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

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

17-10-2021 / 19-10-2021

CITATION

Ma, Chao (2021): SSL-Unet: A Self-Supervised Learning Strategy Base on U-Net for Retinal Vessel Segmentation. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16823236.v1>

DOI

[10.36227/techrxiv.16823236.v1](https://doi.org/10.36227/techrxiv.16823236.v1)

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

This work was supported by the National Natural Science Foundation of China under Grant 11971084, in part by the Science and Technology Innovation Project of Chongqing Municipal Education Commission "Construction of Two-city Economic Circle in Chengdu-Chongqing Region" under Grant KJCX2020024, in part by Chongqing Normal University Postgraduate Research and Innovation Project under Grant YKC21043 and in part by the Chongqing University Innovation Research Group under Grant CXQT20015.(Corresponding author: Jia Lu)

The authors are with the School of Computer and Information Sciences, Chongqing Normal University, 401331, China(email: macy@cqnu.edu.cn; jia-lun@163.com; jju_txs@163.com; wuoyucqnu@163.com).

Fig. 1. Illustrations of data visualization. (a)Supervised;(b)SSL+Supervised.

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.

Fig. 2. Diagram of the proposed SSL-Unet.

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. A 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 Θ 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.

Fig. 3. Illustration of the proposed SSL strategy.

C. A 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} :

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

$$f_{enc}(x_R) = p(R) + n(R) \quad (2)$$

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

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

$$f'_{enc}(x_R) = p(L) + n(R) \quad (4)$$

Fig. 4. The left and right eyes of four different patients are shown. The top row represents the left eye and the bottom row represents the right eye. We can find that the distribution of vessel trunks is roughly similar for the left and right eyes of the same patient, while it varies significantly for the same eye of different patients.

Fig. 5. We feed the left and right eyes together into the model and two pairs of features will be obtained separately via the encoder(multiple convolution layers). $p(\cdot)$ as global features and $n(\cdot)$ as local features. Then the global features (vessel trunks) are cross-fused, and the fused features are fed into the decoder for reconstruction. Due to the similarity of $p(L)$ to $p(R)$, the fused features also have the similarity. Finally, two pairs of decoupled features and reconstructed images are utilized to calculate the loss.

Then, they are sent to the corresponding decoder for the following upsampling step, respectively.

$$z_R = f_{dec}(f'_{enc}(x_R)) \quad (5)$$

$$z_L = f_{dec}(f'_{enc}(x_L)) \quad (6)$$

where z is the recovery map after multiple upsampling and f_{dec} is the decoder. The reconstructed representations of the left and right eyes are validated by maximizing consistency, as shown in Fig. 4.

D. Momentum gradient updating

The quality of the pseudo label affects the performance of the model, and training baseline(a) and baseline(b) separately increases the difference between their predictions, which leads to inaccuracy of the pseudo labels. The momentum updating is able to retain the previous updating direction in the gradient optimization. Thus, the weights of baseline(b) w are dominated by those of baseline(a) q in the training process.

$$w_{t+1} = \mu w_t + (1 - \mu)q \quad (7)$$

where μ 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 μ is, the better the baseline(b) improves

TABLE I
DATASETS

Dataset	DRIVE	CHASEDB1
Quantity	40	28
Train / Test	20/20	22/6
Resolution	700×605	996×960
Patches	2000	2000
Input size	1×48×48	1×48×48
Data augmentation	Slicing&Rotating	

its learning ability, so μ is set to 0.999 as recommended by [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

$$L_{bce} = -\frac{1}{C} \sum_{n=1}^C [g_{n,i} \log p_{n,i} + (1 - g_{n,i}) \log(1 - p_{n,i})] \quad (8)$$

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

$$L_{con} = (p(L) - p(R))^2 + (n(L) - n(R))^2 \quad (9)$$

$$L_{rec} = \frac{\sum ||(G(F_\theta(z_R)), G(F_\theta(z_L)))||_2^2 + \varepsilon}{N + \varepsilon} \quad (10)$$

where F_θ 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, ε is set to 10 in the following experiments. Therefore, the total loss function is defined as

$$L_{total} = \alpha_1 L_{bce} + \alpha_2 L_{con} + \alpha_3 L_{rec} \quad (11)$$

where α 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[25]-[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
COMPARISON OF RESULTS WITH MAINSTREAM METHODS ON THE DRIVE DATASET

Method(Year)	DRIVE			
	F1	SP	ACC	AUC
U-Net[17](2015)	0.8106	0.9776	0.9502	0.9726
DenseNet[18](2016)	0.8159	0.9779	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 <i>et al.</i> [23](2021)	—	0.9811	0.9565	0.9793
PSP-UNet[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
COMPARISON OF RESULTS WITH MAINSTREAM METHODS ON THE CHASEDB1 DATASET

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	0.9784
LadderNet[19](2017)	0.8012	0.9850	0.9591	0.9784
AtteNet[20](2018)	0.8151	0.9869	0.9620	0.9828
CE-Net[22](2019)	—	0.9787	0.9639	0.9832
Wang <i>et al.</i> [23](2021)	—	0.9717	0.9574	0.9750
PSP-UNet[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
ABLATION EXPERIMENTS ON THE CHASEDB1 DATASET

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

TABLE V
COMPARISON OF COMPLEXITY WITH MAINSTREAM METHODS ON THE CHASEDB1 DATASET

Methods	Parameters	FLOPs	Avg inferred time	GPU
U-Net	13.3M	1.09G	3.97s	RTX 3060
DenseNet	7.55M	5.33G	4.30s	RTX 3060
LadderNet	1.38M	0.12G	4.21s	RTX 3060
Ours	34.5M	2.29G	4.05s	RTX 3060

Fig. 6. Segmentation visualization of CHASEDB1 dataset, where (a) Ground truth; (b) U-Net; (c) DenseNet; (d) LadderNet; (e) AtteNet and (f) Ours.

Fig. 7. Comparison of details. (a) Ground truth; (b) U-Net and (c) Ours.

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. Xu, 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. Lin, 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-Learnable 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.
- [23] Y. T. Xie, J. P. Zhang, Z. H. Liao, Y. Xia and C. H. Shen, “Pgl: Prior-guided local self-supervised learning for 3d medical image segmentation,” 2020, arXiv preprint arXiv:2011.12640.
- [24] J. B. Grill *et al.*, “Bootstrap your own latent: A new approach to self-supervised learning,” 2020, arXiv preprint arXiv:2006.07733.
- [25] J. Soares, J. J. Leandro, R. M. Cesar, H. F. Jelinek and M. J. Cree, “Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification,” *IEEE Trans. Med. Imaging*, vol. 25, no. 9, pp. 1214–1222, 2006.
- [26] F. F. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. J. Owen and S. A. Barman, “An ensemble classification-based approach applied to retinal blood vessel segmentation,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2538–2548, 2012.