Retinal Image Segmentation with a Structure-Texture Demixing Network

Shihao Zhang[1,2], Huazhu Fu[3], Yanwu Xu[4(✉)], Yanxia Liu[1], and Mingkui Tan[1(✉)]

[1] South China University of Technology, Guangzhou, China
mingkuitan@scut.edu.cn
[2] Pazhou Lab, Guangzhou, China
[3] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[4] Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
ywxu@ieee.org

Abstract. Retinal image segmentation plays an important role in automatic disease diagnosis. This task is very challenging because the complex structure and texture information are mixed in a retinal image, and distinguishing the information is difficult. Existing methods handle texture and structure jointly, which may lead biased models toward recognizing textures and thus results in inferior segmentation performance. To address it, we propose a segmentation strategy that seeks to separate structure and texture components and significantly improve the performance. To this end, we design a structure-texture demixing network (STD-Net) that can process structures and textures differently and better. Extensive experiments on two retinal image segmentation tasks (*i.e.*, blood vessel segmentation, optic disc and cup segmentation) demonstrate the effectiveness of the proposed method.

Keywords: Retinal image · Optic disc and cup · Vessel segmentation

1 Introduction

Retinal image segmentation is important in automatic disease diagnosis [1,2]. For example, retinal vessels are correlated to the severity of diabetic retinopathy, which is a cause of blindness globally [1]. Moreover, the optic disc (OD) and optic cup (OC) are used to calculate the cup-to-disc-ratio (CDR), which is the main indicator for glaucoma diagnosis [2]. However, retinal image segmentation is often extremely challenging because retinal images often contain complex texture and structure information, which is different from general natural images.

Recently, deep neural networks (DNNs) have shown a strong ability in image segmentation with remarkable improvements [3–7]. However, existing methods are strongly biased toward recognizing textures rather than structures [8] since they handle the two types of information jointly. As a result, tiny structures that

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12265, pp. 765–774, 2020.
https://doi.org/10.1007/978-3-030-59722-1_74

are very similar to textures will be misclassified. Therefore, separately processing the structure and texture information in a retinal image is necessary. Structure-texture demixing is an essential operation in image processing that has been extensively utilized in many computer vision tasks, including image enhancement [9], optical flow [10] and image stylization [11]. However, the application of a structure-texture demixing operation in retinal image segmentation remains an open question.

Existing structure-texture demixing methods cannot adequately distinguish the boundary structures from textures, because they may have similar statistical properties [12,13]. The texture component will inevitably contain structure information. Therefore, the structure information is not fully exploited by these methods, which produces inferior segmentation results.

In this paper, we propose a **Structure-Texture Demixing Network (STD-Net)** that decomposes the image into a structure component and a texture component. Note that, the structure and texture components have different properties and need to be treated differently. We exploit two types of networks to treat them differently. The structure component mainly contains smooth structures, while the texture component mainly contains high-frequency information. Thus the structure component is suitable for processing by representative networks, and the texture component is easily overfitted, a shallower network is a better choice. We conduct extensive experiments for two tasks: vessel segmentation using the DRIVE dataset, and optic disc and cup segmentation using the ORIGA and REFUGE datasets. The results demonstrate the effectiveness of our method.

The contributions of this paper are listed as follows: 1) We propose a segmentation strategy that demix a retinal image into structure and texture components. This strategy can be applied to any segmentation framework to improve its performance. 2) We design a structure-texture demixing network (STD-Net) that can process structures and textures differently and better. 3) Extensive experiments for two retinal image segmentation tasks demonstrate the effectiveness of the proposed strategy.

2 Methodology

We illustrate the overview of our proposed STD-Net in Fig. 1. STD-Net decomposes an input image to a structure component and texture component. The structure component corresponding to the main object (smoothed part), and the texture component contains fine-grained details (almost periodic textures, noise). The segmented object's primary information is contained in the structure component. We choose M-Net [14] to process the structure component. The segmented object's detailed information is contained in the texture component, such as the boundaries. We propose a texture block to process the texture component. Details are provided in the following section.

2.1 Structure-Texture Demixing Loss Function

The structure-texture demixing module decomposes an image into a structure component and texture component by two types of loss functions, namely the

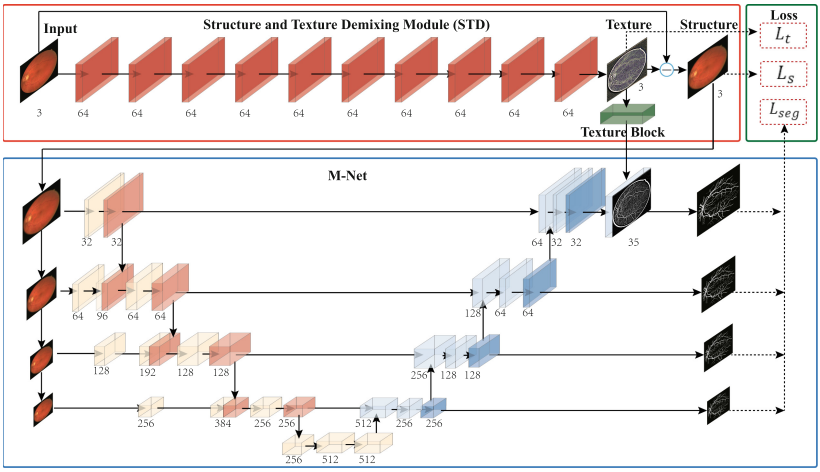


Fig. 1. Overview of the proposed STD-Net. Built on the M-Net [14] as a backbone, STD-Net decomposes the input image into structure and texture components. The structure component serves as the input of M-Net to recover the boundary structures using the texture information extracted by a *texture block* (refer to Fig. 2). The operator $\ominus$ represents the minus operation. The functions $\mathcal{L}_t$, $\mathcal{L}_s$ and $\mathcal{L}_{seg}$ represent the texture loss, structure loss, and segmentation loss, respectively.

structure loss and the texture loss. The structure and the texture loss demix images by penalizing structures and textures differently. The different penalizes are based on statistical priors that structures and textures receive different penalty under some loss functions.

Given the input image $\mathbf{I}$, the structure-texture demixing (STD) module aims to decompose $\mathbf{I}$ into two components: $\mathbf{I} \rightarrow \mathbf{S} + \mathbf{T}$, where $\mathbf{S}$ and $\mathbf{T}$ represent the structure component and texture component, respectively. This decomposition can be formulated as the following optimization problem:

$$\min_{\mathbf{S},\mathbf{T}} \lambda \mathcal{L}_s(\mathbf{S}) + \mathcal{L}_t(\mathbf{T}), \tag{1}$$

where $\mathbf{S} = \mathbf{I} - \mathbf{T}$, $\mathcal{L}_s$ is the structure loss function and $\mathcal{L}_t$ is the texture loss function, which leads $\mathbf{S}$ and $\mathbf{T}$ to different statistical properties, that is, for the structure component $\mathcal{L}_s(\mathbf{S}) \ll \mathcal{L}_t(\mathbf{S})$ and for the texture component $\mathcal{L}_s(\mathbf{T}) \gg \mathcal{L}_t(\mathbf{T})$. The constant λ is the balancing parameter.

The total variation (TV) [15] is one of the most popular structure priors; we exploit it as the structure loss function $\mathcal{L}_s$:

$$\mathcal{L}_s(\mathbf{S}) = \sum_{i,j} ||(\nabla \mathbf{S})_{i,j}||_2, \tag{2}$$

where ∇ is the spatial gradient operator. Using the TV, various demixing methods have been proposed, e.g., TV-$\triangle^{-1}$ [16], TV-L2 [15], and TV-L1 [17]. The

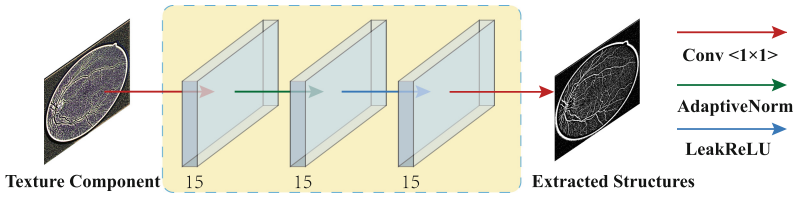


Fig. 2. Architecture of the texture block. The texture block is utilized to recover the falsely demixed structures and reduce the texture influence.

L1-norm is more suitable for structure-texture demixing [16]. Specifically, the texture loss function can be defined as follows:

$$\mathcal{L}_t(\mathbf{T}) = ||\mathbf{T}||_1. \tag{3}$$

We employ the cross-entropy function $\mathcal{L}_{seg}$ as the segmentation loss function. The final loss function $\mathcal{L}_{total}$ is defined as:

$$\begin{aligned} \mathcal{L}_{total}(\mathbf{I}, \mathbf{T}, \mathbf{R}) &= \mathcal{L}_{seg}(\mathbf{R}) + \mu(\mathcal{L}_t(\mathbf{T}) + \lambda\mathcal{L}_s(\mathbf{S})) \\ &= \mathcal{L}_{seg}(\mathbf{R}) + \mu(\mathcal{L}_t(\mathbf{T}) + \lambda\mathcal{L}_s(\mathbf{I} - \mathbf{T})), \end{aligned} \tag{4}$$

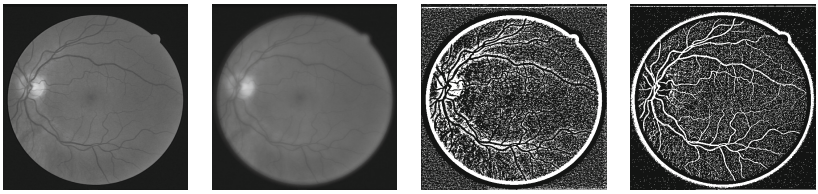
where μ and λ are trade-off parameters, and $\mathbf{R}$ is the segmentation result.

2.2 Structure-Texture Demixing Module

We show the architecture of the proposed Structure-Texture Demixing (STD) Module in Fig. 1. First, we apply STD to extract the texture component. Second, we obtain the structure component by subtracting the texture component from the input image. In this way, we confirm that $\mathbf{I} = \mathbf{S} + \mathbf{T}$. The STD consists of 10 convolutional layers with Leak ReLU to extract texture features. The extracted texture features are also serves as the input of the texture block.

Texture Block: The texture block is a component of STD. Because some structures, especially the boundary structures, may receive similar penalties from the structure loss and texture loss, they may be misclassified as the texture components. While these structures in texture component are important for segmentation, the textures and noises will affect the segmentation performance. To address it, the texture block is designed to extract boundaries and reduce the influence of textures and noises. Considering the limited amount of information in the texture component and a deep model may overfit, we design a very shallow network as the texture block. Figure 2 shows the architecture of the texture block, which contains two convolution layers, an adaptive normalization layer [18] and a leaky ReLU layer.

Figure 3 shows the visualization of the demixed structure, demixed texture, and E-structures extracted by the texture block. To help observe more clearly, we only display the green (G) channels (RGB image). The extracted structure component mainly contains smooth structures and the texture component



(a) Input image (b) Structure map (c) Texture map (d) E-structures

Fig. 3. Visualization of components in STD. Comparing (a) and (b), using texture loss, the demixed structure component maintains most of the smooth structure information and filters out many high-frequency texture noises. As shown in (c), the texture component mainly contains high-frequency information, which is a mixture of textures and boundary structures. Comparing (c) and (d), the texture block clearly helps to extract structures in the texture component, while filtering out high-frequency textures.

mainly contains high-frequency information. With the proposed texture block, we strengthen the structure information in the texture component and reduce the high-frequency textures.

3 Experiments

In this paper, we evaluate our method in vessel segmentation and optic disc/cup segmentation from retina fundus images. We train our STD-Net using Adam with a learning rate of 0.001. The batch size is set to 2. The balancing parameters λ and μ are set to 1 and 0.001 respectively.

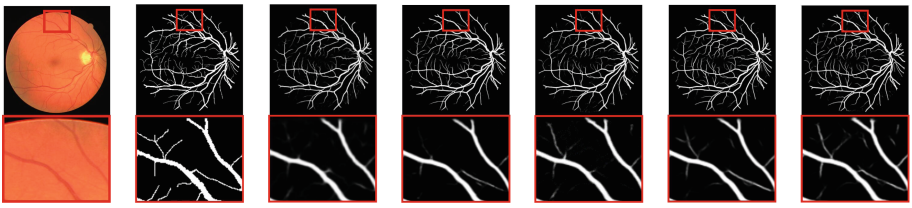
3.1 Vessel Segmentation on DRIVE

We conduct vessel segmentation experiments with DRIVE to evaluate the performance of our proposed STD-Net. The Digital Retinal Images for Vessel Extraction (DRIVE) dataset [19] contains 40 colored fundus images (20 training images and 20 testing images), which are obtained from a diabetic retinopathy screening program in the Netherlands. We resize the original images to 512×512 as inputs. Following the previous work [20], we employ Specificity (Spe), Sensitivity (Sen), Accuracy (Acc), intersection-over-union(IOU), and Area Under ROC (AUC) as measurements.

We compare our STD-Net with several state-of-the-art methods, including Li [21], Liskowski [22], MS-NFN [23],U-Net [3], M-Net [14], and AG-Net [20]. Li [21] redefines the segmentation task as cross-modality data transformation from a retinal image to a vessel map, and outputs the label map of all pixels instead of a single label of the center pixel. Liskowski [22] trains a deep neural network with samples that were preprocessed with global contrast normalization and zero-phase whitening and augmented using geometric transformations and gamma corrections. MS-NFN [23] generates multi-scale feature maps with an 'up-pool' submodel and a 'pool-up' submodel. U-Net [3] applies a contracting

Table 1. Quantitative comparison of segmentation results with DRIVE

Method	Acc	AUC	Sen	Spe	IOU
Li [21]	0.9527	0.9738	0.7569	0.9816	–
Liskowski [22]	0.9535	0.9790	0.7811	0.9807	–
MS-NFN [23]	0.9567	0.9807	0.7844	0.9819	–
AG-Net [20]	0.9692	0.9856	0.8100	0.9848	0.6965
U-Net [3]	0.9681	0.9836	0.7897	0.9854	0.6834
M-Net [14]	0.9674	0.9829	0.7680	**0.9868**	0.6726
STD-Net	**0.9695**	**0.9863**	**0.8151**	0.9846	**0.6995**



(a) Input (b) GT (c) M-Net (d) AG-Net (e) BL (f) BLST (g) Ours

Fig. 4. Example results for DRIVE. M-Net, AG-Net, and BL disregard some edge structures, which are very similar to textures. Conversely, by decomposing structures and textures, BLST gains better discrimination power and detects more tiny structures. Comparing (f) and (g), when adding the texture block, more tiny boundary structures are detected.

path to capture context and a symmetric expanding path to enable precise localization. M-Net [14] introduces multi-input and multi-output to learn hierarchical representations. AG-Net [20] proposes a structure sensitive expanding path and incorporates it into M-Net.

Table 1 shows the performances of different methods for DRIVE. Based on the results, for the four metrics AUC, Acc, Sen, and IOU, the proposed STD-Net achieves the highest value. STD-Net outperforms the backbone M-Net by 0.0021, 0.0034, 0.0471 and 0.0269 in terms of Acc, AUC, Sen, and IOU, respectively. Note that the proposed STD-Net achieves a much higher Sen score than M-Net, which shows that our structure-texture demixing mechanism improves the structure detection ability of models.

We remove the texture block, structure loss $\mathcal{L}_s$, and texture loss $\mathcal{L}_t$ from STD-Net and name the baseline model as **BL**. The model **BLST** is formed by adding the structure-texture loss into BL. Figure 4 shows a test example, including the ground truth vessel (GT) and segmentation results obtained by M-Net, AG-Net, BL, BLST, and the proposed STD-Net. The experimental results of BL and BLST are shown in Table 3.

Table 2. Comparisons of different methods with ORIGA and REFUGE

	ORIGA			REFUGE		
Method	OE_{disc}	OE_{cup}	OE_{total}	OE_{disc}	OE_{cup}	OE_{total}
ASM [24]	0.148	0.313	0.461	–	–	–
SP [25]	0.102	0.264	0.366	–	–	–
LRR [26]	–	0.244	–	–	–	–
U-Net [3]	0.115	0.287	0.402	0.171	0.257	0.428
AG-Net [20]	**0.061**	0.212	0.273	0.178	0.220	0.398
M-Net [14]	0.071	0.230	0.301	0.204	0.231	0.435
STD-Net	0.063	**0.208**	**0.271**	**0.168**	**0.217**	**0.385**

3.2 Optic Disc/Cup Segmentation on ORIGA

Optic Disc/Cup Segmentation is another important retinal segmentation task. In this experiment, we employ the ORIGA dataset, which contains 650 fundus images with 168 glaucomatous eyes and 482 normal eyes. The 650 images are divided into 325 training images and 325 testing images (including 73 glaucoma cases and 95 glaucoma cases, respectively). We crop the OD area and resize it to 256×256 as the input. We compare STD-Net with several state-of-the-art methods, including ASM [24], Superpixel [25], LRR [26], U-Net [3], M-Net [14], and AG-Net [20]. The ASM [24] employs the circular hough transform initialization to segmentation. The superpixel method [25] utilizes superpixel classification to detect the OD and OC boundaries. The method in LRR [26] obtains satisfactory results but only focuses on OC segmentation. AG-Net [20] also strengthens the structure information but is easily influenced by the textures.

Following the setting in [20], we localize the disc center with a pre-trained LinkNet [27] and then enlarge 50 pixels of bounding-boxes in up, down, right and left directions to crop the OD patch as the input image. The polar transformation is also exploited to improve the segmentation performance. We employ overlapping error (OE) as the evaluation metric, which is defined as $OE = 1 - \frac{A_{GT} \bigcap A_{SR}}{A_{GT} \bigcup A_{SR}}$. A_{GT} and A_{SR} denote the ground truth area and segmented mask, respectively. In particular, OE_{disc} and OE_{cup} are the overlapping error of OD and OE. OE_{total} is the sum of OE_{disc} and OE_{cup}.

Table 2 shows the segmentation results. Our method outperforms all the state-of-the-art OC segmentation algorithms, which demonstrates the effectiveness of our model. For OD segmentation, the proposed STD-Net is slightly lower than AG-Net, but STD-Net achieves the best performance on OC segmentation and better performance when considering OC and OD segmentation. Our STD-Net performs much better than the original M-Net, which further demonstrates that our structure-texture demixing method is beneficial for the segmentation performance.

We obtained similar results with the REFUGE dataset [28], which are shown in Table 2. The training set and validation set of REFUGE have distinct appearances due to different shooting equipment, which requires a high generalization

Table 3. Ablation study with DRIVE and ORIGA

	DRIVE					ORIGA		
Method	Acc	AUC	Sen	Spe	IOU	OE_{disc}	OE_{cup}	OE_{total}
BL	0.9678	0.9829	0.7776	**0.9864**	0.6785	0.065	0.217	0.282
BL+$\mathcal{L}_s$	0.9684	0.9842	**0.8236**	0.9827	0.6948	**0.063**	0.211	0.274
BL+$\mathcal{L}_t$	0.9687	0.9841	0.8167	0.9837	0.6951	0.064	0.213	0.277
BLST	0.9691	0.9859	0.8201	0.9837	0.6984	**0.063**	0.210	0.273
STD-Net	**0.9695**	**0.9863**	0.8151	0.9846	**0.6995**	**0.063**	**0.208**	**0.271**

ability to reduce overfitting. Therefore, the results with REFUGE can better demonstrate the ability of structural texture decomposition.

3.3 Ablation Study

We conduct an ablation investigation to further verify the effectiveness of the structure-texture demixing mechanism and texture block. The results for DRIVE are presented in Table 3. We note several interesting observations. First, when BL considers the structure loss $\mathcal{L}_s$ or the texture loss $\mathcal{L}_t$, the results are improved with metrics other than Spe. With the structure loss, BL achieved the highest Sen score, which shows that more vessel structures are detected. Second, when BL considers both the structure loss $\mathcal{L}_s$ and the texture loss $\mathcal{L}_t$, it achieves higher Acc, AUC and IOU scores, which demonstrates the superiority of the structure-texture demixing strategy. Last, when BL further incorporates the texture block (STD-Net), it achieves the highest scores for Acc, AUC, and IOU. This finding demonstrates the effectiveness of the texture block. As shown in Table 3, similar results are obtained for ORIGA.

4 Conclusion

In this paper, we have proposed a trainable structure-texture demixing network (STD-Net) to decompose an image into a structure component and texture component and separately process them. In this way, the segmentation model focuses more on structure information and reduces the influence of texture information. We have also proposed a texture block to further extract the structural information from the texture component, which substantially improves the segmentation results. Extensive experiments for two retinal image segmentation tasks (*i.e.*, blood vessel segmentation, optic disc and cup segmentation) demonstrate the effectiveness of our proposed method.

Acknowledgments. This work was partially supported by National Natural Science Foundation of China (NSFC) 61836003 (key project), Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183, Guangdong Provincial Scientific and Technological Funds under Grant 2018B010107001,

Grant 2019B010155002, Tencent AI Lab Rhino-Bird Focused Research Program (No. JR201902), Fundamental Research Funds for the Central Universities D2191240.

References

1. Jelinek, H., Cree, M.J.: Automated Image Detection of Retinal Pathology. Crc Press, Boca Raton (2009)
2. Hancox, O., Michael, D.: Optic disc size, an important consideration in the glaucoma evaluation. Clin. Eye Vis. Care **11**(2), 59–62 (1999)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2015)
4. Gu, Z., Cheng, J., et al.: Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging **38**(10), 2281–2292 (2019)
5. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L.: Et-net: a generic edge-attention guidance network for medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 442–450. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_49
6. Zhang, Y., et al.: From whole slide imaging to microscopy: deep microscopy adaptation network for histopathology cancer image classification. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 360–368. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_40
7. Zhang, Y., Wei, Y., et al.: Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Trans. Image Process. **29**, 7834–7844 (2020)
8. Geirhos, R., Rubisch, P., et al.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. International Conference on Learning Representations (2019)
9. Guo, X., Li, Y., Ling, H.: Lime: low-light image enhancement via illumination map estimation. IEEE Trans. Image Process. **26**(2), 982–993 (2016)
10. Revaud, J., Weinzaepfel, P., et al.: Epicflow: edge-preserving interpolation of correspondences for optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
11. Gastal, E.S., Oliveira, M.M.: Domain transform for edge-aware image and video processing. ACM SIGGRAPH 2011 papers (2011)
12. Xu, L., Yan, Q., et al.: Structure extraction from texture via relative total variation. ACM Trans. Graph. **31**(6), 1–10 (2012)
13. Kim, Y., Ham, B., Do, M.N., Sohn, K.: Structure-texture image decomposition using deep variational priors. IEEE Trans. Image Process. **28**(6), 2692–2704 (2018)
14. Fu, H., Cheng, J., et al.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging **37**(7), 1597–1605 (2018)
15. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
16. Aujol, J.-F., Gilboa, G., et al.: Structure-texture image decomposition modeling, algorithms, and parameter selection. Int. J. Comput. Vis. **67**(1), 111–136 (2006). https://doi.org/10.1007/s11263-006-4331-z
17. Alliney, S.: A property of the minimum vectors of a regularizing functional defined by means of the absolute norm. IEEE Trans. Signal Process. **45**(4), 913–917 (1997)

18. Ogasawara, E., Martinez, L.C., De Oliveira, D., Zimbrão, G., Pappa, G.L., Mattoso, M.: Adaptive normalization: a novel data normalization approach for non-stationary time series. In: The 2010 International Joint Conference on Neural Networks. IEEE (2010)
19. Staal, J., Abràmoff, M.D., et al.: Ridge-based vessel segmentation in color images of the retina. IEEE Trans. Med. Imaging **23**(4), 501–509 (2004)
20. Zhang, S., et al.: Attention guided network for retinal image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 797–805. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_88
21. Li, Q., Feng, B., et al.: A cross-modality learning approach for vessel segmentation in retinal images. IEEE Trans. Med. Imaging **35**(1), 109–118 (2016)
22. Liskowski, P., Krawiec, K.: Segmenting retinal blood vessels with deep neural networks. IEEE Transactions on Medical Imaging (2016)
23. Wu, Y., Xia, Y., Song, Y., Zhang, Y., Cai, W.: Multiscale network followed network model for retinal vessel segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 119–126. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_14
24. Yin, F., Liu, J., et al.: Model-based optic nerve head segmentation on retinal fundus images. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE (2011)
25. Cheng, J., Liu, J., et al.: Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE Trans. Med. Imaging **32**(6), 1019–1032 (2013)
26. Xn, Y., et al.: Optic cup segmentation for glaucoma detection using low-rank superpixel representation. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 788–795. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_98
27. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing. IEEE (2017)
28. Orlando, J.I., Fu, H., et al.: REFUGE challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)