

Received 9 December 2024, accepted 22 December 2024, date of publication 3 January 2025, date of current version 14 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3525813

RESEARCH ARTICLE

RMHA-Net: Robust Optic Disc and Optic Cup Segmentation Based on Residual Multiscale Feature Extraction With Hybrid Attention Networks

MOHAMMAD J. M. ZEDAN^{1,2}, (Graduate Student Member, IEEE), SITI RAIHANAH ABDANI³, JAESUNG LEE⁴, AND MOHD ASYRAF ZULKIFLEY¹, (Member, IEEE)

¹Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia

²Computer and Information Engineering Department, College of Electronics Engineering, Ninevah University, Mosul 41002, Iraq

³School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam 40450, Malaysia

⁴Department of Artificial Intelligence, Chung-Ang University, Dongjak, Seoul 06974, Republic of Korea

Corresponding author: Mohd Asyraf Zulkifley (asyraf.zulkifley@ukm.edu.my)

This work was supported by the Ministry of Higher Education Malaysia through Fundamental Research Grant Scheme under Grant FRGS/1/2022/TK07/UKM/02/4.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research and Studies Department, Ministry of Health, Iraq under Approval No. 49213, and performed in line with the Declaration of Helsinki.

ABSTRACT Glaucoma is a chronic eye disease that damages the optic nerve, often leading to permanent vision loss. Early screening with automated technology is crucial to assist ophthalmologists in making accurate diagnoses. One of the key technologies for automated diagnosis is the segmentation of the optic disc (OD) and optic cup (OC). In this paper, RMHA-Net is developed using a residual multiscale feature extractor with a hybrid attention mechanism, which is introduced purposely for automated OD and OC segmentation. This network's encoder is designed based on advanced convolutional neural network (CNN) blocks that combine dilated convolution, which allows field-of-view expansion by capturing larger-scale features. In addition, the encoder also embeds residual connections to improve the model capacity in extracting low-level features. This design accurately separates the OD and OC from surrounding retinal tissues, handling complex environmental and anatomical changes. The proposed network is further improved by integrating two modules to enhance the segmentation performance: 1) a multiscale feature extractor module to provide various scales contextual information, and 2) dual attention mechanisms through channel-wise and spatial-wise mechanisms so that irrelevant information or noise can be excluded by assigning lesser weights to irrelevant features. To validate RMHA-Net's effectiveness, extensive experiments were conducted using five public datasets: Drishti-GS, ORIGA, PAPILA, Chaksu, and REFUGE, and one private dataset, Ibn Al-Haitham. The proposed model outperformed seven cutting-edge segmentation models for OD and OC segmentation. The results demonstrate that the network extracts detailed features, offering an efficient framework for future studies.

INDEX TERMS MS attention mechanism, glaucoma screening, multi-scale features, optic cup segmentation, optic disc segmentation, semantic segmentation.

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino¹.

I. INTRODUCTION

Glaucoma is a chronic optical neuropathy that affects millions of individuals by inducing progressive damage to the optic

nerve and eventually results in permanent vision loss [1]. The prevalence of glaucoma varies according to numerous factors such as age, race, and family history [2], [3]. This disease is mainly caused by increased fluid pressure inside the eye due to blockage in the drainage channels for excess fluid, which leads to the deterioration of nerve fibers manifested by an increase in the size of the optic cup compared to the constant-sized optic disc as well as thinning of the retinal nerve fiber layer (RNFL). However, all these changes occur without any obvious symptoms during the early stages of glaucoma, whereby it is difficult to treat once it reaches the advanced stage. Therefore, early detection technology becomes an essential tool for effective glaucoma treatment to prevent adverse effects of the late stage [1], [4], [5]. Nowadays, colour fundus images and optical coherence tomography (OCT) are the most extensively used imaging modalities for monitoring eye diseases including glaucoma [6], [7]. Compared to OCT, fundus images are a cheaper option that is highly accessible in most medical centres. These 2D images provide detailed information about the optic nerve head (ONH), including structural alterations used to indicate glaucoma's presence manually or automatically by calculating the cup-to-optic disc ratio (CDR). The primary goal of evaluating ONH is to assess either binary or multi-stage classifications of glaucoma, based on the CDR ratio, which is the ratio of the vertical cup diameter (VCD) to the vertical disc diameter (VDD). A high ratio (> 0.65) indicates the enlargement of the optic cup, which is clinically considered as a major sign of glaucoma as illustrated in Fig 1. [8], [9], [10].

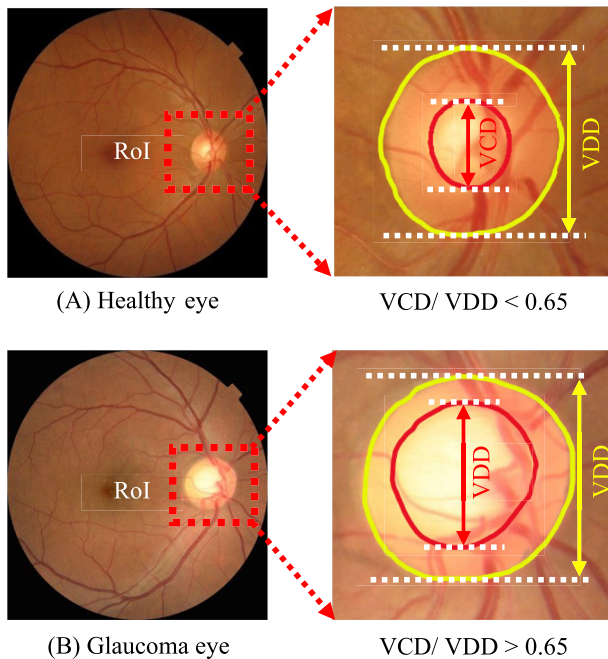


FIGURE 1. (A) Healthy eye ($CDR < 0.65$), (B) Glaucoma eye ($CDR > 0.65$).

These diagnosis processes are complicated and time-consuming due to the mixed morphological features of the

blood vessels and the pathological changes in the anatomy of the retina. Consequently, the process of identifying OD and OC boundaries needs to be as accurate as possible, which propels the development of automatic algorithms to accurately segment these two regions from fundus images [11].

Initially, the attention was geared towards segmenting the optic disc and cup using image-based segmentation methods that centered around extracting image characteristics like colors, textures, gradients, and contrasts by employing several algorithms such as threshold method, region growth, and active contour model. However, there are multiple drawbacks to these methods such as low adaptability to background changes, inconsistency of extracted features under illumination variations, and robustness fluctuation between different fundus imaging. Hence the recent methodologies have focused on applying deep learning-based segmentation algorithms to address the previously mentioned challenges [12], [13], [14], [15].

Semantic segmentation is one of the most widely used approaches for such tasks, whereby the network is built upon deep convolutional blocks utilizing U-shaped benchmark architecture to form an encoder-decoder structure to assign a class label to each pixel in the fundus image. Due to the capability of the deep segmentation network that is able to collect both high-level context and fine-grained features, making it possible to accurately determine the boundaries of the optic disc and cup.

Nevertheless, this deep learning approach faces several challenges and limitations, such as data-hungry issues, class imbalance, heterogeneous datasets, and biasing among algorithms [14], [17], [18].

Additionally, most deep learning methods use only a single scale information for each layer of feature maps, which hinders the extraction of the detailed features of the OD and OC, particularly the fine-detailed patterns. This particular part can be bypassed using multiscale feature learning that integrates convolutional kernels of different sizes to extract multi-scale information simultaneously, hence expanding the effective field of view [19]. In addition, a pyramid pooling structure called Spatial Pyramid Pooling (SPP) can be used to extract semantic features using different scales of a pooling kernel [20]. The pooling kernels occupy the whole, half, and small portions of the image, producing fixed-length outputs which are then merged with the neighboring scale feature maps. In the same context, the Atrous Spatial Pyramid Pooling (ASPP) module can also be employed to gather multiscale context-based data using various dilated convolutional operations. This module allows larger extracted semantic information to be pooled within the same scale of feature maps [19], [22], [23].

Moreover, to reduce irrelevant features, an attention mechanism (channel attention, spatial attention, and mixed attention) can also be applied to increase the focus on the image's target features during the training process, in other words, reducing the training time, improving segmentation precision and model interpretability [24], [25]. Channel attention

allows the model to adaptively focus on significant characteristics across the channel dimension by assigning high distinct weights to channels that include informative regions within the feature maps. As a result, relevant channels gain more influence, while less relevant channels are diminished, resulting in boosting the representational capacity of the network [26]. On the other hand, spatial attention allocates the informative spots such as edges or textures within the feature maps while neglecting the less relevant areas. These two types of attention can be applied jointly as a hybrid module or individually within the segmentation models to boost the performance of the deep model by targeting the informative channel and relevant spots within the visual scene [27], [28].

Based on the aforementioned information, a deep segmentation architecture named “RMHA-Net” is proposed to semantically segment OD and OC. This architecture is further improved by integrating the multiscale features extractor and attention mechanisms, targeted to the OD and OC. Furthermore, RMHA-Net also achieves superior computational efficiency. The main contributions of this paper can be summarized as follows:

- 1) Introduces an encoder-decoder architecture that employs dilated convolutions and residual connections to capture complex features from a broader context;
- 2) Integrates multi-scale feature extractor modules to effectively capture complex spatial relationships and variations within the feature maps;
- 3) Incorporates the channel and spatial attention modules separately to increase the focus on the targeted features;
- 4) Performs comprehensive ablation experiments to investigate the contribution of each module with optimal hyperparameter configurations;
- 5) Demonstrates superior computational efficiency, with notable improvements in FLOPs, inference time, and model size that ensure RMHA-Net suitability for real-time applications;
- 6) Outperforms cutting-edge segmentation models for glaucoma diagnosis over six datasets, and demonstrates superior robustness and adaptability in diverse conditions.

The rest of this paper is organized as follows: Section II discusses some related studies for the U-shaped semantic segmentation models, Section III presents the architecture of the proposed model, Section IV presents the used datasets, data preprocessing operations, implementation environments, evaluation metrics, and experimental findings, and Section V discusses the findings and concludes this study.

II. RELATED WORKS

Glaucoma detection is a challenging task and depends heavily on the segmentation outputs of the optic disc and cup regions. Therefore, researchers set out to assist ophthalmologists by developing methods based on artificial intelligence to facilitate and accelerate the diagnosis of such diseases. Due to the high performance of the CNN model, it has

been considered an essential approach for many popular segmentation/classification tasks. The authors of [29] promoted the basic Fully Convolutional Network structure to be used for pixel-wise predictions that exclude the fully connected layers resulting in segmentation maps of arbitrary size. In this section, we will comprehensively explore several relevant studies, organized according to their proposed techniques to systematically evaluate their works. In the first section, we aim to present a clear and structured summary of the existing literature that relies on the traditional CNN network to extract features of interest, while the second part highlights the more advanced and complex networks that contain added modules to improve the segmentation performance, specifically for glaucoma detection as follows:

A. SEMANTIC SEGMENTATION BASED ON THE CONVENTIONAL CNN

Semantic segmentation [30] is a fundamental step in the detection of many diseases. However, it remains a complicated task that depends heavily on the complexity of the segmented subjects. Veena et.al [31] proposed a deep learning approach, built upon two distinct U-shaped CNN models, each composed of 39 layers, to automatically diagnose glaucoma by segmenting the optic cup and optic disc to extract the cup-to-disc ratio. The proposed models are trained and evaluated using the well-known DRISHTI-GS dataset. The input images are pre-processed by applying morphological operations to enhance the contrast factor, after that the optic nerve areas are detected and localized using Sobel filter and Watershed method. The proposed system achieved an accuracy of 98% for optic disc segmentation and 97% for optic cup segmentation.

However, the authors did not assess the performance of the models on multiple datasets or in practical clinical settings, in which the quality and characteristics of the disease’s features might vary significantly. Similarly, the authors of [8] proposed an improved deep learning U-Net model for optic disc segmentation based on a 19-layer encoder and decoder blocks to extract spatial information and localize the OD precisely. In addition, their proposed model employed morphological techniques to generate ground truth labels from manually annotated fundus images. The proposed system achieved an accuracy level of 99.7% and Intersection over Union (*IoU*) of 87.9% tested on various kinds of public datasets. However, the suggested model only addresses segmenting the optic disc while ignoring the segmentation of the optic cup, which is a key element for estimating the cup-to-disc ratio.

Furthermore, another deep learning framework for segmenting optic disc and cup areas is proposed by [29] which is based on a modified U-Net model. The improvement is achieved by integrating the pre-trained ResNet-34 model in the encoder part while retaining the original decoder part without any modification. In addition, post-processing processes are utilized to determine the vertical Cup-to-Disc Ratio. This model achieves faster convergence and improved

segmentation performance compared to conventional models. The proposed system is primarily trained using the RIGA dataset and achieved a dice value of 97.3% for the optic disc and 87.6% for the optic cup. Although the system was validated using two more datasets, DRISHTI-GS and RIM-ONE V3, they argued that it is important to validate the performance using low-quality images and images with severe disc atrophy conditions. Likewise, the work by [33] describe an automatic framework for glaucoma detection. Their work includes a contrast enhancement step to the input fundus images using histogram-based operations, green channel and ROI region extraction, a modified U-Net called Au-Net with adaptive convolution for segmenting the optic disc and optic cup, which is then used to compute the cup-to-disc CDR ratio. Finally, a random forest classifier is used to categorize the glaucomatous images based on the CDR values. The proposed system achieved an accuracy level of 99%, specificity of 93.5%, and sensitivity of 86.5% for the OC and OD segmentation, whereby the experiments have been simulated on the DRISHTI-GS dataset. However, the authors evaluated the performance of the proposed models using only one dataset which reduces the generalization capability of the model.

On the other hand, the M-Net, which is suggested by [34] is based on a joint optic disc and optic cup automatic end-to-end deep learning segmentation framework using a multi-label deep network. The proposed M-Net contains four sequential stages, the first one is the multiscale input stage that resizes the input images to scale the receptive fields. The second stage then utilizes a modified encoder-decoder U-Net that learns hierarchical representation followed by a side-output stage that works as an early classifier to capture distinctive features to generate the prediction map of the segmentation. The last stage introduces a multi-label loss function based on the dice coefficient, which is commonly used to evaluate segmentation efficiency with accessible ground truth.

Furthermore, polar transformations are applied to the original images to enhance the segmentation performance. The proposed system achieved an area under the curve of 0.85 tested on the ORIGA dataset and 0.90 tested on the SCES dataset. However, the proposed framework relies upon the initial manual identification of the optic nerve area and the subsequent implementation of the polar transformations, which may lead to inaccuracies in recognizing the OD and OC boundaries. Consistent with the previous findings, [35] presented a hierarchical deep learning framework called “Disc-aware Ensemble Network (DENet)” that fuses both global and local information from fundus images for glaucoma screening. The DENet comprises four parallel branches that extract the features simultaneously. The classification branch is based on the ResNet 50, followed by the segmentation branch that is tasked to detect disc region and embedding disc representation, which is based on the U-Net model. The third branch focuses on the local disc region to generate a diagnosis probability based on the disc area level, followed by

the final branch that uses the polar transformation to enlarge the structure of the optic disc and cup by applying some geometrical operations to further enhance the screening performance. The concluded screening outcome is determined by combining the output probabilities from various channels. The proposed system achieved an area under the curve of 0.817 on the SINDI dataset and 0.918 on the SCES dataset. However, applying the proposed complicated ensemble system with multiple stages and modifications for glaucoma diagnosis yields lower results than various simplified algorithms that were known to perform better.

Rather than UNet, SegNet has been used as a baseline model in [36] where a lightweight encoder-decoder framework was proposed by segmenting both the optic disc and optic cup for glaucoma screening. The encoder part is modified by integrating four dense blocks, each with two convolutional layers, that aim to increase the connections between the convolutional layers to preserve semantic feature information while keeping the decoder part simplified for an efficient upscaling process. The proposed model has a built-in bottleneck layer that is integrated within the encoder part after the concatenation of features.

The model also does not utilize pre- and post-processing operations so that less memory is used and reduces the computational cost. The proposed system was trained on DRISHTI-GS, RIM-ONE, and REFUGE datasets. Although the proposed system is devoid of pre- and post-processing, the augmentation processes and cropping of images were used before the training and testing of the model.

B. SEMANTIC SEGMENTATION BASED ON MULTISCALE MODULES AND ATTENTION MECHANISMS

Previous studies have shown that the inclusion of multiscale modules and attention mechanisms can enhance CNN models’ capabilities in extracting suitable features of interest and addressing some of their weaknesses in collecting contextual information by focusing on selective parts of the feature maps [37], [38].

Based on these concepts, the following studies have employed the conventional U-shape segmentation network as its core structure. One of the studies proposed a deep learning framework named “multiscale fused network with additive channel-spatial attention (MSF-ACSA)” [39] that aims at tackling class imbalance issues in medical image segmentation. They also integrated an additive attention mechanism to combine low-level and high-level features to target selective RoI by adjusting the channel’s importance factors of the feature maps. Furthermore, they include a multiscale module to improve the segmentation results by fusing fine-grained details with coarse-grained semantic information across different scales. They also incorporate an alternative loss function, which is based on focal weighted Tversky loss that reduces the effect of class imbalance. To validate the effectiveness of their proposed model, an ablation study was conducted using various module combinations.

The proposed framework was tested and evaluated using three public medical datasets that have class imbalance issues. However, they have not validated the performance for small RoI images.

In a similar vein, Jiang et al. [40] presented a segmentation framework based on an encoder-decoder backbone model, coupled with a residual path for spatial attention to capture spatial intersection relationships between the optic cup and optic disc. In addition to the cropping and augmentation operations, they proposed an integration of a multi-scale color restoration method and polar transformation as pre-processing steps to produce more distinct cup and disc borders, especially for the optic cup. Cropped samples of the Drishti-GS1 dataset are used to assess the model's performance which produced F1 scores of 0.975 for optic disc and 0.901 for optic cup segmentation, respectively. However, the suggested approach has not been validated by other standard metrics for segmentation procedures, as well as the complex pre-processing steps applied may limit the model's adaptability to various clinical or research contexts. Additionally, the authors of [41] proposed a U-shaped model with multi-scale input capability and different attention mechanisms for segmenting the optic cup and disc simultaneously. The input images are scaled to various sizes and the informative features are extracted using an adaptive channel attention module. The extracted features are combined with the scaled input and encoded using convolution blocks with a wide kernel to boost the spatial attention for segmenting the optic cup and disc. The bottleneck layer of their network comprises of a self-attention layer that functions as a bridge between the encoding and decoding sections, which aims to provide more global contextual information. The proposed framework is trained and evaluated using a unique training strategy that includes hybrid cross-entropy and dice loss functions trained over three public datasets (REFUGE, DRISHTI-GS, and RIM-ONE-r3). Yet, the main weakness of this method is its heavy reliance on cropped images, which does reflect the real requirement of an automated application. Additional study by Hua et al. [42] presented the DCAM-NET segmentation architecture to improve the generalization abilities of models for OC and OD segmentation. The design incorporates a Multi-Scale Attention Mechanism (MSA) to detect important features across channels and locations, and a Multi-Region Weight Fusion Convolution (MWFC) module to more effectively extract characteristics from source domain data. Experimental evaluations showed that the combination of the MSA and MWFC modules enhanced segmentation performance on unseen domain data. Despite that, the combination of the MSA and MWFC modules has the drawback of increasing computational complexity, which would make it more challenging to deploy in real-time clinical settings.

Another study by [43] introduced an improved U-Net architecture that incorporates attention mechanisms and residual connections to boost the model's capacity to collect local and global information necessary for precise

segmentation. To figure out which pre-trained backbone model is most suitable for segmentation tasks, the study examines some models, such as ResNet and InceptionV3.

The suggested model shows certain improvements in segmentation accuracy across different datasets, including Drishti-GS, REFUGE, and RIM-ONE-R3. The main constraint is associated with the intricacy of the model and its computing demands. This intricacy might make it more difficult to use the model in real-time clinical situations, especially on devices with little computing power.

Despite these findings, the following three studies focus on integrating the DeepLab to improve the segmentation performance, where Zheng et al. [44] proposed a lightweight encoder-decoder deep learning system with multiscale capabilities for segmenting the optic disc in fundus images. Two baseline models, U-Net and DeepLab-v3 were used as the benchmarked models. The model's encoder extracts low-level features from the optic disc region using dense connection blocks, yielding high informative features, while semantic information is extracted using high-resolution blocks. Additionally, the atrous spatial pyramid pooling (ASPP) module is applied to gather high-level contextual information. Conversely, the decoder leverages the encoder's multi-scale capability to predict the regions of interest. The performance of this model was tested using cropped samples with 93.5% *IoU* for the ORIGA dataset, 91.5% for the RIM-ONE-r1 dataset, and 94.7% for the MESSIDOR dataset. Although the suggested model is designed to be a lightweight model, yet, the authors do not provide any details on memory usage, processing time, and so on. Furthermore, their work focuses only on optic disk segmentation, while no optic cup segmentation has been addressed. For the same reason, Wei et al. [45] proposed an automatic deep learning framework based on a modified encoding and decoding structure to segment the optic disc and optic cup areas for glaucoma diagnosis. The input images are cropped using a basic U-Net structure to specify the RoI, after which, the cropped samples are resized using multiscale input to keep the informative features. After that, the scaled samples are passed through different separable and dilated convolutional blocks in the encoder and decoder layers to extract high-level semantic features, while preventing spatial information loss. The symmetrical layers of the proposed network are linked together by external skip connections to mitigate feature degradation that produces outstanding segmentation performance. The proposed system achieved Dice Coefficients of 97.8% and 91.9% for OD and OC, respectively, tested on the DRISHTI-GS dataset as well as 96.5% and 91.0% for OD and OC, respectively, tested on the REFUGE dataset. Despite the good segmentation performance, the problem of the model's stability still remains due to the diversity of resources of fundus samples, which needs an adaptation system to improve the generalizability and stability of the model.

In addition, the work by [46] introduced a deep learning framework for optic disc segmentation based on modified

DeepLab v3+ and U-Net models. The modification involves the integration of spatial and channel attention mechanisms between the encoder and decoder modules to boost the model's performance. In addition, they integrate conditional random fields strategy to refine the separation lines between different objects. The proposed framework is trained and evaluated using three public medical datasets (DRIONS-DB, RIM-ONE 3, and DRISHTI-GS). The input data are augmented to increase the size of the training dataset and avoid the possibility of overfitting issues.

Although the proposed model outperformed some of the benchmarked studies, it has some limitations, such as not using the network to detect the optic cup, which is considered a challenge itself in addition to the performance degradation and lack of robustness resulted from the variations in data distribution and the boundary delineation issue between the optic disc and the retinal background.

On the other hand, the following studies use a different approach that perform improved semantic segmentation based on the combination of the attention mechanisms or multi-scale techniques with the convolutional networks.

For instance, MSA-UNet proposed by [25] is intended exclusively for the segmentation of the OD and OC. This model combines attention processes with multi-scale image convolution, which together enhance segmentation accuracy and efficiently capture intricate information. The design divides the decoder path into two main branches: one concentrates on employing double convolution for upsampling, and the other on integrating attention gates with multi-kernel convolution. The ORIGA and REFUGE datasets were used to verify the efficacy of the proposed model. Although MSA-UNet looks to be producing good segmentation performance, it requires complex mathematical operations that need huge training data, which might make it difficult to apply in resource-constrained situations.

Similarly, a deep learning approach presented by [47] focuses on enhancing glaucoma diagnosis by learning the interactions between the inner and outer boundary features between the optic disc and cup. The proposed model employs multi-scale techniques to extract the features to alienate the boundaries effectively so that multiscale feature representations can be preserved. Furthermore, a pyramid-shaped attention framework is presented to capture better representations and mutual connections that exist between the optic disc and optic cup. The proposed approach incorporates a multi-scale loss function that resulted in an AUC of 89.4% and 79.6%, tested over Drishti-GS and RIM-ONE datasets, respectively. The performance of the model is tested using pre-processed and cropped samples that focus only on the OD area. It is important to validate the model's performance using raw input images, without much manual intervention.

In a comparable manner, the authors of [48] proposed a hybrid deep learning approach that combines local feature extraction model and multiscale module to fuse multiscale shallow features. The features are extracted from a hybrid

model to improve the segmentation performance of the optic cup and optic disc from retinal fundus images. The baseline convolutional model employed is ResNet-50 which serves as the backbone network to extract the features. Besides, the multiscale model incorporates the ASPP module to capture contextual information on different scales by changing the field of view of the input features by applying filters or pooling operations at various rates to overcome the base model constraints in learning global contextual information required for medical image segmentation. The performance of this model is tested using cropped samples of the REFUGE dataset that produced an *IoU* of 90.8% and 75.5% for the OD and OC, respectively. Despite the application of a hybrid system that includes many approaches and processed images, the segmented performance achievement falls short of expectations.

Despite various contributions from the previous studies, there are still some limitations and gaps that can be explored to further improve the segmentation performance. Some of the weaknesses that can be researched are removing a dedicated focus on segmenting the OD without the OC, reliance on extremely complex systems with limited outcomes, or validating the model's performance using non-standard metrics over limited datasets as well as eliminating complicated pre-processors that do not simulate reality. All of these drawbacks drove us towards proposing a framework of medium complexity, devoid of any image pre-processors, and highly efficient based on residual multiscale feature extractor with a hybrid attention mechanism detected for OD and OC segmentation.

III. METHODOLOGY

The proposed RMHA-Net network is an end-to-end semantic segmentation network that is dedicated to glaucoma disease based on 2D fundus input images. Its network architecture is designed to consist of an encoder, skip connection, and decoder, which is then integrated with strategic modules to extract distinct features of the OD and OC, to significantly improve segmentation accuracy performance. In the following subsections, we will go over the proposed framework in detail, including all the added modules.

A. RMHA-NET ARCHITECTURE

Normally, a basic encoder-decoder architecture is commonly utilized in the segmentation field, which also serves as the foundation for the proposed framework for glaucoma screening through the OC and OD segmentation. In general, the encoder part task is to extract high-dimensional features of the OD/OC sequentially across multiple scales to produce effective feature maps from the image space using a set of convolutional blocks, activation functions, and pooling layers. On the other hand, the decoder part of the network aims to progressively transform the encoded information into label maps by using a set of transposed convolutional blocks, activation functions, and expansion layers.

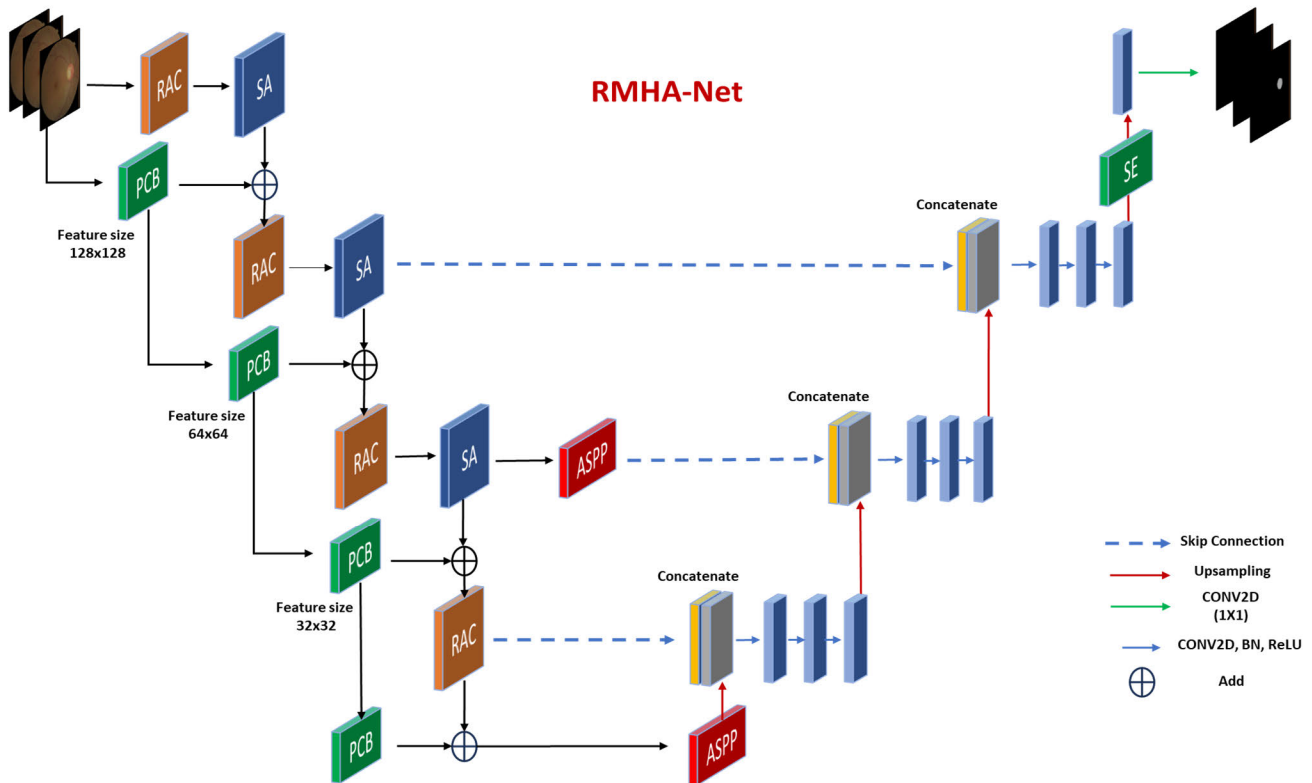


FIGURE 2. The Proposed Network (RMHA-Net).

In more detail, the encoder architecture of the proposed RMHA-Net shown in Fig 2 is made up of four consecutive layers in depth.

For each layer, the input is passed through the residual-atrous-conversion (RAC) module then passed through the spatial attention mechanism (SA) module and then added to the output of the pool-conv-batch (PCB) module. The atrous spatial pyramid pooling (ASPP) module is also included twice in the middle and bottom of the network. Each of these modules has a specific objective that aims to solve some outstanding issues related to glaucoma screening by discovering unique hidden features within the feature maps.

After that, these features are up-sampled within the decoder phase and then combined with the high-level context-dependent data provided by the skip connections. The merged features are passed through three sets of convolutions and activations followed by the Squeeze-and-Excitation (SE) module. Finally, a single 1×1 convolutional block with a sigmoid activation function is used to produce the final output masks. It is a challenging task, where the OC itself is located within the OD, making it difficult to alienate the boundaries clearly due to the features overlapping in the fundus images.

This issue becomes more challenging with the interference of blood vessels that cover the OD as well as the OC, preventing the segmentation network from producing effective high-contrast features.

The mentioned models are adapted to the proposed RMHA-Net in order to extract unique spatial information across different scales, whilst reducing the risk of false positives, hence boosting the segmentation performance. Each module is discussed in detail in the subsequent subsections:

1) RESIDUAL-ATROUS-CONV MODULE (RAC)

The first layer of this module starts with an atrous (dilated) convolution utilizing a 3×3 kernel. The atrous convolutions can enlarge the receptive field significantly so that a broader spatial context can be included while preserving the size of the parameters. A variable set of receptive fields is particularly crucial for identifying OD and OC regions that rely on contextual details from an expanded region of the retina. The dilated convolution is used as an alternative to the conventional down-sampling operation that negatively impacts the extracted information of different spatial resolutions. The dilation rate used in RMHA-Net is precisely tuned to achieve an appropriate combination of contextual richness and ensure the details are carried over throughout the network. The second layer incorporates batch normalization, which is added to keep the learning process more stable across the latter layers. After that, the activation function (ReLU) is added as the third layer that provides the non-linearity needed to allow the network to learn detailed patterns from the input data. By going deeper within the RAC module,

as shown in Fig 3, the first three layers are repeated three times consecutively using standard convolutional layers with a kernel size of 3×3 , utilizing zero padding mode to maintain a consistent output dimension with the input tensor, while contributing to a smooth transition throughout the proposed structure.

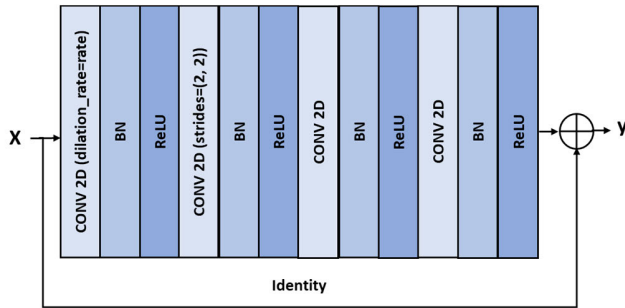


FIGURE 3. Residual-Atrous-Conv module (RAC).

Furthermore, the kernel initializer technique is included within the arrangement of convolutional layers to set the initial values of weights to achieve superior training performance. The last layer consists of a residual connection that combines the input tensor with the last convolutional layer outputs to minimize the network's complexity during the training process. As a result, the network will have a lower likelihood of overfitting, effectively mitigating the vanishing gradient problems, and maintaining a constant learning process throughout the network. This design is particularly useful for segmenting the OD and OC, where tiny changes in color and contour information are significant to diagnosis results, and keeping the information between layers will also help the network to improve the accuracy of these segmentation tasks.

2) POOL-CONV-BATCH MODULE (PCB)

This submodule utilizes several operators, starting with the pooling operation with a 2×2 pool size to execute the max pooling operation to shrink the spatial dimension of the feature maps, hence minimizing the model's sensitivity to overfitting.

The pooled feature maps are fed to two sets of convolutional operations with a 3×3 kernel and zero padding. In addition, this process is combined with an L2 regularization technique to reduce the possibility of overfitting and assure the model's generalization. After that, batch normalization and a ReLU activation function are placed sequentially. This PCB module also includes a dropout layer with a rate of 0.3, ensuring better generalizing ability of the model.

3) SPATIAL ATTENTION MODULE (SA)

The spatial attention module is designed to be an adaptive mechanism used for selecting spatial spots to build a set of focused maps that improve the segmentation performance by assigning higher weights towards informative parts.

This module is tasked to provide more emphasis on the edges, contrasts, and colors of the OD/OC regions, which are essential to accurately segment these regions, even under various dimensions, forms, and locations between different patients. The spatial attention mechanism begins with applying the channel-wise average pooling that receives the contextual information in parallel with the max-pooling operator, whereby these extracted feature maps represent the most salient features of the glaucoma disease. The resulting informative features are concatenated along the channel axis, which is then fed to the convolutional layers with a sigmoid activation function and adaptive kernel sizes that control the amount of information in generating the output feature maps.

The resulting maps are multiplied with the original input tensor to produce the spatial attention maps that will improve the overall model's sensitivity. The entire process can be summarized as modelled in (1):

$$Y = X \cdot \sigma(\text{Conv}(\text{Conc}[\text{AvgPool}(X), \text{MaxPool}(X)])) \quad (1)$$

where:

X and Y : Input and output feature map at position (i, j) in channel c

σ : Sigmoid activation function,

$\text{AvgPool}(Xi, j, k)$: Average pooling operation

$\text{MaxPool}(Xi, j, k)$: Max pooling operation

Conc : Concatenation operation.

Conv : convolution operation.

4) SQUEEZE-AND-EXCITATION MODULE (SE)

This submodule is added to boost the expressive capability of the segmentation model by dynamically adjusting the channel-wise feature outputs to include mutual dependency between channels. This network design is essential to properly capture the finer features of the OD/OC, which aids in better managing the variances in the physical appearance of the retinal such as the border and texture properties. In general, this process can be realized by squeezing the global spatial information using the Global Average Pooling operator to compress each channel of the input tensor into a single numeric value known as channel descriptor. Subsequently, the produced descriptor is passed to a dense convolutional layer to reduce the dimensionality of the channel descriptor using an adaptable ratio to extract the most significant characteristics, objectively, features related to OD and OC. On the contrary, the expansion process is carried out to restore the original dimensions of the channel descriptor by employing a second dense layer along with the sigmoid activation function. In the last step, the produced vector of the channel descriptor that contains the weights is resized to fit the input dimension, which is then multiplied by the original input map so that the network can prioritize the relevant channels based on respective adaptive weights. The entire process can be summarized as modelled

in (2):

$$Y = X \cdot \sigma \left(W2 \cdot \text{ReLU} \left(W1 \cdot \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_{i,j,c} \right) \right) \quad (2)$$

where:

X and Y : Input and output feature map at position (i,j) in channel c

σ : Sigmoid activation function,

$W1$ and $W2$: Weight matrix of the first and second fully connected layer

H and W : Height and width of the input feature map.

5) MIXED ATTENTION MODULE (MA)

An additional submodule is added to the RMHA-Net by incorporating both forms of attention mechanisms in a cascade scenario as illustrated in Fig 4.

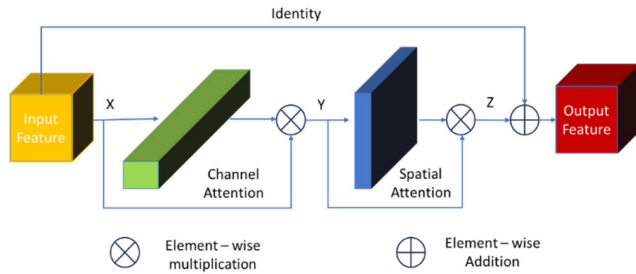


FIGURE 4. Mixed attention module.

The channel attention module aims to concentrate on the most significant bits of the input feature, while the spatial attention module aims to improve the representation of their respective characteristics. Finally, the output results are added with the feed-forwarded identity input to generate the output feature maps.

6) ATRous SPATIAL PYRAMID POOLING (ASPP)

Due to several factors of the glaucoma symptoms that are caused by age differences between the patients, imaging conditions, and ocular health, the relative size of the OD or the OC can vary to a certain range, whereby the segmentation algorithm needs to be able to cater for these scale variations.

Therefore, an ASPP submodule has been embedded into the RMHA-Net to extract these multiscale features to suit various glaucoma conditions for both optic disc and optic cup. The ASPP architecture as shown in Fig 5 is built to effectively deal with features of different scales and spatial contexts with the final aims of improving semantic segmentation performance.

The first stage includes applying a 1×1 convolutional layer to the input tensor to reduce feature dimensionality, hence decreasing computational complexity followed by a batch normalization operator to increase the stability and an activation function to add non-linearity to the resulting

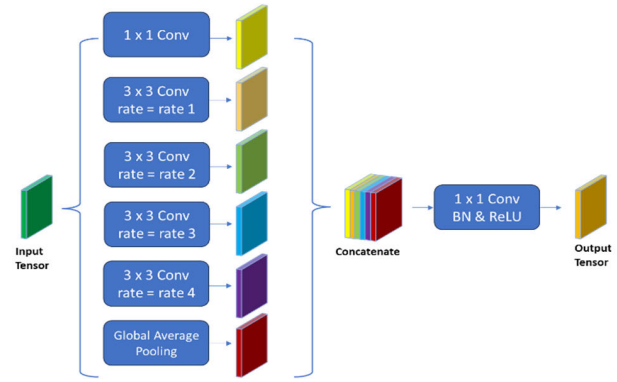


FIGURE 5. Atrous spatial pyramid pooling (ASPP).

feature maps. The second stage involves the addition of a set of atrous convolutional layers with different dilation rates to improve the network's capacity to depict complicated spatial relationships in object size. Every single convolutional path is accompanied by a batch normalization operator and a ReLU activation function. The third layer in this submodule depends on the global average pooling operator to further reduce spatial dimensions and extract the global context from the previous feature maps. The resulting feature maps from these three early stages are concatenated together along the channel axis in order to merge local and global contextual details across multiple scales. Finally, all feature maps produced from the previous layers are fed to a 1×1 convolutional operator, a batch normalization layer, and an activation function to refine the outcomes. By using different dilation rates in the ASPP submodule, the receptive field of view can be adapted to various scales.

For this proposed architecture, we incorporated four distinct dilation rates that match the general sizes and shapes of OD/OC within the retinal images to enhance the network's capability to capture both detailed and contextual information.

7) ENCODER DESCRIPTION

The first stage of the RMHA-Net encoder is composed of applying the RAC module with 64 filters and a unity dilation rate to the input fundus image to extract detailed spatial information while preserving the feature map resolution. The output of the RAC module is fed directly into the SA module to boost the focus on informative features and precisely capture the complex features of the OD/OC. The SA module improves the important characteristics found by the RAC while stifling less important data by analyzing the output of the RAC modules. As a result, segmentation accuracy is increased since only the most important characteristics are carried over to the subsequent layers of the network.

Meanwhile, the PCB module with 64 filters is applied to the input feature maps, again with the aim of reducing the dimension, while extracting a better set of features. The output of this layer is combined with the output of the second layer using an addition operator to improve the representation

of extracted features using different techniques. The second stage is quite similar to the first stage with the exception of an increment in the dilation rate value of the RAC to 3 and the doubling of filter numbers to capture wider spatial contexts.

The SA module is then employed to further refine the feature maps. Getting deeper into the third stage, the dilation rate increases to 4 with 256 filters. The goal of the RAC and SA modules prior to the ASPP is to capture and enhance features at various abstraction levels. By adding multi-scale details to these properties, the ASPP module expands on functionality. In this approach, the model is guaranteed to be able to segment items in the image precisely, independent of their size or position. The ASPP module called low-level ASPP is applied after the SA module using four ascending rates to cope with the multiscale context-related information.

In the last stage of the encoder, the dilation rate is raised again to 5 with 512 filters to enlarge the receptive fields. In addition, the SA module is excluded in this stage to decrease the training burden while integrating the ASPP module once again with different dilation rates (high-level ASPP). This stage serves as a bottleneck for the proposed U-shaped of the RMHA-Net architecture. To ensure that the computational complexity is acceptable, the input fundus images are resized to 256×256 pixels before passing them to the encoder. After each encoder stage, the size of the feature maps is reduced to half compared to the previous stage, resulting in 16×16 pixels feature maps for the shallowest part of the encoder.

8) DECODER DESCRIPTION

At the end of the encoder part, the feature maps will be passed to the decoder part to decode the latent variables through a series of operations that aim to convert the compressed encoded features back into pixel-wise prediction maps. Initially, the highest-level features produced from the high-level ASPP module are up-sampled by applying a bilinear interpolation technique to mimic the original input shape. In the meantime, feature maps resulting from the last convolutional operation in the last encoding stage are resized to match the dimensions of the upsampled feature maps from the ASPP output. These two sets of feature maps are then merged to combine the high-level context-dependent data and related spatial information. The merged features are passed through three sets of convolutions, batch normalizations and ReLU activation functions progressively to enhance the accuracy of the reconstructed feature maps. The upsampling, concatenation, and convolutional sets are designed sequentially to ensure compatibility of input feature maps received from the encoder part by iteratively recovering the spatial dimensions and refining the feature maps. At this point, the feature maps are upsampled to match the input tensor dimensions. Alongside, a channel attention mechanism (SE) is introduced to prioritize useful characteristics over the less important ones. Finally, a single 1×1 convolutional block with a sigmoid activation function is used to produce the final output masks, creating the segmented maps of OD and OC.

9) SKIP CONNECTIONS

To improve the performance of the proposed segmentation framework, skip connections are added among the symmetrical layers between the encoding and decoding parts, allowing a direct flow of information that also mitigates the vanishing gradients problem, while preventing information loss during feed-forward propagation. The first-layer skip connection is excluded in the RMHA-Net design, considering that the early encoder layers primarily capture basic features that may not contribute much to the segmentation task, whereas the deeper layers are designed to have feed-forward layers directly so that more complex and high-level features that are vital to segmentation accuracy can be provided. By avoiding the introduction of potentially redundant low-level details, the model architecture was simplified, leading to improved performance and more efficient learning. This choice is supported by actual findings from implemented experiments, which show that the model functions better without this particular skip connection.

B. GLAUCOMA DIAGNOSIS: CUP-TO-DISC RATIO

The Cup-to-Disc Ratio (CDR), which measures the proportion size of the OC to the OD, is mainly considered an essential indicator in the diagnosis of glaucoma. In particular, the Vertical CDR (VCDR) enhances this diagnostic procedure by focusing on the vertical dimensions, which are more informative in assessing glaucomatous damage because of the vertical elongation of the optic cup. To compute the VCDR, the RMHA-Net segmentation outputs were employed, which produces accurate binary masks for the OC and OD areas. These generated masks are used to determine the largest vertical diameters by identifying the topmost and bottom-most levels of non-zero pixels within the masks. After that, the vertical diameter is determined by applying the formula modelled in (3):

$$VCDR = \frac{\text{Vertical Cup Diameter}(VCD)}{\text{Vertical Disc Diameter}(VDD)} \quad (3)$$

According to clinical standards, a $VCDR > 0.65$ indicates a high risk for glaucoma that requires an additional follow-up with physicians, whereas a $VCDR < 0.65$ indicates a minimal risk of glaucoma. This binary decision-making approach used our segmentation model's output to improve diagnostic functionality by offering useful insights and patterns within the fundus images.

IV. EXPERIMENTS AND RESULTS

This section evaluates the proposed RMHA-Net framework using private and public glaucoma segmentation datasets. Section 4.1 provides the full description of the used datasets and the dataset splitting strategy, while section 4.2 introduces the preprocessing operations applied, and section 4.3 introduces the evaluation metrics employed to assess the segmentation performance. Finally, section 4.4 introduces the implementation details of the RMHA-Net and the respective experimental configurations.

A. DATASETS

To evaluate the proposed RMHA-Net, we conducted experiments using our private dataset known as the Ibn Al-Haitham dataset [49]. The colored fundus images are collected from Ibn Al-Haitham Eye Teaching Hospital in Iraq, taken from patients within the age range of 5 to 80 years. The ground truth segmentation labels of OD and OC are annotated manually using the GIMP-2 platform under the supervision of two experts who produced two individual masks for the optic disc and optic cup as illustrated in Fig 6.

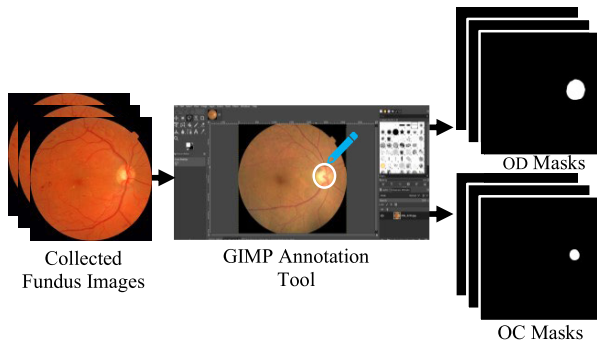


FIGURE 6. Annotation process for Ibn Al-haitham dataset.

These masks are annotated as binary images of black and white pixels only, with the white areas representing the optic disc or cup regions and the remainder of the eye components are labeled as black pixels, corresponding to background information. This dataset consists of 3000 fundus images with a resolution of 512×512 pixels, which aims to boost the training of deep learning-based models for Iraqi glaucoma detection. This dataset is split into 2400 training images and 600 testing images according to the 4:1 split ratio. Moreover, a wide collection of publicly available datasets including, Drishti-GS, ORIGA, PAPILA, Chaksu, and REFUGE, was employed to a comprehensive framework that enables an in-depth analysis of the proposed model's segmentation performance over a range of clinical imaging settings and dataset attributes.

B. PRE-PROCESSING STEPS

Preprocessing methods are critical steps for achieving high-quality segmentation of fundus images that frequently suffer from distortions and noise because of medical instrument limitations and environmental factors. These steps can enhance the image quality by removing the artifacts to improve the training performance of the model. Common techniques for preprocessing include noise reduction, contrast enhancement, and scaling to a constant resolution. In this study, the fundus images of all datasets are preprocessed by applying Contrast-limited Adaptive Histogram Equalization (CLAHE) and median filter, resulting in a notable improvement in the segmentation accuracy for delineating relevant clinical structures, such as the optic disc and optic cup. In addition, all the images are resized to a uniform size of

512×512 pixels to reduce the computational complexity and ensure consistency in input dimensions across the datasets.

The datasets are further processed using data augmentation techniques including rotation, horizontal flipping, and contrast adjustments to improve the dataset's generalization capability, reduce the likelihood of overfitting, and increase the variety of the training data [50]. The first augmentation technique implemented in this paper is horizontal flipping, which aims to introduce additional spatial variations by mirroring both images and masks along the vertical axis. Subsequently, the second augmentation method, random rotation is also conducted by rotating images and the respective masks at an angle ranging from -30 to $+30$ degrees. This technique uses a dedicated rotation matrix to align the image and its mask while keeping spatial consistency. Finally, the dataset is also enriched with the third augmentation technique by performing random brightness and contrast adjustments using a brightness range between -30 and $+30$ and contrast adjustment that is scaled by a random factor between 0.7 and 1.3 . All these three techniques are implemented on each dataset before executing the training phase, which is then saved with distinct filenames indicating the specific augmentation applied.

C. EVALUATION METRICS

In general, a semantic segmentation model is assessed using several complementary metrics, which are critical in measuring the performance of the proposed model and compare it with the benchmarked models. Practically, a combination of five performance metrics that are often used to provide a comprehensive assessment of the semantic segmentation model are used to measure the quality of OD and OC segmentation maps. Two metrics are considered as the primary measures, which are the mean Intersection over Union (mIoU) as formulated in (4) and the Dice Similarity Coefficient (DSC) as formulated in (5). The mIoU provides a statistical summary of overlapping region closeness between the segmented maps and the ground truth maps, while the DSC assesses the similarity measure of the predicted segmentation maps compared to the ground truth maps. Moreover, three more performance metrics are also used to supplement the validation process, which are the accuracy (ACC) as modeled in (6), sensitivity (SEN) or recall as modeled in (7), and Specificity (SPE) as modeled in (8), which will capture the model's effectiveness in different aspects. The accuracy metric represents the percentage of true predictions out of all predictions. The Sensitivity is the percentage measure of the true positives that are accurately segmented. Whereas, the specificity represents the percentage measure of the true negatives that are not properly segmented. These metrics depend on four basic terms generated by the confusion matrix components, which is a visual tool for illustrating a classification model's efficacy. The four terms are: False Positive (FP), False Negative (FN), True Negative (TN), and False Positive (TP).

$$mIoU = \frac{1}{N} \sum_{i=1}^N \left(\frac{TP_i}{TP_i + FP_i + FN_i} \right) \quad (4)$$

where N represents the number of classes.

$$DSC (Dice) = \frac{2 * TP}{(2 * TP + FP + FN)} \quad (5)$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (6)$$

$$SEN (recall) = \frac{TP}{(TP + FN)} \quad (7)$$

$$SPE = \frac{TN}{(FP + TN)} \quad (8)$$

Moreover, the Hausdorff Distance (HDD) [51] is applied as an additional metric to assess the extent of alignment between the predicted area and the ground truth mask. For the HDD metric, a lower score indicates a higher boundary accuracy and segmentation precision. The Hausdorff Distance for the i^{th} test image, is denoted by $HDDi$ as formulated in (9):

$$HDDi = \max \{ \sup_{p \in S_{X_i}} \inf_{q \in S_{Y_i}} \|p - q\|, \sup_{q \in S_{Y_i}} \inf_{p \in S_{X_i}} \|q - p\| \} \quad (9)$$

where:

S_{X_i} : Set of points corresponding to the boundary or non-zero region of the ground truth mask X .

S_{Y_i} : Set of points corresponding to the boundary or non-zero region of the predicted mask Y .

q and p : Points in S_{X_i} and S_{Y_i} , respectively.

$\|q - p\|$: Euclidean distance between points q and p .

$\sup$: Represents the maximum (supremum).

$\inf$: Represents the minimum (infimum).

The mean Hausdorff Distance across all N test images are then computed by the formula modelled in (10):

$$MeanHDD = \frac{1}{N} \sum_{i=1}^N HDDi \quad (10)$$

D. EXPERIMENTAL CONFIGURATIONS

To ensure that RMHA-Net has been designed to advance the performance of OD and OC segmentation, an extensive sequence of experiments has been conducted using the Keras + TensorFlow framework over the Kaggle platform with the support of NVIDIA Tesla GPU T4 $\times$ 2 accelerator that provides up to 29GB memory spaces. A grid search approach is applied to determine the optimal combination of loss function, batch size, learning rate. The results reveal that a crossentropy loss function with an 8-batch size and a learning rate of 0.0001 produced the best hyperparameter settings. Concurrently, we also trained the model using different optimizers that were used to update the weights, whereby the Adam optimizer was found to perform the best.

Furthermore, the ratio in the SE module that represents the scaling factor used to reduce the dimensionality of the channel descriptor is also experimented with to identify the best value.

Some of the possible ratios are in the set of [4], [8], [16], and [32], whereby all these values were systematically tested and the best performance is found to be the ratio of 8 for the

provided data. The same mechanism is repeated with another hyperparameter of the SA module which is kernel size that controls the amount of information involved to generate the feature maps with a possible set value of [3], [5], and [7], where the kernel size of 7 returned the best result. Likewise, the low- and high-level ASPP hyperparameter rates that indicate the used set of dilation rates are also fine-tuned to produce the best configuration to perceive features at multiple scales without losing resolution. Extensive attempts are conducted specifically for these values to ensure the best ranges that are compatible with the network requirements are chosen. All these experimental conditions are maintained consistently for the ablation tests to ensure fair comparison among the tested models.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. ABLATION STUDY

In this paper, an ablation study was carried out to evaluate the performance of the proposed framework with respect to the suggested additional submodules. A set of detailed experiments were conducted to identify the contribution of each added module; separately or in a combination mode. The tested additional submodules include 1) the multiscale module, which is represented by the ASPP, 2) the spatial attention mechanism, which is represented by the SA module, 3) the channel attention mechanism, which is represented by the SE module, and 4) the mixed attention module (MA) mechanism. These modules are added systematically, where each module is embedded into the baseline model by considering all the possibilities for inclusions and the variation of hyperparameters for each module.

Accordingly, the ASPP module was included after each layer of the encoder and decoder parts separately, whereby the simultaneous inclusion of more than one module is also considered at different locations, in addition to different depths and rates of the module. For each experiment, the results obtained for all metrics are recorded for comparison purposes. The same procedures were applied to the SA, SE and MA modules. Moreover, the inclusion of more than one module at a time was also studied to understand the effect of all modules on the proposed framework including the extent of compatibility between them. Table 1 demonstrates the segmentation performance and the obtained results after the inclusion of each module into the proposed baseline model.

The conducted ablation studies comprise three rounds of experiments. The first one is to focus only on the integration of a specific module into the proposed baseline network. After analyzing the results, it is clear that installing a single ASPP module (high-level ASPP) in the bottleneck between the encoder and decoder parts improves the glaucoma segmentation performance by precisely adjusting the field of view using four distinct scales to match the shapes and sizes of OD and OC within the retinal samples. Additionally, the representation of the deepest layer features, generated through the down-sampling process has also been enriched prior to the expansion mode during the decoder phase.

TABLE 1. Segmentation results of the ablation study tested on the Ibn Al-Haitham dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
BL (Proposed)	99.771	0.9291	0.9251	0.9042	0.9503	13.08	99.850	0.8223	0.8095	0.8115	0.8478	22.96
BL + ASPP	99.796	0.9399	0.9351	0.9220	0.9529	12.89	99.868	0.8385	0.8112	0.8122	0.7751	22.12
BL + 2 ASPP	99.800	0.9411	0.9359	0.9210	0.9387	12.45	99.866	0.8401	0.8142	0.8277	0.8012	22.01
BL + SE	99.800	0.9378	0.9347	0.9121	0.9434	12.67	99.867	0.8375	0.8121	0.8423	0.7841	22.45
BL + SA	99.799	0.9397	0.9370	0.9231	0.9535	12.50	99.868	0.8382	0.8145	0.8386	0.7917	22.31
BL + MA	99.797	0.9387	0.9335	0.9209	0.9510	13.06	99.868	0.8371	0.8117	0.8467	0.7794	22.98
BL + 2 ASPP + SE	99.801	0.9420	0.9375	0.9177	0.9409	12.25	99.869	0.8441	0.8258	0.8408	0.7922	21.86
BL + 2 ASPP + SA	99.804	0.9425	0.9380	0.9163	0.9408	12.17	99.865	0.8445	0.8268	0.8132	0.8205	21.75
BL + 2 ASPP + MA	99.800	0.9419	0.9370	0.9184	0.9425	12.32	99.863	0.8428	0.8244	0.8056	0.8234	21.88
BL + 2 ASPP + SE+ SA (RMHA-Net)	99.822	0.9437	0.9383	0.9210	0.9463	12.01	99.843	0.8471	0.8273	0.8117	0.8457	21.16

In addition, the output shows that integrating the low-level ASPP module with distinct scales managed to enhance the extracted features at the intermediate detail stage, while the high-level ASPP provides more in terms of a broader context. These dual modules enable tiny details and unique border patterns to be detected, contributing to better segmentation performance.

Furthermore, several experiments were also carried out to assess the integration of the SE module within the baseline model. The segmentation results generally improved, particularly when inserting the SE module with a precise ratio factor right before the final decoder layer. The added module increased the OD mIoU by 0.94% and the DSC by 1.04%, whereas the OC mIoU and DSC scaled are also improved by 1.85% and 0.26 %, respectively. These results indicate the SE module's capability to prioritize essential features that are crucial for OD segmentation. Meanwhile, multiple experiments were conducted to test the effect of integrating the SA module within the baseline network. It improved mIoU performance by 1.06% and DSC performance by 1.19% for OD segmentation, whereas OC mIoU and DSC also increased by 1.59% and 0.50%, respectively. These findings indicate the SA module's capability to focus on the most relevant parts of feature maps, hence enhancing the segmentation performance by generating more precise contours.

The performance of the proposed mixed attention module (MA) was also evaluated as an extra module that includes both forms of attention in a sequential scenario with direct identity mapping. The added modules enhanced the OD mIoU and DSC by 0.96% and 0.84%, respectively, while the OC mIoU and DSC rose by 1.56% and 0.22%. While the performance improvements indicate positive trends by integrating spatial and channel attentions, they are slightly smaller than the standalone SA or SE module due to the potential of overfitting when examined on various datasets.

The second round of the ablation study focuses on the integration of two types of submodules within the baseline network to better understand their behavior in a combined

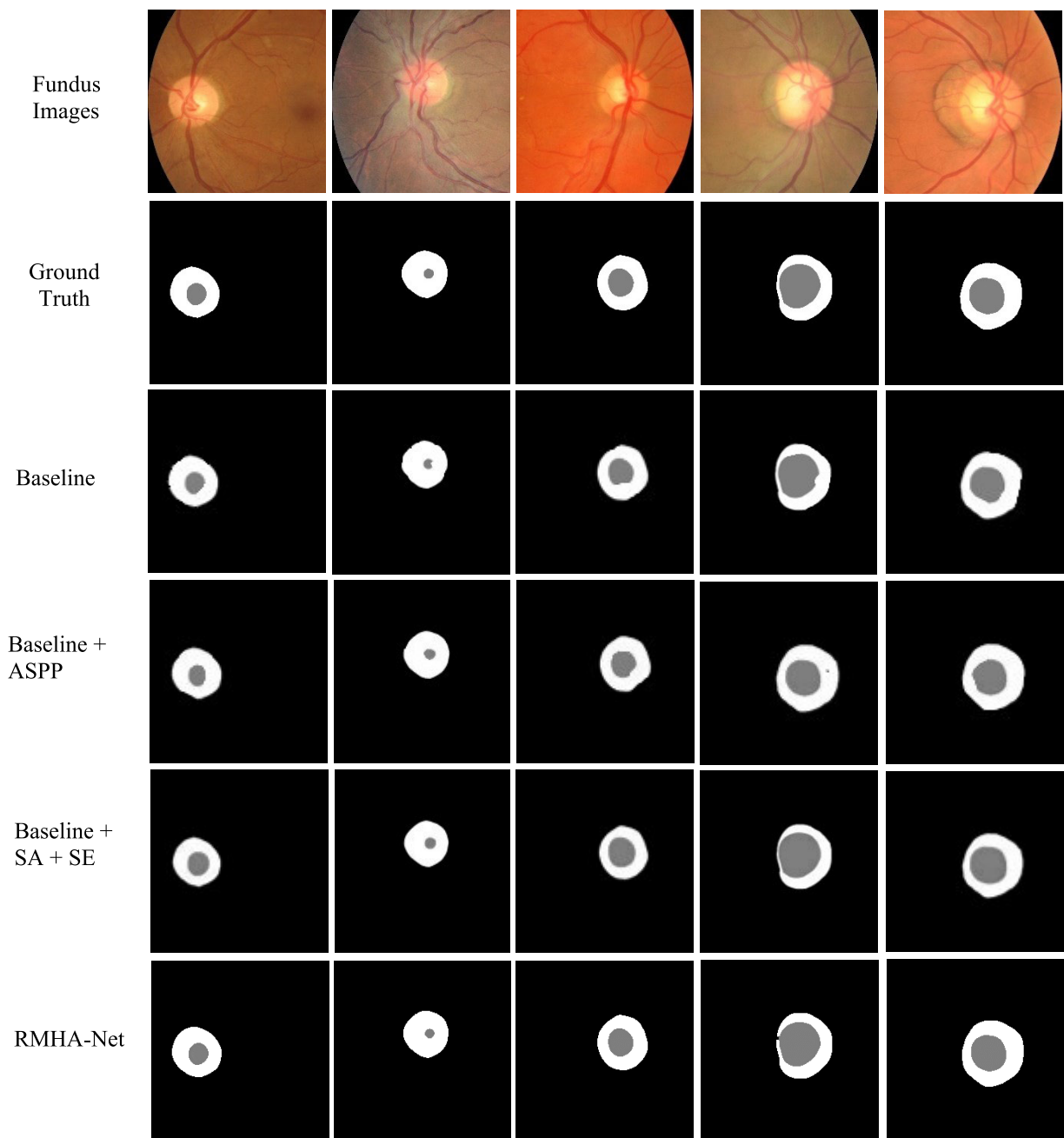
mode. Accordingly, the network performance was tested by combining ASPP modules with the SE module, SA module, and MA module respectively. The combination of the ASPP module with the SA module outperforms the other combinations' performances. This combination managed to improve OD mIoU and DSC by 1.34% and 1.29%, respectively, whereas OC mIoU and DSC have also increased by 2.22% and 0.75%, respectively. This improvement can be attributed to the combined performance of the ASPP's multi-scale capacity and the focused regions of the SA module.

The third-round study consists of concurrent integration of the best three modules into the baseline model, which comprised the integration of the ASPP modules, the SA module, and the SE module. In other words, this optimal combination is the proposed RMHA-Net, the most efficient architecture resulted in the highest performance, with an OD mIoU improvement of 1.46% and DSC of 1.32%, while the OC mIoU improved by 2.48% and DSC by 1.78%. These findings support the synergy of multiscale extraction (ASPP), channel prioritization (SE), and spatial attention (SA) in enhancing segmentation performance.

As is obvious, the integration of each unit has a positive and active impact on the initial baseline. To ensure neutrality, all of these experiments were implemented using the same hyperparameter settings. We believe that this ablation study provides a deeper understanding of the contributions of individual modules and their combined impact, emphasizing their significance in achieving the observed segmentation performance. Table 2 illustrates the performance outputs of the ablation study after integrating the submodules.

B. ABLATION ANALYSIS: LOSS FUNCTION

Further ablation studies were conducted to understand the behavior of loss functions that are best suited for the proposed RMHA-Net. Three different loss functions, which are cross-entropy, Dice, and a hybrid combination of cross-entropy and Dice were evaluated using the Ibn Al-Haitham and REFUGE

TABLE 2. Samples of the segmented maps through the addition of the sub modules on the proposed baseline.

datasets. The quantitative segmentation results from these experiments are presented in Table 3 and Table 4.

According to the findings reported in Tables 3 and 4, the usage of CE Loss produced the highest mIoU of 0.9437, which was 0.96% and 0.97% higher than Dice Loss and combined loss, respectively when tested for the OD segmentation on the Ibn Al-Haitham dataset. Similarly, the usage of CE Loss produced a mIoU of 0.9369 on the REFUGE dataset, outperforming both Dice Loss and the combined loss by 10.63% and 11.05%, respectively. These findings show the effectiveness of CE Loss in handling the comparatively larger

and more distinct OD structure, where accurate segmentation is essential for further analysis. In the case of OC segmentation, which is more difficult because of its smaller size and detailed features, the usage of CE Loss proved to be a more appropriate choice.

When a private dataset is used for validation, the performance of CE Loss outperformed Dice Loss and the combined loss by 4.78% and 2.25%, respectively, with a mIoU equal to 0.8471.

In a comparable manner on the REFUGE dataset, CE scored a mIoU of 0.8857, outperforming Dice Loss and

TABLE 3. Loss functions performance comparison over the Ibn Al-Haitham dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
Dice and Cross-entropy	99.76	0.9346	0.9313	0.8991	0.9659	14.04	99.82	0.8284	0.7940	0.7162	0.8908	25.85
Dice Loss	99.78	0.9381	0.9352	0.9151	0.9561	13.88	99.85	0.8085	0.7646	0.7024	0.8388	29.98
Cross-entropy	99.82	0.9437	0.9383	0.9210	0.9463	12.01	99.84	0.8471	0.8273	0.8117	0.8457	21.16

TABLE 4. Loss functions performance comparison over the Refuge dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
Dice and Cross-entropy	99.44	0.8435	0.8154	0.7303	0.9230	22.17	99.23	0.8447	0.8333	0.8047	0.8174	24.85
Dice Loss	99.59	0.8469	0.8230	0.7136	0.9721	21.81	99.85	0.7533	0.7337	0.7523	0.892	29.17
Cross-entropy	99.72	0.9369	0.9305	0.9071	0.9398	14.20	99.58	0.8857	0.8721	0.8801	0.8990	21.19

the combined loss by 17.58% and 4.85%, respectively. These findings illustrate CE Loss capacity to effectively address class imbalance while maintaining outstanding segmentation accuracy, especially for the complex OC region.

C. BENCHMARKING COMPARISON

To further validate the performance of the RMHA-Net, its performance is compared to seven state-of-the-art semantic segmentation networks. The networks were chosen for comparison purposes due to their excellent segmentation performance in various applications. The chosen networks for the benchmarking purposes UNET [52], TerausNet [53], SegNet [54], DeepLab V2 [55], PSPNet [56], DABNET [57], and FCDenseNet [58].

The proposed RMHA-Net, along with the seven benchmark models, was thoroughly evaluated using five public datasets: Drishti-GS, ORIGA, PAPILA, Chaksu, and REFUGE, and one private dataset, Ibn Al-Haitham. This wide collection of datasets provides a comprehensive framework

that enables an in-depth analysis of the proposed model's segmentation performance over a range of clinical imaging settings and dataset attributes. The reported results in Tables 5–10 show the RMHA-Net's robustness and adaptability in addressing the complexities of optic disc and optic cup segmentation, which ensures consistent performance across several datasets of various clinical situations.

The results listed in each table are further analyzed, with particular emphasis on RMHA-Net performance on various datasets. The results indicate notable gains over baseline models and highlight the proposed model's capacity to handle the difficulties of OD and OC segmentation under various imaging settings and dataset characteristics. The superior accuracy of RMHA-Net is demonstrated in the Drishti-GS dataset findings as summarized in Table 5. RMHA-Net outperforms popular models like UNET and DeepLabV2 by achieving the highest mIoU scores of 0.9723 for OD and 0.8806 for OC. Additionally, it also produces the lowest HDD values, demonstrating its capacity to precisely

TABLE 5. Segmentation performance comparison over the Drishti GS dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TerausNet	99.77	0.9530	0.9522	0.9432	0.9614	14.13	98.73	0.8470	0.8270	0.8479	0.8161	22.24
UNET	99.78	0.9489	0.9527	0.9408	0.9648	15.07	98.69	0.8380	0.8377	0.8302	0.8421	24.21
SegNet	99.76	0.9571	0.9565	0.9429	0.9705	15.46	98.56	0.8266	0.8431	0.8022	0.8801	22.24
DeepLabV2	99.66	0.9411	0.9393	0.9325	0.9463	13.48	98.89	0.8339	0.8319	0.8572	0.8265	25.85
PSPNet	98.54	0.9273	0.9016	0.8723	0.8925	15.74	98.56	0.8138	0.8183	0.8736	0.8703	26.58
DABNET	99.58	0.9307	0.9280	0.9219	0.9341	12.57	98.79	0.8546	0.8233	0.8687	0.8980	25.45
FCDenseNet	99.76	0.9567	0.9561	0.9610	0.9512	13.21	98.86	0.8483	0.8591	0.8168	0.8602	23.85
Proposed (Baseline)	99.83	0.9676	0.9665	0.9706	0.9516	11.65	98.86	0.8716	0.8586	0.8887	0.8701	22.98
(RMHA-Net)	99.89	0.9723	0.9693	0.9786	0.9679	11.20	98.69	0.8806	0.8594	0.8949	0.8786	22.16

TABLE 6. Segmentation performance comparison over the Origa dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TernausNet	99.74	0.9323	0.9290	0.9062	0.9529	14.10	99.69	0.8227	0.7866	0.7582	0.8173	21.65
UNET	99.78	0.9347	0.9326	0.9169	0.9699	15.58	99.78	0.8564	0.8472	0.7966	0.9047	26.45
SegNet	99.66	0.9279	0.9240	0.9012	0.9479	14.76	99.72	0.8428	0.8153	0.7748	0.8602	24.08
DeepLabV2	99.69	0.9221	0.9174	0.9037	0.9314	14.28	99.69	0.8568	0.7891	0.7248	0.8706	22.43
PSPNet	98.17	0.8860	0.8602	0.8438	0.8777	20.78	99.37	0.8167	0.7391	0.8996	0.8199	25.48
DABNET	99.64	0.9152	0.9094	0.8734	0.9486	16.56	99.73	0.8254	0.7904	0.7921	0.7887	25.42
FCDenseNet	99.68	0.9251	0.9208	0.9044	0.9378	16.28	99.71	0.8296	0.7966	0.7897	0.8036	24.28
Proposed (Baseline)	99.71	0.9408	0.9318	0.9586	0.9513	13.28	99.76	0.8533	0.8159	0.8401	0.7931	20.34
(RMHA-Net)	99.83	0.9528	0.9515	0.9531	0.9499	12.19	99.70	0.8675	0.8787	0.8703	0.7045	19.14

TABLE 7. Segmentation performance comparison over the Papila dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TernausNet	99.34	0.9125	0.9115	0.8578	0.9471	11.10	99.61	0.7849	0.7048	0.6635	0.7486	24.66
UNET	99.39	0.9117	0.9100	0.8517	0.9840	10.58	99.67	0.7969	0.7473	0.6681	0.8477	26.48
SegNet	99.17	0.9133	0.9097	0.8476	0.9815	11.76	99.66	0.7834	0.7259	0.6984	0.7556	22.04
DeepLabV2	99.27	0.9156	0.9120	0.8770	0.9500	10.28	99.48	0.7901	0.7402	0.6897	0.7854	25.83
PSPNet	98.13	0.8771	0.8653	0.8762	0.8623	15.73	98.55	0.7211	0.7022	0.6405	0.6890	23.15
DABNET	99.18	0.9126	0.9087	0.8703	0.9505	13.56	99.66	0.7793	0.7191	0.6880	0.7532	23.17
FCDenseNet	99.11	0.8944	0.8880	0.8123	0.9792	14.26	99.66	0.7865	0.7309	0.6552	0.8265	25.86
Proposed (Baseline)	99.45	0.9240	0.9225	0.9212	0.9430	9.21	99.62	0.8114	0.8018	0.7099	0.8740	21.34
(RMHA-Net)	99.55	0.9369	0.9356	0.9229	0.9497	8.10	99.70	0.8241	0.8129	0.7271	0.8877	20.16

TABLE 8. Segmentation performance comparison over the Chákşu dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TernausNet	99.61	0.9272	0.9141	0.8983	0.9514	13.12	99.70	0.8438	0.8170	0.8179	0.8160	26.66
UNET	99.71	0.9195	0.9151	0.9143	0.9220	14.01	99.73	0.8560	0.8470	0.8192	0.8427	30.18
SegNet	99.56	0.9218	0.9189	0.8731	0.9480	13.76	99.69	0.8561	0.8338	0.7922	0.8800	27.24
DeepLabV2	99.57	0.9204	0.9164	0.9141	0.9187	14.28	99.20	0.8565	0.8309	0.7552	0.8265	24.85
PSPNet	98.31	0.8804	0.6625	0.9982	0.8958	15.78	99.52	0.7771	0.7288	0.8936	0.7703	25.18
DABNET	99.52	0.9138	0.9089	0.8758	0.9445	13.57	99.68	0.8341	0.8033	0.8087	0.7980	25.15
FCDenseNet	99.46	0.8970	0.8894	0.8481	0.9348	14.28	99.62	0.8239	0.7891	0.7148	0.8806	26.85
Proposed (Baseline)	99.56	0.9307	0.9258	0.9098	0.9425	11.21	99.72	0.8622	0.8286	0.7989	0.8605	22.38
(RMHA-Net)	99.50	0.9459	0.9335	0.9228	0.9545	11.10	99.78	0.8697	0.8314	0.8244	0.8783	21.56

define the complex structural borders necessary for a reliable glaucoma diagnosis. Additionally, analysis on the ORIGA dataset, which is reported in Table 6, proved the RMHA-Net's adaptability to various situations. It produces mIoU metrics of 0.9528 and 0.8675, respectively, with DSC scores of 0.9515 for OD and 0.8787 for OC. Furthermore, RMHA-Net also reports lower HDD values, indicating its significance for clinical applications and its accurate border precision. Using a comparable set of configurations, the PAPILA dataset, which is recognized for its wide range of imaging conditions

and anatomical complexity, is also used to demonstrate RMHA-Net's flexibility and precision in clinical segmentation tasks, as reported in Table 7. The proposed model outperforms benchmarked models such as UNET and SegNet, with an OD DSC of 0.9356 and an OC DSC of 0.8129, respectively.

RMHA-Net also achieves lower HDD values compared to more sophisticated models such as FCDenseNet and DABNet. These findings show that RMHA-Net can offer accurate segmentation and precise border delineation.

TABLE 9. Segmentation performance comparison over the Refuge dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TernausNet	99.64	0.9062	0.8916	0.8071	0.8477	16.11	99.86	0.8659	0.8547	0.7606	0.8542	26.21
UNET	99.64	0.9040	0.8935	0.8164	0.9268	15.58	99.89	0.8721	0.8674	0.7521	0.8194	25.21
SegNet	99.54	0.8810	0.8784	0.7962	0.9308	16.76	99.82	0.8580	0.8484	0.7675	0.8061	23.24
DeepLabV2	99.61	0.8925	0.8849	0.8072	0.9419	14.28	99.65	0.8466	0.8431	0.8022	0.8801	25.33
PSPNet	97.67	0.9088	0.8918	0.8204	0.9346	15.70	99.12	0.8484	0.8304	0.8021	0.8887	27.58
DABNET	99.55	0.8798	0.8701	0.8156	0.9259	16.50	99.79	0.8347	0.8272	0.7871	0.8074	26.40
FCDenseNet	99.44	0.8882	0.8704	0.8196	0.8216	18.26	99.23	0.8447	0.8333	0.8047	0.8174	24.85
Proposed (Baseline)	99.69	0.9275	0.9187	0.8829	0.9255	14.81	99.85	0.8759	0.8701	0.8523	0.882	22.18
(RMHA-Net)	99.72	0.9369	0.9305	0.9071	0.9398	14.20	99.58	0.8857	0.8721	0.8801	0.8990	21.19

TABLE 10. Segmentation performance comparison over the Ibn Al-Haitham dataset.

Model	OD Segmentation Results						OC Segmentation Results					
	ACC (%)	mIoU	DSC	SPE	SEN	HDD	ACC (%)	mIoU	DSC	SPE	SEN	HDD
TernausNet	99.69	0.8951	0.8991	0.8411	0.9890	15.10	99.81	0.8203	0.7823	0.6769	0.9266	24.24
UNET	99.68	0.9014	0.8906	0.8414	0.8922	14.56	99.82	0.8290	0.7949	0.6895	0.9183	22.58
SegNet	99.61	0.9003	0.8916	0.8126	0.8875	15.70	99.81	0.8191	0.7805	0.6683	0.9181	21.22
DeepLabV2	99.69	0.9010	0.9021	0.8350	0.9712	15.69	99.841	0.8210	0.7842	0.7421	0.8321	28.15
PSPNet	97.47	0.8989	0.8355	0.8664	0.8943	18.78	97.47	0.7992	0.7945	0.7569	0.8988	30.88
DABNET	99.59	0.8896	0.8785	0.8232	0.8918	17.50	99.80	0.7991	0.7802	0.7585	0.8239	27.42
FCDenseNet	99.66	0.9095	0.9025	0.8434	0.9706	14.28	99.76	0.7974	0.7477	0.7266	0.9268	25.85
Proposed (Baseline)	99.77	0.9291	0.9251	0.9042	0.9503	13.08	99.85	0.8223	0.8095	0.8115	0.8478	22.96
(RMHA-Net)	99.82	0.9437	0.9383	0.9210	0.9463	12.01	99.84	0.8471	0.8273	0.8117	0.8457	21.16

Building upon the previous comparison, RMHA-Net has also delivered outstanding segmentation results on the Chaksu dataset. As illustrated in Table 8, RMHA-Net reports mIoU values of 0.9459 for OD and 0.8697 for OC, along with DSC scores of 0.9335 for OD and 0.8314 for OC. These excellent performance values emphasize the model's precision in distinguishing optic disc and cup borders that are used to compute the cup-to-disc ratio (CDR), which serves as an essential indicator in glaucoma detection. An additional experiment is further conducted using the REFUGE dataset, whereby the findings are reported in Table 9. This dataset poses a challenging segmentation difficulty because of its varied lighting conditions and structural variations. RMHA-Net obtains notable mIoU scores of 0.8857 for OC and 0.9369 for OD. These outcomes are better than those of multi-scale models including PSPNet, which offers mIoU scores of 0.9088 for OD and 0.8484 for OC, and DeepLabV2, which offers mIoU scores of 0.8925 for OD and 0.8466 for OC.

Finally, RMHA-Net was evaluated on a private dataset to determine its capability to adapt to a variety of clinical settings. The results obtained from the Ibn Al-Haitham dataset, listed in Table 10, demonstrate the model's performance for medical applications. RMHA-Net achieved a mIoU of 94.37% for OD and 84.71% for OC, outperforming SegNet (90.03% for OD and 81.91% for OC) and

TernausNet (89.51% for OD and 82.03% for OC). The findings support the claim of RMHA-Net adaptability and dependability for clinical applications, especially under the tough and variable conditions offered by privately gathered datasets. In conclusion, the comprehensive evaluation of six datasets demonstrates RMHA-Net's effectiveness and adaptability compared to the benchmarked models. Its architecture, which incorporates dual ASPP modules and attention mechanisms, enables accurate multiscale feature extraction while maintaining computing efficiency. These features enable RMHA-Net to accurately segment the optic disc and cup, which is crucial for reliable glaucoma diagnosis. Moreover, the model consistently outperforms other deep models under a variety of imaging scenarios.

D. FEATURE EXTRACTION AND COMPUTATIONAL EFFICIENCY

To compare the proposed model's performance with the benchmarked segmentation models that also utilize similar multiscale techniques, extensive experiments were performed, focusing on the feature extraction and computational efficiency aspects. In terms of feature extraction, RMHA-Net is compared to DeepLabv3+ since both models provide superior performance using multiscale context aggregation. In comparison, DeepLabv3+ provides superior segmentation

results based on a single ASPP module which employs fixed filter sizes (256) and dilation rates (6, 12, 18, 24). However, it lacks the adaptive recalibration and spatial refinement of RMHA-Net, which makes the model's size larger and more computationally demanding.

RMHA-Net, on the other hand, uses a hierarchical technique to integrate the Residual Atrous Blocks alongside an optimized set of dilation rates of (1, 3, 4, and 5) and bigger filter size configurations (64 to 512). It also combines channel and spatial attention strategies to regularly recalibrate features and improve relevant spatial region feature extraction. In addition, the dual-ASPP strategy with distinct dilation rates captures fine details and broader context more effectively, enabling the proposed RMHA-Net to achieve superior multiscale representation with greater computational efficiency.

With regards to the computational performance of the multiscale module of the RMHA-Net compared to the DeepLabv3, comprehensive experiments were carried out using both the Ibn Al-Haitham and REFREGE datasets as listed in Table 11.

TABLE 11. Computational efficiency.

Model	FLOPs (M)	Ibn Al-Haitham		REFREGE
		Model Size (MB)	Inference Time (ms)	Inference Time (ms)
DeepLabV3	89.256	90.02	78.79	77.68
DeepLabV3+	92.235	94.36	80.94	79.72
(Baseline)	80,386	79.76	74.65	73.68
(RMHA-Net)	87,998	85.01	76.73	75.72

Based on the results, RMHA-Net outperforms other multiscale segmentation approaches in terms of computational efficiency and effectiveness. It requires 4.6% fewer FLOPs compared to DeepLabV3+ and 1.4% fewer FLOPs compared to basic DeepLabV3, which can be attributed to its optimized multiscale module architecture. Furthermore, the model size of the RMHA-Net is 9.9% less compared to DeepLabV3+ and 5.6% smaller than DeepLabV3, indicating better parameters utilization, which consequently produced a lighter model with fewer compute-intensive processes. Based on the Ibn Al-Haitham and REFREGE datasets, RMHA-Net reaches inference times of 5.2% faster than DeepLabV3+ and 2.6% faster than DeepLabV3, indicating its suitability for real-time applications. Despite the proposed model having the lowest computational requirement, with 13.3% fewer FLOPs and a 7.8% faster inference time than DeepLabV3+, RMHA-Net produced the best compromise between the model's efficiency and performance. This optimal tradeoff is reflected through the improved OD and OC segmentation performance for glaucoma detection, as well as increased precision in Cup-to-Disc Ratio computation for computer-aided glaucoma diagnosis.

VI. DISCUSSION

The experimental results show that the proposed RMHA-Net, which is based on a U-shaped architecture has outperformed all state-of-the-art architectures for optic cup and disc segmentation when tested on the private and public datasets.

The architecture utilizes the dilated convolutions and residual connections to collect more information from a wider context. The dilated convolutions can significantly expand the receptive field to include a broader spatial context while preserving the parameters' size and the respective feature map dimensions. This strategy is particularly critical for identifying OD and OC regions, which are heavily dependent on the contextual details of neighboring regions of various scales. In addition, the utilization of the residual connections also allows the network to capture early layer information directly, especially the lower-level features like edges and textures, which are critical for accurate boundary alienation, boosting the model's capacity to capture critical feature information.

This architecture was also further improved by incorporating two types of attention mechanisms, in addition to the multiscale feature extraction capabilities, resulting in smooth, uniform, and rounded segmentation boundaries for both OD and OC regions. The attention mechanisms are added to emphasize more on the strategic regions, especially regions with unique edges, contrasts, and colors for the OD and OC.

This step is critical for an effective segmentation model because the respective regions might vary significantly in proportions, shapes, and locations across different patients. In addition, two sets of four distinctive dilation rates are employed within the ASPP modules, so that the field of view can be adjusted across various scales to match the sizes and shapes of OD and OC, which will improve the network's capacity to acquire both detailed and contextual information effectively.

During the development of the RMHA-Net, extensive systematic experiments were conducted over three rounds, covering all possible submodule placement locations as well as the optimal hyperparameter values used in the network to deliver the best segmentation model. The values of the spatial attention's kernel size and the channel attention's division ratio were properly explored and selected to ensure the features are focusing on the informative spots such as edges or textures while putting less consideration on the less relevant areas, which happened to help in minimizing the training time.

The same extensive experimental procedures were also implemented to determine the rates and depths of the ASPP modules to best fit the OD and OC conditions. These experiments were conducted under standardized settings, utilizing the Grid search approach to find the optimal values for the batch size, learning rate, and loss function. The results were also evaluated using several complementary metrics that have been widely used for the semantic segmentation task.

In summary, the three rounds of the development process of the RMHA-Net are crucial to ensure the extracted features fit the goal of segmenting OD and OC. The first round of experiments focused on adding a single submodule to the baseline network, and the results showed an improvement, whereas the second round of experiments focused on adding two different types of submodules, which produced better performance compared to single submodule addition in the first-round experiments. However, the best results were obtained during the third round of experiments, which comprised the addition of three distinct types of submodules of the SE, ASPP, and SA modules, that has been proven to outperform the baseline model with a mIoU improvement of 1.46% and DSC of 1.32%, while the OC mIoU improved by 2.48% and DSC by 1.78%. both were tested using the Ibn Al-Haitham dataset. These ablation experiments need to be conducted carefully as the addition of any submodule may bring adverse effects on the other components, which calls for optimal balancing between the benefits of adding extra submodules and the available computational resources.

Generally, increasing the computational complexity of the network could lead to overfitting issues, particularly when the dataset size is homogenous to certain demographic conditions, as in our case (Iraqi patients). Despite the better segmentation maps, these results can be further improved by employing some operations like image cropping to address the problem of imbalance dataset distribution in the fundus image, whereby the regions of interest are relatively small in size compared to background areas. Another improvement that can be explored is by applying post-processing processes to fine-tune the segmentation maps by removing small holes and artefacts as well as smoothing the edges of the segmented OD and OC regions.

VII. CONCLUSION

This work introduced a developed encoder-decoder framework called RMHA-Net for automated optic disc and cup segmentation used in automated glaucoma diagnosis. The RMHA-Net consists of four main components, which are the residual atrous convolutional modules, spatial attention modules, ASPP modules and channel attention modules. The input tensor is encoded through a series of convolutional blocks with a set of dilated kernels, whereby the informative regions within the feature maps are emphasized using the spatial attention modules. Consequently, the resulting feature maps are passed through to two levels of multiscale feature extractor submodules that utilize a set of unique dilation rates to effectively deal with the complex feature extraction as the network becomes deeper. After that, the rich latent features are decoded through a series of convolutional operations, which are prioritized using the SE module to produce pixel-wise segmentation labeling. The RMHA-Net has been validated using five public datasets, and one private dataset, in which the segmentation results have returned better performance compared to the benchmarked state-of-the-art models. On the Ibn Al-Haitham dataset, RMHA-Net

achieved the best segmentation results, recording the highest mIoU (94.37%) and DSC (93.83%) for OD segmentation. These performance results exceeded those of UNet, which achieved an mIoU of 90.14% and a DSC of 89.06%, as well as DeepLabV2, with an mIoU of 90.10% and a DSC of 90.21%. For OC segmentation, RMHA-Net continued to outperform other models, achieving the highest mIoU (84.71%) and DSC (82.73%), compared to SegNet (mIoU: 81.91%, DSC: 78.05%) and FCDenseNet (mIoU: 79.74%, DSC: 74.77%). Similar performance patterns are observed when RMHA-Net is tested on the other five datasets, confirming RMHA-Net's capacity to deliver precise delineation for both segmentation tasks. The model also demonstrates superior computational efficiency, with notable improvements in FLOPs, inference time, and model size that ensure RMHA-Net suitability for real-time applications. Compared to DeepLabV3+, RMHA-Net reduces FLOPs by approximately 4.6%, model size by 9.9%, and inference time by 5.2% when tested on the Ibn Al-Haitham dataset.

In summary, the proposed framework shows its highly competitive capability in OD and OC segmentation through the effective application of dilated convolutions and attention modules, which produces a strong foundation for future research on this topic. In the future, we plan to validate the RMHA-Net using different medical datasets to analyze the model's generalization capability. Besides that, a further tweak can be done by applying a generative adversarial network to produce synthetic augmented data, especially for challenging cases.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

- [1] A. Haider, M. Arsalan, M. B. Lee, M. Owais, T. Mahmood, H. Sultan, and K. R. Park, "Artificial intelligence-based computer-aided diagnosis of glaucoma using retinal fundus images," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117968, doi: [10.1016/j.eswa.2022.117968](https://doi.org/10.1016/j.eswa.2022.117968).
- [2] M. Zedan, M. Zulkifley, A. Ibrahim, A. Moubark, N. Kamari, and S. Abdani, "Automated glaucoma screening and diagnosis based on retinal fundus images using deep learning approaches: A comprehensive review," *Diagnostics*, vol. 13, no. 13, p. 2180, Jun. 2023, doi: [10.3390/diagnostics13132180](https://doi.org/10.3390/diagnostics13132180).
- [3] H. Xiong, F. Long, M. S. Alam, and J. Sang, "Multi-GlaucNet: A multi-task model for optic disc segmentation, blood vessel segmentation and glaucoma detection," *Biomed. Signal Process. Control*, vol. 99, Jan. 2025, Art. no. 106850, doi: [10.1016/j.bspc.2024.106850](https://doi.org/10.1016/j.bspc.2024.106850).
- [4] A. Q. Saeed, S. N. H. Sheikh Abdullah, J. Che-Hamzah, and A. T. Abdul Ghani, "Accuracy of using generative adversarial networks for glaucoma detection: Systematic review and bibliometric analysis," *J. Med. Internet Res.*, vol. 23, no. 9, Sep. 2021, Art. no. e27414, doi: [10.2196/27414](https://doi.org/10.2196/27414).
- [5] B. Goutam, M. F. Hashmi, Z. W. Geem, and N. D. Bokde, "A comprehensive review of deep learning strategies in retinal disease diagnosis using fundus images," *IEEE Access*, vol. 10, pp. 57796–57823, 2022, doi: [10.1109/ACCESS.2022.3178372](https://doi.org/10.1109/ACCESS.2022.3178372).

- [6] N. A. Mohamed, M. A. Zulkifley, W. M. D. W. Zaki, and A. Hussain, "An automated glaucoma screening system using cup-to-disc ratio via simple linear iterative clustering superpixel approach," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101454, doi: [10.1016/j.bspc.2019.01.003](https://doi.org/10.1016/j.bspc.2019.01.003).
- [7] T. M. Khan, M. M. Mehmood, S. S. Naqvi, and M. F. U. Butt, "A region growing and local adaptive thresholding-based optic disc detection," *PLoS ONE*, vol. 15, no. 1, Jan. 2020, Art. no. e0227566, doi: [10.1371/journal.pone.0227566](https://doi.org/10.1371/journal.pone.0227566).
- [8] E. S. Kumar and C. S. Bindu, "Two-stage framework for optic disc segmentation and estimation of cup-to-disc ratio using deep learning technique," *J. Ambient Intell. Humanized Comput.*, vol. 2, pp. 1–26, Mar. 2021, doi: [10.1007/s12652-021-02977-5](https://doi.org/10.1007/s12652-021-02977-5).
- [9] D. Parashar and D. Agrawal, "2-D compact variational mode Decomposition- based automatic classification of glaucoma stages from fundus images," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021, doi: [10.1109/TIM.2021.3071223](https://doi.org/10.1109/TIM.2021.3071223).
- [10] M. T. Islam, S. T. Mashfu, A. Faisal, S. C. Siam, I. T. Naheen, and R. Khan, "Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation," *IEEE Access*, vol. 10, pp. 2828–2841, 2022, doi: [10.1109/ACCESS.2021.3139160](https://doi.org/10.1109/ACCESS.2021.3139160).
- [11] S. Pathan, P. Kumar, R. M. Pai, and S. V. Bhandary, "An automated classification framework for glaucoma detection in fundus images using ensemble of dynamic selection methods," *Prog. Artif. Intell.*, vol. 12, no. 3, pp. 287–301, Sep. 2023, doi: [10.1007/s13748-023-00304-x](https://doi.org/10.1007/s13748-023-00304-x).
- [12] S. Iqbal, T. M. Khan, K. Naveed, S. S. Naqvi, and S. J. Nawaz, "Recent trends and advances in fundus image analysis: A review," *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106277, doi: [10.1016/j.combiomed.2022.106277](https://doi.org/10.1016/j.combiomed.2022.106277).
- [13] W. S. Shalaby, O. M. Ahmed, M. Waisbourd, and L. J. Katz, "A review of potential novel glaucoma therapeutic options independent of intraocular pressure," *Surv. Ophthalmology*, vol. 67, no. 4, pp. 1062–1080, Jul. 2022, doi: [10.1016/j.survophthal.2021.12.003](https://doi.org/10.1016/j.survophthal.2021.12.003).
- [14] B. Toptas, M. Toptas, and D. Hanbay, "Detection of optic disc localization from retinal fundus image using optimized color space," *J. Digit. Imag.*, vol. 35, no. 2, pp. 302–319, Apr. 2022, doi: [10.1007/s10278-021-00566-8](https://doi.org/10.1007/s10278-021-00566-8).
- [15] W. Zhou, Y. Yi, Y. Gao, and J. Dai, "Optic disc and cup segmentation in retinal images for glaucoma diagnosis by locally statistical active contour model with structure prior," *Comput. Math. Methods Med.*, vol. 2019, pp. 1–16, Nov. 2019, doi: [10.1155/2019/8973287](https://doi.org/10.1155/2019/8973287).
- [16] S. Virbukaite, J. Bernatavičienė, and D. Imbrasienė, "Glaucoma identification using convolutional neural networks ensemble for optic disc and cup segmentation," *IEEE Access*, vol. 12, pp. 82720–82729, 2024, doi: [10.1109/ACCESS.2024.3412185](https://doi.org/10.1109/ACCESS.2024.3412185).
- [17] Y. Meng, H. Zhang, Y. Zhao, X. Yang, Y. Qiao, I. J. C. MacCormick, X. Huang, and Y. Zheng, "Graph-based region and boundary aggregation for biomedical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 690–701, Mar. 2022, doi: [10.1109/TMI.2021.3123567](https://doi.org/10.1109/TMI.2021.3123567).
- [18] P. B. Singh, P. Singh, and H. Dev, "Optimized convolutional neural network for glaucoma detection with improved optic-cup segmentation," *Adv. Eng. Softw.*, vol. 175, Jan. 2023, Art. no. 103328, doi: [10.1016/j.advengsoft.2022.103328](https://doi.org/10.1016/j.advengsoft.2022.103328).
- [19] Z. Shi, T. Wang, Z. Huang, F. Xie, Z. Liu, B. Wang, and J. Xu, "MD-net: A multi-scale dense network for retinal vessel segmentation," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 102977, doi: [10.1016/j.bspc.2021.102977](https://doi.org/10.1016/j.bspc.2021.102977).
- [20] E. Elizar, M. A. Zulkifley, R. Muharar, M. H. M. Zaman, and S. M. Mustaza, "A review on multiscale-deep-learning applications," *Sensors*, vol. 22, no. 19, p. 7384, Sep. 2022.
- [21] M. M. Stofa, M. A. Zulkifley, M. A. A. M. Zainuri, and A. A. Ibrahim, "U-net with atrous spatial pyramid pooling for skin lesion segmentation," in *Lecture Notes in Electrical Engineering*. Berlin, Germany: Springer-Verlag, 2022.
- [22] X. Jia, T. Shen, and Y. He, "Improving the semantic segmentation algorithm of DeepLabv3+," in *Proc. IEEE 6th Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, vol. 6, Feb. 2023, pp. 1730–1734, doi: [10.1109/ITNEC56291.2023.10082715](https://doi.org/10.1109/ITNEC56291.2023.10082715).
- [23] X. Song, X. Fang, X. Meng, X. Fang, M. Lv, and Y. Zhuo, "Real-time semantic segmentation network with an enhanced backbone based on atrous spatial pyramid pooling module," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 107988, doi: [10.1016/j.engappai.2024.107988](https://doi.org/10.1016/j.engappai.2024.107988).
- [24] S. V. Lim, M. A. Zulkifley, A. Saleh, A. H. Saputro, and S. R. Abdani, "Attention-based semantic segmentation networks for forest applications," *Forests*, vol. 14, no. 12, p. 2437, Dec. 2023, doi: [10.3390/f14122437](https://doi.org/10.3390/f14122437).
- [25] C. Meas, W. Guo, and M. H. Miah, "Multi-scale attention U-Net for optic disc and optic cup segmentation in retinal fundus images," in *Proc. 2nd Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, May 2024, pp. 760–765.
- [26] J. Zhao, X. Hou, M. Pan, and H. Zhang, "Attention-based generative adversarial network in medical imaging: A narrative review," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105948, doi: [10.1016/j.combiomed.2022.105948](https://doi.org/10.1016/j.combiomed.2022.105948).
- [27] I. A. Butt, L. Bo, M. H. F. Butt, M. T. Tassew, and M. U. Bin Akhtar Khan, "Enhancing hyperspectral image classification: A hybrid CNN approach with spatial and channel attention mechanisms," in *Proc. 20th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2023, pp. 1–5, doi: [10.1109/iccwamtip60502.2023.10387100](https://doi.org/10.1109/iccwamtip60502.2023.10387100).
- [28] G. Mao, G. Liao, and H. Zhu, "Object detection algorithm based on channel-spatial fusion cascade attention," in *Proc. IEEE 8th Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2022, pp. 210–214, doi: [10.1109/CCIS7298.2022.10016395](https://doi.org/10.1109/CCIS7298.2022.10016395).
- [29] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017, doi: [10.1109/TPAMI.2016.2572683](https://doi.org/10.1109/TPAMI.2016.2572683).
- [30] S. Basu, "Image segmentation by semantic method," *Pattern Recognit.*, vol. 20, no. 5, pp. 497–511, Jan. 1987.
- [31] H. N. Veena, A. Muruganandham, and T. Senthil Kumaran, "A novel optic disc and optic cup segmentation technique to diagnose glaucoma using deep learning convolutional neural network over retinal fundus images," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6187–6198, Sep. 2022, doi: [10.1016/j.jksuci.2021.02.003](https://doi.org/10.1016/j.jksuci.2021.02.003).
- [32] S. Yu, D. Xiao, S. Frost, and Y. Kanagasigam, "Robust optic disc and cup segmentation with deep learning for glaucoma detection," *Computerized Med. Imag. Graph.*, vol. 74, pp. 61–71, Jun. 2019, doi: [10.1016/j.compmedimag.2019.02.005](https://doi.org/10.1016/j.compmedimag.2019.02.005).
- [33] P. Shanmugam, J. Raja, and R. Pitchai, "An automatic recognition of glaucoma in fundus images using deep learning and random forest classifier," *Appl. Soft Comput.*, vol. 109, pp. 1–29, Sep. 2021, doi: [10.1016/j.asoc.2021.107512](https://doi.org/10.1016/j.asoc.2021.107512).
- [34] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018, doi: [10.1109/TMI.2018.2791488](https://doi.org/10.1109/TMI.2018.2791488).
- [35] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2493–2501, Nov. 2018, doi: [10.1109/TMI.2018.2837012](https://doi.org/10.1109/TMI.2018.2837012).
- [36] M. Tabassum, T. M. Khan, M. Arsalan, S. S. Naqvi, M. Ahmed, H. A. Madni, and J. Mirza, "CDED-net: Joint segmentation of optic disc and optic cup for glaucoma screening," *IEEE Access*, vol. 8, pp. 102733–102747, 2020, doi: [10.1109/ACCESS.2020.2998635](https://doi.org/10.1109/ACCESS.2020.2998635).
- [37] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for fast image restoration and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1934–1948, Feb. 2023, doi: [10.1109/TPAMI.2022.3167175](https://doi.org/10.1109/TPAMI.2022.3167175).
- [38] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 1, pp. 121–130, Jan. 2021, doi: [10.1109/JBHI.2020.2986926](https://doi.org/10.1109/JBHI.2020.2986926).
- [39] C. Gao, H. Ye, F. Cao, C. Wen, Q. Zhang, and F. Zhang, "Multiscale fused network with additive channel-spatial attention for image segmentation," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106754, doi: [10.1016/j.knsys.2021.106754](https://doi.org/10.1016/j.knsys.2021.106754).
- [40] Y. Jiang, Z. Ma, C. Wu, Z. Zhang, and W. Yan, "RSAP-net: Joint optic disc and cup segmentation with a residual spatial attention path module and MSRCR-PT pre-processing algorithm," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–11, Dec. 2022, doi: [10.1186/s12859-022-05058-2](https://doi.org/10.1186/s12859-022-05058-2).
- [41] S. Yan, X. Pan, and Y. Wang, "MRSNet: Joint consistent optic disc and cup segmentation based on large kernel residual convolutional attention and self-attention," *Digit. Signal Process.*, vol. 145, Feb. 2024, Art. no. 104308, doi: [10.1016/j.dsp.2023.104308](https://doi.org/10.1016/j.dsp.2023.104308).
- [42] K. Hua, X. Fang, Z. Tang, Y. Cheng, and Z. Yu, "DCAM-NET: A novel domain generalization optic cup and optic disc segmentation pipeline with multi-region and multi-scale convolution attention mechanism," *Comput. Biol. Med.*, vol. 163, Sep. 2023, Art. no. 107076, doi: [10.1016/j.combiomed.2023.107076](https://doi.org/10.1016/j.combiomed.2023.107076).

- [43] A. U. Alam, S. P. Islam, S. M. Mahedy Hasan, A. Y. Srizon, M. F. Faruk, M. A. Mamun, and M. R. Hossain, "Optic disc and cup segmentation via enhanced U-Net with residual and attention mechanisms," in *Proc. 6th Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2024, pp. 329–334, doi: [10.1109/iceeict62016.2024.10534436](https://doi.org/10.1109/iceeict62016.2024.10534436).
- [44] X. Zheng, Y. He, H. Yuan, Y. Jiang, and Y. Xu, "DHA-net: An encoder-decoder network fusing multi-scale features for optic disc segmentation," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2023, pp. 1–6.
- [45] Z. Wei, P. Yuhan, J. Jianhang, Y. Jikun, B. Weiqi, Y. Yugen, and W. Wenle, "RMSDSC-net: A robust multiscale feature extraction with depthwise separable convolution network for optic disc and cup segmentation," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 11482–11505, Dec. 2022, doi: [10.1002/int.23051](https://doi.org/10.1002/int.23051).
- [46] B. J. Bhatkalkar, D. R. Reddy, S. Prabhu, and S. V. Bhandary, "Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields," *IEEE Access*, vol. 8, pp. 29299–29310, 2020, doi: [10.1109/ACCESS.2020.2972318](https://doi.org/10.1109/ACCESS.2020.2972318).
- [47] G. Sun, Z. Zhang, J. Zhang, M. Zhu, X.-R. Zhu, J.-K. Yang, and Y. Li, "Joint optic disc and cup segmentation based on multi-scale feature analysis and attention pyramid architecture for glaucoma screening," *Neural Comput. Appl.*, vol. 35, no. 22, pp. 16129–16142, Aug. 2023, doi: [10.1007/s00521-021-06554-x](https://doi.org/10.1007/s00521-021-06554-x).
- [48] X. Xia, Z. Huang, Z. Huang, L. Shu, and L. Li, "A CNN-transformer hybrid network for joint optic cup and optic disc segmentation in fundus images," in *Proc. Int. Conf. Comput. Eng. Artif. Intell. (ICCEAI)*, Jul. 2022, pp. 482–486, doi: [10.1109/ICCEAI55464.2022.00106](https://doi.org/10.1109/ICCEAI55464.2022.00106).
- [49] M. J. M. Zedan, M. A. Zulkifley, B. A. Saeed, and S. R. Abdani, "Deep learning-based model benchmarking of glaucoma segmentation using a novel ibn al-haitham fundus image dataset," in *Proc. International Conference on Smart Computing, IoT and Machine Learning*, 2024, pp. 244–249, doi: [10.1109/SIML61815.2024.10578202](https://doi.org/10.1109/SIML61815.2024.10578202).
- [50] H. R. Khajeha, M. Fateh, and V. Abolghasemi, "Diagnosis of glaucoma using multi-scale attention block in convolution neural network and data augmentation techniques," *Eng. Rep.*, vol. 6, no. 10, p. 12866, Oct. 2024, doi: [10.1002/eng2.12866](https://doi.org/10.1002/eng2.12866).
- [51] D. Karimi and S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," 2019, *arXiv:1904.10030*.
- [52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Artificial Intelligence and Lecture Notes in Bioinformatics*. Berlin, Germany: Springer-Verlag, 2015, pp. 234–241.
- [53] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation," 2018, *arXiv:1801.05746*.
- [54] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: [10.1109/TPAMI.2016.2644615](https://doi.org/10.1109/TPAMI.2016.2644615).
- [55] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: [10.1109/TPAMI.2017.2699184](https://doi.org/10.1109/TPAMI.2017.2699184).
- [56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239, doi: [10.1109/CVPR.2017.660](https://doi.org/10.1109/CVPR.2017.660).
- [57] G. Li and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," in *Proc. 30th Brit. Mach. Vis. Conf.*, 2020, pp. 1–18.
- [58] S. Jégou, M. Drozdal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1175–1183, doi: [10.1109/CVPRW.2017.156](https://doi.org/10.1109/CVPRW.2017.156).



interests include deep learning for medical applications, embedded systems, MCU, and the IoT applications.



SITI RAIHANAH ABDANI received the bachelor's degree in biomedical sciences and the master's degree from Universiti Kebangsaan Malaysia. Her main research interest includes deep learning applications in biomedical applications.



theory, machine learning, multilabel learning, model selection, and neural architecture search applications. He is an Associate Editor of the journal *Earth, Moon, and Planets*.



Asian Science Diplomat Award, and also a member of Young Scientists Network—Academy Sciences of Malaysia.

...