RetinaLiteNet: A Lightweight Transformer based CNN for Retinal Feature Segmentation

Mehwish Mehmood
Queen's University Belfast
United kingdom
mmehmood01@qub.ac.uk

Majed Alsharari
Queen's University Belfast
United kingdom
malsharari01@qub.ac.uk

Shahzaib Iqbal
Abasyn University Islamabad
Pakistan
shahzeb.iqbal@abasynisb.edu.pk

Ivor Spence
Queen's University Belfast
United kingdom
i.spence@qub.ac.uk

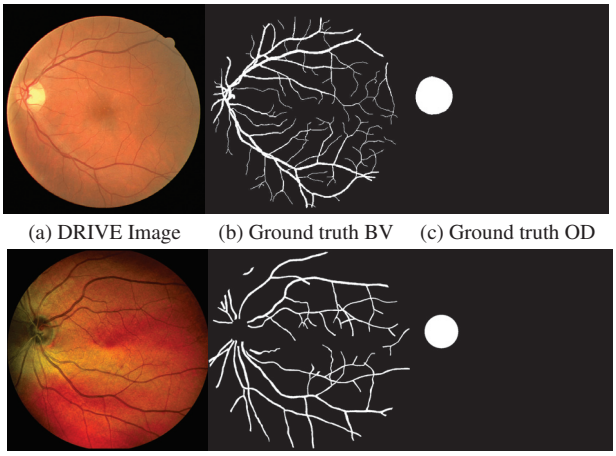
Muhammad Fahim
Queen's University Belfast
United kingdom
m.fahim@qub.ac.uk

Abstract

Retinal image analysis plays a pivotal role in diagnosing diseases like glaucoma, diabetic retinopathy, neurodegenerative disorders, and cardiovascular diseases. The recent advancement of artificial intelligence (AI) can assist the practitioners to analyze the images accurately. In this research, a lightweight deep learning model is proposed which is based on multitask learning to segment the retinal images including retinal vessels and optic disc for further analysis by clinicians. The proposed model has encoder-decoder framework, where the encoder has convolutional layers with multi-head attention that captures both local details and long range dependencies effectively. The resulting features from convolutional layers and multi-head attention are fused together to make the model more efficient and resilient for segmentation tasks. To further refine the features, the skip connections are implemented along with the convolutional block attention module (CBAM) in the decoder. The model's efficiency is validated on two publicly available datasets (i.e., IOSTAR and DRIVE) to confirm the lightweight aspects and robustness. It achieved the F1_scores of 80.6% and 93.3% on DRIVE and 80.1% and 85.4% on IOSTAR dataset for simultaneous segmentation of blood vessels and optic disc, respectively. The empirical evaluations show ***0.25 MB*** *of memory,* ***0.066 million*** *parameters, and a FLOPs estimation of* ***2.46 GFLOPs****, which is better than existing models.*

1. Introduction

The use of computerized tools to analyze retinal images to diagnose several diseases is one of the emerging research fields. Retinal images are routinely utilized to diagnose diseases related to blood vessels and other conditions, such as diabetic retinopathy (DR), neurodegenerative disorders, glaucoma, age-related macular degeneration (AMD), multiple sclerosis (MS) and cardiovascular disease [10, 30].



(a) DRIVE Image (b) Ground truth BV (c) Ground truth OD

(a) IOSTAR Image (b) Ground truth BV (c) Ground truth OD

Figure 1. Sample images from DRIVE and IOSTAR datasets; presents the original images paired with their ground truths of BV and OD.

Diagnosing these diseases without any computer-aided system demands expert knowledge, time, dedication, and monetary resources. Furthermore, the chances of inaccuracy are high. Therefore, efficient segmentation of retinal images is crucial for screening the disease early to ensure deterrence or timely cures [33]. Introducing AI in healthcare has the potential to revolutionize autonomous

disease detection [13]. Our research focuses on retinal image segmentation due to a growing demand for accurate and effective diagnostic tools.

Recent studies highlight the importance of retinal image segmentation for observing signs of neurodegenerative disorders, particularly dementia. Among the various forms of dementia, alzheimer's disease accounts for 60-80% of all cases. Significant changes in the optic nerve and retinal vessels have been observed in subjects with alzheimer's [4, 5, 12].

In this research, we proposed RetinaLiteNet, a novel lightweight multitask learning (MTL) technique. Unlike conventional methods, RetinaLiteNet is designed for simultaneous segmentation of retinal features, focusing on blood vessels (BV) and optic disc (OD) to enhance disease screening efficiency.

We have used two public datasets, IOSTAR and DRIVE, to train and evaluate our model. Figure 1 displays the sample images from two public datasets, DRIVE and IOSTAR, along with their ground truths of BV and OD. The performance of our model is evaluated using performance metrics such as F1(dice)_Score, Jaccard index(IoU), area under the curve (AUC), sensitivity, and specificity. Furthermore, the model's complexity is evaluated using estimation of floating-point operations (FLOPs), trainable parameters count, and the memory utilization. The efficiency of our model is then validated by comparing the results with existing state-of-the-art (SOTA) segmentation models.

The primary contributions of this paper are outlined below:

- We presented a MTL for segmentation of BV and OD from retinal images, facilitating the model to retrieve multiple features simultaneously while improving efficiency.
- We fuse convolutional neural network (CNN) with a multi-head attention block (MHA) [31], a vital transformer component, to detect local and long-range dependencies. This fusion is critical since it ensures that fine-grained and broader patterns are detected, thus improving segmentation.
- We integrate skip connections with the CBAM [34]. Combining these features allows the model to extract them more precisely and ensure that no significant information is overlooked.

The organization of the paper is given below: In section 2, we delve into the background and relevant literature. The proposed methodology is discussed in section 3 and the experiments and results in section 4. Lastly, conclusion is reported in section 5.

2. Related Work

Deep learning (DL) methods have been implemented to segment retinal features like OD and BV. These techniques have outstanding accuracy in retinal image segmentation that suppressed traditional segmentation methods. In the last few years, several SOTA models have been proposed for segmenting retinal vessels and disc separately. Moreover, numerous multi-task learning techniques have been proposed to perform segmentation on various medical images.

Retinal Blood Vessel Segmentation. U-shaped DL networks such as U-Net [27], RefineNet [16], Mask R-CNN [7] and DeconvNet [23] are the popular techniques because of their outstanding achievements in retinal image segmentation. For medical image segmentation, U-Net++ [38] is proposed to use feature fusion and pruning technology to solve the low efficiency of U-Net. Encoder Enhanced Atrous(EEA)-UNet [28] enhanced receptive field by adopting dilated convolution. In this technique, the number of kernels in the encoder and decoder are equal, with layers added in each stage at the encoder side for feature enhancement.

Furthermore Miu et al. [18] proposed Wave-Net for retinal vessel segmentation in which pixel-wise retinal vessel extraction. This model uses detailed enhancement and denoising (DED) block to replace simple skip connection from U-net and multi-scale feature fusion(MFF) to achieve high accuracy. Multi-scale attention-guided fusion network [14] is designed for retinal vessel segmentation. It combines feature enhancement, attention-guided fusion, hybrid feature pooling, and multi-scale attention block. It achieves good accuracy and F1_score at the cost of high inference time. M3U-CDVAE [36] is another lightweight refinement network for retinal vessels segmentation. It uses the first 13 layers of mobilenet-V3 as a network encoder backbone. The model operates in three stages: pre-segmentation, segmentation, and refinement. It achieves good accuracy and F1_score with less number of parameters for single task learning. Liu et al. [17] proposed Residual depth-wise over-parameterized(ResDO)-UNet that uses ResDO-conv with multiple pooling operations, a pooling and attention fusion blocks to implement non-linear pooling and multi-scale feature. A minimal U-Net variant to segment retinal vessels is introduced by Jingfei et al. [8], known as Salient U-Net (S-UNet). To address concerns with data imbalance, it is a bridge-style U-Net design with a saliency mechanism that employs a cascading technique and adds foreground elements. It is a lightweight model with 0.21M parameters, outperforms SOTA methods on

benchmark datasets, DRIVE [1], CHASE_DB1 [24], and the TONGREN clinical dataset [9].

Li et al. [15] presented a dual-directional lightweight model with an attention block that models long-range dependencies and deals with intra-class variations. This block used pooling methods in vertical and horizontal manner to create an attention map to collect global contextual details from semantically similar regions or parts of the same object class. Moreover, the Selective Kernel (SK) unit is used instead of traditional convolution to get multiple features with distinctive sizes of receptive fields, which are impacted by soft attention. The proposed model can accurately detect various shapes and sizes of retinal vessels. Aurangzaib et al. [2] used a lightweight DL technique called ColonSegNet to segment BV from retinal images. Three online available datasets (CHASE_DB1, DRIVE and STARE) are utilized to assess its performance. This model is lightweight with five million trainable parameters appropriate for deployment on lower-end edge devices. This an efficient system for single task learning i.e., segmentation of retinal BV. Shahzaib et al. presented G-Net light [11], which is a lightweight model obtained by modification of Google net for segmentation of retinal vessels.

Optic Disc Segmentation. An automated optic disc (OD) segmentation technique is introduced by Yinghua et al. [6] that synergizes the design for irregular fundus images. The spatial correlation based probability bubble is designed between retinal BV and OD. Its outcome is integrated into the U-net's output layer by evaluating joint probabilities. Souvik et al. [20] introduced a DL technique to segment the OD from fundus images using an enhanced convolutional network. In this approach, VGG16 serves as the encoder, while its symmetric counterpart functions as the decoder, ensuring more efficient object segmentation. Furthermore, the Convolutional LSTM is integrated within the encoder block.

Mehwish et al. [21] proposed EDDense-Net for segmentation of OD and OC jointly. In this network, each block contains dense layer along with grouped convolution. It employs dice pixel classification to address class imbalance issues. Xia et al. [35] presented a network that combines CNN and transformer for optic disc (OD) and cup (OC) segmentation from fundus images. Firstly, it utilized CNN to acquire local features and combined the ASPP module to gather multi-scale data. Then, the transformer is used to access global features. The model is tested on REFUGE dataset, resulting in improvement of model's efficiency by merging the two sets of features. A. Sevastopolsky [29] performed a transformation of the original U-Net CNN to segment OD, where the image dimensions of the given image are expanded by going through the contracting and expansive network paths with up-sampling layers. This approach produces high-quality OD segmentation with less inference time.

Multi-Task Learning (MTL). Vengalil et al.[32] proposed a DL model based on MTL, designed for simultaneous segmentation of the OD, BV, exudates, and macula. The entire image serves as the foundation for both training and prediction. The proposed model comprises a modified U-Net architecture, which exclusively employs convolutional and de-convolutional layers. In term of MTL, a hybrid CNN-Transformer encoder is proposed by Cheng et al. [3] to tackle task correlation and heterogeneity for the MTL model based on the MRI dataset. The proposed model comprises a transformer and CNN to obtain spatial and global features. A loss function for MTL is developed by combining classification and segmentation loss with random weights. This method is assessed on public datasets from multiple institutions. The proposed multi-task model reveals outstanding results compared to single-task learning models and other SOTA approaches. The semi-supervised MTL structure for unlabeled datasets improves the efficiency of glioma extraction and isocitrate dehydrogenase(IDH) genotyping. Although the above mentioned techniques provide outstanding accuracy, researchers are still focusing on balancing performance and computational complexity.

3. Methodology

The proposed model is based on MTL to simultaneously segment retinal BV and OD. It comprises an encoder-decoder framework, embedding convolutional layers and MHA within the encoder. Feature fusion is performed at the bottleneck, integrating the resulting features derived from both the CNN and MHA block. This fusion enables the model to gain local and global information by utilizing the strengths of CNN and MHA. Additionally, the decoder is enriched with the CBAM and ConvTranspose, paired with two distinct outputs. It ensures that both tasks can learn common representations while allowing each task to learn its specific features. Figure 2 depicts the system diagram of our proposed model.

3.1. Preprocessing

In the pre-processing stage, we performed data augmentation to cope with minor dataset issues in medical images. We applied three primary augmentation techniques to both the DRIVE and IOSTAR datasets, namely: CLoDSA[1], IMGAUG[2], and Albumentations[3].

[1] https://github.com/joheras/CLoDSA
[2] https://github.com/aleju/imgaug
[3] https://github.com/albumentations-team/albumentations

Contrastive Data Augmentation for Image Segmentation(CLoDSA) is a comprehensive data augmentation technique for image segmentation that includes several transformations. We used the following techniques to increase the size of DRIVE and IOSTAR datasets: Dropout (to set some pixels from the image to 0 for training), white noise, gamma correction, histogram equalization, blurring, elastic deformation, flipping, shearing, sharpening, and boosting saturation.
IMGAUG is a robust and comprehensive image augmentation technique that provides a variety of operations that might applied to images. We use augmentation techniques such as rotation, JPEG compression, zooming, contrast enhancement, and shifting.
Albumentations is a fast image augmentation library widely used in DL techniques. It performs various computer vision tasks, including semantic segmentation, and is tuned for optimal speed and performance. It employed random crop, padding, rotation, grid distortion, optical distortion, CLAHE (Contrast Limited Adaptive Histogram Equalization), and random brightness.

3.2. Encoder

The encoder in neural networks distills and encodes the features of an input that facilitates subsequent layers in the model. It primarily serves to extract and represent high-level features from the image. In the propose model, it captures both spatial hierarchies and complex patterns in retinal images for further processing. The encoder comprises convolution blocks along with a MHA block to capture local and long-range dependencies. Consider I is an input retinal image with 3 channels and 512x512 image size.

$$I \in \mathbb{R}^{512\times512\times3}$$

The retinal image is pass to convolution blocks with ReLU activation function, max-pooling and batch normalization layers. Each convolution block can be expressed as:

$$\text{Conv\_Block}(i) = \text{BN}(\text{MaxPool}(\text{ReLU}(\text{conv}(K_i * I + b_i))))$$

Where, i indicates the number for each convolution block, K_i represents the kernel of the convolutional layer, b_i symbolizes the bias and the operation $*$ denotes convolution.

It employs ReLU [26] as an activation function to learn complex features, as it introduces non-linearity in the network.

Max Pooling operates on a matrix by selecting the maximum element from the sub-matrix at each position defined by the pooling window [22].

BN refers to batch normalization that normalizes the resulting output g by normalising and scaling with respect to the mean μ and standard deviation σ respectively, with a small constant ϵ added for numerical stability. The result is then scaled by γ and shifted by β, where γ and β are learnable parameters.

$$\text{BN}(g) = \gamma\left(\frac{I-\mu}{\sqrt{\sigma^2+\epsilon}}\right) + \beta$$

In this expression, g is the output of max pooling operation $\text{MaxPool}(f)_{ij}$.

The convolutional layers are integrated with MHA by reshaping the output of final convolutional layer. The MHA comprises four attention heads and a key dimension d_k of 32, to capture long-range dependencies in retinal images. This concept is precisely defined by following expressions.

$$\mathbf{C_3} \in \mathbb{R}^{n\times h\times w\times c} \xrightarrow{\text{reshape}} \mathbf{S} \in \mathbb{R}^{n\times(h\cdot w)\times c}$$

where, C_3 is the third convolution block in the encoder, S is reshaped feature sequence, n indicates the batch size, h and w signifies the height and width, and c indicates the total channels .

The output feature map, represented as t_m, is produced by processing the reshaped sequence S using attention mechanism given below.

$$t_m = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}}\right) V_m$$

Where, $Q_m = s_m W_Q$ represents the query projection of the m-th reshaped feature s_m and $K_m = s_m W_K$ represents the key projection of the m-th reshaped feature s_m. $V_i = s_m W_V$ denotes the value projection of the m-th reshaped feature s_m. W_Q, W_K, and W_V symbolizes weight matrices that map the reshaped features to the query, key, and value projections, respectively. Finally, d_k represents the dimension of the key projections.

The individual outputs '$t_1, t_2, ...t_L$' are then concatenated for the final output of the MHA.

$$\text{MHA}(T) = \text{Concat}\,(t_1, t_2, \ldots, t_L)$$

The characteristics of CNN and the MHA are combined by fusing the outputs of these components, which comprise local spatial features and global context information. The fusion makes the model more powerful and flexible, potentially improving the model's performance on the segmentation tasks.

The aim is to tile the output of MHA T with dimensions $1 \times 1 \times D_T$, to match the spatial dimensions of convolution

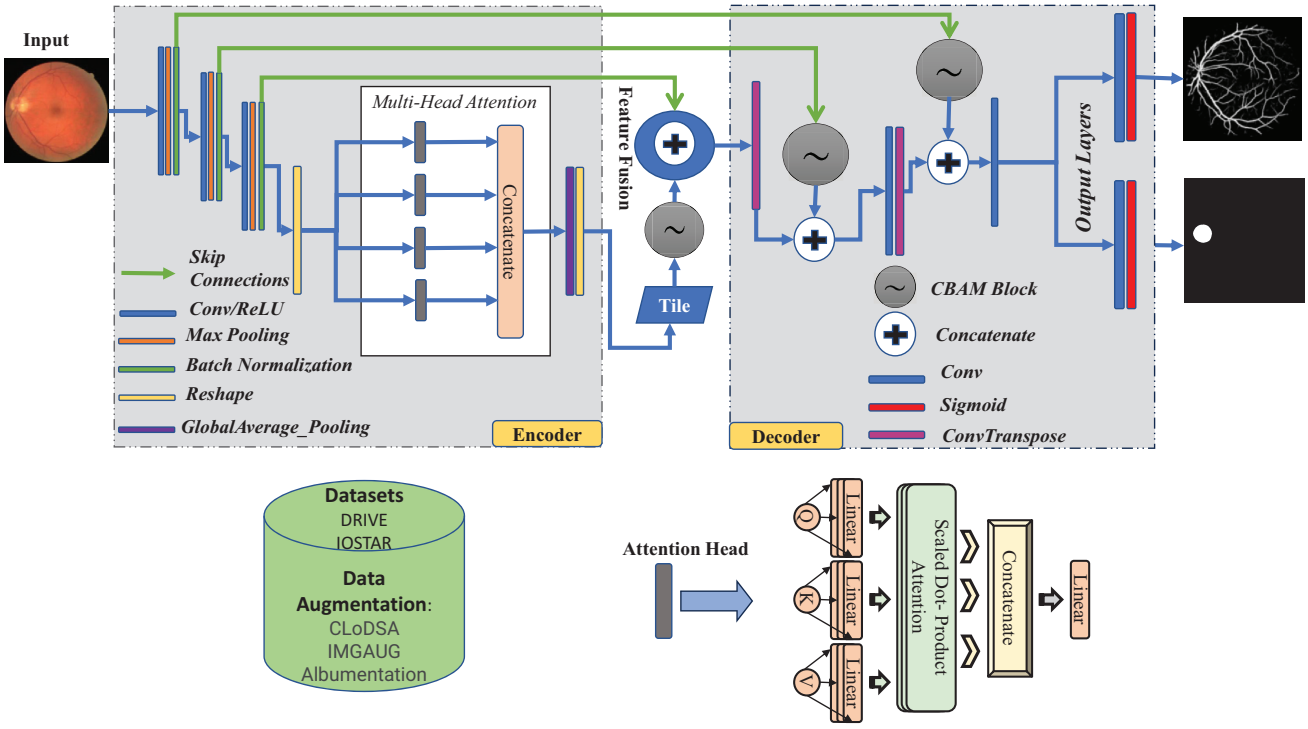


Figure 2. RetinaLiteNet architecture

block's output C with dimensions $H \times W \times D_C$, and then concatenate them. The tiling operation is defined as:

$$T' = \text{Tile}(T, [H, W, 1])$$

This results in T' with dimensions $H \times W \times D_T$. Finally, the feature fusion through concatenation is:

$$F = \text{Concat}(C, T')$$

Where, F has dimensions $H \times W \times (D_C + D_T)$, combining the strengths of both the CNN and MHA features.

3.3. Decoder

The decoder is employed for reconstructing the feature representations from the encoder back to the spatial resolution of the original input. It primarily reverses the spatial compression performed by the encoder. The decoder comprises CBAM and ConvTranspose layers to obtain the original image size. Skip connections are applied by processing up-sampled feature maps with the CBAM block and then concatenated with the encoder's feature maps. CBAM facilitates the model's focus on more relevant and detailed features and makes it more efficient. It performs two stages of operation: channel attention (CA) and spatial attention (SA). This procedure is repeated throughout the decoder blocks, gradually increasing spatial resolution while reducing the depth. Each upsampling block concatenates the ConvTranspose layer with CBAM.

At the end, two separate convolutional layers with sigmoid activation function are implemented to get the segmentation masks for the retinal BV and OD. . The representation of output block is:

$$Out_{BV} = Sigmoid(\text{Conv}(\text{Upsample}(R)))$$
$$Out_{OD} = Sigmoid(\text{Conv}(\text{Upsample}(R)))$$

where, R is the number of last upsampling block in the decoder and it's value is 3, sigmoid is the activation function, implied in the output layers, is suitable for a binary segmentation task as it ensures that the output values lie within the range of 0 and 1.It is defined as:

$$sigmoid(k) = \frac{1}{1 + e^{-k}}$$

Furthermore, the Adam optimizer is employed for training our model, and a custom loss function is developed with the combination of the dice loss and iou loss for each task.

$$DiceLoss = 1 - F1\_Score$$
$$IoULoss = 1 - IoU$$
$$CombinedLoss = DiceLoss + IoULoss$$

Where, formulas for F1_Score and IoU are mentioned in section 4.1. This combined loss function encourages the model to improve the overlap between the predicted and actual segmentation masks for both tasks.

3.4. Datasets

We have used two datasets for our model evaluation i.e. DRIVE and IOSTAR. The DRIVE dataset [25] encompasses 40 colored retinal images, each of 565 x 584-pixel resolution (8 bits per channel). A 45-degree viewing field Canon CR5 non-mydriatic 3CCD camera was used to acquire these images. The dataset splits into two subsets: a training set and a test set, each comprising 20 images. Notably, while one expert annotation is available in the training set, the test set benefits from two distinct expert annotations. The IOSTAR dataset [37] comprises 30 images, each with a 1024 x 1024-pixel resolution. The distinct feature of these images is that they stem from scanning laser ophthalmoscopy (SLO), which yields high-resolution retinal images. All images within the IOSTAR collection were taken using an EasyScan device produced by i-Optics Inc., which integrates the SLO method and provides a 45° Field of View (FOV). A team of retinal image analysis experts diligently annotated all vessels and disc in this dataset.

4. Experiments and Results

4.1. Performance Metrics

Standard evaluation metrics, including F1_score, Jaccard (IoU), specificity, sensitivity and AUC, are used to assess the model's performance on the publicly available datasets. The selected evaluation metrics are as given below:

$$F1\_Score = \frac{TP + TP}{TP + TP + FP + FN}$$

$$Jaccard(IoU) = \frac{TP}{FN + FP + TP}$$

$$Sen = \frac{TP}{FN + TP}$$

$$Spe = \frac{TN}{FP + TN}$$

$$AUC = \frac{1}{2 \times TP \times TN} \sum_{i=1}^{TP} \sum_{j=1}^{TN} \left(1 + \frac{FN}{FP}\right)$$

where, FP, TP, FN, TN represent false positive, true positive, false negative, and true negative, respectively. Additionally, specificity and sensitivity are denoted by Spe and Sen, respectively. Moreover, the model's complexity is evaluated using several criteria, including FLOPs, inference time, trainable parameters count, and the memory needed to save the parameters.

4.2. Implementation Details

The experiments were performed using TensorFlow and Keras and the model was trained on Google Colab Pro, featuring 32 GB of RAM and an NVIDIA Tesla P100 GPU. It was tested on a single CPU, specifically an 11th Gen Intel(R) Core(TM) i9-11900KF, with a maximum frequency of 5300 MHz for inference time.

Table 1. Experimental results comparing model complexity: A comparison of parameters count, FLOPs, inference time, and memory usage among UNet, UNet++, Attention UNet, and RetinaLiteNet models.

Model	Param (M)	FLOPs (G)	Infer. Time (s)	Memory (MB)
UNet [27]	7.76	96.682	0.685	29.60
UNet++ [38]	9.04	238.52	1.667	34.49
Attention UNet [19]	9.25	371.68	2.679	35.33
RetinaLiteNet	**0.066**	**2.4614**	**0.384**	**0.25**

Moreover, our model was evaluated for the MTL, which performs retinal BV and OD segmentation simultaneously. The images used for network training are 565x584 and 1024x1024 pixels for DRIVE and IOSTAR dataset, respectively. These images were resized into 512x512 pixels to train it on GPU with 300 epochs and batch size 16. We employed 400 images for model training from each dataset and for testing, we utilized twenty images from DRIVE and ten images from IOSTAR.

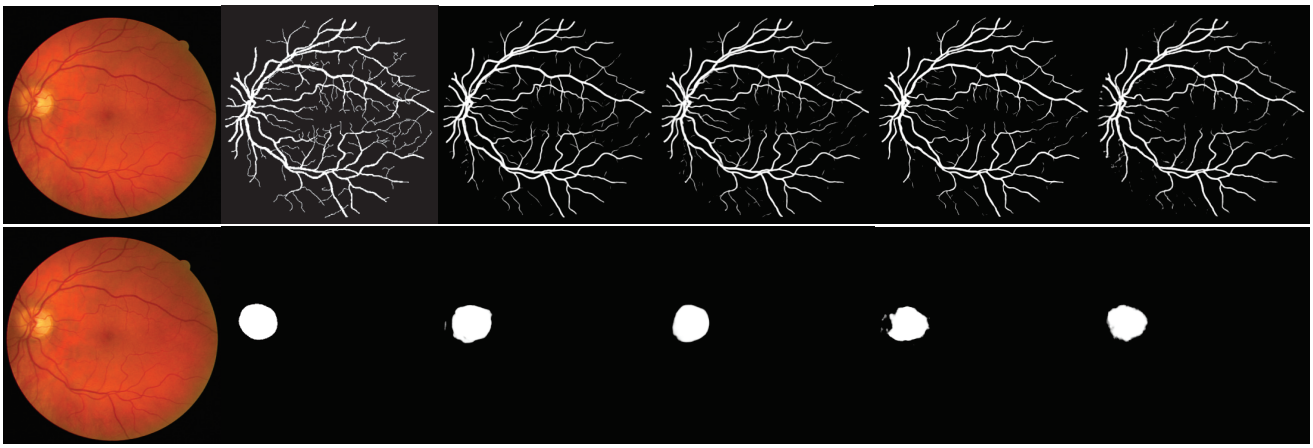
4.3. Results

To prove the lightweight characteristics and robustness of our model, we compared it with previous edge-cutting models designed for retinal feature segmentation and reported them in Table 1. It can be visualized that UNet, UNet++, and attention Unet have 7.76, 9.04, and 9.25 million parameters, respectively, however our model has only 0.066 million parameters. It requires only 0.25 MB of memory whereas above mentioned models require 29.60,34.49, and 35.33 MB, respectively, which could be better for low memory hardware deployment. Similarly, the proposed model has fewer FLOPs estimation and inference time relative to its counterparts, having 96.68, 238.52, and 9.25 GFLOPs with inference time of 0.68, 1.66,2.67 seconds respectively. In contrast, our model has only 2.46 GFLOPs with an inference time of 0.38 seconds. Such attributes render it lightweight and suitable for deployment on resource-constrained hardware platforms. The existing models were trained and tested under the same conditions as the proposed model to ensure a fair comparison.

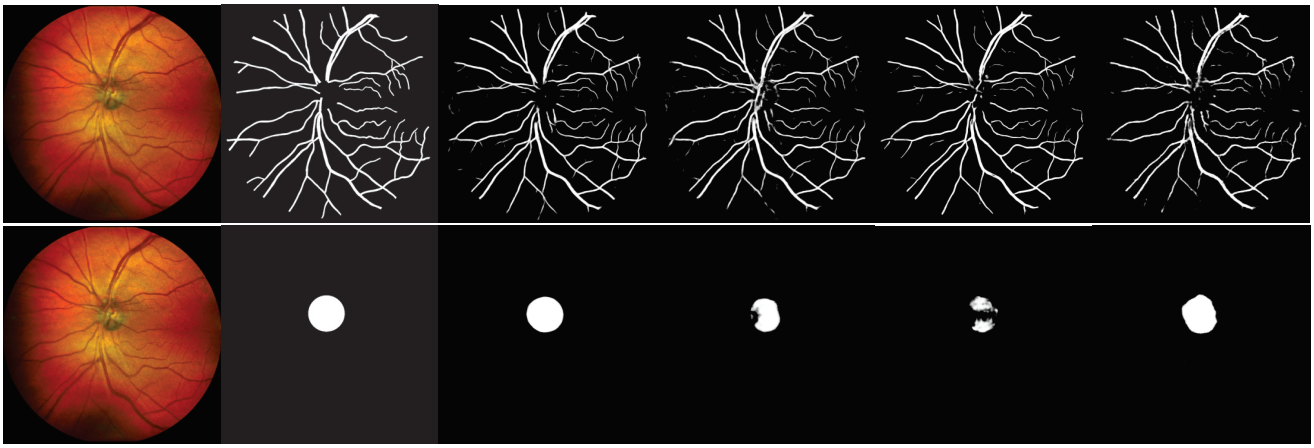
Table 2 presents the comparison of evaluation metrics of the existing SOTA models for MTL with our proposed model using the DRIVE and IOSTAR datasets to confirm if the system has achieved comparable performance with a smaller number of parameters. The results show that our model achieved a higher F1_score and specificity in terms of BV and a higher F1_score, Jaccard, and AUC

Table 2. Performance matrix comparison of RetinaLiteNet with other retinal feature segmentation models on the DRIVE and IOSTAR Datasets.

	DRIVE										IOSTAR									
	Blood Vessels					Optic Disc					Blood Vessels					Optic Disc				
Model	**F1**	**Jac.**	**Sen.**	**Spe.**	**AUC**	**F1**	**Jac.**	**Sen.**	**Spe.**	**AUC**	**F1**	**Jac.**	**Sen.**	**Spe.**	**AUC**	**F1**	**Jac.**	**Sen.**	**Spe.**	**AUC**
UNet	80.5	67.5	87.5	95.7	98.0	86.0	76.6	92.1	99.5	99.0	79.1	65.5	72.3	98.0	97.0	90.9	83.6	85.3	99.7	98.0
UNet++	80.0	66.8	89.4	95.1	98.0	89.6	76.2	92.5	94.6	97.0	78.0	64.1	78.3	95.7	96.0	87.6	78.5	82.1	99.8	98.0
Att_UNet	80.5	67.6	77.8	97.8	97.0	84.3	81.3	93.1	99.6	99.0	80.0	66.6	79.1	98.5	93.0	77.3	64.1	64.9	99.7	94.0
Ours	**80.6**	67.5	78.4	**98.0**	97.0	**93.3**	**88.0**	**94.0**	97.0	99.0	**80.1**	**67.6**	77.5	97.2	**98.0**	85.4	74.9	76.5	**99.9**	**99.0**



(a) DRIVE Image (b) Ground Truth (c) Unet Results (d) Unet++ Results (e) Att._Unet Results (f) RetinaLiteNet Results



(a) IOSTAR Image (b) Ground Truth (c) Unet Results (d) Unet++ Results (e) Att._Unet Results (f) RetinaLiteNet Results

Figure 3. Visual representation of segmented BV and OD along with the original images and their ground truths, for UNet, UNet++, Attention_UNet and RetinaLiteNet on DRIVE and IOSTAR datasets.

in terms of OD segmentation compared to other models on the DRIVE dataset. Similarly, it achieved a higher F1_score, Jaccard, and AUC for BV segmentation and higher specificity and AUC for OD segmentation on the IOSTAR dataset. Visual representation of the results, comparing our model with existing SOTA models, is presented in Figure 3. Consequently, our model has comparable performance with existing models while retaining its advantage of being lightweight.

4.4. Ablation Studies

We evaluate our proposed model on the DRIVE dataset with several variations in the number of heads and key dimension (d_k). d_k represents the dimensionality of the key vectors in each attention head and is adjustable during training. Careful selection of d_k is essential to prevent underfitting with a small d_k value due to insufficient information and overfitting with a large d_k value by learning redundant information. Our results, detailed in Table 3, explore configurations

Table 3. Performance matrix for retinal feature segmentation on DRIVE Dataset with several combination of number of heads and key dimension (d_k).

Head	d_k	Blood Vessels %				Optic Disc %			
		F1	Jacc.	Sen.	Spe.	F1	Jacc.	Sen.	Spe.
2	32	78.9	65.8	77.1	98.0	90.5	82.4	96.5	97.2
3	32	79.2	66.2	78.0	97.9	91.1	83.2	96.3	97.5
4	16	79.3	66.4	78.6	97.8	89.1	81.9	95.7	97.4
4*	32	80.6	67.5	78.4	98.0	93.3	88.0	94.0	97.0
5	32	78.3	65.1	72.2	98.6	89.9	84.6	95.9	97.4

Table 4. Computational complexity of the model with several combination of number of heads and key dimension (d_k).

Head	d_k	Memory (MB)	FLOPs (G)	Param (M)
2	32	0.22	2.40	0.057
3	16	0.21	2.38	0.055
3	32	0.24	2.38	0.061
4	16	0.22	2.45	0.057
4*	32	0.25	2.46	0.066
5	32	0.27	2.55	0.070

with 2, 3, 4, and 5 attention heads coupled with d_k of 16 and 32. Our model achieves the best performance using 4 attention heads with a d_k of 32. Moreover, Table 4 shows a slight difference in the computational performance of our model with different numbers of heads and d_k.

Table 5. Performance metrics for retinal feature segmentation with positional encoding on DRIVE dataset.

P.E	Blood Vessels %				Optic Disc %			
	F1	Jac.	Sen.	Spe.	F1	Jac.	Sen.	Spe.
With	80.8	67.8	78.8	98.2	93.4	88.2	94.4	97.4
Without*	80.6	67.5	78.4	98.0	93.3	88.0	94.0	97.0

Our model is also evaluated with and without positional encoding (PE) to assess the impact of positional encoding (PE) on its performance, as shown in Table 5. As we are using MHA in the bottleneck of the encoder, it gives comparable results without including PE. Therefore, we opted to exclude PE from our model to remove extra overhead.

We accessed our model performance on benchmark datasets, i.e, DRIVE and IOSTAR, across various epochs and batch sizes to determine the optimal settings for full convergence. The experimental results showcase the significant impact of batch size and training duration on model efficiency and effectiveness.The findings presented in Tables 6 and 7 depict that the best performance achieved at 300 epochs with 16 batch size.

Table 6. RetinaLiteNet performance matrix for retinal feature segmentation on DRIVE Dataset with different epochs and batch sizes.

Ep.	BS	Blood Vessels %				Optic Disc %			
		F1	Jacc.	Sen.	Spe.	F1	Jacc.	Sen.	Spe.
200	32	79.0	65.0	77.0	98.0	91.0	85.5	95.0	97.5
250	32	77.4	65.0	76.0	99.0	91.0	85.5	91.0	97.8
300	32	79.8	65.0	78.0	98.5	89.4	84.0	93.5	97.5
300	48	79.9	64.0	74.0	98.0	90.0	83.0	91.0	97.7
300*	16	80.6	67.5	78.4	98.0	93.3	88.0	94.0	97.0

Table 7. RetinaLiteNet performance matrix for retinal feature segmentation on IOSTAR Dataset with different epochs and batch sizes.

Ep.	BS	Blood Vessels %				Optic Disc %			
		F1	Jacc.	Sen.	Spe.	F1	Jacc.	Sen.	Spe.
200	32	79.4	63.4	71.5	98.0	81.5	71.8	81.2	99.0
250	32	78.0	62.9	39.0	99.6	84.0	72.0	81.9	98.8
300	32	78.5	65.0	72.0	98.5	84.5	73.5	83.2	99.6
300	48	75.2	61.5	66.0	99.0	82.3	71.7	80.4	98.0
300*	16	80.0	67.0	79.6	98.0	86.0	75.1	83.5	99.7

5. Conclusion

In this paper, we present a lightweight deep learning model designed for retinal feature segmentation, specifically blood vessels and optic disc. Our model comprises an encoder-decoder framework, in which convolutional layers and multi-head attention mechanism embedded within the encoder that retrieves both local and global details from the images. The resulting features are then fused together at the bottleneck of the encoder by integrating the results from convolutional layer and attention heads, enhancing the model's efficiency . To refine the features further, we have incorporated skip connections along with CBAM in the decoder. This approach fine-tunes the focus of the model on useful features, boosting its efficiency and effectiveness in segmentation tasks. Extensive evaluation on the DRIVE and IOSTAR datasets produced promising results, yielding F1_scores of 80.6% and 93.3% on DRIVE and 80.1% and 85.4% on IOSTAR for simultaneous segmentation of blood vessels and optic disc, respectively. Our model requires a memory of 0.25 MB, 66,000 parameters, and a computational cost of 2.46 GFLOPs, which satisfies its lightweight aspects. These findings demonstrate that efficient medical image analysis is possible even with limited hardware resources.

*Indicates the parameters chosen for our model in section 4.4.

References

[1] Ridge-based vessel segmentation in color images of the retina, author=Staal, Joes and Abràmoff, Michael D and Niemeijer, Meindert and Viergever, Max A and Van Ginneken, Bram. *IEEE transactions on medical imaging*, 23 (4):501–509, 2004. 3

[2] Khursheed Aurangzeb, Rasha Alharthi, Syed Irtaza Haider, and Musaed Alhussein. An efficient and Light Weight Deep Learning Model for Accurate Retinal Vessels Segmentation. *IEEE Access*, 2022. 3

[3] Jianhong Cheng, Jin Liu, Hulin Kuang, and Jianxin Wang. A Fully Automated Multimodal MRI-Based Multi-Task Learning for Glioma Segmentation and IDH Genotyping. *IEEE Transactions on Medical Imaging*, 41:1520–1532, 2022. 3

[4] Carol Y Cheung, Vincent Mok, Paul J Foster, Emanuele Trucco, Christopher Chen, and Tien Yin Wong. Retinal imaging in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(9):983–994, 2021. 2

[5] Carol Y. Cheung, An Ran Ran, Shujun Wang, Victor T.T. Chan, Kaiser Sham, Saima Hilal, Narayanaswamy Venketasubramanian, Ching Yu Cheng, Charumathi Sabanayagam, Yih Chung Tham, Leopold Schmetterer, Gareth J. McKay, Michael A. Williams, Adrian Wong, Lisa W.C. Au, Zhihui Lu, Jason C. Yam, Clement C. Tham, John J. Chen, Oana M. Dumitrascu, Pheng Ann Heng, Timothy C.Y. Kwok, Vincent C.T. Mok, Dan Milea, Christopher Li Hsian Chen, and Tien Yin Wong. A deep learning model for detection of Alzheimer's disease based on retinal photographs: a retrospective, multicentre case-control study. *The Lancet Digital Health*, 4:e806–e815, 2022. 2

[6] Yinghua Fu, Jie Chen, Jiang Li, Dongyan Pan, Xuezheng Yue, and Yiming Zhu. Optic disc segmentation by U-net and probability bubble in abnormal fundus images. *Pattern Recognition*, 117:107971, 2021. 3

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[8] Jingfei Hu, Hua Wang, Shengbo Gao, Mingkun Bao, Tao Liu, Yaxing Wang, and Jicong Zhang. S-UNet: A Bridge-Style U-Net Framework With a Saliency Mechanism for Retinal Vessel Segmentation. *IEEE Access*, 7:174167–174177, 2019. 2

[9] Jingfei Hu, Hua Wang, Zhaohui Cao, Guang Wu, Jost B Jonas, Ya Xing Wang, and Jicong Zhang. Automatic artery/vein classification using a vessel-constraint network for multicenter fundus images. *Frontiers in cell and developmental biology*, 9:659941, 2021. 3

[10] Shahzaib Iqbal, Tariq M. Khan, Khuram Naveed, Syed S. Naqvi, and Syed Junaid Nawaz. Recent trends and advances in fundus image analysis: A review. *Computers in Biology and Medicine*, 151:106277, 2022. 1

[11] Shahzaib Iqbal, Syed S. Naqvi, Haroon A. Khan, Ahsan Saadat, and Tariq M. Khan. G-Net Light: A Lightweight Modified Google Net for Retinal Vessel Segmentation. *Photonics*, 9:923, 2022. 3

[12] Ashir Javeed, Ana Luiza Dallora, Johan Sanmartin Berglund, Arif Ali, Liaqata Ali, and Peter Anderberg. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. *Journal of medical systems*, 47(1): 1–25, 2023. 2

[13] Shalini Kapoor and Sameep Mehta. *AI for You: The New Game Changer*. Bloomsbury Publishing, 2022. 2

[14] Jianyong Li, Ge Gao, Yanhong Liu, and Lei Yang. MAGF-Net: A multiscale attention-guided fusion network for retinal vessel segmentation. *Measurement: Journal of the International Measurement Confederation*, 206, 2023. 2

[15] Kaiqi Li, Xingqun Qi, Yiwen Luo, Zeyi Yao, Xiaoguang Zhou, and Muyi Sun. Accurate Retinal Vessel Segmentation in Color Fundus Images via Fully Attention-Based Networks. *IEEE Journal of Biomedical and Health Informatics*, 25:2071–2081, 2021. 3

[16] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2

[17] Yanhong Liu, Ji Shen, Lei Yang, Guibin Bian, and Hongnian Yu. ResDO-UNet: A deep residual network for accurate retinal vessel segmentation from fundus images. *Biomedical Signal Processing and Control*, 79:104087, 2023. 2

[18] Yanhong Liu, Ji Shen, Lei Yang, Hongnian Yu, and Guibin Bian. Wave-Net: A lightweight deep network for retinal vessel segmentation from fundus images. *Computers in Biology and Medicine*, 152:106341, 2023. 2

[19] Yan Lv, Hui Ma, Jianian Li, and Shuangcai Liu. Attention guided U-Net with atrous convolution for accurate retinal vessels segmentation. *IEEE Access*, 8:32826–32839, 2020. 6

[20] Souvik Maiti, Debasis Maji, Ashis Kumar Dhara, and Gautam Sarkar. Automatic detection and segmentation of optic disc using a modified convolution network. *Biomedical Signal Processing and Control*, 76:103633, 2022. 3

[21] Mehwish Mehmood, Khuram Naveed, Haroon Ahmed Khan, and Syed S Naqvi. EDDense-Net: Fully Dense Encoder Decoder Network for Joint Segmentation of Optic Cup and Disc. *arXiv preprint arXiv:2308.10192*, 2023. 3

[22] Jawad Nagi, Frederick Ducatelle, Gianni A Di Caro, Dan Cireşan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE international conference on signal and image processing applications (ICSIPA)*, pages 342–347. IEEE, 2011. 4

[23] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[24] Christopher G Owen, Alicja R Rudnicka, Robert Mullen, Sarah A Barman, Dorothy Monekosso, Peter H Whincup, Jeffrey Ng, and Carl Paterson. Measuring retinal vessel tortuosity in 10-year-old children: validation of the computer-assisted image analysis of the retina (CAIAR) program. *Investigative ophthalmology & visual science*, 50(5):2004–2010, 2009. 3

[25] Touseef Ahmad Qureshi, Maged Habib, Andrew Hunter, and Bashir Al-Diri. A manually-labeled, artery/vein classified benchmark for the DRIVE dataset. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 485–488. IEEE, 2013. 6

[26] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 4

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 6

[28] Vallikutti Sathananthavathi and G Indumathi. Encoder enhanced atrous (EEA) unet architecture for retinal blood vessel segmentation. *Cognitive Systems Research*, 67:84–95, 2021. 2

[29] Artem Sevastopolsky. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. *Pattern Recognition and Image Analysis*, 27(3):618–624, 2017. 3

[30] João VB Soares, Jorge JG Leandro, Roberto M Cesar, Herbert F Jelinek, and Michael J Cree. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Transactions on medical Imaging*, 25(9):1214–1222, 2006. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[32] Sunil Kumar Vengalil, Bharath Krishnamurthy, and Neelam Sinha. Simultaneous segmentation of multiple structures in fundal images using multi-tasking deep neural networks. *Frontiers in Signal Processing*, 2, 2023. 3

[33] Jingyi Wen, Dong Liu, Qianni Wu, Lanqin Zhao, Wai Cheng Iao, and Haotian Lin. Retinal image-based artificial intelligence in detecting and predicting kidney diseases: Current advances and future perspectives. *View*, page 20220070, 2023. 1

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[35] Xue Xia, Zhuxiang Huang, Zijian Huang, Lei Shu, and Lin Li. A CNN-Transformer Hybrid Network for Joint Optic Cup and Optic Disc Segmentation in Fundus Images. pages 482–486. Institute of Electrical and Electronics Engineers Inc., 2022. 3

[36] Yang Yu and Hongqing Zhu. M3U-CDVAE: Lightweight retinal vessel segmentation and refinement network. *Biomedical Signal Processing and Control*, 79, 2023. 2

[37] Jiong Zhang, Behdad Dashtbozorg, Erik Bekkers, Josien PW Pluim, Remco Duits, and Bart M ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE transactions on medical imaging*, 35(12):2631–2644, 2016. 6

[38] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2, 6