Multi-Level Attention Network for Retinal Vessel Segmentation

Yuchen Yuan, Lei Zhang, *Senior Member, IEEE*, Lituan Wang, and Haiying Huang

***Abstract*—Automatic vessel segmentation in the fundus images plays an important role in the screening, diagnosis, treatment, and evaluation of various cardiovascular and ophthalmologic diseases. However, due to the limited well-annotated data, varying size of vessels, and intricate vessel structures, retinal vessel segmentation has become a long-standing challenge. In this paper, a novel deep learning model called AACA-MLA-D-UNet is proposed to fully utilize the low-level detailed information and the complementary information encoded in different layers to accurately distinguish the vessels from the background with low model complexity. The architecture of the proposed model is based on U-Net, and the dropout dense block is proposed to preserve maximum vessel information between convolution layers and mitigate the over-fitting problem. The adaptive atrous channel attention module is embedded in the contracting path to sort the importance of each feature channel automatically. After that, the multi-level attention module is proposed to integrate the multi-level features extracted from the expanding path, and use them to refine the features at each individual layer via attention mechanism. The proposed method has been validated on the three publicly available databases, i.e. the DRIVE, STARE, and CHASE _ DB1. The experimental results demonstrate that the proposed method can achieve better or comparable performance on retinal vessel segmentation with lower model complexity. Furthermore, the proposed method can also deal with some challenging cases and has strong generalization ability.**

***Index Terms*—Retinal vessel segmentation, deep learning, efficient channel attention, multi-level attention.**

I. Introduction

RETINAL vessel segmentation plays an important role in the screening, diagnosis, treatment, and evaluation of various cardiovascular and ophthalmologic diseases such as diabetic retinopathy, hypertension, arteriosclerosis, cardiovascular disease, and stroke [1]. The segmented vascular tree can be used to delineate the morphological attributes of retinal blood vessels, such as length, width, branching pattern and angles, which can be used as a basis for diagnosis of vascular-related diseases [2]. Moreover, the segmented vascular tree is useful for biometric identification because of its uniqueness in each individual [3]. Thus, it is important to accurately segment vessels from retinal fundus images for the diagnosis and intervention of vascular-related diseases.

In clinical practice, the retinal fundus images are manually delineated by the ophthalmologists. However, manual segmentation of retinal blood vessels is a laborious, tedious and highly skilled task. Therefore, automatic segmentation of retinal vessels is urgently required in clinical practice to reduce the annotation time and workload on ophthalmologists [4]. However, automatic retinal vessel segmentation is a long-standing challenge due to the following challenges: (1) the intensity values, shapes, and sizes of vessels can vary hugely within a fundus retinal image. (2) Retinal images contain various structures, such as vessel crossing, branching, retinal boundary and centerline reflex, which makes it difficult to accurately segment the vessels. (3) High noise, poor contrast and low resolution of the fundus images exacerbate the aforementioned problems. To solve these problems, numerous methods [5] are proposed to satisfy the demand for automatic retinal vessel segmentation. All these methods can be classified into two categories: traditional methods and deep neural network based methods.

Traditional methods [6], [8] usually segment the vessels using hand-crafted features, which requires professional knowledge. Hence, these methods are complicated and lack robustness, which may lead to poor generalization [9]. Furthermore, the segmentation accuracy and real time performance need to be further improved.

With the advent of deep convolutional neural networks (DCNNs), the remarkable successes have been achieved in computer vision system on many high-level problems, including image classification [10], [11], object detection [12]. These successes can be partially attributed to the built-in invariance of DCNNs to local image transformations, which is the basis to learn hierarchical abstractions of data [13]. While this invariance is

Manuscript received December 29, 2020; revised April 12, 2021 and May 12, 2021; accepted June 10, 2021. Date of publication June 15, 2021; date of current version January 5, 2022. This work was supported by the National Natural Science Fund for Distinguished Young Scholar under Grant 62025601, and the General Program of National Natural Science Foundation of China under Grant 61772353, as well as the National Natural Science Foundation of China under Grant 62006164. (*Corresponding author: Lituan Wang.*)

Yuchen Yuan, Lei Zhang, and Lituan Wang are with the Machine Intelligence Laboratory, College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yuanyuchen@stu.scu.edu.cn; leizhang@scu.edu.cn; lituanwang@scu.edu.cn).

Haiying Huang is with the Information Management Department of West China Second University Hospital, Sichuan University, Chengdu, Sichuan 610041, China (e-mail: xuxu2110@163.com).

Digital Object Identifier 10.1109/JBHI.2021.3089201

desirable for high-level vision tasks, it can hamper low-level tasks, such as pose estimation [14] and image segmentation, where the spatial accuracy and precise localization are needed rather than abstraction of spatial details. One way to mitigate this problem is to use skip-layers to add the features at shallow layers to the ones at deep layers, such as fully convolutional networks (FCNs) [15] and U-Net [16]. An alternative approach is utilizing atrous convolution with a higher sampling rate to replace the last two max pooling operations to prevent resolution reduction and a fully-connected CRF to refine the segmentation results, which is clearly introduced in deeplab [17]. For the retinal vessel segmentation task, Many methods based on these well-known architectures have been proposed and achieved improved performance on the retinal vessel segmentation [18]–[20]. For example, Wu *et al.* [18] proposed two parallel submodels which both consist of two identical encoder-decoder networks. The first network converts an image patch into a probabilistic retinal vessel map, and the following network further refines the map. In [19], a dual encoding U-Net was proposed to capture both spatial and semantic information with a spatial encoding path and a context path. However, as the width and structure can vary hugely within a fundus image, correctly distinguishing the vessels from the background requires features with different receptive fields.

To satisfy this demand, one kind of strategy [21], [22] is to use multiple dilated convolutional blocks to capture the multi-scale features. For instance, PSPNet [22] aggregated feature maps generated by multiple dilated convolutional blocks to exploit contextual information at different scales. In DeepLabv2 [23], the proposed ASPP module is able to capture objects as well as image context at multiple scales by employing multiple parallel filters with different rates. However, in the retinal vessel segmentation task, each pixel needs query-specific contextual dependencies for different categories (i.e. vessel and non-vessel). Moreover, it is difficult and inflexible to manually design these multi-context representations to adapt all the vessels with different widths. An alternative way is to combine the features at different levels. Mo *et al.* [9] proposed a novel deep supervised fully convolutional network that fuse multi-level features with different receptive fields to obtain precise segmentation results. Although these methods have increased the segmentation performance of the retinal images, there still exist some problems: (i) there are only dozens of well-annotated retinal images, a DCNN model designed for retinal vessel segmentation is prone to be over-fitting. (ii) Features in the contracting path are usually treated as equally important to be concatenated to the features in the expanding path, which makes effective information not be fully exploited and brings much redundant and irrelevant information to the high-level features. This may lead to poor segmentation on the capillaries and bring subtle noise to the segmentation results. (iii) To obtain a more precise segmentation result, most existing methods usually exploit features at each level by directly integrating them to form the final probability vessel map. However, the integrated features inevitably incorporate subtle noise from shallow layers and irrelevant semantic information from deep layers, which leads to unsatisfactory results.

Therefore, to address these aforementioned problems, a novel retinal vessel segmentation network named as AACA-MLA-D-UNet [1] is proposed in this work. The architecture of the proposed network is constructed based on the U-Net model and trained end-to-end. Since the number of the annotated images is limited, features at shallow layers should be fully exploited to improve the robustness of the segmentation model. Dense connectivity proposed in the DenseNet [24] can ensure maximum information flow between layers in the network. Thus, to make more information flow propagate between layers, dense block is applied to replace the original convolution blocks in the U-Net in this work. In the contracting path, to provide the high-level features with more discriminative detailed information, the adaptive atrous channel attention module is proposed to emphasize effective features and suppress redundant or irrelevant ones without increasing the model complexity. In the expanding path, the multi-level attention module is proposed to selectively leverage the complementary characteristic of the detailed information at low levels and semantic information at high levels with attention mechanism. These two attention modules are designed to work together to make the network recover more detailed information lost in the downsampling process and suppress noise from irrelevant regions. Furthermore, to mitigate the over-fitting problem for retinal vessel segmentation, dropout [25] is added in each dense block.

In summary, the contributions of our work are as follows:

1) The AACA-MLA-D-UNet is proposed based on U-Net, and dropout dense block (DDB) is proposed to replace the original convolution block in U-Net to preserve maximum vessel information between convolution layers and mitigate the over-fitting problem.
2) The adaptive atrous channel attention (AACA) module is proposed to efficiently select more effective detailed information to improve the segmentation accuracy of capillaries.
3) The multi-level attention (MLA) module is developed, which applies attention mechanism to leverage the complementary advantage of multi-level information integrated from different layers to learn more discriminative representation for accurate and robust vessel segmentation.

II. Methods

A. Overview of the Network Architecture

U-Net has been widely used in many medical image segmentation tasks owing to its specific U-shape structure with skip-connection. In this work, the overall architecture of the proposed AACA-MLA-D-UNet is also constructed based on U-Net. As shown in Fig. 1, the 3-channel input images are fed into the proposed AACA-MLA-D-UNet architecture for vessel segmentation. A 3×3 convolution is first applied to the original images to increase the number of channels, obtaining the primary feature maps. Then, these features are fed into the dropout dense block and the adaptive atrous channel attention

[1] https://github.com/IsYuchenYuan/retinal-vessel-segmentation

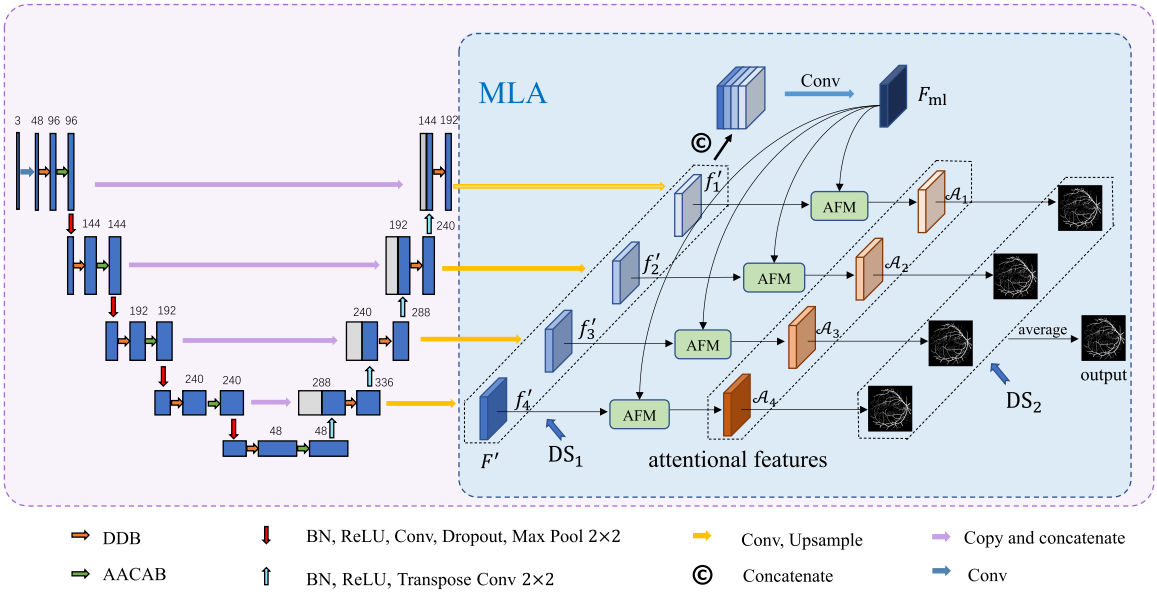

Fig. 1. The detailed architecture of AACA-MLA-D-UNet for vessel segmentation. AACA-MLA-D-UNet contains two path, i.e., the contracting path and expanding path. Each layer in the contracting path includes a dropout dense block (DDB) (orange right arrows) and an adaptive atrous channel attention module (AACAB) (green right arrows), followed by a downsampling operation (red down arrows). Each layer in the expanding path includes a DDB and a transposed convolution layer for unsampling (blue up arrows). There are skip connections between the contracting and expanding path to concatenate the features (purple right arrows). The features extracted from the expanding path are enlarged by bilinear interpolation (yellow right arrows) and feed into the multi-level attention (MLA) module (shallow blue).

block. To complete down sampling, a series operations including batch normalization, ReLU activation [26] and a 1×1 convolutional layer followed by a 2×2 max pooling layer are adopted. Subsequently, several cascade layers with the same operations as the first layer (except the first convolution) are used to learn the characteristics of the images. In the following expanding path, a transposed convolution is used for upsampling, feature maps with 48 channels are generated and concatenated with the feature maps from the parallel contracting path. The combined features are then fed into a dropout dense block. To get more discriminative features, features at each layer of the expanding path are further refined by the multi-level attention module. Then, the segmentation maps obtained from these refined features are averaged to generate the final probability map.

Vessel segmentation task can be treated as a per-pixel classification problem. Formulaically, let $D = \{(x_i, y_i); i = 1, \ldots, M\}$ denotes the training set of the network. The proposed architecture is designed to learn a segmentation model $f(y|x, \theta)$ that maps each pixel of the input image to its corresponding pixel in label space. Here, $\theta = \{\theta^d, \theta^a, \theta^m\}$ denotes the trainable parameters in the proposed architecture.

Next, details about the dropout dense block, the adaptive atrous channel attention module, and the multi-level attention module will be introduced separately. For convenience, the notations that used in this work are listed in Table I.

TABLE I
TABLE OF NOTATIONS

Symbol	Description
u_l	the output of the DDB of the lth layer in the contracting path
c_l	the output of the AACA of the lth layer in the contracting path
f_l	the output of the DDB of the lth layer in the expanding path
$f_l^{'}$	the enlarged feature map at the lth layer in the expanding path
L	the number of the layers of the proposed network
Cat	the concatenate operation across the channels
$C1D$	$1D$ atrous convolution operation
$Conv$	convolution operation
θ^d	all the trainable parameters in the DDB
θ^a	all the trainable parameters in the AACA module
θ^m	all the trainable parameters in the MLA module
δ	ReLU activation
σ	sigmoid activation

B. Dropout Dense Block

As the number of the annotated fundus images is limited, features extracted by each convolution layer are supposed to be fully exploit. To preserve maximum vessel information between convolution layers and mitigate the over-fitting problem, the dense blocks with dropout are employed to replace the original convolution blocks in U-Net, and the formed architecture is named as D-UNet. As seen in Fig. 2, the dth layer of the dropout dense block has an output with j feature maps and an input with $F + (d-1) \times j$ feature maps, where F is the number of the features maps of the initial input of the dropout dense block. At each layer, features are learned through a series operations, including *BatchNormalization*, *ReLU*, *a* 3×3 *Convolution* and *Dropout*. To keep the dimension of the feature maps consistent and preserve the edge information of the feature maps, a zero padding of size 1 is applied to each 3×3 convolution in the dropout dense block.

Specifically, in this work, each dropout dense block contains 4 layers, and j is set to 12. Then, the final output of each dropout

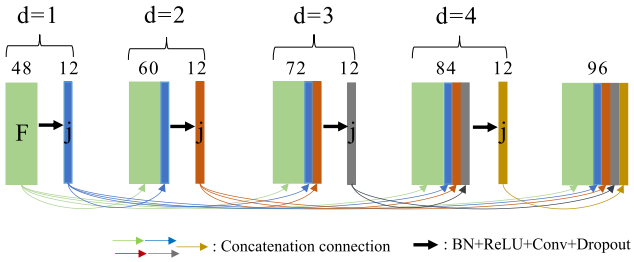

Fig. 2. Four-layer dropout dense block with j=12 and F=48. Feature maps of the previous layers are concatenated to form the input of the following layers.

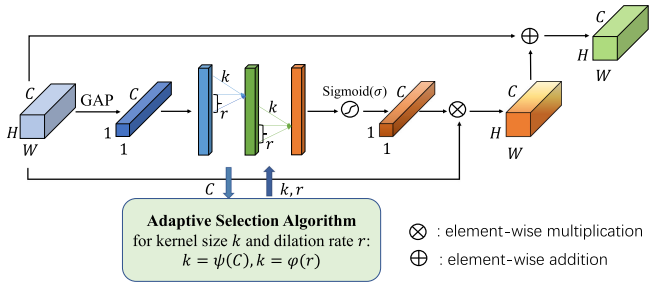

Fig. 3. Diagram of our adaptive atrous channel attention (AACA) module. Given the features aggregated through the global average pooling (GAP), AACA generates recalibrated weights with two consecutive $1D$ convolutions of the same size k and dilation rate r, where k and r are adaptively determined via the *Adaptive Selection Algorithm* given the channel dimensions C.

dense block has $F + 48$ feature maps. Specially, to prevent our model from high computational complexity, in the bottleneck layer, only new generated feature maps of each dense layer are collected to form the final output of this (i.e., 48 feature maps).

C. Adaptive Atrous Channel Attention

Features in the contracting path play an important role in determining the segmentation accuracy. On one hand, the subsequent features are based on these features to extract higher level semantic information. On the other hand, features in the contracting path provide the ones in the expanding path with detailed information via skip connection. However, we have found that the traditional segmentation network usually treat the features in the expanding path as equally important, which makes effective information not be fully exploited and preserves much redundant and irrelevant information. This may hamper the accurate segmentation of capillaries. Thus, to sort the importance of each feature channel automatically, a channel attention module named as adaptive atrous channel attention module is embedded to the network without increasing the model complexity.

To compute the channel attention effectively and efficiently, the proposed attention module is expected to satisfy three demands. Firstly, it must has the ability of capturing long-range dependencies which exceed the local context across the feature channels. Secondly, it should be flexible, i.e. it should be able to capture the nonlinear interaction between channels. At last, the whole operation would not bring higher model complexity or suffer from heavier computational burden.

To this end, the adaptive atrous channel attention module is developed and the architecture is illustrated in Fig. 3. In this module, a spatial context descriptor: $F_{avg} \in \mathbb{R}^{1\times1\times C}$ is firstly generated by aggregating spatial information of the input feature map $u_l \in \mathbb{R}^{H\times W\times C}$ using global average pooling (GAP). The descriptor is then forwarded to the two sequential atrous convolution operations and the ReLU activation to produce the channel attention map $M \in \mathbb{R}^{1\times1\times C}$. In this way, the module can capture the long-range contextual information across the channel dimension of the feature maps without increased computational complexity, and is able to capture the nonlinear interaction between channels. Then, the channel attention map is used to refine the input feature map via the element-wise multiplication. Finally, in order to keep the discriminative information encoded in both refined features and original features, the refined feature map is combined with the original feature map using element-wise addition to form the final output c_l. The overall procedure of the adaptive atrous channel attention module can be formulated as follows:

$$F_{avg} = GAP(u_l) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_l(i,j), \tag{1}$$

$$M = \sigma\left(C1D\left(\delta\left(C1D(F_{avg}; k, r, \theta_1^a)\right); k, r, \theta_2^a\right)\right), \tag{2}$$

$$c_l = u_l + M u_l. \tag{3}$$

Here, k and r represent the kernel size and dilation rate used in the two $1D$ atrous convolutions, respectively; θ_1^a and θ_2^a represent the trainable parameters for the first and the second atrous convolution, respectively. δ refers to the ReLU activation; σ refers to the sigmoid activation, which make the value of the channel attention map locate between 0 and 1. k and r determine the coverage of the local cross-channel interaction, i.e., how many adjacent channels participate in the reweighting of one channel and how far apart they are. It is obvious that high-dimensional (low-dimensional) channels require long-range (short-range) information to better model their cross-channel dependencies. To make the coverage of interaction (i.e., k and r) adaptively fit in the different channel dimensions, we assume that kernel size k and dilation rate r are both proportional to the channel dimension C. In this work, we propose the *adaptive selection algorithm* to automatically choose the k and r for different channel dimensions. Since k is proportional to the C, the simplest mapping ϕ between kernel size k and channel dimension C can be represent by a linear function: $C = \phi(k) = \omega * k - q$. Here, ω and q are two constants, which indicate the slope and the y-intercept of the linear function, respectively. However, linear function is too limited to model the relationship between the kernel size and channel dimension. On the other hand, channel dimension C is generally set to the power of 2. Thus, the mapping ϕ can be possibly represent as [27]:

$$C = \phi(k) = 2^{\omega * k - q}. \tag{4}$$

Given the channel dimension C, kernel size k can be adaptively calculated as follows.

$$k = \psi(C) = \left| \frac{log_2(C)}{\omega} + \frac{q}{\omega} \right|_{odd}, \tag{5}$$

where $|t|_{odd}$ indicates the nearest odd number of t. If t is an odd number, $|t|_{odd}$ indicates itself, otherwise, if t is an even number, $|t|_{odd}$ indicates the smaller one of the two nearest odd number of t. In this work, ω and q are set to 2 and 1. In terms of dilation rate r, we model a mapping φ between k and r:

$$k = \varphi(r). \tag{6}$$

Since k and r are both proportional to channel dimension C, we choose a linear function to model the relationship, i.e.,

$$\varphi(r) = \lambda r - m, \tag{7}$$

here, λ and m are set to 2 and 1, respectively.

Differences with SE block and ECA block: The proposed AACA module, SE block [28] and ECA block [27] are all designed to capture the cross-channel interaction, but there are certain differences between them. SE block consists of two FC layers and a ReLU activation layer, long-range dependencies and non-linear interaction between channels are guaranteed. However, the fully-connected operations bring high model complexity. While dimensionality reduction can reduce the model complexity, it destroys the direct corresponding between channel and its weight. To reduce the model complexity, ECA block utilizes a $1D$ convolution to replace the two FC layers. However, the number of the channels to participate in the attention prediction of one channel is limited by the kernel size, which discards dependencies of other channels beyond the receptive field. Moreover, the non-linear interaction between channels can not be guaranteed by just one convolution operation. Different from the two methods, the proposed AACA module leverages two atrous convolution layers and a ReLU activation layer between them to model the long-range dependencies. Compared with one convolution layers used in ECA, the proposed AACA module not only can exceed the local context to capture the global information between channels through atrous convolution operations, but also can learn the non-linear interaction between channels. Furthermore, the AACA module would not bring extra model complexity, which reduces the space complexity from $O(C \times C)$ to $O(1)$ when compared with the SE block.

D. Multi-Level Attention

Multi-level features are essential for accurate and robust retinal vessel segmentation. Thus, inspired by the method in [29], the multi-level attention module (see the right part of Fig. 1) is proposed to fully exploit the complementary features at each level in a more effective way. The proposed multi-level attention module includes three processes. Firstly, feature maps at different layers of the expanding path are enlarged and concatenated to form the integrated multi-level feature. Then, the integrated multi-level feature is used to refine the features at each individual layer through an attentional feature module. For convenience, the refined feature maps are named as the attentional features.

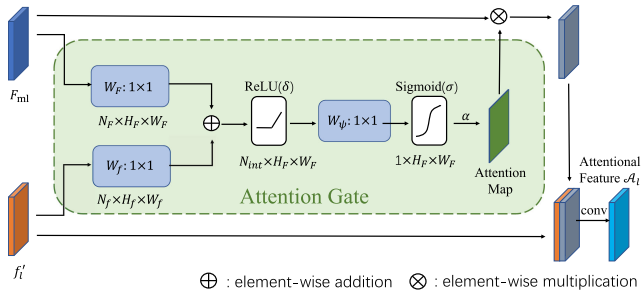

Fig. 4. The illustration of the attentional feature module.

At last, the segmentation results generated by each attentional feature are fused to obtain the final predicted segmentation result. The following describes the details of the three processes.

Letting $F = \{f_1, \ldots, f_l, \ldots, f_{L-1}\}$ denotes the generated feature maps in the expanding path at different layers (features of the bottleneck layer are ignored due to the memory cost). Feature maps at shallow layers include abundant detail information, while feature maps at deep layers encode high-level semantic information. The high-level semantic information can help to locate the main structure of the vessels, and the detail information can detect the fine structure such as capillaries. To concatenate these feature maps along the channel dimension, we enlarge these feature maps with different resolutions to the size of the original input image by using bilinear interpolation. At last, these enlarged feature maps $F' = \{f_1', \ldots, f_l', \ldots, f_{L-1}'\}$ are concatenated, and followed by a convolutional layer to form the integrated multi-level feature map F_{ml}. The F_{ml} is formulated as:

$$F_{ml} = Conv(Cat(f_1', \ldots, f_l', \ldots, f_{L-1}'), \theta_1^m). \tag{8}$$

Here, θ_1^m is the trainable parameters of the convolution operation. In this way, F_{ml} encodes detail information from shallow layers as well as high-level semantic information from deep layers. And then the generated integrated multi-level feature map is further fed into an attentional feature module (see Fig. 4) to refine the enlarged feature map at each layer (refer to as single-level feature) for better representations.

The refining process is shown in Fig. 4, the integrated multi-level feature map F_{ml} and the single-level feature f_l' are first linearly mapped to the same lower dimensional intermidiate space, then additive attention is used to obtain the attention map $a_l = \{a_{l,i}\}_{i=1}^n$, where $a_{l,i} \in [0, 1]$, n is the number of pixels in the single-level feature. The attention map is calculated as follows:

$$q_l^{att} = \psi^T(\delta(W_f^T f_l' + W_F^T F_{ml} + b_F)) + b_\psi \tag{9}$$

$$a_l = \sigma(q_l^{att}(f_l', F_{ml}; \theta_2^m)), \tag{10}$$

where θ_2^m consists of: three linear transformations $W_f \in \mathbb{R}^{N_f \times N_{int}}$, $W_F \in \mathbb{R}^{N_F \times N_{int}}$, $\psi \in \mathbb{R}^{N_{int} \times 1}$ and two bias terms $b_F \in \mathbb{R}^{N_{int}}$, $b_\psi \in \mathbb{R}$. The linear transformation is implemented by the 1×1 convolution. After obtaining the attention map, we multiply it with the integrated multi-level feature map element-by-element to generate the refined feature map, which

encodes the complementary information of the single-level feature. Then, the new refined feature map is concatenated with the single-level feature and followed by two consecutive 3×3 convolution operation and a 1×1 convolution operation to produce the final attentional feature $\mathcal{A}_l$ for each layer. At last, the segmentation maps generated from the attentional features $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_l, \ldots, \mathcal{A}_{L-1}\}$ are fused to form the final probability vessel map.

Different from the work in [29], we utilize the attention gate model instead of a series convolution operations to generate the attention map. The attention gate model not only can learn to suppress irrelevant regions of the integrated multi-level feature that are not needed by each individual single-level feature while highlighting features which the single-level feature want, but also maintain much lower computational overhead and model complexity.

E. Loss Function

In this work, to fully leverage the multiple branch outputs, and improve the performance of segmentation, deep supervision [9] is incorporated in the proposed method. A loss function that is a summation of the loss from the final predicted segmentation result, and branch losses associated with the auxiliary classifiers is designed as follows.

$$L = \sum_{l=1}^{L-1} \alpha_l L_l^{coarse} + \sum_{l=1}^{L-1} \beta_l L_l^{refined} + \gamma_f L_f, \qquad (11)$$

where α_l and L_l^{coarse} represent the weight and the loss of the segmentation map at lth layer before refinement, respectively (i.e., DS $_1$ in Fig. 1); β_l and $L_l^{refined}$ represent the weight and the loss of the segmentation map refined by our attentional feature module at l th layer (i.e., DS_2 in Fig. 1); L_f and γ_f refer to the loss of the final output map averaged by all the refined predicted maps, and the corresponding weight, respectively. All the loss functions utilized in this work are cross-entropy loss. Finally, the parameters θ of the proposed network are optimized in an end-to-end way by minimizing the total loss function L via Adam optimizer [30].

The overall learning process of the proposed architecture is shown in Algorithm 1.

Algorithm 1 The learning process of the proposed architecture.

Input: The training dataset: $D = \{(x_i, y_i); i = 1, \ldots, M\}$, learning rate lr and the corresponding decay rate lr_{decay}, the predefined epoch number to decay learning rate pre_{epoch}, and the maximum epoch max_{epcoh}

Output: The learned parameters in the proposed architecture $\theta = \{\theta^d, \theta^a, \theta^m\}$

```
for epoch=1:max_epoch do
  for i=1:M do
    Extracting multi-level features F = {f_1, ..., f_l, ..., f_{L-1}} from each input image x_i
    Enlarging the extracted feature maps F to generate the single-level features: F' = {f'_1, ..., f'_l, ..., f'_{L-1}}
    Integrating the single-level features F' to form the integrated multi-level feature F_ml
    Leveraging the integrated multi-level feature F_ml to refine each single-level feature F' to generate the attentional features A = {A_1, ..., A_l, ..., A_{L-1}}
    Obtaining the final predicted segmentation result by averaging the segmentation results generated from the attentional features A
  end for
  Computing the cost L and the gradient of L with respect to θ
  Updating the parameters θ using the Adam optimizer
  if epoch is a multiple of pre_epoch then
    lr ← lr × lr_decay
  end if
end for
```

III. Experiments

A. Database

The proposed method is evaluated on three databases, DRIVE, CHASE _ DB1 and STARE. The DRIVE database [31] consists of 40 color retinal images, which is divided into the training set and the test set, and each subset contains 20 images. Each image is captured at 565×584 pixels and equipped with a manually vessel segmentation map as the ground truth and a binary field of view (FOV) mask. The CHASE _ DB1 database [32] contains 28 retinal images with the resolution of 999×960 pixels. There are two manual annotations by observers, the first one is normally used as the ground truth. Since no official data split is available for CHASE _ DB1, we divide these images into 2 sets: the first 20 images are used for training and the rest 8 images are used for testing [33]. The STARE database [34] contains 20 retinal images, and each image is captured at 700×605 pixels and equipped with two manual segmentation maps by experts. And the first one is taken as the ground truth. There is also no specific separated training and test sets available for this database, we adopt the 'leave-one-out'[31] technique to train our model for a fair comparison. Because that there are no binary FOV masks in the CHASE _ DB1 and STARE databases, the masks of these two databases are manually generated similar as the method used in [8].

B. Pre-Processing & Patch Extraction

As shown in Fig. 5(a), the hue and saturation can vary hugely within each color retinal image. To normalize such variation, each original image is converted into an intensity image, and then the intensities of the generated image are normalized to zero mean and unit variance. At last, the normalized intensities are rescaled to the range of 0 to 255. An example of the normalized image is shown in Fig. 5(b). It can be seen that the processed images also suffer from low intensity contrast.

Fig. 5. Pre-processing of the fundus retinal images. (a) Original fundus retinal image. (b) Image after graying and normalization. (c) Image after CLAHE and gamma adjusting on the basis of (b).

To solve this problem, the contrast limited adaptive histogram equalization (CLAHE) algorithm [35] is adopted to improve image contrast and gamma adjusting algorithm is employed to adapt the processed images to the human vision. The CLAHE algorithm can not only focus on the local area to avoid the loss of light or dark details, but also can overcome the problem of excessive amplification of the noise in near regions. In this work, we divide the original image into 8×8 patches and the histogram is calculated for each patch. The clip limit is set to 2.0, value of histogram bin larger than the clip limit will be clipped and distributed uniformly to other bins before the histogram equalization is applied. An example of the generated image using the aforementioned pre-processing is illustrated in Fig. 5(c).

Since there are very few training data available for all these three databases, to overcome the over-fitting problem during the training, data augmentation and patch-based segmentation are applied. Various transformations including flipping and rotation are applied on each image in the training set to generate ten augmented images. Multiple overlapped patches with the size of size 256×256 are extracted from the augmented images for model training. For each image in the training set of DRIVE, 20 different patches are extracted. While for the images in STARE and CHASE _ DB1, because of the image sizes are larger when compared with DRIVE, we cropped 30 and 40 different patches from each image of these two training databases, respectively.

C. Training and Implementation Details

The proposed model is trained using mini-batch Adam optimizer with the batch size of 4. During training, the learning rate is initialized as 0.0005, and decay to 1/10 each 20 epochs, and the momentum is set to 0.9, weight decay is set to =0.0001. Furthermore, we empirically set all the hyperparameters in the loss function. i.e., α_l, β_l, γ_f as 1.

In the test stage, the same data pre-processing methods introduced in Section III-B were applied to each test image. Firstly, we extract overlapped patches of size 256 with a stride of 20 from each pre-processed test image and feed them into the well-trained model for segmentation. Then, the obtained segmented vessel maps are composed together in the order of the cropping operation and averaged in the overlapping region. The predicted segmentation result is a vessel probability map, in which each value represents the probability of each pixel belonging to the vessel class. To obtain the binary vessel segmentation, we employ a threshold of 0.5 to the probability of each pixel in the binary segmentation map to determine whether it belongs to vessels or the background.

To evaluate whether there exists significant difference between the two correlated ROC curves of UNet and the proposed method, the Wilcoxon's rank-sum test [36] is conducted. The detailed steps are as follows, firstly, a bootstrapping approach is employed to randomly sample 20 test samples from the test datasets of the DRIVE databases for 1000 times. Then, for each time, the 20 randomly sampled test samples are fed into the well-trained proposed method and UNet to test. Finally, two sets of 1000 testing results are fed into the Wilcoxon's rank-sum test algorithm to calculate the p value with the statistical significance level set to $\alpha = 0.05$. The null hypothesis is that there exist significant differences between the sample sets if $p < 0.05$, and no significant differences if $p > 0.05$.

D. Evaluation Metrics

Since retinal vessel segmentation is a binary segmentation problem, we employ the evaluation metrics Accuracy (Acc), Sensitivity (Sen) (also known as Recall (Re)), Specificity (Spe), F1-score ($F1$), the area under the receiver operating characteristic (ROC) curve

(area under curve $= AUC$) and Matthews correlation coefficient (MCC), which are based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), to evaluate the segmentation results. The definitions of these metrics are shown as follows:

$$Sen = \frac{TP}{TP + FN}, Spe = \frac{TN}{TN + FP}, Acc = \frac{TP + TN}{N},$$

$$MCC = \frac{TP/N - S \times P}{\sqrt{P \times S \times (1 - S) \times (1 - P)}},$$

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re},$$

where $N = TP + FP + TN + FN$, $S = (TP + FN)/N$, $P = (TP + FP)/N$ and $Pr = TP/(TP + FP)$. The MCC are often used in the performance analysis of segmentation results on class imbalance datasets. The MCC returns a value between -1 and 1, with -1 indicating a completely incorrect prediction, and 1 indicating a perfect prediction. For all the above metrics, higher values indicate the better segmentation performance.

IV. Results

A. Vessel Segmentation Results

In this subsection, both the qualitative and quantitative experiments are conducted on the DRIVE, CHASE _ DB1, and STARE databases, and the results are shown in Fig. 6 and Table II, respectively. Fig. 6 shows the test images from the three databases, as well as the final binary segmentation results and the ground truths. It can be seen that the structure of vessels is well preserved by our model and the connectivity of vessel tree is guaranteed. Moreover, most vessels can be segmented clearly from the background by using the proposed method, including

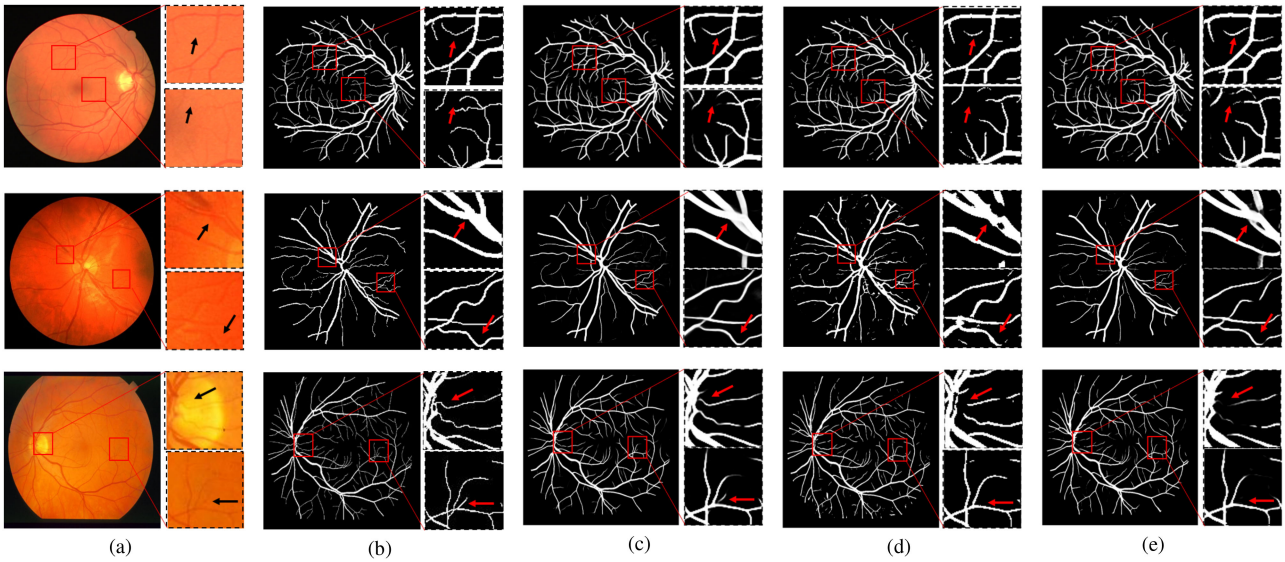

Fig. 6. Three example test images from the DRIVE (top), CHASE _ DB1 (middle), STARE (bottom) respectively, with two enlarged rectangles beside each sub-figure. (a) Original fundus retinal images, (b) the ground truths, (c) binary output of the proposed method, (d) binary output of Joint Loss [45]. (e) binary output of Jin *et al.* [46].

TABLE II
AVERAGE PERFORMANCE MEASURES FOR DRIVE, STARE AND CHASE _ DB1

Database	Method	Spe	Sen	Acc	AUC	$F1$	MCC
DRIVE	2nd human observer	0.9724	0.7760	0.9472	-	0.7891	-
	Proposed method	0.9805	0.8046	0.9581	0.9827	0.8303	0.8070
STARE	2nd human observer	0.9384	0.8952	0.9349	-	0.7427	-
	Proposed method	0.9870	0.7914	0.9665	0.9864	0.8276	0.8116
CHASE_DB1	2nd human observer	0.9711	0.8105	0.9545	-	0.7787	-
	Proposed method	0.9801	0.8402	0.9673	0.9874	0.8248	0.8070

some difficult cases such as vessels with variable widths and vessels in the presence of the central vessel reflex, which are enlarged and present beside each figure.

Table II demonstrates the quantitative segmentation results of the proposed method on the three databases. It can be seen that our method outperforms the second human observer for all three databases on the metrics of accuracy, specificity, sensitivity, $F1$ and MCC values. Furthermore, the metric of AUC values achieved by our method are greater than 98.25% for all three databases.

B. Comparison to the State-of-The-Arts Methods

We compare the proposed method with other state-of-the-art methods in recent years on DRIVE, CHASE _ DB1, and STARE databases. The experimental results are shown in Table III. It can be seen that the proposed method achieves the highest $F1$ score for the three databases and the highest sensitivity and accuracy for the STARE and CHASE databases. For the DRIVE database, the method proposed by Srinidhi *et al.* [41] achieves better results than our method in term of sensitivity and accuracy metrics, but the proposed method outperforms this method in terms of specificity, AUC and $F1$, which are improved by 1.38%,1.26% and 6.96%, respectively. Moreover, compared with the same method [41], our method improves the accuracy, AUC and $F1$ values by 1.99%, 3.03% and 10.59% for the CHASE database and 1.63%, 1.94% and 5.48% for the STARE database. This is mainly due to the method in [41] devotes to detecting more vessels but much more pixels in the background are misclassified as the vessel pixels, while the proposed method can distinguish vessels from the background more accurately and has the stronger generalization ability for different databases. In addition, the proposed method achieves AUC values of 98.27%, 98.74%, and 98.64% for the three databases. These values are higher than in all other methods on the DRIVE and CHASE _ DB1 databases. While on the STARE database, although our method achieves slightly lower AUC values when compared with the method in [33], it can achieve better performance on all the other three metrics. The improvement of the proposed methods are 0.26%, 1.88%, and 0.37% on the metrics of Spe, Sen, and Acc, respectively. Since retinal datasets are class imbalance, we further compare our method with the existing approaches in terms of average MCC on the tree databases. Results in Table IV show that our method

TABLE III
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DRIVE, CHASE_ DB1, AND STARE DATABASES

		DRIVE					CHASE_ DB1					STARE				
Methods	Year	*Spe*	*Sen*	*Acc*	*AUC*	*F1*	*Spe*	*Sen*	*Acc*	*AUC*	*F1*	*Spe*	*Sen*	*Acc*	*AUC*	*F1*
Azzopardi [37]	2015	0.9704	0.7655	0.9442	0.9614	-	0.9587	0.7585	0.9387	0.9487	-	0.9701	0.7716	0.9497	0.9563	-
Li [33]	2015	0.9816	0.7569	0.9527	0.9738	-	0.9793	0.7507	0.9581	0.9716	-	0.9844	0.7726	0.9628	**0.9879**	-
Zhang [38]	2016	0.9712	0.7861	0.9466	0.9703	0.7953	0.9716	0.7644	0.9502	0.9706	0.7581	0.9729	0.7882	0.9547	0.9740	0.7815
Liskowski [39]	2016	0.9807	0.7811	0.9535	0.9790	-	-	-	-	-	-	0.9754	0.7868	0.9566	0.9785	-
Orlando [40]	2017	0.9684	0.7897	0.9507	0.7857	-	0.9712	0.7277	-	0.9524	0.7332	0.9738	0.7680	-	-	0.7644
Srinidhi [41]	2017	0.9667	**0.8644**	**0.9589**	0.9701	0.7607	0.9663	0.8297	0.9474	0.9571	0.7189	0.9746	**0.8325**	0.9502	0.9670	0.7698
Yan [42]	2018	0.9820	0.7631	0.9538	0.9750	-	0.9806	0.7640	0.9607	0.9776	-	0.9857	0.7735	0.9638	0.9833	-
ResUnet [43]	2018	**0.9820**	0.7726	0.9553	0.9779	0.8149	0.9820	0.7726	0.9553	0.9779	0.7800	-	-	-	-	-
R2U-Net [44]	2018	0.9813	0.7799	0.9556	0.9784	0.8171	0.9820	0.7756	0.9634	0.9815	0.7928	-	-	-	-	-
Joint Loss [45]	2018	0.9818	0.7653	0.9542	0.9752	-	0.9809	0.7633	0.9610	0.9781	-	0.9846	0.7581	0.9612	0.9801	-
MS-NFN [18]	2018	0.9819	0.7844	0.9567	0.9807	-	**0.9847**	0.7538	0.9637	0.9825	-	-	-	-	-	-
Jin [46]	2019	0.9800	0.7963	0.9566	0.9802	0.8237	0.9752	0.8155	0.9610	0.9804	0.7883	**0.9878**	0.7595	0.9641	0.9832	0.8143
DEU-Net [19]	2019	0.9816	0.7940	0.9567	0.9772	0.8270	0.9821	0.8074	0.9661	0.9812	0.8037	-	-	-	-	-
VesselNet [47]	2019	0.9802	0.8038	0.9578	0.9821	-	0.9814	0.8132	0.9661	0.9860	-	-	-	-	-	-
Proposed	2020	0.9805	0.8046	0.9581	**0.9827**	**0.8303**	0.9801	**0.8402**	**0.9673**	**0.9874**	**0.8248**	0.9870	0.7914	**0.9665**	0.9864	**0.8276**

TABLE IV
PERFORMANCE COMPARISON WITH OTHER EXISTING APPROACHES IN TERMS OF AVERAGE MCC

Methods	Year	DRIVE	CHASE	STARE
Azzopardi [37]	2015	0.7475	0.6802	0.7335
Orlando [40]	2016	0.7556	0.7046	0.7417
Zhang [38]	2016	0.7673	0.7324	0.7608
Srinidhi [41]	2017	0.7421	0.6927	0.7398
Proposed	2020	**0.8070**	**0.8070**	**0.8116**

perform better on MCC metric, indicating that the proposed method is robust to class imbalance dataset.

To further evaluate the effectiveness of the proposed method on the three databases, the sample qualitative results of our method and other two state-of-the-art methods are present in Fig. 6. From the figure, it can be seen that compared with the methods proposed by Jin *et al.* [46] (column e), our method (column c) can segment more vessels from the background including the thin vessels with lower contrast and vessels in the presence of central vessel reflex. The method proposed by Yan *et al.* [45] proposed a joint loss function that pays more attention to detecting thin vessels, but many background pixels are misclassified to the vessel pixels. Our method not only can detect thin vessels, but also performs better in avoiding misclassifying the background as vessels.

C. Ablation Studies

The proposed AACA-MLA-D-UNet has three unique modules, i.e. the dropout dense block, the adaptive atrous channel attention module, and the multi-level attention module. To evaluate the impact of each component in the proposed method, we perform extensive ablation studies. We compare our AACA-MLA-D-UNet with (1) Baseline (U-Net [16]), (2) D-UNet, (see Section II-B), (3) AACA-D-UNet (D-UNet + adaptive atrous channel attention), and (4) MLA-D-UNet (D-UNet + multi-level attention). Furthermore, for the dropout dense block, to evaluate how much improvement comes from the change of Baseline (original convolution block vs. dense block), and how the dropout layer influences the performance, we conduct ablation studies on the Baseline, D-UNet without dropout layer (D-UNet w/o dropout) and D-UNet. And to prove the proposed atrous channel attention module is better than the SE block and ECA block in this work, ablation studies on AACA-D-UNet, SE-D-UNet (D-UNet + SE block) and ECA-D-UNet (D-UNet + ECA block) are performed. As DRIVE database has the specified training and test sets, the following ablation analyses are conducted on this database. For fair comparison, all the comparison methods are trained with the same settings. All the results and model sizes in the ablation studies are shown in Table V.

From Table V, it can be seen that directly replacing the original convolution blocks with dense block (Baseline vs. D-UNet w/o dropout) improves the Sen and $F1$ by 0.43% and 0.12%, respectively, while with much lower model complexity. It indicates that the dense block is able to preserve much more vessel information especially low-level one for segmentation. And introducing the dropout layer to the dense block can further improve the $F1$ value by 0.16%. In terms of channel attention model, as discussed in Subsection II-C, compared with SE block, the proposed adaptive atrous channel attention module avoids dimensionality reduction that may affect the performance of the model, while reducing the model complexity. The better results and lower model complexity obtained from the AACA-D-UNet demonstrate the efficiency and effectiveness of the AACA module. And compared with ECA block, the key difference is that the AACA module learns longer range dependencies and non-linear interaction between channels while ECA block does not, and consequently AACA-D-UNet achieves higher results on all the metrics except slightly lower specificity, while maintaining the same model complexity. The results indicates that the proposed AACA module is superior to the two similar channel attention methods in this task. Moreover, it can be seen that the MLA-D-UNet seems to focus slightly more on sensitivity, which achieves better values on Sen and $F1$ than other methods. As the specificity has a tradeoff relationship with sensitivity, the MLA-D-UNet achieves a lower specificity than the other methods. The results indicate that the multi-level

TABLE V
PERFORMANCE MEASURES WITH ABLATION STUDIES

Methods	Spe	Sen	Acc	AUC	$F1$	MCC	#.Param.
Baseline	0.9805	0.7907	0.9564	0.9804	0.8218	0.7978	4.32 M
D-UNet w/o dropout	0.9809	0.7950	0.9572	0.9814	0.8230	0.8020	1.29 M
D-UNet (ours)	0.9818	0.7890	0.9573	0.9818	0.8246	0.8014	1.29 M
SE-D-UNet	**0.9827**	0.7843	0.9574	0.9820	0.8243	0.8015	1.34 M
ECA-D-UNet	0.9824	0.7857	0.9574	0.9819	0.8243	0.8013	1.29 M
AACA-D-UNet (ours)	0.9822	0.7892	0.9577	0.9822	0.8259	0.8030	1.29 M
MLA-D-UNet (ours)	0.9776	**0.8204**	0.9575	0.9823	**0.8310**	0.8056	2.03 M
AACA-MLA-D-UNet (ours)	0.9805	0.8046†	**0.9581**†	**0.9827**†	0.8303†	**0.8070**†	2.03 M

† represents these results have significant difference from the Baseline.

Fig. 7. A test patch from the DRIVE database and the binary segmentation results of the different modules in the proposed method. (a) Pre-processed image, (b) output of Baseline, (c) output of D-UNet, (d) output of AACA-D-UNet, (e) output of MLA-D-UNet, (f) output of AACA-MLA-D-UNet. (g) The ground truth.

TABLE VI
PERFORMANCE MEASURES WITH ABLATION STUDIES ON DEEP SUPERVISION

	Spe	Sen	Acc	AUC	$F1$
w/out DS	0.9791	0.7883	0.9548	0.9772	0.8163
with DS_1	**0.9815**	0.7965	0.9579	0.9825	0.8282
with DS_2	0.9804	0.8001	0.9575	0.9819	0.8273
Proposed	0.9805	**0.8046**	**0.9581**	**0.9827**	**0.8303**

TABLE VII
MODEL COMPLEXITY COMPARISON WITH STATE-OF-THE-ART METHODS

	U-Net	ResUnet	R2U-Net	**Ours**
#.Param.	4.32M	32.61M	39.09M	**2.03M**

attention module has the ability of learning more discriminative information especially detail information, which can make the model detect more vessels from the background. Combining these sub-modules above, the proposed method can achieve the best performance on AUC, Acc and MCC, of which the values are 98.27%, 95.81% and 80.70%, respectively. Compared with the Baseline (UNet), the proposed method achieves higher results on accuracy, AUC, sensitivity, $F1$ score and MCC, and these results are significantly different from UNet according to the Wilcoxon's rank sum test with the statistical significance level set to 0.05. Furthermore, as shown in Table V, the proposed method achieves higher sensitivity (80.46%) in comparison with the UNet (79.07%) with the same specificity, inferring that the proposed method has the better ability of detecting thin vessels than the UNet, and the visual results in column (b) and (d) of Fig. 7 also verify this merit of our method. In addition, the model complexity of the proposed method is less than half that of UNet.

To visualize the segmentation results of each model, Fig. 7 shows a patch containing micro vessels from the fundus image after pre-processing, the segmentation results obtained by the Baseline, each proposed model, and the ground truth, respectively. These figures show that, separately introducing the proposed module to the Baseline in turn can further detect more thin vessels with low contrast and preserve the vessel structures better.

Furthermore, we also evaluate the impact of the deep supervision (DS) mechanism in our proposed model. Experimental results are reported in Table VI. It reveals that model with DS (i.e. proposed) can achieve far better performance compared to model without DS (i.e. w/out DS). Furthermore, to evaluate the impact of DS applied in different stages, we respectively adopt only DS_1 and DS_2 to our model. The results show that the performance obtained by fusing multi-scale information with both DS_1 and DS_2 (i.e. proposed) is significantly better than that obtained only with single stage supervision (i.e. with DS_1 or DS_2).

D. Complexity

In this subsection, we compare our model with other state-of-the-art models in both efficiency (i.e., network parameters) and effectiveness (i.e., Acc, $F1$ and AUC) on DRIVE database. The results of model performance and complexity are shown in TableS VII, III, respectively. It can be seen that the proposed method achieves the best results on AUC, Acc and $F1$ with the lowest model complexity when compared with the other four methods. That is to say, our model performs good on both the effectiveness and efficiency.

V. DISCUSSION

A. Dealing With Challenging Cases

Retinal vessel segmentation has been a challenging task due to the presence of central vessel reflex, low contrast thin vessels and pathological lesions such as exudates, hemorrhages and microaneurysms. Fig. 8 shows the qualitative performance of the proposed method on various challenging cases. As discussed before, the proposed method can extract multi-level information, which can learn more discriminative representation. As a result,

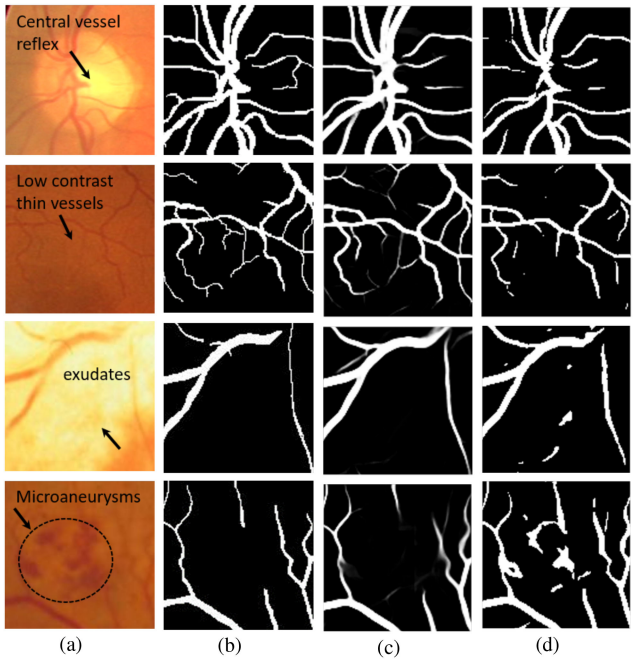

Fig. 8. Some challenging cases of retinal vessel segmentation. (a) Original fundus retinal images, (b) the ground truths, (c) binary output of the proposed method, (d) binary output of Joint Loss [45].

the proposed method show good performance on segmenting thin vessels at low contrast regions (second row). With the presence of central vessel reflex (first row), the proposed method is able to preserve the complete vessel structure. Furthermore, compared to the method proposed by [45], the proposed method can better distinguish the vessels from the pathological lesions within the fundus image patches (third and fourth row). Hence, our method is robust and reliable for retinal vessel segmentation.

B. Generalization

The generalization ability is highly desirable for a CAD system in real application. We evaluate the generalization ability of the proposed method in a cross-training way [33], [45]. Specifically, we apply the model trained on one dataset to test on another dataset for vessel segmentation without fine-tuning. Table VIII gives the performance of the proposed method and other four existing methods, which are trained on the DRIVE database but test on the STARE database and vice versa. It shows that the proposed method achieves the highest AUC, accuracy and specificity for testing on the DRIVE database. However, as the proposed method perform good in the detection of thin vessels and the manual annotations in the DRIVE database contain more thin vessels, applying the model trained on the DRIVE database to the STARE database achieves slightly lower specificity than other methods, but it can achieve the highest AUC and sensitivity for testing on the STARE database. These results indicate that the proposed method has the generalization ability to retinal vessel segmentation.

TABLE VIII
PERFORMANCE OF FOUR EXISTING METHODS AND THE PROPOSED METHOD ON THE DRIVE AND STARE DATABASES USING CROSS-TRAINING STRATEGY

Database	Methods	Spe	Sen	Acc	AUC
DRIVE (trained on STARE)	Fraz [48]	0.9792	0.7242	0.9456	0.9697
	Li [33]	0.9810	0.7273	0.9486	0.9677
	Yan [42]	0.9802	0.7014	0.9444	0.9568
	Yan [45]	0.9815	**0.7292**	0.9494	0.9599
	Proposed	**0.9846**	0.7098	**0.9497**	**0.9731**
STARE (trained on DRIVE)	Fraz [48]	0.9770	0.7010	0.9495	0.9660
	Li [33]	0.9828	0.7027	0.9545	0.9671
	Yan [42]	0.9840	0.7319	**0.9580**	0.9678
	Yan [45]	**0.9840**	0.7211	0.9569	0.9708
	Proposed	0.9732	**0.8079**	0.9559	**0.9735**

VI. CONCLUSION

In this paper, the AACA-MLA-D-UNet is proposed for retinal vessel segmentation. The dropout dense block is used to replace the original convolution block to preserve maximum vessel information between convolution layers and mitigate the over-fitting problem. To sort the importance of each feature channel automatically, the adaptive atrous channel attention module is embedded in the contracting path. To fully exploit the features extracted from the multiple layers of the expanding path, the multi-level attention module is proposed to integrate these multi-level features and use the integrated features to further refine the features at each individual level. Experiments conducted on the DRIVE, STARE and CHASE databases indicate that the proposed method is able to perform better on retinal vessel segmentation when compared with other existing methods. The ablation studies on the DRIVE database demonstrate the effectiveness of the proposed dropout dense blocks, adaptive atrous channel attention and multi-level attention module. And experimental results of the complexity comparison indicate that the proposed method has much lower model complexity compared with other methods. Furthermore, the proposed method can also perform good on some challenging cases and has the generalization ability. In our future work, we will attempt to develop a more discriminative model which can detect the vessels explicitly and further boost the connectivity of the vessel structures.

REFERENCES

[1] M. M. Fraz *et al.*, "Blood vessel segmentation methodologies in retinal images - A survey." *Comput. Methods Programs Biomed.*, vol. 108, no. 1, pp. 407–433, 2012.
[2] J. V. B. Soares, J. J. G. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D gabor wavelet and supervised classification." *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214–1222, Sep. 2006.
[3] C. Köse and C. Ikibas, "A personal identification system using retinal vasculature in retinal fundus images." *Expert Syst. Appl.*, vol. 38, no. 11, pp. 13670–13681, 2011.
[4] C. Kirbas and F. Quek, "A review of vessel extraction techniques and algorithms," *ACM Comput. Surv. (CSUR)*, vol. 36, no. 2, pp. 81–121, 2004.
[5] C. L. Srinidhi, P. Aparna, and J. Rajan, "Recent advancements in retinal vessel segmentation," *J. Med. Syst.*, vol. 41, no. 4, pp. 1–22, 2017.

[6] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Trans. Med. Imag.*, vol. 8, no. 3, pp. 263–269, Sep. 1989.

[7] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Proc.Int. Conf. Med. Image Comput. Comput.-Assist. interv.*. Springer, 1998, pp. 130–137.

[8] J. V. Soares, J. J. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214–1222, Sep. 2006.

[9] J. Mo and L. Zhang, "Multi-level deep supervised networks for retinal vessel segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 12, pp. 2181–2193, 2017.

[10] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2246–2255, Jun. 2020.

[11] L. Wang, L. Zhang, M. Zhu, X. Qi, and Z. Yi, "Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks," *Med. Image Anal.*, vol. 61, p. 101665, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520300311

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2014, pp. 580–587.

[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Euro. Conf. Comput. Vis.* Springer, 2014, pp. 818–833.

[14] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 1736–1744.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

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

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representat.*, 2015.

[18] Y. Wu, Y. Xia, Y. Song, Y. Zhang, and W. Cai, "Multiscale network followed network model for retinal vessel segmentation," in *Proc.Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2018, pp. 119–126.

[19] B. Wang, S. Qiu, and H. He, "Dual encoding U-Net for retinal vessel segmentation," in *Proc.Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer. 2019, pp. 84–92.

[20] L. Mou *et al.*, "CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging," *Med. Image Anal.*, vol. 67, p. 101874, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520302383

[21] L. Xie, L. Zhang, T. Hu, H. Huang, and Z. Yi, "Neural networks model based on an automated multi-scale method for mammogram classification," *Knowl.-Based Syst.*, vol. 208, p. 106465, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705120305943

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Ana. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

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

[25] K. Wan, S. Yang, B. Feng, Y. Ding, and L. Xie, "Reconciling feature-reuse and overfitting in densenet with specialized dropout," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell.*, 2019, pp. 760–767.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[27] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 534–11 542.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[29] Y. Wang *et al.*, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2768–2778, Dec. 2019.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[31] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[32] C. G. Owen *et al.*, "Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (CAIAR) program," *Invest. Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2004–2010, 2009.

[33] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109–118, Jan. 2016.

[34] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 203–210, Mar. 2000.

[35] A. W. Setiawan, T. R. Mengko, O. S. Santoso, and A. B. Suksmono, "Color retinal image enhancement using clahe," in Proc. IEEE *Int. Conf. ICT Smart Soc.*, 2013, pp. 1–3.

[36] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.

[37] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Med. Image Anal.*, vol. 19, no. 1, pp. 46–57, 2015.

[38] J. Zhang, Y. Chen, E. Bekkers, M. Wang, B. Dashtbozorg, and B. M. ter Haar Romeny, "Retinal vessel delineation using a brain-inspired wavelet transform and random forest," *Pattern Recognit.*, vol. 100, no. 69, pp. 107–123, 2017.

[39] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2369–2380, Nov. 2016.

[40] J. I. Orlando, E. Prokofyeva, and M. B. Blaschko, "A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 16–27, Jan. 2017.

[41] C. L. Srinidhi, P. Aparna, and J. Rajan, "A visual attention guided unsupervised feature learning for robust vessel delineation in retinal images," *Biomed. Signal Process. Control*, vol. 44, pp. 110–126, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809418301010

[42] Z. Yan and Yang, X. and Cheng, Kwang-Ting, "A three-stage deep learning model for accurate retinal vessel segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1427–1436, 2018.

[43] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[44] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[45] Z. Yan, X. Yang, and K.-T. Cheng, "Joint segment-level and pixel-wise losses for deep learning based retinal vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1912–1923, Sep. 2018.

[46] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119301984

[47] Y. Wu *et al.*, "Vessel-Net: Retinal vessel segmentation under multi-path supervision," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv,*Springer, 2019, pp. 264–272.

[48] M. M. Fraz *et al.*, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, Sep. 2012.