RetFiner: A Vision-Language Refinement Scheme for Retinal Foundation Models

Ronald Fecso[1,3(✉)][0009−0006−2423−7071], José Morano[1,2,3][0000−0003−3785−8185], Ursula Schmidt-Erfurth[4][0000−0002−7788−7311], and Hrvoje Bogunović[1,2,3][0000−0002−9168−0894]

[1] Institute of Artificial Intelligence, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria
[2] Christian Doppler Lab for Artificial Intelligence in Retina, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria
[3] Comprehensive Center for AI in Medicine, Medical University of Vienna, Austria
[4] OPTIMA Lab, Dept. of Ophthalmology, Medical University of Vienna, Austria
{ronald.fecso,jose.moranosanchez,hrvoje.bogunovic}@meduniwien.ac.at

Abstract. The rise of imaging techniques such as optical coherence tomography (OCT) and advances in deep learning (DL) have enabled clinicians and researchers to streamline the assessment of retinal diseases. A popular DL approach is self-supervised learning (SSL), where models learn from vast amounts of unlabeled data, avoiding costly annotation. SSL has allowed the development of foundation models (FMs), large models that can be used for a variety of downstream tasks. However, existing FMs for OCT, trained solely on image data, lack a comprehensive and robust semantic understanding of images, as evidenced by their downstream performance (especially for complex tasks), and thus require supervised fine-tuning (which may be unfeasible) to better adapt to specific applications and populations. To address this, we propose RetFiner, an SSL vision-language refinement scheme that improves the representations of existing FMs and enables their efficient and direct adaptation to specific populations for improved downstream performance. Our method uses a diverse set of training objectives which take advantage of the rich supervisory signal found in textual data. We tested RetFiner on the retinal FMs RETFound, UrFound, and VisionFM, showing significant improvements in linear probing performance on seven highly diverse OCT classification tasks, with an average increase of 5.7, 3.9, and 2.1 percentage points over their baselines, respectively.

Keywords: Vision-language modeling · Foundation models · OCT

1 Introduction

Ocular and systemic diseases affecting the eye represent an important health concern. Optical coherence tomography (OCT) has become the standard imaging technique to assess and diagnose several retinal diseases such as age-related macular degeneration (AMD) [12]. With the emergence of OCT and other advanced

imaging modalities, medical artificial intelligence (AI) offers great potential to accelerate the diagnostic process [32]. However, traditional AI methods, mostly based on deep learning (DL), rely on large amounts of labeled data, which requires costly manual annotation. Recently, self-supervised learning (SSL) has gained popularity because it allows models to learn meaningful features from unlabeled data [6]. The combination of SSL techniques and large datasets and DL architectures has enabled the development of *foundation models* (FMs) [2], generalizable models that can be efficiently adapted to several applications.

A common SSL approach is masked modeling (MM), which randomly masks part of the input and tasks the model with reconstructing the missing data, thus learning meaningful data representations. The most common MM method for images is Masked Autoencoding (MAE) [10], based on Vision Transformer (ViT) [5]. RETFound [32] applied MAE to develop separate FMs for retinal OCT and fundus images, demonstrating strong performance on diagnostic tasks. UrFound [30] trained an FM using joint MAE and masked language modeling (MLM). Despite the promising results, some studies [7,1] have shown that MAE-based methods produce suboptimal representations for perceptual, highly semantic tasks. In contrast, VisionFM [20] used a self-distillation approach, which consists of feeding two different image views to two encoders, and mapping one to the other using a predictor, to develop separate FMs for 8 ophthalmic modalities, resulting in improved performance over RETFound.

Regardless of the approach, existing FMs for OCT are still limited by the relatively small size of the pretraining dataset compared to FMs for computer vision [19,32,20], which leads to data bias, lower generalizability and, in some cases, the need for supervised tuning to specific populations or applications.

Another popular SSL approach is contrastive language-image pretraining (CLIP) [21], which consists of training a vision-language model (VLM) by aligning visual and textual representations from different encoders using an image-text contrastive (ITC) loss. CLIP has gained popularity in the medical field due to the common availability of paired image and Electronic Health Record (EHR) data [24,31]. FILIP [28] proposes a VLM which uses fine-grained contrastive learning with a cross-modal late interaction mechanism, but has no specific language modeling losses. FLAIR [24] performed CLIP on 37 classification datasets of color fundus images by converting class labels into descriptions. However, this approach is not readily adaptable to unstructured data (e.g., EHRs) as it requires non-trivial decisions about how to create class labels and their descriptions. Moreover, CLIP alone is suboptimal for medical data because there exists high semantic overlap (e.g., patients with the same diseases or biomarkers) [23]. This causes unpaired examples to be pushed apart in the embedding space regardless of their semantic similarity, resulting in false negatives [16]. Also, CLIP-based approaches usually struggle to distinguish subtle pathological patterns in medical images, as they rely only on global features [11].

To mitigate these issues, other works [15,29,3] have proposed to combine CLIP-like ITC losses with MM and image–text matching (ITM) objectives. ALBEF [15] trains a 3-encoder network (image, text, and multimodal) using ITC,

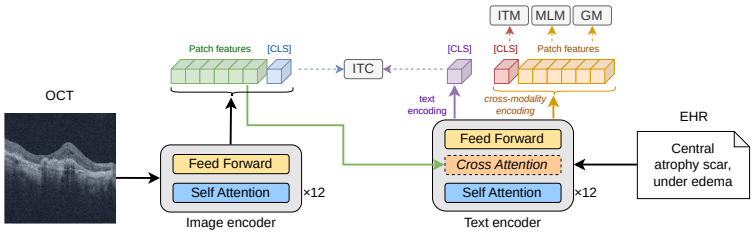

Fig. 1. RetFiner method. Cross-attention layers are activated only during the forward passes for ITM, MLM, and GM. An example of an OCT image and report is shown.

ITM and MLM losses and a momentum model. BLIP-2 [14] proposes a unified VLM by tuning specialized modules that interact with frozen, separately trained vision and text encoders. In the medical domain, PTUnifier [3] used a similar setting but with a single model by unifying text and image inputs via prompts. CoCa [29] achieved state-of-the-art (SOTA) zero-shot classification performance of natural images by jointly training a VLM on contrastive and captioning tasks. Despite strong downstream performance, the use of these advanced approaches to develop or improve retinal FMs has not yet been explored.

Contribution. In this work, we present RetFiner (Fig. 1), an efficient vision-language refinement scheme for retinal foundation models. Our approach consists of training a VLM composed of a vision encoder based on an arbitrary retinal FM and a separate language model using a set of diverse training objectives focused on exploiting the EHRs as supervisory signals for improving visual representations. To validate our method, we refined the retinal FMs RETFound [32], UrFound [30], and VisionFM [20] with our scheme using an in-house dataset of 100k pairs of OCTs and associated EHRs. Running RetFiner on this dataset for an FM requires less than 10 epochs. Linear probing of the refined vision encoders on six public and one in-house OCT classification datasets demonstrates the effectiveness of the proposed approach for both improving the semantic understanding of the models and adapting them to targeted populations. Our RetFiner models set a new benchmark for OCT image analysis, with potential applications where high-level semantics are required, such as visual question answering. To facilitate research progress, we have published the code and model weights at https://github.com/ronnief1/RetFiner.

2 Methods

RetFiner (Fig. 1) employs a simple architecture comprised of a ViT [5] vision encoder and a Transformer text encoder [26], allowing for single-modality and cross-modality embedding. Cross-attention (CA) layers are added between the self-attention and feed-forward layers of the text encoder. They are utilized for cross-modality encoding and generation, but deactivated for uni-modal encoding. Thus, the text encoder acts as both a uni- and cross-modal encoder and decoder.

The model is trained to optimize four losses, as shown in Fig. 1: image-text contrastive (ITC), image-text matching (ITM), masked language modeling (MLM), and generative modeling (GM) (captioning). Such a combination effectively enhances the model's cross-modality alignment, understanding, and generation capabilities, ultimately improving visual representations. The total loss is a direct sum of all the losses equally weighted:

$$\mathcal{L} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{GM}}. \tag{1}$$

To efficiently use the model for downstream classification tasks, we propose a novel feature pooling strategy that integrates both global and local features. In particular, we concatenate the CLS token and the average pool of the patch tokens. Then, these features are fed into a trainable linear layer.

2.1 Training Objectives

ITC. We employ an ITC loss to align the vision and text encoders in the embedding space in order to learn better cross-modal representations [3]. We use the InfoNCE loss from [18] to bring pairs of images and texts together in the embedding space while pushing apart negative pairs.

ITM. We use an ITM loss along with ITC to further align the image and text encoders and learn more powerful, discriminative multi-modal features [3]. To challenge the model during training, we use hard negative mining, inspired by [15]. In this scheme, for a given image, the model must predict whether another report from the same batch is part of the same pair. The candidate report is sampled from other reports in the batch, where reports with higher cosine similarity in the embedding space to the given image have a higher chance of being sampled. This loss is also calculated with images and reports swapping places. Such a loss forces the model to learn to differentiate highly similar samples and in turn better model fine-grained image-text interactions.

MLM. Despite the success of instance-level tasks such as ITC and ITM, they may be limited by the semantic overlap in medical images [31]. To address this, we include two reconstruction tasks: MLM [4] and GM. These non-contrastive tasks act as regularizers for ITC and ITM, while improving the model's semantic understanding and generation. [15] demonstrate MLM's value in improving language understanding from a mutual information perspective. Specifically, MLM aims to predict randomly masked input tokens from reports in a bidirectional manner using Cross-Entropy (CE) loss.

GM. To supplement the MLM task and enhance the model's reconstruction and generation abilities, we add a generative text modeling task. Given the context of an OCT and previous report tokens, the model is trained to auto-regressively predict the next masked report token using CE as loss function. Such a task improves language understanding and thus downstream performance [29].

2.2 Development Data

For model development, we used an in-house dataset of 100k paired OCT images and EHRs, from which we extracted the text describing the OCT scan. The

reports cover a range of retinal conditions such as cataracts, choroidal neovascularization (CNV), age-related macular degeneration (AMD), retinal vein occlusion (RVO), and glaucoma. In addition, we used an extra 160k images with no EHRs to train a MAE baseline for the ablation studies. All scans were collected between 2007 and 2021 at the Medical University of Vienna, Austria. Images were taken with Cirrus and Spectralis devices. In line with previous work [32], only the central B-scans of the 3D OCT volumes were used in this study.

2.3 Implementation Details

The pipeline is implemented in Python 3.10 using PyTorch and trained on a NVIDIA A100 GPU (80GB). The AdamW optimizer was used with a learning rate of 10^{-4} and a batch size of 128. Early stopping was triggered when the validation loss did not decrease for three epochs. The vision encoder is based on the ViT architecture [5] and text encoder is a pre-trained BERT [4]. During the forward pass for the ITC loss, cross-attention was turned off in the text encoder. The CLS tokens from each encoder were projected down to a dimension of 512 with a linear layer, then L2-normalized before being passed into the ITC loss. For the remaining losses, cross-attention was activated between the self-attention and feed forward layers of each block. The patch tokens from the vision encoder were passed into the text encoder as the hidden states to perform cross-attention. For MLM, 15% of the report tokens are randomly masked, as in BERT [4]. For the GM loss, 60% of the report tokens are masked with causal attention masks.

3 Experiments and Results

Experimental Setup. To validate our approach, we applied RetFiner on our OCT–text dataset (100k) to three SOTA FMs: RETFound [32], UrFound [30], and VisionFM [20], resulting in RetFiner-[R,U,V] refined models. In addition, for the ablation studies, to ensure that the efficacy of our approach is not derived from our imaging data alone and to discard data bias, we used the full dataset (260k OCTs) to pretrain a ViT-Base model using MAE [10], as in [32], and then applied RetFiner for comparison. The performance of the refined models was then compared with that of the out-of-the-box models and two SOTA general-purpose vision models: CLIP [21] and DINOv2 [19]. Models were evaluated via linear probing with our concatenation pooling strategy on seven retinal disease classification datasets: OCTDL [13], OCTID [9], GAMMA [27], Harvard Glaucoma [17], NEHUT [25], Noor Eye Hospital [22], and an in-house dataset [8]. These datasets cover a range of demographics, devices, and diseases, including AMD, glaucoma, diabetic retinopathy and diabetic macular edema. With linear probing, all weights of the model except for the final linear layer remain frozen. This evaluation is less dependent on downstream optimization than full fine-tuning, and better reflects how discriminative the extracted features are.

Each experiment was run five times with different seeds. Models were evaluated using balanced accuracy (BAcc), area under receiver operating characteristic curve (AUROC), average prevision (AP), and, for ablation, also F1.

Table 1. Average linear probing performance over all downstream datasets. We compare the best metric out of all the models (bolded) with the best metric out of the base models (underlined) to measure if there was a statistically significant difference using the Wilcoxon signed-rank test (**: $p < 0.01$, ***: $p < 0.001$). Values in parentheses represent change in performance compared to their baseline counterpart.

Model	BAcc (%)	AUROC (%)	AP (%)
CLIP	73.6±12.3	90.6±8.6	84.5±9.9
DINOv2	74.5±12.5	90.4±9.7	84.7±11.7
RETFound	78.1±11.6	<u>92.9</u>±7.7	88.6±9.5
UrFound	78.6±11.1	92.7±8.5	88.5±9.5
VisionFM	<u>81.4</u>±13.4	92.8±9.1	<u>89.3</u>±10.7
RetFiner-R	**83.8*****±12.0 (+5.7)	94.6±6.7 (+1.7)	91.2±9.5 (+2.6)
RetFiner-U	82.5±12.3 (+3.9)	**94.7****±7.7 (+2.0)	**92.3*****±8.4 (+3.8)
RetFiner-V	83.5±12.2 (+2.1)	94.4±7.5 (+1.6)	91.2±10.2 (+1.9)

State-of-the-art Comparison. Table 1 shows the average performance of our method and the SOTA retinal and general-purpose FMs across the different tasks. Per-dataset performances are shown in Table 2. As shown in the tables, the top three performances across all metrics come from models refined using our proposed approach, with RetFiner-R and RetFiner-U performing the best in terms of BAcc and AUROC and AP, respectively.

Improvement Analysis. Tables 1 and 2 show a significant improvement in downstream classification performance for RetFiner models compared to their off-the-shelf counterparts. These results demonstrate the effectiveness of our approach for improving existing FMs. This is more remarkable considering our method requires less than ten epochs to refine a model. Also importantly, our method significantly improves all retinal FMs on our complex in-house dataset (with 9 classes), which represents pathologies of high clinical relevance in general and in our clinic in particular. This demonstrates our methods's ability to mitigate the data bias found in retinal FMs while also leveraging their powerful representations, allowing us to create a model suitable for our applications using in-house data, with no need for manual annotation or data processing.

Baseline Comparison and Ablation Study. We compared our method to existing VLM approaches and tested the effect of the different losses on downstream performance to demonstrate their positive effect. In both cases, we used our MAE-pretrained model as the base model and passed it through our scheme using the loss combinations listed in Table 3. CLIP and UrFound baselines are analogous to training using only ITC and MLM losses, respectively. The resulting models were evaluated by linear probing on our in-house dataset. As shown in Table 3, all losses result in significantly higher downstream performance. We also tested the effect of our token pooling strategy for linear probing. During refinement, the ITC loss uses only the CLS token, while the other losses use only the patch tokens. We asserted that using both led to a better exploitation

Table 2. Linear probing performance on downstream datasets. We compare SOTA FMs for OCT with our RetFiner-refined versions of them. Performance differences are shown in parentheses. For each metric, we compare the best overall result (bold) with the best result of the base models (underlined) using Student's t-test (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). The number of test cases (n) and classes (C) are also listed.

Dataset	Model	BAcc (%)	AUROC (%)	AP (%)
In-house $n = 640$ staging & diagnosis $C = 9$	RETFound	75.3±2.3	98.7±0.2	92.0±0.7
	UrFound	76.3±0.5	98.6±0.1	92.5±0.4
	VisionFM	80.1±1.4	98.3±0.1	91.5±0.5
	RetFiner-R	**84.3*****±1.1 (+9.0)	**99.1*****±0.1 (+0.4)	**95.2*****±0.1 (+3.2)
	RetFiner-U	81.9±1.2 (+5.6)	99.0±0.1 (+0.4)	94.2±0.3 (+1.7)
	RetFiner-V	82.2±1.0 (+2.1)	98.4±0.1 (+0.1)	93.0±0.5 (+1.5)
GAMMA $n = 20$ staging & diagnosis $C = 3$	RETFound	54.7±9.3	80.2±0.5	69.3±2.5
	UrFound	57.0±3.9	82.3±0.7	72.0±3.3
	VisionFM	53.6±6.0	77.0±4.2	68.1±5.1
	RetFiner-R	58.9±7.5 (+4.2)	84.6±1.9 (+4.4)	71.8±3.9 (+2.6)
	RetFiner-U	**60.0**±6.1 (+3.0)	**88.2***** ±0.9 (+5.9)	**80.1*****±1.7 (+8.1)
	RetFiner-V	59.4±2.9 (+5.8)	82.2±2.2 (+5.2)	70.4±3.6 (+2.3)
Harvard Glaucoma $n = 400$ diagnosis $C = 2$	RETFound	74.6±1.8	82.0±0.8	80.8±0.5
	UrFound	71.2±1.4	77.3±0.6	76.6±0.7
	VisionFM	74.5±0.7	81.2±0.6	80.7±0.6
	RetFiner-R	**77.7****±0.6 (+3.1)	**83.8****±0.4 (+1.8)	83.1±0.4 (+2.3)
	RetFiner-U	70.1±0.6 (–1.1)	78.6±0.4 (+1.3)	79.0±0.3 (+2.4)
	RetFiner-V	74.5±1.3 (0.0)	83.4±0.8 (+2.2)	**83.3*****±0.9 (+2.6)
NEHUT $n = 135$ diagnosis $C = 3$	RETFound	84.7±1.2	95.2±0.5	91.9±0.8
	UrFound	84.8±1.3	95.2±0.4	90.9±1.1
	VisionFM	88.2±1.4	95.3±0.7	91.8±1.5
	RetFiner-R	**89.5***±0.6 (+4.8)	97.6±0.2 (+2.4)	95.6±0.3 (+3.7)
	RetFiner-U	89.5±0.6 (+4.7)	**97.7*****±0.2 (+2.5)	**96.0*****±0.5 (+5.1)
	RetFiner-V	88.0±0.8 (–0.2)	97.5±0.1 (+2.2)	95.5±0.3 (+3.7)
Noor Eye Hospital $n = 30$ diagnosis $C = 3$	RETFound	88.0±3.8	97.6±0.2	96.2±0.4
	UrFound	88.0±5.1	97.6±0.4	96.0±0.6
	VisionFM	94.0±2.8	98.8±0.1	98.0±0.1
	RetFiner-R	**95.3**±1.8 (+7.3)	97.7±0.3 (+0.1)	96.5±0.4 (+0.3)
	RetFiner-U	92.0±1.8 (+4.0)	99.8±0.2 (+2.2)	99.6±0.4 (+3.6)
	RetFiner-V	93.3±0.0 (–0.7)	**99.8*****±0.1 (+1.0)	**99.6*****±0.1 (+1.6)
OCTDL $n = 332$ diagnosis $C = 7$	RETFound	80.3±1.2	97.9±0.3	93.6±0.6
	UrFound	84.4±0.9	99.0±0.0	95.4±0.2
	VisionFM	87.6±3.4	99.2±0.1	96.5±0.6
	RetFiner-R	87.9±2.0 (+7.6)	99.5±0.0 (+1.6)	97.1±0.2 (+3.5)
	RetFiner-U	90.6±1.3 (+6.2)	**99.5*****±0.0 (+0.5)	**98.4*****±0.1 (+3.0)
	RetFiner-V	**90.9***±1.5 (+3.3)	99.4±0.2 (+0.2)	97.7±0.3 (+1.2)
OCTID $n = 174$ diagnosis $C = 5$	RETFound	88.8±2.0	99.0±0.1	96.7±0.2
	UrFound	88.7±2.5	98.9±0.2	96.0±0.5
	VisionFM	91.8±0.9	99.7±0.1	98.6±0.2
	RetFiner-R	93.2±1.8 (+4.4)	99.7±0.1 (+0.7)	98.9±0.3 (+2.2)
	RetFiner-U	93.4±1.3 (+4.7)	99.8±0.1 (+0.9)	99.0±0.3 (+3.0)
	RetFiner-V	**96.3*****±1.1 (+4.5)	**99.8*****±0.0 (+0.1)	**99.2*****±0.1 (+0.6)

Table 3. Linear probing performance of combinations of losses on our in-house dataset.

Losses	BAcc (%)	AUROC (%)	AP (%)	F1-score (%)
ITC (∼ CLIP [21,24])	78.4±0.7	98.4±0.1	91.6±0.4	85.6±1.0
MLM (∼ UrFound [30])	77.0±1.5	98.4±0.1	92.4±0.6	84.8±0.7
ITC+GM	80.0±0.8	98.6±0.1	93.4±0.1	86.5±0.3
ITC+GM+MLM	80.1±0.4	98.6±0.1	92.7±0.1	86.7±0.6
ITC+GM+MLM+ITM	**82.0*****±0.7	**98.7****±0.1	**93.5**±0.1	**88.2*****±0.3

Table 4. Linear probing performance of each pooling strategy on our in-house dataset.

Model	BAcc (%)	AUROC (%)	AP (%)	F1-score (%)
CLS token	79.2±1.3	98.5±0.1	92.6±0.6	86.4±1.3
Patch features	79.3±0.2	98.8±0.0	93.3±0.1	87.3±0.3
All tokens	79.5±0.9	**98.8**±0.0	93.3±0.1	87.5±0.4
Concatenation	**82.0*****±0.7	98.7±0.1	**93.5****±0.1	**88.2****±0.3

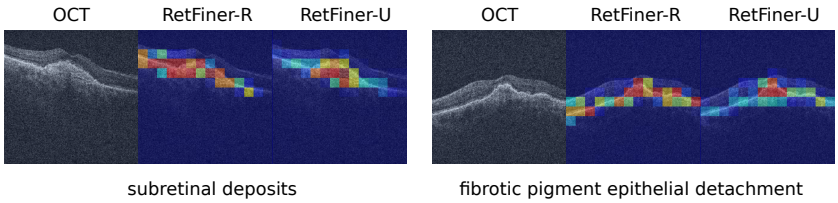

Fig. 2. Examples of RetFiner-R and RetFiner-U attention maps for disease cases.

of the models' features. We compared this strategy with other common pooling strategies: CLS token, average pooling of patch tokens, and average pooling of all tokens (patch+CLS). Table 4 shows that the concatenation technique results in a statistically significant increase in balanced accuracy, AP, and F1-score.

Explainability. Fig. 2 shows examples of the attention rollout maps for our refined models RetFiner-R and RetFiner-U, based on RETFound and UrFound, respectively. The attention maps are generated by calculating the cross-attentions in the text encoder between the image features and the text features. As shown in the images, RetFiner's maps highlight the retinal layers and activate more strongly around the lesion biomarkers indicated in the text, regardless of the base FM. This demonstrates the effective semantic understanding of OCT images by our RetFiner models and their explanatory capabilities.

4 Conclusion

We introduce RetFiner, a vision-language refinement scheme that enhances the semantic understanding of retinal FMs through SSL on paired OCT images and EHRs. By combining multiple training objectives, RetFiner leverages textual data to refine visual representations without manual annotation. Evaluated on

7 diverse classification tasks, RetFiner significantly improved the linear probing performance of SOTA retinal FMs (RETFound, UrFound, VisionFM). Notably, RetFiner demonstrated strong adaptability to our complex in-house dataset, highlighting its utility for population-specific adaptation. Ablation studies confirmed the effectiveness of each training objective and our new feature pooling strategy. Furthermore, attention visualizations revealed that RetFiner models focus on clinically relevant biomarkers, enhancing explainability. With efficient training (under 10 epochs) and compatibility with existing FMs, RetFiner offers a practical solution to adapt models to local data distributions while improving overall semantic understanding and performance. Notably, our scheme is not specific to ophthalmic data and could be easily applied to other medical domains.

Acknowledgments. This research was funded in part by the Austrian Science Fund (FWF) Grant-DOI:10.55776/FG9, Christian Doppler Research Association, Austrian Federal Ministry of Economy, Energy and Tourism, and the National Foundation for Research, Technology and Development. For open access purposes, the author has applied a CC BY public copyright license to any author-accepted manuscript version.

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

References

1. Balestriero, R., LeCun, Y.: How learning by reconstruction produces uninformative features for perception. In: International Conference on Machine Learning (2024)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2021)
3. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. In: IEEE/CVF International Conference on Computer Vision. pp. 23403–23413 (2023)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Ericsson, L., Gouk, H., Loy, C.C., Hospedales, T.M.: Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine **39**(3), 42–62 (2022)
7. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., et al.: EVA: Exploring the limits of masked visual representation learning at scale. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)
8. Gerendas, B.S., Sadeghipour, A., Michl, M., Goldbach, F., Mylonas, G., Gruber, A., Alten, T., Leingang, O., Sacu, S., Bogunovic, H., Schmidt-Erfurth, U.: Validation of an automated fluid algorithm on real-world data of neovascular age-related macular degeneration over five years. RETINA **42**(9) (2022)

9. Gholami, P., Roy, P., Kuppuswamy Parthasarathy, M., Lakshminarayanan, V.: OCTID: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (01 2020)
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
11. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: IEEE/CVF International Conference on Computer Vision. pp. 3922–3931
12. Keenan, T.D.L., Cukras, C.A., Chew, E.Y.: Age-Related Macular Degeneration: Epidemiology and Clinical Aspects, pp. 1–31. Springer International Publishing, Cham (2021)
13. Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., et al.: OCTDL: Optical coherence tomography dataset for image-based deep learning methods. Sci. Data **11**(1), 365 (Apr 2024)
14. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023)
15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
16. Liu, B., Lu, D., Wei, D., Wu, X., Wang, Y., Zhang, Y., Zheng, Y.: Improving medical vision-language contrastive pretraining with semantics-aware triage. IEEE Transactions on Medical Imaging **42**(12), 3579–3589 (2023)
17. Luo, Y., Shi, M., Tian, Y., Elze, T., Wang, M.: Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. pp. 20414–20425 (10 2023)
18. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al.: DINOv2: Learning robust visual features without supervision (2024)
20. Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., et al.: Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. NEJM AI **1**(12), AIoa2300221 (2024)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Rasti, R., Rabbani, H., Mehridehnavi, A., Hajizadeh, F.: Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Transactions on Medical Imaging **37**(4), 1024–1034 (2018)
23. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Zhang, L., Liang, T., Yang, L., Ye, X., et al.: Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In: International Conference on Learning Representations (2025)
24. Silva-Rodríguez, J., Chakor, H., Kobbi, R., Dolz, J., Ben Ayed, I.: A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (Jan 2025)

25. Sotoudeh-Paima, S., Jodeiri, A., Hajizadeh, F., Soltanian-Zadeh, H.: Multi-scale convolutional neural network for automated amd classification using retinal OCT images. Computers in Biology and Medicine **144**, 105368 (2022)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30 (2017)
27. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., et al.: GAMMA challenge: Glaucoma grading from multi-modality images. Medical Image Analysis **90**, 102938 (2023)
28. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained interactive language-image pre-training (2021)
29. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
30. Yu, K., Zhou, Y., Bai, Y., Soh, Z.D., Xu, X., Goh, R.S.M., et al.: UrFound: Towards universal retinal foundation models via knowledge-guided masked modeling. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) International Conference on Medical Image Computing and Computer Assisted Interventions. pp. 753–762. Springer Nature Switzerland, Cham (2024)
31. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., et al.: CLIP in medical imaging: A comprehensive survey (2024)
32. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)