PseudoRET: A General Retinal Segmentation Framework with Multi-target Pseudo Embeddings

Zhonghua Wang[1,†], Sijia Li[1,†], Lie Ju[1,3], Wei Feng[1], Sijin Zhou[1], Ming Hu[1], James Huang[4], Anthony Huang[4], Li Dong[2], and Zongyuan Ge[1]

[1] AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, Australia
[2] Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China
[3] Institute of Ophthalmology, University College London, London, UK
[4] Airdoc LLC., Beijing, China
zongyuan.ge@monash.edu
[†] These authors contributed equally to this work.

Abstract. The Segment Anything Model has demonstrated strong cross-domain generalization for image segmentation, making it a promising foundation for medical imaging. However, deploying SAM in retinal fundus analysis remains difficult because existing datasets are fragmented and task-specific, typically annotated for vessels, lesions, or optic disc/cup only, preventing effective multitask training. In addition, many retinal targets are spatially diffuse, making manual prompt engineering impractical in real workflows. We present PseudoRET, a semi-supervised framework that unifies disjoint fundus annotations to enable generalizable, multitask segmentation. PseudoRET adopts a task-agnostic training strategy and introduces a pseudo memory that aggregates and propagates supervision across tasks to mitigate annotation sparsity. To improve robustness to label noise, we further propose a confidence-aware pseudo loss that adaptively reweights pseudo labels stored in the memory, retaining reliable signals while suppressing noisy ones during training. We validate PseudoRET on 13 public datasets and 4 relabeled datasets covering vessel, lesion, and optic disc/cup segmentation. PseudoRET consistently outperforms state-of-the-art baselines across tasks and datasets. Ablation studies show that both the task-agnostic paradigm and CAP Loss contribute substantially, and their combination yields the strongest gains. Our work provides a recipe for training a SAM-style multi-target model under sparse annotations, and we release code and relabeled datasets at https://github.com/Wzhjerry/PseudoRET.

Keywords: Semi-supervised Learning · Multi-target Segmentation · Retinal Segmentation

1 Introduction

Accurate segmentation of retinal structures and pathological findings in fundus images supports diagnosis, monitoring, and risk assessment for ocular and

systemic diseases [22][11]. Foundation segmentation models, such as the Segment Anything Model (SAM) [17], have motivated interest in using a single model across segmentation settings. However, fundus datasets are fragmented and task-specific. Most provide annotations for only one target and follow different label definitions, complicating unified model training. In addition, SAM relies on prompts, but many fundus targets are small, diffuse, or numerous, making prompt-based workflows difficult to scale [5]. These issues are further affected by domain shifts and severe class imbalance for small targets [7,8,10,6,9].

Semi-supervised and self-supervised learning are common ways to use unlabeled data when annotations are limited. Semi-supervised learning uses techniques such as consistency regularization and pseudo-labeling, and has been applied to disease classification [34] and single-structure segmentation [26]. Temporal Ensembling [19] and Mean Teacher [28] enforce teacher–student consistency over time, but within a fixed single-task label space. Self-supervised learning uses pretext objectives such as masked reconstruction and contrastive learning [13][4][29], which can improve representations under sparse labels. However, most existing methods are developed for a single task with a fixed label space, and they usually assume that the available annotations partially cover the same target. In fundus imaging, a more practical setting is cross-task annotation incompleteness: different datasets annotate different targets and may use different label protocols. This setting can be viewed either as multi-target learning with incomplete annotations or as semi-supervised learning along the task-label axis, where the unlabeled dimension is which task is annotated rather than which pixels. Directly mixing such datasets leads to missing labels and biased supervision, for example when unlabeled regions are treated as background. In addition, in multitask settings, pseudo labels can introduce noise and the errors may accumulate across tasks and datasets [15][25]. Therefore, we need methods that can combine disjoint annotations while controlling pseudo-label noise.

To address these issues, we propose **PseudoRET**, a segmentation framework with pseudo embedding for fundus image analysis. PseudoRET aims to train a single model with disjoint annotations across tasks and datasets by using dynamically weighted pseudo supervision. In contrast to these single-task consistency methods, PseudoRET shares pseudo supervision *across tasks* for the same image to address cross-dataset label-space mismatch. Our approach includes:

- A **task-agnostic pseudo-embedded training** scheme that uses a semi-supervised pipeline and a **pseudo embedding memory** to store and reuse pseudo labels, together with available ground-truth supervision.
- A **confidence-aware dynamic pseudo loss** that re-weights pseudo labels in the pseudo embedding memory based on spatial reliability and prediction stability, reducing the influence of low-confidence regions.
- A **curated benchmark** with 4 relabeled public datasets covering 3 segmentation tasks and 7 classes, which supports systematic evaluation under fragmented annotations.

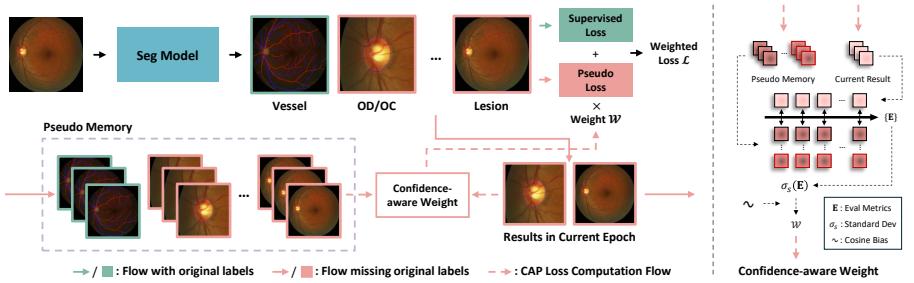

Fig. 1. Overall framework of proposed PseudoRET. Left shows the task-agnostic training paradigm. Right shows the confidence-aware weight for computing the loss.

2 Method

Figure 1 illustrates the proposed PseudoRET. PseudoRET combines a task-agnostic training paradigm with pseudo embedded supervision and a confidence-aware dynamic pseudo loss to address incomplete fundus annotations. We introduce a pseudo-memory that stores model predictions from the most recent n epochs as a rolling cache of pseudo labels. During training, current predictions are checked against the pseudo-memory to emphasize consistent anatomical patterns and suppress transient errors. The memory is updated with a FIFO policy, replacing the oldest entries with predictions from the current epoch. To reduce error propagation, we apply a confidence-based dynamic weight to the pseudo loss so that high-certainty pseudo labels contribute more while the learning signal from ground-truth labels remains dominant. Pseudo embedding means storing recent predictions and using their temporal agreement as auxiliary supervision.

2.1 Task-agnostic Training with Pseudo Embedding

A common approach to incomplete annotations trains task-specific models and generates static pseudo labels for missing targets, which can introduce dataset-specific bias and does not update pseudo supervision during training. We instead use self-ensembling with online pseudo embedding.

For each image, we separate tasks with ground-truth annotations from those without. For each unannotated task, the pseudo memory stores probability maps from the most recent n epochs and updates them in FIFO order. Their average, $\widetilde{\mathbf{Y}}^{(c)} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{Y}}_i^{(c)}$, is combined with supervised labels in the training loss as

$$\mathcal{L}(f_\theta|\mathcal{I}) = -\frac{1}{N}\left(\mathbf{Y}_0 \log\left(\hat{\mathbf{Y}}_0\right) + \sum_{c=0}^{C}\sum_{i=1}^{I}\widetilde{\mathbf{Y}}_i^{(c)} \log\left(\hat{\mathbf{Y}}_i^{(c)}\right)\right), \tag{1}$$

where f_θ denotes the model parameters, $\mathcal{I}$ is the input image set, N is the number of pixels, $\mathbf{Y}_0$ is the ground-truth label, and $\widetilde{\mathbf{Y}}$ denotes pseudo labels retrieved from the pseudo memory. C is the number of tasks, and I is the number of stored pseudo labels per task. We use pixel-wise cross-entropy for both supervised

and pseudo terms and ignore pixels without supervision for a given task. This formulation uses available ground-truth supervision while incorporating recent model predictions as additional supervision for unlabeled tasks.

2.2 Confidence-aware Pseudo Loss

Pseudo labels can be noisy at early epochs, and using them with a fixed weight can harm training. We introduce a Confidence-aware Pseudo Loss (CAP Loss) that adjusts the contribution of pseudo supervision based on reliability estimated from the pseudo memory. Unlike entropy- or confidence-based reweighting of a single prediction, CAP Loss uses the temporal dispersion (σ_s) of pseudo labels in the FIFO memory to gauge their stability across recent epochs.

We compute a dynamic weight by measuring the dispersion of pseudo-label quality over the last n stored predictions. We evaluate pseudo labels in the memory with an error measure $\mathbf{E}$ and summarize the variability with σ_s. In practice, $\mathbf{E}$ can be Dice dissimilarity or cross-entropy between memory predictions and the current prediction, and σ_s denotes the standard deviation over indices from 1 to n. The weight is defined as

$$\mathcal{W} = \exp\left(-\frac{1}{C}\sum_{c=0}^{C}\sigma_s\left(\left\{\mathbf{E}\left(\widetilde{\mathbf{Y}}_i^{(c)}, \hat{\mathbf{Y}}_i^{(c)}\right) \;\middle|\; 1 \le i \le n\right\}\right)\right) + \frac{1}{3}\cos\left(\frac{\pi E}{10}\right), \quad (2)$$

where $\widetilde{\mathbf{Y}}_i^{(c)}$ and $\hat{\mathbf{Y}}_i^{(c)}$ follow Eq. 1, $\mathbf{E}$ denotes the evaluation metric, and E is the training epoch index. The cosine term adds a small bias to avoid near-zero weights in early training. The weight is computed per image, averaged over tasks, and used to scale the pseudo term in Eq. 1 as

$$\mathcal{L}(f_\theta|\mathcal{I}) = -\frac{1}{N}(\mathcal{L}_s + \mathcal{W} \times \mathcal{L}_p), \quad (3)$$

where $\mathcal{L}_s$ is the supervised loss on ground-truth labels and $\mathcal{L}_p$ is the loss computed from pseudo labels in the pseudo memory. This weighting balances pseudo supervision and ground-truth supervision during training.

3 Experimental Results

3.1 Datasets

We evaluate PseudoRET on 13 public fundus datasets covering three segmentation tasks, including retinal vessels, lesions, and optic disc and cup. MAPLES-DR is used only for evaluation on its test split and is not included in training. For datasets with official predefined splits, we follow the provided partitions. For datasets without standardized splits, we perform patient-level random splits into 70% training, 10% validation, and 20% test, with no patient overlap across splits. To enable multi-task evaluation under disjoint annotations, we further relabel four datasets, DRIVE [27], HRF [2], GAMMA [30], and IDRiD [24],

Table 1. Dataset details in PseudoRET (train/val/test splits). √: relabeled datasets.

Task	Datasets & Train/Val/Test Splits & Relabel			
Vessel	DRIVE [27] (√)	17/3/20	FIVES[16]	420/180/200
	HRF[2] (√)	31/4/10	STARE[14]	14/1/5
OD/OC	G1020[1]	704/101/205	GAMMA[30] (√)	70/10/20
	ORIGA[33]	455/64/131	Papila[18]	340/48/100
	Refuge[23]	400/400/400		
Lesion	DDR[21]	383/149/225	FGADR[35]	1289/184/369
	IDRiD[24] (√)	43/11/27		
General	Maples-DR[20]	120/18/60		

by adding annotations for the tasks that are missing in their original releases. For lesion segmentation, we annotate hard exudates and hemorrhages, while microaneurysms and soft exudates are not annotated due to their very limited frequency in these datasets. These added annotations are used only for evaluation. Training uses only the original single-task labels of the other 12 datasets, so the relabeled datasets and MAPLES-DR form a unified multi-target evaluation benchmark and never serve as cross-task training supervision. For relabeling, all datasets followed one annotation protocol with consistent class definitions. The data were split into three disjoint subsets, each annotated by one of three retinal specialists and cross-checked by the other two, with disagreements resolved by consensus. The annotations and inter-rater agreement statistics are released with the public dataset. Table 1 summarizes the original dataset splits, annotation targets, and task distributions.

3.2 Experimental Setting

PseudoRET is implemented with SwinUNETR [12]. It uses no SAM backbone and is "SAM-style" only because one prompt-free model produces all task outputs. Images are resized to 640×640. We apply random flipping, rotation, and intensity jitter during training. We use AdamW with a learning rate of 1×10^{-4}, weight decay of 0.01, and cosine decay. We train for 100 epochs with cross-entropy for both supervised and pseudo losses. Pseudo memory updates start at epoch 10, and pseudo-loss weight updates start at epoch 15. We report Dice Similarity Coefficient and Jaccard Index.

3.3 Comparisons with SOTAs

We compare PseudoRET with recent methods from three categories: SAM-based approaches including Medical-SAM-Adapter [31], MedSAM2 [37], and SAM3 [3]; a fully supervised baseline SAM2-UNet [32]; and a retina-specific foundation model RETFound [36]. For fair comparison, we re-implement all methods from their official repositories and align training details and evaluation protocols as

Table 2. Quantitative comparison of PseudoRET with SOTAs on four relabeled datasets. Full-training denotes fully supervised training on the relabeled datasets. PseudoRET-Base/-Large use SwinUNETR with Swin ViT-B/-L. SAM-family methods use official repository configurations without modification.

Task	Vessel		OD/OC				Lesion			
	Vessel		OD		OC		EX		HE	
Metrics	DSC	JAC	DSC	JAC	DSC	JAC	DSC	JAC	DSC	JAC
Full - supervision	0.836	0.719	0.851	0.750	0.721	0.591	0.550	0.401	0.530	0.371
MedSAM2 [37]	0.637	0.469	0.859	0.784	0.449	0.304	0.336	0.274	0.260	0.181
Medical-SAM-Adapter [31]	0.612	0.444	0.826	0.711	0.727	0.599	0.299	0.228	0.315	0.217
SAM3 [3]	0.668	0.529	0.862	0.795	0.738	0.612	0.354	0.275	0.363	0.257
SAM2-UNet [32]	0.677	0.541	0.855	0.796	0.723	0.605	0.372	0.301	0.385	0.278
RETFound [36]	0.602	0.449	0.841	0.787	0.692	0.571	0.279	0.178	0.312	0.206
PseudoRET - Base (ours)	0.713	0.565	**0.891**	**0.807**	0.747	0.629	0.397	0.267	0.437	0.296
PseudoRET - Large (ours)	**0.739**	**0.576**	0.887	0.802	**0.795**	**0.669**	**0.432**	**0.303**	**0.493**	**0.336**

closely as possible. All models are trained using only the original single-task labels of the 12 training datasets. The relabeled multi-target annotations and all MAPLES-DR annotations are reserved exclusively for evaluation. We then evaluate them on the four relabeled benchmark datasets using direct inference without additional fine-tuning, and report results for vessel, lesion, and optic disc and cup segmentation with the same metrics. Table 2 summarizes the quantitative comparisons. PseudoRET obtains the best performance across all four relabeled datasets for all three tasks. The largest gains are observed on lesion segmentation, followed by vessel and optic disc and cup segmentation. Prompt-based SAM-style inference is less stable on fundus images when targets are small and spatially scattered, which can increase false positives. In contrast, PseudoRET learns from disjoint annotations through pseudo memory and uses confidence-aware weighting to reduce the impact of low-quality pseudo labels. Figure 2 provides qualitative comparisons on representative cases. Across the three tasks, PseudoRET produces more accurate masks with clearer boundaries and more complete structures. For lesions, it reduces scattered false positives and better preserves lesion extent. For vessels, it maintains continuity of thin branches with fewer breaks. For optic disc and cup, it provides more consistent cup-to-disc separation. Importantly, PseudoRET handles multiple targets in a single model and produces all task outputs for the same image, rather than optimizing for one task at the expense of others. The qualitative results are consistent with the quantitative results in Table 2, showing that the model maintains stable performance across tasks without a clear drop in any single task.

3.4 Ablation Study

We conduct ablation studies on the 12 single-task public datasets with original annotations, without using the relabeled benchmark. We study two questions. First, we test whether adding pseudo supervision is useful for training a single multi-target model on fundus segmentation under incomplete labels. Second, we

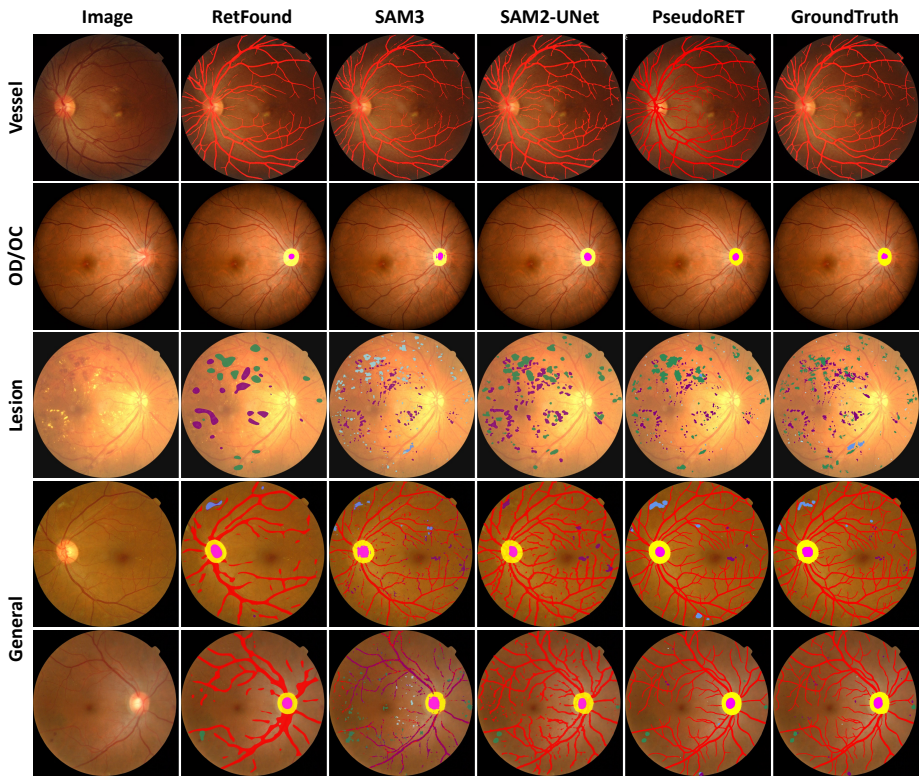

Fig. 2. Qualitative segmentation comparisons of PseudoRET against other SOTA methods. The segmentation results are overlaid on the images.

examine how the pseudo-memory size n affects both accuracy and the dynamic weighting behavior.

Effect of task-agnostic pseudo embedding and CAP Loss. We compare three variants. The first uses task-specific pseudo embedding, where pseudo labels are generated and reused only within the same task. This variant is a single-task temporal-ensemble baseline within our framework. The second uses task-agnostic pseudo embedding, where the pseudo memory is shared across tasks and provides auxiliary supervision for missing targets. The third adds CAP Loss on top of task-agnostic pseudo embedding to down-weight unreliable pseudo supervision. Figure 3 summarizes the results, reporting per-task DSC on each dataset's own test split. Task-specific pseudo embedding gives only small gains. A likely reason is that pseudo labels in fundus images can be noisy for small targets and low-contrast regions, and directly reusing them can introduce inconsistent supervision. Task-agnostic pseudo embedding improves the average DSC by 2.1%, suggesting that sharing supervision across tasks provides additional training signal when labels are missing. Adding CAP Loss further improves DSC by 1.2% on average. This is consistent with the design of CAP Loss, which

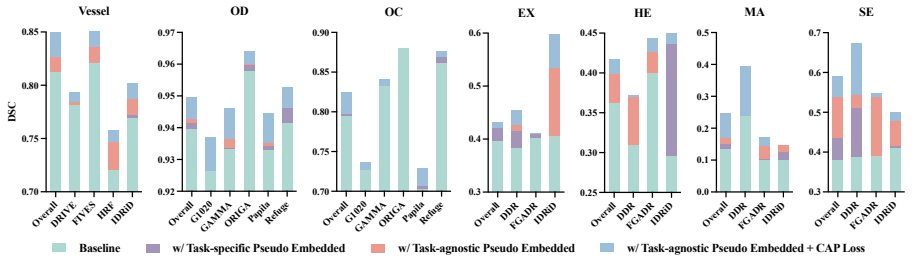

Fig. 3. Per-task DSC of the baseline, task-specific pseudo embedding, task-agnostic pseudo embedding, and the full model with CAP Loss, on the 12 single-task datasets across 3 downstream tasks.

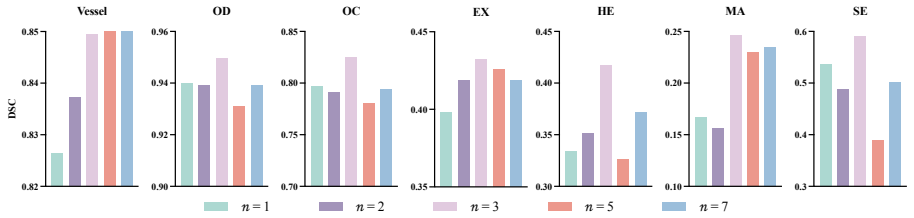

Fig. 4. Results of segmentation performance on downstream tasks with combined test set using different value of n for pseudo memory.

reduces the contribution of pseudo labels when the pseudo memory shows large variation across recent epochs. Together, task-agnostic sharing and CAP Loss improve the average DSC by 3.3% over the baseline. The small gain from the task-specific temporal-ensemble variant helps distinguish cross-task sharing from a generic single-task SSL effect.

Impact of pseudo-memory size. We then analyze the pseudo-memory size n, which controls how many past predictions are retained for computing the confidence weight. Figure 4 reports results for vessel, lesion, and optic disc and cup segmentation. For visualization, we aggregate test sets across datasets within each task. We observe a non-monotonic trend. Performance peaks at $n = 3$. Increasing n beyond 3 leads to lower accuracy, with up to 3.3% drop in average DSC. When n is large, the memory includes older predictions that may not match the current model state, and their disagreement increases the noise in the pseudo supervision and in the confidence estimation. When n is small, the memory provides limited temporal context, and the confidence weight becomes sensitive to short-term fluctuations. Using $n = 3$ provides a stable estimate of consistency while keeping the memory close to the current model. Training cost also increases with n because more stored predictions are involved in the weighting computation. Based on the accuracy and cost trade-off, we use $n = 3$ for all other experiments.

4 Conclusion

We propose PseudoRET, a unified multi-target segmentation framework for fundus images under sparse and disjoint annotations. PseudoRET addresses cross-task annotation incompleteness using task-agnostic pseudo embedded training and a confidence-aware pseudo loss to down-weight unreliable pseudo labels. Experiments on 13 public datasets and a four-dataset relabeled benchmark show consistent improvements across vessel, lesion, and optic disc and cup segmentation. Future work will extend PseudoRET to finer-grained targets and downstream clinical reasoning.

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

References

1. Bajwa, M.N., Singh, G.A.P., Neumeier, W., Malik, M.I., Dengel, A., Ahmed, S.: G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
2. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G.: Robust Vessel Segmentation in Fundus Images. International Journal of Biomedical Imaging (2013). https://doi.org/10.1155/2013/154860
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
5. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al.: Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
6. Feng, W., Ge, Z.: Generalized category discovery under domain shift: A frequency domain perspective. Advances in Neural Information Processing Systems **38**, 111721–111749 (2026)
7. Feng, W., Jiang, Y., Zhou, S., Ge, Z.: Beyond the static world: Continual category discovery under visual drift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25032–25042 (2026)
8. Feng, W., Jiang, Y., Zhou, S., Qi, Z., Xu, Z., Wang, Z., Tang, F., Ge, Z.: Seeing through the shift: Causality-inspired robust generalized category discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17766–17775 (2026)
9. Feng, W., Ju, L., Wang, L., Song, K., Zhao, X., Ge, Z.: Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 623–631 (2023)

10. Feng, W., Zhou, S., Jiang, Y., Ge, Z.: Prism: Progressive robust learning for open-world continual category discovery. In: The Fourteenth International Conference on Learning Representations (2026)
11. Fernandez-Granero, M., Sarmiento, A., Sanchez-Morillo, D., Jiménez, S., Alemany, P., Fondón, I.: Automatic cdr estimation for early glaucoma diagnosis. Journal of healthcare engineering **2017**(1), 5953621 (2017)
12. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
14. 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)
15. Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)
16. 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)
17. 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. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data **9**(1), 291 (2022)
19. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017)
20. Lepetit-Aimon, G., Playout, C., Boucher, M.C., Duval, R., Brent, M.H., Cheriet, F.: Maples-dr: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy. Scientific Data **11**(1), 914 (2024)
21. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
22. Niemeijer, M., Xu, X., Dumitrescu, A.V., Gupta, P., Van Ginneken, B., Folk, J.C., Abramoff, M.D.: Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. IEEE Transactions on medical imaging **30**(11), 1941–1950 (2011)
23. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
24. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
25. Saha, O., Sathish, R., Sheet, D.: Learning with multitask adversaries using weakly labelled data for semantic segmentation in retinal images. In: International Conference on Medical Imaging with Deep Learning. pp. 414–426. PMLR (2019)

26. Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R.: Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 75–82. Springer (2017)
27. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE transactions on medical imaging **23**(4), 501–509 (2004)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1195–1204 (2017)
29. Wang, Z., Lyu, J., Tang, X.: autosmim: automatic superpixel-based masked image modeling for skin lesion segmentation. IEEE Transactions on Medical Imaging **42**(12), 3501–3511 (2023)
30. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
31. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
32. Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., Li, G.: Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. Visual Intelligence **4**(1), 2 (2026)
33. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual international conference of the IEEE engineering in medicine and biology. pp. 3065–3068. IEEE (2010)
34. Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2079–2088 (2019)
35. Zhou, Y., Wang, B., Huang, L., Cui, S., Shao, L.: A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. IEEE transactions on medical imaging **40**(3), 818–828 (2020)
36. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)
37. Zhu, J., Qi, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)