SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation

Changlu Guo[1,*], Márton Szemenyei[1], Yugen Yi[2,*], Wenle Wang[2], Buer Chen[1], Changqi Fan[1]
[1]Budapest University of Technology and Economics, Budapest, Hungary
Email: clguo.ai@gmail.com
[2]Jiangxi Normal University, Nanchang, China
Email: yiyg510@jxnu.edu.cn

***Abstract*— The precise segmentation of retinal blood vessel is of great significance for early diagnosis of eye-related diseases such as diabetes and hypertension. In this work, we propose a lightweight network named Spatial Attention U-Net (SA-UNet) that does not require thousands of annotated training samples and can be utilized in a data augmentation manner to use the available annotated samples more efficiently. SA-UNet introduces a spatial attention module which infers the attention map along the spatial dimension, and then multiply the attention map by the input feature map for adaptive feature refinement. In addition, the proposed network employees a kind of structured dropout convolutional block instead of the original convolutional block of U-Net to prevent the network from overfitting. We evaluate SA-UNet based on two benchmark retinal datasets: the Vascular Extraction (DRIVE) dataset and the Child Heart and Health Study (CHASE_DB1) dataset. The results show that our proposed SA-UNet achieves the state-of-the-art retinal vessel segmentation accuracy on both datasets.**

Keywords—Segmentation; retinal blood vessel; SA-UNet; U-Net; spatial attention

I. INTRODUCTION

Many diseases can be fairly diagnosed and tracked by observing the fundus vascular system, because these diseases such as diabetes and hypertension can cause morphological changes in the blood vessels of the retina. Systemic microvascular and small vessel disease are common pathological changes caused by diabetes, especially the fundus retinal vascular disease is the most vulnerable. Diabetic retinopathy (DR) is caused by diabetes [1]. If swelling of the blood vessels in the retina of a diabetic patient is observed, special attention is required. Patients with long-term hypertension may observe blood vessel curvature due to increased arterial blood pressure or vascular stenosis, which is called hypertensive retinopathy (HR) [2]. Retinal vessel segmentation is a key step in the quantitative analysis of fundus images. By segmenting the retinal blood vessels, we can obtain the relevant morphological information of the retinal blood vessel tree (such as the curvature, length, and width of the blood vessels) [3]. Moreover, the vascular tree of retinal vessels has unique characteristics and can also be applied to biometric recognition [4] [5]. Therefore, accurate segmentation of retinal blood vessels is of great significance.

However, the retinal blood vessels have a large number of small and fragile blood vessels, and the blood vessels are closely connected, so the retinal blood vessel tree structure is very complicated. In addition, the difference between the blood vessel area and the background is not obvious, and the fundus image is also susceptible to uneven lighting and noise. The above reasons cause retinal blood vessel segmentation still to be a challenging task.

In the past few decades, a large number of retinal blood vessel segmentation methods have been proposed, mainly divided into manual segmentation and computer algorithm automatic segmentation. The former is time-consuming and labor-intensive and requires extremely high professional skills of practitioners. The latter can reduce the burden of manual segmentation, so the research on the automatic segmentation algorithm is of great significance. With the development of deep learning in recent years, it has gradually become the mainstream technology of retinal segmentation.

In the field of medical image segmentation, U-Net [6] is a common and well-known backbone network. Basically, U-Net consists of a typical downsampling encoder and upsampling decoder structure and a "skip connection" between them. It combines local and global context information through the encoding and decoding process. Due to the excellent performance of U-Net, many recent methods for retinal blood vessel segmentation are based on U-Net. Wang et al. [7] reported the Dual Encoding U-Net (DEU-Net) that remarkably enhances network's capability of segmenting retinal vessels in an end-to-end and pixel-to-pixel way. Wu et al. [8] proposed Vessel-Net, which first time uses a strategy that combines the advantages of the initial method and the residual method to perform retinal vessel segmentation. Although these U-Net variants perform well, they inevitably make the network more complex and less interpretable.

In order to address these problems, we introduce spatial attention in U-Net and propose a lightweight network model, which we named Spatial Attention U-Net (SA-UNet). Inspired by SD-Unet [9], using DropBlock [10] can effectively prevent overfitting of the network, so even small sample datasets such as retinal fundus images also can be well trained. In addition, bacth normalization (BN) can improve the convergence speed

* Corresponding authors
This work is supported by the National Natural Science Foundation of China under Grants 61602221 and 61672150.

Manuscript 501 submitted to 2020 25th International Conference on Pattern Recognition. Received March 16, 2020.

of the network [11]. Therefore, SA-UNet first employees a kind of structure dropout convolutional block integrating DropBlock and batch normalization (BN) to replace the original U-Net convolutional block. More importantly, the difference between vascular and non-vascular features in the retinal fundus image is not obvious, especially the small and marginal vascular areas. With the introduction of a small amount of additional parameters, spatial attention can enhance important features (such as vascular features) and suppress unimportant features, thereby improving the network's representation ability. We evaluate SA-UNet on two public retinal fundus image datasets: DRIVE and CHASE_DB1. We first evaluate the newly introduced part of the network through ablation experiments. The experimental results show that the structured dropout convolutional block and spatial attention we introduced are effective, and compared with the original U-Net, our proposed SA-UNet is very lightweight. Finally, compared with other existing state-of-the-art methods for retinal vascular segmentation, our proposed SA-UNet achieves state-of-the-art performance.

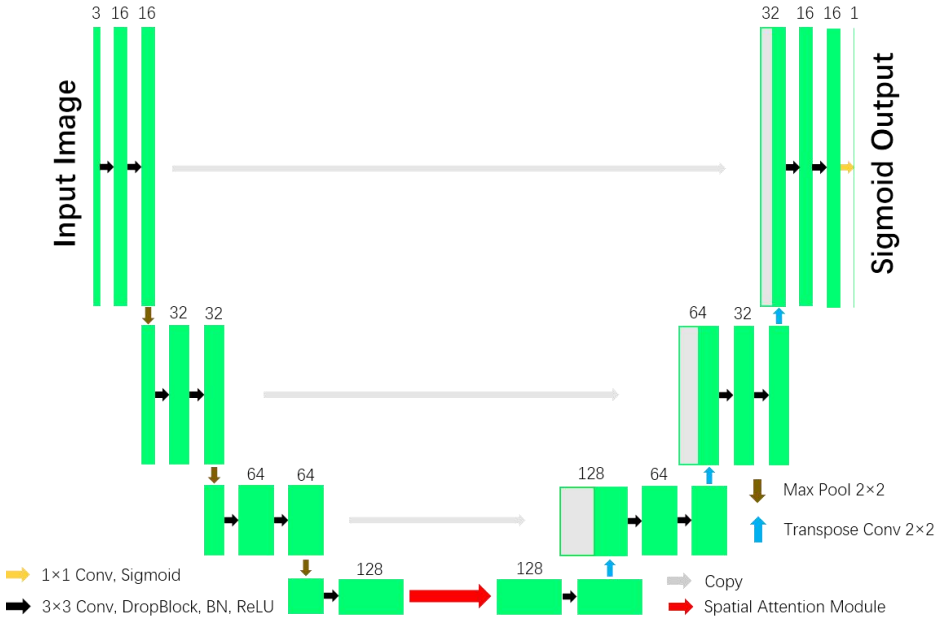

Fig. 1. Diagram of the proposed SA-UNet.

II. Methodology

A. Network Architecture

Figure 1 shows our proposed SA-UNet with a U-shaped encoder (left side)-decoder (right side) structure. Every step of the encoder includes a structured dropout convolutional block and a 2×2 max pooling operation. The convolutional layer of each convolutional block is followed by a DropBlock, a layer of batch normalization (BN) and a rectified linear unit (ReLU), and then the max pooling operation is utilized for down-sampling with a stride size of 2. In each down-sampling step, we double the number of feature channels. Each step in the decoder includes a 2×2 transposition convolution operation for up-sampling and halving the number of feature channels, a concatenation with the corresponding feature map from the encoder, and then followed by one structured dropout convolutional block. Between the encoder and the decoder, we add the spatial attention module. At the final layer, we use 1×1 convolution and Sigmoid activation function to get our output segmentation map.

B. Structured Dropout Convolutional Block

Although we do data augmentation for the original datasets, serious overfitting is still observed during U-Net training, as shown in the upper row of Fig. 2. Dropout is a very common and successful method of preventing network from overfitting. Its main feature is that it can randomly discard some features during training. Although this feature makes dropout effective at the fully connected layer, the effect at the convolutional layer is not significant because of the spatial correlation of the active cells. Therefor, in general, dropout was primarily employed at the fully connected layers of the convolutional nenetworks [12] [13]. In Fig. 2., we also show the training situation of U-Net with dropout. It is easy to observe that the overfitting problem still exists. So we argue that U-Net as a fully convolutional network requires a more appropriate regularization method to prevent overfitting.

DropBlock, a structured form of dropout, can effectively prevent over-fitting problems in convolutional networks [10]. Its primary difference from dropout is that it discards contiguous areas from a feature map of a layer instead of dropping independent random units. Based on this, we construct a structured dropout convolutional block, that is, each convolutional layer is followed by a DropBlock, a layer of batch normalization (BN) and a ReLU activation unit, as shown in the right side of Fig. 3. Different from convolutional block of SD-Unet, as shown in the middle of Fig. 3., the structured dropout convolutional block introduces batch normalzation (BN) to accelerate network convergence. We

Manuscript 501 submitted to 2020 25th International Conference
on Pattern Recognition. Received March 16, 2020.

employee this structured dropout convolutional block instead of original convolutional block of U-Net to build a U-shaped network as our "Backbone". Compared to the 23-layer convolutional layer of the original U-Net, our Backbone has only 18 convolutional layers, and as shown in Fig. 2., the overfitting problem is alleviated.

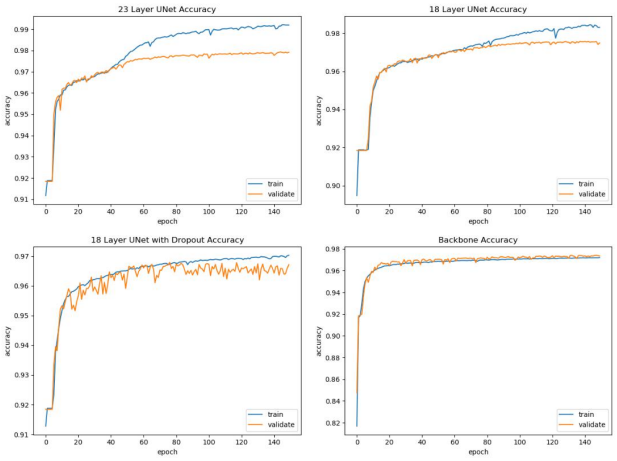

Fig. 2. Comparison of different models training 150 epochs on DRIVE.

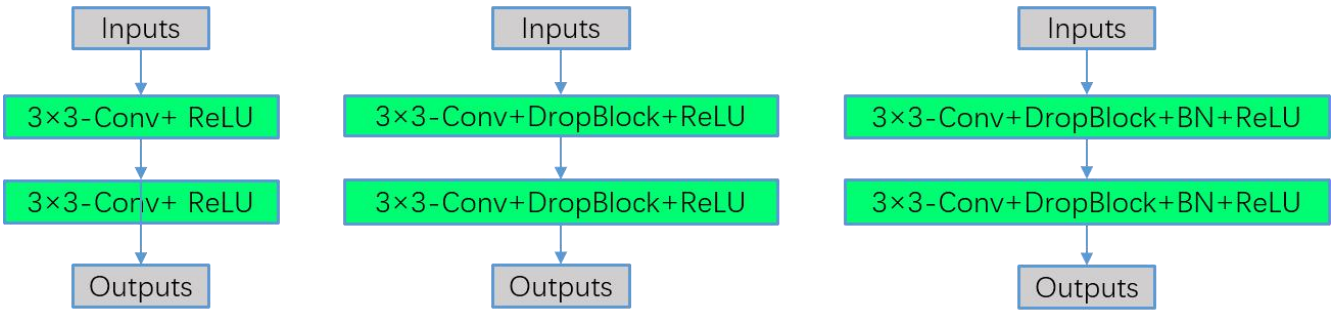

Fig. 3. Original U-Net block (left),SD-Unet block (middle), Structured dropout convolutional block(right)

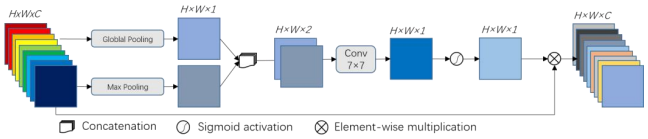

Fig. 4. Diagram of the Spatial Attention

III. EXPERIMENTS

A. Datasets

We evaluate our proposed SA-UNet on two public retinal fundus image datasets: DRIVE and CHASE DB1. The specific information of the two datasets is given in Table I. It should be noted that the original size of the two datasets is not suitable for our network, so we adjusted its size by zero padding around it, but the size will be crop to the initial size during evaluation. To augment the data, we adopt four data augmentation methods shown in the last column of Table I for both datasets, each of which generated three new images from an original image, that is, we augment the two original datasets from the original 20 training images to 256 images.

C. Spatial Attention Module (SAM)

Spatial Attention Module(SAM) was introduced as a part of the c*onvolutional block attention module* for classification and detection [14]. SA uses the spatial relationship between features to produce a spatial attention map. To calculate spatial attention, SA first applies max-pooling and average-pooling operations along the channel axis and concatenate them to produce an efficient feature descriptor. Formally, input feature $F \in R^{H\times W\times C}$ through the channel-wise max-poling and average-pooling generate $F^s_{mp} \in R^{H\times W\times 1}$ and $F^s_{ap} \in R^{H\times W\times 1}$, respectively. Then a convolutional layer followed by the sigmoid activation function on the concatenated feature descriptor is used to generate a spatial attention map $M^s(F) \in R^{H\times W\times 1}$. In short, the output feature $F^s \in R^{H\times W\times C}$ of spatial attention module is calculated as:

$$\begin{aligned} F^s &= F \cdot M^s(F) \\ &= F \cdot \sigma(f^{7\times 7}([MaxPool\ (F); AvgPool\ (F)])) \\ &= F \cdot \sigma(f^{7\times 7}([F^s_{mp}; F^s_{ap}])) \end{aligned} \tag{4}$$

Where $f^{7\times 7}(\cdot)$ denotes a convolution operation with a kernel size of 7 and $\sigma(\cdot)$ represents the sigmoid function.

B. Evaluation Metrics

In order to evaluate our model, we compare the segmentation results with the corresponding ground truth and divide the results of each pixel comparison into true positive (TP), false positive (FP), negative (FN), and true negative (TN) . Then, the sensitivity (SE), specificity (SP), F1-score (F1), and accuracy (ACC) are used to evaluate the performance of the model. In retinal vessel segmentation, only 9%-14% of the pixels belong to the blood vessel, while other pixels are considered background pixels. The Matthews Correlation Coefficient (MCC) is suitable for performance measurement of binary classifications for two categories with different sizes. Therefore, the MCC value can help find the optimal setting for the vessel segmentation algorithm. MCC is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{7}$$

The area under the ROC curve (AUC) can be used to measure the performance of the segmentation. If the AUC value is 1, it means perfect segmentation.

Manuscript 501 submitted to 2020 25th International Conference on Pattern Recognition. Received March 16, 2020.

TABLE I. THE SPECIFIC INFORMATION OF **DRIVE** AND **CHASE_DB1** DATASETS

Datasets	**DRIVE**	**CHASE_DB1**
Obtain from	Dutch Diabetic Retinopathy Screening Program	Child Heart and Health Study
Total number	40	28
Train/Test number	20/20	20/8
Resolution (pixel)	584x565	999x960
Resize (pixel)	592x592	1008x1008
Augmentation methods	(1) Random rotation; (2) adding Gaussian noise; (3) color jittering; (4) orizontal, vertical and diagonal flips.	

C. Implementation Details

In order to monitor whether our network is overfitting, we randomly select 26 and 13 images in the DRIVE and CHASE DB1 augmented datasets as the validation set. As mentioned earlier, Fig. 1 shows the case of training 150 epochs on the DRIVE dataset. We train SA-UNet from scratch using the augmented training set. For both datasets, we employ Adam as our optimizer and binary cross entropy as our loss function, and in order not to make the parameter amount too large, we set the number of channels after the first convolutional layer to 16. The size of the discard blocks of DropBlock is set to 7.

For DRIVE, the batch size of the training is set to 4 and the maximum training epochs is 200. The learning rate of the first 150 epochs is 0.001, and the last 50 epochs is 0.0001. In order to reach the best performance, we set the dropout rates of DropBlock to 0.18.

For CHASE DB1, the batch size of the training is set to 2 and the maximum training epochs is 150 . The learning rate of the first 100 epochs is 0.001, and the last 50 epochs is 0.0001. In order to reach the best performance, we set the dropout rates of DropBlock to 0.13.

In the ablation experiments, SD-Unet and Backbone use the same configuration as SA-UNet.

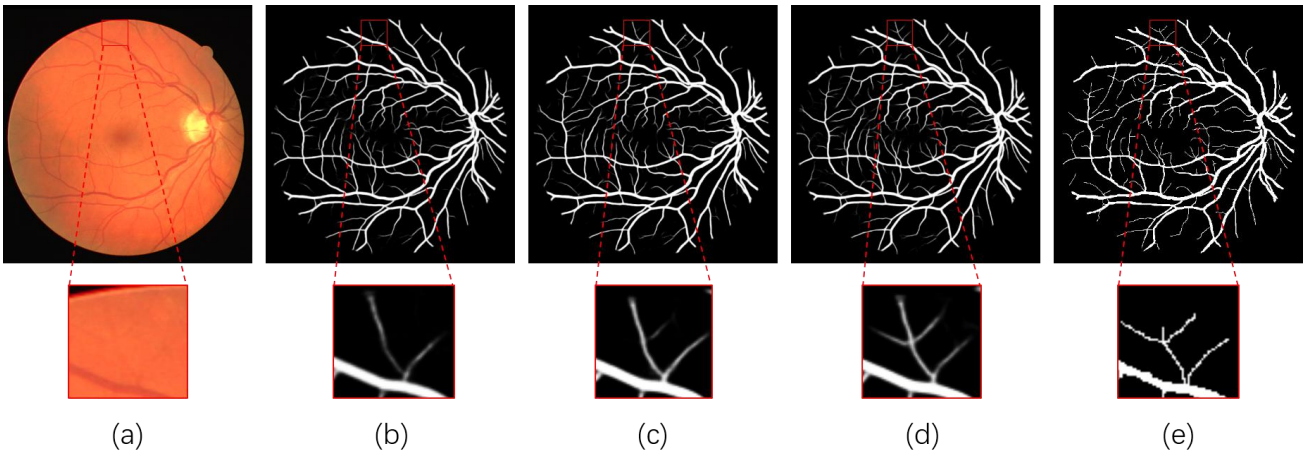

Fig. 5. (a) A test image from DRIVE dataset; (b) Segmentation result by SD-Unet; (c) Segmentation result by Backbone; (d) Segmentation result by SA-UNet; (e) Ground truth segmentation.

IV. RESULTS

A. Ablation Experiments

In order to prove that each component of the proposed SA-UNet can improve the performance of retinal vascular segmentation, ablation experiments were performed on DRIVE and CHASE_DB1 respectively. Tables II and III show the segmentation performance of U-Net, SD-Unet, Backbone (i.e. SD-Unet + BN), and SA-UNet (i.e. Backbone + SA) from top to bottom, respectively. In addition, Table IV shows the parameter quantities of different models. The results show that, in DRIVE and CHASE_DB1, (1) Backbone has better performance compared with the original SD-Unet, although the parameter amount is increased slightly, which shows that adding the batch normalization (BN) can improve the network performance to a certain extent. Meanwhile, it demonstrates that our strategy of adopting the newly constructed structured dropout convolutional block to build Backbone is effective. (2) With only 98 parameters added, the AUC, F1, and MCC of SA-UNet are 0.05% / 0.08%, 0.29% / 0.31%, and 0.34% / 0.36% higher than Backbone, respectively, which proves the strategy of the introduction of spatial attention is effective. (3) Compared with the original U-Net with 23 convolutional layers, our SA-UNet has a much smaller amount of parameters, so for the task of retinal blood vessel segmentation, SA-UNet is a lightweight and effective network .

In Fig. 4., we show a test example on the DRIVE dataset, including the segmentation results obtained by SD-Unet, Backbone and the proposed SA-UNet, and the corresponding ground truth. SD-Unet ignores some edge and small vascular

Manuscript 501 submitted to 2020 25th International Conference
on Pattern Recognition. Received March 16, 2020.

structures. Backbone produces more accurate small vessel segmentation than the original SD-Unet, which proves the effectiveness of Backbone constructed using structured dropout blocks with batch normalization (BN). Compared with Backbone, SA-UNet proposed in this paper can produce more accurate segmentation results for small border blood vessels and retain more retinal blood vessel spatial structure, which proves that the spatial attention mechanism can highlight blood vessels and reduce the influence of background. In order to better observe the test results, we show more segmentation examples of SD-Unet, Backbone, and SA-UNet on DRIVE and CHASE_DB1 in Fig. 6. and Fig. 7., respectively.

TABLE II. ABLATION STUDIES ON **DRIVE** DATASET. (*THE RESULTS ARE OBTAINED FROM [19])

Methods	SE	SP	ACC	AUC	F1	MCC
U-Net [6]*	0.7897	0.9854	0.9681	0.9836	-	-
SD-Unet [9]	0.7806	**0.9875**	0.9694	0.9846	0.8172	0.8016
Backbone	**0.8312**	0.9816	0.9685	0.9855	0.8219	0.8047
SA-UNet	0.8231	0.9834	**0.9694**	**0.9860**	**0.8248**	**0.8081**

TABLE III. ABLATION STUDIES ON **CHASE_DB1** DATASET. (*THE RESULTS ARE OBTAINED FROM [19])

Methods	SE	SP	ACC	AUC	F1	MCC
U-Net [6]*	0.7715	**0.9858**	0.9723	0.9837	-	-
SD-Unet [9]	0.8297	0.9854	**0.9756**	0.9897	0.8109	0.7981
Backbone	0.8422	0.9844	0.9755	0.9897	0.8123	0.7997
SA-UNet	**0.8573**	0.9835	0.9755	**0.9905**	**0.8153**	**0.8033**

B. Comparisons with state-of-the-art methods

Finally, we compare the performance of SA-UNet with other state-of-the-art methods currently applied in retinal vessel segmentation task. In Tables V, we summarize the release year of different methods and the performance on DRIVE and CHASE_DB1 datasets. From the results in the table, it can be concluded that SA-UNet has achieved the best performance on both DRIVE and CHASE_DB1. It achieves the highest sensitivity of 0.8225/0.8573, the highest accuracy of 0.9694/0.9755, the highest AUC of 0.9694/0.9905, and specificity are comparable. The above results show that our proposed SA-UNet achieves state-of-the-art performance in the retinal vessel segmentation challenge.

TABLE IV. AMOUNT OF PARAMETERS ON DIFFERENT MODELS.

Parameters	23 Layer U-Net	18 Layer U-Net	SD-Unet	Backbone	SA-UNet
Total	2,158,705	535,793	535,793	538,609	538,707
Trainable	2,158,705	535,793	535,793	537,201	537,299
Non-trainable	0	0	0	1,408	1,408

TABLE V. RESULTS OF SA-UNET AND OTHER METHODS ON **DRIVE** AND **CHASE_DB1** DATASETS.

Datasets	DRIVE					CHASE_DB1			
Metrics	Year	SE	SP	ACC	AUC	SE	SP	ACC	AUC
Liskowski et .al. [15]	2016	0.7811	0.9807	0.9535	0.9790	0.7816	0.9836	0.9628	0.9823
Orlando et. al. [16]	2017	0.7897	0.9684	0.9454	0.9507	0.7277	0.9712	0.9458	0.9524
Yan et. al. [17]	2018	0.7653	0.9818	0.9542	0.9752	0.7633	0.9809	0.9610	0.9781
MS-NFN [18]	2018	0.7844	0.9819	0.9567	0.9807	0.7538	0.9847	0.9637	0.9825
DEU-Net [7]	2019	0.7940	0.9816	0.9567	0.9772	0.8074	0.9821	0.9661	0.9812
Vessel-Net [8]	2019	0.8038	0.9802	0.9578	0.9821	0.8132	0.9814	0.9661	0.9860
AG-Net [19]	2019	0.8100	**0.9848**	0.9692	0.9856	0.8186	**0.9848**	0.9743	0.9863
SA-UNet	**2020**	**0.8225**	0.9834	**0.9694**	**0.9860**	**0.8573**	0.9835	**0.9755**	**0.9905**

V. CONCLUSION

For retinal fundus image datasets, most of them are typical small sample datasets without a large number of annotated training samples. We first do data augmentation in an ambitious way, and we use lightweight U-Net, but overfitting is still observed. Inspired by the successful application of DropBlock and Bactch Normalization on convolutional neural networks, we replace the convolutional block of U-Net with a structured dropout convolutional block that integrates DropBlock and Batch Normalization as our Backbone. In addition, In the retinal fundus image, the difference between the blood vessel area and the background is not obvious, especially the edges and small blood vessels. Therefore, we add a spatial attention module between the encoder and decoder of Backbone and propose Spatial Attention U-Net (SA-UNet). The spatial attention can help the network to focus on important features and suppress unnecessary ones to achieve the goal of improving the network's representation capability. We evaluate SA-UNet on two publicly available retinal fundus image data including DRIVE and CHASE DB1. The experimental results demonstrate that our strategy of using structured dropout of convolutional blocks and the introduction of spatial attention are effective, and by comparing with other state-of-the-art methods for retinal vessel segmentation, our lightweight SA-UNet achieves state-of-the-art performance. Because the vascular structure characteristics of the retinal image are

Manuscript 501 submitted to 2020 25th International Conference
on Pattern Recognition. Received March 16, 2020.

similar, it is believed that SA-UNet is a general one and can be applied to other retinal vessel segmentation tasks.

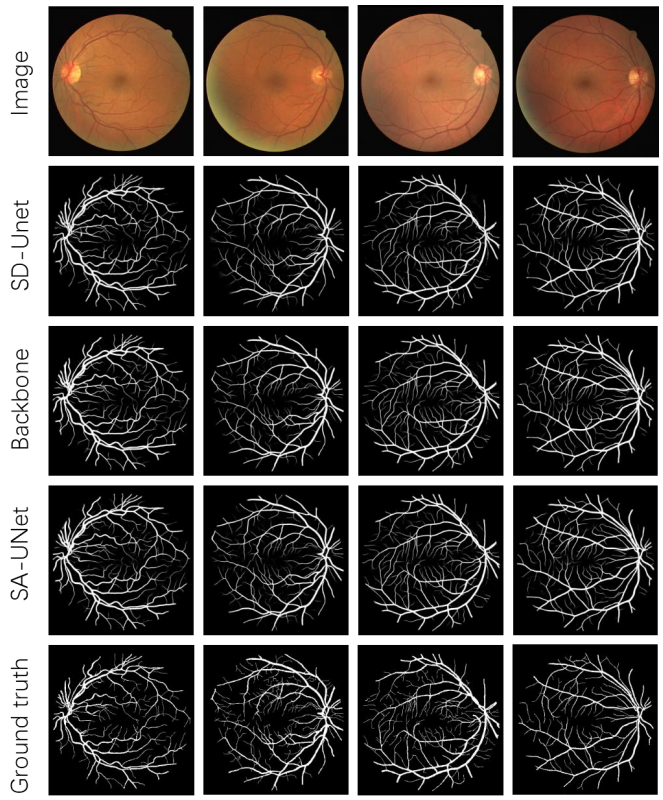

Fig. 6. Segmentation results on **DRIVE**

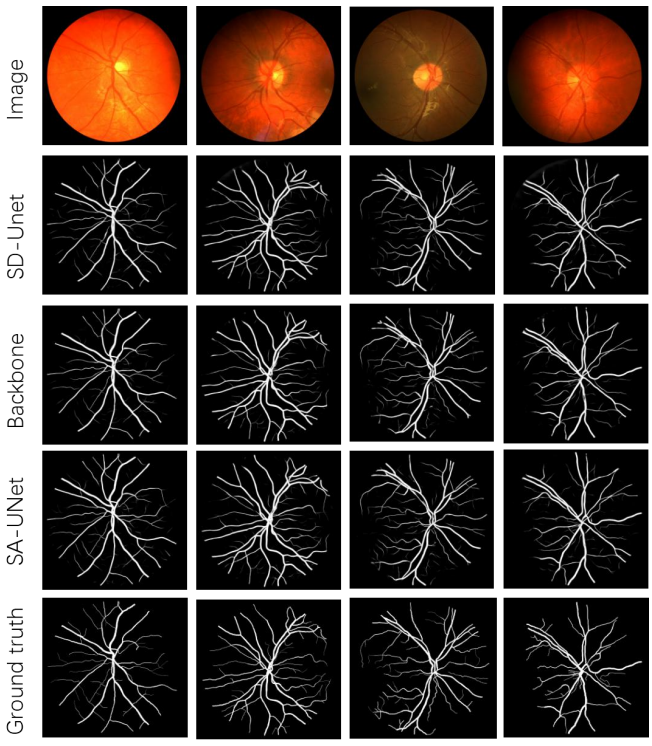

Fig. 7. Segmentation results on **CHASE_DB1**

REFERENCES

[1] Q. Guo, S. P . Duffy, K. Matthews, A. T. Santoso, M. D. Scott, and H. Ma, ‘‘Microfluidic analysis of red blood cell deformability,’’ J. Biomech., vol. 47, no. 8, pp. 1767–1776, Jun. 2014.

[2] K. Kipli, M. E. Hoque, L. T. Lim, M. H. Mahmood, S. K. Sahari, R. Sapawi, N. Rajaee, and A. Joseph, ‘‘A review on the extraction of quantitative retinal microvascular image feature,’’ Comput. Math. Methods Med., vol. 2018, pp. 1–21, Jul. 2018.

[3] Jin, Qiangguo, et al. "DUNet: A deformable network for retinal vessel segmentation." Knowledge-Based Systems 178 (2019): 149-162.

[4] Marcos Ortega, M.G. Penedo, J. Rouco, N. Barreira, M.J. Carreira, "Personal verification based on extraction and characterization of retinal feature points."Journal of Visual Languages & Computing 20.2 :80-90, 2009.

[5] Simon and I. Goldstein. A new scientific method of identification. New York State Journal of Medicine, 35(18):901–906, Sept. 1935.

[6] Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham, 2015.

[7] Wang B., Qiu S., He H. (2019) Dual Encoding U-Net for Retinal Vessel Segmentation. In: Shen D. et al. (eds) MICCAI 2019. Lecture Notes in Computer Science, vol 11764. Springer, Cham, 2019.

[8] Y Wu, Y Xia, Y Song, D Zhang, D Liu, C Zhang, W Cai. (2019) Vessel-Net: Retinal Vessel Segmentation Under Multi-path Supervision. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science, vol 11764. Springer, Cham, 2019.

[9] C. Guo, M. Szemenyei, Y. Pei, Y. Yi and W. Zhou, "SD-Unet: A Structured Dropout U-Net for Retinal Vessel Segmentation," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 2019, pp. 439-444, 2019.

[10] G. Ghiasi, T.-Y. Lin, and Q. V. Le. DropBlock: A regularization method for convolutional networks. In Neural Information Processing Systems, 2018.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[12] Karen Simonyan and Andrew Zisserman. V ery deep convolutional networks for large-scale image recognition. Advances in Neural Information Processing Systems, 2015

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent V anhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In CVPR, 2015.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In The European Conference on Computer Vision (ECCV), 2018.

[15] Liskowski, P., Krawiec, K.: Segmenting retinal blood vessels with deep neural networks. TMI 35, 2369–2380, 2016.

[16] Orlando, J.I., Prokofyeva, E., Blaschko, M.B.: A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. IEEE Trans. Biomed. Eng. 64(1), 16–27, 2017.

[17] Yan, Z., Yang, X., Cheng, K.T.: Joint segment-Level and pixel-Wise losses for deep learning based retinal vessel segmentation. IEEE Trans. Biomed. Eng. 65(9), 1912–1923, 2018.

[18] Y, Wu., Y, Xia., Y, Song., Y, Zhang., W, Cai.: Multiscale network followed network model for retinal vessel segmentation. In: Frangi, A.,Schnabel, J., Davatzikos, C., Alberola-Lopez, C., Fichtinger,G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 119–126. Springer, Heidelberg , 2018.

[19] S, Zhang., H, Fu., Y, Yan., Y, Zhang., Q, Wu., M, Yang., M, Tan. Attention Guided Network for Retinal Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science, vol 11764. Springer, Cham, 2019.