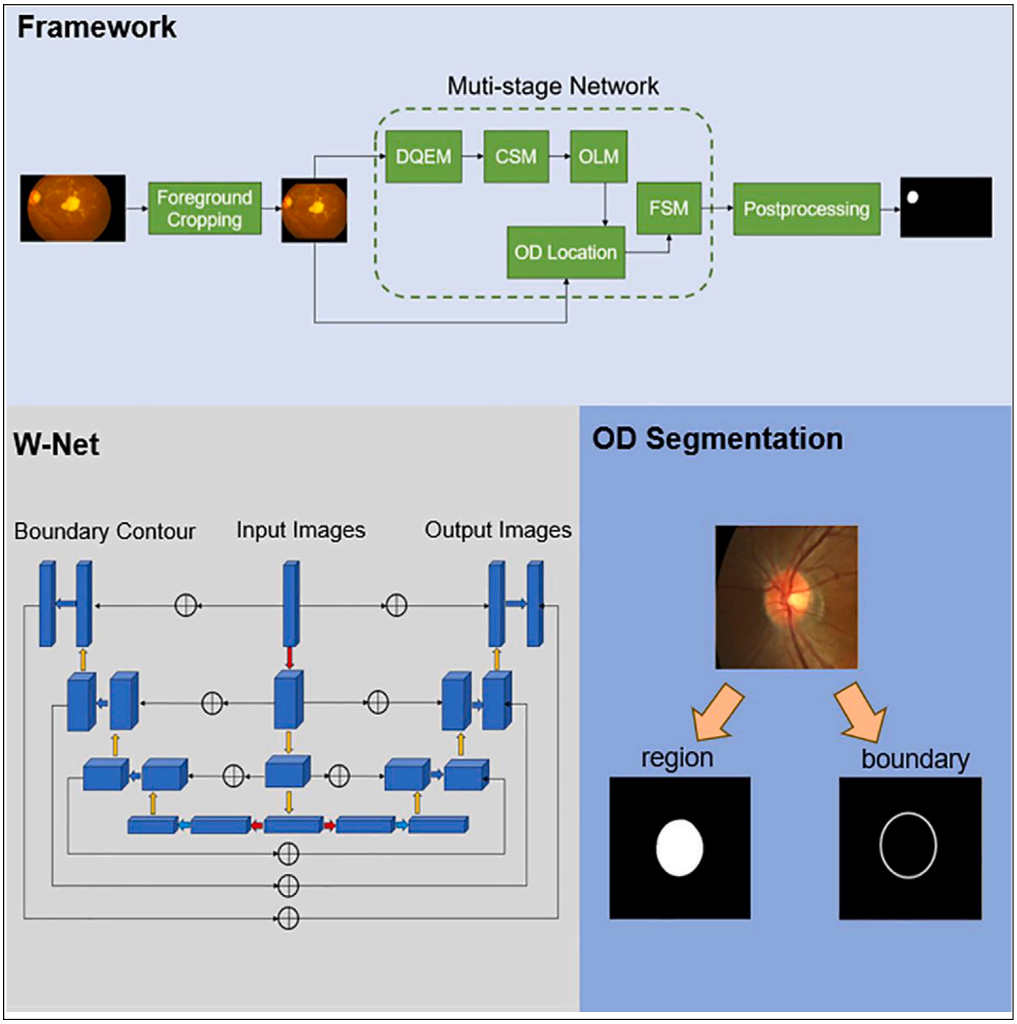
iScience

CellPress
OPEN ACCESS

Article

W-Net: A boundary-aware cascade network for robust and accurate optic disc segmentation



Shuo Tang, Chongchong Song, Defeng Wang, Yang Gao, Yuchen Liu, Wang Lv

dfwang@buaa.edu.cn (D.W.)
yanggao@buaa.edu.cn (Y.G.)

Highlights

A multi-stage segmentation network is proposed to enhance the generalization performance

The W-Net module is proposed for the first time in the multi-stage segmentation framework

The boundary loss is introduced in W-Net to improve OD segmentation performance

Our method shows outstanding OD segmentation ability on the other three mature datasets

Tang et al., iScience *27*, 108247
January 19, 2024 © 2023
https://doi.org/10.1016/j.isci.2023.108247

iScience

CellPress
OPEN ACCESS

Article

W-Net: A boundary-aware cascade network for robust and accurate optic disc segmentation

Shuo Tang,[1] Chongchong Song,[1] Defeng Wang,[1,*] Yang Gao,[1,*] Yuchen Liu,[1] and Wang Lv[1,2]

SUMMARY

Accurate optic disc (OD) segmentation has a great significance for computer-aided diagnosis of different types of eye diseases. Due to differences in image acquisition equipment and acquisition methods, the resolution, size, contrast, and clarity of images from different datasets show significant differences, resulting in poor generalization performance of deep learning networks. To solve this problem, this study proposes a multi-level segmentation network. The network includes data quality enhancement module (DQEM), coarse segmentation module (CSM), localization module (OLM), and fine segmentation stage module (FSM). In FSM, W-Net is proposed for the first time, and boundary loss is introduced in the loss function, which effectively improves the performance of OD segmentation. We generalized the model in the REFUGE test dataset, GAMMA dataset, Drishti-GS1 dataset, and IDRiD dataset, respectively. The results show that our method has the best OD segmentation performance in different datasets compared with state-of-the-art networks.

INTRODUCTION

Accurate optic disc (OD) segmentation has a great significance for computer-aided diagnosis of different types of eye diseases. For example, accurate OD segmentation can locate retinal blood vessels, which is one of the important steps to calculate central retinal artery equivalent (CRAE) and central retinal vein equivalent (CRVE), and these two parameters are important for early detection and diagnosis of diabetes and hypertension.[1] In addition, OD segmentation also helps to establish a retinal frame, which can be used to determine the location of many retinal abnormalities, such as exudates, edema, microaneurysms, and hemorrhages.[2]

Recently, convolutional neural networks (CNN) for medical image segmentation have been more satisfactory than traditional human-based feature extraction methods. The fully convolutional neural network (FCN)[3] solves the problem of pixel-by-pixel classification (i.e., semantic segmentation) using CNN. The emergence of U-Net[4] has contributed significantly to the wide application of CNNs in medical image segmentation. To extend the use of U-Net, Zhou et al. added a multi-layer adaptive depth mechanism to U-Net and proposed U-Net++.[5] Considering that the receptive field limits the model, Google's team employed a segmentation framework of the DeepLab series using atrous convolution.[6–9] Inspired by the U-Net, DeepLab v3+[7] adopted a fusion strategy of multi-scale features to advance the segmentation performance. In the field of OD segmentation of CFP, Hua zhu et al. presented a multi-stage segmentation method that firstly localizes the OD, then performs spatial transformation using polar coordinate transformation, and finally uses M-Net for segmentation.[10]

Despite the demonstrated utility of OD segmentation, there is no universally accepted method capable of segmenting accurately and efficiently across a wide variety of datasets. A fundamental assumption of deep learning is that the distribution of samples is independent and identical. That is, the model will perform well if the data distribution between the training and testing dataset is similar to the appearance of the images. However, due to different acquisition devices and methods, the acquired images vary significantly in resolution, size, contrast, sharpness, etc. We refer to this difference between datasets as domain shift. In this study, we use four OD datasets: GAMMA dataset, Drishti-GS1 dataset, and IDRiD dataset. The differences between these datasets are meticulously presented and analyzed in Figure 1 and Table 1. Current methods are usually trained on a particular dataset and can only perform well on the corresponding test set. However, in cases where the differences between the test and training set images are more pronounced, or the image contrast is poor, the inference ability of the model is often unsatisfactory. For example, M-Net[10] shows outstanding segmentation performance on the REFUGE dataset, yet when the model is used to infer other datasets, its performance becomes unsatisfying, as shown in Figure 2.

In order to obtain a model with good generalization performance and accurate OD segmentation results on different datasets, we propose a multi-level segmentation model, which is trained on a single dataset and has good OD segmentation results on different datasets. Some researchers have proposed many effective methods, which can be divided into two categories based on whether the majority of the network is generative adversarial networks (GAN)[11–13] or semi-supervised.[14–16]

[1] School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China
[2] Lead contact
*Correspondence: dfwang@buaa.edu.cn (D.W.), yanggao@buaa.edu.cn (Y.G.)
https://doi.org/10.1016/j.isci.2023.108247

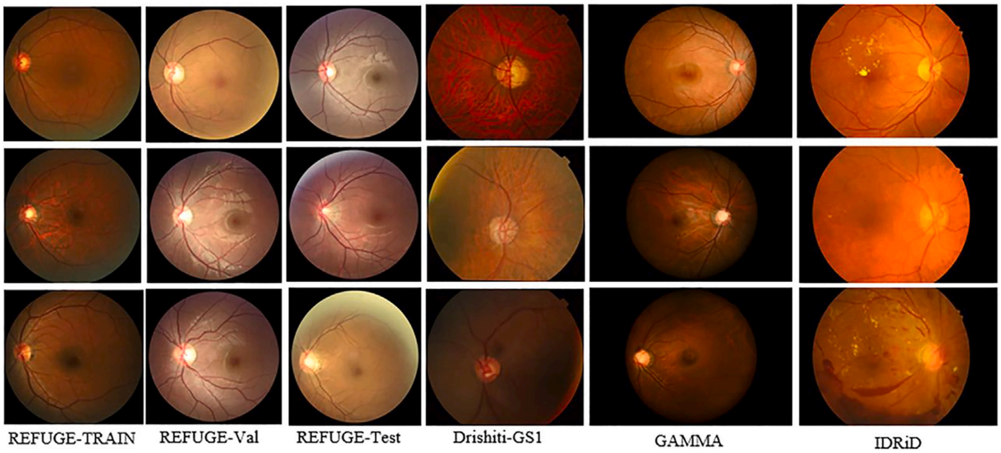


Figure 1. Examples of different datasets

From left to right are REFUGE's training dataset, validation dataset, test dataset, Drishiti-GS1 dataset, and IDRiD dataset.

Generative adversarial networks

The fundamental idea of the GAN methods is to learn the image distribution from the source and target domains. In the inference phase, the images in the target domain are transformed into the source domain, and the transformed images are segmented. In clinical practice, the generator adds extra and false information to the transformed images, which does not fully reflect the real situation of the patient. In addition, this false information will also affect the segmentation accuracy.

Semi-supervised networks

The semi-supervised approach obtains a segmentor by gold-standard learning of the source domain data, then applies the model to images in the target domain to obtain pseudo-labels. Again, researchers will fine-tune the model under the supervision of pseudo-labels to solve the problem of the data domain transfer. The deficiencies of these methods can already be reflected in principle. If we want to generalize the model on a new dataset, we need to retrain or fine-tune the model with the new dataset. Accordingly, the generalization performance of this model will be limited. Zhu et al. developed a generic OD and OC segmentation network for multi-device CFP to address these issues.[17] The authors mixed CFPs from different datasets and then trained and tested them on the mixed dataset. This approach performs well on the hybrid dataset but may not work well for images without similar distribution in the hybrid dataset.

RESULTS

Evaluation metric

In this study, the segmentation performance is evaluated by Dice Coefficient, IOU, and Hausdorff Distance, respectively:

$$Dice(pred, gt) = \frac{2 \times (pred \cap gt)}{pred + gt} \quad \text{(Equation 1)}$$

Table 1. An overview of the datasets

Dataset	Number of images	Resolution	R	G	B
Train	400	2056 × 2124	27.37 ± 7.10	42.79 ± 9.51	71.86 ± 15.18
Val	400	1634 × 1634	54.88 ± 9.77	68.92 ± 10.88	104.37 ± 12.79
Test	400	1634 × 1634	56.51 ± 9.15	69.03 ± 10.48	102.99 ± 12.48
Drishti-GS1	101	(1741–1845) × (2046–2468)	12.95 ± 5.82	40.80 ± 12.09	85.61 ± 21.31
GAMMA	100	2000 × 2992 1934 × 1956	10.23 ± 8.47	33.08 ± 10.58	66.73 ± 16.16
IDRiD	81	2848 × 4288	17.01 ± 14.63	57.83 ± 10.34	118.32 ± 15.03

Train, Val, and Test are three subsets of the REFUGE dataset. Drishti-GS1 and IDRiD contain Train and Test, respectively, in the corresponding datasets. GAMMA contains only the Train subset. (1741–1845) × (2046–2468) indicates that in the Drishti-GS1 dataset, the horizontal resolution is distributed from 1741 to 1845, and the vertical resolution is distributed from 2046 to 2468.

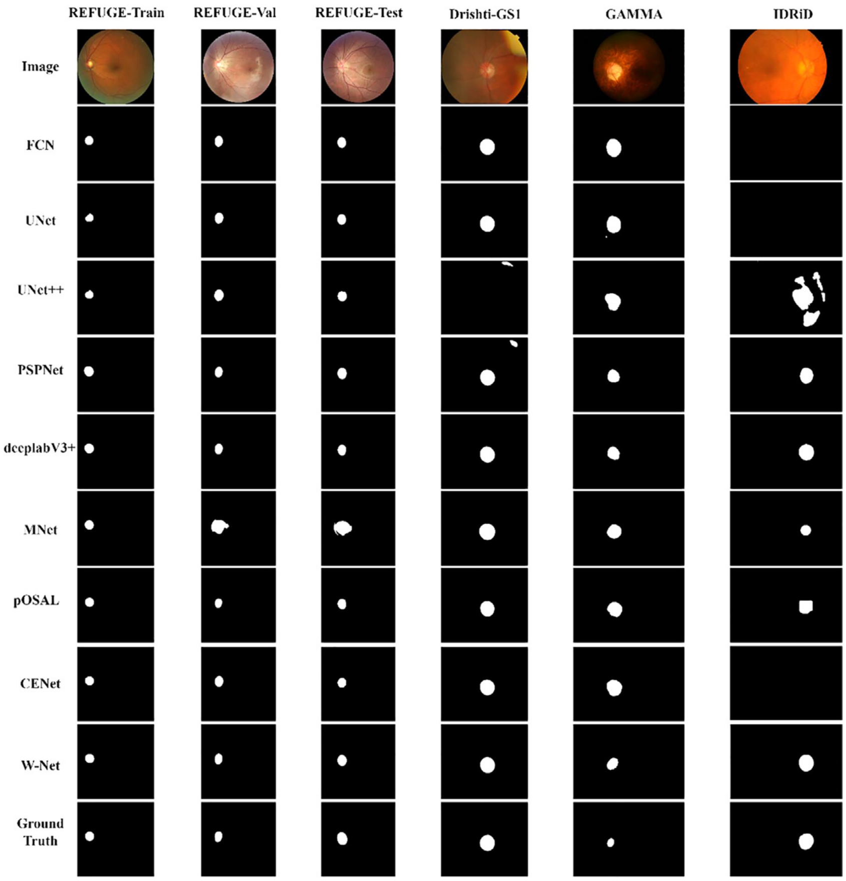


Figure 2. Segmentation performance of M-Net in different OD datasets

$$IOU(pred, gt) = \frac{pred \cap gt}{pred \cup gt} \tag{Equation 2}$$

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{Equation 3}$$

where *pred* is the prediction result, and *gt* is the annotation information of the dataset. In our experiments, *pred* and *gt* are binary images. We use the average Dice and the minimum Dice to represent the generalization performance. The average Dice can reflect the segmentation performance for most cases, whereas the minimum Dice can reflect the performance for the worst quality images. In a sense, the minimum Dice represents the generalization performance more correctly.

Table 2. The average Dice of ablation experiments

Avg Dice	REFUGE Train	REFUGE Val	REFUGE Test	Drishiti-GS1	GAMMA	IDRiD
FCN	0.9596	0.9577	0.9543	0.9534	0.858	0.8095
UNet	0.9532	0.954	0.9534	0.9423	0.8723	0.8173
Unet++	0.9525	0.9488	0.9459	0.4247	0.6197	0.7963
PSPNet	0.9593	0.9554	0.955	0.9469	0.8894	0.8403
deeplabV3+	0.9563	0.956	0.9519	0.8808	0.8967	0.8605
Mnet	0.9625	0.9619	0.7953	0.9498	0.898	0.739
pOSAL	0.9696	0.9536	0.9601	0.9533	0.9383	0.913
CE-Net	0.9796	0.9519	0.9512	0.9217	0.8862	0.9077
W-Net	0.9783	0.9749	0.9611	0.9534	0.941	0.9525

Table 3. The average IOU of ablation experiments

Avg IOU	REFUGE Train	REFUGE Val	REFUGE Test	Drishiti-GS1	GAMMA	IDRiD
FCN	0.9232	0.9196	0.9137	0.9124	0.7942	0.7385
UNet	0.9124	0.9129	0.9119	0.8946	0.81	0.756
Unet++	0.9108	0.9041	0.9008	0.3324	0.6326	0.7029
PSPNet	0.9224	0.9154	0.9147	0.9014	0.8153	0.7699
deeplabV3+	0.9175	0.9165	0.9094	0.8331	0.8345	0.7905
Mnet	0.9286	0.9272	0.676	0.907	0.8306	0.5925
pOSAL	0.9414	0.9123	0.924	0.9118	0.8968	0.8533
CE-Net	0.9602	0.9097	0.9079	0.8721	0.804	0.8535
W-Net	0.9568	0.9517	0.9512	0.9137	0.8898	0.9078

Comparison experiments results

On each dataset, we trained FCN, UNet, UNet++, PSPNet, deeplabv3+, MNet, CE-Net, pOSAL, and W-Net, respectively. In this study, we produce the average Dice Coefficient, Hausdorff Distance, and IOU of W-Net and each SOTA model to compare the average segmentation performance of the models. In addition, we also produce the minimum Dice Coefficient to compare the worst segmentation performance of the models. In addition, we also calculate the minimum Dice Coefficient to compare the worst segmentation performance of the models. The larger the Dice Coefficient and IOU, the better the segmentation performance, and the smaller the Hausdorff Distance, the better the segmentation performance.

It can be seen from Table 2 that W-Net is only slightly lower than CE-Net on the training set of REFUGE and does not exceed 0.7 percentage points. However, in the Drishiti-GS1, GAMMA, and IDRiD datasets, the segmentation performance of W-Net is better than other SOTA models. Especially on the IDRiD dataset, W-Net is 3.95% higher than the best performing pOSAL model. In addition, as shown in Tables 3 and 4, we also calculated the average Hausdorff Distance and average IOU, further verifying that the average segmentation performance of W-Net in the other three datasets is better than other SOTA models.

More importantly, W-Net's OD segmentation performance is significantly better than other SOTA models when segmenting OD images with large differences in training datasets. It can be seen from Table 5 that when the image is segmented with a large difference from the training dataset, some SOTA models will appear in the extreme case of dice = 0, and the OD image cannot be segmented. At this time, the minimum Dice Coefficient of W-Net can still get 64.57%, which is 18.62% higher than the best performing Mnet model, and its OD segmentation performance is significantly better than other SOTA models. In summary, compared with other SOTA models, the average segmentation performance of W-Net trained with a single dataset in the other three datasets is slightly improved, and it has obvious advantages in the segmentation of pictures with large differences. Therefore, the model has better generalization performance.

Ablation experiments results

To verify the effectiveness of the multi-level segmentation framework and boundary loss function, two ablation experiments are conducted. Because the backbone of W-Net is CE-Net,[23] CE-Net is used to verify the effectiveness of the multi-level segmentation framework. CE-Net1 was OD segmentation only by CE-Net, and CE-Net2 is OD segmentation using CE-Net as FEM in the multi-level segmentation framework. The ablation results are shown in Tables 6 and 7. By comparing the average Dice of CE-Net1 and CE-Net2, we can conclude that the multi-stage segmentation model is less effective than the single stage in the case of a similar image (test dataset). Nevertheless, the advantages of the multi-stage model can be fully revealed by generalizing images from other datasets. We believe that there are two reasons for the

Table 4. The average Hausdorff Distance of ablation experiments

Avg Hausdorff	REFUGE Train	REFUGE Val	REFUGE Test	Drishiti-GS1	GAMMA	IDRiD
FCN	0.2357	0.1891	0.2348	0.4252	1.1667	7.3951
UNet	0.4243	0.2515	0.2578	0.7203	1.5006	13.7276
Unet++	0.4107	0.6188	1.0824	55.0164	16.2824	38.4777
PSPNet	0.2346	0.2027	0.225	0.5845	1.4459	2.3588
deeplabV3+	0.3427	0.2147	0.2689	2.2018	1.8171	14.4152
Mnet	0.2171	0.1741	4.873	0.5499	1.9571	7.9059
pOSAL	0.1202	0.2238	0.1678	0.3594	4.6172	1.5848
CE-Net	0.0584	0.2663	0.3309	1.9567	2.9993	5.844
W-Net	0.0675	0.0688	0.0711	0.3835	0.555	0.5354

Table 5. The minimum Dice of ablation experiments

Min Dice	REFUGE Train	REFUGE Val	REFUGE Test	Drishiti-GS1	GAMMA	IDRiD
FCN	0.6891	0.8651	0.7836	0.7375	0	0
UNet	0.4485	0.7702	0.7291	0.7615	0	0
Unet++	0.5713	0.6642	0.1111	0	0	0
PSPNet	0.8148	0.8693	0.7533	0.7676	0	0
deeplabV3+	0.6474	0.8238	0.6968	0	0	0
Mnet	0.6877	0.8131	0.3834	0.7661	0.4595	0.5418
pOSAL	0.8304	0.8212	0.8177	0.8135	0	0
CE-Net	0.9396	0.7161	0.7822	0.0006	0.3937	0
W-Net	0.9326	0.9259	0.8722	0.8654	0.6457	0.856

composition of this phenomenon. On the one hand, information reduction is caused by cropping and resampling. On the other hand, the interpolation method adopted cannot be adapted to the original image in the postprocess. By comparing the minimum Dice of CE-Net1 and CE-Net2, we can clearly find that the multi-stage segmentation model can handle relatively difficult images to segment in the OD region. This fully reflects the effectiveness of this method in improving the robustness of the model.

In order to verify the effectiveness of the loss function, an ablation experiment of boundary loss is carried out. It can be seen from Equation 15 that the loss function of the left branch in W-Net consists of boundary loss and Dice loss. When setting $\alpha = 1$, $\beta = 0.5$, there is only dice loss and no boundary loss in the loss function. The experimental results are shown in the "W-Net-no-BL" rows of Tables 6 and 7. When setting $\alpha = \beta = 0.4$, the loss function of W-Net consists of boundary loss and dice loss and achieves the best generalization performance. The experimental results are shown in the "W-Net" row of Tables 6 and 7.

Influence of super parameter

In order to make our proposed method work with better performance, we conducted the following experiments. First, we analyzed the parameters α and β, and the results are shown in Table 8. The generalization performance is best and most stable when $\alpha = 0.4$ and $\beta = 0.4$. At this point, the output is equivalent to being influenced by the left and right branches. Our explanation for this is that the nature of W-Net is a voting mechanism, and the performance of the left and right branches is comparable, so it can achieve good results under the same influence. Secondly, we wanted to know which strategy facilitates generalization in the case of using boundary loss, fine-tune training, or direct training. We set the same parameters to direct training and fine-tune training, and the experimental results are shown in Table 8. With these experiments, we can conclude similarly to that in[18]: when using boundary loss, we first need to learn the region information and then fine-tune the model under boundary supervision.

DISCUSSION

In this study, we combine the prior knowledge of manual segmentation of OD and develop a multi-stage segmentation framework to accommodate the problem of data domain shift. This framework can be divided into four modules: data quality enhancement module (DQEM), OD coarse segmentation module (CSM), OD localization module (OLM), and fine segmentation stage module (FSM). Our proposed model's average Dice and minimum Dice metrics are more advanced in all four datasets than the state-of-the-art (SOTA) method. Particularly, for the FSM module, we innovatively proposed W-Net. With multiple branches, W-Net learned boundary information while learning the region information, and the multi-branch structure can also achieve the purpose of ensemble learning.

More importantly, we implement cross-dataset testing in this paper. The comparative experiments show that our proposed method can achieve comparable performance with the SOTA method with good-quality images and far outperforms the SOTA method in images with poor contrast. This can be clearly seen in our comparative experiments that the robustness of the multi-stage W-Net model is far superior to other SOTA methods. These experiments show that our proposed method dramatically improves robustness without retraining on new datasets.

Table 6. The average Dice of ablation experiments

Method	Train	Val	Test	Drishiti-GS1	GAMMA	IDRiD
CE-Net1	0.9796	0.9519	0.9512	0.9217	0.8862	0.9077
CE-Net2	0.9735	0.9751	0.9464	0.9699	0.9264	0.9450
W-Net-no-BL	0.9782	0.9767	0.9602	0.9546	0.9386	0.9516
W-Net	0.9783	0.9749	0.9611	0.9534	0.941	0.9525

Table 7. The minimum Dice of ablation experiments

Method	Train	Val	Test	Drishti-GS1	GAMMA	IDRiD
CE-Net1	0.9396	0.7161	0.7822	0.0062	0.3937	0
CE-Net2	0.9302	0.9386	0.8218	0.8372	0.5844	0.7149
W-Net-no-BL	0.9178	0.877	0.8764	0.8044	0.6413	0.8304
W-Net	0.9326	0.9259	0.8722	0.8654	0.6457	0.8560

Limitations of the study

Our work can effectively improve the robustness of OD segmentation; unfortunately there are two main limitations. On the one hand, the multi-stage cascading method has defects, and on the other hand, our method cannot effectively improve the segmentation performance of images with better quality.

In the multi-stage cascading method, we only used CycleGAN as the DQEM module and used two data sets as the training sets of DQEM. In order to further improve the generalization of the model, we can use multiple advanced generative adversarial networks, such as MedGAN,[27] stylegan,[28–31] Pixel2style2pixel GAN,[32] and Encoder4Editing[33] to perform data enhancement on the four data sets respectively. Through adversarial learning, synthetic images are generated from the source domain to the target domain and from the target domain to the source domain, and the synthetic images are added to the training set for training together. Ultimately, the types of annotated images can be greatly increased, thereby improving the generalization of the model.

There are three main reasons why our model cannot improve the segmentation performance of good-quality images: the performance ceiling of OD segmentation, the accuracy loss caused by the resampling process of our method, and the limitation of the way the model extracts features. The limitation caused by the second reason can be improved by directly training the segmentation network through the GAN method described earlier instead of multi-stage method. For the limitation of the way the model extracts features, inspired by ViT,[25] it is necessary to have a deeper understanding of the reasons why Transformer[34] is successful in natural language processing. Models like Swin Transformer,[35] SegFormer,[36] SegNeXt, etc. can be used for feature extraction.

Future expectations

One of our future research focuses on replacing our CFM with the newly proposed self-attention-based structure[24–26] to enhance the performance of OD localization, which may combine positional information to determine the location of OD. We propose a generalized OD multi-stage segmentation framework and W-Net. Furthermore, we will explore other segmentation tasks in the future using the W-Net proposed in this paper, such as retinal blood vessel segmentation and arterial and vein segmentation in CFP.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Data quality enhancement module (DQEM)

Table 8. The average Dice of parameter setting experiments

	Fine-tune	α	β	Train	Val	Test	Drishti-GS1	GAMMA	IDRiD
1	False	0.7	0.3	0.9771	0.9780	0.9593	0.9507	0.9377	0.9506
2	False	0.6	0.4	0.977	0.9763	0.9594	0.9564	0.9383	0.9510
3	False	0.5	0.5	0.9782	0.9767	0.9602	0.9546	0.9388	0.9525
4	False	0.4	0.6	0.9772	0.9774	0.9599	0.9538	0.9384	0.9512
5	True	0.45	0.45	0.9783	0.9767	0.9604	0.9559	0.9388	0.9526
6	True	0.4	0.4	0.9783	0.9749	0.9611	0.9534	0.941	0.9525
7	False	0.45	0.45	0.9733	0.9715	0.9597	0.9431	0.9386	0.9505
8	False	0.4	0.4	0.9672	0.9652	0.9589	0.9303	0.9403	0.9487

- ○ Coarse segmentation module (CSM)
- ○ OD localization module (OLM)
- ○ Fine segmentation module (FSM)
- ○ DAC module and RMP module
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

We received financial support from the Beijing Municipal Natural Science Foundation (No. 4212012).

AUTHOR CONTRIBUTIONS

Methodology: Shuo Tang, Chongchong Song, Defeng Wang, Yang Gao;

Investigation: Chongchong Song, Yang Gao, Defeng Wang, Shuo Tang;

Software: Shuo Tang, Wang Lv, Chongchong Song, Yuchen Liu;

Resources: Yang Gao, Defeng Wang;

Writing: Shuo Tang, Chongchong Song, Yuchen Liu.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 29, 2022
Revised: March 14, 2023
Accepted: October 16, 2023
Published: October 18, 2023

REFERENCES

1. Dashtbozorg, B., Mendonça, A.M., Penas, S., and Campilho, A. (2014). RetinaCAD, a system for the assessment of retinal vascular changes. In 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE).
2. Lu, S. (2011). Accurate and efficient optic disc detection and segmentation by a circular transformation. IEEE Trans. Med. Imag. *30*, 2126–2133.
3. Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation, pp. 3431–3440.
4. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, and W.M. Wells, et al., eds. (Springer International Publishing), pp. 234–241.
5. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation[C]//Deep Learning. In Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4 (Springer International Publishing), pp. 3–11.
6. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. *40*, 834–848.
7. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, pp. 801–818.
8. Chen, L.C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.05587.
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. (2016). Semantic image segmentation with deep convolutional nets and fully connected CRFs. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.7062.
10. Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., and Cao, X. (2018). Joint Optic Disc and Cup Segmentation Based on Multi-label Deep Network and Polar Transformation. IEEE Trans. Med. Imaging *37*, 1597–1605.
11. Zhu, J.Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks, pp. 2223–2232.
12. Kamnitsas, K., Baumgartner, C., Ledig, C., Virginia, N., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al. (2017). Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. In Information Processing in Medical Imaging, M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.T. Yap, and D. Shen, eds. (Springer International Publishing), pp. 597–609.
13. Zhao, H., Li, H., Maurer-Stroh, S., Guo, Y., Deng, Q., and Cheng, L. (2019). Supervised Segmentation of Un-Annotated Retinal Fundus Images by Synthesis. IEEE Trans. Med. Imaging *38*, 46–56.
14. Liu, P., Kong, B., Li, Z., Zhang, S., and Fang, R. (2019). CFEA: Collaborative feature ensembling adaptation for domain adaptation in unsupervised optic disc and cup segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22 (Springer International Publishing), pp. 521–529.
15. Lei, H., Liu, W., Xie, H., Zhao, B., Yue, G., and Lei, B. (2022). Unsupervised Domain Adaptation Based Image Synthesis and Feature Alignment for Joint Optic Disc and Cup Segmentation. IEEE J. Biomed. Health Inform. *26*, 90–102.
16. Chen, C., Liu, Q., Jin, Y., Dou, Q., and Heng, P.A. (2021). Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24 (Springer International Publishing), pp. 225–235.
17. Zhu, Q., Chen, X., Meng, Q., Song, J., Luo, G., Wang, M., Shi, F., Chen, Z., Xiang, D., Pan, L., et al. (2021). GDCSeg-Net: general optic disc and cup segmentation network for

multi-device fundus images. Biomed. Opt Express *12*, 6529–6544.

18. Orlando, J.I., Fu, H., Barbosa Breda, J., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al. (2020). REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. *59*, 101570.
19. Sivaswamy, J., Krishnadas, S.R., Joshi, G.D., Jain, M., and Tabish, A.U.S. (2014). Drishti-GS: Retinal Image Dataset for Optic Nerve head(ONH) segmentation. In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 53–56.
20. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. 3 (Data, Multidisciplinary Digital Publishing Institute), p. 25.
21. Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., and Ayed, I. B. (2019). Boundary loss for highly unbalanced segmentation. In Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning (PMLR), pp. 285–296.
22. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1512.03385.
23. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., and Liu, J. (2019). CE-Net: Context Encoder Network for 2D Medical Image Segmentation. IEEE Trans. Med. Imaging *38*, 2281–2292.
24. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A., and Zhou, Y. (2021). Transunet: transformers make strong encoders for medical image segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.04306.
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.
26. Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with Relative Position representations. Preprint at arXiv *2*.
27. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., and Yang, B. (2020). MedGAN: Medical image translation using GANs. Comput. Med. Imaging Graph. *79*, 101684.
28. Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410.
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119.
30. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. Adv. Neural Inf. Process. Syst. *33*, 12104–12114.
31. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks[J]. Adv. Neural Inf. Process. Syst. *34*, 852–863.
32. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. (2021). Encoding in Style: A Stylegan Encoder for Image-To-Image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2287–2296.
33. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., and Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. ACM Trans. Graph. *40*, 1–14.
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., Vaswani, A., Shazeer, N., et al. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 1–15.
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. *34*, 12077–12090.
37. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., and Hu, S.M. (2022). SegNeXt: rethinking convolutional attention design for semantic segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2209.08575.
38. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al. (2022). Gamma challenge: glaucoma grading from multi-modality images. Preprint at arXiv *06511*. https://doi.org/10.48550/arXiv.2202.06511.
39. Guo, C., Szemenyei, M., Yi, Y., Wang, W., Chen, B., and Fan, C. (2021). In Sa-unet: spatial attention U-Net for retinal vessel Segmentation. 2020 (IEEE).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
REFUGE	Orlando J I et al.[18]	https://refuge.grand-challenge.org/
Drishti-GS1	Sivaswamy J[19]	http://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/enter.php
GAMMA	Wu J et al.[38]	https://aistudio.baidu.com/aistudio/competition/detail/90
IDRiD	Porwal P[20]	https://idrid.grand-challenge.org/
Software and algorithms		
Cycle-GAN	Zhu J-Y et al.[11]	https://github.com/junyanz/CycleGAN
Resnet −101	He K et al.[22]	https://github.com/zhanghang1989/ResNeSt
SA-Unet	Guo, Changlu et al.[39]	https://github.com/clguo/SA-UNet
CE-Net	Gu Z et al.[23]	https://github.com/Guzaiwang/CE-Net
W-net	This paper	https://github.com/ts66666/OD
Pytorch	Version1.11.0	https://pytorch.org/docs/1.11/
Python	Version 3.8.0	https://www.python.org/downloads/release/python-380/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Defeng Wang (Dfwang@buaa.edu.cn).

Materials availability

This study did not generate any new unique materials.

Data and code availability

- The code used for reproducing our analysis result are made available on GitHub (https://github.com/ts66666/OD).
- The four optic disc datasets (REFUGE, Drishti-GS1, GAMMA, and IDRiD) are all public datasets and can be downloaded from the website. The DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All experiments in this paper were run with python3.8.0 coding on ubuntu 20.04. The framework used for implementation is Pytorch on NVIDIA GTX1080Ti with 11GB memory. Our training process consists of 4 modules, and the training details are shown in below table. It should be noted that in the FSM, we adopted the mechanism of early termination to avoid overfitting. That is, the training will be terminated if the overall loss does not drop in 20 epochs or the learning rate for 10 epochs is less than $5e^{-7}$.

The training settings in our work

Module	Network	Optimizer	Epochs	Learning rate	Loss function
QSEM	Cycle-GAN	Adam betas: (0.5, 0.999)	200	Strategy: poly learning rate Initial: 0.0002 Decay iters: 50	Adversarial loss Cycle loss
CSM	SA-UNet	SGD momentum = 0.9	100	Strategy: plateau Patience: 50 Factor: 0.9 Initial: 0.01	Dice loss

(Continued on next page)

Continued

Module	Network	Optimizer	Epochs	Learning rate	Loss function
OLM	ResNet101	SGD momentum = 0.9	100	Strategy: plateau Patience: 50 Factor: 0.9 Initial: 0.0001	Cross-entropy loss
FSM	W-Net	Adam	1000	Strategy: poly learning rate Initial: 0.0002 Power: 0.9	Boundary loss Dice loss

The training sets used in DQEM are Val and IDRiD. The training set of domain A is IDRiD, and domain B is IDRiD. The input image is the original image cropped out of the retinal foreground region and resampled to 256 × 256. Moreover, the IDRiD dataset is used here as the training set for domain B because the data contrast in this dataset is deficient. We can use any dataset, even a clinically collected dataset, to replace it.

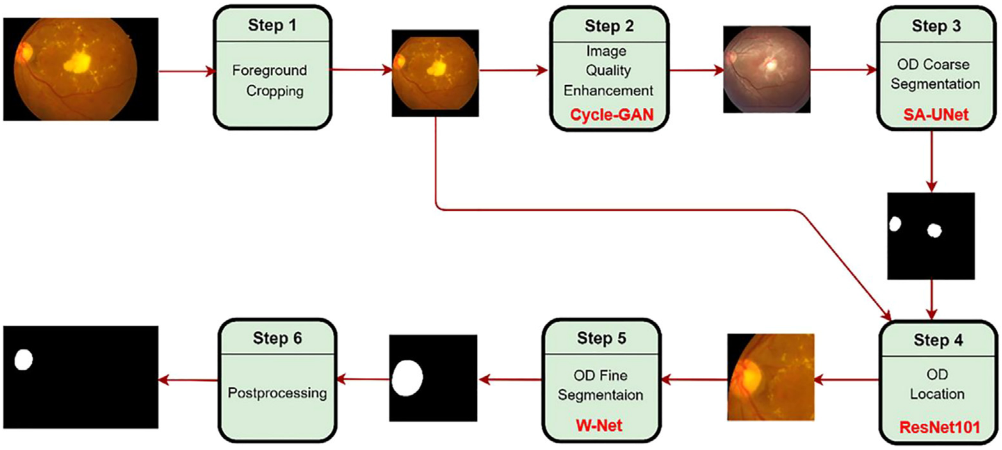
We collect training sets from Train and Val. First, we translate the images into the target domain using the trained model in DQEM. Our training data consit of both translated and augmented images (three times random rotations, three times random color transformation, and three times random noise added to translated images). The input size of CSM is 256 × 256.

The training datasets of OLM are from Train and Val. Firstly, the foreground regions of these two datasets are cropped. We augment these foreground images with the data augmentation of CFM, and we call these images the original images. Then we resize the original image to 128 × 128. Afterward, we crop the images with a kernel (stride 8 and size 32). Next, we perform an inverse transformation with the cropped position information to get the corresponding patches from the original image. Then we resample these patches to 224 × 224. We definite the patch containing the entire OD area as positive examples and the other patches as negative examples. Finally, we randomly remove negative examples so that the number of positive samples is comparable to negative ones.

The training data sources for FSM are Train and Val. We follow the steps in the flowchart in Figure 2 to obtain images that only contain the whole OD and their annotation results. The input of this module is 256 × 256.

METHOD DETAILS

The multi-stage OD segmentation is presented in below figure, which maintains four modules: data quality enhancement module (DQEM), OD coarse segmentation module (CSM), OD localization module (OLM), and fine segmentation module (FSM). In FSM, we propose the W-net structure, as shown in below figure.



Multi-stage OD segmentation in CFP flowchart

In the order of the arrows, the sequence is preprocessing, data quality enhancement, coarse optic disc segmentation, optic disc positioning, optic disc fine segmentation, and postprocessing.

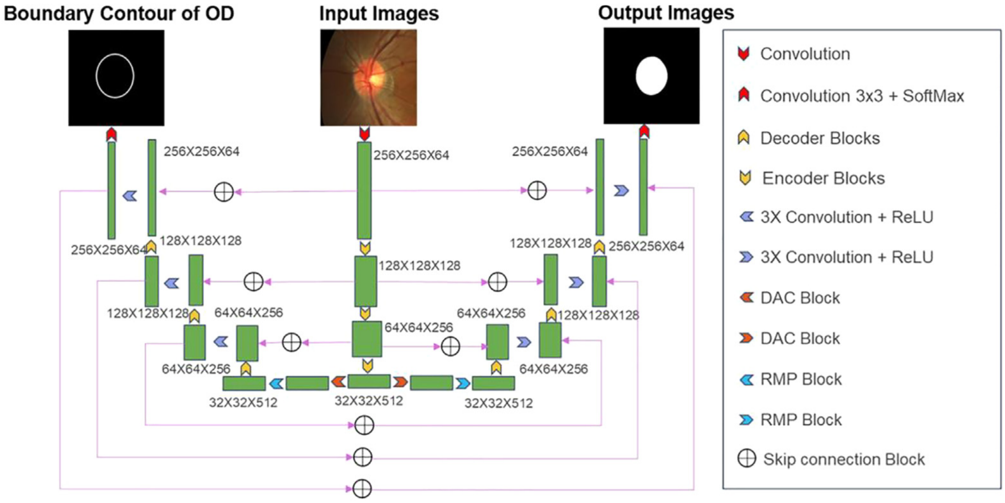


Illustration of the proposed W-Net

The left branch is to learn boundary information of OD. The right branch is to learn region information. The backbone is similar to CE-Net. The last down-sampling layer adopts a dense atrous convolution module (DAC) and residual multi-kernel pooling module (RMP).

For the convenience of description, we denote the image to be segmented as $I \in \mathcal{R}^{h\times w\times 3}$, where h and w represent the height and width of each input scan, respectively.

Data quality enhancement module (DQEM)

To apply OD segmentation to different data domains and improve model inference's robustness, we first use Cycle-GAN[11] for data quality enhancement. The DQEM module transforms various CFPs into similar image domains, thus improving the contrast of the OD region of the CFPs. In this way, we achieve the purpose of improving the robustness of the CSM. In this process, the foreground needs to be cropped initially, and we denote it as $I_{crop} \in \mathcal{R}^{256\times 256\times 3}$. The output of this module is $I_{cycle} \in \mathcal{R}^{256\times 256\times 3}$. In general, the computational procedure of DQEM can be summarized as:

$$I_{cycle} = GA\left(I_{crop}\right) \tag{Equation 4}$$

$$I_{crop} = R\left(I * I_{mask}\right) \tag{Equation 5}$$

Here, $R(\cdot)$ stands for resampling function, which interpolates the input image to 256 × 256×3 by bilinear interpolation. I_{mask} is an image generated by thresholding the grayscale image of image I that only contains the retinal foreground. When the gray level of the gray image of image I is greater than 5, the pixel value of I_{mask} is 1, otherwise it is 0. Then we use the largest connected area as I_{mask}.

It is worth noting that the images generated by Cycle-GAN may change the original features, such as generating structures that are not present in the input.

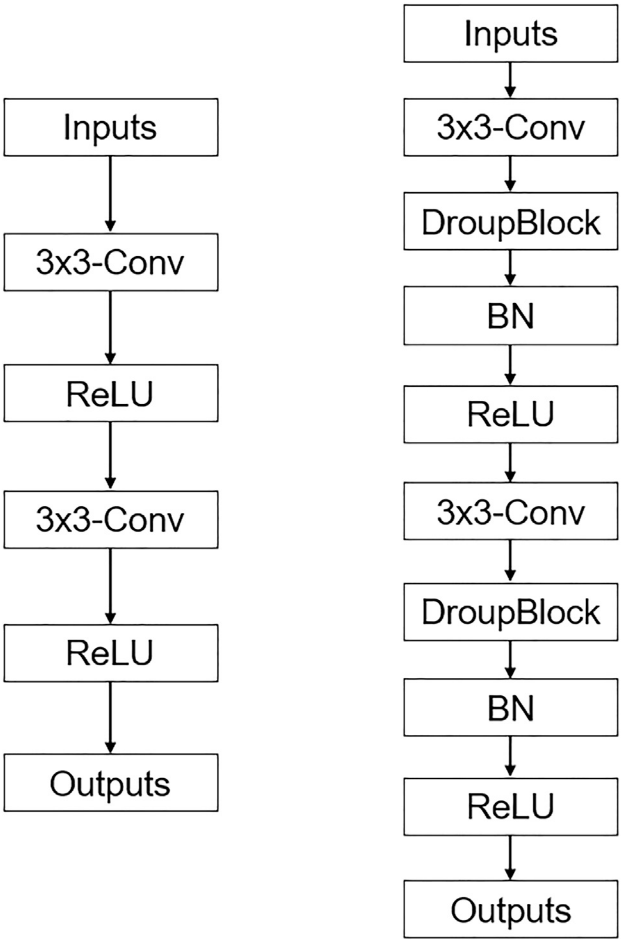
This false information generated by GANs will affect the segmentation accuracy. The reason we can use a GAN is that our next stage of segmentation is a coarse segmentation whose purpose is to localize the OD region rather than segmentation.

Coarse segmentation module (CSM)

CSM segments the image I_{cycle} generated by DQEM. CSM does not need to produce fine segmentation, i.e., we allow the model to recognize non-OD regions as OD regions and tolerate rough boundaries as well. In other words, we do not pursue accuracy, but only the recall of CSM. We use SA-UNet to perform a coarse segmentation of OD. As shown in the below figure, compared with the U-Net block, the SA-UNet block adds DropBlock and batch normalization (BN) layer,[39] which can enhance important features and suppress unimportant features, thereby improving the network's representation ability. In addition, Spatial Attention Module (SAM) is introduced into SA-UNet, aggregates the channel information of a feature map by using the maximum pooling and average pooling, generates two 2D maps F^{s}_{mp}, and F^{s}_{mp}, and then concatenated and convolved by a standard convolution layer, producing our 2D spatial attention map. Spatial attention can enhance important features and suppress unimportant ones, thus improving the representation ability of the network, the final output of SAM is calculated as:

$$\begin{aligned} P_{CS} &= F \cdot M^S(F) \\ &= F \cdot \sigma\left(f^{7\times 7}([MaxPool(F); AvgPool(F)])\right) \\ &= F \cdot \sigma\left(f^{7\times 7}\left(\left[F^S_{mp}; F^S_{ap}\right]\right.\right. \end{aligned}$$

where the input image is $F \in R^{H\times W\times C}$ and the output image is $P_{CS} \in R^{H\times W\times C}$. $f^{7\times 7}(\cdot)$ denotes a convolution operation with a kernel size of 7 and σ (·) represents the Sigmoid function.



Original U-Net block (left), SA-Unet block (right)

We denote the probability map of the SA-UNet outputs as P_{cs}, and the output of this module as OD_{cs}.

$$OD_{cs} = \begin{cases} 1 & P_{cs} \geq T \\ 0 & P_{cs} < T \end{cases} \qquad \text{(Equation 6)}$$

where T is the threshold value. In this paper, $T = 0.3\times\max(P_{cs})$. This way, this stage's output can contain at least one OD region.

OD localization module (OLM)

Due to the setting method of T, there are many false positive regions in OD_{cs}, and the role of OLM is to determine which region contains OD. The positional between the OD_{cs} and image I has a corresponding mapping relationship. Therefore, for each connected region R_i of OD_{cs}, the corresponding region $R0_i$ can be found in image I. According to the actual situation, we crop out $IP_i \in \mathcal{R}^{h_i\times w_i\times 3}$ in the image I as the corresponding region R_i. After that, we use ResNet101[22] to divide IP_i into patches that contain the whole OD and patches that do not contain the

OD, and keep only the IP_i that contains the whole OD, which we denote as IP. The cropping strategy is to crop the $h_i \times w_i$ region on the image I with the center of each connected region as the center. In this paper, $h_i = w_i = 0.4 \times \min(h, w)$, and we can express this process as

$$IP = \sum_i IP_i * \varnothing_i \qquad \text{(Equation 7)}$$

$$\varnothing_i = \begin{cases} 1 & P_{OL} \geq T \\ 0 & P_{OL} < T \end{cases} \qquad \text{(Equation 8)}$$

where i is the index of the connected region. P_{OL} is the output probability of the OD localization model ResNet101. T is the threshold value, $T = \max(P_{OL})$.

Fine segmentation module (FSM)

Figure in method details shows W-net with a W-shaped decoder (left branch) -encoder (middle) -decoder (right branch) structure, which consists of two branches similar to CE-Net[23] and shares the same down sampling feature maps. In the encoder module, we use the first four feature extraction blocks of the pre-trained ResNet-34[22] without the average pooling layer and the fully connected layer. Between the encoder and decoder is the context extractor module, which is composed of DAC block and RMP block. This module is described in detail in DAC module and RMP module. In each step of the decoder includes a 1 × 1 convolution, 3 × 3 transpose convolution and 1 × 1 continuous convolution, the number of feature channels is halved. Then, the corresponding feature map of the encoder is connected by skip connection block, the final output is the same size as the original input mask. Prior knowledge suggests that the boundary of OD is circular or elliptical. Accordingly, we use the boundary loss inspired by[21] to make the model learn the edge information of OD. For the left branch, boundary loss[21] is used as the loss function so that the network can learn more edge information. For the right branch, dice loss is used to get more region information. Combining the boundary features and region features of both left and right branches, the loss functions of the left and right branches are added together as the final loss function, and finally the output image of OD segmentation is obtained.

DAC module and RMP module

The backbone of the network is borrowed from CE-Net.[23] The dense atrous convolution module (DAC) and the residual multi-kernel pooling module (RMP) are used in the last down-sampling layer. As shown in below figure, the DAC module has three branches. The atrous rate fields of the convolution kernel of each branch are 3, 7, 9, and 19. Finally, a 1 × 1 convolution in each branch is used for feature channel sorting, while the other branch does not perform any operation. Using a large atrous rate broadens the atrous rate field of the model without enlarging the computational effort, allowing more comprehensive features to be extracted, while using a smaller atrous rate convolution allows more subtle features. We can express it as,

$$F_{out}[j] = \sum\nolimits_{k=0}^{4} Conv(F_{in}[j]) \qquad \text{(Equation 9)}$$

where for the jth input feature map $F_{in}[j]$, its output $F_{out}[j]$ is the sum of the five branches. $Conv(\bullet)$ is a cascaded atrous convolution operation performed on each branch.

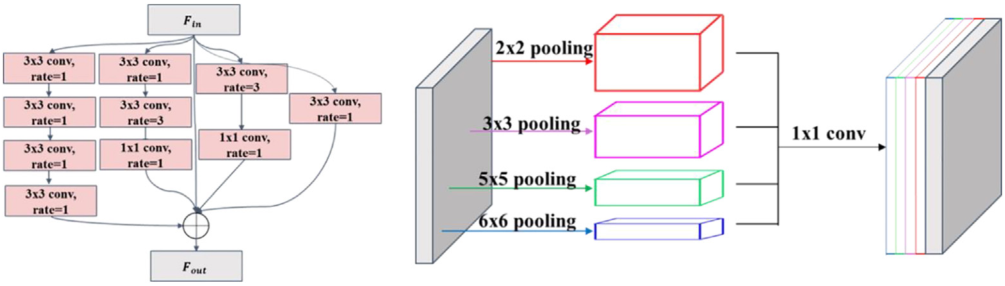


Illustration of DAC and RMP module

On the left is the DAC module, which has 5 branches: 4 atrous convolution branches and 1 feature branch. The DAC module improved the receptive field without increasing the number of parameters. On the right is the RMP module, which has 4 pooling kernels (2 × 2, 3 × 3, 5 × 5, 6 × 6) and a 1 × 1 convolution layer. The RMP module improved the ability to learn features of different scales.

In semantic segmentation, pooling layers can broaden[37] the atrous rate field of convolution kernels, but reduce the accuracy. To solve this problem, we use an RMP module like CE-Net. The RMP module has multiple pooling kernels that can extract different levels of contextual information. And we use 1 × 1 convolution at each layer of the pooling operation to reduce the dimensionality of the feature map to 1/N of the original dimensionality, where N is the number of channels of the original feature map. Then, we up-sample the low-dimensional feature map and obtain features of the same size as the original feature map through bilinear interpolation. Finally, we concatenate the original features with the up-sampled feature map as output. This process can be expressed as follows,

$$F_{pool}[i] = up\left\{conv1\left[Maxpool(F_{in})[i]\right]\right\} \tag{Equation 10}$$

$$F_{out} = F_{pool}[0], F_{pool}[1], \bullet\bullet\bullet, F_{pool}[k] \tag{Equation 11}$$

where $F_{pool}[i]$ represents the output corresponding to the i-pooling kernel. $up(\bullet)$ means to interpolate in a bilinear fashion. $conv1(\bullet)$ is a 1*1 convolution operation. K is the number of pooling kernels, which is taken as k = 4 in this paper. The sizes of the pooling kernels are 2 × 2, 3 × 3, 5 × 5, and 6 × 6, respectively.

The decoding structure uses 1 × 1 convolution-3×3 deconvolution-1×1 convolution. The skip connection part is performed with the residual connection module in CE-Net, i.e., the output feature is the sum of different input features.

QUANTIFICATION AND STATISTICAL ANALYSIS

To accommodate the circular structure of the OD, in the left boundary branch, we add boundary loss[21] to the training process of W-Net. The theory of adding boundary loss has been verified in.[21] The computational procedure can be formulated as follows

$$BL(y, y) = \sum_q \varphi_y(q) S_\theta(q) \tag{Equation 12}$$

$$\varphi_y(q) = \begin{cases} D_y(q) & q \notin y \\ -D_y(q) & q \in y \end{cases} \tag{Equation 13}$$

where $S_\theta(q)$ is the forward output result of the network model y_l at point q, and $D_y(q)$ is the distance between point q and the closest point $Z_{\partial G}(q)$ on the contour ∂y.

Thus, the loss of our left branch can be expressed as,

$$Dice(y, y) = 1 - 2 \times \frac{y \times y}{y + y} \tag{Equation 14}$$

$$L_L(y, y_l) = \beta Dice(y, y_l) + (0.5 - \beta) \times BL(y, y_l) \tag{Equation 15}$$

where y_l is the output of the left branch and y is the label corresponding to the input. β is a hyperparameter that will be discussed in the influence of super parameter section. In the right branch, the loss we use is Dice loss, which is,

$$L_R(y, y_r) = Dice(y, y_r) \tag{Equation 16}$$

where y_r is the output of the left branch and y is the label corresponding to the input. We use the weighted loss as our final loss function:

$$L(y_l, y_r, y) = L_L(y, y_l) + \alpha L_R(y, y_r) \tag{Equation 17}$$

where α is a hyperparameter, which we will explain in detail in the experiments and results section.