Towards Robust Retinal Vessel Segmentation via Reducing Open-set Label Noises from SAM-generated Masks

Minqing Zhang[1 [0000−0002−7214−0569]], Mengxian He[1 [0009−0005−7585−374X]], and Wu Yuan[1 ✉ [0000−0001−9405−519X]]

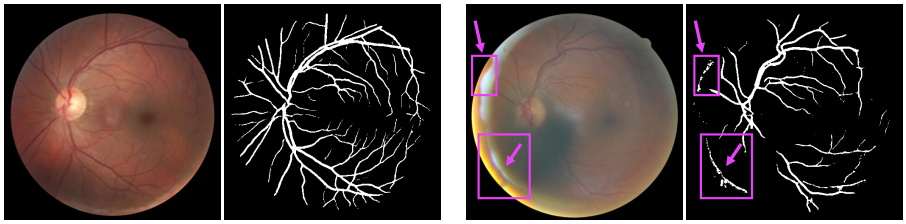
The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong
wyuan@cuhk.edu.hk

Abstract. Retinal vessel segmentation from fundus images is an important task in intelligent ophthalmology. Because vessel annotation is particularly challenging, the scarcity of training labels hinders the model robustness for real-world scenarios. Recent research has shown that SAM, a foundation model for natural image segmentation, demonstrates impressive performance on medical images after few-shot fine-tuning. Therefore, fine-tuned SAM holds promise as a pseudo label generator to alleviate the label scarcity problem in vessel segmentation. However, the limited labeled data fails to represent real-world distribution, fine-tuned SAM might produce erroneous predictions in unseen image patterns, which is known as open-set label noise. In this work, we propose SAM-OSLN to reduce open-set label noises and improve the quality of generated pseudo masks. Firstly, we introduce the prototype technique to perform open-set aware SAM fine-tuning and identify open-set label noises accordingly. Subsequently, we design an explicit label denoising method and an implicit training strategy to jointly mitigate the impact of open-set label noises. Extensive experiments demonstrate that SAM-OSLN outperforms previous state-of-the-art methods on multiple fundus datasets under synthetic and real-world scenarios.

Keywords: Robust retinal vessel segmentation · Segment anything model · Open-set label noise · Label-efficient · Domain generalization.

1 Introduction

Retinal vessel segmentation (RVS) from fundus images is one of the important medical image analysis applications, as morphological changes of vessels serve as indicative markers for the detection of various ophthalmic disorders, such as diabetic retinopathy (DR) [12], and hypertensive retinopathy [1]. Although deep neural networks (DNNs) have considerably promoted the field of medical image processing, challenges persist in the RVS task. This is because DNNs typically rely on a large amount of labeled data to achieve satisfactory performance. However, the complex structures and intricate branching patterns make vessel



(a) without open-set label noise (b) with open-set label noise

Fig. 1. A comparison of pseudo masks from fine-tuned SAM to illustrate the open-set label noise issue, which is indicated by the bounding boxes and arrows in purple.

annotation a labor-intensive and time-consuming process. Therefore, the scarcity of available labels for training has become a major bottleneck in the RVS task.

The recent surge of various large models [11] has shifted researchers' focus from merely advancing model architectures to improving the quality and quantity of training data. Therefore, it is imperative to rethink the RVS task from a data-centric AI [16] perspective. In real-world scenarios, RVS models are expected to possess strong domain generalization (DG) capabilities to handle the huge image diversity that may arise from device or operation variations. However, the aforementioned label scarcity issue, resulting in an inadequate representation of the real-world data distribution, severely constrains the DG capability. The availability of several DR screening datasets provides an opportunity to access large-scale fundus images (without vessel annotation). In summary, the key to advancing the RVS field lies in how to leverage minimal labeled data along with abundant unlabeled data to improve the DG capability.

Segment anything model (SAM) [11], a foundation model trained by billion-scale masks, has recently revolutionized the natural image segmentation field. Despite lacking consideration for medical images, previous research [15] has revealed that SAM can generate satisfactory pseudo labels for medical tasks after a few-shot fine-tuning, which can potentially alleviate the label scarcity issue in RVS. However, we observed that SAM is likely to generate erroneous predictions for unfamiliar image patterns due to the domain shifts between limited fine-tuning data and real-world samples. The fine-tuned SAM produces accurate pseudo labels for a clear fundus image (Fig. 1(a)), but the quality might be moderate for images with unseen patterns, such as bright spots (Fig. 1(b)). This is called open-set label noises. Therefore, reducing the impact of these label noises will further unleash SAM's potential to leverage large-scale unlabeled data.

In this work, we propose SAM-OSLN to identify open-set pixels and mitigate open-set label noises using explicit and implicit denoising strategies. Our method effectively alleviates the open-set label noise issue of SAM and further improves the DG capability of RVS. Our contributions are summarized as: 1) We rethink RVS from the perspective of data-centric AI, and highlight the importance of leveraging abundant unlabeled data. 2) We found that although SAM generates reasonable pseudo labels, it suffers from open-set label noises when handling

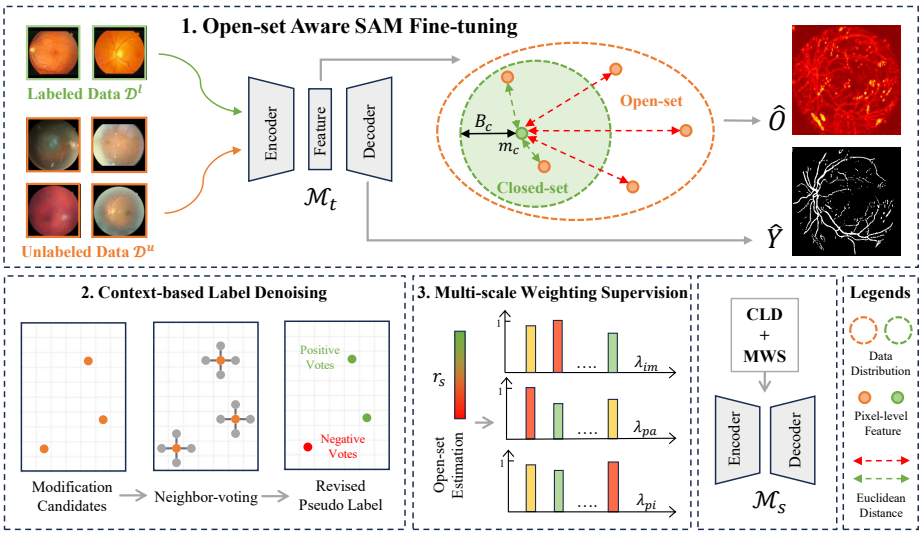


Fig. 2. The overview of our method SAM-OSLN.

extensive real-world data. 3) We propose SAM-OSLN to mitigate the open-set label noise issue of SAM and more effectively leverage large-scale unlabeled data. 4) Our approach outperforms previous state-of-the-art (SOTA) methods on various RVS datasets under synthetic and real-world open-set scenarios.

2 Method

As illustrated in Fig. 2, we utilize a teacher-student framework [8] with SAM acting as the teacher model $\mathcal{M}_t$. By utilizing pseudo labels and suppressing the open-set label noises, the knowledge is distilled into the student model $\mathcal{M}_s$. Concretely, we first conduct open-set aware SAM fine-tuning using a very limited amount of labeled data $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^N$. Subsequently, for large-scale unlabeled images $\mathcal{D}^u = \{x_j^u\}_{j=1}^M$, we can generate initial pseudo masks $\hat{Y} = \{\hat{y}_j\}_{j=1}^M$ and simultaneously identify possible open-set pixels $\hat{O} = \{\hat{o}_j\}_{j=1}^M$. Since those pixels might introduce open-set label noises in pseudo masks, we devised a context-based label denoising strategy to obtain revised pseudo labels $\overline{Y} = \{\overline{y}_j\}_{j=1}^M$. In addition, considering that the presence of open-set samples increases training difficulty, we designed a multi-scale weighting strategy to effectively utilize the large-scale revised labels.

2.1 Open-set Aware SAM Fine-tuning

Our method is initialized with fine-tuning SAM on $\mathcal{D}^l$. To mitigate open-set label noises, we introduce prototype techniques [26] for possible open-set pixel recognition. Specifically, by forward feeding the training images into feature extractor

$f_\theta(\cdot)$ of the fine-tuned SAM, the pixel-level features of training samples could be collected to build feature pools $\mathcal{P}_c$ for each class c according to Eqt. 1, where $c = 0$ (background) or 1 (vessel) and h, w indicates height, width indices of an image. The features in these pools represent closed-set image patterns and have the potential to identify open-set label noises. To find the anchor point for each class, we calculate the prototype m_c by averaging all the features within each pool following Eqt. 2. We also define the decision boundary B_c for each class using the farthest distance from pool features to the prototype (see Eqt. 3).

$$\mathcal{P}_c = \{f_\theta(x_i^l)[h,w] | x_i^l \in \mathcal{D}^l, y_i^l = c\} \tag{1}$$

$$m_c = \frac{\sum_{p_c \in \mathcal{P}_c} p_c}{N_c}, \text{where } \; N_c = size(\mathcal{P}_c) \tag{2}$$

$$B_c = \max_{p_c \in \mathcal{P}_c} \frac{m_c \cdot p_c}{\|m_c\| \, \|p_c\|} \tag{3}$$

By determining whether each pixel-level feature falls within the corresponding decision boundary of any prototype, we can identify possible open-set pixels $\hat{o}_j$ for each unlabeled image x_j^u according to Eqt. 4. So far, for an unlabeled image x_j^u, its initial pseudo label $\hat{y}_j$ and possible open-set recognition $\hat{o}_j$ could be obtained based on our open-set aware SAM fine-tuning.

$$\hat{o}_j[h,w] = \begin{cases} 0, \text{ if } \; \frac{m_c \cdot x_j^u[h,w]}{\|m_c\| \left\| x_j^u[h,w] \right\|} < B_c \;\; \text{for each class} \;\; c \\ 1, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{4}$$

2.2 Context-based Label Denoising (CLD)

The initial pseudo labels $\hat{Y}$ suffer from the open-set label noise issue, i.e., SAM easily misinterprets some open-set pixels as vessels. We observed that certain open-set pixels (such as bright spot areas) occur outside vessel boundaries, while others might lie within vessel regions. Therefore, simply considering all open-set pixels as background could be an inaccurate strategy. To effectively suppress open-set label noise, we devised a context-based label denoising approach that leverages spatial information for label modification. For all the pixels predicted as vessels in the initial pseudo labels, our goal is to eliminate the open-set label noises and preserve the true vessel labels. Specifically, we designate pixels that meet two criteria as modification candidates $\mathcal{S}$: 1) predicted as vessels in the initial pseudo labels, and 2) identified as possible open-set pixels, based on Eqt. 5. For each candidate pixel, we proposed a neighbor-voting approach given by Eqt. 6 to extract its contextual information, where $N^*(h,w)$ denotes the neighborhood of pixel (h,w). A pixel lacking nearby closed-set vessel predictions is more likely to be an isolated open-set label noise and should be reclassified as background. Otherwise, the candidate pixel should be retained as a vessel. By applying this process to each candidate pixel, we could obtain revised pseudo labels $\overline{Y} = \{\overline{y}_j\}_{j=1}^M$ where open-set label noises have been suppressed.

$$\mathcal{S} = \{x_j^u[h,w] | x_j^u \in \mathcal{D}^u, \hat{y}_j[h,w] = 1, \hat{o}_j[h,w] = 1\} \tag{5}$$

$$\overline{y}_j[h,w] = \begin{cases} 0, \text{ if} \sum_{h',w' \in N^*(h,w)} \hat{y}_j[h',w'] \cdot (1 - \hat{o}_j[h',w']) > 0 \\ 1, \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{6}$$

2.3 Multi-scale Weighting Supervision (MWS)

Since accurately identifying open-set pixels is fairly challenging, the revised pseudo labels might not be completely noise-free. Furthermore, those severe open-set samples could intensify the training difficulties. Therefore, simply employing the revised labels to train the student model is sub-optimal. We observed that open-set label noises show inconsistent distribution across multiple scales, with some images and regions exhibiting a higher prevalence of such label noises. Therefore, we devised a multi-scale weighting strategy to further mitigate the impact of label noise based on the distribution of open-set pixels. In general, our strategy involves evaluating the degree of open-set pixels at multi-scale to obtain weights at the image-, patch-, and pixel- levels, i.e., λ_{im}, λ_{pa}, λ_{pi}, which are subsequently used to weight the loss function for training the student model. We assign $\lambda_{pi} = 1$ to all closed-set pixels, and determine the weights for open-set pixels based on the distance to the prototypes according to Eqt. 7.

$$dist(x_j^u[h,w]) = \frac{1}{C} \sum_{c=0}^{C} \frac{m_c \cdot x_j^u[h,w]}{\|m_c\| \left\|x_j^u[h,w]\right\|} \tag{7}$$

$$\lambda(r_s) = \frac{2N_s - r_s - 1}{2(N_s - 1)} \tag{8}$$

After collecting and ascendingly sorting distances of all open-set pixels, we obtain the pixel-level ranking. Patch-level ranking is determined by the ascending order of the number of open-set pixels in each patch. Similarly, we can also have the image-level ranking. After collecting pixel-, patch-, and image- rankings, we can then determine the weights for each scale based on Eqt. 8, where r_s and N_s represent the ranking and total sample size of each scale, respectively. Based on the modified labels and corresponding multi-scale weights, we can compute the loss for each unlabeled data x_j^u during the student model training using Eqt. 9.

$$\mathcal{L}(x_j^u) = \sum_{h=1,w=1}^{H,W} \lambda_{im} \cdot \lambda_{pa} \cdot \lambda_{pi} \cdot |\mathcal{M}_s(x_j^u)[h,w], \overline{y}_j[h,w]|_{loss} \tag{9}$$

3 Experiments

3.1 Summary of Open-source Fundus Datasets

We collected representative open-source datasets for fundus images and provided an analysis from the perspective of data-centric AI. As mentioned in Sec. 1, due

to annotation challenges, RVS datasets typically contain only a small number of vessel labels. Apart from FIVES [10] with 800 annotations, the majority of those datasets contain only a few dozen labels. Furthermore, there are obvious domain shifts across different RVS datasets. Since fundus photography is an economical and non-invasive imaging tool widely used in DR screening [6], some DR grading datasets like EyePACS [7], while lacking pixel-level vessel annotation, typically contain a large number of fundus images collected from real-world scenarios. These images reflect various interferences that arise during the imaging process, such as bright spot artifacts, inappropriate exposure intensity, and motion blur. Therefore, we aim to investigate more effective ways of utilizing the limited annotated data to learn vessel patterns and leveraging abundant unlabeled data to enhance DG capability, achieving robust RVS. Our method would be investigated on those datasets under both synthetic and real-world settings.

3.2 Implementation Details

Our experiments were conducted using a GeForce RTX 3060 GPU and PyTorch 1.12.1. We followed SAM-Adapter [3] to fine-tune the ViT-B version of SAM on $\mathcal{D}^l$. TransUnet [2] was adopted as our student model. The 400 training epochs, initial learning rate of 0.001, an ADAM optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), batch size of 3, and cross-entropy loss were leveraged. Data augmentations were random horizontal/vertical flipping, and rotation. All images were resized to a unified resolution of 512. The patch size was 64. N^* was set to 5×5 neighbor pixels.

3.3 Comparison on Synthetic Open-set Interference

We herein investigate the capability of our method to generate pseudo labels under varying synthetic open-set interference. First, we randomly selected only 10 clean labels from FIVES [10] to train models, and the remaining 790 samples were introduced with open-set interference for evaluation. We utilized CVC-ClinicSpec [19] dataset, which consists of 58 colonoscopy images with pixel-level masks of bright spot areas, as the source of open-set interference. By extracting the specular highlight regions from CVC-ClinicSpec and merging them into evaluating samples of FIVES, we were able to synthesize fundus samples with open-set interference. It is worth noting that with a 1:1 intensity blending ratio, the open-set interference is only introduced to the image and does not alter the semantic labels of vessels. To comprehensively assess the robustness of our method in generating pseudo labels for synthetic open-set images, we implemented three levels of open-set interference based on the ratio α of open-set pixels to vessel pixels: mild $\alpha = 0.15$, moderate $\alpha = 0.30$, and severe $\alpha = 0.45$.

As shown in Tab. 1, we compared the quality of pseudo labels generated by our method with the classical self-training [13] and fine-tuning SAM [3] under multiple levels of open-set interference using Dice score and intersection over union (IoU). Experimental results indicate that due to the generalization capability of the foundation model, fine-tuned SAM exhibits stronger robustness and produces better pseudo labels with limited training labels, resulting in a 12.7%

Table 1. Comparisons of the quality of pseudo labels with several methods under various open-set interferences on FIVES. The best performance is marked in bold.

Methods	Clean($\alpha = 0$)		Mild($\alpha = 0.15$)		Moderate($\alpha = 0.30$)		Severe($\alpha = 0.45$)	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)
Self-training [13]	64.5	48.9	56.4	40.6	50.5	34.9	45.6	30.5
Fine-tuned SAM [3]	77.2	61.7	72.5	55.6	68.5	51.1	64.5	46.6
SAM-OSLN (w/o MWS)	**79.1**	**65.6**	**75.5**	**60.4**	**71.8**	**56.0**	**68.1**	**51.7**

Dice increase on clean samples than self-training. As the level of open-set interference increases, fine-tuned SAM tends to make erroneous predictions leading to a performance decline. Our method, however, explicitly revises open-set label noises, consistently improving pseudo label qualities regardless of the interference degree. Compared to fine-tuned SAM, our method improves by 3.9 IoU on clean samples and 5.1 IoU under severe interference, respectively.

3.4 Comparison with SOTA Methods on Real-world Settings

An RVS benchmark was constructed to simulate real-world scenarios, as reported in Tab. 2. We included only 10 STARE [9] data as a limited label source along with 35,126 unlabeled data from EyePACS [7] as large-scale fundus images. Six RVS datasets were utilized to evaluate the algorithms' DG capability and robustness based on mean and minimum Dice scores, respectively.

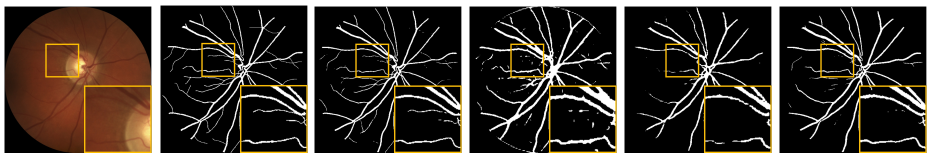
We included SOTA methods of several kinds and explored how to achieve superior RVS performance in real-world scenarios, as shown in Tab. 3. First, we trained U-Net [18] with the 10 labels as the baseline. We then compared SOTA RVS methods, that utilize techniques like vessel structure priors [5, 20] and learnable data augmentation [14]. These methods achieved improvements over the baseline, such as a mean Dice of 63.0 from AADG [14] and a min. Dice of 46.5 from DRIS-GP [5] compared to the baseline. Constrained by the limited involvement of training samples, RVS methods do not yield considerable improvements in DG performance. Therefore, we compared semi-supervised learning (SSL) methods which utilize abundant unlabeled data. CPS [4] achieving 65.3 mean Dice highlights the value of large-scale data. However, the lack of consideration for the scarcity of labels and the domain shift problem of unlabeled data led to inconsistent improvements in SSL methods. For example, due to the poor quality of generated pseudo labels, Unimatch [24] performed worse than the baseline. To this end, we also trained noisy label learning (NLL) approaches using pseudo labels from fine-tuned SAM. Attributed to the improved quality of pseudo labels, P&L [27] increased mean Dice to 67.0. However, as the NLL

Table 2. The details of our RVS benchmark for evaluation under real-world scenarios.

	Labeled	Unlabeled	Domain Generalization Evaluation Datasets					
Dataset	STARE	EyePACS	IOSTAR	ORVS	DR-HAGIS	Les-AV	RETA	TREND
Number	10	35,126	30	49	40	22	54	82

Table 3. Quantitative comparisons of SAM-OSLN with SOTA methods on the real-world benchmark. The best results are marked in bold. Our method outperforms the comparative algorithms, especially with a considerable improvement in min. Dice.

SOTAs		Multi-domain Evaluation						Dice Scores (%)	
Type	Methods	IOSTAR	ORVS	HAGIS	Les-AV	RETA	TREND	Mean↑	Min.↑
	Baseline	53.3	41.0	69.1	69.0	72.0	56.9	60.2	41.0
RVS	DRIS-GP[5]	66.5	52.2	54.4	64.4	66.0	46.5	58.3	46.5
	SkelCon[20]	69.1	57.1	64.5	71.0	69.7	46.1	62.9	46.1
	AADG[14]	68.6	65.1	59.7	69.6	70.1	45.0	63.0	45.0
SSL	MT[21]	68.6	39.0	68.9	78.1	76.8	52.3	63.9	39.0
	Co-T[17]	65.5	39.6	69.8	77.5	76.7	54.0	63.8	39.6
	CPS[4]	70.1	40.7	69.9	77.2	78.1	55.6	65.3	40.7
	U^2PL[22]	68.7	59.4	69.9	71.5	68.7	43.1	63.6	43.1
	Unimatch[24]	63.5	41.8	64.4	71.8	65.6	51.1	59.7	41.8
NLL	P&L[27]	72.9	40.5	73.0	80.7	76.3	58.3	67.0	40.5
	CL+SLSR[25]	72.0	40.9	73.1	80.5	75.1	58.2	66.6	40.9
	MTCL[23]	68.4	40.0	71.3	79.9	76.5	56.6	65.5	40.0
Ours	SAM-OSLN	76.6	69.3	75.5	83.5	77.5	62.9	**74.2**	**62.9**
	w/o CLD	74.2	68.9	74.7	82.8	73.3	60.7	72.5	60.7
	w/o MWS	75.9	68.2	70.5	81.2	72.6	54.7	70.5	54.7



(a) Image (b) GT (c) Ours (d) AADG[14] (e) CPS[4] (f) P&L[27]

Fig. 3. Qualitative illustrations of SAM-OSLN versus previous SOTA methods.

strategies are usually designed for manual annotation rather than automatically generated ones, these methods still did not achieve satisfactory performance in robustness. Through leveraging massive amounts of unlabeled data and addressing open-set label noise issues, SAM-OSLN achieved the best mean Dice of 74.3 and minimum Dice of 63.2, surpassing previous SOTA methods. The ablation study demonstrates that the explicit label denoising method CLD and the implicit loss weighting strategy MWS both contributed positively to our approach. Qualitative comparisons on the Les-AV dataset in Fig. 3 show the superior segmentation performance of our method compared to previous SOTAs.

4 Conclusion

We provide a rethinking for the RVS task from the perspective of data-centric AI and argue that effectively utilizing a large amount of unlabeled data in scenarios with extremely limited labeled data is crucial for improving DG capability. To address this challenge, we propose a solution that leverages SAM as a high-quality pseudo label generator and mitigates the open-set label noise issue. Comparative experiments, both under synthetic and real-world settings, demonstrate the superior performance of our method compared to previous SOTA approaches.

Although our current effort is concentrated on alleviating the detrimental effects of open-set label noise during the exploitation of unlabeled data for model robustness, several other aspects remain unaddressed. Specifically, the deficiency of fine-scale vessels and the utilization of prior knowledge of vessel morphology are areas that are worth exploration in future research.

Acknowledgement. This work is supported by the Research Grants Council (RGC) of Hong Kong SAR (GRF14216222, GRF14201824), the Innovation and Technology Fund (ITF) of Hong Kong SAR (ITS/252/23), and the Science, Technology, and Innovation Commission (STIC) of Shenzhen Municipality (SGDX20220530111005039).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akbar, S., Akram, M.U., Sharif, M., Tariq, A., Khan, S.A.: Decision support system for detection of hypertensive retinopathy using arteriovenous ratio. Artificial intelligence in medicine **90**, 15–24 (2018)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything? – sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more (2023)
4. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
5. Cherukuri, V., Bg, V.K., Bala, R., Monga, V.: Deep retinal image segmentation with regularization under geometric priors. IEEE Transactions on Image Processing **29**, 2552–2567 (2019)
6. Dai, L., Sheng, B., Chen, T., Wu, Q., Liu, R., Cai, C., Wu, L., Yang, D., Hamzah, H., Liu, Y., et al.: A deep learning system for predicting time to progression of diabetic retinopathy. Nature Medicine pp. 1–11 (2024)
7. Emma Dugas, Jared, Jorge, Will Cukierski: Eyepacs. https://kaggle.com/competitions/diabetic-retinopathy-detection (2015)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000)
10. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9**(1), 475 (2022)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

12. Klein, R., Lee, K.E., Danforth, L., Tsai, M.Y., Gangnon, R.E., Meuer, S.E., Wong, T.Y., Cheung, C.Y., Klein, B.E.: The relationship of retinal vessel geometric characteristics to the incidence and progression of diabetic retinopathy. Ophthalmology **125**(11), 1784–1792 (2018)
13. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
14. Lyu, J., Zhang, Y., Huang, Y., Lin, L., Cheng, P., Tang, X.: Aadg: automatic augmentation for domain generalization on retinal image segmentation. IEEE Transactions on Medical Imaging **41**(12), 3699–3711 (2022)
15. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)
16. Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H.R., et al.: Dataperf: Benchmarks for data-centric ai development. Advances in Neural Information Processing Systems **36** (2024)
17. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the european conference on computer vision (eccv). pp. 135–152 (2018)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
19. Sánchez, F.J., Bernal, J., Sánchez-Montes, C., de Miguel, C.R., Fernández-Esparrach, G.: Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos. Machine Vision and Applications **28**(8), 917–936 (2017)
20. Tan, Y., Yang, K.F., Zhao, S.X., Li, Y.J.: Retinal vessel segmentation with skeletal prior and contrastive loss. IEEE Transactions on Medical Imaging **41**(9), 2238–2251 (2022)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
22. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4248–4257 (2022)
23. Xu, Z., Lu, D., Luo, J., Wang, Y., Yan, J., Ma, K., Zheng, Y., Tong, R.K.Y.: Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation. IEEE Transactions on Medical Imaging **41**(11), 3062–3073 (2022)
24. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7236–7246 (2023)
25. Zhang, M., Gao, J., Lyu, Z., Zhao, W., Wang, Q., Ding, W., Wang, S., Li, Z., Cui, S.: Characterizing label errors: Confident learning for noisy-labeled image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 721–730. Springer (2020)

26. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2582–2593 (2022)
27. Zhu, H., Shi, J., Wu, J.: Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 576–584. Springer (2019)