CiLA: Knowledge-Preserving Self-Supervised Adaptation of CLIP for Fundus Imaging

Yijin Huang[1,2], Pujin Cheng[1,3,4], Roger Tam[2], and Xiaoying Tang[1,4*](✉)

[1] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
[2] School of Biomedical Engineering, The University of British Columbia, Vancouver, Canada
[3] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
[4] Jiaxing Research Institute, Southern University of Science and Technology, Jiaxing, China

Abstract. Adapting general-purpose CLIP to medical imaging is often done by continued pre-training on in-domain unlabeled data. However, self-supervised re-pretraining can distort or overwrite the useful knowledge encoded in the original CLIP, leading to representation drift and degraded transfer to downstream tasks. We propose CiLA, a knowledge-preserving framework for in-domain CLIP adaptation on fundus images. CiLA leverages ophthalmic prior knowledge by pre-defining a set of clinically meaningful fundus attributes and obtaining attribute-related pseudo labels from the original CLIP. During self-supervised learning, the adapted model is trained not only with standard contrastive or masked-image objectives, but also to (i) predict the predefined attributes and (ii) maintain consistency with the original CLIP predictions, thereby anchoring training to the desired semantics and mitigating catastrophic forgetting. Since general-purpose CLIP predictions can exhibit dataset-specific bias in medical domains, we further introduce a Dynamic Debias Module that adaptively calibrates pseudo labels to reduce biased guidance during adaptation. CiLA is plug-and-play and can be integrated with multiple self-supervised paradigms. Experiments on diverse retinal datasets and architectures demonstrate that CiLA consistently improves downstream fundus classification performance over direct in-domain self-supervised re-pretraining and prior adaptation baselines. The source code is available online.

Keywords: Lesion attribute learning · Self-supervised learning · CLIP

1 Introduction

Self-supervised learning (SSL) [12,4,5,15,14,32] has become a key approach for medical image analysis, where large-scale expert annotations are expensive and

* Corresponding Author (tangxy@sustech.edu.cn)

often unavailable. By learning transferable representations from unlabeled data and then fine-tuning on limited labels, SSL has shown strong potential in retinal fundus imaging [2,3,16,21] for downstream tasks such as disease classification. Despite their success on natural images, these generic objectives can be suboptimal for fundus images, because they primarily emphasize global appearance and augmentation invariances rather than clinically meaningful evidence [16].

In parallel, vision-language models such as Contrastive Language-Image Pre-training (CLIP) [27] provide a complementary capability: they associate images with text, enabling zero-shot, text-guided recognition without task-specific training. CLIP has been explored for medical imaging via prompting, pseudo-labeling [19,23,25] or concept-injection [22]. However, directly applying a general-purpose CLIP to fundus disease classification can be unreliable when the diagnostic differences are fine-grained and require lesion-level evidence [33]. A common strategy to improve domain suitability is to adapt CLIP by re-pretraining (or continuing pre-training) on in-domain fundus data using SSL. Yet, this adaptation introduces a critical risk: *self-supervised re-pretraining can drift away from the original CLIP knowledge*, weakening the capabilities that are valuable for downstream tasks (i.e., representation drift during domain adaptation).

To address this issue, we propose **C**LIP-**i**nformed **L**esion **A**ttribute Learning (CiLA), a knowledge-preserving framework for in-domain SSL adaptation. CiLA is motivated by a simple principle: when adapting a general-purpose model with SSL, we should anchor the training to the semantics we want to retain. Concretely, leveraging prior ophthalmic knowledge, we pre-define a set of clinically meaningful fundus attributes (e.g., lesion type, location, and severity-related characteristics). We then query the original CLIP to obtain attribute-related predictions that serve as pseudo labels. During SSL, CiLA asks the target encoder to (i) learn to predict these predefined attributes and (ii) maintain consistency with the original CLIP outputs, thereby preserving useful CLIP knowledge while adapting to the fundus domain. This design uses structured attribute supervision as a bridge between generic visual pre-training and fine-grained clinical evidence, while explicitly preventing destructive drift during SSL adaptation.

A remaining challenge is that general-purpose CLIP predictions can be biased or miscalibrated on medical datasets, potentially providing skewed pseudo-label guidance. If such biased pseudo labels are used as anchors, they may constrain adaptation in undesirable directions. Therefore, CiLA further introduces a Dynamic Debiasing Module (DDM) to mitigate bias in pseudo-label assignments by dynamically calibrating and balancing pseudo-label distributions during training [31]. Importantly, CiLA is designed as a plug-and-play component that can be integrated into diverse SSL paradigms, including SimCLR [4], MoCo [5], MAE [14], and fundus-specific SSL such as SSiT [16]. We evaluate CiLA by incorporating it into multiple SSL frameworks and fine-tuning on various fundus datasets for classification. Results demonstrate that CiLA consistently improves downstream performance, yielding an average quadratic kappa improvement of up to 4.73% across five retinal datasets.

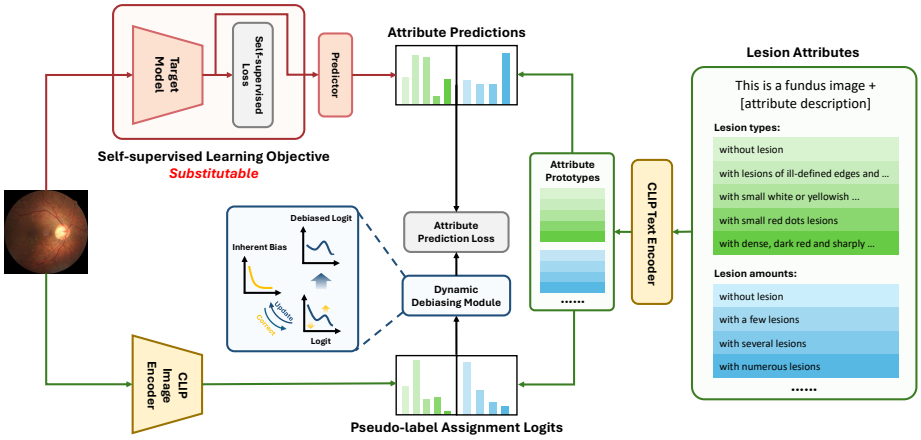

Fig. 1: The proposed CiLA framework.

The primary contributions of this work are: (1) We introduce **CiLA**, a knowledge-preserving in-domain adaptation framework that anchors SSL re-pretraining with CLIP-informed, attribute-specific pseudo labels and explicitly maintains consistency with the original CLIP predictions; (2) We propose a **Dynamic Debiasing Module (DDM)** to mitigate bias in CLIP pseudo labels, enabling more reliable guidance during adaptation; (3) We show that CiLA is **plug-and-play** across multiple SSL paradigms and conduct extensive experiments on five retinal datasets, where CiLA consistently enhances downstream classification performance over strong SSL baselines.

2 Method

2.1 CLIP-informed Lesion Attribute Learning

CiLA is designed for knowledge-preserving in-domain adaptation of CLIP with SSL. When re-pretraining on unlabeled fundus images, standard SSL objectives may drive the representation toward generic invariances and domain statistics, potentially causing drift from the semantics encoded in the original CLIP. CiLA mitigates this issue by introducing an attribute-anchored supervision signal derived from (i) ophthalmic prior knowledge and (ii) the predictions of the original CLIP. Instead of using CLIP to pseudo-label disease grades, which are often subtle and unreliable for a general-purpose model [33], CiLA focuses on lesion attributes described in a structured and fine-grained manner. Such attribute descriptions provide a more explicit bridge between fundus pathology and text, enabling CLIP to produce more meaningful guidance. For example, rather than using a coarse label like "hard exudate", we use a descriptive prompt such as "a fundus image with small white or yellowish lesions with sharp margins" [20], which better reflects the visual evidence required in retinal diagnosis.

Attribute set and CLIP pseudo labels. As illustrated in Fig. 1, we first pre-define an attribute set $\mathcal{A} = \{A_1, \ldots, A_N\}$, where each attribute A_i has C_i categorical states with corresponding textual descriptions $\{t_{i,1}, \ldots, t_{i,C_i}\}$. For instance, the lesion type attribute can include five categories: no lesion, hard exudate, soft exudate, hemorrhage, and microaneurysm [20]. Given a pre-trained CLIP [27] with a frozen image encoder E_I and text encoder E_T, we encode each description into an attribute prototype:

$$P_i = \{p_{i,1}, \ldots, p_{i,C_i}\}, \quad p_{i,j} = E_T(t_{i,j}) \in \mathbb{R}^D, \tag{1}$$

where D is the embedding dimension. For an unlabeled fundus image I, CLIP extracts the image embedding $z = E_I(I) \in \mathbb{R}^D$ and computes the probability (logit-normalized similarity) that I belongs to the j-th category of attribute A_i:

$$a_{i,j}^z = \frac{\exp(\text{sim}(z, p_{i,j}))}{\sum_{k=1}^{C_i} \exp(\text{sim}(z, p_{i,k}))}, \tag{2}$$

where $\text{sim}(\cdot)$ denotes cosine similarity. The pseudo label for attribute A_i is then assigned as $\hat{y}i = \arg\max j \in [C_i] a_{i,j}^z$, indicating the index of the most confident class, and is subsequently converted into a one-hot target vector $\hat{y}_{i,j}$.

Attribute prediction as an anchor during SSL adaptation. Let M denote the target encoder being adapted with SSL, and F be an attribute predictor implemented by a linear projector. For the same image I, we compute the target representation $\tilde{z} = F(M(I)) \in \mathbb{R}^D$ and train the model to predict the CLIP-informed attribute pseudo labels via:

$$\mathcal{L}_{\text{attr}} = -\sum_{i=1}^{N} \sum_{j=1}^{C_i} \hat{y}_{i,j} \log a_{i,j}^{\tilde{z}}. \tag{3}$$

This objective provides structured, clinically grounded supervision while explicitly anchoring the adapted model to the semantic behavior of the original CLIP. The overall training objective combines the employed SSL loss (contrastive [4,5,15] or reconstruction-based [14,32]) with the attribute-anchoring loss:

$$\mathcal{L} = \mathcal{L}_{\text{ssl}} + \lambda \mathcal{L}_{\text{attr}}, \tag{4}$$

where λ balances SSL adaptation and CLIP-informed attribute preservation. In this way, CiLA turns in-domain SSL re-pretraining into a guided adaptation process: the model learns from unlabeled fundus images while maintaining the clinically relevant semantics encoded in the original CLIP, yielding representations that transfer more reliably to downstream retinal diagnosis tasks.

2.2 Dynamic Debiasing Module

CiLA relies on pseudo labels produced by the original CLIP to anchor SSL adaptation. However, CLIP zero-shot predictions are known to be intrinsically

biased: due to inter-class similarity and prompt-induced priors, CLIP tends to over-assign certain categories, leading to persistent label imbalance even when the underlying data distribution is not extremely skewed [31]. When such biased pseudo labels are used as supervision, the bias is amplified in two ways: (i) the target model is trained to imitate the biased assignments, and (ii) the anchoring signal may over-constrain the representation toward majority categories, hurting minority attributes that are often clinically important (e.g., rare lesions). Since pseudo labels serve as anchors to preserve CLIP knowledge, they must be calibrated to avoid anchoring the model to biased semantics.

A probabilistic view. For each attribute A_i with categories $\{1, \ldots, C_i\}$, CLIP assigns pseudo-label probabilities by a softmax over similarities,

$$a_{i,j}^{z} = \frac{\exp(\mathrm{sim}(z, p_{i,j}))}{\sum_{k=1}^{C_i} \exp(\mathrm{sim}(z, p_{i,k}))}, \tag{5}$$

which can be interpreted as a log-linear model with logits $\ell_{i,j} = \mathrm{sim}(z, p_{i,j})$. If CLIP exhibits an implicit class prior (or preference) $\pi_{i,j}$ for category j (caused by training data bias, prompt bias, and intrinsic prototype similarity), then the predicted distribution can be viewed as proportional to:

$$p_{\mathrm{CLIP}}(y = j \mid z, A_i) \;\propto\; \exp(\ell_{i,j})\, \pi_{i,j}. \tag{6}$$

In this view, bias corresponds to a non-uniform prior $\pi_{i,j}$ that systematically increases the probability of certain categories, independent of the actual evidence in z. A natural correction is to estimate this prior and divide it out in the logit space, analogous to prior-adjusted (balanced) softmax and debiased pseudo labeling [31]. This motivates subtracting $\log \pi_{i,j}$ from the logits.

Estimating bias via running assignment statistics. Since the true prior $\pi_{i,j}$ is unknown, we estimate it online using CLIP's assignment statistics. For a mini-batch $\{I_b\}_{b=1}^{B}$, let $z_b = E_I(I_b)$ be CLIP image embeddings and $a_{i,j}^{z_b}$ be the corresponding assignment probabilities for attribute A_i. We compute the batch-wise mean assignment

$$s_{i,j} = \frac{1}{B} \sum_{b=1}^{B} a_{i,j}^{z_b}, \tag{7}$$

and maintain an exponential moving average (EMA) to obtain a stable estimate of the assignment tendency:

$$\hat{s}_{i,j} \leftarrow \alpha \hat{s}_{i,j} + (1 - \alpha) s_{i,j}, \quad \alpha \in (0, 1], \tag{8}$$

where $\hat{s}_{i,j}$ is initialized uniformly (or as zeros and warmed up), and α controls the temporal smoothing. We interpret $\hat{s}_{i,j}$ as an online estimator of CLIP's implicit prior $\pi_{i,j}$ for attribute A_i (up to a proportionality constant), capturing systematic over-/under-assignment [31].

Prior-corrected (debiased) assignment. We then form a debiased pseudo-label distribution by subtracting a scaled log-prior term from the logits:

$$\tilde{\ell}_{i,j} = \mathrm{sim}(z, p_{i,j}) - \beta \log(\hat{s}_{i,j} + \epsilon), \tag{9}$$

where $\beta \geq 0$ controls the strength of debiasing and ϵ is a small constant for numerical stability. The debiased assignment probability is

$$\hat{a}^z_{i,j} = \frac{\exp(\tilde{\ell}_{i,j})}{\sum_{k=1}^{C_i} \exp(\tilde{\ell}_{i,k})} = \frac{\exp(\mathrm{sim}(z, p_{i,j}))(\hat{s}_{i,j} + \epsilon)^{-\beta}}{\sum_{k=1}^{C_i} \exp(\mathrm{sim}(z, p_{i,k}))(\hat{s}_{i,k} + \epsilon)^{-\beta}}. \tag{10}$$

Eq. (10) makes the mechanism explicit: categories that CLIP over-predicts (large $\hat{s}_{i,j}$) are down-weighted by $(\hat{s}_{i,j})^{-\beta}$, while under-predicted categories are relatively up-weighted. The pseudo label for attribute A_i is then obtained as $\hat{y}_i = \arg\max_{j \in [C_i]} \hat{a}^z_{i,j}$ (or using $\hat{a}^z_{i,j}$ directly as a soft target).

3 Experiments

3.1 Datasets

The EyePACS [10] dataset is used for pre-training in this work, comprising 88,702 fundus images with annotations removed. We evaluate the pre-trained models on five retinal disease classification datasets: Messidor-2 [8] (DR), IChallenge-AMD [11] (AMD), Retinal [26] (RD), ODIR5k [1] (ODIR), and MPOS [30], covering over eight retinal diseases. For ODIR, we removed multi-label samples to align with the evaluation setting. These datasets vary in complexity, with the number of disease categories ranging from 2 to 8 and sample sizes between 400 and 6,392. We use official dataset splits when available. Otherwise, we employ a random partition of 70%/10%/20% for training, validation, and testing, respectively.

3.2 Experiment Setup

Pre-training setup. In all experiments, we adopt Vision Transformer [9] (ViT-B) as the backbone model. A commonly used pre-trained CLIP model [28] with a ViT-B backbone from OpenCLIP [17] is employed in CiLA, and the image encoder of the pre-trained CLIP is used to initialize the target model across all SSL methods for a fair comparison purpose. We implement SSL methods using the Solo-Learn library [7]. All models are trained for 300 epochs with a batch size of 512, except for MAE, which is pre-trained for 400 epochs following the original paper's recommendation [14] for improved performance. For CiLA-specific settings, we pre-define lesion type, size, amount, and distribution in the attribute set. Please refer to the released code for detailed descriptions of each attribute. For hyper-parameters, we set $\lambda = 1$, $\alpha = 0.999$, and $\beta = 1$. All other SSL hyper-parameters remain unchanged, whether with or without CiLA, ensuring a controlled experimental setup.

Table 1: Retinal disease classification performance using different SSL methods. †SSiT is an SSL method specifically designed for fundus imaging.

Pre-training methods	Retinal datasets					Average κ	$\triangle$ κ
	DR	AMD	RD	ODIR	MPOS		
Fine-tuning evaluation							
MoCo [5]	68.26	78.40	54.23	45.09	67.07	62.61	-
MoCo + CiLA	**70.18**	**79.81**	**56.61**	**46.82**	**78.04**	**66.29**	**+3.68**
SimCLR [4]	71.14	76.92	55.87	46.13	72.34	64.48	-
SimCLR + CiLA	**73.48**	**78.76**	**60.11**	**51.61**	**80.43**	**68.88**	**+4.40**
MAE [14]	65.89	74.92	45.28	41.15	65.54	58.56	-
MAE + CiLA	**67.53**	**80.47**	**52.09**	**41.88**	**74.48**	**63.29**	**+4.73**
SSiT† [16]	71.97	76.72	56.35	45.15	77.67	65.57	-
SSiT† + CiLA	**74.30**	**82.39**	**60.05**	**46.21**	**78.92**	**68.37**	**+2.80**
Linear evaluation							
MoCo [5]	57.95	69.49	31.82	37.40	67.27	52.79	-
MoCo + CiLA	**60.23**	**70.33**	**37.12**	**39.67**	**73.88**	**56.25**	**+3.46**
SimCLR [4]	63.42	64.72	37.19	36.14	66.20	53.53	-
SimCLR + CiLA	**66.81**	**65.50**	**43.29**	**38.54**	**74.81**	**57.79**	**+4.26**
MAE [14]	19.97	36.98	2.83	27.02	10.91	19.54	-
MAE + CiLA	**23.27**	**50.63**	**3.01**	**30.06**	**11.21**	**23.64**	**+4.10**
SSiT† [16]	64.82	71.03	43.73	41.42	65.51	56.90	-
SSiT† + CiLA	**65.81**	**72.65**	**44.46**	**43.11**	**66.25**	**58.46**	**+1.56**

Evaluation setup. Following previous works [4,5], we adopt two common evaluation protocols: fine-tuning evaluation and linear evaluation. In fine-tuning evaluation, the model used for the downstream task is initialized with the pre-trained parameters and then fully trained under a supervised learning setup. In linear evaluation, the training procedure remains the same, except that all parameters of the pre-trained model are frozen, and only a linear classification head is trained for the downstream task. Keeping in line with previous works [13,16], we adopt quadratically weighted kappa score κ [6] to evaluate the classification performance. All downstream tasks are trained for 20 epochs with a batch size of 16. To ensure reliability, grid search is performed on the learning rate and weight decay, and we report the best performance for each pre-trained model. All experiments are repeated five times, and the mean performance is reported.

3.3 CiLA Integration with Different SSL Methods

We evaluate CiLA as a plug-and-play knowledge-preserving adaptation module by integrating it with representative SSL methods, including MoCo-v3, SimCLR, MAE, and SSiT (a fundus-specific method). Table 1 reports downstream classification results on five retinal datasets under fine-tuning and linear evaluation.

In the fine-tuning evaluation, CiLA consistently improves performance across all SSL methods. Integrating CiLA with MoCo and SimCLR yields average kappa

Table 2: Fine-tuning evaluation with different components.

DD	LAD	DDM	Average κ
			64.48
✓			65.32
✓		✓	65.91
	✓		67.58
	✓	✓	**68.88**

Table 3: Fine-tuning evaluation with different attribute compositions.

Type	Size	Amount	Distribution	Avg. κ
	✓	✓	✓	66.60
✓		✓	✓	67.33
✓	✓		✓	67.53
✓	✓	✓		68.60
✓	✓	✓	✓	**68.88**

gains of 3.68% and 4.40%, respectively, demonstrating that anchoring SSL adaptation with CLIP-informed attributes enhances contrastive pre-training. Notably, MAE integrated with CiLA achieves a 4.73% improvement, indicating that CiLA also benefits reconstruction-based SSL by preventing representation drift during in-domain adaptation. For SSiT, which already incorporates fundus-specific design, CiLA still provides a 2.80% average kappa increase, suggesting that attribute-anchored guidance complements specialized SSL objectives. CiLA also improves performance in the linear evaluation setting. Across all SSL methods, CiLA leads to consistent gains, with SimCLR+CiLA achieving the largest improvement of 4.26%. For SSiT, CiLA adds a further 1.56% boost, implying that the proposed attribute anchoring encourages more transferable and discriminative representations even without end-to-end fine-tuning.

3.4 Ablation Study

Impact of components. We analyze CiLA components based on SimCLR. Table 2 reports different configurations, where DD and LAD denote Disease Description and Lesion Attribute Description, respectively. DD replaces attribute-level descriptions with disease-level descriptions covering all diseases in the considered datasets. The baseline (SimCLR) achieves 64.48% average κ. Using DD yields only a modest gain (+0.84%), indicating that coarse disease prompts provide weak supervision. Adding DDM to DD further improves by +0.59%, suggesting debiasing helps but cannot fully address noisy disease-level pseudo labels. Replacing DD with LAD increases performance to 67.58%, confirming that structured lesion attributes provide stronger guidance during adaptation.

Impact of attribute description. We further evaluate the contribution of individual lesion attributes by removing specific attribute descriptions. We consider four attributes (type, size, amount, and distribution) and report fine-tuning results in Table 3. Removing lesion type causes the largest drop (from 68.88% to 66.60%), highlighting its central role in providing diagnostic anchoring signals. Removing distribution has the smallest effect, suggesting it is complementary but less influential than type, size, or amount in our setting. Using all four attributes yields the highest κ (68.88%), indicating that combining diverse attributes provides the most effective supervision.

4 Conclusion

In this work, we proposed **C**LIP-**i**nformed **L**esion **A**ttribute Learning (CiLA), a knowledge-preserving framework for in-domain self-supervised adaptation in fundus imaging. CiLA mitigates representation drift in SSL re-pretraining by anchoring training with CLIP-informed, attribute-specific pseudo labels derived from clinically meaningful lesion attributes. To address bias in CLIP pseudo labels [31], we introduced a Dynamic Debiasing Module that calibrates assignments and provides more reliable guidance. CiLA is plug-and-play across diverse SSL paradigms and consistently improves downstream retinal disease classification under both fine-tuning and linear evaluation on five datasets. Future work will extend the evaluation of CiLA to newer vision-language models [29,24,18], enabling a more comprehensive assessment of its generalization ability.

Acknowledgments. This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the National Natural Science Foundation of China (T2422012); the Guangdong Basic and Applied Basic Research (2024B1515020088); the High Level of Special Funds (G030230001, G03034K003); the Guangdong Key Research and Development Program (2025B1111080001); the SUSTech Fang Keng Faculty Award.

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

References

1. Odir5k. https://odir2019.grand-challenge.org/ (2015), odir-2019 competition
2. Cai, Z., Lin, L., He, H., Cheng, P., Tang, X.: Uni4eye++: A general masked image modeling multi-modal pre-training framework for ophthalmic image classification and segmentation. IEEE Transactions on Medical Imaging (2024)
3. Chen, H., Wang, R., Wang, X., Li, J., Fang, Q., Li, H., Bai, J., Peng, Q., Meng, D., Wang, L.: Unsupervised local discrimination for medical images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
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. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
6. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
7. da Costa, V.G.T., Fini, E., Nabi, M., Sebe, N., Ricci, E.: solo-learn: A library of self-supervised methods for visual representation learning. Journal of Machine Learning Research **23**(56), 1–6 (2022), http://jmlr.org/papers/v23/21-1155.html
8. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J.R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology pp. 231–234 (2014)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection (2015), kaggle
11. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., et al.: Adam challenge: Detecting age-related macular degeneration from fundus images. IEEE transactions on medical imaging **41**(10), 2828–2847 (2022)
12. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9052–9071 (2024). https://doi.org/10.1109/TPAMI.2024.3415112
13. He, A., Li, T., Li, N., Wang, K., Fu, H.: Cabnet: Category attention block for imbalanced diabetic retinopathy grading. IEEE Transactions on Medical Imaging **40**(1), 143–153 (2020)
14. 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)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
16. Huang, Y., Lyu, J., Cheng, P., Tam, R., Tang, X.: Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading. IEEE Journal of Biomedical and Health Informatics (2024)
17. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773
18. Isztl, D., Spitznagel, T., Somfai, G.M., Santos, R.: When do domain-specific foundation models justify their cost? a systematic evaluation across retinal imaging tasks. arXiv preprint arXiv:2511.22001 (2025)
19. Lai, Z., Vesdapunt, N., Zhou, N., Wu, J., Huynh, C.P., Li, X., Fu, K.K., Chuah, C.N.: Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16155–16165 (2023)
20. Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., Fu, H.: Applications of deep learning in fundus images: A review. Medical Image Analysis **69**, 101971 (2021)
21. Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L.: Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Transactions on Medical Imaging **39**(12), 4023–4033 (2020)
22. Mehta, D., Jiang, Y., Jan, C., He, M., Jadhav, K., Ge, Z.: Interpretable few-shot retinal disease diagnosis with concept-guided prompting of vision-language models. In: International Conference on Information Processing in Medical Imaging. pp. 263–277. Springer (2025)
23. Menghini, C., Delworth, A., Bach, S.: Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. Advances in Neural Information Processing Systems **36**, 60984–61007 (2023)
24. MH Nguyen, D., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al.: Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. Advances in Neural Information Processing Systems **36**, 27922–27950 (2023)

25. Mo, S., Kim, M., Lee, K., Shin, J.: S-clip: Semi-supervised vision-language learning using few specialist captions. Advances in Neural Information Processing Systems **36** (2024)
26. Padia, N., Hirpara, A., Jani, D., Patel, S.: Cataract dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset (2015), kaggle
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Samir Yitzhak Gadre, G.I., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
29. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
30. Wang, H., Xing, Z., Wu, W., Yang, Y., Tang, Q., Zhang, M., Xu, Y., Zhu, L.: Non-invasive to invasive: Enhancing ffa synthesis from cfp with a benchmark dataset and a novel network. In: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. pp. 7–15 (2024)
31. Wang, X., Wu, Z., Lian, L., Yu, S.X.: Debiased learning from naturally imbalanced pseudo-labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14647–14657 (2022)
32. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
33. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., Shen, D.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)