An Encoder-decoder Deep Learning Model Combining Mixed Attention Mechanism and Asymmetric Convolution for Automation of Retinal Vessels Segmentation

Jiajia Cao, Qin Zhou, Leo Yi Chen*, *Senior Member, IEEE*, Lin Yin and Fei Zhang

***Abstract*—The segmentation of the retinal vascular tree is the fundamental step for diagnosing ophthalmological diseases and cardiovascular diseases. Most existing vessel segmentation methods based on deep learning give the learned features equal importance. Ignored the highly imbalanced ratio between background and vessels (the majority of vessel pixels belong to the background), the learned features would be dominantly guided by background, and relatively little influence comes from vessels, often leading to low model sensitivity and prediction accuracy. The reduction of model size is also a challenge. We propose a mixed attention mechanism and asymmetric convolution encoder-decoder structure(MAAC) for segmentation in Retinal Vessels to solve these problems. In MAAC, the mixed attention is designed to emphasize the valid features and suppress the invalid features. It not only identifies information that helps retinal vessels recognition but also locates the position of the vessel. All square convolutions are replaced by asymmetric convolutions because it is more robust to rotational distortions and small convolutions are more suitable for extracting vessel features (based on the thin characteristics of vessels). The employment of asymmetric convolution reduces model parameters and improve the recognition of thin vessel. The experiments on public datasets DRIVE, STARE, and CHASE_DB1 demonstrated that the proposed MAAC could more accurately segment vessels with a global AUC of 98.17%, 98.67%, and 98.53%, respectively. The mixed attention proposed in this study can be applied to other deep learning models for performance improvement without changing the network architectures.**

***Index Terms*—Deep learning, retinal vessel segmentation, attention mechanism, asymmetric convolution**

I. INTRODUCTION

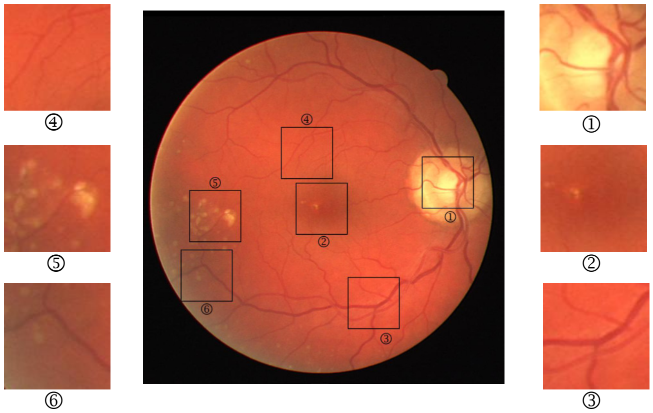


Fig. 1. A fundus retinal image from the DRIVE database. ① Bright area. ② Macular area. ③ Thick vessels with high contrast. ④ Thin vessels with low contrast. ⑤ Erea of lesion. ⑥ Dark area.

Fundus photography is a non-invasive method of fundus examination. It is commonly used to diagnose retinal diseases, such as macular degeneration, diabetic retinopathy, and glaucoma. It is also used to diagnose cardiovascular diseases, such as hypertension. The significant causes of vision loss are age-related macular degeneration, diabetic retinopathy, and glaucoma [1]. Early diagnosis of these diseases can prevent blindness. Hypertension is one of the most common chronic diseases and the most important risk factor for inducing various cardiovascular and cerebrovascular diseases. The earlier the hypertension is controlled, the less harm to cardiovascular and cerebrovascular diseases [2]. Studies have shown that microaneurysm and vessel diameter are two important biomarkers for diagnosing diabetic retinopathy [3]. Retinal blood vessel ratio is an important clinical parameter for the diagnosis of glaucoma [4]. There are three changes in the blood vessels of the fundus of patients with hypertension: 1) classic retinal vascular changes in response to blood pressure (hypertensive retinopathy), 2) changes in retinal vascular caliber, and 3) changes in global geometric patterns of the retina [5]. Analysis of the retinal blood vessels is a meaningful way to diagnose ophthalmological diseases such as diabetes, and glaucoma [6], [7]. And the segmentation of the vessel tree is a basic step to extract these quantitative features. In the past, vascular tree segmentation was carried out by professional ophthalmologists, which was very tedious and time-consuming. With the development of artificial intelligence, the automatic segmentation of retinal vessels has been realized, which saves doctors' diagnosis time. However, due to the inconsistency in the size, shape, contrast, and intensity level of retinal vessels in different local areas(see Fig. 1), the automatic segmentation of retinal vessels is still recognized as a challenging task [8].

In recent years, many algorithms for automatic retinal vessel segmentation have been proposed. In general, retinal vessel segmentation algorithms can be summarized into six categories:(1)methods based on vessel tracking, (2)methods based on matched filtering, (3)methods based on morphological operations, (4)methods based on deformation models, (5)methods based on traditional machine learning, (6)methods based on deep learning [4]. Given the outstanding performance of deep learning in the field of computer vision, and has gradually become a dominative method in medical image processing, this paper

This work was supported by the Ministry of Science and Technology of China (Grant No. 2018AAA0101300), the Guangdong Regular Institutions of characteristic innovation project (No.2017KTSCX176), the DGUT innovation center of robotics and intelligent equipment of China (No.KCYCXPT2017006), KEY Laboratory of Robotics and Intelligent Equipment of Guangdong Regular Institutions of Higher Education (No.2017KSYS009), The National Natural Science Foundation of China (No. 62073091), the Key fields (new generation information technology) special projects of universities in Guangdong Province (No.2020ZDZX3042), the Open Fund of Hunan Provincial Key Laboratory of Mechanical Equipment Health Maintenance(No.21903), and the Royal Society International Exchanges Scheme (IEC-NSFC-201029).

Jiajia Cao and Qin Zhou are with School of Computer Science and Technology, Dongguan University of Technology, Dongguan, 523808 China (e-mail: jiajia.cao@i4ai.org, qin.zhou@i4ai.org).

Leo Yi Chen* is with School of Engineering, Newcastle University Newcastle upon Tyne, NE1 7RU, United Kingdom (email: leo.chen@newcastle.ac.uk).

Lin Yin and Fei Zhang are with School of Mechanical Engineering, Dongguan University of Technology, Dongguan, 523808 China (e-mail: yinl@dgut.edu.cn, zhangfei@dgut.edu.cn).

mainly reviews the retinal vessel segmentation methods based on deep learning in the past five years.

U-Net model [9] was initially proposed to be used to segment cells in electron microscope images. It consists of an encoder and a decoder. The encoder is used to extract features of the shallow layer, while the decoder is responsible for the recovery of images and the extraction of abstract features. In addition, a cascade structure is proposed, which can realize the fusion of the shallow feature map and the deeper feature map to obtain more comprehensive context information. Because the encoding and decoding parts are symmetrical, like a U shape, it is called U-Net. Compared with other deep learning models, it has the advantages of fewer training parameters, faster calculation speed, less annotated data needed, and acceptable input of any shape. In particular, the advantages of low demand for annotated data and simple and lightweight are consistent with the characteristics of the small sample and simple semantic information of medical data (the amount of annotated data is limited due to the privacy and professionalism of medical data). Therefore, U-NET is widely used in medical image segmentation. To eliminate the impact of subjective factors, Li et al. [8] treat the segmentation problem as a modal conversion problem from fundus images to vessel images. They used the deep neural network to learn the transform function and achieved the segmentation results with considerable accuracy without artificial design features and image preprocessing. Literature et al. [10], [11] realize that there are two problems in the traditional deep-learning-based methods that regard the vessels segmentation problem as pixel classification. Firstly, the non-local connections between individual pixels or image blocks are ignored. The second is that both the training and testing phases require much computation. To obtain a rich hierarchical representation, the author innovatively transforms the segmentation problem into an edge detection task. Firstly, the convolutional layer is used to extract the multi-scale discriminant features. Secondly, the side output layer is used together with the early convolutional layer to generate local output. Finally, the CRF(Conditional Random Field) layer is used to process the non-local connection between pixels. Because the non-local information between pixels is considered, the segmentation accuracy is improved. Dai et al. [12] considering the limitations of the inherent form of traditional convolution kernel, replaced the traditional convolution kernel with a deformable module based on U-NET, so that kernel size of convolution and the receiving field can be adjusted adaptively according to the size and shape of the object so that more complex vessel structures can be detected. Compared with U-NET, it can extract more tiny blood vessels, but at the cost of increased training and reasoning time [13]. To solve the problem of imbalanced pixels distribution in fundus images, a series of studies [14]–[16] focused on multi-scale feature extraction to achieve local feature extraction and global feature extraction, thereby achieve segment more tiny vessels. Yan et al. [17] proposed a three-stage network model to solve the imbalance between thick and thin vessels and the characteristic differences. The first step is to segment the thick vessels, then segment the thin vessels, and fuse the two segmentation results. Each model has its loss function, which prevents the problem of loss dominated by thick vessels. In the end, low-contrast vessels can also be extracted effectively, and vessels not annotated by human experts are also extracted [17]. Cherukuri et al. [18] to solve the problems that current deep-learning-based methods rely heavily on the quality of training data, ignore the processing of background noise similar to vessels and the segmentation of thin vessels. A segmentation network combining geometric features prior is proposed. Firstly, a representation network is applied to learn the geometric features of the retinal image. In order to promise the actual effectiveness of the REPRESENT filter, Cherukuri et al. proposed a directional constraint and an adaptive noise regularizer to ensure the geometric diversity of the curve. Then a task network with residual module and multi-scale representation is used to classify the previous feature representation results. This architecture can detect thin vessels accurately, while the network size is very simple and the inference time is fast. In summary, the segmentation of fundus microvessels is still the key challenge for the segmentation of fundus vessels. The continuity of blood vessels is another problem that remains to be solved, and there are few studies focus on improving the sensitivity of vessel segmentation.

In this paper, we propose a new retinal vessel segmentation model based on encoding and decoding structure. This model incorporates hybrid attention and asymmetric convolution. Mixed attention is consists of channel attention and spatial attention, which are joined by the design method. Asymmetric convolution is introduced to reduce the number of model parameters. Motivation one there is an imbalance between coarse and fine vessels, background and target. Previous methods did not notice such imbalances and treated all features equally, which will lead to the neglect of valuable features learned from a small sample. We applied attention to focus on effective features and suppress invalid features so that useful features are emphasized and preserved. In addition, without a separate positioning module, the proposed method can also locate the region of interest. Another motivation is that the model needs to ensure the lightness of the model and the inference speed is fast for practical application. In order to maintain the lightness of the model, an asymmetric convolution module is introduced to reduce the number of model parameters without losing precision. The main contributions of this paper include:

(1)We propose a novel model for retinal vessel segmentation that incorporates the attention mechanism, in which attention gives different weights to different features. The model can achieve higher sensitivity and prediction accuracy for retinal vessels segmentation.
(2)We propose an improved hybrid attention mechanism that can be integrated into any convolutional neural network (CNN). It combines the features extracted by the deep layer to provide more accurate channel attention and spatial attention.
(3)We introduced the asymmetric convolution module, which can save lots of training parameters.

The organization of this paper is as follows. Section II describes the details of the proposed method. Section III contains details of the data and experiments. In Section IV, we show the results of the proposed method on three benchmark datasets with figures and tables. The effectiveness of the proposed method is also demonstrated by comparing it with the existing methods and constructing ablation experiments and cross-training experiments. Finally, we analyzed the results and shortcomings of the current method and looked forward to the future research trends and our next research work in SectionV.

II. METHODOLOGY

The proposed mixed attention mechanism and asymmetric convolution encoder-decoder structure, which we called MAAC encoder-decoder, is novel in the mixed attention mechanism with signal and the use of asymmetric convolution module(see Fig.2). The backbone used in the MAAC model is based on the U-NET, which has an encoder-decoder structure. The mixed attention module is arranged in the skip connection to refine the image features learned by the shallow layer, i.e., the mixed attention module helps to focus on the meaningful features along the channel and spatial dimensions and suppress unnecessary ones. All the square convolution kernels of the backbone were replaced with asymmetric convolution kernels to reduce training parameters, reduce calculation, and enhance the convolution kernel's robustness.

A. The Framework of MAAC Encoder-decoder

The backbone used in the MAAC encoder-decoder is composed of encoder and decoder(see Fig.2). In the encoding phase, the features of the input image were extracted and refined. In the decoding phase, the feature maps were restored to their original size layer by layer through the up-sampling layers. The whole MAAC encoder-decoder consists of five sub-modules. Each sub-module contains an asymmetric group composed of two asymmetric convolution layers with kernel size 3*1 and 1*3 connected in series, the strides of each asymmetric convolution kernel is 1, and padding is set to 'same.' A major advantage of asymmetric convolution kernel is that it can reduce the number of training parameters of models and increase the robustness of convolution. Then, the asymmetric convolution groups are followed by a batch normalization layer (BN) [19], a ReLu [20] activation function layer, a dropout layer with a dropout rate of 20%, and identical asymmetric convolution groups, a BN layer, and a ReLu activation function layer. The batch normalization was used to accelerate model convergence and alleviate the "gradient dispersion" problem in the deep network. We made the arrangement of activation layer followed by BN layer because the experiment proved that such arrangement would produce better results than placing activation function before BN layer. The dropout layer was adopted to prevent model overfitting. It is worth noting that each submodule in the decoding phase has a cascade layer before the convolution layer. The channels of the convolution layer in each submodule are 32, 64, 128, 64, 32. In addition, the encoder contains two 2*2 max-pooling with a stripe of 2 to reduce the feature map, and the decoder uses two 2*2 up-sampling with a stripe of 2 to restore the size of the feature maps. In addition, there are skip connections between the submodules of the encoder and the corresponding submodules of the decoder. For the segmentation task, spatial information plays an important role. But with deepens of the network level, the spatial information of the feature map and the detailed features of the image will be gradually lost, which is not beneficial to the accurate segmentation of the image. Moreover, with the refinement and abstraction of features layer by layer, lots of information to assist in image restoration will be missing. Skip connections can combine the detailed features extracted from the shallow layer with the abstract features learned from the deep layer, and provide multi-scale and multi-level information for the image recovery during up-sampling, so more fine segmentation results can be obtained. A mixed attention module is placed on the corresponding skip connection. For more details on mixed attention, see the following subsection, Mixed Attention Mechanism.

B. Mixed Attention Mechanism

Woo et al. proposed a Convolutional Block Attention Module(CBAM) and validated that utilizing a combination of spatial attention and channel attention outperforms using only the channel attention independently [21]. Oktay et al. [22]point out that a grid-based gating can allow attention coefficients to be more specific to the local region and can get better performance than gating based on a global feature vector. Inspired by Woo et al. and Oktay et al., we improved the Convolutional Block Attention Block proposed in [21] by introducing grid-based gating. We named the improved CBAM "Gating-triggered Convolutional Block Attention Module(GCBAM)," as shown in Fig 3. Given an intermediate feature map F $\in R^{C*H*W}$ and a trigger signal g $\in R^{C*H*W}$ as input, they are first processed by the channel attention module to infer a channel refined feature $F' \in R^{C*H*W}$ which is the product of channel attention coefficient $M_c \in R^{C*1*1}$ and F, and then input F' and signal g into spatial attention module and get $F'' \in R^{C*H*W}$ which is the product of spatial attention coefficient $M_s \in R^{1*H*W}$ and F'. The mathematical expression of the whole processing process is as follows:

$$\begin{aligned} F' &= M_c(F, g) \otimes F, \\ F'' &= M_s(F', g) \otimes F'. \end{aligned} \tag{1}$$

Where $\otimes$ denotes element-wise multiplication. $M_c()$ refers to the channel attention processing while $M_s()$ means the spatial. F'' is the final output of the GCBAM. The following describes the details of each submodule in the GCBAM.

1) Channel Attention Module: In a convolutional neural network, the number of channels in the feature map is equal to the number of learned features, and the 2D map of each channel represents one learned feature. In general, research on image segmentation treats each channel of feature maps equally, neglecting the different importance of different features to the segmentation task, which is not conducive to further improving the accuracy of image segmentation. In the task of blood vessel segmentation, the features that are beneficial to vessel recognition should be emphasized, and the background information and so on that is not beneficial to vessel recognition should be suppressed. Channel attention is to sort each channel by assigning different weights representing importance to each channel, calculated by backpropagation. The essence of channel attention is to focus on" what" is meaningful of an input image. The gating signal g comes from the deeper layer and has rich contextual information, which can assist in selecting meaningful channels of the intermediate feature map that comes from the shallow layer. In addition, the gating signal g aggregates multi-scale image information to improve the resolution of the feature map after attention refinement. This operation is helpful to improve the segmentation performance.

The detailed computation process is shown in Fig. 4. First, max-pooling and average-pooling are performed on intermediate feature F and gating signal g, respectively. Generating two spatial context vector F^c_{max} and F^c_{avg} about F and two spatial context vector g^c_{max} and g^c_{avg} about g respectively. Woo et al. [21] point out use both average-pooled and max-pooled features simultaneously is helpful to gather object features that are used to infer channel-wise attention. Next, the sum of F^c_{max} plus g^c_{max} is fed into the multi-layer perceptron (MLP), and the sum of F^c_{avg} plus g^c_{avg} as well. The MLP that has one hidden layer with C/16 neurons is shared by them. After shared MLP processing, two output feature vectors are generated and merged by element-wise summation. Finally, our channel attention map $M_c \in R^{C*1*1}$ is obtained after the Sigmoid activation function is applied to the merge vector. The channel attention calculation process is expressed as:

$$\begin{aligned} M_c(F, g) =& S(MLP(AvgPool(F) \oplus AvgPool(g)) \oplus \\ & MLP(MaxPool(F) \oplus MaxPool(g))) \\ =& S(MLP(F^c_{avg} \oplus g^c_{avg}) \oplus MLP(F^c_{max} \oplus \\ & g^c_{max})) \end{aligned} \tag{2}$$

where S refers to the Sigmoid activation function. The $\oplus$ means element-wise summation. The MLP denotes the multi-layer perceptron, which has one hidden layer followed by a ReLu activation function. The grid signal g needs additional processing to be consistent with the F dimension. Firstly, g is upsampled by 2*2, and then the output of upsampled is convolved with a convolution whose kernel size is 3*3, and the number of channels is equal to F. Finally, the final output gating signal g is obtained after normalization of the BatchNormalization layer and activation of ReLu function.

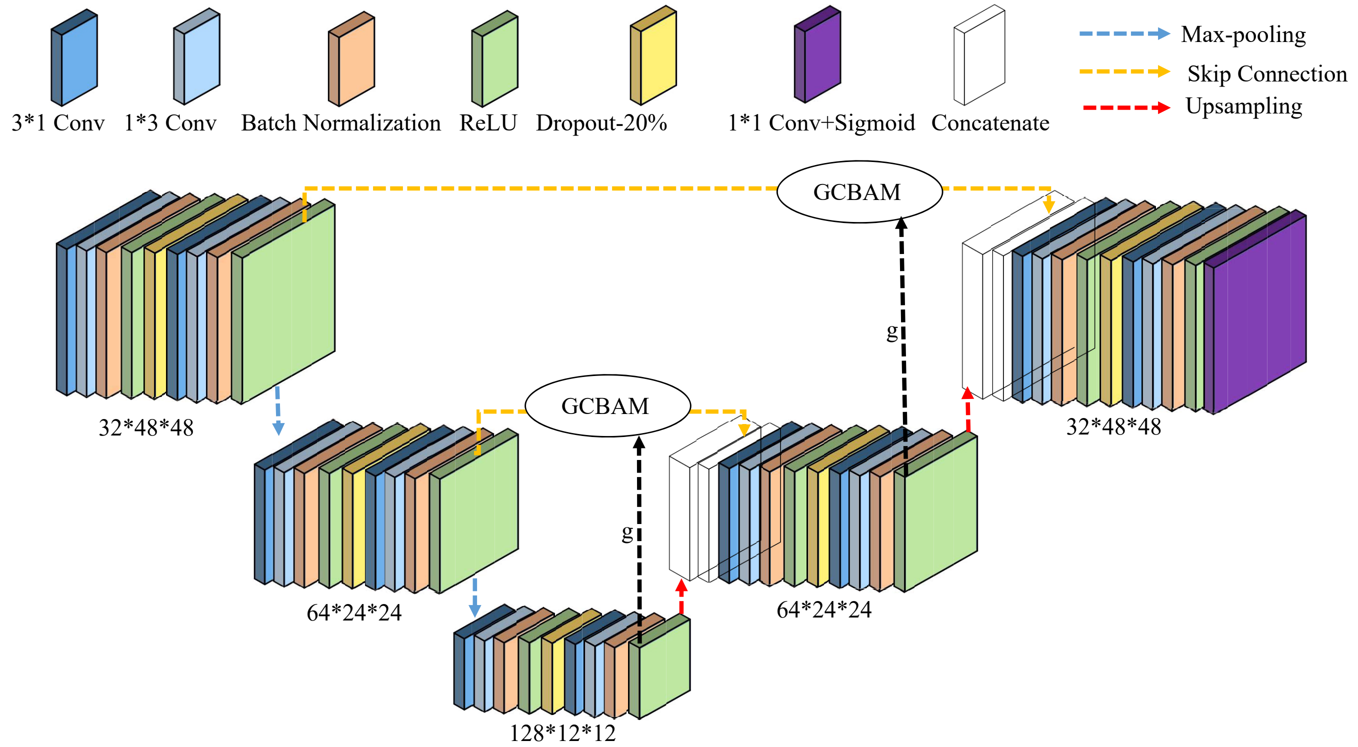


Fig. 2. The overview of the proposed MAAC encoder-decoder deep learning framework. The left side of the network is the encoder, which gradually reduces input image by 2x down-sampling. On the right is the decoder, which mainly for restoring the image. The GCBAM generates attention coefficients by combining channel and spatial attentions.

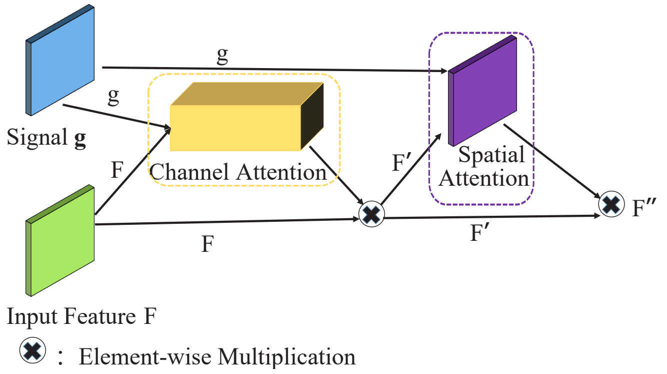


Fig. 3. The overview of GCBAM. The module is composed of spatial attention module and channel attention module. Input features are refined with attention coefficients computed in GCBAM. Attention informations are supplemented by analysing both the activations and contextual information provided by the gating signal (g) which is collected from a coarser scale.

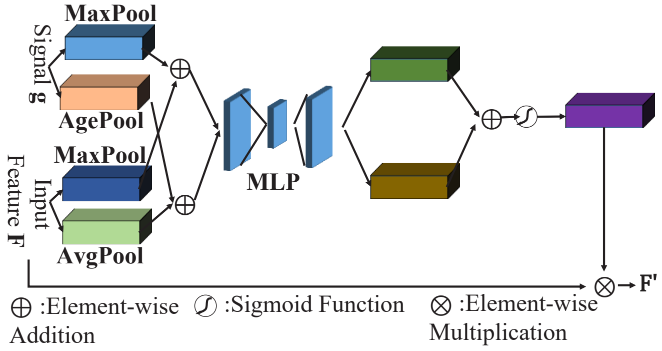


Fig. 4. Diagram of channel attetion. It contains max-pooling, average-pooling, and a shared network.

2) Spatial Attention Module: Unlike channel attention, spatial attention focus on ”where” is important of an input image. For the image segmentation task, semantic segmentation is the separation of the target region from the background. So, the location of the target area will contribute to the improvement of segmentation performance. In previous studies, they would set additional ROI(region of interest) extraction and object localization modules to improve the accuracy of the segmentation task. Different from the previous studies, the spatial attention used in this paper is soft attention, which can be calculated without additional modules. The detail of the spatial attention generation process(see Fig. 5) is described in the following paragraphs.

Firstly, we apply max-pooling and average-pooling operation along the channel axis to aggregate spatial information for each pixel in the output F' of channel attention and do the same for gating signal g. The gating signal g is equal to the g in channel attention. After the above operation, generating two spatial feature map $F^s_{max} \in R^{1*H*W}$ and $F^s_{avg} \in R^{1*H*W}$ about F' and two spatial feature map $g^s_{max} \in R^{1*H*W}$ and $g^s_{avg} \in R^{1*H*W}$ about g. And then the maps about F' are concatenated, and the maps about g are concatenated. After, they are added by element-wise summation and convolved by a standard convolution layer. The convolution layer is followed by a Batch Normalization layer. Finally, spatial attention $M_s(F', g) \in R^{1*H*W}$ was obtained Sigmoid function activation.

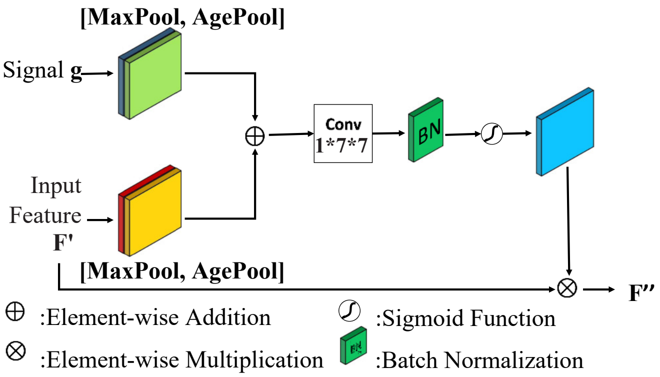


Fig. 5. Diagram of spatial attetion. It takes the cascading sum of the max-pooling and average-pooling outputs of the signal g and the input features F, and passes them to the convolution layer.

The spatial attention is calculated as follows:

$$M_s(F',g) = S(BN(f^{7*7}([MaxPool(F');AvgPool(F')]\oplus [MaxPool(g);AvgPool(g)]))) \quad (3)$$
$$= S(BN(f^{7*7}([F^s_{max};F^s_{avg}] \oplus [g^s_{max};g^s_{avg}]))),$$

where S denotes the sigmoid and BN denotes the Batch Normalization with a momentum equal to 0.01 and epsilon equal to 1e-5. Where $\oplus$ refers to element-wise summation. And f^{7*7} represent the convolution layer with the kernel size of 7*7.

In summary, the proposed spatial attention has two advantages. One is to help target positioning by suppressing unrelated background responses. The second is that there is no need to train the additional network module used to extract ROI.

C. Asymmetric Convolution

In convolutional neural networks, the model's performance is evaluated by time complexity T and space complexity S. T determines the model's training time and prediction time. The larger the T, the longer the training and testing time of the model. S determines the number of parameters of the model. The more parameters of the model, the more data is needed to train the model. Mathematically, the calculation of T is given by:

$$T \sim O(\sum_{j=0}^{D} M_j^2 \cdot K_j^2 \cdot C_{j-1} \cdot C_j), \quad (4)$$

$$M = (X - K + 2 * Padding)/Stride + 1, \quad (5)$$

where D is the total number of convolution layers in the model, M_j is the side length of the feature map output by the jth convolutional layer, K_j is the size of the jth convolution kernel, C_{j-1} is the number of channels for the output of the (j-1)th convolution kernel, C_j is the number of channels for the j convolution kernel. Among them, the calculation of M is shown in (5), where Padding is the padding size of original image, Stride is the step size of the convolution kernel. It can be inferred from (4) and (5) that T is related to the size of the convolution kernel, and the larger the convolution kernel, the larger T.

The space complexity S is determined by the total parameters of the model(P) and the size of the feature map output by each layer of the model during operation(F). The calculation of S is given by:

$$S \sim O(P + F),$$
$$S \sim O(\sum_{j=0}^{D} K_j^2 \cdot C_{j-1} \cdot C_j + \sum_{j=0}^{D} M_j^2 * C_j), \quad (6)$$

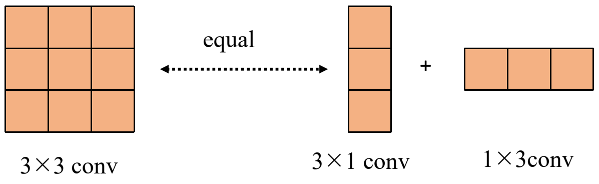


Fig. 6. The square convolution is replaced by asymmetric convolution.

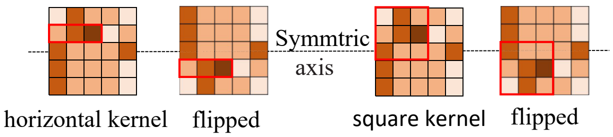


Fig. 7. If the image is flipped upside down, the horizontal 1*3 convolution can still learn the same features at a symmetrical position, but the 3*3 square convolution learns features at this position different from the original ones.

the meaning of the parameter is the same as in T. Refer to (6), P depends only on the size of the convolution kernel, the number of channels and the number of layers D. F depends only on the feature map M and the number of channels C. According to (4), (5) and (6), it is not difficult to conclude that, on the premise that other variables remain unchanged, the smaller the convolution kernel, the smaller the time and space complexity of CNN. Thus, 3 × 3 convolution kernel is adopted by this model.

[23] pointed out that the result of using n*1 and 1*n convolution cascade to do a convolution operation on a 2D image is equivalent to the result of using n*n convolution to convolve in the same 2D image. Asymmetric convolution is the decomposition of a square n*n convolution into a sequence of two layers with n * 1 and 1 * n kernels(see Fig.6). Inspired by [23], all the standard 3*3 convolutions in the model were replaced with 3*1 and 1*3 convolution cascade, except the attention module and the gating signal g processing module. This operation can further improve computational efficiency and reduce the parameters of the model. Based on the formula above (6), if 3*3 square convolution, the number of parameters is $9C^2$, if 3*1 and 1*3 convolution cascade, the number of parameters is $6C^2$. The number of parameters required is saved by ($9C^2$ - $6C^2$)/$9C^2$ = 33% if the asymmetric convolution's input and output filter channels are equal to the original square convolution. Similarly, based on (4), asymmetric convolution can reduce computation by 33%, but only if the padding is equal to the same. In addition, we assume that the employment of asymmetric convolution can enhance the model's robustness to rotational distortions [24]. As shown in Fig.7, when the model inputs an upside-down image, the 3*3 convolution will extract meaningless features, but the horizontal filter will produce the same output as the original image at the axisymmetric position. Same thing with vertical convolution.

III. EXPERIMENTS

A. Database

The proposed method was evaluated on three public datasets, digital retinal images for vessel extraction(DRIVE), structured analysis of the retina(STARE), and the CHASE_DB1. The following describes the detail of each dataset listed above:

1) DRIVE: The DRIVE dataset was first constructed by Staal et al. [25] for automated segmentation of retinal vessels. It is now the benchmark for comparison of all segmentation methods of the blood vessel in retinal images. The dataset consists of 40 fundus images randomly selected from a diabetic retinopathy screening program in the Netherlands. The screening population was between the ages of 25 and 90. Of the 40 images, only 7 show signs of early diabetic retinopathy, while the remaining 33 did not show any sign of diabetic retinopathy. All images were taken with a Canon CR5 Nonmydriatic 3CCD camera at 45° field of view. Each image has a resolution of 565 by 584 pixels, with 8 bits per color channel. Each image is cropped around the field of view(FOV) of approximately 540 pixels in diameter, and each provides a corresponding binary FOV mask. The 40 images are officially divided into a training set and a test set, each containing 20 images. For the training set, only one manual vascular segmentation is provided as ground truth. For the test set, two manual vascular segmentation are provided, one of which is used as the gold standard for segmentation, and the other is used to compare automatic and human segmentation. All the observers who manually segmented the vessels were trained by an ophthalmologist.

2) STARE: The STARE project was first created by Michael Goldbaum, M.D., in 1975. Many scientists have contributed to the project since it was established. Among them, Hoover et al. [26] selected 20 images for manual labeling as the ground truth vessel segmentation. And, these 20 pictures were used for the blood vessel segmentation work. So this dataset contains a total of 20 images, 10 of which do not have pathology, and the other 10 have pathology. Each image was captured using 24 bits per pixel at 700 by 605 pixels. The slides were acquired using a TopCon TRV-50 fundus camera with a 35° field of view. For this dataset, two manual segmentations marked by two experts with rich experience in image processing and retinal image analysis are provided. Generally, the segmentation of the first expert is used as ground truth [27], and the segmentation of the second expert is used as a reference for performance comparison [26]. In addition, the dataset does not provide an official data split, nor does it provide a binary FOV mask for each image.

3) CHASE_DB1: CHASE_DB1 was provided by FRaz et al. [28] and is a subset of the Child Heart and Health Study in England (CHASE) dataset, which was initiated by Owen et al. [29]. CHASE_DB1 contains 28 retinal digital images collected from 28 eyes of 14 children (aged ten years). The resolution of each image is 999 by 960 pixels. Each image of the child's eye was acquired using the handheld fundus camera(NM-200-D; Nidek Co., Ltd., Gamagori, Japan) with a 30-degree field of view (FOV). There are two sets of ground-truth vessel annotations. In the study of retinal vessel segmentation, the first set is generally selected for training and testing, and the second set is generally used as a "human" baseline. Like the STARE dataset, the CHASE_DB1 dataset also neither provides an official data split nor binary FOV mask.

B. Preprocessing

In this paper, to simplify the calculation, reduce noise and enhance the segmentation target, all retinal images(both the training set and the test set) are preprocessed as follows.

First, all color images were converted to grayscale images to reduce the impact of light and reduce the computation required for subsequent processing. Second, image standardization was provided to implement centralized processing. Image normalization followed immediately to convert the data to between 0 and 1. The image details are often unclear in the actual photographing process due to equipment, shooting Angle, and other reasons. Contrast Limited Adaptive Histogram Equalization(CLAHE) was adopted to enhance the image's contrast and make the details clearer. Here, tileSize is set to 8*8, and the clipLimit is set to 2. And then, the gamma correction with a gamma value of 1.2 was adopted to highlight dark field color details, make image brightness consistent, and increase contrast. Finally, all images were divided by 255 to reduce the pixel value to [0,1]. The example of the preprocessed image of DRIVE is shown in Fig. 8.

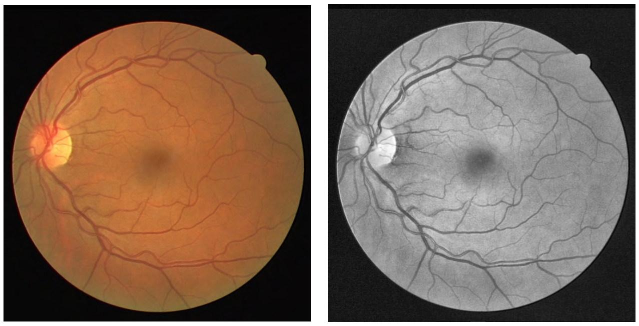

(a) Original image (b) Pre-processed image

Fig. 8. An example of pre-processed image from DRIVE

As we all know, training a deep neural network needs many data. But as described in Section Database, the number of densely annotated images in the dataset is relatively small. To solve this issue, we adopted six data augmentation methods. That includes translation, flipping along the horizontal and vertical axis, and rotating by an angle from [90°, 180°, 270°]. After each image is augmented, there are 7 in total, including the original image. Note that the data augmentation is performed on the entire image and only on the image of the training set.

Just as STARE and CHASE_DB1 have no official data split, we must separate the training and test sets ourselves. For STARE, there are two split methods used in the general study. In this paper, we use the method called "leave-one-out," in which each image is tested, and training is conducted on the remaining 19 samples [30]. That is, the training-testing cycle will be repeated 20 times. As for CHASE_DB1, we randomly selected 20 images as a training set and the remaining eight images as a test set, as most studies do. Since STARE and CHASE_DB1 did not provide the mask of FOV, we apply the color threshold to generate the mask.

C. Patch Extraction

To further increase the training samples, patches of the preprocessed images are adopted to train the neural network. In this paper, the dimension of a patch is 48*48. Each patch is randomly generated by randomly selecting its center point in the whole augmented image. For the neural network to distinguish the FOV border from blood vessels, we select patches that are partially or wholly outside the FOV. The patches extraction rules of the training set on the three datasets are the same. The DRIVE and CHASE_DB1 datasets both are extracted 448000 patches from their respective training sets. That is, randomly extracting 3200 patches in each of the 140 their respectively training images(including augmented and original images). For the STARE dataset, a total of 425600 patches were randomly extracted from 133 training images(including augmented and original images), i.e., 3200 patches were extracted per image, and 20 cycles were performed. The remaining unaugmented image is used as the test image.

The patch is also used for testing, with the same dimension as the patch in the training set. However, different from the training set's patch extraction method, the test set extracts patches sequentially, convenient for image recovery. We sample the center points of the patch with a stride of 5 pixels along the width and height to obtain multiple consecutive overlapping patches for each image. Note that if the image does not exactly extract an integer number of patches, the image will be filled. Then, the average multiple prediction probability of a pixel is calculated as the final prediction result of the pixel. This operation helps improve segmentation performance.

D. Implementation Details

For the DRIVE, we use the official data split. For the STARE and CHASE_DB1, we use the method described in the Preprocessing section to split the data. The number of patches extracted for training from each dataset is described in the Patches extraction section. Of these, 90% of the training set is used for training, and the remaining 10% is used for validation. We train our model with a batch size of 32, with a maximum number of iterations of 50. The optimizer selected the Adam optimizer with the default setting. The binary cross-entropy loss function with a threshold of 0.5 is used as the loss function. To demonstrate the effectiveness of the proposed model, the above Settings are valid for all three datasets.

We were coding in python 3.6.9. The neural network is developed on Keras with TensorFlow 2.3.0 as the backend. The IDE(Integrated Development Environment) is Spyder. The experiment is run on Tesla v100 GPU on HPC(Hign Performance Computing).

E. Evaluation Metrics

The blood vessel segmentation task is a two-category classification problem, and the pixels in the image either belong to the positive or negative category. After the segmentation task is completed, each pixel has four possible classifications: true positives(TP), false positives(FP), true negatives(FN), and false negatives(TN). TP refers to the number of correctly classified samples as positive examples, and FP refers to the number of samples incorrectly classified as positive examples. TN denotes the number of correctly classified samples as negative examples, and FN denotes the number of samples incorrectly classified as negative examples. Sensitivity(SE), specificity(SP), accuracy(ACC), precision(PR) are the most commonly used evaluation indicators for blood vessel segmentation tasks. Calculated as follow:

$$SE = \frac{TP}{TP+FN}, \qquad SP = \frac{TN}{TN+FP}$$
$$ACC = \frac{TP+TN}{TP+TN+FP+FN}, \qquad PR = \frac{TP}{TP+FP}. \tag{7}$$

In theory, the higher the ACC, the better the classifier. The PR represents the accuracy of the prediction in the positive sample results. Sensitivity measures the classifier's ability to recognize positive examples, while specificity measures the classifier's ability to recognize negative examples. Under normal circumstances, we hope that the values of SE and SP are both high, but in fact, we will find a balance point between SE and SP. At this time, the receiver operating characteristic(ROC) curve was introduced to express this process. We also adopt the area under the ROC curve(AUC) metric to clarify which classifier is better. Other evaluation metrics like the precision-recall curve, area under the precision-recall curve(AUPR), the harmonic mean of precision, and recall f1-score were also adopted.

IV. RESULTS

To verify the effectiveness of the proposed method, we evaluate the proposed method in DRIVE, STARE, and CHASE_DB1 by calculating the ACC, SE, SP, AUC, AUPR, F1-SCORE, and Precision. The proposed method is also compared with the existing techniques to observe the advantages and disadvantages of the proposed method intuitively. In addition, to prove the effectiveness and generalization of the proposed method, we also conducted ablation experiments and cross-training.

TABLE I

PERFORMANCE OF OUR MAAC MODEL ON DRIVE, STARE, CHASE_DB1

Dataset	AUC	AUPR	F1_ SCORE	ACC	SEN	SPE	PR
DRIVE	0.9817	0.9169	0.8289	0.9574	0.8118	0.9786	0.8469
STARE	0.9867	0.9192	0.8204	0.9667	0.7813	0.9874	0.8768
CHASE _DB1	0.9853	0.9192	0.8342	0.9656	0.8071	0.9846	0.8632

A. Vessel Segmentation

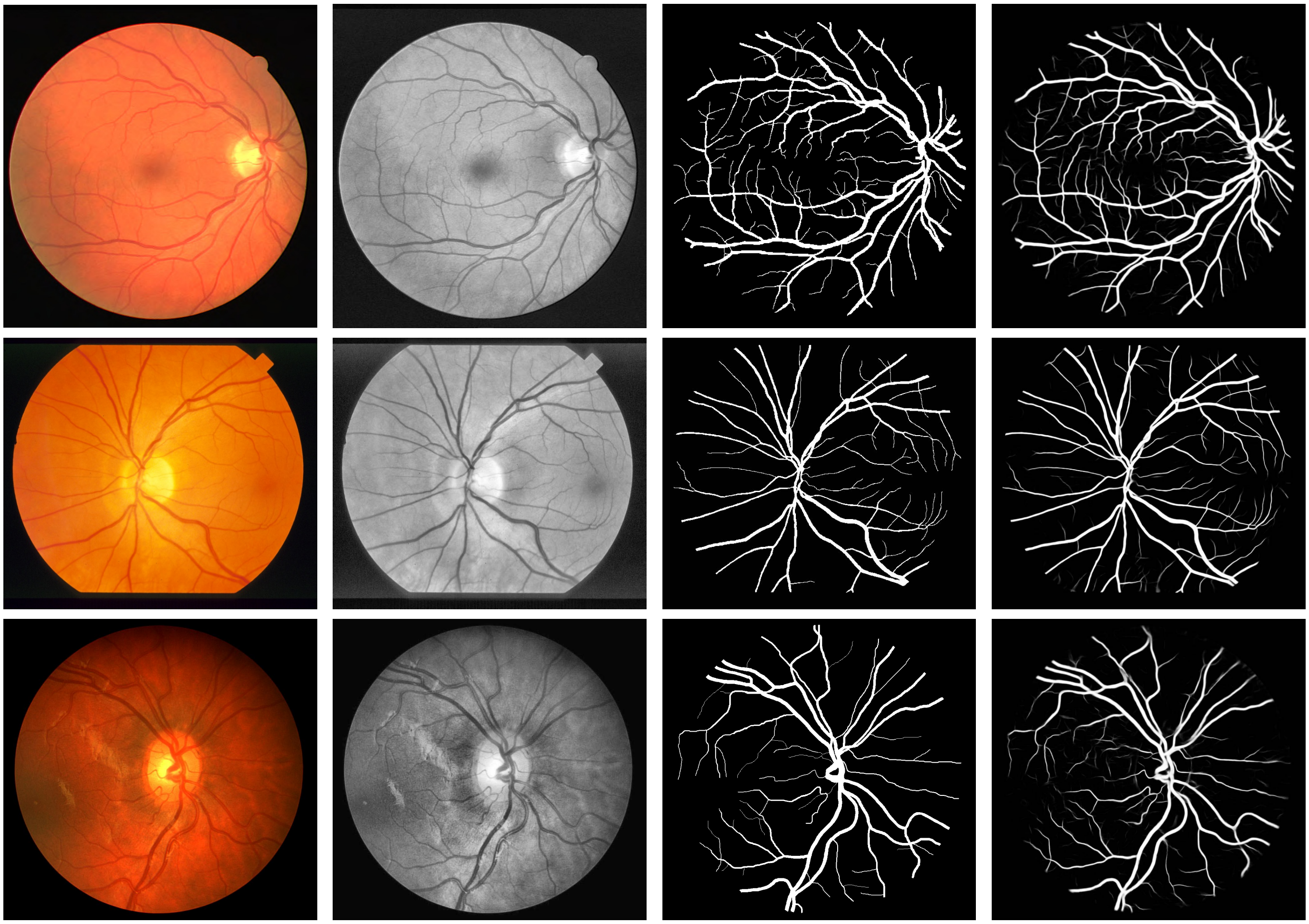
The model trained on the corresponding training set was used to predict the blood vessel segmentation results of the corresponding test set. The three binary vessel segmentation samples of DRIVE, STARE, and CHASE_db1 in their respective test sets are shown in Fig. 9. The DRIVE, STARE, and CHASE_DB1 evaluation metrics calculated from the corresponding segmentation results are shown in the Table I. For the DRIVE dataset, the AUC, SEN, SP, and ACC values are 0.9817, 0.8118, 0.9786, and 0.9574. For the STARE dataset, the AUC, SEN, SP, and ACC values are 0.9867, 0.7813, 0.9874, and 0.9667. For the CHASE_DB1 dataset, the AUC, SEN, SP, and ACC values are 0.9853, 0.8071, 0.9846, and 0.9656. The AUC of the three datasets all reached above 0.98, and the STARE dataset reached 0.9867 at the highest. The values of SP of all three datasets were higher than the values of SE. The performance of the proposed method on the DRIVE dataset is slightly inferior to that of the other two datasets.

B. Comparison to Existing Methods

We compare the results of the proposed method with those of seven existing methods. Results from existing methods are extracted from their respective published papers. The comparison results are shown in Table II. According to Table II, the AUC, ACC, and F1-score of the proposed method in DRIVE are all higher than those of the seven existing methods used for comparison. The SE ranked second among all methods of comparison. In CHASE_DB1, compared with the existing seven methods, the method proposed in this paper is the best in each evaluation metrics. The SP values are higher than those seven existing methods on STARE, but the rest of the evaluation metrics are in the medium range. In general, compared with the existing seven methods, the proposed method achieves the optimal results for DRIVE and CHASE_DB, and the performance is acceptable even though it is not optimal for STARE.

C. Ablation Studies

The proposed method has two highlights, one is a mixed attention mechanism, and the other is an asymmetric convolution module. To verify the effectiveness of the two innovation points and the optimal combination of the proposed method, we conducted an ablation experiment on DRIVE. Because the DRIVE dataset provides an official data split, it can provide a unified reference for different studies, which is more convincing. We constructed four sets of experiments, including backbone only, backbone+gcbam, backbone+asymmetric convolution and backbone+gcbam+asymmetric convolution(proposed). The experimental results are shown in Table III. The table shows that the AC_GCBAM_Unet model achieves the highest AUC values compared with the other three structures. After



(a) Original image (b) Pre-processed image (c) Ground Truth (d) Binary output of MAAC model

Fig. 9. Three exemplar vessel segmentation results from DRIVE(top), STARE(middle), CHASE DB1(bottom). From left to right:(a)original fundus images, (b)pre-processed images, (c)the ground truth, and (d)binary output of MAAC model.

TABLE II

COMPARISON WITH STATE-OF-ART METHODS ON DRIVE, STARE, AND CHASE_DB1

Y¿X

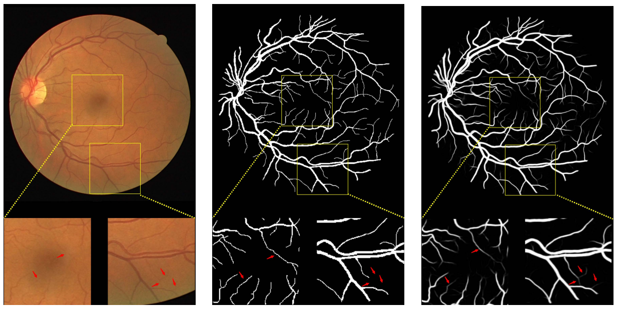
X[0]p

Dataset	Methods	Year	AUC	ACC	F1_SCORE	SEN	SPE
DRIVE	Azzopardi et al. [31]	2015	0.9614	0.9442	-	0.7655	0.9704
	Liskowski and Krawiec. [32]	2016	0.9790	0.9535	-	0.7811	0.9807
	Li et al. [8]	2016	0.9738	0.9527	-	0.7569	0.9816
	Orlando et al. [33]	2017	0.9507	-	0.7857	0.7897	0.9684
	Alom, M. Z., et al. [34]	2019	0.9784	0.9556	0.8171	0.7792	0.9813
	Yan et al. [17]	2019	0.9752	0.9542	-	0.7653	**0.9818**
	Tang,X., et al. [35]	2020	0.9769	0.9551	0.8155	**0.9682**	-
	proposed	2021	**0.9817**	**0.9574**	**0.8289**	0.8118	0.9786
STARE	Azzopardi et al. [31]	2015	0.9563	0.9497	-	0.7716	0.9701
	Liskowski and Krawiec. [32]	2016	**0.9928**	**0.9729**	-	0.8554	0.9862
	Li et al. [8]	2016	0.9879	0.9628	-	0.7726	0.9844
	Orlando et al. [33]	2017	-	-	0.7644	0.7680	0.9738
	Alom, M. Z., et al. [34]	2019	0.9914	0.9712	**0.8475**	0.8298	0.9862
	Yan et al. [17]	2019	0.9801	0.9612	-	0.7581	0.9846
	Tang,X., et al. [35]	2020	0.9883	0.9687	0.8312	**0.9745**	-
	proposed	2021	0.9867	0.9667	0.8204	0.7813	**0.9874**
CHAS_DB1	Azzopardi et al.	2015	0.9487	0.9387	-	0.7585	0.9587
	Liskowski and Krawiec. [31]	2016	-	-	-	-	-
	Li et al. [32]	2016	0.9716	0.9581	-	0.7507	0.9793
	Orlando et al. [8]	2017	0.9524	-	0.7332	0.7277	0.9712
	Alom, M. Z., et al. [34]	2019	0.9815	0.9634	0.7928	0.7756	0.9820
	Yan et al. [17]	2019	0.9781	0.9610	-	0.7633	0.9809
	Tang,X., et al. [35]	2020	-	-	-	-	-
	proposed	2021	**0.9853**	**0.9656**	**0.8342**	**0.8071**	**0.9846**

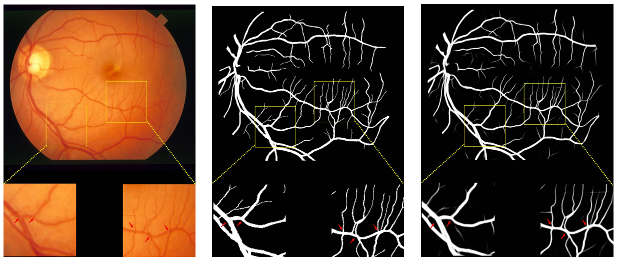
the backbone is integrated into GCBAM, the values of indicators such as AUC, AUPR, ACC, and F1_SCORE have been improved, compared with backbone. The backbone + Asymmetric Convolution not only maintains the backbone performance but also reduces the total parameters of the model. Furthermore, the total model parameters are reduced by 153152 compared with the backbone. Because AC_GCBAM_Unet is integrated with asymmetric convolution, compared with backbone + GCBAM, it can maintain good performance and reduce the model parameters. Compared with the backbone, it not only improves the segmentation performance but also has fewer model parameters.

D. Cross-Training

To verify the generalization of the proposed method, we carried out cross-training constructed as shown in [8]. Cross-training refers to training the model on one dataset and testing the model on another. Unlike [8], we do not need to retrain the model but directly apply the pre-trained model of the previous experiment. Note that when training on the STARE dataset (annotated by the first observers as the ground truth), all images in the data set are used to train the model. The cross-training results of the two datasets are shown in Table IV. The results show that the segmentation performance decreases on both DRIVE and STARE datasets when trained on STARE and DRIVE datasets, respectively. When the model trained on the STARE dataset is tested on the DRIVE dataset, AUC, ACC, and SE values dropped to 0.9742, 09501, and 0.6659, respectively. Compared with all methods, SPE achieves the best performance. Although AUC and ACC are slightly lower than Wu et al. [27], they are better than other methods. When the model trained on the DRIVE dataset is tested on the STARE dataset, AUC, ACC, and SPE values dropped to 0.9603, 09507, and 0.9721, respectively. SE achieved the best performance, but other indicators did not achieve optimal results compared with all methods.



(a) Original image and presence of macular and low contrast (b) Ground Truth and presence of macular and low contrast (c) Binary output of MAAC model and presence of macular and low contrast



(d) Original image and the presence of furcation and intersection (e) Ground Truth the and presence of furcation and intersection (f) binary output of MAAC model and the presence of furcation and intersection

Fig. 10. Examples of detail segmentation rerults from DRIVE(top) and STARE(bottom) with two enlarged rectangles showing beneath each subfigure:(a)original imag and two enlarged recteangles showing the presence of macular and low contrast, (b)ground truth, (c)binary output of MAAC model, (d)original image and two enlarged recteangles showing the presence of furcation and intersection, (e)ground truth, (c)binary output of MAAC model.

V. DISCUSSION AND CONCLUSION

In this paper, we have proposed a new convolutional neural network for retinal vessel segmentation. The network is based on the traditional encoding and decoding structure and further incorporates the attention mechanism and asymmetric convolution module. The proposed network can more effectively focus on the features that contribute to the segmentation of blood vessels and suppress the features that are ineffective in the segmentation of blood vessels, thus more helpful to the segmentation of thin blood vessels. Furthermore, the proposed network reduces a large number of trainable parameters by using asymmetric convolution.

Results of blood vessel segmentation on DRIVE, STARE, and CHASE_DB1 datasets show that the proposed method can effectively improve the performance of blood vessel segmentation, especially the AUC evaluation index, with values above 0.98 in all three data sets. Moreover, the accuracy of the three datasets is also more than 0.95, so we can conclude that the proposed method has robust performance in classifying blood vessels and background. The specificity of the three datasets is higher than the sensitivity. Our analysis is due to the limitation of the fundus image itself: the background pixel in the fundus image is much more than the blood vessel pixel. Therefore, there will be more background data in the training process, which leads to the model learning more background features, so the SP is higher than the SE in the end. By observing Fig. 10, we can see that the structure of the vascular tree of the segmentation results of the three datasets is complete, with almost no disconnection and strong continuity of the vessels. As shown in the area highlighted by the yellow box in Fig. 10(b), the small blood vessels in the low-contrast area in the DRIVE are accurately segmented, and even the small blood vessels that the observer does not annotate are also segmented. According to the ophthalmologist, these are the correct blood vessels. The other is the area around the macula (see Fig. 10(c)), it has low image intensity, which is similar to the strength of blood vessels in the green channel [8]. However, the thin blood vessels in this area can also be correctly classified, and the macula is not classified as blood vessels. Both results indicate that the proposed method has a solid ability to segment thin blood vessels; As highlighted in the yellow box in Fig. 10(f), it is the intersection point of blood vessels, and the intersection point of blood vessels is an indicator to judge whether the blood vessels of fundus oculars are healthy. The blood vessel at the intersection is easy to disconnect, but the intersection of the blood vessel is segmented clearly, and the surrounding blood vessels are not disconnected; As highlighted in the yellow box in Fig. 10(e), it is the furcation area of the blood vessel. The bifurcation point of the blood vessel is another indicator for judging whether the fundus blood vessel is healthy. The blood vessels in this area are also clearly segmented. Therefore, the proposed method is not influenced by special positions. In general, the proposed method shows advantages in the segmentation of small vessels and can also be used to segment vessels in special regions. It is a method with high accuracy and stable performance.

In the DRIVE dataset, previous research methods can generally achieve the AUC index of segmentation results above 0.97 but less than 0.98. The AUC of the proposed method is higher than 0.98, which is better than the existing methods, and the accuracy is also the

TABLE III

COMPARISON OF DIFFERENT MODELS ON DRIVE

	-	AUC	AUPR	F1_SCORE	ACC	SEN	SPE	PR	Total params
	Backbone	0.9809	0.9154	0.8252	0.9573	0.7925	**0.9813**	**0.8607**	473537
	Backbone+GCBAM	0.9814	**0.9170**	**0.8298**	**0.9578**	0.8081	0.9796	0.8526	567123
	Backbone+Asymmetric Convolution	0.9811	0.9158	0.8276	0.9573	0.8044	0.9796	0.8521	**320385**
X[0]p	**AC_GCBAM_Unet**[a] (Proposed)	**0.9817**	0.9169	0.8289	0.9574	**0.8118**	0.9786	0.8469	413971

All the models are adjusted to the highest performance. Except for the dropout rate of Backbone + GCBAM which is different from the description in the experiment part, the experimental Settings of other models and Backbone + GCBAM are consistent with the description in the experiment part.

[a] The dropout rate of the model is 0.3.

TABLE IV

RESULT OF THE CROSS-TRAINING EVALUATION

	Testing	Training	Mothods	AUC	ACC	SPE	SE
	DRIVE	STARE	Fraz et al. [28](2012a)	0.9697	0.9456	0.9792	0.7242
			Li et al. [8](2016)	0.9677	0.9486	0.9810	0.7273
			Yan et al. [17](2018b)	0.9568	0.9444	0.9802	0.7014
			Yan et al. [36](2018a)	0.9599	0.9494	0.9815	**0.7292**
			Wu, Y., et al. [27](2020)	**0.9761**	**0.9538**	0.9881	0.7187
X[0]p			Proposed	0.9742	0.9501	**0.9915**	0.6659
	STARE	DRIVE	Fraz et al. [28](2012a)	0.9660	0.9495	0.9770	0.7010
			Li et al. [8](2016)	0.9671	0.9545	0.9828	0.7027
			Yan et al. [17](2018b)	0.7027	**0.9580**	**0.9840**	0.7319
			Yan et al. [36](2018a)	**0.9708**	0.9569	**0.9840**	0.7211
			Wu, Y., et al. [27](2020)	0.9635	0.9540	0.9785	0.7378
			Proposed	0.9603	0.9507	0.9721	**0.7616**

highest among all methods. The sensitivity is 0.8118, which is lower than Tang et al. [35], ranking second. Combining the above three evaluation metrics shows that the proposed method is superior to all previous methods and can detect more blood vessel pixels accurately. In the Chase _DB1 dataset, all indexes of the proposed method are optimal compared with the previous methods, which further proves that the proposed method is superior to the previous methods and has generalization. In the STARE dataset, the performance of the proposed method is not optimal but compared with Azzopardi et al. [31], AUC and ACC have increased by 0.04 and 0.017, respectively, indicating that the segmentation capabilities of the proposed method have indeed improved. However, the value of SP is the highest among all methods, indicating that on the STARE dataset, the proposed method has a stronger ability to segment the background. In addition to the limited fundus image itself, another reason is that we use the first expert annotation as ground truth, while the first expert focused more on thick blood vessels. Hence, many thin blood vessels were not annotated. There are three reasons why the results on the STARE dataset are not optimal. One is that the diseased cases in the STARE dataset are more severe than DRIVE. The second is that the images of the STARE dataset are of low quality; and the third is that we directly use unified experimental parameters, such as thresholds are 0.5, while the data characteristics of the STARE data set indicated that a smaller threshold would have a better effect, and we did not adopt fine-tuning as in the previous method. In the future, we will use datasets with more pathological images to train our model and further explore how to counter the interference of pathological images to achieve a segmentation model that is less affected by pathology.

In the ablation experiment, the AUC, ACC, and SEN of the backbone+GCBAM model are increased to 0.9814, 0.9578, and 0.8081, respectively, compared with the backbone, which shows that the GCBAM module can effectively help the model to classify the blood vessel pixels and improve the recognition accuracy. The backbone+Asymmetric Convolution model has improved other indicators except for SPE and PR compared with the backbone model. Although the improvement is not very big, the parameters of the model are reduced to 320385. Compared with the backbone+GCBAM model, the various evaluation indicators of the backbone+Asymmetric Convolution model are slightly reduced, but its model parameters are reduced by 246738 compared with the backbone+GCBAM, indicating that the model parameters are few, but it has achieved relatively impressive results. Compared with the backbone+GCBAM model, the AUC and SEN of the MAAC encoder-decoder are increased from 0.9814 and 0.8081 to 0.9817 and 0.8118. And the training parameters are reduced from 567123 to 413971. In summary, it shows that the combination of the proposed methods can detect more blood vessels pixels with higher classification accuracy, and the model has fewer parameters, which can save computing resources.

In cross-training, the AUC, ACC, and SPE of the proposed method outperform all the comparison methods for the DRIVE dataset. The AUC of the Yan et al.'s [36] method decreased from 0.9752 to 0.9568, a total decrease of 0.0184, while ours decreased from 0.9817 to 0.9742, a decrease of 0.0075 for the proposed method. The above shows that the proposed method is more robust than previous methods. The robust is required for applying the method to practice because in practice, the trained model will not be retrained, and it has to face many cases that have never been encountered. Since the first observer in the STARE dataset mainly annotates thick blood vessels, when the model trained on the STARE dataset is applied to the DRIVE dataset, the model's ability to detect thin blood vessels will decrease, so the SE on DRIVE is lower. The proposed method's AUC, ACC, and SPE were slightly lower than those of previous methods in the STARE dataset. Because when DRIVE is used as a training set, there is a lack of pathological feature data at the same level as STARE, which leads to a slightly more decrease in AUC when tested on the STARE dataset. However, the SE of the proposed method is the highest among all methods because the manual annotation of the DRIVE dataset is very detailed, and most blood vessels are marked, including thick and thin blood vessels, so when the model trained in DRIVE is applied to the STARE dataset, there will be strong blood vessel detection capabilities.

The computation time only needs 30 iterations to train the model on the DRIVE, STARE and CHASE_DB1 data sets. The model is trained and tested on NVIDIA Tesla v100 16GB GPU. The time

TABLE V
TIME COST OF PERFORMING MODEL TRAINING AND INFERENCING TEST IMAGE BY OUR MAAC MODEL.

-	DRIVE	STARE	CHASE_DB1	
Training	310.38m	197.2m	254.92m	
Testing	1.32m	0.03m	0.82m	

consumed by each dataset is shown in Table V. It can be seen that the training of the model is very time-consuming. Each model needs to be trained for more than 1 hour, but the inference speed is fast. For example, it only takes 1.32 minutes to infer 20 images in the DRIVE test set, which is much faster than manual annotation. In practice, we can train the model offline and use the trained model in clinical practice.

In conclusion, we improved a hybrid attention module and proposed a new retinal vessel segmentation model on this basis. This model can segment more blood vessels, achieve higher accuracy than previous methods, and have the potential for medical applications. And in theory, our approach is not limited to retinal vessel segmentation but can also be extended to other types of image segmentation. It means the approach has generalization.

ACKNOWLEDGMENT

The authors would like to thank J. J. Staal, A. Hoover, M.M. Fraz, and their colleagues created datasets and made them publicly available. The authors would also like to thank Professor Li from Shenzhen Institute for Advanced Study, UESTC, for providing academic guidance and equipment support. Finally, the authors sincerely thank Doctor Xie from Heping County People's Hospital of Heyuan for providing guidance on retinal images.

REFERENCES

[1] J. B. Saaddine, K. V. Narayan, and F. Vinicor, "Vision loss: a public health problem?" *Ophthalmology*, vol. 110, no. 2, pp. 253–254, Feb. 2003.

[2] C. L. Schwartz and R. J. McManus, "What is the evidence base for diagnosing hypertension and for subsequent blood pressure treatment targets in the prevention of cardiovascular disease?" *BMC Med.*, vol. 13, no. 1, pp. 1–9, Oct. 2015, 10.1186/s12916-015-0502-5.

[3] M. K. Ikram *et al.,* "Retinal vessel diameters and risk of impaired fasting glucose or diabetes: the rotterdam study," *Diabetes*, vol. 55, no. 2, pp. 506–510, Feb. 2006, 10.2337/diabetes.55.02.06.db05-0546

[4] D. Fu, Y. Liu, and Z. Huang, "A review of retinal vessel segmentation and artery/vein classification," in *Proc. CISC,* Mudanjiang, CN, 2017, pp. 727–737.

[5] P. Zhu *et al.,* "The relationship of retinal vessel diameters and fractal dimensions with blood pressure and cardiovascular risk factors," *PloS one*, vol. 9, no. 9, p. e106551, Sep. 2014, 10.1371/journal.pone.0106551.

[6] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, Dec. 2010, 10.1109/RBME.2010.2084567.

[7] H. Jelinek and M. J. Cree, "Automated image detection of retinal pathology," Boca Raton, FL, USA: CRC Press, 2009.

[8] Q. Li *et al.,* "A cross-modality learning approach for vessel segmentation in retinal images," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 109–118, July. 2015, 10.1109/TMI.2015.2457891.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI,* Munich, GER, 2015, pp. 234–241.

[10] H. Fu *et al.,* "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in *Proc. MICCAI,* Athens, GRE, 2016, pp. 132–139.

[11] H. Fu *et al.,* "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *Proc. ISBI,* Prague, CZ, 2016, pp. 698–701.

[12] J. Dai *et al.,* "Deformable convolutional networks," in *Proc. ICCV,* Venice, IT, 2017, pp. 764–773.

[13] M. M. U. Islam and M. Indiramma, "Retinal vessel segmentation using deep learning–a study," in *Proc. ICOSEC,* Thottiam Taluk, IND, 2020, pp. 176–182.

[14] J. Guo, S. Ren, Y. Shi, and H. Wang, "Automatic retinal blood vessel segmentation based on multi-level convolutional neural network," in *Proc. CISP-BMEI,* Beijing, CN, 2018, pp. 1–5.

[15] K. Hu *et al.,* "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function," *Neurocomputing*, vol. 309, pp. 179–191, Oct. 2018, 10.1016/j.neucom.2018.05.011

[16] R. Xu *et al.,* "Retinal vessel segmentation via multiscaled deep-guidance," in *Proc. PCM,* Hefei, CN, 2018, pp. 158–168.

[17] Z. Yan, X. Yang, and K.-T. Cheng, "A three-stage deep learning model for accurate retinal vessel segmentation," *IEEE J Biomed Health Inform*, vol. 23, no. 4, pp. 1427–1436, July. 2018, 10.1109/JBHI.2018.2872813.

[18] V. Cherukuri *et al.,* "Deep retinal image segmentation with regularization under geometric priors," *IEEE Trans Image Process*, vol. 29, pp. 2552–2567, Oct. 2019, 10.1109/TIP.2019.2946078.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML,* Lille, FRA, 2015, pp. 448–456.

[20] G. F. Montúfar, "Universal approximation depth and errors of narrow belief networks with discrete units," *Neural Comput*, vol. 26, no. 7, pp. 1386–1407, July. 2014, 10.1162/NECO_a_00601.

[21] S. Woo *et al.,* "Cbam: Convolutional block attention module," in *Proc. ECCV,* Munich, GER, 2018, pp. 3–19.

[22] O. Oktay *et al.,* "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[23] C. Szegedy *et al.,* "Rethinking the inception architecture for computer vision," in *Proc. CVPR,* Las Vegas, NV, USA, 2016, pp. 2818–2826.

[24] X. Ding *et al.,* "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proc. ICCV,* Seoul, KR, 2019, pp. 1911–1920.

[25] J. Staal *et al.,* "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004, 10.1109/TMI.2004.825627.

[26] 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, 10.1109/42.845178.

[27] Y. Wu *et al.,* "Nfn+: A novel network followed network for retinal vessel segmentation," *Neural Netw.*, vol. 126, pp. 153–162, June. 2020, 10.1016/j.neunet.2020.02.018.

[28] 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, 10.1109/TBME.2012.2205687.

[29] 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, May. 2009, 10.1167/iovs.08-3018.

[30] J. V. Soares *et al.,* "Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification," *IEEE Trans. Med. Imag.*, vol. 25, no. 9, pp. 1214–1222, Aug. 2006, 10.1109/TMI.2006.879967

[31] G. Azzopardi *et al.,* "Trainable cosfire filters for vessel delineation with application to retinal images," *Med Image Anal*, vol. 19, no. 1, pp. 46–57, Jan. 2015, 10.1016/j.media.2014.08.002.

[32] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Trans. Med. Imag.,* vol. 35, no. 11, pp. 2369–2380, Mar. 2016, 10.1109/TMI.2016.2546227.

[33] 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. 2016, 10.1109/TBME.2016.2535311.

[34] M. Z. Alom *et al.,* "Recurrent residual u-net for medical image segmentation," *JMI*, vol. 6, no. 1, p. 014006, Mar. 2019, 10.1117/1.JMI.6.1.014006.

[35] X. Tang *et al.,* "Multi-scale channel importance sorting and spatial attention mechanism for retinal vessels segmentation," *Appl. Soft Comput.*, vol. 93, p. 106353, Aug. 2020, 10.1016/j.asoc.2020.106353.

[36] 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, Sept. 2018, 10.1109/TBME.2018.2828137.