ISSN: 1813-4890

A Prototype Network for Few-Shot Retinal Vessel Segmentation Based on Nested Structures

Jiqiang Zhu, Ming Gao

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China

Abstract

Retinal vessel images contain rich structural information about blood vessels, which can assist doctors in diagnosing and screening for diseases such as diabetes, glaucoma, and hypertension. In recent years, deep learning technologies have rapidly developed and are widely applied in the field of medical image processing. Convolutional neural networks demonstrate excellent segmentation performance, providing valuable assistance to clinicians in clinical diagnosis. However, convolutional neural networks require a large number of annotated samples for training, and the collection and annotation of medical image data is a hard task. This difficulty in obtaining a sufficient amount of annotated medical image data limits the application of deep neural networks To address this problem, this paper proposes a few-shot retinal vessel segmentation network based on prototype networks and nested structures. Experimental results on the publicly available DRIVE, STARE, and CHASE_DB1 datasets demonstrate that our prototype network, trained with the Fetoscopy Placenta Data, effectively segments retinal vessel images. In comparison with other advanced models, our approach exhibits advantages in segmentation performance.

Keywords

Few-Shot;Unet++;Prototype Network;Retinal Vessel Segmentation.

1. Introduction

Eye diseases caused by conditions such as glaucoma and diabetic retinopathy are significant contributors to vision problems in the elderly. Many retinal conditions are asymptomatic until they reach an advanced stage, making early detection and diagnosis crucial for both doctors and patients. However, retinal vessel segmentation poses considerable challenges [1]. The complex structure of retinal vessels, along with the low contrast between some small vessels and the background, makes vessel extraction difficult. Additionally, issues such as uneven image contrast and brightness distribution further complicate vessel segmentation, thereby affecting the accuracy of retinal vessel segmentation results [2].

In recent years, segmentation methods based on deep learning have been extensively researched and widely applied in the field of medical image segmentation. Shelhamer et al. [3] proposed an end-to-end fully convolutional network, which employs an encoder-decoder architecture capable of receiving features from input images of arbitrary sizes and inferring results. Badrinarayanan et al. [4] introduced the SegNet model, which utilizes pooling indices computed during the corresponding encoder pooling process to perform non-linear upsampling in the decoder, reducing the number of parameters and computations required for deconvolution. Ronneberger et al. [5] presented the Unet model, widely used in medical image segmentation, which improves upon the FCN model by using skip connections to fuse spatial and semantic information at different depths between the encoder and decoder. This model demonstrates excellent performance in the field of medical image segmentation. Building upon

ISSN: 1813-4890

this, Zhou et al. [6] proposed the Unet++, which nests different depths of Unet models. This network is a nested network with deep supervision capability and multiple output ports, enabling pruning operations on subnetworks of different depths. The denser skip connections further narrow down the semantic gap between the encoder and decoder.

Due to reasons such as patient privacy and hospital policies, medical image samples are often limited, making it challenging to meet the training requirements of deep neural networks. Few-shot learning has emerged as a new research topic in recent years, aiming to achieve good performance with limited supervision in scenarios with a small number of samples. Among them, the idea of prototype learning focuses on training the network to extract prototypes rather than directly fitting the current task, effectively improving the generalization of small-sample segmentation models. Current few-shot learning models include the Attention-based Multi-Context Guiding network (A-MCG) [7], Prototype Alignment Network (PANet) [8], Self-Regularized Prototypical Network (SRP) [9], and others. While these models excel in segmentation tasks where pixels are densely clustered, they face challenges in effectively recovering fine vessel details lost during the encoding process in retinal vessel images. As a result, their performance in segmenting vessel-like targets is limited.

To better segment structurally complex foreground objects, this paper proposes a nested multi-prototype segmentation network for segmenting retinal vessel images. The network's feature extraction layer incorporates multiple attention mechanisms to better preserve the feature information of fine structures. Through the nested structure of Unet++, multiple prototype vectors are extracted at different stages and fused with weights. Finally, the segmentation head combines two metric learning methods to output the segmentation results. Experimental results demonstrate that our model reduces the difficulty of network training, effectively improves the recall rate of segmentation results, preserves more fine vessels, and assists doctors in diagnosis more effectively..

2. Theoretical Framework

2.1. Network Framework

The idea of prototype networks is to extract prototype vectors for corresponding foreground objects from limited samples and then calculate the distance between the feature maps of the query set and the prototype vectors using metric learning to segment the pixels at the current spatial position. This process can be divided into two steps: In the first step, prototype vectors p_s are extracted using masked average pooling(MAP), then the distance between p_s and the query set is calculated to obtain the predicted probability map of the query set, and the loss L_q of query set is computed. In the second step, the query set probability map is used as a mask to extract the prototype vectors p_q of the query set. Then, the distance between p_q and the feature maps of the support set is calculated to obtain the predicted probability map of the support set, and the loss L_s of support set is computed. Finally, the network is trained using L_s and L_q. The proposed network architecture is illustrated in Figure 1.

Early-stage features contain more detailed information, which can effectively complement vessel feature details. The prototype network proposed in this paper extracts multiple prototype vectors at different depths through the feature extraction layer and integrates them through the multi-prototype fusion block (MFB) to further enhance the expression ability of the prototype vectors. Then, different metric distances are fused and normalized to output the predicted probability map through fusion normalization block (FNB).

ISSN: 1813-4890

2.2. Feature Extraction Layer

In encoder-decoder structured network models, the early encoding stage's feature maps have higher resolutions, enabling the preservation of more detailed information. In this paper, we introduce the Unet++ model to extract image features. The nested structure of Unet++ effectively restores the resolution of shallow-level feature maps and output multiple feature maps with the same mask size.. The structure of the feature extraction layer is illustrated in Figure 2.

The retinal blood vessels in fundus images exhibit complex branching structures and are characterized by numerous small vessels. These small vessels have relatively small diameters, making the segmentation of retinal blood vessels more dependent on high-resolution spatial details compared to segmentation tasks involving other organs. This is because this part of the information is more complete at the spatial scale and does not suffer from the loss of detailed information caused by multiple pooling operations. The nested structure of Unet++ allows for more comprehensive utilization of this information, thereby enhancing the network's segmentation capability for small blood vessels.

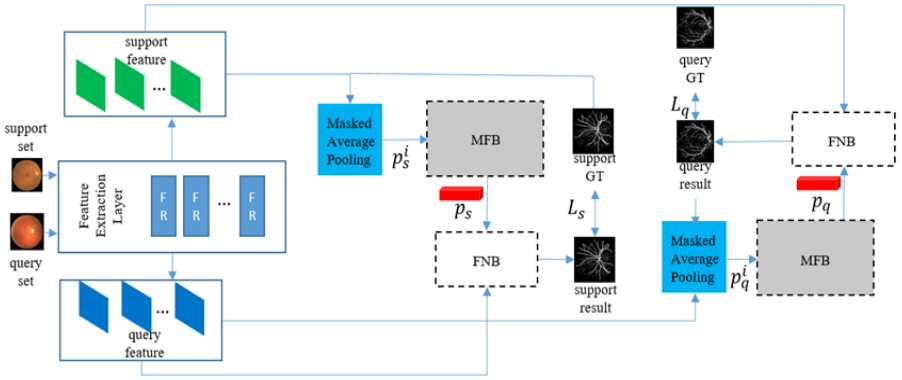

Fig1. Network Framework

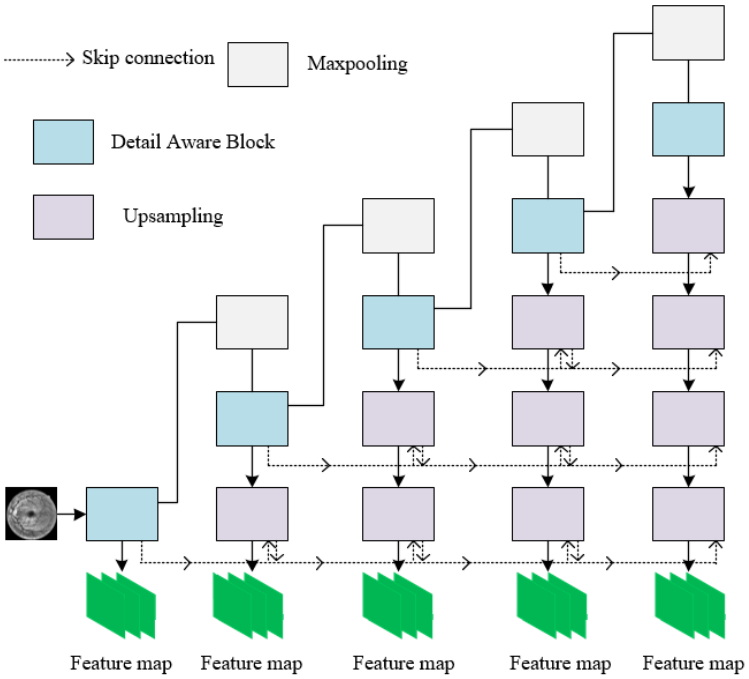

Fig2. Feature Extraction Layer

However, the double convolutional modules in Unet++ cannot effectively extract detailed features of vascular structures. In this paper, we introduce the Detail Aware Block (DAB) to optimize the encoding process. The structure of DAB is illustrated in Figure 3.

ISSN: 1813-4890

Firstly, we expand the number of channels of the input feature map using 1×1 convolutions, and then preliminary feature extraction is performed using one 3×3 convolution followed by another 1×1 convolution. Subsequently, channel attention is applied to enhance the weights of important channels while suppressing irrelevant information [10]. The channel attention module can be represented by Formula 1.

$$F_c = M_c(F) = F * \sigma(MLP(F_{Avg}^c) + MLP(F_{Max}^c)) \quad (1)$$

Where F represents the input feature map, *denotes element-wise multiplication, σ stands for the Sigmoid function, *MLP* refers to a multi-layer perceptron with shared weights, and F_{Avg}^c and F_{Max}^c indicate the feature map outputs after spatial dimension-wise average pooling and max pooling, respectively.

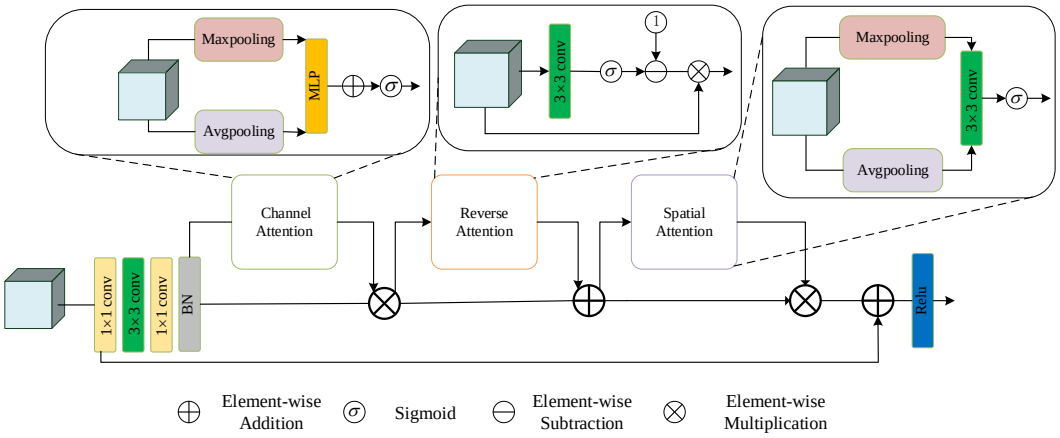

Fig3. DAB

Through the reverse attention module, significant vessel information in the input feature map F_c can be explicitly erased to further enhance the network's focus on subtle vascular details, followed by element-wise addition for information supplementation. The reverse attention module can be represented by the following equation:

$$F_r = M_r(F_c) = F_c + F_c * (1 - \sigma(f^{3\times3}(F_c))) \quad (1)$$

Where $f^{3\times3}$ represents 3×3 convolution.

Subsequently, the spatial attention module is employed to further enhance the feature representation capability along the channel dimension and suppress irrelevant information. The spatial attention module can be represented by Equation 3:

$$F_s = M_s(F_r) = \sigma(f^{3\times3}([F_{Avg}^s; F_{Max}^s])) \quad (3)$$

Where F_{Avg}^s and F_{Max}^s represent the feature map outputs of F_r after average pooling and max pooling along the channel dimension, respectively, followed by an activation through a ReLU.

2.3. Multi-prototype Fusion Block（MFB）

To effectively represent the feature information of small blood vessels, a single prototype may not suffice. In order to further enhance the expression capability of prototype vectors, this paper proposes a multi-prototype fusion module (MFB). The structure of MFB is illustrated in Figure 4.

ISSN: 1813-4890

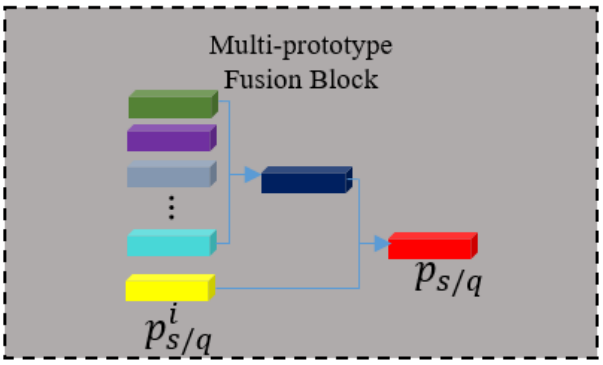

Fig4. MFB

The process of extracting prototype vectors $\{p^1_{s/q}, p^2_{s/q},..., p^i_{s/q},..., p^n_{s/q}\}$ from multiple feature maps outputted by the feature extraction layer can be represented by the equation:

$$p^i_{s/q} = \frac{1}{K}\sum_{j=1}^{K}\frac{\sum_{a,b} F^{(i,j,a,b)} y^{(i,j,a,b)}}{\sum_{a,b} y^{(i,j,a,b)}} \tag{4}$$

Where (i,j,a,b) denotes the index, $F^{(i,j,a,b)}$ FF represents the feature map outputted by the feature extraction layer, and $y^{(i,j,a,b)}$ represents the ground truth. Due to the early-stage features containing more detailed information, as well as possibly more background noise, while the deeper features contain richer semantic information, the integration of prototype vectors from different depths is performed through weighting to enhance the expression capability of prototype vectors. The multi-prototype fusion module can be represented by Equation 5:

$$p_{s/q} = \frac{\frac{\sum_{i=1}^{n-1} p^i_{s/q}}{n-1} + p^n_{s/q}}{2} \tag{5}$$

To better segment foreground target pixels, this paper measures the feature vectors at each spatial position using cosine similarity and distance fidelity. The cosine similarity ranges from -1 to 1 and is unaffected by the magnitudes of the elements in the feature vectors, making it effective in distinguishing foreground pixels from background pixels. The cosine similarity can be represented by Equation 6:

$$D_c(F^{(a,b)}_{s/q}, p_{s/q}) = \frac{F^{(a,b)}_{s/q} \cdot p_{s/q}}{|F^{(a,b)}_{s/q}||p_{s/q}|} \tag{6}$$

Distance fidelity represents the distance between the density matrices of features and prototype vectors, where the density matrix is defined as the product of a vector and its conjugate transpose. Since the distribution range of distance fidelity is greater than 0, it complements the negative part of cosine similarity, thereby aiding in further integrating critical pixel information near the threshold. Distance fidelity can be represented by Equation 7:

$$D_N(F^{(a,b)}_{s/q}, p_{s/q}) = F^{(a,b)}_{s/q} \cdot (p^T_{s/q} \cdot p_{s/q}) \cdot (F^{(a,b)}_{s/q})^T \tag{7}$$

The structure of the fusion normalization block is depicted in Figure 5. Since deep feature maps contain more semantic information, the last layer's output feature map is utilized for measurement.

ISSN: 1813-4890

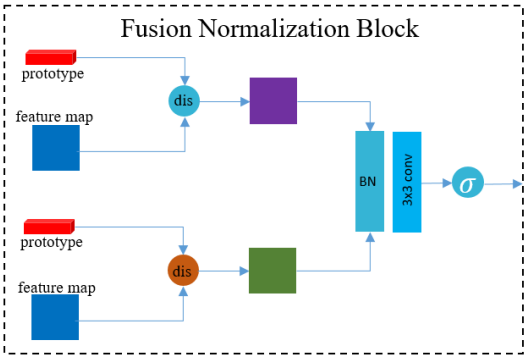

Fig5. FNB

FNB initially computes the cosine similarity and non-parametric distance fidelity between the feature map and prototype vectors. Subsequently, the resulting distance map is concatenated and subjected to batch normalization (BN) before being adjusted through a convolutional layer. Finally, the prediction probability map is generated by applying the Sigmoid function.

2.4. Loss Function

This paper employs binary cross-entropy (BCE) as the loss function. The BCE loss LL for the support set/query set can be represented by Equation 8:

$$L_{s/q} = \frac{1}{N}\sum_{1}^{N}[y_{s/q}^{i}\log(\hat{y}_{s/q}^{i}) + (1-y_{s/q}^{i})\log(1-\hat{y}_{s/q}^{i})] \quad (8)$$

3. Experiments

3.1. Dataset and Preprocessing

In order to comprehensively evaluate the generalization performance of the model, this study selects 84 fetal vascular images from the fetal placenta dataset [11] for model training. For the trained model, no adjustments are made, and it is directly utilized for segmentation tasks on the DRIVE [12], STARE [13], and CHASE_DB1 [14] datasets of retinal fundus vascular images. The image samples from the four datasets are depicted in Figure 6.

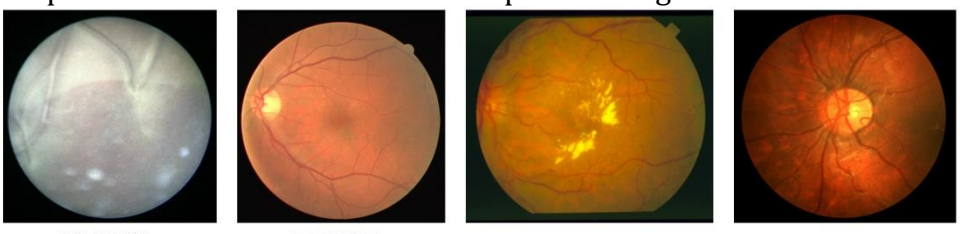

Fig6. Datasets

The size of each group of images in fetal placenta dataset varies. Firstly, the images are preprocessed by grayscale conversion and contrast-limited adaptive histogram equalization (CLAHE). Subsequently, the image size is reduced to 128*128 through interpolation. Every 4 fetal placenta images are concatenated to form large image blocks of size 256*256, resulting in a total of 21 concatenated image blocks. Among these, 5 blocks are designated as the support set, while the remaining 16 blocks serve as the query set, with no overlap between the support and query sets. The preprocessing process is illustrated in Figure 7.

ISSN: 1813-4890

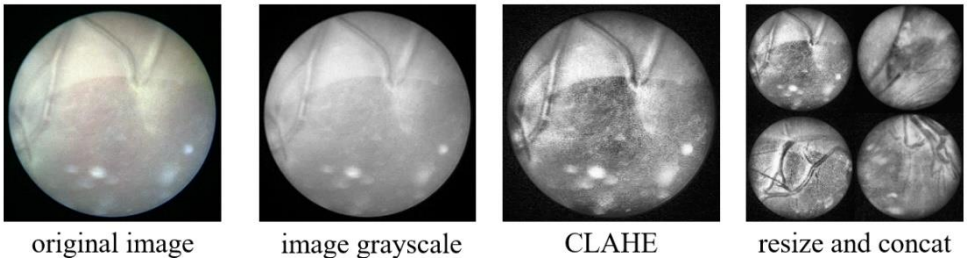

Fig7. Processing of Train Set

The DRIVE dataset comprises 40 color retinal fundus images with original dimensions of 584*565 pixels. Preprocessing involves grayscale conversion and contrast-limited adaptive histogram equalization (CLAHE). Among these, 5 images are designated as the support set, while the remaining 35 images serve as the query set, with no overlap between the support and query sets.

The STARE dataset consists of 20 color retinal fundus images with original dimensions of 700*605 pixels. Preprocessing follows the same method as DRIVE. Among these, 5 images are designated as the support set, while the remaining 15 images serve as the query set, with no overlap between the support and query sets.

The CHASE_DB1 dataset comprises 28 pediatric retinal fundus images with original dimensions of 996*960 pixels. Preprocessing also follows the aforementioned method. Among these, 5 images are designated as the support set, while the remaining 23 images serve as the query set, with no overlap between the support and query sets. The preprocessing process is illustrated in Figure 8.

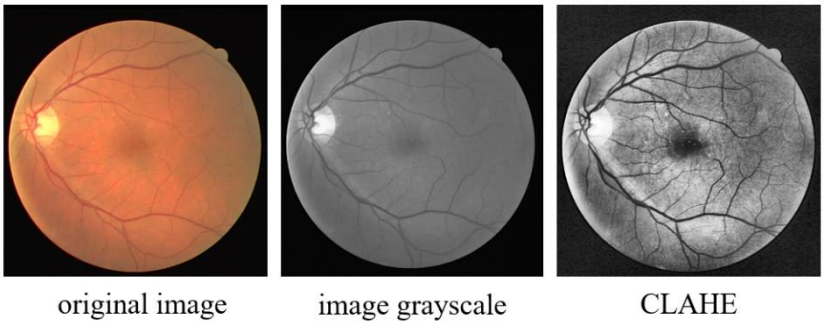

Fig8. Processing of Test Set

3.2. Experimental Results

The experiments of this study were conducted on an RTX 3090 graphics processing unit (GPU) with 24GB of memory. PyTorch framework was employed for implementation. Adam optimizer was utilized with an initial learning rate of 0.0006 and β1 and β2 values set to 0.5 and 0.9, respectively. The total number of training epochs was set to 100. To evaluate the generalization performance of the model, training was performed using the fetal placenta dataset. Subsequently, without fine-tuning, the trained model was directly applied to the segmentation tasks of three retinal fundus vascular datasets. To demonstrate the superiority of our model, comparative experiments were conducted with several existing small-sample segmentation models. The experimental results are presented in Table 1-3:

Table 1 Experimental Results on DRIVE

Method	SE	SP	ACC	AUC
PANet[8]	0.6626	0.8487	0.8323	0.8416
PU-Net[15]	0.6604	0.9600	0.9339	0.9114
SRPNet[9]	0.7206	0.9554	0.9347	0.9428
Ours	0.8517	0.9446	0.9363	0.9609

ISSN: 1813-4890

Table 2 Experimental Results on STARE

Method	SE	SP	ACC	AUC
PANet[8]	0.6142	0.9097	0.8855	0.8622
PU-Net[15]	0.5363	0.9776	0.9411	0.8923
SRPNet[9]	0.6883	0.9070	0.8887	0.9155
Ours	0.7826	0.9479	0.9340	0.9345

Table 3 Experimental Results on CHASE_DB1

Method	SE	SP	ACC	AUC
PANet[8]	0.5328	0.8907	0.8610	0.7923
PU-Net[15]	0.6378	0.9653	0.9382	0.8895
SRPNet[9]	0.7032	0.9410	0.9215	0.9220
Ours	0.7271	0.9703	0.9502	0.9220

From Table 1-3, it can be observed that our model exhibits significant advantages across multiple metrics, particularly in terms of SE (Sensitivity). Specifically, on the DRIVE, STARE, and CHASE_DB1 datasets, our model achieved SE values of 0.8517, 0.7826, and 0.7271, respectively. This indicates that our model is capable of correctly segmenting more foreground vessel pixels. Additionally, the ACC (Accuracy) values on the three datasets were 0.9363, 0.9340, and 0.9502, respectively, suggesting that our model's segmentation results also possess high accuracy.

In summary, the prototype network proposed in this paper demonstrates good generalization and segmentation performance.

3.3. Visualization

To further demonstrate the superiority of our model, Figure 9 illustrates the segmentation results of different models. From Figure 9, it can be observed that compared to other models, the segmentation results of our method are clearer, with fewer vessel fragments and better vessel connectivity. Our method is able to retain more fine vessel pixels, demonstrating a clear advantage over the other models.

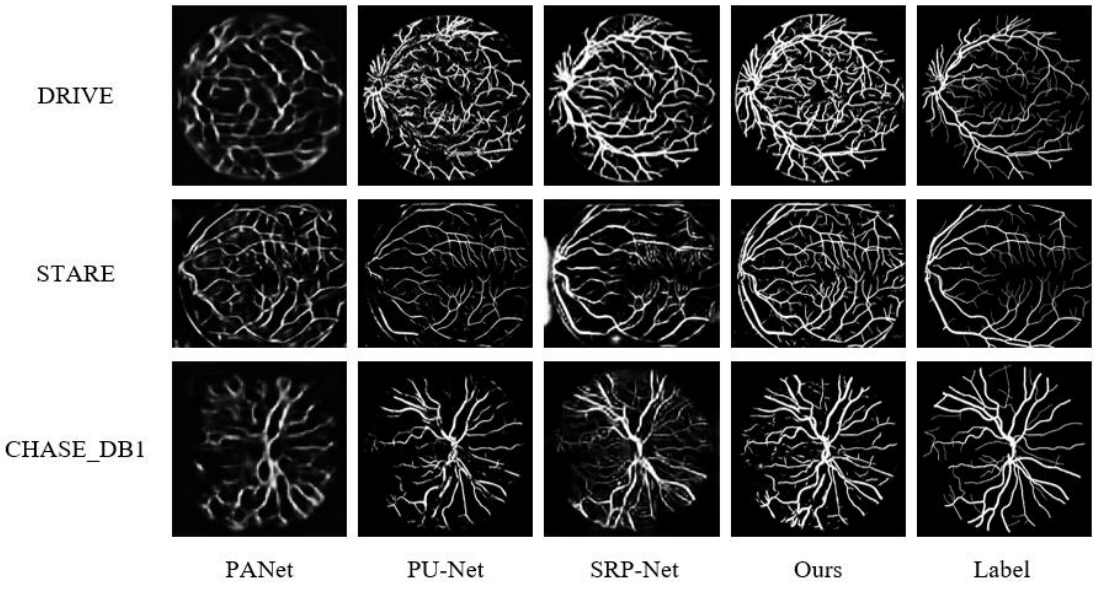

Fig9. Experimental Results of Different Methods

In conclusion, through quantitative analysis and qualitative evaluation, the prototype network proposed in this paper demonstrates effective training of prototype extraction capability through the fetal placenta image segmentation task and achieves segmentation of retinal vessel images. The model exhibits good generalization and effectively improves the recall and

precision of foreground vessel pixels. Furthermore, it outperforms other few-shot models in comparison.

4. Conclusion

This paper proposes a prototype network for retinal fundus vascular image segmentation, trained on the fetal placenta dataset and applied to segment the DRIVE, STARE, and CHASE_DB1 datasets, effectively alleviating the generalization issues caused by limited samples. Comparative experiments and visualization results demonstrate that the proposed prototype network can effectively improve the recall of foreground vessel pixels while maintaining accuracy, indicating good generalization performance. The main contributions of this paper are summarized as follows:

1.The encoding process of the Unet++ nested structure has been optimized by introducing DAB instead of double convolution layers. Furthermore, the network's ability to extract vessel features has been enhanced by concatenating multiple attention modules.

2. A prototype network for small-sample retinal fundus vascular image segmentation is proposed in this paper. The network leverages the nested structure of Unet++ to extract multiple prototype vectors and integrates them using MFB. Additionally, FNB is designed by fusing cosine similarity and fidelity to output segmentation probability maps with high accuracy.

Experimental results demonstrate that the proposed prototype network effectively mitigates the generalization issues caused by insufficient sample size and outperforms other few-shot models in comparative experiments.

References

[1] Z. Zhang, R. Srivastava, H. Liu, X. Chen, L. Duan, D.W. Kee Wong, C.K. Kwoh, T.Y. Wong, J. Liu,: A survey on computer aided diagnosis for ocular diseases, BMC Medical Informatics and Decision Making, Vol. 14 (2014) No.1, p.80.

[2] J. Li, G. Gao, L. Yang, Y. Liu: GDF-Net: A multi-task symmetrical network for retinal vessel segmentation, Biomedical Signal Processing and Control, Vol. 81 (2023), p. 104426.

[3] E. Shelhamer, J. Long, T. Darrell: Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39 (2017) No.4, p. 640—651.

[4] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39 (2017) No.12, p. 2481–2495.

[5] O. Ronneberger, P. Fischer, T. Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation, Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention(Munich, Germany, October 5-9, 2015). Vol.9351, p. 234–241.

[6] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang: UNet++: A Nested U-Net Architecture for Medical Image Segmentation, Proc. International Conference on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support(Granada, Spain, September 20, 2018), Vol.11045, p. 3–11.

[7] HU T，YANG P，ZHANG C，et al.: Attention-based multi-context guiding for few-shot semantic segmentation. Proc. the 33rd AAAI Conference on Artificial Intelligence (Palo Alto, 2019). Vol.33, p. 8441-8448.

[8] WANG K，LIEW J H，ZOU Y，et al.: PANet：few-shot image semantic segmentation with prototype alignment. Proc. the 2019 IEEE/CVF International Conference on Computer Vision(Seoul, Korea, October 27-November 2, 2019).Vol.2019, p. 9196-9205.

[9] Ding H, Zhang H, Jiang X.: Self-regularized prototypical network for few-shot semantic segmentation. Pattern Recognition, Vol.133, (2023), p. 109018.

[10] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon: CBAM: Convolutional Block Attention Module. Proc. European Conference on Computer Vision (Munich, Germany, September 8-14, 2018), Vol. 11211, p. 3–19.

[11] Bano S, Vasconcelos F, Shepherd L M, et al.: Deep placental vessel segmentation for fetoscopic mosaicking. Proc. Medical Image Computing and Computer Assisted Intervention(Lima, Peru, October 4–8, 2020), Vol.12263, p. 763-773.

[12] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman: An Ensemble Classification-Based Approach Applied to Retinal Blood Vessel Segmentation. IEEE Transactions on Biomedical Engineering, Vol.59 (2012) No.9, p. 2538–2548.

[13] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken: Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging, Vol.23 (2004) No.4, p. 501–509.

[14] J.V.B. Soares, J.J.G. Leandro, R.M. Cesar, H.F. Jelinek, M.J. Cree: Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. IEEE Transactions on Medical Imaging, Vol.25 (2006) No.9, p. 1214–1222.

[15] DONG Yang, PAN Haiwei, CUI Qianna, BIAN Xiaofei, TENG Teng, WANG Bangju. Few-shot segmentation method for multi-modal magnetic resonance images of brain tumor. Journal of Computer Applications, Vol.41 (2021) No.4, p. 1049-1054.