Compound Scaling Encoder-Decoder (CoSED) Network for Diabetic Retinopathy Related Bio-marker Detection

Dewei Yi*, Petar Baltov, Yining Hua, Sam Philip, and Pradip Kumar Sharma, *Senior Member, IEEE*

***Abstract*— Biomedical image segmentation plays an important role in Diabetic Retinopathy (DR)-related biomarker detection. DR is an ocular disease that affects the retina in people with diabetes and could lead to visual impairment if management measures are not taken in a timely manner. In DR screening programs, the presence and severity of DR are identified and classified based on various microvascular lesions detected by qualified ophthalmic screeners. Such a detection process is time-consuming and error-prone, given the small size of the microvascular lesions and the volume of images, especially with the increasing prevalence of diabetes. Automated image processing using deep learning methods is recognized as a promising approach to support diabetic retinopathy screening. In this paper, we propose a novel compound scaling encoder-decoder network architecture to improve the accuracy and running efficiency of microvascular lesion segmentation. In the encoder phase, we develop a lightweight encoder to speed up the training process, where the encoder network is scaled up in depth, width, and resolution dimensions. In the decoder phase, an attention mechanism is introduced to yield higher accuracy. Specifically, we employ Concurrent Spatial and Channel Squeeze and Channel Excitation (scSE) blocks to fully utilise both spatial and channel-wise information. Additionally, a compound loss function is incorporated with transfer learning to handle the problem of imbalanced data and further improve performance. To assess performance, our method is evaluated on two large-scale lesion segmentation datasets: DDR and FGADR datasets. Experimental results demonstrate the superiority of our method compared to other competent methods. Our codes are available at https://github.com/DeweiYi/CoSED-Net.**

***Index Terms*— diabetic retinopathy, fundus image, lesion segmentation, retinal screening, compound scaling, attention mechanism.**

I. Introduction

The anomalous changes in retina serve as a bio-marker for identification of Diabetic Retinopathy (DR) [1]. For patients with type 1 and type 2 diabetes, DR could lead to irreversible visual impairment in later stages if disease is not identified early and treatments are not instituted in a timely manner [2].

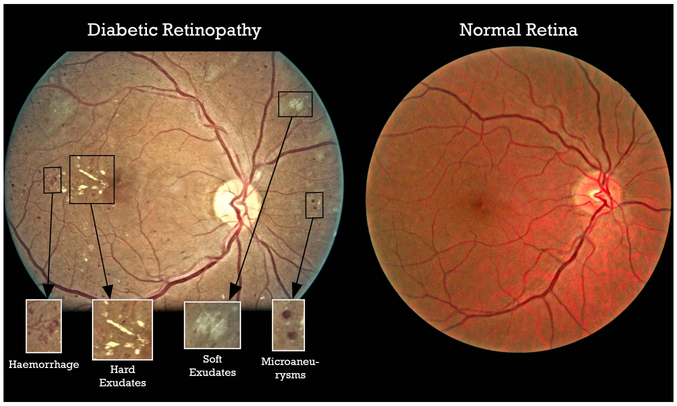


Fig. 1: Comparison of two photographic retinal images. Left retina is with four distinct microvascular lesions (HE, EX, SE, MA) of DR and right retina is a normal retina.

According to the international protocol [3, 4, 5], the severity of DR is graded into 5 different stages (0-4): No Retinopathy (0), Mild Non-Proliferative DR (NPDR) (1), Moderate NPDR (2), Severe NPDR (3) and Proliferative DR (4). This grading scale is identified based on anatomical features of the retina and microvascular lesions, which can be photographically detected by ophthalmologists [6]. Biomedical image segmentation can visualise these lesions, extract quantitative clinical measurements, and aid in the treatment planning, which can improve the interpretability of DR severity made by machine learning models [7]. There are four common microvascular lesions types, which are Hard Exudates (EX), Soft Exudates (SE) also known as Cotton-Wool Spots, Microaneurysms (MA) and Haemorrhages (HE) [8]. These lesions play a critical role in identifying the severity of DR. Given that MA is the first clinically visible evidence of diabetic retinopathy [9], it is an important feature that needs to be identified during DR screening. In the Moderate NPDR (2) stage DR, other features of DR (EX and/or HE) can be observed as presented in Fig. 1. Trained human graders working under the supervision of ophthalmologists are responsible for the identification of different microvascular lesions types and the severity of DR according to their local screening guidelines. Patients who have disease of severity that needs further assessment are referred for appropriate medical and ophthalmic treatments.

Established screening programmes around the world increasingly use digital fundus photography as take one or

This work was supported by Cancer Research UK (CRUK) under Grant EDDPJT-May23/100001.(Corresponding author*: Dewei Yi)

Dewei Yi, Petar Baltov, Yining Hua, and Pradip Kumar Sharma are with the Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK (e-mail: dewei.yi, yining.hua, pradip.sharma, @abdn.ac.uk; peter.baltov99@gmail.com).

Sam Philip is with Department of Diabetes and Endocrinology, NHS Grampian, Aberdeen AB25 2ZN, UK (e-mail: sam.philip@nhs.scot).

more images of each eye with or without routine pupillary dilation [10]. Because microvascular lesions and proportion of those screened with referable disease are small, manual DR identification has a higher risk of misdiagnosis, and to determine the most optimal treatment, the examinations can be time-consuming and labour-intensive [11]. In addition, the localization and type identification of different lesions are challenging tasks because other objects within retinal fundus images may have similar appearances as lesions, such as red dots and blood vessels [12]. Thus, the development of automatic detection systems for DR classification and detection has been area of growing interest to help ease the workload and deliver timely screening to the growing number of people with diabetes worldwide. With the consideration of powerful feature extraction ability, deep learning (DL)-based approaches have attracted great attention from other areas of medical image analysis [13]. Therefore, DL is also recognised as a promising solution for lesion segmentation of diabetic retinopathy [14]. As pointed in [14], since 2015, there is a sharp increasing amount of research output of various DL models proposed for DR domain. As one of the most well-known DL approaches, convolutional Neural Networks (CNNs) transform inputs by convolution filters, and allow weights being shared spatially among the different layers of a given neural network model [14]. As proposed in [15], U-Net is modified to segment different retinal microvascular lesions. Although these conventional proposals have proved the feasibility of detecting retinal microvascular lesions through CNNs, achieving accurate localisation and type identification of lesions in an automatic and efficient manner is still a challenging task.

With the affect of global epidemics, remote healthcare monitoring has attracted great attention due to urgent requirement of accelerating deployment to practical applications [16]. However, it is challenging to make remote healthcare services more affordable, accessible, and effective. Most of deep learning models of medical image applications are large-size models which need to be deployed in high cost computer. Moreover, large-size models are normally deployed in high-performance centre server which brings the challenge of accessibility. Once centre server is attacked, it may lead to service outage for all patients so it is hard to guarantee the sufficient quality of healthcare services [17]. Motivated by the above observations, we propose a novel compound scaling encoder-decoder network for the lesion segmentation of DR, especially in segment small and sparse microvascular lesions. Specifically, the proposed method carries out subtle innovations in the encoder and decoder phases. In the encoder phase, we develop a lightweight encoder architecture utilising a compound scaling coefficient. In the decoder phase, an attention mechanism is introduced, which is realised by using spatial and excitation blocks to improve performance. Moreover, transfer learning is applied to further improve performance when there are insufficient data available. The main contributions of this paper are summarised as below.

- A novel compound scaling encoder-decoder architecture is proposed for microvascular lesion segmentation, which consists of a lightweight encoder and attention-based decoder. The lightweight encoder introduces a compound scaling coefficient, and the attention-based decoder incorporates with Squeeze and Excitation block.
- We introduce a compound loss, which combines dice loss with cross entropy to deal with the class imbalance issue. Moreover, to further improve segmentation performance, transfer learning is applied to alleviate the effect of insufficient data.
- To manifest the superiority of our proposed method, a comprehensive experimental comparison is conducted against other state-of-the-art methods. All of methods are evaluated on two lesion segmentation datasets, including the DDR dataset and the FGADR Dataset. An ablation study is provided to identify the contributions of each innovative component in our proposed method.
- A real-world case study is conducted, where our proposed model is incorporated into a real-time, web-based application that provides information on both lesion segmentation and DR severity classification. The application is based on a server-client architecture, and deployed on a low-cost embedded unit, Jeston NX[1]. With the real-time and low-cost features, our proposed model can assist the screening and assessment of diabetic retinopathy, with the potential to reduce the burden and improve conventional screening efficiency.

II. Related Work

A. Diabetic Retinopathy and Lesion Segmentation

The severity of diabetic retinopathy is assessed on the basis of features on a retinal fundus image. The detection process takes the image as the input, and produces the corresponding severity grade according to the ETDRS protocol [3]. The approach proposed in [18] highlights the most important features by Gradient-weighted Class Activation Mapping (Grad-CAM). More specifically, Grad-CAM produces a coarse localization map. Based on this approach, [19] develops a explainable neural network model to visualise how the DR is located and classified with a given retinal fundus image.

In DR detection, lesion segmentation plays a vital role in determining the severity grade. Some studies are carried out to push the field of semantic segmentation for microvascular lesions, which can assign pixel-wise prediction labels of various microvascular lesions, including HE, EX, SE, MA. They are all characteristics or signs associated with DR, a complication of diabetes that affects the eyes. DR is a condition that damages the blood vessels in the retina, the light-sensitive tissue at the back of the eye. In [15], a modified U-Net architecture is proposed to segment the different types of lesions in retinal images. The modifications are made in both encoder and decoder phases of the U-Net architecture. In the encoder phase, the convolutional layers are changed to 3×3 ResNet convolutional blocks followed by leaky Rectified Linear Unit (ReLU) and batch normalisation layers. In the decoder phase, a deconvolution layer is adopted to replace the

[1] Jeston-NX-`https://developer.nvidia.com/embedded/jetson-xavier-nx-devkit`

normal U-Net upscaling convolutional operation. With the help of these modifications, Dice scores can be improved in HE and EX segmentation. In [5], DeepLabV3+ architecture is adopted to segment four types of microvascular lesions as illustrated in Fig. 1. Similar to U-Net architecture, DeepLabV3+ is also based on an encoder-decoder structure. Different from U-Net, DeepLabV3+ utilises the Atrous convolution layer which enables a further control of over the resolution of the produced feature map. To capture generalised multi-scale information, Atrous convolution layer can also adjust the field-of-view of the filter [20]. However, DeepLabV3+ is struggling to segment Hard Exudates because of the small size of lesions [5]. Inspired by [21], [4] develops the 2-D Dense U-Net architectures for lesions segmentation, which utilises the DenseNet-161 architecture [22] as the encoder of the U-Net for extracting more useful features.

In this paper, all the aforementioned methods, including DeepLabV3, Dense U-Net, U-Net and U-Net++, are implemented and compared with our proposed method on the task of microvascular lesion segmentation.

B. Encoder-Decoder Architectures

Building upon the breakthroughs made by FCNs [23, 24], a U-Net structure is proposed in [25], which attracts great attention on biomedical image segmentation. In biomedical image segmentation, there are lower data availability and pixel-wise prediction is demanded to locate different classes within a biochemical image rather than only classifying an image. The U-Net proposes an encoder-decoder architecture to replace the FCN architecture for enhancing segmentation performance. In U-Net, the network architecture includes a contracting path (i.e., encoder) and a symmetric expanding path (i.e., decoder). The encoder is to extract the most relevant features from an image. The final output of this path is a rich feature map but it is vastly smaller than the original image given that the input has been passed through multiple pooling layers. Once the final layer from the contracting path is reached, the features map is passed to the decoder. The decoder is to process upsampling the learned features. Such a symmetrical architecture matches the total number of layers in both contracting and expanding paths and therefore forms a distinctive U-shape structure. To address the issues of losing important feature information when upsampling a feature map, Ronneberger et al. [25] concatenates the feature map of each layer within the contracting path with its respective symmetric counterpart within the expanding path. This process is also known as a *skip connection* [26]. As a result, such a design preserves both higher and lower resolution features so that better pixel-wise prediction and better localization of different classes can be obtained for an input image.

Inspired by U-Net architecture, [27] proposes a improved version of U-Net to minimise the semantic gap of the encoder and decoder feature map. This is achieved by redesigning the aforementioned skip connections and introducing a series of nested, dense skip pathways that further improve the semantic segmentation outcomes for medical images. The improved U-Net architecture is called U-Net++ [27]. With the help of the modified skip pathways, U-Net++ is able to capture more fine-grained details from an input image, where high-resolution feature maps from the encoder path are gradually enriched prior to fusion with the corresponding feature maps from the decoder path. By learning from interconnected nature of the DenseNet architecture [22], the skip pathways are also introduced in U-Net++, where all layers are connected with each other. Each layer obtains additional inputs from all preceding layers and passes its own feature-maps to all subsequent layers. Specifically, [27] improves skip pathways by concatenating the features extracted from the subsequent layers within the encoder path with the previous layers by up-sampling them.

To further improve the segmentation performance and model efficiency, lightweight encoder and attention-based decoder are proposed and integrated into our method.

C. Model Scaling for Deep Neural Network

As clarified in [28, 29], they find that the performance of network can increase along with scaling up its model, where a network can be scaled up by scaling its depth, width, and resolution. The depth can be scaled up by increasing the number of its layers. The width can be scaled up by growing neurons for each layer. The resolution can be scaled up by enlarging the dimensional size of an input image. Although larger and more complex models can yield better performance, it also significantly increases the number of trainable parameters and lead to higher cost of computational resource. Balancing the scaling of network width, depth, and resolution can achieve better accuracy and efficiency. Such a design enables to easily scale up a baseline model to any target resource constraints so it maximises the running efficiency in deployed devices. That is, it is adaptive strategy to deploy models. Therefore, when the model is deployed in a low-cost device, the design can optimise the use of computational resource. To achieve this, [29] propose a Compound Scaling method. This method efficiently scales up a model by using a *compound coefficient* ϕ. The compound coefficient is regarded as a hyperparameter and it is used to determine how many more resources are available for model scaling. Moreover, [29] transforms the numerical values associated with the aforementioned depth, width and image size into variables. These variables are assigned as α for depth, β for width and γ for image size. The compound coefficient ϕ is applied to these variables as their power terms. This can efficiently balance depth, wight, resolution of a network by scaling each of them with a constant ratio. Therefore, [29] propose a neural network architecture that takes advantage of this compound coefficient and it is called EfficientNet. This network architecture builds upon MnasNet [30]. Its fundamental building block is a mobile-inverted bottleneck convolutional layer (MBConv) that has been adapted from the neural network architecture proposed in [31], where the normal convolutional layers are changed to a depth-wise spatial convolutions with the addition of inverted residual bottleneck layers. Moreover, [29] add a squeeze-and-excitation operation [32]. The pooling and fully-connected layers of network are the same as a traditional CNN architecture. All these modifications have been utilised in the creation of the baseline model called EfficientNet-b0 and the subsequent EfficientNet-b1 to b7 [29] have been proposed based on changes made to the compound coefficient ϕ and the variables α, β, γ. The number of parameters of EfficientNet-b0 to EfficientNet-b7 is ranging from 5.3M to 66M.

To further improve the efficiency, [33] proposes an improve version of EfficientNet. This neural network model improves on its predecessors with faster training and 6.4 times fewer number of trainable parameters. To achieve this, [33] identifies three problems associated with their previous architecture and they are as follows. The first problem is related to the resolution (size) of the input image. It finds that larger input sizes considerably decelerates the training process of the network. They have identified that smaller input size leads to fewer computational operations and enables larger batch sizes to be applied in the training process, thus improvements are made on the training speed by up to 2.2 times. The second problem is associated to the aforementioned depth-wise convolutional layers. It finds that these types of convolutional layers considerably slow down the training process of the model in the early stages but are regarded as relatively effective in later stages of training. To address this problem, a Fused-MBConv is proposed which is modified MBConv Layer [33]. Such a design combines the depth-wise convolutional layer with a normal 3x3 convolutional layer. The third problem is related to uniformity of scaling the aforementioned variables α, β, γ with the compound coefficient ϕ. It is more efficient to scale this parameter in a non-uniform strategy meaning that not all variables would be changed at the same time. Furthermore, [33] restricts the resolution variable(γ) to a smaller numeric value to combat the first encountered problem of input size. These modifications derives EfficientNetV2. This architecture is composed of both Fused-MBConv and normal MBconvs layers. Fused-MBConv layers have been utilised in the early stages of the model and MBconvs have been applied in later stages so as to deliver better performance.

Taking into the strengths of EfficientNet and its variants, we introduce compound scaling scheme into the encoder of our method so as to develop a lightweight encoder.

D. Attention Mechanism in Deep Neural Networks

Attention mechanisms are helpful to capture the fine-grained features in the diabetic retinopathy detection. In [34], a Category Attention Block is proposed which can significantly improve the performance of diabetic retinopathy grading. In addition, [32] proposes a channel-wise attention mechanism, where a "Squeeze-and-Excitation" (SE) block is introduced into network. The SE block adaptively re-calibrates channel-wise feature responses by explicitly assigning additional weights for the different input image channels. Such a design can provide more attention and prioritise one channel over the other. To combat the locally learned features and further improve the model's performance, the Squeeze-and-Excitation block first squeezes the global spatial information represented by the dimensions H×W×C (H-height, W-width, C-channels) into a channel-wise vector with a size of 1×1×C (C is the total number of channels). This is achieved through an aggregation strategy, where global average pooling produces a 1×1×C vector, which is the average numerical value for every channel. Then, the vector is excited by being passed through two fully connected layers with ReLU as activation function. The first fully connected layer is to reduce the dimensionality of the vector by the parameter r. The output of the first fully connected layer is passed through the ReLU operation and then to the second fully connected layer which increases the dimensionality of the vector to its original size. After that, the vector is passed to a sigmoid activation operation to normalise values within the vector between 0 and 1 [32]. Finally, the vector is multiplied with the initial H×W×C feature map.

Inspired by [32], a improved version of SE block is proposed in [35], which introducing an additional convolutional operation with a 1×1 filter creating an additional feature map. The produced feature map is passed through a sigmoid activation operation like SE block for further adding weights to the different spatial regions within the input image and so is called Spatial Squeeze and Excitation Block (sSE). Moreover, sSE block is further re-calibrated and concatenated with the output of the conventional channel-wise SE (cSE) block to form scSE Block, which is called Spatial and Channel "Squeeze and Excitation" Block.

To fully use spatial information along with channels, our method introduces Spatial and Channel Squeeze & Excitation Block into the decoder part to recalibrate feature maps both channel-wise and spatially so as to enhance the segmentation performance.

III. Compound Scaling Encoder-Decoder Network (CoSED-Net)

A. Architecture of CoSED-Net

This section provides a comprehensive and detailed explanation of our proposed neural network architecture, its building blocks, layers and further modifications. The overview of the proposed network architecture is provided in Fig. 2, where an encoder-decoder structure is adopted in lesion segmentation. More specifically, a lightweight encoder is proposed to relieve the problem of insufficient data. This achieves by introducing model scaling, where a compound scaling scheme is adopted in the encoder. In addition, an attention-based decoder is adopted to extract both spatial and channel information in a more efficient way, which is realised by introducing spatial and channel SE block. Moreover, a compound loss function is proposed to tackle the data imbalance problem. Furthermore, Transfer Learning is also used in our method to further enhance the performance.

In Fig. 2, it provides the flow of processing a retina image throughout our proposed architecture. The input image is with size of H×W×C, which is resized to 512×512×3 after data preprocessing. Then, the first convolutional layer lowers the feature map by the stride factor of 1 and the rest of following convolutional layers are to lower the feature map by the stride factor of 2. The number of channels is increased with passing more convolutional layers. The size of the final feature map is 16×16×256. For the attention-based decoder,

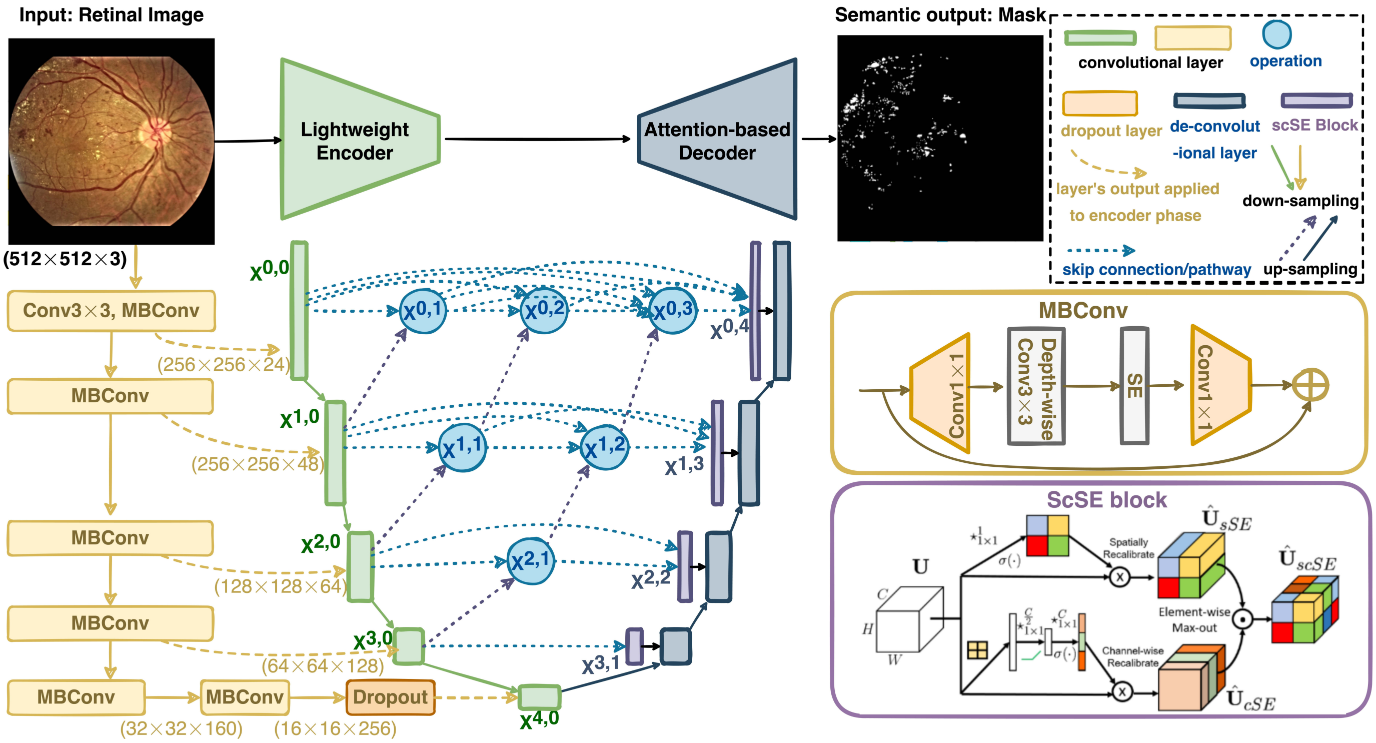


Fig. 2: The overview of the our CoSED-Net architecture. On the left-hand side, the Lightweight Encoder accepts the input image and on the right-hand side the attention-based decoder architecture is shown and outputs the given segmentation mask.

the dimensions of the feature map are up-scaled by a factor of 2 and concatenated with the feature from the scSE blocks on different levels as seen in Figure 2. With the execution of the first deconvolution layer and final encoder layer, the feature map is upscaled by a factor of 4. The final output of the encoder-decoder architecture has the dimensions of 512×512×1 to provide pixel-wise prediction in a mask.

B. Lightweight Encoder

Inspired by [25, 27], there is also an encoder in our network. Different from [25, 27], a lightweight encoder is proposed into our network, which is achieved by introducing compound scaling scheme and mobile inverted bottleneck Convolutional (MBConv) network. The lightweight encoder composite of MBConv, Conv, and dropout layers. To achieve better efficiency and accuracy, we carry out compounding scaling for network. Different from previous studies which mainly focuses a single dimension to scale up network, all three dimensions of our encoder network are scaled up, including depth, width, and resolution. Scaling the depth of encoder is to extract richer and more complex feature so as to generalise well on new tasks. Scaling the width of encoder is to obtain more fine-grained features, which also make train network easier. Scaling the resolution of encoder is to capture more fine-grained patterns. Following [29], our lightweight encoder can be defined as:

$$\begin{gathered} N(d,w,r) = \bigodot_{i=1,\ldots,s} F_i^{d\cdot L_i}(X_{<r\cdot H_i, r\cdot W_i, w\cdot C_i>}) \\ d=\alpha^{\phi}, w=\beta^{\phi}, r=\gamma^{\phi} \end{gathered} \tag{1}$$

where $F_i^{L_i}$ represents that the layer of F_i is duplicated L_i times with i-th stage and $\bigodot$ represents repeated product. $< H_i, W_i, C_i >$ is the input size of X for i-th layer. w, d, r are coefficients for scaling the width, depth, and resolution, respectively. Specifically, to scale up the width, depth, and resolution of network in an homogeneous way, our method conducts compound scaling as follows.

$$\begin{aligned} &\forall\ \alpha^{\phi}, \beta^{\phi}, \gamma^{\phi} \ni \alpha\cdot\beta^2\cdot\gamma^2 \approx 2 \\ &where \quad \alpha \geq 1,\ \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{2}$$

where the depth, width, and resolution are presented by α^{ϕ}, β^{ϕ}, and γ^{ϕ}, respectively. The α, β, and γ are constants that can be determined by a small grid search and ϕ is a user-specified coefficient. $\ni$ is the symbol standing for "subject to". Intuitively, ϕ controls how many more resources are available for model scaling and α, β, γ specify how to assign these extra resources to the depth, depth, and resolution of network.

In addition, MBConv [29] is adopted to composite our lightweight encoder. The operation of MBConv can be expressed as a composition of three operators as follow.

$$\begin{aligned} F(x) = [A \circ N \circ B]x& \\ A: R^{s\times s\times k} &\to R^{s\times s\times n} \\ N: R^{s\times s\times n} &\to R^{s'\times s'\times n} \\ B: R^{s'\times s'\times n} &\to R^{s'\times s'\times k'} \end{aligned} \tag{3}$$

where A and B are linear transformations. N is a channel-based non-linear transformation.

Such a design can significantly decrease the number of network parameters which also help address the issues of deploying the architecture in a limited resource environment such as a mobile or server platform [33]. As seen in Fig. 2, the feature map from the different intermediate layers of the lightweight encoder is highlighted by the arrow symbol and applied in the encoder. Additionally, the encoder is concatenated with an additional dropout layer to avoid overfitting and improve generalisation error, where neuron outputs are dropped out randomly. After every convolutional layer, ReLU activation is applied with Batch-Normalization to rescale and standardise the output of a layer by maintaining the mean of a batch close to 0 and its standard deviation close to 1, which can avoid to get stuck in local minima. Once the input is passed though all the convolutional layers of our lightweight encoder, feature map is passed to decoder.

C. Attention-based Decoder

In the decoder, feature map is upscaled and the intermediate features of the skip pathways are concatenated until the final convolutional layer is reached. In the final covolutional layer, there is a 1x1 filter preforming pixel-wise classification on the final feature maps so as to create the final binary segmentation mask. In our method, an attention-based mechanism is adopted, where skip pathways are modified by introducing the Spatial and Channel "Squeeze and Excitation" (scSE) Block [35]. Such a block can provide the model with an additional attention mechanism, which prioritises different spatial regions alongside different channels. As mentioned in [35], scSE block can substantially increase network performance by only a slight increase of model complexity. scSE block is consisted of cSE block and sSE block. Here, the input feature map combines channels together, which is denoted as $F = [f_1, f_2, \dots, f_C]$ and $f_i \in \mathbb{R}^{H \times W}$. In cSE block, spatial squeeze is performed by a global average pooling layer to produce below vector z of k-th element.

$$z_k = \frac{\sum_i^H \sum_j^W F_k(i,j)}{H \times W} \tag{4}$$

where u_k is a feature map and (i, j) is a spatial location for a channel. The H and W are its the height and width of the feature map. This operation is to attain spatial features globally. With $z \in \mathbb{R}^{1 \times 1 \times C}$, the vector z_k can be changed to $\hat{z}$ by $W_1(\delta(W_2^z))$ and $W_1 \in \mathbb{R}^{C \times \frac{C}{b}}$ represents the weights between fully-connected layers The operator of ReLU is denoted as $\delta(\cdot)$. The bottleneck of channel excitation is determined by b to encode the dependencies of channels. Referring to [35], r is set to 2. The value of the activation of $\hat{z}$ dynamically changes within [0, 1] with using a sigmoid layer $\sigma(\hat{z})$. To recalibrate or excite U, the resultant vector in Eq. (5) is used.

$$\hat{F}_{cSE} = [\sigma(\hat{z_1}) f_1, \sigma(\hat{z_2}) f_2, \dots, \sigma(\hat{z_C}) f_C] \tag{5}$$

where the importance of the i-th channel are specified by the activation $\sigma(\hat{z}_i)$, which is adaptively tuned to pay attention more important channels.

In sSE block, the feature map F is squeezed along the channels and excites spatially. This is achieved by introducing the channel squeeze and spatial excitation block. The spatial squeeze operation is significant for fine-grained lesion segmentation. The slicing of feature map F is denoted as follows.

$$F = [f^{1,1}, f^{1,2}, \dots, f^{i,j}, \dots, f^{H,W}] \tag{6}$$

where $f^{i,j}$ is the feature of the spatial location (i, j) and $f^{i,j} \in \mathbb{R}^{1 \times 1 \times C}$. The spatial squeeze operation is achieved through a convolution

$$q = W_{sq} * F, \quad W_{sq} \in \mathbb{R}^{1 \times 1 \times C \times 1} \tag{7}$$

where W_{sq} is the weight of spatial squeeze. q is a project tensor generated by sSE block, which belongs to $\mathbb{R}^{H \times W}$. The projection $q_{i,j}$ is a linear combination of all channels C with regard to the spatial location of (i, j). A sigmoid layer $\sigma(\cdot)$ is used to generate the projection so as to recalibrate or excite F spatially within the range of [0, 1] as below.

$$\hat{F}_{sSE} = [\sigma(q_{1,1}) f^{1,1}, \dots, \sigma(q_{i,j}) f^{i,j}, \dots, \sigma(q_{H,W}) f^{H,W}] \tag{8}$$

where the relative importance of a spatial information is denoted by $\sigma(q_{i,j})$. (i, j) represents the location of feature map. Such type SE block ensures that more attentions are paid to relevant spatial location.

In addition, we adopt the concatenation aggregation to concatenate both outputs from the Spatial SE and channel SE layers to further enhance segmentation performance, where the two outputs are concatenated along with the channel index. The operation of concatenation aggregation can be defined by

$$\hat{F}_{scSE} = concat(\hat{F}_{cSE}, \hat{F}_{sSE}) \tag{9}$$

where $\hat{F}_{cSE}$ represents spatial squeeze and channel excitation block and $\hat{F}_{sSE}$ represents channel squeeze and spatial excitation block. Such an aggregation strategy can avoid information loss. As clarified in [27], we find that dense skip pathways bring the intermediate feature maps of both encoder and decoder paths on different layers to be semantically similar. Therefore, we also adopt dense skip pathways to simplify network optimisation and strengthening prediction accuracy.

D. Loss Functions

To optimise the network weights, Dice loss and cross entropy loss (CE) are two commonly used loss in medical image due to its compelling performance in image segmentation. For Dice loss function [36], it is based on the Dice score coefficient. This metric calculates overlapping between predictions with the ground truth. Dice loss shows a great performance on medial MRI image datasets while it has the nature of non-convex. In a non-convex problem, it is easy to be stuck at locally optimisation.

Taking this into account, CE loss is introduced into lesion segmentation as well [4]. When Cross-Entropy Loss is applied in the pixel-wise classification of an image, each pixel is classified based on a given ground truth segmentation mask. In lesion segmentation, Cross-Entropy Loss calculates error between the pixel-wise predictions made from the model with the pixel-level ground truth masks. However, CE Loss cannot perform well when data has a higher class imbalance [37]. In lesion segmentation, there are much less foreground data (i.e.,the specific lesion outlines) compared to background data. Therefore, a compound loss function is introduced, which combines both Dice loss and CE loss for achieving a synergy effect. The compound loss is defined as follows.

$$\begin{aligned} L_{DCE} &= L_{DL}(y,p) + L_{CE}(y,p) \\ &= 1 - \frac{2yp+1}{y+p+1} - (y\log(p) + (1-y)\log(1-p)) \end{aligned} \tag{10}$$

where the variable y represents the pixel-wise targets and p represents the predicted values produced by the neural network with the addition of log operations that further contributes to the final error calculations. The difference comes with the addition of the number 1 in both the numerator and denominator of the division to mitigate the edge case of zero division. To achieve a harmonic effect of DL and CE, the weights of these losses are identical and both are set to 1.

E. Transfer Learning

Transfer learning is used in our method to further address the low data availability problem in DR lesion segmentation. As clarified in [38], transfer learning is able to provide tremendous improvements in diabetic retinopathy detection. Different from conventional transfer learning which transfer the knowledge of different datasets but on the same task, our transfer learning is to transfer the knowledge between different tasks, i.e., from image classification to semantic segmentation. Specifically, we transfer a pre-trained EfficientNet model, which is trained on the ImageNet [39, 40] of image classification, to the task of DR lesion segmentation. Following [4], transfer learning is conducted by removing the last classification layer from the pre-trained models so that only the fine-grained feature maps are extracted and the final feature map from the EfficientNet structure is passed to the attention-based decoder.

IV. Experimental Evaluation

This section conducts experiments for lesion segmentation. The experiments are evaluated by two large-scale diabetic retinopathy data: DDR dataset and FGADR dataset. These two datasets are used to train, validate and test our proposed network architecture and other compared methods. The details of these two datasets are described in Section IV-A. The evaluation metrics are presented in Section IV-B. The implementation details are discussed in Section IV-C, where the settings of our network are provided along with the explanation of platforms and configurations. To quantitatively and qualitatively evaluate lesion segmentation performance, our CosedNet method is compared against other methods with regard to IoU, Dice Score, and mIoU in Section IV-D.

A. Datasets

1) DDR Dataset: In DDR dataset [5], the image data were collected from 147 hospitals between the period of 2016 to 2018. These images are provided by 9598 patients whose ages ranged from 1 to 100, with an average age of 54.13 with a relatively equal split between male and female contribution. The dataset provides pixel-level lesion annotations and bounding-box annotations of lesion segmentation. We utilises the images and their relative pixel-level annotation masks for the task of lesion segmentation.

For lesion segmentation of DDR dataset, there are 757 fundus images with their annotated masks that represented the four different lesions such as EX, SE, MA and HE. Six annotators perform pixel-level annotations for the four lesions, using annotation software and an additional annotation tool has been used to automatically identify the lesions. Furthermore, the annotators manually annotate lesions that are not identified by the annotation tool and verify if the different lesions are correctly outlined by removing false positives. The DDR dataset is randomly split into training, validation and testing set in the ratio of 5:2:3 respectively.

2) FGADR Dataset: FGADR dataset [4] is another dataset for DR lesion segmentation to evaluate the models' performance. In the FGADR dataset, it consists of 1842 images with their corresponding pixel-level annotation masks. The range of lesions includes EX, SE, MA, HE but also intra-retinal microvascular abnormalities (IRMA) and neovascularization (NV). With considering uniformity and consistency, we focus on the four aforementioned lesions (EX, SE, MA, HE). This dataset focuses on higher quality images with a goal of building a diverse range of lesion representation. Thus, three board-certified ophthalmologists with intra-rater consistency manually annotated above images with strict quality control.

Following the settings of DDR dataset, we also split the FGADR dataset into a ratio of 5:2:3. That is, there are 50%, 20%, and 30% of randomly split into the training set, validation set, and test set.

B. Evaluation Metrics

In this section, various metrics are used to assess the performance of lesion segmentation, including Accuracy, Precision, Recall, Dice Score, and Intersection over Union (IoU). The IoU is also called Jaccard Index, which is defined as the area of intersection between the predicted segmentation mask and the ground truth mask, divided by the area of the union between the two maps [41]. The range of this metrics is between 0 and 1. Dice Score is defined as twice the number of overlapping area of the predicted and ground-truth maps divided by the total number of pixels [41].

The additional used metrics are accuracy, precision and recall. The precision metric indicates what fraction of the total prediction made by the model are correct, and the recall metrics quantifies the number of correct positive predictions made out of all positive predictions [12]. The following formulas show how these fractions are used to calculate the metrics, note that for the task of lesion segmentation, Dice Score can be expressed via Precision and Recall, which is defined as follows.

$$Dice-score = \frac{2|A \cap B|}{|A|+|B|} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

TABLE I: The performance of lesion segmentation on validation set of DDR dataset with regard to IoU, Dice Score, and mIoU.

Model Name	**Hard Exudates (EX)**		**Haemorrhages (HE)**		**Microaneurysms (MA)**		**Soft Exudates (SE)**		**mIoU**
	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	
Cosed-Net (Ours)	45.56	62.60	40.31	57.46	13.98	24.53	30.10	46.28	32.49
Dense U-Net [21]	38.39	55.49	19.33	32.40	6.16	11.60	24.79	39.73	22.17
U-Net++ [27]	32.69	49.27	17.81	30.24	3.85	7.42	22.56	36.81	19.98
U-Net [25]	29.70	45.80	13.02	23.04	4.36	8.36	24.29	39.08	17.84
DeepLabV3+ [20]	29.10	-	28.19	-	4.29	-	28.19	-	22.29
HED [5]	9.48	-	21.83	-	2.04	-	21.83	-	9.29

TABLE II: The performance of lesion segmentation on test set of DDR dataset with regard to IoU, Dice Score, and mIoU.

Model Name	**Hard Exudates (EX)**		**Haemorrhages (HE)**		**Microaneurysms (MA)**		**Soft Exudates (SE)**		**mIoU**
	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	
Cosed-Net (Ours)	39.87	57.01	24.10	38.84	10.34	18.74	20.88	34.54	23.80
Dense U-Net [21]	35.67	52.58	13.03	23.05	6.84	12.8	20.68	34.27	19.06
U-Net++ [27]	27.58	43.23	11.93	21.32	3.85	7.42	14.88	25.90	14.56
U-Net [25]	29.90	46.03	13.39	23.61	3.53	6.82	14.59	25.47	14.60
DeepLabV3+ [20]	31.18	-	14.25	-	3.25	-	22.95	-	17.91
HED [5]	18.74	-	5.24	-	1.10	-	7.82	-	8.83

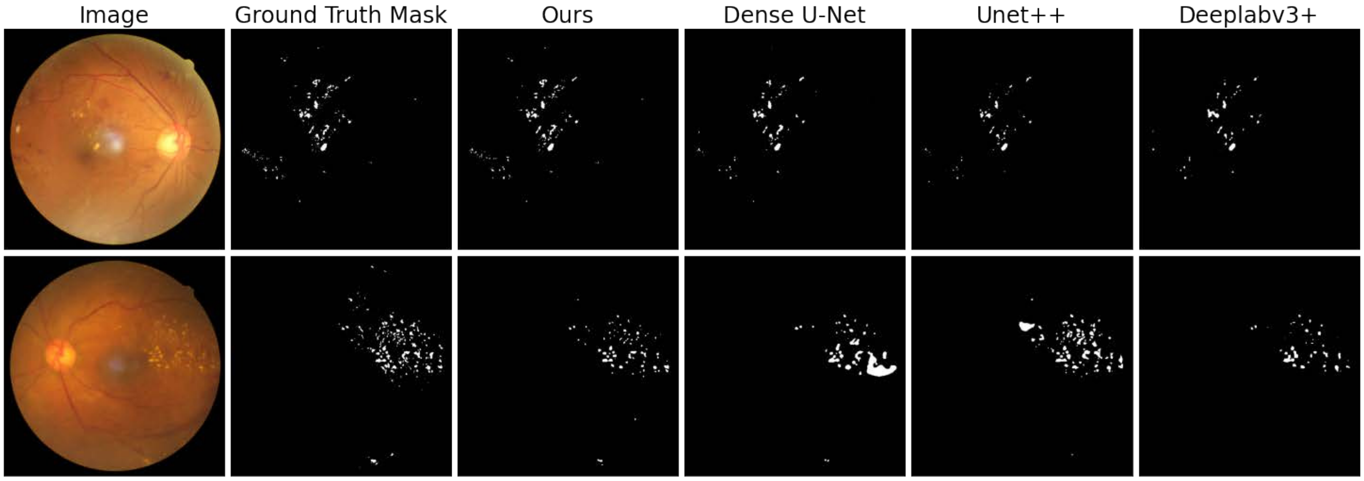


Fig. 3: Qualitative Comparison between our proposed method and other advanced methods on the test set of DDR dataset.

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{12}$$

$$IoU = Jaccard(A, B) = \frac{area(A \cap B)}{area(A \cup B)} \tag{13}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. A is predicted segmentation mask and B is the ground truth mask.

C. Implementation Details

In this section, we cover the different aspects of implementing the different experiments such as different frameworks, libraries, dependencies and further hardware specifications. Additionally, we demonstrate the different stages, including the training, validation and testing stages alongside the data pipeline, data pre-processing and further augmentations. All models are implemented under PyTorch Lightning[2], which is an open-sourced machine learning framework. The framework improves the PyTorch framework by providing additional scalability, optimisation and better development of different neural network architectures by abstracting and organising a majority of the boilerplate code related to the PyTorch framework. Furthermore, it provides further utilities and better organisation of the training, validation and testing loops.

In the experiments, data augmentation is carried out before training lesion segmentation model. As suggested in [11], various pre-processing techniques on the fundus retinal images are evaluated for combating the problem of non-uniformed data collection based on the use of different equipment for the screening process of diabetic retinopathy. These techniques can reduce noise by applying adaptive noise removing filters, illustrate correction via homomorphic filtering, and techniques for contrast enhancement. As clarified in [11], contrast limited adaptive histogram equalization (CLAHE), which is a contrast

[2] PyTorch Lightning-https://www.pytorchlightning.ai/

enhancement technique, can significantly improve models' performance on the task of vessel segmentation. After preprocessing data, image augmentation is conducted where we applied CLAHE [11] to augment raw image data with considering the balance between training speed and performance. After this, the data is trained in batches and training batch size is set to 8. The network predicts the segmentation masks. The loss function calculates the error between the predicted masks and the ground truth masks. Such a error is passed to optimizer for updating the network weights. To have a fair comparison among different networks, we configure the same hyperparameters for all the neural networks when they are trained on the DDR and FGADR datasets. Following [4], ADAM is used to optimise the network and initial learning rate is set to 0.001. Moreover, `ReduceLROnPlateau` is used as learning rate scheduler to dynamically adjust the learning rate. When model performance stop improving or plateaus for a certain number of epochs during training, the learning rate is reduced by a factor. This reduction helps the model to fine-tune its parameters more delicately and potentially escape local minima. In this paper, all schedulers are set to monitor the training loss and for every 1 epoch by setting corresponding the patience parameter. The decreasing factor of learning rate is set to 0.1. The minimum learning rate for every model is set to 8e-05. The probability of dropout within the dropout layer has been set to 0.5. The maximum epoch number is set to 300 for training DDR and FGADR datasets. Transfer learning is also performed by applying a pre-trained EfficientNet, that have been previously trained on the ImageNet dataset [39].

D. Comparisons With the State-of-The-Art Methods

To evaluate the performance, our method is compared with other advanced methods on a segmentation task. Specifically, four advanced methods are compared, including DeepLabV3 [20], U-Net [25], U-Net++ [27], and Dense U-Net [21]. These methods are evaluated on both DDR and FGADR datasets on four the lesions segmentation tasks, i.e., EX, HE, MA, and SE. Then, experimental results on DDR and FGADR datasets and their corresponding observations is provided as follows.

For the performance of lesion segmentation on the DDR dataset, the result of our method is significantly better than other methods. The qualitative results of lesion segmentation are illustrated in Fig. 3 and the quantitative comparison is summarised in Table. I and II. From Table I, Table II, and Fig. 3, the following observations can be drawn:

- Our proposed method can achieve the best performance with regard to mIoU. where its mIoU can reach 32.49% and 23.80% for validation set and test set, respectively.
- For the lesion segmentation tasks of EX, HE, MA, and SE, our proposed method almost provides the best results for each of them in terms of Dice score and IoU on both the validation set and test set of DDR dataset.
- For the lesion segmentation results of test set, we find that our method significantly outperforms other advanced methods in terms of IoU and Dice Score. Our method can achieve 39.87% of IoU and 57.01% of Dice Score for EX, 24.10% of IoU and 38.84% of Dice Score for HE, 10.34% of IoU and 18.74% of Dice Score for MA, and 20.88% of IoU and 34.54% of Dice Score for SE.

To have a comprehensive comparison, the experiments on the FGADR dataset are also carried out. For the following experiments, our method is compared to other four advanced methods on the FGADR dataset. All of the four lesion segmentation tasks are considered to evaluate the performance on diabetic retinopathy, including EX, HE, MA, and SE. As seen in Table III, Table IV, and Fig. 4, the following observations can be drawn:

- Our proposed method also outperforms other methods in terms of mIoU. Specifically, its mIoU of EX, HE, MA, and SE can achieve 23.36% and 24.81% for validation and test sets, significantly better than other methods.
- For the lesion segmentation tasks of EX, HE, MA, and SE, our proposed method provides the best results for each them in terms of Dice score and IoU on both validation set and test set of FGADR dataset.
- For the much more challenging segmentation task of MA and SE, our method can perform much better compared to the other methods, where the IoU of MA and SE can reach 2.84% and 22.21%, and the Dice score of MA and SE can reach 5.52% and 36.35% on FGADR test set.

E. Ablation Study

In the ablation study, we analyse the contributions of each innovative components in our proposed method. Since FDADR dataset twice larger than DDR dataset, Extensive experiments are conducted on FDADR dataset to resolve the how much does each design component contribute on the improvement of overall performance. The experiments for ablation study are focused on Hard Exudates for the sake of brevity. By adopting one more components at each stage, the improvement of IoU and Dice Score (Dice) is presented in Table V, where the network structure of encoder-decoder is denoted as ED, lightweight encoder is denoted as LE, the compound loss of dice loss and cross entropy is denoted by DCE, transfer learning is denoted by TL, and Attention-based Decoder is denoted as AD.

According to Table III and IV, we can find that U-Net++ provided better results than U-Net and DeepLabV3+. Therefore, the encoder-decoder structure of the U-Net++ network is adopted as baseline model. To derive a lightweight encoder, the encoder phase of the U-Net++ architecture is modified by introducing compound scaling scheme. Specifically, MBconv layers are integrated into encoder. We observe significant increase in both IoU and Dice score, where the performance gain of IoU and Dice score are 1.71% and 2.07%.

Moreover, a compound loss function DCE is adopted to optimise the network weights which combines Dice loss and cross entropy to achieve a synergy effect. Specifically, such a design can solve the issue of imbalance class, alleviate to be trapped into local optimization, and further enhance the segmentation performance. The performance gain of adopting DCE is 2.04% and 2.41% for IoU and Dice, respectively. Then, Transfer Learning is used to accelerate the training time and overcome the issue of insufficient data. It yields a significant

TABLE III: Lesion segmentation performance on validation set of FGADR dataset with regard to IoU, Dice Score, and mIoU.

Model Name	Hard Exudates (EX)		Haemorrhages (HE)		Microaneurysms (MA)		Soft Exudates (SE)		mIoU
	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	
Cosed-Net (Ours)	36.73	53.72	36.97	53.99	1.79	3.52	17.94	30.43	23.36
Dense U-Net [21]	28.57	44.44	28.21	44.01	0.76	1.51	12.08	21.55	17.42
U-Net++ [27]	26.58	41.99	28.74	44.64	0.72	1.44	12.70	22.54	17.15
U-Net [25]	29.25	45.26	25.63	40.81	0.37	0.74	11.13	20.03	16.60
DeepLabV3+ [20]	32.32	48.85	24.90	39.88	0.09	0.18	11.01	19.83	17.08

TABLE IV: Lesion segmentation performance on test set of FGADR dataset with regard to IoU, Dice Score, and mIoU.

Model Name	Hard Exudates (EX)		Haemorrhages (HE)		Microaneurysms (MA)		Soft Exudates (SE)		mIoU
	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	IoU	Dice Score	
Cosed-Net (Ours)	39.55	56.68	34.62	51.43	2.84	5.52	22.21	36.35	24.81
Dense U-Net [21]	34.19	50.96	26.49	41.89	1.38	2.72	15.83	27.33	19.47
U-Net++ [27]	27.66	43.33	27.75	43.44	1.98	3.89	17.94	30.42	18.83
U-Net [25]	29.09	45.07	24.79	39.73	1.39	2.75	16.47	28.28	17.94
DeepLabV3+ [20]	31.25	47.62	25.06	40.07	0.79	1.58	14.42	25.2	17.88

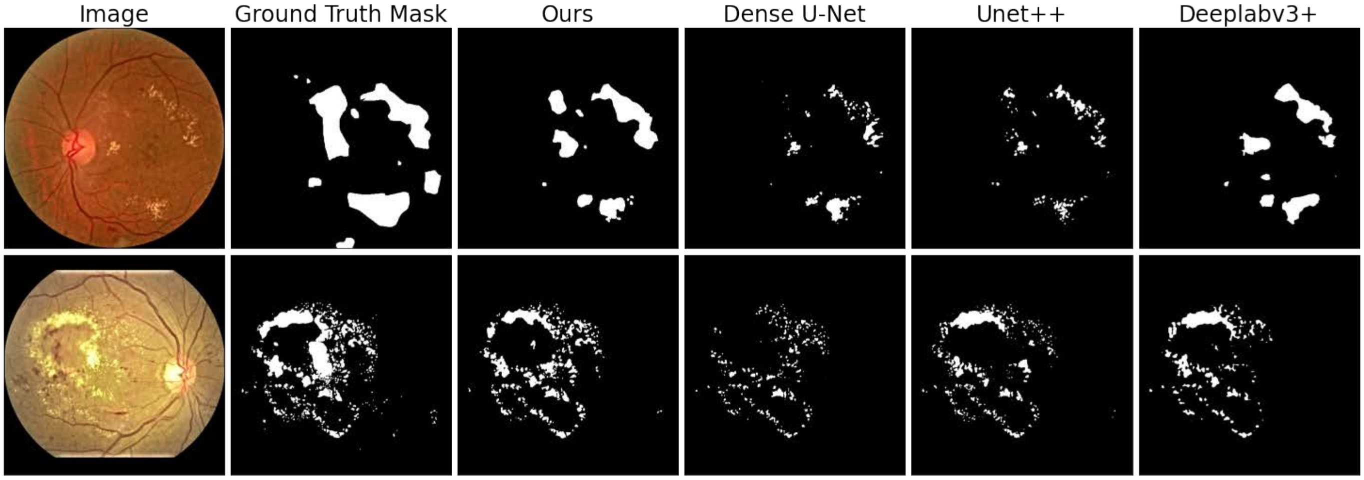


Fig. 4: Qualitative Comparison between our proposed method and other advanced methods on the test set of FGADR dataset.

improvement with using TL, where the performance gains of IoU and Dice score are 5.27% and 5.86%, respectively.

Furthermore, attention-based decoder (AD) is proposed to further enhance the performance by introducing spatial and channel Squeeze-and-Excitation (scSE) Attention Mechanism. More specifically, the skip connection within the U-Net++ is modified by integrating the scSE layers in the decoder. Such a modification can provide a further performance gain and the gains of IoU and Dice score are 2.87% and 3.01%.

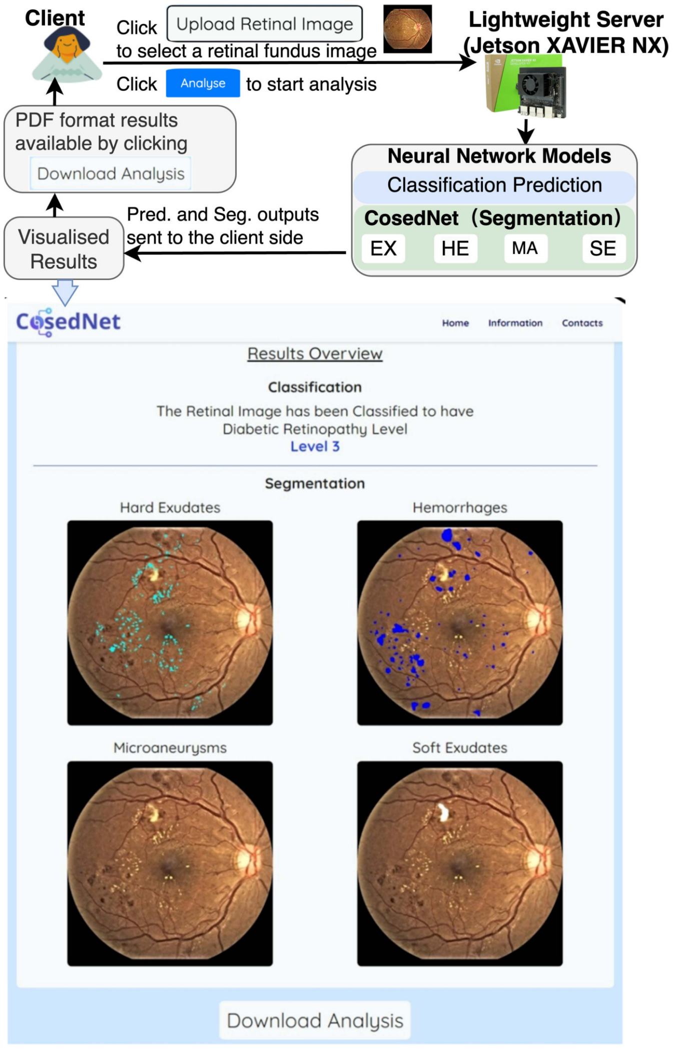
TABLE V: Ablation Study of Hard Exudates based U-Net++ Network Backbone on FGADR Dataset (ED: encoder-decoder, LE: lightweight encoder, DCE: the compound loss of dice loss and cross entropy, TL: transfer learning, and AD: Attention-based Decoder).

Method	ED	LE	DCE	TL	AD	IoU	Dice
baseline	✓					27.66	43.33
+LE	✓	✓				29.37	45.40
+DCE	✓	✓	✓			31.41	47.81
+TL	✓	✓	✓	✓		36.68	53.67
+AD	✓	✓	✓	✓	✓	39.55	56.68

Overall, our method can improve the performance of IoU from 27.66% to 39.55% and the performance of Dice score of 43.33% to 56.68%, which significantly outperforms baseline.

V. Case Study: A Web-based Application

Recent technological advances made smartphone-based biomedical imaging systems make it possible to capture retina images at home and enforce small-sized, low-powered, and affordable DR screening in diverse environment. DR detection is a time-consuming task and requires an intensive effort because lesions are small-sized and is short of contrast. In this section, a web-based prototype is conducted, where our model is deployed in a low-cost device to support screening programmes with particularly in resource constrained settings. Such a service can help reduce the risk of misclassification. To achieve the digitalisation of healthcare systems, it is important to deploy healthcare services in IoT (internet of things) devices, which are located closer to the end users [42]. A server-client architecture is adopted to realise digitalisation. By adopting this architecture, the system has the benefits of availability, scalability and cost efficiency. Furthermore, by addressing the complex evaluation process that ophthalmologists need to go through to reach a concrete detection, we only keep the graphical user interface on client-side and also migrate the additional inference operations of the neural networks to the server-side. More implementation details of our web-based application are provided as follows.



The Retinal Image has been Classified to have Diabetic Retinopathy Level
Level 3

Fig. 5: Actions are taken within the client and server sides of the web-based application. Numerical representation and arrow symbols are for flow and clarity. The Jetson XAVIER NX is 384-core GPU with 48 Tensor Cores and 8GB memory.

Once a retina image is uploaded and sent to the server-side, FastAPI handles the request and further decryption of the Base64 images is performed. After the decryption is finished, the retinal image is passed to our networks for lesion segmentation and diabetic retinopathy grading. For the task of lesion segmentation, our method segments four types of lesions within the retinal image, including EX, HE, MA and SE. The predicted segmentation masks for four lesions are highlighted in the original image so that the ophthalmologist can identify the different lesions much easier. For the task of diabetic retinopathy grading, the network architecture of EfficientNetV2-S [33] is chosen to classify the image with considering its high running efficiency and limited computation resource available in an IoT device. The inferred severity labels is combined with the four segmented and coloured images under the form of a JSON object and it is sent to the client-side where it is handled by the different React components within the Results page. Further detailed visualisation of the processes within this web-based automatic detection system is shown in Fig. 5. Our web-based application is launched and available at our CosedNet Website [3].

VI. Conclusion and Future Work

Biological features can be efficiently extracted from biomedical images by semantic segmentation. This paper proposes a novel compound scaling encoder-decoder network architecture for biomedical image segmentation to detect biomarkers of diabetic retinopathy. Such a network architecture includes the encoder and decoder phases. In the encoder phase, a lightweight encoder is developed by introducing a compound scaling network architecture. In the decoder phase, the skip pathway is realised by introducing scSE Blocks, which is an attention mechanism that prioritises different spatial regions of the given image. To further improve the performance, a compound loss function is proposed to optimise the network, along with the utilisation of transfer learning technique to tackle the class imbalance issue.

Our proposed method is assessed on both DDR and FGADR datasets. For the segmentation tasks, all four types of lesions are evaluated, including EX, HE, MA, and SE. The experimental result shows that our proposed method outperforms other competing methods on lesion segmentation both in IoU and Dice Score. We also conduct an ablation study to investigate how each innovative component in our proposed method contributed to the improvement in performance.

In the future, a multi-scale based compound loss function will be developed to optimise network so as to thoroughly solve the class imbalance issue. We will further evaluate our model in local collected data, such as, the data collected from the Scottish national screening programme, to explore the generalisability of the proposed models. For example, the features learned from the domain of diabetic retinopathy are expected to be transferred to the wider field of other ocular diseases, such as glaucoma or cataracts.

Acknowledgment

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Y. Niu, L. Gu, Y. Zhao, and F. Lu, “Explainable diabetic retinopathy detection and retinal image generation,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 44–55, 2021.

[2] N. Eladawi *et al.*, “Early diabetic retinopathy diagnosis based on local retinal blood vessel analysis in optical coherence tomography angiography (octa) images,” *Medical physics*, vol. 45, no. 10, pp. 4582–4599, 2018.

[3] E. Ophthalmoscopy, “International clinical diabetic retinopathy disease severity scale detailed table,” *ETDRS*, 2002.

[3] https://sites.google.com/view/cosednet/home

[4] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 818–828, 2020.

[5] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," *Inf. Sci.*, vol. 501, pp. 511–522, 2019.

[6] G. A. Williams *et al.*, "Single-field fundus photography for diabetic retinopathy screening: a report by the american academy of ophthalmology," *Ophthalmology*, vol. 111, no. 5, pp. 1055–1062, 2004.

[7] A. Chakravarty and J. Sivaswamy, "Race-net: a recurrent neural network for biomedical image segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1151–1162, 2018.

[8] A. W. Stitt *et al.*, "The progress in understanding and treatment of diabetic retinopathy," *Prog. Retinal Eye Res.*, vol. 51, pp. 156–186, 2016.

[9] M. U. Akram, S. Khalid, and S. A. Khan, "Identification and classification of microaneurysms for early detection of diabetic retinopathy," *Patt. Recogn.*, vol. 46, no. 1, pp. 107–116, 2013.

[10] D. Taraprasad et al., "Recently updated global diabetic retinopathy screening guidelines: commonalities, differences, and future possibilities," *Eye*, vol. 35, no. 10, pp. 2685–2698, 2021.

[11] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Inform. Med. Unlocked*, vol. 20, p. 100377, 2020.

[12] N. Asiri, M. Hussain, F. Al Adel, and N. Alzaidi, "Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey," *Artif Intell Med*, p. 101701, 2019.

[13] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[14] S. Stolte and R. Fang, "A survey on medical image analysis in diabetic retinopathy," *Medical image analysis*, vol. 64, p. 101742, 2020.

[15] N. Sambyal, P. Saini, R. Syal, and V. Gupta, "Modified u-net architecture for semantic segmentation of diabetic retinopathy images," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 3, pp. 1094–1109, 2020.

[16] U. Ahmed, J. C.-W. Lin, and G. Srivastava, "Towards early diagnosis and intervention:: An ensemble voting model for precise vital sign prediction in respiratory disease," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[17] F. Ullah, G. Srivastava, H. Xiao, S. Ullah, J. C.-W. Lin, and Y. Zhao, "A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging," *IEEE J. Biomed. Health Inform.*, 2023.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, 2019.

[19] Q. Meng, Y. Hashimoto, and S. Satoh, "How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3351–3361, 2020.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Euro. Conf. on Comp. Vis.*, 2018, pp. 801–818.

[21] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, 2018.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 4700–4708.

[23] H. Zhao, Z. Fang, J. Ren, C. MacLellan, Y. Xia, S. Li, M. Sun, and K. Ren, "Sc2net: A novel segmentation-based classification network for detection of covid-19 in chest x-ray images," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4032–4043, 2022.

[24] J. Ren, Y. Yan, H. Zhao, P. Ma, J. Zabalza, Z. Hussain, S. Luo, Q. Dai, S. Zhao *et al.*, "A novel intelligent computational approach to model epidemiological trends and assess the impact of non-pharmacological interventions for covid-19," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3551–3563, 2020.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput. Assist. Int.*, 2015, pp. 234–241.

[26] J. S. Suri, S. Agarwal, S. K. Gupta, A. Puvvula, K. Viskovic, N. Suri, A. Alizad, A. El-Baz, L. Saba, M. Fatemi *et al.*, "Systematic review of artificial intelligence in acute respiratory distress syndrome for covid-19 lung patients: a biomedical imaging perspective," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 11, pp. 4128–4139, 2021.

[27] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support.* Springer, 2018, pp. 3–11.

[28] Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu *et al.*, "Gpipe: Efficient training of giant neural networks using pipeline parallelism," *Advances in Neural Inf. Process. Syst.*, vol. 32, 2019.

[29] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[30] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 2820–2828.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 4510–4520.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 7132–7141.

[33] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Int. Conf. Mach. Learn.*, 2021, pp. 10 096–10 106.

[34] A. He, T. Li, N. Li, K. Wang, and H. Fu, "Cabnet: category attention block for imbalanced diabetic retinopathy grading," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 143–153, 2020.

[35] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, 2018.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[37] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging. Graph.*, vol. 95, p. 102026, 2022.

[38] L. Dai *et al.*, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Ieee, 2009, pp. 248–255.

[40] Z. Fang, J. Ren, C. MacLellan, H. Li, H. Zhao, A. Hussain, and G. Fortino, "A novel multi-stage residual feature fusion network for detection of covid-19 in chest x-ray images." *IEEE Trans Mol Biol Multiscale Commun*, pp. 17–27, 2022.

[41] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[42] U. Ahmed, J. C.-W. Lin, and G. Srivastava, "Graph attention-based curriculum learning for mental healthcare classification," *IEEE Journal of Biomedical and Health Informatics*, 2023.