SSL-Unet: A Self-Supervised Learning Strategy Base on U-Net for Retinal Vessel Segmentation

Chao Ma, Jia Lu, Xinshuai Teng, Ruoyu Wu

Abstract—Retinal vessel segmentation is a crucial non-destructive detection technology for fundus disease, and supervised methods are commonly applied in it and achieve a good performance. However, supervised methods wholly rely on expert-provided labels as the only learning pattern in the training process. This over-reliance on direct semantic supervision has several issues: (1) Structural information of whole raw data is much richer than that of labeled data. (2) Supervised learning can only deal with a single segmentation task on the guidance of the given labels. Nevertheless, there is more helpful information within unlabeled data such as spatial structure, association properties. Actually, they are an instructive and significant supplement to supervised learning. (3) Labeling low-contrast and complex structured vascular images by clinical experts is an onerous task, which would result in a smaller dataset and expensive cost. So a self-supervised learning U-Net model (SSL-Unet) is presented in this paper, which requires just a few numbers of labeled data and some unlabeled data. A self-supervised module is introduced to help the model to study inter-relation between data according to a low-entropy priori task, thus, both labeled data and unlabeled data can be effectively exploited. Accordingly, the potential features of data can be learned by self-supervised module. Two different self-supervised learning strategies and a pixel-level optimization function are designed for two datasets with different structures. Experimental results on two public retinal datasets of the DRIVE dataset and the CHASEDB1 dataset demonstrate that the proposed model further improves the performance of vessel segmentation, and it can be generalized to different backbone networks.

Index Terms—Retinal Vessels; Image Segmentation; U-Net; Self-supervised Learning

I. INTRODUCTION

Retinal vascular abnormalities are closely associated with common diseases such as diabetes, cataracts, and atherosclerosis[1]. Vessel segmentation is an essential basis for computer-aided diagnosis of retinal diseases[2], and the efficient and accurate segmentation of retinal vessels has become an urgent requirement for clinical diagnosis.

In recent years, many supervised learning methods have shown outstanding performance on retinal vessel segmentation[3]-[6]. However, a prerequisite for the success of supervised methods is the availability of a large amount of annotated data. In reality, the task of labeling medical images is high-cost and laborious, especially for retinal vascular images[7]-[10]. Chen et al. [11] proposed a semi-supervised method combined with a generative adversarial network (GAN) to correct pseudo labels by discriminators, which can achieve decent segmentation results with a small amount of labeled data. Xu et al. [12] proposed a partially supervised framework with an active learning strategy to label the patches with the most informative, reducing the dependence on labeled data. Unfortunately, while supervised learning can leverage unlabeled data through pseudo labels, the inner-relation between vessel images is not considered. Ma et al. [13] proposed a self-supervised approach to learn the semantic similarity of vascular images by training attention-guided generators and dividers. In this paper, self-supervised learning (SSL) is used to mitigate this problem. SSL can learn the data representation information without annotations through pretext tasks[14]-[19]. We find that this information is positive for supervised methods, which can enhance the model’s understanding of the inherent correlations in the unlabeled data and compensate for the shortcomings of supervised methods that rely excessively on ground truths during segmentation, as shown in Fig.1.

Based on the above analysis, the main innovations of this paper are as follows:

1) A segmentation model incorporating the self-supervised module (SSL-Unet) is proposed, allowing both labeled and unlabeled images to be fed into the network.
2) Two self-supervised training strategies are proposed for two datasets, and they effectively exploit the unlabeled data for the sake of improving the segmentation performance of the model.
3) A regularized reconstruction loss is introduced to minimize the pixel difference between unlabeled data.
II. Method

A. Model Architecture

A SSL-Unet model consisting of two baseline networks is proposed, and unlabeled and labeled images are considered as input data. It is worth noting that the two networks are the identical structure of encoder-decoder, as shown in Fig. 1. Specifically, the labeled data are fed into baseline(a), and the loss of supervision is calculated by the vessel probability map obtained from baseline(a) and the ground truths. The unlabeled data are taken as the input to the self-supervised module, which contains the baseline(a) and the baseline(b). By constructing a pretext task to train the self-supervised module, a binary probability map comes into being by baseline(a), and the pseudo labels are obtained by baseline(b). The pseudo labels are used to supervise the SSL module to learn, accordingly, the reconstruction loss can be got. Finally, both reconstruction loss and supervised loss are utilized to optimize the model.

B. SSL strategy based on the DRIVE datasets

Generally, SSL needs to design specific prerequisite tasks such as cropping[20], scaling[21], and jigsaw puzzles[22], which can enable the network to be fully trained on the different datasets. The above transformations will not change the semantic information of the data but the spatial information. Inspired by [23], a self-supervised strategy is proposed to be applied to the DRIVE dataset. Given an unlabeled data $X$, the geometric transformation operation $\Theta$ can transform $X$ into $X_1$ and $X_2$, where slicing and rotation are used in a sequential order. Here, rectangles of the same colors in $X_{rec1}$ and $X_{rec2}$ represent the same local area. By predicting the images patches, the baselines can capture local information of different views, and the consistency of the local feature will be maximized for pixel-wise comparison to train the network, as shown in Fig. 2.

C. SSL strategy based on the CHASEDB1 dataset

The structure of the human organ is basically symmetrical, such as lungs, kidneys, eyes, and arms. We observe that the distribution of vascular trunks in the left and right eyes of the same patient happens to be almost concurrent in the CHASEDB1 dataset, as shown in Fig. 3. Therefore, a decoupled-rebuild strategy is introduced to aim at modeling the relationship between the local and global representations of blood vessels and focuses on the prediction of the encoder-decoder baselines at the pixel level. Suppose $x_L$ is data of the left eyes, $x_R$ is data of the right eyes, $f_{enc}$ is the encoder and $f_{dec}$ is the decoder. $x_L$ and $x_R$ are respectively decoupled as inputs to the encoder $f_{enc}$:

$$f_{enc}(x_L) = p(L) + n(L)$$  \hspace{1cm} (1)

$$f_{enc}(x_R) = p(R) + n(R)$$  \hspace{1cm} (2)

where $p(\cdot)$ represents the trunks of the vessel by attention branch, $n(\cdot)$ represents the tiny vessel information. As can be seen from Fig. 3, the vascular trunks of the left and right eyes of the same patient are approximately similar, assuming that $p(L) \approx p(R)$. The vascular trunks information of the left eye and the right one are cross-fused to obtain new features as follows

$$f'_{enc}(x_L) = p(R) + n(L)$$  \hspace{1cm} (3)

$$f'_{enc}(x_R) = p(L) + n(R)$$  \hspace{1cm} (4)
Concatenation

where \( \mu \) is the momentum coefficient, \( w_t \) is the weights before the baseline(b) updating, and \( w_{t+1} \) is the weights after updating. The larger \( \mu \) is, the better the baseline(b) improves its learning ability, so \( \mu \) is set to 0.999 as recommended by \cite{24}, and is updated every \( T \) epochs, \( T \) is set to 2.

**E. Loss function**

The loss function of the SSL-Net model is comprised of two components: loss of supervised module and loss of self-supervised module. For the segmentation task under supervision, the cross-entropy loss function is defined as

\[
L_{bce} = -\frac{1}{C} \sum_{n=1}^{C} [g_{n,i} \log p_{n,i} + (1 - g_{n,i}) \log(1 - p_{n,i})]
\]

where \( C \) represents the number of categories, \( p_{n,i} \) is the predicted probability value when pixel \( i \) belongs to the \( n \)th category, \( g_{n,i} \) is the true label corresponding to pixel \( i \). Since the self-supervised module has no labeled data involved in the loss calculation, two sets of losses are designed respectively: contrast loss and regularized reconstruction loss, and both of them are denoted as

\[
L_{con} = (p(L) - p(R))^2 + (n(L) - n(R))^2
\]

\[
L_{rec} = \frac{\sum ||G(F_\theta(z_R), G(F_\theta(z_L)))||^2 + \varepsilon}{N + \varepsilon}
\]

where \( F_\theta \) is a predictor which has the same AdaptiveAvgPool as the projector. \( G \) is \( L_2 \) regularization along the channel axis and \( N \) is the number of input data. In view of the difference in magnitude between the optimization functions, the Laplace smoothing factor is added, which is able to speed up the convergence. Here, \( \varepsilon \) is set to 10 in the following experiments. Therefore, the total loss function is defined as

\[
L_{total} = \alpha_1 L_{bce} + \alpha_2 L_{con} + \alpha_3 L_{rec}
\]

where \( \alpha \) is the weighting parameter. \( \alpha_1=0.8, \alpha_2=0.1, \alpha_3=0.1 \). Note that \( \alpha_1=0.8,\alpha_2=0, \alpha_3=0.2 \) on the DRIVE dataset.

**III. Experiments**

**A. Datasets and Evaluation Metrics**

The experiments on two public datasets\cite{25}-\cite{26} are conducted. Unlabeled data are obtained after the labels of labeled data are removed. The detailed information is shown in Table I. ACC, SP, F1 and AUC are utilized to evaluate the performance of the model.
TABLE II

| Method (Year)      | DRIVE       |
|-------------------|-------------|
|                  | F1          | SP          | ACC         | AUC         |
| U-Net [17] (2015) | 0.8106      | 0.9776      | 0.9502      | 0.9726      |
| DenseNet [18] (2016) | 0.8159      | 0.9797      | 0.9532      | 0.9759      |
| LadderNet [19] (2017) | **0.8279**  | 0.9791      | 0.9553      | **0.9811**  |
| AtteNet [20] (2018) | 0.8162      | 0.9816      | 0.9567      | 0.9772      |
| CE-Net [22] (2019) | 0.9747      | 0.9545      | 0.9779      |             |
| Wang et al. [23] (2021) | –           | 0.9811      | 0.9565      | 0.9793      |
| PSP-U-Net [24] (2021) | –           | 0.9810      | 0.9556      | 0.9780      |
| Backbone          | 0.8233      | 0.9808      | 0.9567      | 0.9799      |
| SSL-Net (Ours)    | **0.8258**  | **0.9830**  | **0.9575**  | **0.9811**  |

TABLE III

| Method (Year)      | CHASEDB1    |
|-------------------|-------------|
|                  | F1          | SP          | ACC         | AUC         |
| U-Net [17] (2015) | 0.7783      | 0.9701      | 0.9578      | 0.9772      |
| DenseNet [18] (2016) | 0.8012      | 0.9850      | 0.9591      | 0.9764      |
| LadderNet [19] (2017) | 0.8012      | 0.9850      | 0.9591      | 0.9764      |
| AtteNet [20] (2018) | 0.8151      | 0.9869      | 0.9620      | 0.9828      |
| CE-Net [22] (2019) | –           | 0.9787      | 0.9639      | 0.9832      |
| Wang et al. [23] (2021) | –           | 0.9717      | 0.9574      | 0.9750      |
| PSP-U-Net [24] (2021) | –           | 0.9727      | 0.9590      | 0.9784      |
| Backbone          | 0.8258      | 0.9817      | 0.9575      | 0.9811      |
| SSL-Net (Ours)    | **0.8316**  | **0.9914**  | **0.9661**  | **0.9886**  |

B. Implementation Details for Training

The proposed model is experimented using the Pytorch framework. The batch size is set to 8, the Adam optimizer with default parameters is applied, combined with the cosine annealing method, the initial learning rate is set to 0.0005, and the number of epochs is 50.

TABLE IV

| Methods            | F1          | SP          | ACC         | AUC         |
|--------------------|-------------|-------------|-------------|-------------|
| Backbone           | 0.7783      | 0.9701      | 0.9578      | 0.9772      |
| Backbone+pseudo    | 0.8302      | 0.9897      | 0.9654      | 0.9863      |
| Backbone+pseudo+Moment+MSE | 0.7794      | 0.9869      | 0.9561      | 0.9714      |
| Backbone+pseudo+Moment+$L_{rec}$ | **0.8316**  | **0.9914**  | **0.9661**  | **0.9886**  |

C. Experimental Results and Analysis

The SSL-Unet is compared to several mainstream methods on DRIVE and CHASEDB1 datasets. The results are summarized in Table II and Table III. It can be concluded that SSL-Unet achieves the best performance on most metrics. The visual comparison is shown in Fig. 3 and Fig. 4. Besides, the results of ablation experiments illustrate that the generation of pseudo labels and reconstruction loss can significantly improve the segmentation performance on the backbone network in Table IV. In the end, we also compare the complexity of several different methods in Table V. It can be seen that although the number of parameters of the SSL-Unet are larger than that of other methods, lower inferred time is more beneficial for practical applications.

IV. Conclusion

We propose a SSL-Unet model for retinal vascular segmentation as well as two self-supervised training strategies. The strategy can help the self-supervised module to learn pseudo labels for improving the segmentation performance. Moreover, the fusion of both self-supervised and supervised paradigms is applied to retinal segmentation for the first time. Meanwhile, it can also be extended to any segmentation network.
REFERENCES

[1] M. Abràmoff et al., “Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle,” JAMA Ophthalmol., vol. 131, no. 3, pp. 351–357, Mar. 2013.

[2] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen and S. A. Barman, “Blood vessel segmentation methodologies in retinal images—a survey,” Comput. Methods Programs Biomed., vol. 108, no. 1, pp. 407–433, Oct. 2012.

[3] O. Ronneberger, P. Fischer and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Med. Image Comput. Comput.-Assisted Intervention, Munich, Germany, 2015, pp. 234–241.

[4] G. Huang, Z. Liu, L. Der Maaten and K. Q. Der Maaten, “Densely connected convolutional network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Hawaii, American, 2017, pp. 4700–4708.

[5] J. Zhuang, “LadderNet: Multi-path networks based on U-Net for medical image segmentation,” 2018, arXiv preprint arXiv:1810.07810.

[6] B. Wang, S. Qiu and H. G. He, “Dual encoding u-net for retinal vessel segmentation,” in Proc. Med. Image Comput. Comput.-Assisted Intervention, Shenzhen, China, 2019, pp. 84–92.

[7] T. J. Yang, T. T. Wu, L. Li and C. H. Zhu, “SUD-GAN: deep convolution generative adversarial network combined with short connection and dense block for retinal vessel segmentation,” J. Digit. Imaging, vol. 33, no. 4, pp. 946–957, 2020.

[8] Z. W. Gu, J. Cheng, H. Z. Fu, K. Zhou, H. Y. Hao, Y. T. Zhao, T. Y. Zhang, S. H. Gao and J. Liu, “Ce-net: Context encoder network for 2d medical image segmentation,” IEEE Trans. Med. Imaging, vol. 38, no. 10, pp. 2281–2292, 2019.

[9] X. F. Du and J. S. Wang and W. Z. Sun, “Densely Connected U-Net Retinal Vessel Segmentation Algorithm Based on Multi-scale Feature Convolution Extraction,” Med. Phys., vol. 66, no. 17, pp. 175013, 2021.

[10] X. F. Du and J. S. Wang and W. Z. Sun, “UNet retinal blood vessel segmentation algorithm based on improved pyramid pooling method and attention mechanism,” Phys. Med. Biol., vol. 66, no. 17, pp. 175013, 2021.

[11] A. Lahiri, V. Jain and A. Mondal, P. K. Biswas, “Retinal Vessel Segmentation Under Extreme Low Annotation: A Gan Based Semi-Supervised Approach,” in Proc. IEEE Conf. Image Process., ADNEC, UAE, 2020, pp. 418–422.

[12] Y. Y. Xu, X. X. Wu, L. Jin, S. H. Gao, R. S. M. Goh, D. S. W. Ting and Y. L., “Partially-Supervised Learning for Vessel Segmentation in Ocular Images,” in Proc. Med. Image Comput. Comput.-Assisted Intervention, Strasbourg, France, 2021, pp. 271–281.

[13] Y. X. Ma, Y. Hua, H. M. Deng, T. Song, H. Wang, Z. G. Xue, H. Cao, R. H. Ma and H. B. Guan, “Self-Supervised Vessel Segmentation via Adversarial Learning,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 7536–7545.

[14] N. Araslanov and S. Roth, “Self-supervised Augmentation Consistency for Adapting Semantic Segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 15384–15394.

[15] O. Ciga, A. L. Martel and T. Xu, “Self supervised contrastive learning for digital histopathology,” 2020, arXiv preprint arXiv:2011.13971.

[16] S. Azizi et al., “Big self-supervised models advance medical image classification,” 2021, arXiv preprint arXiv:2101.05224.

[17] K. M. He, H. Q. Fan, Y. X. Wu, S. N. Xie and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Hawaii, American, 2020, pp. 9729–9738.

[18] J. K. Liu, Q. L. Cai and M. Y. Lin, “Multi-Label Classification of Fundus Images With Graph Convolutional Network and Self-Supervised Learning,” IEEE Signal Process. Lett., vol. 28, pp. 454–458, Feb. 2021.

[19] L. Xu, Y. X. Chen, L. J. Wen, L. L. Pan and Z. L. Xu, “Boosting Few-Shot Classification with View-Learned Contrastive Learning,” in Proc. IEEE Conf. Multimedia Expo, Shenzhen, China, 2021, pp. 1–6.

[20] K. Yan, J. Z. Cai, D. K. Cai, S. Miao, A. P. Harrison, D. Z. Guo, Y. B. Tang, J. Xiao, J. J. Lu and L. Lu, “Self-supervised learning of pixel-wise anatomical embeddings in radiological images,” 2020, arXiv preprint arXiv:2012.02383.

[21] K. Chaitanya, E. Erdil, N. and K. Ender, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” 2020, arXiv preprint arXiv:2006.10511.

[22] C. Wei, L.X. Xie, X. T. Ren, Y. D. Xia, C. Su, J. Y. Liu, Q. Tian and A. L. Yuille, “Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Los Angeles, American, 2019, pp. 1910–1919.