localized features extracted by the atrous convolutions. The use of batch normalization and non-linear activations further improves the network's convergence stability and generalization. Finally, the projection layer helps reduce computational complexity while effectively merging the aggregated features into a unified representation, making the module both powerful and efficient.

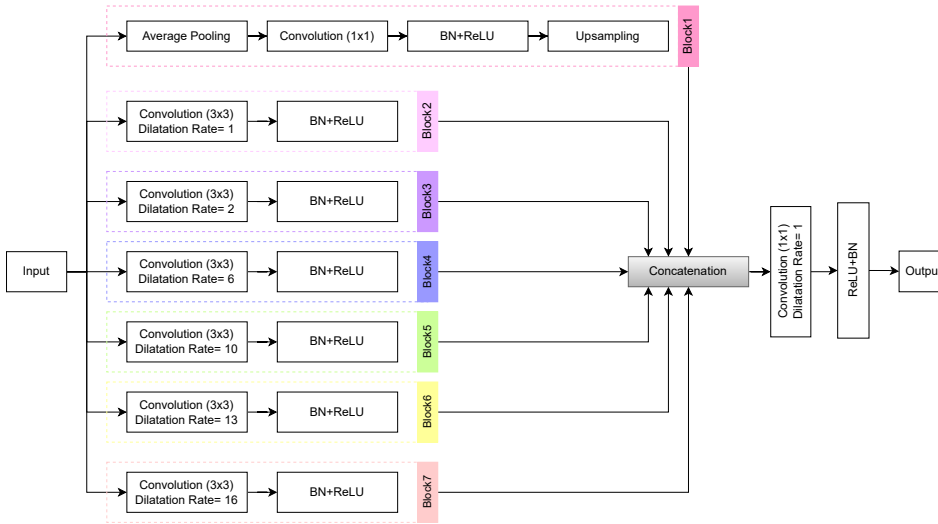


Figure 3. Utilized Atrous Spatial Pyramid Pooling

3.3.4 Squeeze and Excitation (SE)

Squeeze and Excitation is a channel-wise attention mechanism used in CNNs to improve their performance by selectively emphasizing informative feature's channels and suppressing irrelevant ones. Traditional CNN used convolution operator to extract spatial and channel-wise information from the images at each layer and the network weights each channel in the resulting feature maps equally, which consists of a lot of parameters and increases the computational complexity. SE overcomes this problem by decomposing each feature channel into a single numeric value through two main components of SE operations, the first used to extract the global information from each feature map channel and decompose each feature (4D tensor: Batch

size × Height × Width × Number of channels) channel into a single numeric value to decrease the number of parameters and thus reduce the computational complexity. This operation is performed with the global average pooling (GAP) that instead of using a window to compute the average of each channel and producing a feature map with the same number of channels with reduced spatial dimensions like traditional average pooling, the GAP computes the average of all the values in each channel of the input feature map and provides a single value for each channel.

The second component used a fully connected Multi-Layer Perceptron (MLP) with a bottleneck structure to scale the activations of each channel of the feature map adaptively by generation a set of weights (If the value is close to 1 means the channel is important). This step takes a 1D vector obtained from the squeeze step.

The three main components of SE represented in Figure 4 are detailed as follow:

Squeeze: To aggregate global spatial information into a compact channel descriptor, we apply Global Average Pooling (GAP). For each channel c, the squeezed scalar is computed as:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c(i,j), \quad \text{for } c = 1, 2, \ldots, C, \tag{34}$$

where $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ represent the output of a convolutional layer, H, W, and C are the spatial dimensions and the number of channels, respectively. This results in a descriptor vector $\mathbf{z} \in \mathbb{R}^C$ that captures global context per channel.

Excitation: The vector $\mathbf{z}$ is passed through two fully connected (FC) layers with a bottleneck structure and non-linear activations to learn channel-wise dependencies:

$$\mathbf{s} = \sigma\left(\mathbf{W}_2 \cdot \delta\left(\mathbf{W}_1 \cdot \mathbf{z}\right)\right), \tag{35}$$

where $\delta(\cdot)$ is the ReLU function, $\sigma(\cdot)$ is the sigmoid activation, and $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are the weight matrices of the FC layers. The reduction ratio r controls the dimensionality of the bottleneck.

Scaling: The resulting gating vector $\mathbf{s} \in [0,1]^C$ is used to recalibrate the original feature map by applying channel-wise multiplication:

$$\mathbf{X}'_c = s_c \cdot \mathbf{X}_c, \quad \forall c \in \{1, 2, \ldots, C\}. \tag{36}$$

From a theoretical standpoint, the SE block performs a low-rank, nonlinear transformation that captures inter-channel dependencies and enables dynamic feature selection. By conditioning the response of each channel on the global distribution of features, it enhances the network's capacity to model informative and discriminative representations, while maintaining computational efficiency.

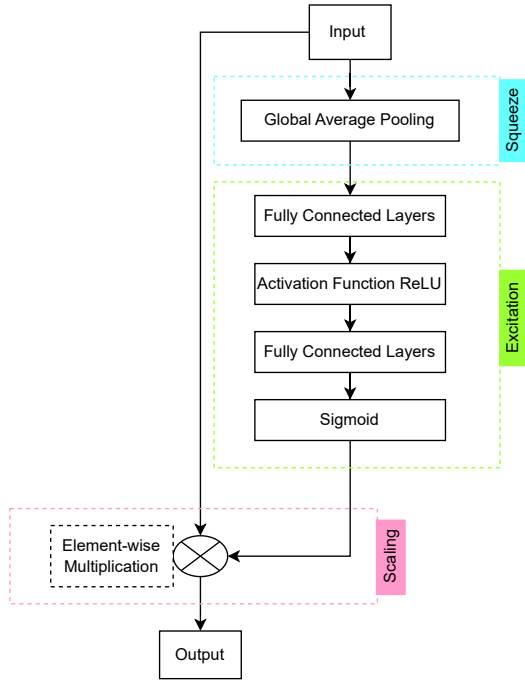


Figure 4. Squeeze and Excitation module

3.3.5 Multi-Kernel Residual Convolution Module (MKRC)

The introduction of the MKRC module, as developed by [32], marks a significant advancement in network architecture. The MKRC module illustrated in Figure 5 is strategically integrated into the network's bottlenecks to mitigate issues related to gradient saturation and degradation during the learning process. This innovative MKRC module enriches the field of view representation of diverse features, enhancing the model's learning efficiency and robustness. It achieves this by incorporating multiple parallel convolution layers with varying kernel sizes. The expansion of kernel sizes in these convolution layers empowers the network to extract more resilient feature representations across a range of multi-scale receptive fields, thereby modulating feature learning uniquely for each block. Following each convolution layer, there is a sequential application of a batch normalization layer and a ReLU activation function. Subsequently, all four feature maps are concatenated, providing comprehensive information spanning every relevant receptive field.

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be the input feature map. The module first applies four parallel convolution operations with kernels of size 1×1, 3×3, 5×5, and 7×7,

each followed by batch normalization and a ReLU activation:

$$\mathbf{X}_{1\times1} = \delta(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{X}))), \tag{37}$$

$$\mathbf{X}_{3\times3} = \delta(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{X}))), \tag{38}$$

$$\mathbf{X}_{5\times5} = \delta(\mathrm{BN}(\mathrm{Conv}_{5\times5}(\mathbf{X}))), \tag{39}$$

$$\mathbf{X}_{7\times7} = \delta(\mathrm{BN}(\mathrm{Conv}_{7\times7}(\mathbf{X}))). \tag{40}$$

The outputs from all branches are concatenated:

$$\mathbf{X}_{\mathrm{multi}} = \mathrm{Concat}(\mathbf{X}_{1\times1}, \mathbf{X}_{3\times3}, \mathbf{X}_{5\times5}, \mathbf{X}_{7\times7}). \tag{41}$$

A 1×1 convolution is then used to fuse the concatenated features and reduce the dimensionality:

$$\mathbf{X}_{\mathrm{fused}} = \delta(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{X}_{\mathrm{multi}}))). \tag{42}$$

Simultaneously, a residual connection is applied by projecting the original input through a 1×1 convolution:

$$\mathbf{X}_{\mathrm{res}} = \mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{X})). \tag{43}$$

Finally, the fused features and the residual projection are concatenated and passed through a ReLU activation to produce the output:

$$\mathbf{Y} = \delta(\mathrm{Concat}(\mathbf{X}_{\mathrm{fused}}, \mathbf{X}_{\mathrm{res}})). \tag{44}$$

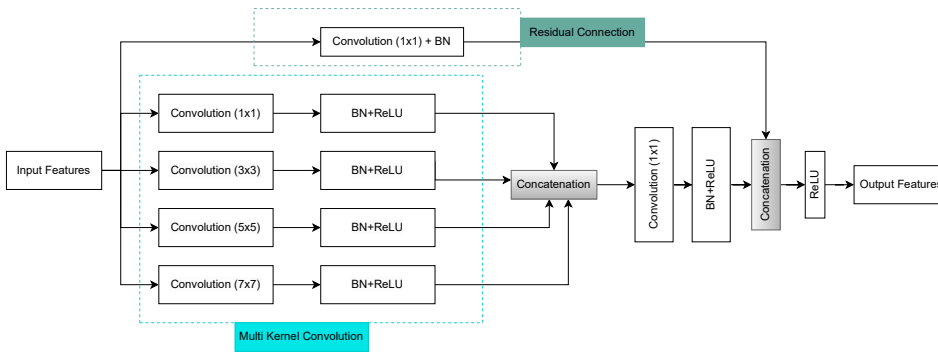


Figure 5. Multi-Kernel Residual Convolution (MKRC) module

3.3.6 Channel and Spatial Attention Mechanism

The incorporation of the Spatial Attention Module (SAM) into convolutional neural networks, as an integral part of the attention mechanism, has yielded favorable results in tasks related to detection and classification [34]. SAM directs its attention

towards the positional details within images, effectively capturing the spatial interplay among input features. The SAM focuses on the spatial dimensions (height and width) of the feature tensor. It learns to assign different weights to each spatial location, based on how much it contributes to the final output. This allows the model to attend to the most relevant spatial locations and suppress the less relevant ones, enabling the model to capture fine-grained features, thereby enhancing its precision in localization.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, spatial attention is computed through a sequence of operations. First, average pooling and max pooling are applied along the channel dimension to aggregate spatial descriptors:

$$\mathbf{M}_{\text{avg}}(i,j) = \frac{1}{C}\sum_{k=1}^{C}\mathbf{X}(i,j,k), \tag{45}$$

$$\mathbf{M}_{\max}(i,j) = \max_{k\in\{1,\ldots,C\}}\mathbf{X}(i,j,k). \tag{46}$$

The resulting descriptors are concatenated to form a two-channel feature map:

$$\mathbf{M}_{\text{spatial}} = \text{Concat}(\mathbf{M}_{\text{avg}}, \mathbf{M}_{\max}) \in \mathbb{R}^{H\times W\times 2}. \tag{47}$$

Next, a convolution operation with a 7×7 kernel is applied, followed by a sigmoid activation to produce the spatial attention map:

$$\mathbf{A}_{\text{spatial}} = \sigma\left(\text{Conv}_{7\times 7}(\mathbf{M}_{\text{spatial}})\right) \in [0,1]^{H\times W\times 1}. \tag{48}$$

This attention map is then broadcasted and multiplied element-wise with the original feature map to enhance or suppress responses at each spatial location:

$$\mathbf{X}'(i,j,k) = \mathbf{A}_{\text{spatial}}(i,j)\cdot\mathbf{X}(i,j,k), \quad \forall k\in\{1,2,\ldots,C\}. \tag{49}$$

The channel attention mechanism, first introduced in [35], is a critical component in deep neural networks. This mechanism enables networks to adaptively emphasize informative channels within feature maps while attenuating less relevant ones. By capturing channel-wise dependencies, the channel attention mechanism significantly enhances the representational power of CNNs. Its integration into diverse computer vision applications, such as image segmentation, classification and object detection, has become pervasive, leading to substantial improvements in model performance and accuracy.

Given an input feature map $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, channel attention is computed through a sequence of operation. We first apply global average pooling and global

max pooling along the spatial dimensions to produce two distinct descriptors:

$$z_c^{\text{avg}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c(i,j), \tag{50}$$

$$z_c^{\max} = \max_{i,j} \mathbf{X}_c(i,j) \tag{51}$$

These two vectors $\mathbf{z}^{\text{avg}}, \mathbf{z}^{\max} \in \mathbb{R}^C$ are then passed through a shared multi-layer perceptron (MLP) composed of two fully connected layers with ReLU activation:

$$\mathbf{s}^{\text{avg}} = \mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \mathbf{z}^{\text{avg}}), \tag{52}$$

$$\mathbf{s}^{\max} = \mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \mathbf{z}^{\max}), \tag{53}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, and r is the reduction ratio. The final channel attention vector is computed by element-wise addition followed by a sigmoid activation:

$$\mathbf{s} = \sigma(\mathbf{s}^{\text{avg}} + \mathbf{s}^{\max}). \tag{54}$$

This attention vector $\mathbf{s} \in [0,1]^C$ is used to recalibrate the input features via channel-wise multiplication:

$$\mathbf{X}'_c = s_c \cdot \mathbf{X}_c, \quad \forall c \in \{1, 2, \ldots, C\}. \tag{55}$$

The two spatial and channel attention mechanisms are illustrated in Figures 6 and 7, respectively.

3.3.7 Triple Attention Module (TAM)

The hybrid Triple Attention Module (TAM) described in [32] combines elements from CBAM [34] and Focus U-Net to refine features. As illustrated in Figure 8, TAM makes use of parallel processing through both squeeze and excitation networks, adapted spatial attention mechanisms, as well as channel-based attention. TAM enhances channel inter-dependencies, selecting relevant information and reducing noise, with a focus on improving channel-based attention. DoubleU-NetPlus extends CBAM's ideas by using global average and max pooling operations along the channel axis, concatenated and activated with sigmoid, for effective feature descriptors. Spatial attention highlights target regions using two channel contexts and spatial recalibration. Attempts to include global pooling in spatial attention didn't improve performance, so the original CBAM implementation was retained.

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ be the input feature map. The input is simultaneously processed by three attention branches:

Channel Attention (CA) computes a channel-wise attention vector $\mathbf{s}_{\text{CA}} \in [0,1]^C$ and produces:

$$\mathbf{X}_{\text{CA}} = \mathbf{s}_{\text{CA}} \odot \mathbf{X}. \tag{56}$$

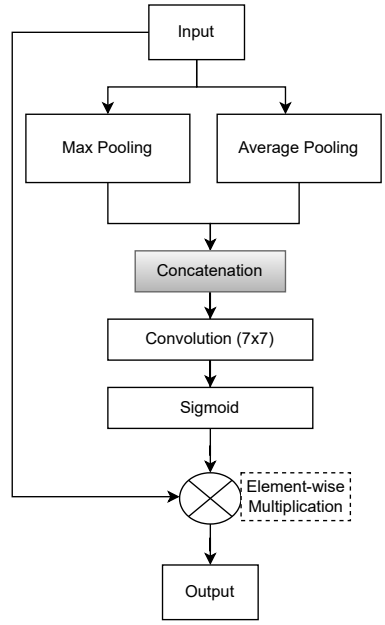


Figure 6. Spatial Attention Mechanism components

Spatial Attention (SA) generates a spatial attention map $\mathbf{a}_{\text{SA}} \in [0,1]^{H\times W}$ and computes:

$$\mathbf{X}_{\text{SA}}(i,j,k) = \mathbf{a}_{\text{SA}}(i,j) \cdot \mathbf{X}(i,j,k). \tag{57}$$

Squeeze-and-Excitation (SE) produces another channel-wise attention vector $\mathbf{s}_{\text{SE}} \in [0,1]^{C}$ and outputs:

$$\mathbf{X}_{\text{SE}} = \mathbf{s}_{\text{SE}} \odot \mathbf{X}. \tag{58}$$

The outputs of these three branches are then concatenated to form a unified attention-enhanced representation:

$$\mathbf{X}_{\text{out}} = \text{Concat}(\mathbf{X}_{\text{CA}}, \mathbf{X}_{\text{SA}}, \mathbf{X}_{\text{SE}}). \tag{59}$$

4 MATERIALS AND EVALUATION METHODS

4.1 Datasets and Pre-Processing

In this section, we provide a concise overview of the datasets used in this study, along with details of their preprocessing. To assess the effectiveness of our proposed model, we employed three distinct datasets with different modalities: Kvasir-SEG [36] and the ISIC-2017 Lesion Boundary Segmentation dataset [37, 38]. Fig-

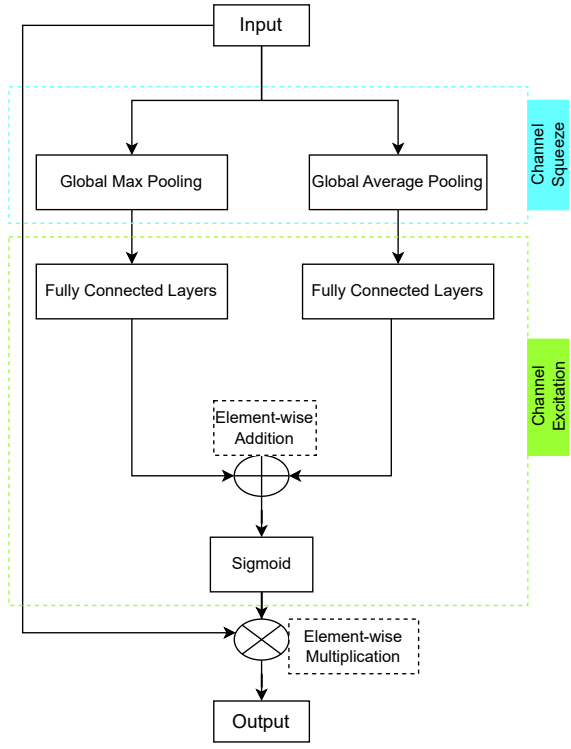


Figure 7. Channel Attention Mechanism components

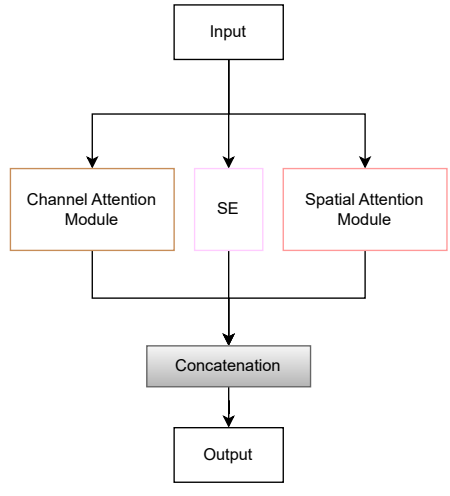
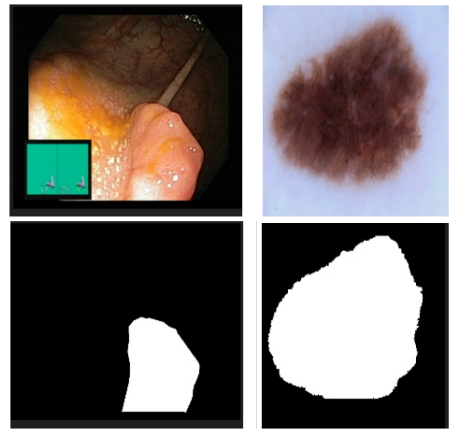


Figure 8. Triple Attention Module

ure 9 illustrates a representative image and its corresponding mask from each of these datasets.



a) Kvasir-SEG b) ISIC-2017

Figure 9. Images and corresponding segmentation masks in the dataset

4.1.1 Datasets

ISIC-2017 Lesion Boundary Segmentation dataset

The ISIC-2017 dataset [37], presented by the International Skin Imaging Collaboration, is considered to be a benchmark for dermatology and computer vision. It provides diverse skin lesion images along with their detailed boundary segmentation masks, with 2 000 training, 150 validation, and 600 test images.

We used the ISIC-2017 dataset in two different ways:

First Approach: To show how well our model works with small datasets, we trained it using just 100 images. With data augmentation, this grew to 2 600 images.

Second Approach: We went bigger here, training on 2 594 images. After applying augmentation, the dataset expanded to 67 444 images.

Details of the data sets used can be found in Table 1.

Kvasir-SEG dataset

The Kvasir-SEG dataset [36] is a specialized subset of the Kvasir dataset. It comprises 1 000 annotated endoscopy images highlighting key regions of interest, such as polyps and ulcers. It was created specifically for medical image segmentation in the field of gastrointestinal endoscopy.

4.1.2 Pre-Processing

Researchers rely on large image databases to overcome the problems of overfitting and class imbalance in recent deep learning trends, but most existing datasets have only a few samples, making it difficult to learn DL models on these datasets, therefore data augmentation methods are seen as a useful solution to these problems [39].

To prepare the acquired dataset, several essential pre-processing steps are performed. Since the dataset comprises images from various sources with varying resolutions, it is imperative to resize them to a uniform input size in order to improve convolution speed and save computing resources. In addition, all images and their corresponding masks are subject to this resizing procedure. In addition, the pixel values in the images are normalized to fall within the range of 0 to 255.

To guarantee effective and robust model learning on three datasets, we used a total of twenty-five data augmentation techniques, including Center Crop, Random Rotate 90 degree, Transpose, Elastic Transform, Grid Distortion, Optical Distortion, Vertical Flip, Horizontal Flip, Grayscale, Grayscale Vertical Flip, Grayscale Horizontal Flip, Grayscale Center Crop, Random Brightness and Contrast adjustments, Random Gamma correction, Hue Saturation and Value adjustments, Random Brightness and Contrast adjustment, Motion Blur, Median Blur and Gaussian Blur with a specified blur limit, Gaussian Noise, Channel Shuffling, Coarse Dropout with a maximum number of holes, height, and width specifications and finally RGB Shift to increase image variability during the learning process.

Dataset	Images Shape	Modality	Training/Validation/Test
Kvasir	256×256	Colonoscopy	20 800/100/100
ISIC-2017	192×256	Dermoscopy	67 444/100/14
ISIC-2017 (small)	192×256	Dermoscopy	2 600/100/14

Table 1. Summary of experiment datasets

4.2 Experiment Setting and Implementation Details

In our study, we deployed six different network models, namely U-Net, Attention UNet, Attention Residual UNet, FF-UNet, ColonSegNet and ResCAM-Net (our contribution). These models were trained and tested on three separate datasets in Nvidia $2 \times$ RTX4090. For the sake of consistency, we have kept identical training configurations and methodologies for all models. The specifics of our training parameters for all datasets are summarized in Table 2. The choice of the loss function is a crucial factor in enhancing model performance [40].

In our experiments, we employed the Dice Loss, as illustrated in Equation (60).

$$\text{Dice Loss} = 1 - \frac{2\sum(P \cdot T) + \epsilon}{\sum P^2 + \sum T^2 + \epsilon}, \tag{60}$$

where P represents the predicted mask, T represents the true mask and ϵ is a small constant added to avoid division by zero.

Method	Epochs	Batch Size	Learning Rate	Loss	Optimizer
All Methods	20	16	1e−4	Dice Loss	Adam

Table 2. Hyperparameters used for baseline segmentation methods

4.3 Evaluation Metrics

Accuracy is a commonly used and effective evaluation metric for assessing the performance of classification models, but its applicability differs when it comes to segmentation tasks. In our specific case, we encounter a class imbalance, where one class represents the background, and the other class represents the object of interest. The challenge arises because the background class dominates the images, particularly when small objects occupy only a tiny portion of the overall image. In such scenarios, the model can yield a high accuracy percentage, indicating a large number of correctly classified pixels, even if it results in a black image. However, this can be misleading because it might disregard critical details represented by a small object [41].

To address this issue in our study, we have chosen not to rely on accuracy as the primary performance metric. Instead, we utilize the Dice Coefficient (DSC) as defined in Equation (61) and the Intersection over Union (IoU) as defined in Equation (62) to provide a more accurate and informative assessment of our model's performance.

$$\mathrm{DSC} = \frac{2 \cdot |P \cap T|}{|P| + |T|}, \tag{61}$$

where P is the set of pixels in the predicted mask, T is the set of pixels in the true mask, $|P|$ represents the cardinality (number of elements) of P, $|T|$ represents the cardinality of T, and $|P \cap T|$ represents the cardinality of the intersection of P and T.

$$\mathrm{IoU} = \frac{|P \cap T|}{|P \cup T|}. \tag{62}$$

where $|P \cap T|$ represents the cardinality of the intersection of P and T, and $|P \cup T|$ represents the cardinality of the union of P and T.

5 EXPERIMENTAL RESULTS

In our in-depth evaluation of segmentation models, we provide both quantitative and qualitative results acquired from two distinct medical image dataset modalities, conducting a comparison with other state-of-the-art methods (SOTA). To ensure that the proposed model outperforms or is at the same level as other SOTA methods with a fair comparison, we applied the same configuration during training on all

models (number of epochs, batch size, learning rate, ...) as well as the same number of test, training and validation data.

Model	Loss	IoU	DSC	N. Parameters
FF-UNet [42]	0.1346	0.7627	0.8653	3 942 930
U-Net [13]	0.3745	0.4551	0.6255	31 402 570
Attention U-Net [43]	0.3475	0.4843	0.6525	37 334 734
AResU-Net [44]	0.2093	0.6538	0.7907	39 090 446
ColonSegNet [1]	0.2471	0.6036	0.7528	4 929 026
ResCAM-Net (ours)	**0.1015**	**0.8156**	**0.8984**	**9 916 453**

Table 3. Performance evaluation on the ISIC-2017 small dataset

In this comparative analysis of various segmentation models on the **ISIC-2017 small dataset**, the performance of each model was evaluated on the basis of IoU and DSC, a measure of segmentation accuracy. Among the models evaluated as demonstred in Table 3, our proposed model ResCAM-Net emerged as the best-performing model with an impressive DSC score of 0.8984, demonstrating excellent segmentation accuracy. Attention residual U-Net (AResU-Net) followed closely behind with a DSC score of 0.7907, demonstrating robust segmentation results. FF-UNet also performed remarkably well, with a DSC score of 0.8653, indicating high segmentation accuracy. ColonSegNet delivered satisfactory results with a DSC score of 0.7528. Attention U-Net obtained an acceptable DSC score of 0.6525, while U-Net, with a DSC score of 0.6255, achieved the lowest segmentation accuracy in this comparison.

Figure 10 illustrates the segmentation results of various models applied to the ISIC-2017 small dataset. The first column of the figure displays the original image to be segmented, the second column presents the corresponding ground truth mask, and the last column exhibits the predicted mask produced by the applied models. Notably, the results depicted in the figure showcase the superior performance of our model, FF-UNet, and AResU-Net compared to the other evaluation models. The effectiveness of these models in achieving better segmentation outcomes can be attributed to their advanced capabilities in capturing the crucial textual information needed for precise delineation of borders and boundaries. It is essential to note that the limitations observed in the segmentation results are primarily attributed to the inherent challenges posed by a relatively small dataset. This dataset lacks the diversity and volume of data necessary for a comprehensive model training, impacting the performance of the tested models.

For the **ISIC-2017 dataset**, the multidimensional analysis based on DSC and Number of Parameters (N Parameters) provides an overview of model performance and computational efficiency. AResU-Net and U-Net excelled with the highest DSC score of 0.8932 and 0.8919, indicating outstanding segmentation accuracy, albeit with a higher number of parameters. FF-UNet achieved excellent segmentation performance with a remarkable DSC score of 0.8407 and an effective parameter count of almost 4 Million (M). Attention U-Net struck a balance between 0.8870

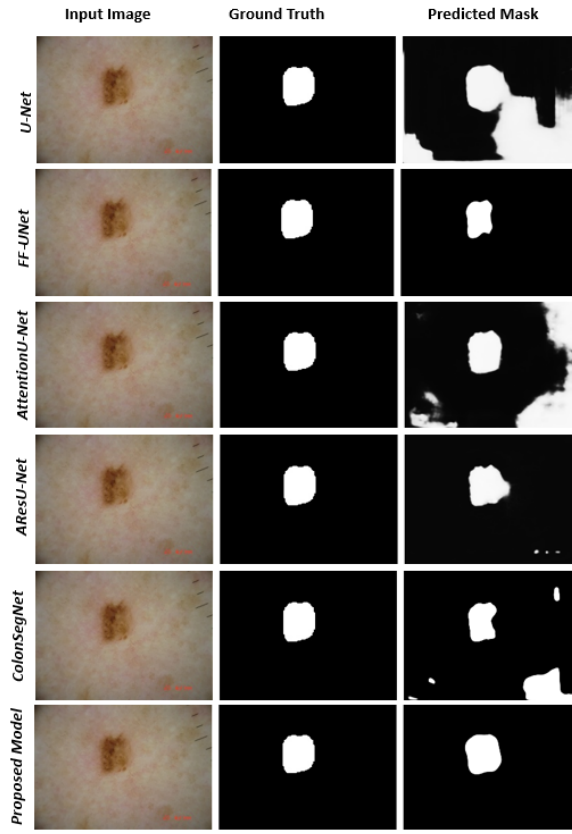


Figure 10. Qualitative segmentation results using six models on ISIC-2017 small dataset

accuracy and computational load. Our model achieved a remarkable DSC score of 0.8739 with a moderate parameter count of almost 10 M. Finally, ColonSegNet demonstrated a satisfactory segmentation accuracy of 0.7977 while maintaining an efficient use of computational resources of almost 5M parameters. Table 4 shows the results achieved.

Model	Loss	IoU	DSC	N. Parameters
FF-UNet [42]	0.1592	0.7252	0.8407	3 942 930
U-Net [13]	0.1080	0.8050	0.8919	31 402 570
Attention U-Net [43]	0.1129	0.7969	0.8870	37 334 734
AResU-Net [44]	0.1067	0.8070	0.8932	39 090 446
ColonSegNet [1]	0.2022	0.6635	0.7977	4 929 026
ResCAM-Net (ours)	**0.1260**	**0.7761**	**0.8739**	**9 916 453**

Table 4. Performance evaluation on the ISIC-2017 dataset

The ISIC-2017 dataset revealed both qualitative and quantitative outcomes, emphasizing that U-Net and its variants demand an extensive volume of training data for favorable results. Their convergence is time-consuming due to the huge number of trainable parameters involved. In contrast, our model exhibited promising performance in terms of convergence time, resource efficiency, and the quality of the predicted masks.

Figure 11 displays the results of each model on the ISIC-2017 dataset. In comparative evaluations, U-Net and its variants consistently deliver outstanding results, establishing them as top-performing models. Concurrently, the proposed model exhibits remarkable performance, achieving impressive outcomes in its own right.

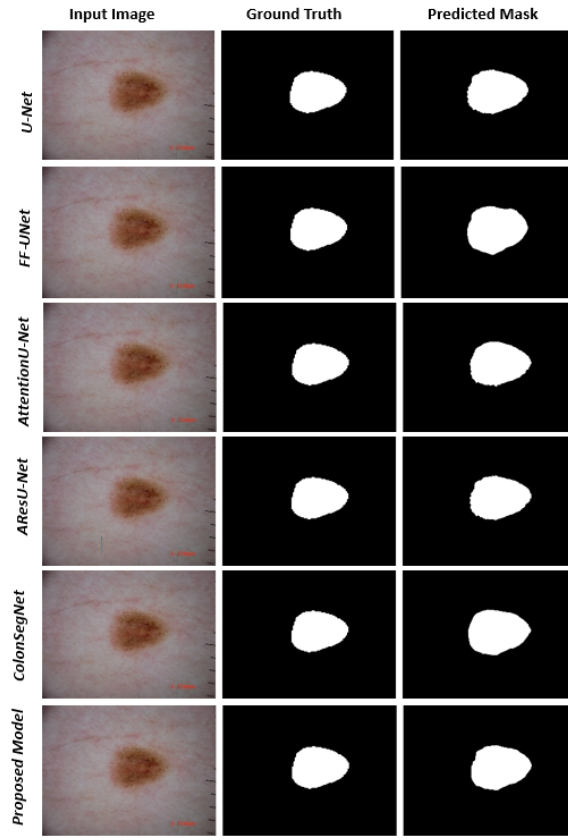


Figure 11. Qualitative segmentation results using six models on ISIC-2017 dataset

For **Kvasir-SEG dataset**, our model ResCAM-Net emerged as the best-performing model, achieving an outstanding DSC score of 0.8500, signifying exceptional segmentation accuracy. Attention Residual U-Net (AResU-Net) followed closely behind with an equally impressive DSC score of 0.8495, demonstrating robust segmen-

tation performance. FF-UNet also achieved outstanding results with a DSC score of 0.8350, indicating its competence in precise segmentation and reduced execution time. U-Net demonstrated high segmentation accuracy, achieving a DSC score of 0.8373. Attention U-Net demonstrated good segmentation accuracy with a remarkable DSC score of 0.8238, while ColonSegNet achieved a satisfactory DSC score of 0.8118. Table 5 shows the results achieved.

Model	Loss	IoU	DSC	N Parameters
FF-UNet [42]	0.1649	0.7290	0.8350	3 942 930
U-Net [13]	0.1626	0.7240	0.8373	31 043 586
Attention U-Net [43]	0.1761	0.7048	0.8238	37 334 734
AResU-Net [44]	0.1504	0.7411	0.8495	39 090 446
ColonSegNet [1]	0.1881	0.6891	0.8118	4 929 026
ResCAM-Net (ours)	**0.1499**	**0.7429**	**0.8500**	**9 916 453**

Table 5. Performance evaluation on the Kvasir-SEG dataset

Figure 12 displays the results of each model on the Kvasir-SEG dataset. The first column of the figure displays the original image to be segmented, the second column presents the corresponding ground truth mask, and the last column exhibits the predicted mask produced by our proposed model. The results shown in the figure demonstrate that the performance of FF-UNet and our proposed model is superior to that of the other tested models.

These results offer valuable guidance for selecting the most suitable model for image segmentation tasks using the Kvasir-SEG and ISIC-2017 datasets, considering precision and the number of parameters, which will have an impact on model convergence time.

6 DISCUSSION AND CONCLUSION

The semantic segmentation of medical images plays a crucial role in the field of medical image analysis. This paper introduces a robust segmentation network, an enhancement of the ColonSegNet model. The primary objective in refining the ColonSegNet architecture is to create a robust segmentation model applicable not only to colonoscopy images but also to a broad spectrum of medical images. Our contribution ResCAM-Net integrates various structural enhancements, including the integration of residual functions in both encoder and decoder components, the introduction of a multi-kernel residual convolution module, a SE-ASPP module for multi-scale feature recalibration, and a triple attention module positioned at the bottleneck of the network. This unique configuration prioritizes relevant regions and eliminates superfluous elements in skip connection features, facilitating the acquisition of feature maps with rich semantic content and discrimination, all while retaining crucial spatial details. The results from our experiments, assessed across three benchmark datasets of varying modalities, underscore the superiority of our model over state-of-the-art segmentation methods in medical image segmentation

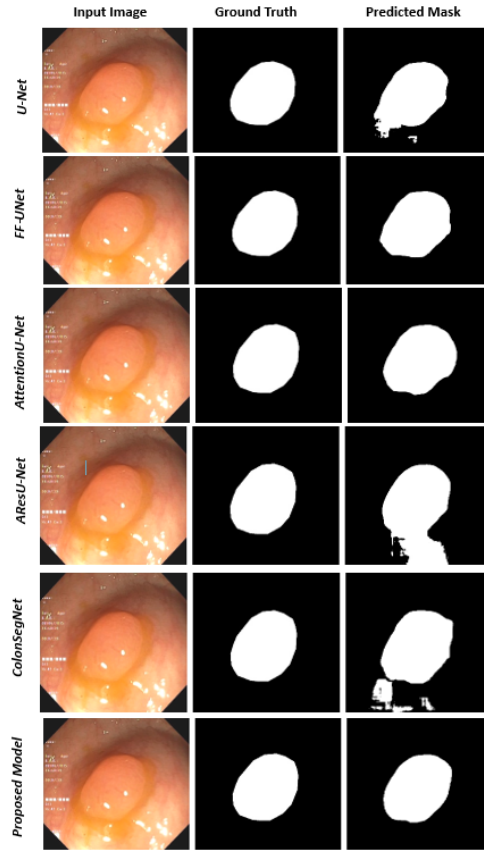


Figure 12. Qualitative segmentation results using six models on Kvasir-SEG dataset

tasks. Our model succeeds in producing comparable results for both large and small databases, while significantly reducing the number of trainable parameters. Notably, one challenge with this architecture is the selection of appropriate hyperparameters for each dataset. Our future research involves reducing the number of parameters and computational complexity by automating hyperparameter selection through optimization methods. We also intend to implement an Attention-Gate-Based model to capture spatial and fine-grained information that may otherwise be lost within skip connections.

REFERENCES

[1] Jha, D.—Ali, S.—Tomar, N. K.—Johansen, H. D.—Johansen, D.—Rittscher, J.—Riegler, M. A.—Halvorsen, P.: Real-Time Polyp Detection,

Localization and Segmentation in Colonoscopy Using Deep Learning. IEEE Access, Vol. 9, 2021, pp. 40496–40510, doi: 10.1109/ACCESS.2021.3063716.

[2] Wang, W.—Liang, D.—Chen, Q.—Iwamoto, Y.—Han, X. H.—Zhang, Q.—Hu, H.—Lin, L.—Chen, Y. W.: Medical Image Classification Using Deep Learning. In: Chen, Y. W., Jain, L. C. (Eds.): Deep Learning in Healthcare: Paradigms and Applications. Springer, Cham, Intelligent Systems Reference Library, Vol. 171, 2020, pp. 33–51, doi: 10.1007/978-3-030-32606-7_3.

[3] Cherukuri, V.—Ssenyonga, P.—Warf, B. C.—Kulkarni, A. V.—Monga, V.—Schiff, S. J.: Learning Based Segmentation of CT Brain Images: Application to Postoperative Hydrocephalic Scans. IEEE Transactions on Biomedical Engineering, Vol. 65, 2018, No. 8, pp. 1871–1884, doi: 10.1109/TBME.2017.2783305.

[4] Li, W.—Jia, F.—Hu, Q.: Automatic Segmentation of Liver Tumor in CT Images with Deep Convolutional Neural Networks. Journal of Computer and Communications, Vol. 3, 2015, No. 11, pp. 146–151, doi: 10.4236/jcc.2015.311023.

[5] Song, T. H.—Sanchez, V.—ElDaly, H.—Rajpoot, N. M.: Dual-Channel Active Contour Model for Megakaryocytic Cell Segmentation in Bone Marrow Trephine Histology Images. IEEE Transactions on Biomedical Engineering, Vol. 64, 2017, No. 12, pp. 2913–2923, doi: 10.1109/TBME.2017.2690863.

[6] Fu, H.—Cheng, J.—Xu, Y.—Wong, D. W. K.—Liu, J.—Cao, X.: Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation. IEEE Transactions on Medical Imaging, Vol. 37, 2018, No. 7, pp. 1597–1605, doi: 10.1109/TMI.2018.2791488.

[7] Chen, C.—Qin, C.—Qiu, H.—Tarroni, G.—Duan, J.—Bai, W.—Rueckert, D.: Deep Learning for Cardiac Image Segmentation: A Review. Frontiers in Cardiovascular Medicine, Vol. 7, 2020, Art. No. 25, doi: 10.3389/fcvm.2020.00025.

[8] Onishi, Y.—Teramoto, A.—Tsujimoto, M.—Tsukamoto, T.—Saito, K.—Toyama, H.—Imaizumi, K.—Fujita, H.: Multiplanar Analysis for Pulmonary Nodule Classification in CT Images Using Deep Convolutional Neural Network and Generative Adversarial Networks. International Journal of Computer Assisted Radiology and Surgery, Vol. 15, 2020, No. 1, pp. 173–178, doi: 10.1007/s11548-019-02092-z.

[9] Bottou, L.—Cortes, C.—Denker, J. S.—Drucker, H.—Guyon, I.—Jackel, L. D.—LeCun, Y.—Muller, U. A.—Sackinger, E.—Simard, P.—Vapnik, V.: Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition. Proceedings of the 12th IAPR International Conference on Pattern Recognition, IEEE, Vol. 2, 1994, pp. 77–82, doi: 10.1109/ICPR.1994.576879.

[10] Krizhevsky, A.—Sutskever, I.—Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C. J., Bottou, L., Weinberger, K. (Eds.): Advances in Neural Information Processing Systems 25 (NIPS 2012). Curran Associates, Inc., 2012, pp. 1097–1105, https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[11] Simonyan, K.—Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014, doi: 10.48550/arXiv.1409.1556 (arXiv Preprint arXiv:1409.1556).

[12] Long, J.—Shelhamer, E.—Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[13] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F. (Eds.): Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, Lecture Notes in Computer Science, Vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[14] Badrinarayanan, V.—Kendall, A.—Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, 2017, No. 12, pp. 2481–2495, doi: 10.1109/TPAMI.2016.2644615.

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

[16] Akbari, M.—Mohrekesh, M.—Nasr-Esfahani, E.—Soroushmehr, S. M. R.—Karimi, N.—Samavi, S.—Najarian, K.: Polyp Segmentation in Colonoscopy Images Using Fully Convolutional Network. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 69–72, doi: 10.1109/EMBC.2018.8512197.

[17] Xiao, Y.—Zhao, J.—Yu, Y.—Ding, X.—Liu, S.—Bao, W.—Wen, S.—Zhou, X.: SimpleCNN-Unet: An Optic Disc Image Segmentation Network Based on Efficient Small-Kernel Convolutions. Expert Systems with Applications, Vol. 256, 2024, Art. No. 124935, doi: 10.1016/j.eswa.2024.124935.

[18] Yuan, C.—Zhao, D.—Agaian, S. S.: UCM-Net: A Lightweight and Efficient Solution for Skin Lesion Segmentation Using MLP and CNN. Biomedical Signal Processing and Control, Vol. 96, Part B, 2024, Art. No. 106573, doi: 10.1016/j.bspc.2024.106573.

[19] Liu, X.—Gao, P.—Yu, T.—Wang, F.—Yuan, R. Y.: CSWin-Unet: Transformer UNet with Cross-Shaped Windows for Medical Image Segmentation. Information Fusion, Vol. 113, 2025, Art. No. 102634, doi: 10.1016/j.inffus.2024.102634.

[20] Wan, Y.—Zhou, D.—Wang, C.: UMF-Net: A UNet-Based Multi-Branch Feature Fusion Network for Colon Polyp Segmentation. Biomedical Signal Processing and Control, Vol. 99, 2025, Art. No. 106851, doi: 10.1016/j.bspc.2024.106851.

[21] Karthikha, R.—Jamal, D. N.—Rafiammal, S. S.: An Approach of Polyp Segmentation from Colonoscopy Images Using Dilated-U-Net-Seg – A Deep Learning Network. Biomedical Signal Processing and Control, Vol. 93, 2024, Art. No. 106197, doi: 10.1016/j.bspc.2024.106197.

[22] Yin, W.—Zhou, D.—Nie, R.: DI-Unet: Dual-Branch Interactive U-Net for Skin Cancer Image Segmentation. Journal of Cancer Research and Clinical Oncology, Vol. 149, 2023, No. 17, pp. 15511–15524, doi: 10.1007/s00432-023-05319-4.

[23] Targ, S.—Almeida, D.—Lyman, K.: Resnet in Resnet: Generalizing Residual Architectures. 2016, doi: 10.48550/arXiv.1603.08029 (arXiv Preprint arXiv:1603.08029).

[24] Diakogiannis, F. I.—Waldner, F.—Caccetta, P.—Wu, C.: ResUNet-A: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 162, 2020, pp. 94–114, doi: 10.1016/j.isprsjprs.2020.01.013.

[25] Jiang, S.—Li, J.—Hua, Z.: GR-Net: Gated Axial Attention ResNest Network for Polyp Segmentation. International Journal of Imaging Systems and Technology, Vol. 33, 2023, No. 5, pp. 1531–1548, doi: 10.1002/ima.22887.

[26] Liu, S.—Wang, P.—Lin, Y.—Zhou, B.: SMRU-Net: Skin Disease Image Segmentation Using Channel-Space Separate Attention with Depthwise Separable Convolutions. Pattern Analysis and Applications, Vol. 27, 2024, No. 3, Art. No. 93, doi: 10.1007/s10044-024-01307-7.

[27] Rumelhart, D. E.—Hinton, G. E.—Williams, R. J.: Learning Representations by Back-Propagating Errors. Nature, Vol. 323, 1986, No. 6088, pp. 533–536, doi: 10.1038/323533a0.

[28] Glorot, X.—Bordes, A.—Bengio, Y.: Deep Sparse Rectifier Neural Networks. In: Gordon, G., Dunson, D., Dudík, M. (Eds.): Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research (PMLR), Vol. 15, 2011, pp. 315–323, https://proceedings.mlr.press/v15/glorot11a.

[29] LeCun, Y.—Boser, B.—Denker, J. S.—Henderson, D.—Howard, R. E.—Hubbard, W.—Jackel, L. D.: Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, Vol. 1, 1989, No. 4, pp. 541–551, doi: 10.1162/neco.1989.1.4.541.

[30] LeCun, Y.—Bottou, L.—Bengio, Y.—Haffner, P.: Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, Vol. 86, 1998, No. 11, pp. 2278–2324, doi: 10.1109/5.726791.

[31] Ioffe, S.—Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Bach, F., Blei, D. (Eds.): Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR), Vol. 37, 2015, pp. 448–456, https://proceedings.mlr.press/v37/ioffe15.html.

[32] Ahmed, M. R.—Ashrafi, A. F.—Ahmed, R. U.—Shatabda, S.—Islam, A. K. M. M.—Islam, S.: DoubleU-NetPlus: A Novel Attention and Context-Guided Dual U-Net with Multi-Scale Residual Feature Fusion Network for Semantic Segmentation of Medical Images. Neural Computing and Applications, Vol. 35, 2023, No. 19, pp. 14379–14401, doi: 10.1007/s00521-023-08493-1.

[33] Chen, L. C.—Papandreou, G.—Kokkinos, I.—Murphy, K.—Yuille, A. L.: DeepLab: Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, 2018, No. 4, pp. 834–848, doi: 10.1109/TPAMI.2017.2699184.

[34] Woo, S.—Park, J.—Lee, J. Y.—Kweon, I. S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Yair, W. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11211, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[35] Hu, J.—Shen, L.—Sun, G.: Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[36] Jha, D.—Smedsrud, P. H.—Riegler, M. A.—Halvorsen, P.—de Lange, T.—Johansen, D.—Johansen, H. D.: Kvasir-SEG: A Segmented Polyp Dataset. In: Ro, Y. M., Cheng, W. H., Kim, J. et al. (Eds.): MultiMedia Modeling (MMM 2020). Springer, Cham, Lecture Notes in Computer Science, Vol. 11962, 2020, pp. 451–462, doi: 10.1007/978-3-030-37752-6_42.

[37] Codella, N. C. F.—Gutman, D.—Celebi, M. E.—Helba, B.—Marchetti, M. A.—Dusza, S. W.—Kalloo, A.—Liopyris, K.—Mishra, N.—Kittler, H.—Halpern, A.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.

[38] Tschandl, P.—Rosendahl, C.—Kittler, H.: The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. Scientific Data, Vol. 5, 2018, Art. No. 180161, doi: 10.1038/sdata.2018.161.

[39] Shorten, C.—Khoshgoftaar, T. M.: A Survey on Image Data Augmentation for Deep Learning. Journal of Big Data, Vol. 6, 2019, No. 1, Art. No. 60, doi: 10.1186/s40537-019-0197-0.

[40] Sudre, C. H.—Li, W.—Vercauteren, T.—Ourselin, S.—Cardoso, M. J.: Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: Cardoso, M. J., Arbel, T., Carneiro, G. et al. (Eds.): Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA 2017, ML-CDS 2017). Springer, Cham, Lecture Notes in Computer Science, Vol. 10553, 2017, pp. 240–248, doi: 10.1007/978-3-319-67558-9_28.

[41] El Abassi, F.—Darouichi, A.—Ouaarab, A.: Images Segmentation Using Deep Learning Algorithms and Metaheuristics. 2022 8th International Conference on Optimization and Applications (ICOA), IEEE, 2022, pp. 1–6, doi: 10.1109/ICOA55659.2022.9934130.

[42] Iqbal, A.—Sharif, M.—Khan, M. A.—Nisar, W.—Alhaisoni, M.: FF-Unet: A U-Shaped Deep Convolutional Neural Network for Multimodal Biomedical Image Segmentation. Cognitive Computation, Vol. 14, 2022, No. 4, pp. 1287–1302, doi: 10.1007/s12559-022-10038-y.

[43] Oktay, O.—Schlemper, J.—Folgoc, L. L.—Lee, M.—Heinrich, M.—Misawa, K.—Mori, K.—McDonagh, S.—Hammerla, N. Y.—Kainz, B.—Glocker, B.—Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. 2018, doi: 10.48550/arXiv.1804.03999 (arXiv Preprint arXiv:1804.03999).

[44] Zhang, J.—Lv, X.—Zhang, H.—Liu, B.: AResU-Net: Attention Residual U-Net for Brain Tumor Segmentation. Symmetry, Vol. 12, 2020, No. 5, Art. No. 721, doi: 10.3390/sym12050721.

Fouzia El Abassi received her Ph.D. degree in computer science and artificial intelligence from the Cadi Ayyad University, Marrakech, Morocco. She is currently Professor at the Moroccan School of Engineering Sciences (EMSI), Marrakech. She is affiliated with the Computer and Systems Engineering Laboratory (L2IS), Cadi Ayyad University, and the LAMIGEP Laboratory at EMSI, Marrakech, Morocco. Her research interests include artificial intelligence, deep learning, medical image analysis, image segmentation, convolutional neural networks, and metaheuristic optimization.

Aziz Darouichi received his Ph.D. degree in applied mathematics and computer science. He is currently Professor with the Faculty of Sciences and Techniques, Cadi Ayyad University, Marrakech, Morocco, and a member of the Laboratory of Computer and Systems Engineering (L2IS). His current research interests include artificial intelligence, data science, computer vision, image processing, medical image analysis, deep learning, bio-inspired algorithms, and healthcare applications.

Aziz Ouaarab received his Ph.D. degree in computer science. He is currently Professor with the Department of Computer Science, Faculty of Sciences and Techniques, Cadi Ayyad University, Marrakech, Morocco, and a member of the Laboratory of Computer and Systems Engineering (L2IS). His current research interests include artificial intelligence, computational intelligence, metaheuristic optimization, swarm intelligence, combinatorial optimization, and deep learning.