IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Advancing Retinal Vessel Segmentation with Diversified Deep Convolutional Neural Networks

TANZINA AKTER TANI[1], (Student Member, IEEE), and JELENA TEŠIĆ[1], (Senior Member, IEEE)
[1]Department of Computer Science, Texas State University, San Marcos, TX 78666, USA
Corresponding author: Tanzina Akter Tani (e-mail: tanzinatani@txstate.edu).

ABSTRACT Retinal vessel segmentation is crucial for the diagnosis and monitoring of ophthalmic illnesses. Deep learning algorithms have been extensively utilized in automated segmentation to improve accuracy and efficiency. In this paper, we introduce the use of DeepLabV3+ architecture to segment retinal blood vessels and enhance its performance by applying six different deep neural network backbones: ResNet50, DenseNet121, MobileNetV2, Xception, Xception with lower features (XceptionLF), and Xception lower features with overlapping regions (XceptionLFOR) patches. We also demonstrate the robustness of placing the Swin Transformer into the DeepLabV3+ model. The integration of XceptionLF and XceptionLFOR into the pipeline enhances the semantic segmentation of retinal images by enabling the merging of global and patch-specific features along with features from both lower and higher resolutions. The enhancements enable our proposed best model, XceptionLFOR to obtain a (98.76%) accuracy and (89.23%) dice score, which represents a significant advancement in applying advanced deep-learning techniques for medical imaging. Moreover, XceptionLFOR model achieves a higher performance and better $F1$ score (0.49%) over the state-of-the-art for the FIVES benchmark evaluation despite using lower image resolution (256 resolution patches from 512 resolution images). The use of lower resolution balances computational efficiency with enhanced accuracy, enabling faster processing and deployment in resource-constrained environments. The findings in this paper point in the right direction in improving semantic segmentation for retinal vessel images, and they highlight the potential to improve early diagnosis and treatment outcomes for ocular illnesses.

INDEX TERMS DeepLabV3+, DCNN, FIVES dataset, Retinal vessel segmentation, Swin Transformer.

I. INTRODUCTION

The segmentation of retinal blood vessels is essential for the early detection of ophthalmic illnesses, such as diabetic retinopathy, hypertension, muscular degeneration, and glaucoma. It is critical for reducing eyesight impairment and improving patient outcome [1]. Blood vessel segmentation allows the quantitative analysis of retinal blood vessels, such as vessel diameter, branch pattern, and changes over time. This information is helpful in following the progression of disease and determining the efficacy of treatment [1]. Retinal blood vessel analysis can also reveal information about one's overall cardiovascular health. Changes in vessel characteristics may signal the development of some cardiovascular illnesses, making it an essential diagnostic tool [2]. Manual blood vessel segmentation is a time-consuming and labor-intensive procedure, whereas automated semantic segmentation approaches lessen the burden on medical practitioners while offering robustness and objectivity. Image segmentation is a computer vision task that groups pixels in an image. Semantic segmentation assigns semantic labels to the pixel groups to further identify the shapes and objects in the image [3]. The profound convolutional neural network breakthrough helped advance the field in the past couple of years; as outlined in Section II, there is room for improvement as the deep Convolutional Neural Networks (CNNs) struggle to deal with multi-scale information and explain missing vessels [4]; and the thin, low-contrast vessels provide additional challenges since typical segmentation approaches may lose spatial information [4].

Contributions: Medical imaging frequently uses a variety of semantic segmentation models, including U-Net, SegNet, FCNs, and DeepLabV3+. In these segmentation models, backbones are crucial because they serve as feature extractors by converting input images into high-dimensional representations using the spatial and contextual information needed for accurate segmentation. Therefore, in this paper, we propose to modify and advance the retinal semantic segmentation with different backbones based on the DeepLabV3+ methodology.

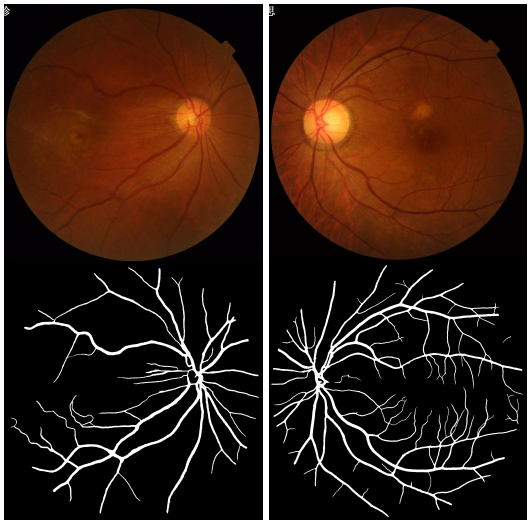

FIGURE 1. Two samples from the FIVES dataset (top) and the ground truth (bottom).

Selecting and fine-tuning several backbones allowed us to leverage their varied strengths, which improved the overall segmentation accuracy and robustness. Paper contributions are:

Introduced the novel *XceptionLF* backbone that combines Xception's low features with a convolutional block to extract features at different levels of semantic resolution.

Introduced the novel *XceptionLFOR*, which extends the XceptionLF by incorporating patches from overlapping regions of images, thus including both global and grid local features in the segmentation process.

Integrated four DCNN backbones, DCNNs-ResNet50, MobileNetV2, DenseNet121 and Xception and demonstrated their robustness in the DeepLabV3+ integration for the task.

Demonstrated the effectiveness and robustness of the Swin Transformer within DeepLabV3+ on the latest benchmark for the retinal vessel segmentation task for the first time.

Our proposed XceptionLFOR model outperforms the state-of-the-art on the FIVES benchmark [5], achieving a 0.49% higher $F1$ score despite using lower-resolution images (256 resolution patches from 512 resolution images). This balance of computational efficiency and accuracy enables faster processing and deployment. The supplementary code can be found in the GitHub repository ''RetinaVseg'' [6].

Paper Overview: Related work and state of the art are outlined in Section II, the proposed methodology is outlined in Section III, proof of concept setup is presented in Section IV, experimental results and discussion is presented in Section V and finally the conclusion and future work in Section VI.

II. RELATED WORK

Researchers have integrated the U-Net architecture with Residual Networks and RCNN [7] to apply in different benchmarks. From the results, R2U-Net showed the best performance in vessel segmentation without increasing parameters. Next, the incorporation of the inception-residual convolutional blocks into a U-like encoder-decoder architecture has been shown to improve the feature representation of vessel images in Vessel-Net [8]. VesselNet also has four supervision paths, including multi-scale supervision, to maintain rich features during optimization. SVSN [9] is a lightweight CNN using an encoder-decoder structure with spatial pyramid pooling inspired by DeepLabV3+ architecture. SVSN model captures multi-scale contextual information without pre- or post-processing, effectively segmenting both large and tiny retinal vessels. SA-UNet is a lightweight network for vessel segmentation that incorporates a spatial attention module and structured dropout convolutional blocks to improve feature refinement and prevent overfitting [10]. FR-UNet is a deep learning approach for segmenting thin, low-contrast vessels using full image resolution and a multi-resolution convolution interactive mechanism [11]. The dual-threshold iterative method (DTI) improved vessel connectivity by identifying weak vessel pixels. Next, pre-processing techniques like gray-scale transformation, CLAHE, normalization, and gamma transformation, along with data augmentation to prevent overfitting, improved the segmentation of retinal vessels when authors integrated Bi-FPN network into U-net [1]. Liu, Renyuan, et al. proposed the DA-Res2UNet model utilizing Res2blocks for multi-scale information extraction and dual attention for better focus [2]. DA-Res2UNet uses a GAN-based image generator to explain the segmentation process and identify errors. DeepLabV3+ network with image CLAHE pre-processing to enhance blood vessels in the image for better training and refines the output using morphological closing operations [12]. The MDUNet model is transformer-based, and it combines cross-dimensional transformation and self-attention mechanisms. It utilizes an encoder-decoder structure with Dense Blocks, HR Blocks, and ASPP modules for rich feature extraction and fusion [13]. The dual encoder system that retains edge information with a dynamic channel graph convolutional network improves feature synthesis across channels [14]. The proposed approach enhanced the fine detail of vessel segmentation and outperformed current methods on numerous benchmarks. A new deep learning architecture called AACA-MLA-D-UNet was also presented for segmenting retinal vessels. It used multi-level attention and adaptive atrous channel attention to improve segmentation accuracy [15]. Another model named FES-Net [16] was introduced that processed input images using four prompt convolutional blocks (PCBs) and a shallow up-sampling approach, bypassing conventional image enhancement. Its architecture balanced performance and computational cost by reducing trainable parameters.

Further, BFMD SN U-net with GCI-CBAM [17] was proposed to improve segmentation of retinal vessels by incorporating switchable normalization for faster convergence, block feature map distortion to prevent overfitting, and GCI-CBAM for better feature refinement. Another Vision Transformer-based model, G2ViT [18], was proposed for efficient vessel segmentation. CNN, GNN, and Vision Transformer were integrated into the model. It also employed a U-Net encoder

and graph-based representation, with MEFA and MLF2 modules enhancing edge information and feature fusion. HT-Net [19], a hybrid Transformer network that combines CNN and Transformer techniques, was also introduced for retinal vascular segmentation. It included an effective self-attention mechanism as well as novel Feature Fusion and Refinement Blocks, all of which significantly enhanced microvessel identification accuracy. Additionally, the research [20] presented a novel deep ensemble learning architecture that improves retinal vascular segmentation by combining Pyramid Vision Transformer (PVT) and FCN-Transformer models. The method effectively captures discriminative features, resulting in higher performance and robustness across numerous datasets than prior approaches. Another transformer-based model TiM-Net [21] was proposed which incorporated a transformer architecture into the M-Net model to improve retinal vessels segmentation. The employment of a dual-attention mechanism for noise reduction and a Transformer for capturing long-range relationships resulted in significantly superior performance than existing baselines. Chen, Danny, et al. proposed PCAT-UNet [22], a U-Net-based model that combined convolution branches with patches-based transformers to improve retinal vessel segmentation. This hybrid technique benefited from both local and global feature extractions, resulting in improved segmentation performance across many datasets.

To date, most of the retinal vessel works are done based on U-Net or the variation of U-Net, where only one paper applied the DeepLabV3+ model for retinal blood vessel segmentation. Also, only two research works utilized the FIVES [5] benchmark comparison in the related work. The first study [4] that used the FIVES dataset introduced SCOPE, a graph-based neural network that preserved continuity and connectivity in vessel segmentation. The second study [23] proposed SGAT-Net, a hybrid model that uses CNNs and transformers. Key modules in this paper included the Stimulus-Guided Adaptive Module (SGA-Module) for extracting detailed features, the SGAP-Former for enriching contextual embeddings, and the SGAFF for effective feature fusion. Evaluations on datasets including FIVES [5], DRIVE [24], CHASEDB1 [25], and STARE [26] showed SGAT-Net's superior results in accuracy and robustness.

III. METHODOLOGY

In this methodology section, we introduce the existing components and their use in our proposed modeling pipelines. Also, we enhance the backbone of the semantic segmentation architecture by using modified Deep Convolutional Neural Networks (DCNN) and incorporating Swin Transformers into the encoder. These enhancements represent novel approaches for retinal vessel segmentation.

A. STATE OF THE ART COMPONENTS

DeepLabV3+ [27], introduced in 2018, is an advanced semantic segmentation model that improves performance by integrating numerous components from prior generations. It starts with an encoder using a backbone network, which extracts higher features from input images utilizing pre-trained weights from massive datasets like ImageNet. The model effectively captures multi-scale contextual information by using atrous (dilated) convolutions that expand the receptive field without increasing parameters. The Atrous Spatial Pyramid Pooling (ASPP) module is a critical component of the architecture, as it performs parallel atrous convolutions at varying rates and adds global context via image-level features. To improve segmentation results, particularly object boundaries, DeepLabV3+ includes a decoder module that combines low-level features from the backbone with high-level ASPP features, followed by a series of 3x3 convolutions and up-sampling to the original image scale. Owing to its capacity to extract precise feature representations from datasets, pre-trained **DCNN models** like ResNet, Xception, DenseNet, and MobileNet are frequently applied as backbone models in segmentation tasks. The robust feature extraction capabilities of these backbones, which were first trained on big datasets like ImageNet, significantly enhance the performance of segmentation models. For example, ResNet's [28] deep architecture and residual connections capture detailed and hierarchical features, making it a familiar candidate for segmentation methods like U-Net and DeepLab. By utilizing depth-wise separable convolutions, Xception [29] offers rich feature maps and efficient computation that improves segmentation accuracy in models like DeepLabV3+. Known for its dense connection, DenseNet [30] enhances gradient flow and feature reuse, yielding parameter-efficient and effective segmentation models. MobileNet [31], designed for lightweight and efficient performance, is ideal for real-time applications and resource-constrained deployment. Commonly, it is used as a backbone for models like DeepLabV3+. The **Swin Transformer** [32] can be used to improve performance when it serves as the backbone of segmentation models. With this transformer method, local and global contexts can be captured effectively. The self-attention mechanism employed in the Swin Transformer is defined in Eq. 1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where Q is query, K is key, V is values, and dk is the critical dimension. Thus, the Swin Transformer model computes self-attention over progressively larger areas due to the hierarchical structure formed by combining windows.

B. PROPOSED MODELING PIPELINES

The DeepLabV3+ model maintains efficacy while reducing computing complexity by using depth-wise separable convolutions. The model produces cutting-edge results on benchmark datasets such as PASCAL VOC 2012 and Cityscapes while balancing accuracy and efficiency [27]. The pre-trained backbones outlined in Section III-A can be used to ensure high-quality and reliable segmentation results across a range of applications by reducing training time and processing

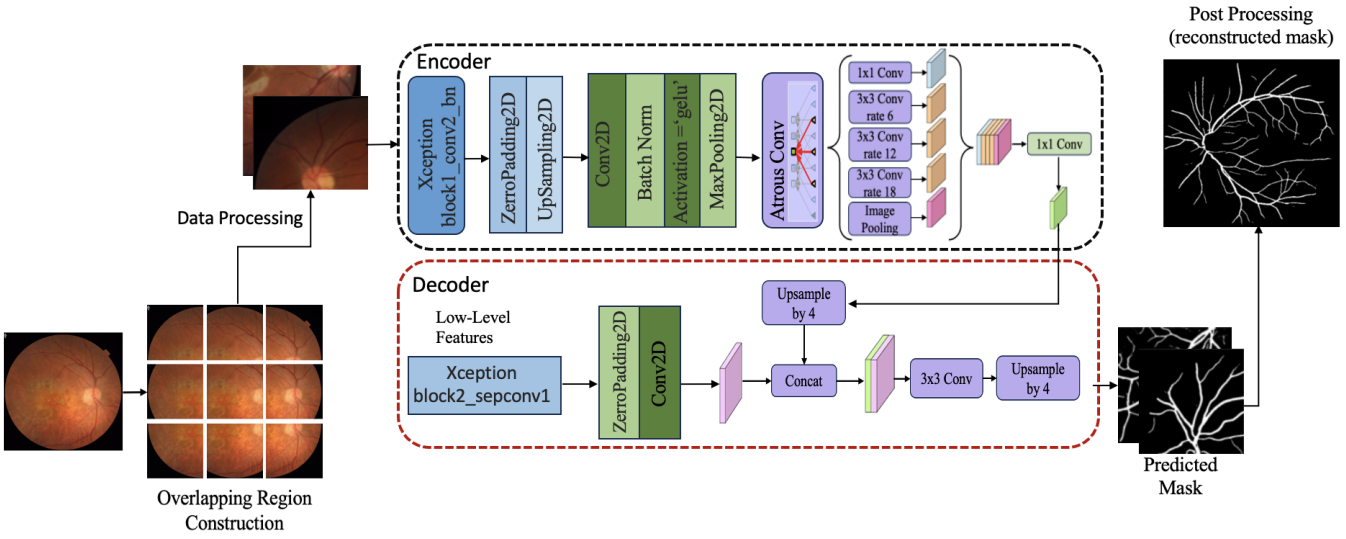

FIGURE 2. Modified DeepLabV3+ Architecture: Integration of Overlapping 256-Resolution Patches from 512-Resolution Images with Xception Low-Level Features and Additional Convolution Block, Followed by Post-Processing for Enhanced Retinal Segmentation

resources while also improving performance, particularly in scenarios with limited data.

The deep learning segmentation model DeepLabV3+ with different deep convolutional neural network (DCNN) backbones was employed as a baseline. Each of the following DCNNs is instantiated without the top layer, utilizing pre-trained ImageNet weights. It was customized to receive input through a specific input tensor using a particular size tensor 256×256×3.

Baseline backbones: DeepLabV3+ model with ResNet50 backbone in the encoder extracts high-level feature from the 'conv4_block6_2_relu' layer outputs. The Atrous Spatial Pyramid Pooling (ASPP) block then processes the features to capture extensive semantic information. Conversely, the decoder processed lower-level features from the 'conv2_block3_2_relu' layer of ResNet50 and then passed through a convolutional block, which used 48 filters with a 1x1 kernel size to refine and integrate textural and edge details. In the DeepLabV3+ model with MobileNetV2 backbone, the encoder collected high-level features from the 'block_4_project_BN' layer, which were subsequently processed using Atrous Spatial Pyramid Pooling (ASPP). Concurrently, the decoder extracted lower-level features from the 'block_2_project_BN' layer and passed through a convolution block equipped with 50 filters of 1x1 kernel. In the DenseNet121-based DeepLabV3+ model, the encoder leveraged higher-level features from the 'conv5_block6_1_conv' layer, while the decoder refined lower-level features from the 'conv2_block4_1_conv' layer. Next, the convolution block with 48 filters of size 1x1 enhances the detail and the texture captured by the features. The encoder extracted higher-level features from the 'block11_sepconv1' layer of the Xception backbone. In contrast, the decoder processed lower-level features from the 'block3_sepconv1' layer, enhanced by ZeroPadding2D, and refined through a convolutional block with 50 filters of 1x1 size.

C. IMPROVING THE BACKBONE DCNN

The main contribution to the DeepLabV3+ backbone is the introduction of the Xception backbone with Lower Features, *XceptionLF*, and the Xception backbone with Lower Features & Overlapping Regions, *XceptionLFOR*. In the XceptionLF backbone, only the lower-level features from Xception were used in both the encoder and decoder. In the encoder, lower-level features were initially extracted from the 'block1_conv2_bn' layer as illustrated in Figure. 2. These features were then first padded using ZeroPadding2D and then upscaled with UpSampling2D to enhance and restore detail. They were further processed through a Conv2D layer with 50 filters of size 3x3, followed by batch normalization and GELU activation, with a subsequent MaxPooling2D step to refine the feature representation. After processing, the Atrous Spatial Pyramid Pooling (ASPP) module processes the features. In parallel, in the decoder, 'block2_sepconv1' outputs lower-level features as an input to a convolution block using 48 filters of size 1x1. This systematic approach to optimizing lower-level features ensures detailed and effective segmentation across the model. The model training with *XceptionLFOR* backbone experiment used the previously described Xception backbone with Lower Features. However, rather than using whole images as inputs, patches from overlapping regions of size 256 were employed at first. Section IV outlines the patch processing steps. The model's predictions on the test set were also performed using these patches. During post-processing, the patches were reassembled into complete images. Figure. 2 illustrates the processing pipeline.

D. SWIN TRANSFORMER AS ENCODER

We integrated the Swin Transformer into the encoder backbone of the DeepLabV3+ pipeline, as illustrated in Figure. 3.

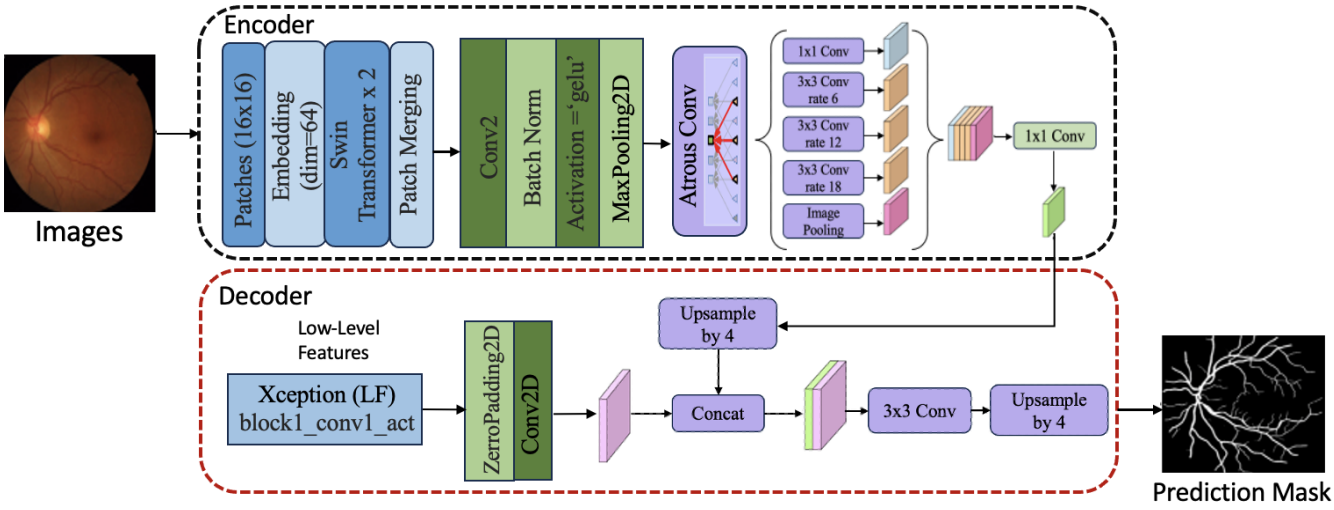

FIGURE 3. Proposed Model Architecture: Integration of Swin Transformer Blocks with Additional Convolution Block in DeepLabV3+ Encoder for Enhanced Retinal Segmentation

The encoder processed the input image by dividing it into 16x16 patches and embedding them. Then, we added two blocks of Swin Transformer before merging the patches. Swin Transformer, implemented from scratch, could capture local and global features through hierarchical self-attention. A block consisting of Conv2D-BatchNorm-GELU activation-Maxpooling layers further refined the features before the decoder extracted low-level features from a pre-trained Xception block. Next, the pipeline merged the features with the encoder output and upscaled through convolutional and upsampling layers. The final output is a prediction mask for segmenting retinal vessels. Combining both approaches leveraged their strengths: the Swin Transformer captures long-range dependencies and multi-scale features, while the Xception decoder efficiently extracts features and ensures precise segmentation.

In summary, we have introduced seven different backbones, four of which followed the DeepLabV3+ architecture and utilized the higher features in the encoder and lower features in the decoder. Our novel contributions include the XceptionLF model, which integrated lower features of Xception with an additional convolution block, and the XceptionLFOR model, an extension of XceptionLF that incorporated patches from overlapping regions to enhance performance. Additionally, we experimented with Swin Transformer with an additional convolution block as a new variant encoder for the DeepLabV3+ model in context to retinal vessel segmentation. These innovations enhance feature extraction and segmentation performance, demonstrating significant improvements in accuracy and efficiency.

IV. PROOF OF CONCEPT

A. BENCHMARKS AND DATA PROCESSING

Figure. 1 illustrates the samples of FIVES datasets. The experiments were conducted using "A Fundus Image Dataset for AI-based Vessel Segmentation," introduced in 2022 [5]. FIVES is one of the most extensive benchmarks for retinal blood vessel image segmentation. The dataset consists of 800 RGB fundus images with pixel-wise annotation. The dataset provider divided the original photos and their corresponding ground truth images into training and testing sets using a 75:25 split. Each image in the dataset has a resolution of 2048x2048 and is categorized into one of four classes with an equal number of samples: healthy, Age-related Macular Degeneration (AMD), Diabetic Retinopathy (DR), and glaucoma. In our experiments, we focused on binary segmentation instead of segmenting based on individual classes. We also evaluate our approach without re-training on the DRIVE dataset [24] to see how it performs in a completely new dataset. The 20 RGB fundus training set of DRIVE was resized to 256×256 to match the training sizes.

We split the training dataset of FIVES [5] into train and validation sets in a ratio of 80:20. As a result, the train set has 480 images, the test set has 200 images and the validation set has 120 images. We then normalized the images and masks to [0,1]. We also applied a few augmentation techniques, such as random horizontal and vertical flips and random rotation within the -30 to 30 degrees range. These pre-processing steps are carried out in all the approaches. However, all the processed images and masks were resized to 256×256 resolution for all experiments except for XceptionLFOR.

We resize the image input to 512×512 pixels for XceptionLFOR. Then, we extract the smaller patches of 256×256 pixels using a stride of 128 pixels to ensure that each patch overlaps half of its predecessor, creating overlapping regions in both horizontal and vertical directions. This results in $\frac{512-256}{128} + 1 = 3$ patches per dimension and a total of nine patches per image. This method deliberately captures sufficient context around the edges of each patch through these overlapping regions, which is crucial for tasks like image seg-

mentation, where preserving edge details is paramount. Both images and their corresponding masks underwent the same procedure. This strategy ensures that the training and validation data are well-prepared, allowing the model to recognize patterns across different scales and conditions efficiently, as illustrated in the initial step of Figure. 2.

B. MODEL EVALUATION METRICS

Several quantitative performance metrics are considered for model evaluations. The definitions are outlined below. The formulas denote the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the evaluation set.

Dice Score is an essential metric for comparing the similarity of two segmentations as it measures the overlap between the predicted and ground truth masks. The formula is:

$$\text{Dice} = \frac{2 \cdot |\text{Ground Truth} \cap \text{Predicted Mask}|}{|\text{Ground Truth}| + |\text{Predicted Mask}|}$$

Precision measures the ratio of accurately predicted positive observations to the total predicted positives. **Sensitivity** is the proportion of accurately predicted positive observations to total observations in the actual class. The precision and sensitivity are two critical metrics for medical image analysis. Higher precision reduces the false positives, which is essential to avoid any false alarms and give more reliable clinical diagnoses. On the other hand, higher sensitivity reduces the false negative, which is crucial for not missing diseases.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}$$

F1 score is the harmonic mean of precision and sensitivity, which balances the trade-off between them:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Accuracy measures the ratio of correctly predicted observation over total observation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

MCC: The Matthew Correlation Coefficient (MCC) balances binary classification by considering all the confusion matrix categories. The MCC provides a more comprehensive evaluation for vessel segmentation as the pixels of vessels are a lot fewer than background pixels.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

C. HYPER PARAMETER TUNING

To get the best hyper-parameter, exhausted hyper-parameter searches are applied in FIVES [5] dataset with different batch sizes (8,16), learning rates (0.0001, 0.001), and optimizers (Adam, AdamW, RMSprop) using the ResNet50 model with 50 epochs. We have obtained the best dice score for batch size 8, Adam optimizer, and 0.001 learning rate. Next, we conduct all experiments with 100 epochs using these hyper-parameters with the dice loss and binary cross-entropy loss combined as the loss function.

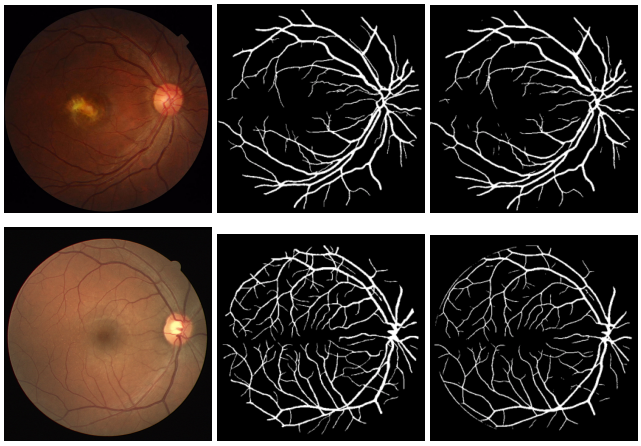

FIGURE 4. Examples of the prediction mask when compared to ground truth for the FIVES (top) image and the DRIVE (bottom) image.

D. COMPUTING

The experiments were conducted using the Google Colab Pro service with 53 GB of system RAM and 22.5 GB of dedicated GPU RAM (L4 GPU configuration).

TABLE 1. Training Time of Different Approaches

Model	Training Time (minutes)
ResNet50	24.25
MobileNetV2	23.63
DenseNet121	23.17
Xception	23.86
SwinTransformer	91.27
XceptionLF	48.95
XceptionLFOR	234.62

The evaluation of training times across different backbones in the DeepLabV3+ segmentation model given in Table 1 indicated substantial differences influenced by their architectural complexities. Base models such as ResNet50, MobileNetV2, DenseNet121, and Xception demonstrated moderate performance with short and comparable training durations—24.25, 23.63, 23.17, and 23.86 minutes, respectively—highlighting their efficiency with the low computational load. The *XceptionLFOR* pipeline used patches from overlapping regions and thus requires a considerably longer training time of almost 4 hours. Still, the XceptionLFOR approach showed the best results on the FIVES benchmark, which confirms that more complex computation does lead to improvement in modeling. Compared to SwinTransformer, which took around 91 minutes to train, the *XceptionLF* modeling displayed a more balanced approach by keeping the training under 50 minutes without compromising its performance. This efficiency made the XceptionLF particularly

appealing in circumstances where resource constraints are a concern.

These findings illustrate a crucial trade-off in deep learning architectures: while advanced features can enhance model capabilities, they also demand more computational resources. Therefore, selecting an appropriate backbone architecture is essential for effectively balancing performance with resource constraints.

V. EXPERIMENTAL RESULTS AND DISCUSSION

All the experimental methods were evaluated on the FIVES [5] and DRIVE [24] test datasets with different metrics. Each experiment in the FIVES test dataset employed a batch size of 8, except for XceptionLFOR, where a batch size of 1 is applied due to the need to reconstruct the prediction patch mask as a whole. Test results on the FIVES dataset are given in Table 2.

TABLE 2. Performances of different approaches(FIVES dataset)

Backbone	Accuracy	Precision	Sensitivity	MCC	Dice
ResNet50	97.46	84.43	78.90	80.26	81.57
MobileNetV2	97.32	83.26	78.10	79.21	80.54
DenseNet121	97.55	86.72	77.53	80.70	81.86
Xception	97.71	88.23	78.33	81.93	83.01
SwinTransformer	98.06	90.21	81.61	84.78	85.62
XceptionLF	98.24	89.60	85.10	86.38	87.21
XceptionLFOR	**98.76**	**90.48**	**91.53**	**90.34**	**89.23**

MobileNetV2, ResNet50, and DenseNet121 gave the lowest performances among all the experiments of different backbones, scoring approximately 81% dice score. Compared to all DCNN backbones with higher features in the encoder and low features in the decoder, Xception performed the best with 97.71% accuracy and 83.01% dice score. The Swin Transformer in the encoder as the backbone is comparable to the XceptionLF and XceptionLFOR with the 85.62% dice score and an accuracy score of 98.06%, almost the same as XceptionLF. The reason for the improved results is that the Swin Transformer can capture both local and global contexts, and the self-attention mechanism can help focus on the essential features of vessels. The performance improved a lot after using Xception with lower features and an additional convolution block (XceptionLF) instead of higher features in the encoder. Dice score and MCC gave an improvement of 4.2% and 4.45%, respectively.

There are several reasons for providing better results for retinal vessel segmentation. Lower features capture fine-grained information such as edges, textures, and small structures that are more critical for segmenting the thin and intricate structures of retinal vessels accurately. On the other hand, high-level features are adequate for understanding the overall image context, but they cannot capture the fine-grained information essential for the vessel segmentation task. Furthermore, low features can preserve more spatial information. The additional convolution block in the encoder also helped to further process and refine those lower features by enhancing their representation. In contrast, higher features usually involve pooling and other operations that can degrade spatial resolution, resulting in the loss of critical details required for precise segmentation. Nevertheless, utilizing low features helped to develop a lightweight model that reduces the computation complexity of training.

The image patching input to the XceptionLF model resulted in a 2.02% improvement in dice score for the test set. Both precision and sensitivity improved a lot for XceptionLFOR by providing 90.48% and 91.53% scores, respectively. Also, the model showed a 90.34% MCC score by outperforming all the other approaches, which indicates that this model is handling imbalance problems more effectively. The reason for the excellent performance is that patch-based training enabled more extensive and varied augmentation, which can help the model reduce over-fitting. The model performed better in segmentation since it is more adaptable to changes in the appearance of the vessel due to its robustness. The 256-resolution patching module captures the local details better to segment delicate vessel structures.

TABLE 3. Performances of different approaches (DRIVE dataset)

Backbone	Accuracy	Precision	Sensitivity	MCC	Dice
ResNet50	95.64	79.99	66.01	70.37	72.33
MobileNetV2	95.38	77.11	66.15	68.95	71.21
DenseNet121	95.60	79.71	65.73	70.01	72.05
Xception	95.70	81.39	65.07	70.54	72.32
SwinTransformer	95.89	78.18	**72.67**	73.14	75.33
XceptionLF	96.03	81.83	69.51	73.32	75.17
XceptionLFOR	**96.39**	**84.06**	71.69	**75.72**	**77.34**

Table 3 outlines the results of the applied methods on the DRIVE test dataset. Base models like ResNet50, MobileNetV2, and DenseNet121 gave competitive accuracy and precision but lag in sensitivity and dice score. On the other hand, the XceptionLFOR model outperforms all other models across most metrics, achieving the highest accuracy 96.39%, precision 84.06%, MCC 75.72%, and dice score 77.34%. It also demonstrates a sensitivity of 71.69%, second only to the SwinTransformer, which has the highest sensitivity of 72.67% but lower performance in other metrics. The model showed over 5.01% dice score improvement against the ResNet50 backbone baseline. The XceptionLF model also performs well in accuracy 96.03% and MCC 73.32%. Overall, the XceptionLFOR stands out for its superior performance in most evaluation metrics, indicating its robustness for the task.

The evaluation of the DRIVE dataset showed that the model is capable of performing well in different unseen retinal blood vessel datasets. Figure. 4 illustrates the output visualization result of FIVES and DRIVE using the best experiment model, XceptionLFOR. Indicate that the predicted mask is efficient in segmenting vessels from images.

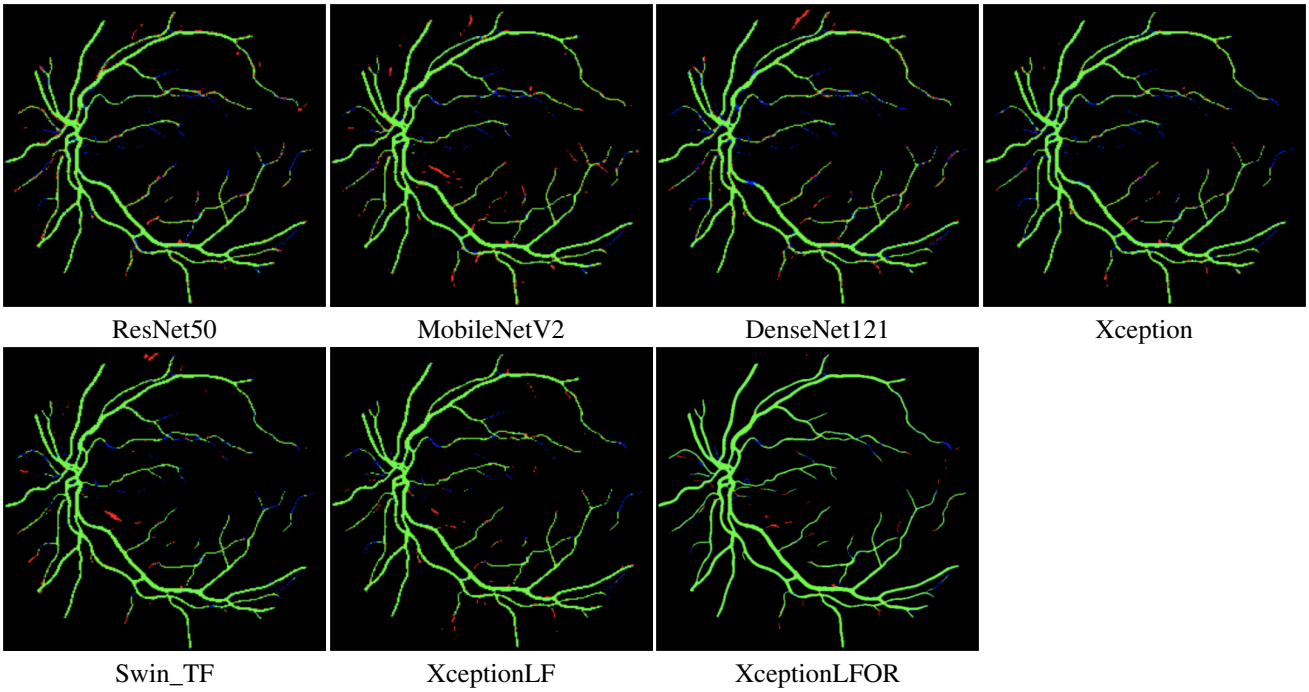

FIGURE 5. TP(green)-vessel correctly classified, FN (blue)-vessel classified as background, FP (red)-non-vessel classified as a vessel with different backbones.

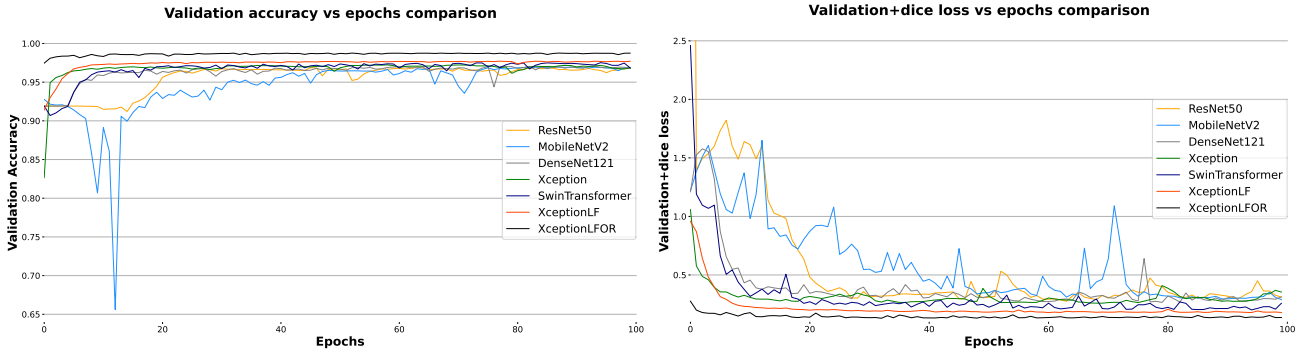

FIGURE 6. Validation accuracy (y-axis) vs. epochs (x-axis) on the left, and validation+dice loss (y-axis) vs. epochs (x-axis) on the right, for proposed approaches (legend).

A. MODEL PERFORMANCE

Figure. 6 shows the accuracy and loss of validation data for the proposed methods for 100 epochs. The DenseNet121 and MobileNetV2 backbone models showed some fluctuations in the plots, which means for some batches, it led to a noisier training update. Therefore, these two models need more fine-tuning. For other experimental models, the validation accuracy and loss have shown more stability, which indicates the model is training well and effectively generalizing the validation set, and the model does not overfit.

We used the color map depicted in Figure. 5 to provide a visual comparison of the predicted test results from our experimental models- ResNet50, MobileNetV2, DenseNet121, Xception, Swin_TF, XceptionLF, and XceptionLFOR. For each model, we use three color codes: green for vessels correctly classified, blue for vessels mistakenly classified as background (false negatives, FN), and red for non-vessels incorrectly classified as vessels (false positives, FP). We observe a noticeable reduction in FP and FN errors as we move from the first image (ResNet50) to the last image (XceptionLFOR) of Figure. 5, indicating an improvement in model performance. The ResNet50 and MobileNetV2 model shows a considerable amount of red and blue, suggesting a higher rate of misclassification. DenseNet121 show a reduction in these errors but still exhibit a significant number of misclassified pixels. The Xception and Swin_TF models further reduce these errors, with fewer red and blue pixels visible. XceptionLF shows even more improvement, but it is the XceptionLFOR model that provides the most refined results. The model significantly minimizes FP and FN errors, with a higher proportion of green pixels indicating correctly classified vessels.

Despite this improvement, some samples still show blue and red pixels, highlighting areas where vessel classification could be enhanced. Overall, the visualization clearly demonstrates the superior performance of the XceptionLFOR model in detecting retinal vessels accurately but also points to the ongoing challenge of perfecting vessel segmentation.

B. COMPARISON WITH STATE OF THE ART

From the related work reviews, we found only two papers that experimented with vessel segmentation using the FIVES dataset. Table 4 shows the evaluation results comparing our best three proposed experimental models with these two papers.

TABLE 4. Comparative analysis with related works

Reference	Approach	ACC	SN	PRE	F1	Dice
Yeganeh et al. [4]	SCOPE	–	85	90	–	85
J. Lin et al. [23]	SGAT-NeT	**98.86**	**91.62**	–	90.51	–
Proposed Model1	SwinTransformer	98.06	81.61	90.21	85.70	85.62
Proposed Model2	XceptionLF	98.24	85.10	89.60	87.29	87.21
Proposed Model3	XceptionLFOR	98.76	**91.53**	**90.48**	**91.00**	**89.23**

The first paper [4] used 512×512 resolution to experiment with their SCOPE models. Their sensitivity score is higher than our Swin Transformer and XceptionLF model; however, our proposed models outperformed other metrics, such as precision and dice scores. Also, our XceptionLFOR outperformed their SCOPE model by a 4.23% dice score. The second paper [23] used 512×512 size of patches from whole images with 2048x2048 resolution to train their SGAT-Net model, which is higher image resolution than our experiment model training. However, our model still could give a higher F1 score (91%) than their model.

Note that our proposed trained models with the FIVES dataset have also been tested with the DRIVE dataset. Still, our models showed effectiveness in this unseen dataset. One related [12] paper applied the DeepLabV3+ model with ResNet18 backbone on the DRIVE dataset and achieved a 66% dice score, which our ResNet50 backbone outperformed by giving a 72.33% dice score. They applied CLAHE for preprocessing and morphological operations for post-processing, which proved less effective compared to our fine-tuned complex model utilizing RGB images. Another paper [9] introduced a lightweight model SVSN, adapted the idea from DeepLabV3+ and achieved 96% accuracy on the DRIVE dataset. Our XpeptionLFOR model achieved nearly the same accuracy despite not training with the dataset. The overall comparison with related works showed our proposed experimental models can efficiently segment blood vessel images.

VI. CONCLUSION AND FUTURE WORK

Retinal blood vessel segmentation is a valuable application in diagnosing ophthalmic diseases. However, due to limited sample availability and image complexity, it remains challenging to achieve efficient automation results using deep learning approaches. To address these challenges, researchers are continuously experimenting with different techniques. In our study, we have experimented with one of the most extensive vessel datasets, FIVES [5], and proposed seven different backbones utilizing the deep semantic segmentation model DeepLabV3+. The six models utilized the DCNN backbones for effective feature extraction. Also, from the related work reviews, it is found none of the studies applied a Swin Transformer with a DeepLabV3+ model for vessel segmentation. The Swin Transformer integrated into the encoder without pre-training delivered comparatively good results. Our best model, XceptionLFOR, with lower features for both encoder and decoder, achieved 98.76% accuracy, 90.34% MCC, and 89.23% dice score. The model also performed well with the unseen dataset DRIVE. Therefore, our proposed methodologies can efficiently segment retinal blood vessels, which will aid in the diagnosis of early eye diseases for patients. In the future, we plan to experiment with different transformers with pre-training and attention mechanisms to evaluate their efficiency in retinal vascular segmentation. We have also applied minimal data pre-processing techniques. The proposed models can test different augmentation approaches and pre-processing procedures.

REFERENCES

[1] Y. Yeganeh, et al., "Scope: Structural continuity preservation for medical image segmentation," arXiv preprint *arXiv:2304.14572*, 2023.

[2] K. Ren, et al., "An improved U-net based retinal vessel image segmentation method," *Heliyon,Heliyon*, vol. 8, no. 10, e11187, 21 Oct, 2022.

[3] R. Liu, et al., "DA-Res2UNet: Explainable blood vessel segmentation from fundus images," *Alexandria Engineering Journal* vol. 68, pp. 539-549, 2023.

[4] R. Wang, et al., "Medical image segmentation using deep learning: A survey." *IET image processing*, vol. 16, no. 5, pp. 1243-1267, 2022.

[5] K. Jin, et al., "Fives: A fundus image dataset for artificial Intelligence based vessel segmentation," *Scientific data*, vol. 9, no. 1, pp. 475, 2022.

[6] T. Akter Tani, "Retinal vessel segmentation," Available: https://github.com/DataLab12/RetinaVseg, Accessed on: June 30, 2024.

[7] M. Z. Alom, et al., "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," arXiv preprint *arXiv:1802.06955*, 2018.

[8] Y. Wu, et al., "Vessel-Net: Retinal vessel segmentation under multi-path supervision," *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China*, October 13–17, 2019, Proceedings, Part I 22, pp. 264-272. Springer International Publishing, 2019.

[9] T. M. Khan, F. Abdullah, S. S. Naqvi, M. Arsalan and M. A. Khan, "Shallow Vessel Segmentation Network for Automatic Retinal Vessel Segmentation," *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1-7.

[10] C. Guo, et al., "Sa-unet: Spatial attention u-net for retinal vessel segmentation," *2020 25th international conference on pattern recognition (ICPR)*, pp. 1236-1242. IEEE, 2021.

[11] W. Liu, et al., "Full-Resolution Network and Dual-Threshold Iteration for Retinal Vessel and Coronary Angiograph Segmentation," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4623-4634, Sept. 2022.

[12] M. C. S. Tang, S. S. Teoh and H. Ibrahim, "Retinal Vessel Segmentation from Fundus Images Using DeepLabv3+," 2022 *IEEE 18th International Colloquium on Signal Processing & Applications (CSPA)*, Selangor, Malaysia, pp. 377-381, 2022.

[13] A. Jayachandran, S. Ratheesh Kumar, and T. Sudarson Rama Perumal, "Multi-dimensional cascades neural network models for the segmentation of retinal vessels in colour fundus images," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42927-42943, 2023.

[14] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang and T. Zhang, "Dual Encoder-Based Dynamic-Channel Graph Convolutional Network With Edge En-

hancement for Retinal Vessel Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 41, no. 8, pp. 1975-1989, Aug. 2022.
[15] Y. Yuan, L. Zhang, L. Wang and H. Huang. "Multi-Level Attention Network for Retinal Vessel Segmentation," in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 312-323, Jan. 2022.
[16] T. M. Khan, et al., "Feature Enhancer Segmentation Network (FES-Net) for Vessel Segmentation," *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 160-167, IEEE, Nov 28. 2023.
[17] S. Deari, I. Oksuz and S. Ulukaya, "Block Attention and Switchable Normalization Based Deep Learning Framework for Segmentation of Retinal Vessels," in *IEEE Access*, vol. 11, pp. 38263-38274, 2023.
[18] H. Xu, and W. Yun, "G2ViT: Graph Neural Network-Guided Vision Transformer Enhanced Network for retinal vessel and coronary angiograph segmentation," *Neural networks : the official journal of the International Neural Network Society*, vol. 176, p. 106356, 2024.
[19] X. Hu, L. Wang, and Y. Li, "HT-Net: A hybrid transformer network for fundus vessel segmentation," *Sensors*, vol. 22, no. 18, p. 6782, 2022.
[20] L. Du, et al., "Deep ensemble learning for accurate retinal vessel segmentation," *Computers in Biology and Medicine* vol. 158, p. 106829, 2023.
[21] H. Zhang, et al., "TiM-Net: Transformer in M-Net for Retinal Vessel Segmentation," *Journal of Healthcare Engineering 2022*, no. 1, p. 9016401.
[22] D. Chen, et al., "PCAT-UNet: UNet-like network fused convolution and transformer for retinal vessel segmentation," *PloS one*, vol. 17, no. 1, p. e0262689, 2022.
[23] J. Lin, et al., "Stimulus-guided adaptive transformer network for retinal blood vessel segmentation in fundus images," *Medical Image Analysis*, vol. 89, p. 102929, 2023.
[24] M. Niemeijer, et al., "Comparative study of retinal vessel segmentation methods on a new publicly available database," *Medical imaging 2004: image processing*, vol. 5370, pp. 648-656. SPIE, 2004.
[25] M. M. Fraz et al., "An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation," in *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538-2548, Sept. 2012
[26] A. D. Hoover, V. Kouznetsova and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," in *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 203-210, March 2000.
[27] LC. Chen, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Proceedings of the European conference on computer vision (ECCV)*,pp. 801-818, 2018.
[28] K. He, et al., "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
[29] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258, 2017.
[30] G. Huang, Z. Liu, L. Van Der Maaten, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708, 2017.
[31] K. Dong, et al., "MobileNetV2 model for image classification," *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, pp. 476-480. IEEE, 2020.
[32] Z. Liu, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012-10022, 2021.

TANZINA AKTER TANI received her B.Sc. degree in Computer Science & Engineering from East West University, Bangladesh, 2021. She is currently pursuing her Ph.D. in Computer Science at Texas State University, USA. Her research interests are computer vision, deep learning, and image processing.

JELENA TEŠIĆ , Ph.D. is an Associate Professor at Texas State University. Prior to that, she was a research scientist at Mayachitra (CA) and IBM Watson Research Center (NY). She received her Ph.D. (2004) and M.Sc. (1999) in Electrical and Computer Engineering from the University of California Santa Barbara, CA, USA, and Dipl. Ing. (1998) in Electrical Engineering from the University of Belgrade, Serbia. Dr. Tešić served as Area Chair for ACM Multimedia 2019-present and IEEE ICIP and ICME conferences; she has served as Guest Editor for IEEE Multimedia Magazine for the September 2008 issue and as a reviewer for numerous IEEE and ACM Journals. She has authored over 60 peer-reviewed scientific papers and holds six US patents. Her research focuses on advancing the analytic application of EO remote sensing, namely object localization and identification at scale.