Continual Retinal Image Modality Learning via Representation Geometry

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Multimodal models can improve biomedical diagnosis, but training them is limited by the scarcity of paired data and the cost of maintaining a separate encoder per modality. We study *Continual Modality Learning*: adapting a single pretrained unimodal encoder to a new imaging modality without forgetting the original, using only limited paired data. Our method, CoMoGe, preserves the pairwise geometry of the pretrained representation during continual retraining, regularizing how far representations may drift and thereby preventing catastrophic forgetting. We adapt a pretrained retinal fundus encoder to three new modalities: OCT (optical coherence tomography), OCTA (OCT angiography), and chest X-rays. We show that CoMoGe retains performance on the original fundus images while learning each new modality, including the unpaired chest X-rays, for which no original fundus-image pairs are available.

Keywords: Continual Learning · Multimodal Learning · Retinal Imaging

1 Introduction

Multimodal data is important in medical applications. Combining different modalities not only provides additional information to improve diagnostic performance, but also acts as a complementary supervision signal. The development of multimodal models for biomedical data, including various imaging and biosensor data as well as genetic modalities, is increasingly important and has the potential to enable broad applications such as personalized medicine, digital clinical trials, and virtual health coaches [1].

While multimodal learning has shown great promise on natural datasets [17], a central challenge in translating these methods to biomedical applications is the availability of sufficient total and paired data. One solution is to train separate encoders for each modality and combine them, but given the large number of biomedical modalities, and applications such as oculomics [25] that span the whole body, this quickly becomes impractical and computationally costly. To address limited data, self-supervised pre-training and transfer learning are common, but the resulting foundation models are typically single-modality. Instead, we integrate a new modality directly into an existing encoder through continual

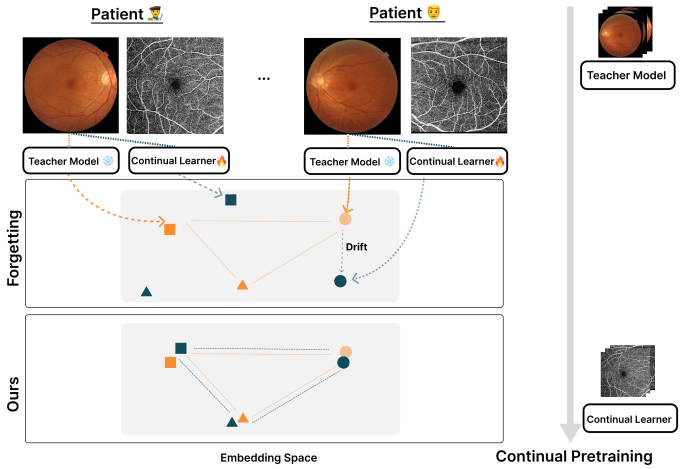

Fig. 1. Overview of our proposed method for continual pre-training for learning new data modalities. We continually train a pretrained model on fundus images, on new modalities including OCTA and OCT images. We assume that given a paired fundus image and new modality, such as OCTA as shown in the figure, the corresponding distances between these two modalities for a given patient should be preserved to prevent forgetting. A pretrained teacher model is used throughout training to provide reference representations to ensure pairwise relationships are maintained throughout pre-training.

pre-training, letting one model accumulate complementary signals across modalities without per-modality encoders.

In continual pre-training, one major challenge is avoiding catastrophic forgetting. Current methods focus on weight regularization, replay or distillation for example to prevent forgetting [19]. Much of a model's transferable knowledge is carried by the relational geometry of its representation, the pairwise relationships between samples, rather than representations of individual samples. This leads to the central question of our work: **Can we leverage pretrained representations to continually learn new data modalities while preventing catastrophic forgetting?**

In this paper, we address these challenges by proposing a method for learning new data modalities called CoMoGe, **Co**ntinual **Mo**dality Learning with Representation **Ge**ometry. We focus on continually learning new imaging modalities given a pretrained encoder with a masked autoencoder (MAE) [7] objective. To avoid catastrophic forgetting, we assume that distances between data pairs from a patient's pretrained modality and new modality are preserved. This assumption allows us to include a regularization term that prevents the original representation from deviating too far from the pretrained representation. We also explore the setting in which paired data from the pretrained modality are not available to examine CoMoGe's ability to retain pretrained knowledge. We

assess CoMoGe's ability on downstream tasks for both the pretrained and new modalities.

Our main contributions are as follows:

- We show that a single pretrained encoder can continually learn a new imaging modality without forgetting the original, by regularizing the pairwise geometry of the pretrained representation during continual pre-training.
- We show this holds across modalities that differ substantially from the pretraining domain, from retinal OCT and OCTA to chest X-ray, and across multiple downstream tasks.
- Using CKA (centered kernel alignment), we characterize how the representation drifts during continual pre-training and show that our regularizer retains knowledge better than naive continual MAE.

2 Related Work

2.1 Representation Learning

While our work focuses on masked autoencoders (MAEs) as the pre-training objective for continual pre-training, there have been many successes in both unimodal and multimodal representation learning, including masked objectives for multiple data modalities [12]. Contrastive objectives such as SimCLR [4] and CLIP [18] align modalities by minimizing distances between positive pairs while maximizing distances between negative pairs.

Our work builds on methods that regularize the relational geometry of representations, the pairwise structure between samples, rather than individual feature values. This is motivated in part by the Platonic Representation Hypothesis [9], which posits that representations of the same underlying reality converge across modalities, suggesting a pretrained modality's relational geometry captures stable, transferable structure. STRUCTURE [6] preserves multi-scale relational geometry to align modalities under limited data. We re-purpose this operator to *preserve* rather than *align* geometry in the continual pre-training setting. Similarly, Relational Knowledge Distillation (RKD) [15] constrains representation geometry by preserving pairwise distance-and angle-wise structure between samples across training stages. We show that preserving this geometry on the original modality is an effective anti-forgetting signal when continually learning a new modality.

2.2 Continual Learning

Continual learning aims to acquire new knowledge without forgetting previously learned knowledge [19]. Existing approaches include regularization-, data-, and representation-based methods [20]. Our work lies at the intersection of representation and regularization-based continual learning in the modality-incremental setting [23].

Regularization-based methods constrain changes in important *parameters* (EWC [10]) or in the model's *outputs* and *features* (LWF, CaSSLe [5], [13]). CaSSLe uses a predictor network to map the current state of the representation to past states to enable continual self-supervised learning. Ye and colleagues [22] use a knowledge-distillation approach to continually learn new data modalities, including clinical reports, X-rays, and CT scans. RetCoP [21] continually pretrains across retinal imaging modalities using image-text pairs and an off-diagonal information distillation.

3 Method

3.1 Problem Setting

CoMoGe learns a new data modality while preserving pretrained representation geometry. We consider a paired dataset, $\{(x_A^i, x_B^i)\}$ on modality A and B, and a model f_θ pretrained A. Our goal is to adapt f_θ into $f_{\hat{\theta}}$ so that it does not forget pretrained representations on A while learning a new data modality B. This allows $f_{\hat{\theta}}$ to perform well on downstream tasks with either pretrained or new data modalities. Representations learned by deep learning models often encode meaningful distances between samples, and much of a model's transferable knowledge lives in this relational geometry [15,11]. We view catastrophic forgetting as a distortion of this geometry on the original modality when the encoder adapts to B, and we prevent it by preserving the pairwise geometry of the pretrained representations on A throughout the continual pre-training, using just modality A independent of the new modality.

3.2 Preserving Representation Geometry

To both learn new data modalities and preserve learned representations, we leverage STRUCTURE [6] to minimize differences between pairwise similarity across samples.

Given mini-batch of embeddings $Z_A = \{z_A^i\}_{i=1}^N$ and $Z_B = \{z_B^i\}_{i=1}^N$, where $z_X^i = f_\theta(x_X^i)$. We first ℓ_2-normalized and center the embeddings before computing the similarity matrices, obtaining $\tilde{z}_A^i$ and $\tilde{z}_B^i$. We then stack these embeddings into $\mathrm{E}_X = [\tilde{z}_X^1, \ldots, \tilde{z}_X^N]$. Following STRUCTURE [6], we compute similarity matrices with temperature $\tau > 0$:

$$\mathrm{S}_A = \frac{\mathrm{E}_A \mathrm{E}_A^T}{\tau}, \mathrm{S}_B = \frac{\mathrm{E}_B \mathrm{E}_B^T}{\tau} \tag{1}$$

Applying a row-wise softmax to S_A and S_B gives transition matrices: $\mathrm{P}_A = \mathrm{softmax}(\mathrm{S}_A)$ and $\mathrm{P}_B = \mathrm{softmax}(\mathrm{S}_B)$. Therefore, entry ij in $\mathrm{P}_A^{(l)}$ represents the sum of probabilities of distinct walks of length l from i to j. Jensen-Shannon divergence can then be applied:

$$\mathrm{JS}(\mathrm{P}_A^{(l)}, \mathrm{P}_B^{(l)}) = \frac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathrm{P}_A^{(l)} \mid\mid \mathrm{M}^{(l)}) + \frac{1}{2}\mathrm{D}_{\mathrm{KL}}(\mathrm{P}_B^{(l)} \mid\mid \mathrm{M}^{(l)}) \tag{2}$$

where $\mathrm{M}^{(l)} = \frac{1}{2}(\mathrm{P}_A^{(l)} + \mathrm{P}_B^{(l)})$.

The final regularizer averages over walk lengths:

$$\mathcal{L}_S^{(L)}(Z_A, Z_B) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{l}\mathrm{JS}(\mathrm{P}_A^{(l)}, \mathrm{P}_B^{(l)}) \tag{3}$$

3.3 Learning Objective

With $Z_A^{\hat{\theta}} = \{f_{\hat{\theta}}(x_A^i)\}_{i=1}^N, Z_B^{\hat{\theta}} = \{f_{\hat{\theta}}(x_B^i)\}_{i=1}^N$ and $Z_A^{\theta} = \{f_{\theta}(x_A^i)\}_{i=1}^N$, the complete training objective is:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{MAE}}(X_B)}_{\text{Continual Pre-training}} + \underbrace{\lambda\mathcal{L}_S^{(L)}(Z_A^{\theta}, Z_A^{\hat{\theta}})}_{\text{Prevent Forgetting}} \tag{4}$$

The MAE term learns the new data modality, and the second term prevents forgetting by constraining representation drift from the original modality, through regularizing multi-levels, or the number of distinct walks, between any two samples within a mini-batch.

4 Experiments

To evaluate CoMoGe's ability to learn new data modalities while preventing forgetting, our experiments show that:

1. CoMoGe avoids catastrophic forgetting (Fig. 2).
2. CoMoGe learns useful representations of the new modality, reflected in downstream performance (Table 1).

Since we prevent forgetting by constraining representations rather than reconstructing the original modality, we can compute the reference original-modality representations from an independent set of images that need not be paired with the new modality. We focus on fundus as the pretrained data modality with pretrained foundation model RETFound [24] with two settings. First, we are given data tuples for which we have access to the original and new (paired) data modalities, such as the same patient's fundus images (the pretrained modality), OCTA images, and OCT images. The second setting is one in which we are given unpaired data, which is more similar to the real-world setting, where we have access to the original data modality and a different unpaired data modality, such as chest X-ray images.

Baselines: We compare against RETFound with no continual training, naive continual MAE with no anti-forget term, and RKD [15], a relational-distillation regularizer preserving pairwise distances and angles between samples. CoMoGe and RKD differ only in the geometry-preserving operator, isolating the effect of the multi-scale relational objective.

Training Details We continually train fundus pretrained RETFound on different data modalities with MAE, RKD regularization, or our CoMoGe (Eq. 4) objective sequentially on OCT, OCTA, or chest x-ray images (dataset details described below). All models are trained for 100 epochs for pre-training, with 10 epochs of linear warmup, base learning rate $1.0e-4$ with cosine learning rate scheduler, weight decay of $5e-2$ and batch size of 64. We finetune encoders for downstream task evaluation, with batch size of 64, learning rate $1.25e-3$, cosine learning rate scheduler with 10 epochs of linear warmup and layer decay of 0.65 for 50 epochs.

Pre-training Datasets: In both settings, we are given a foundation model trained on retinal fundus images, and the goal is to extend this model to different imaging modalities, including OCT, OCTA, and chest X-rays. None of the datasets used for our experiments were used to train the foundation model. We leverage the AI-READI [3] [2] dataset, which provides paired fundus, OCTA, and OCT images for learning new data modalities. VinDr-CXR [14] is used as a new modality with chest X-ray images that are not paired with images from the original modality, where we sample a random mini-batch without replacement of fundus images from AI-READI for each epoch. We follow the recommended data splits (train, validation, test sets) for both datasets.

Downstream Tasks: We finetune the 9 encoders resulting from continual training on one downstream task, diabetic retinopathy (DR) detection using IDRiD (a fundus-image only dataset [16]) to evaluate forgetting. Furthermore, to evaluate learning new modalities, we assess performance on two downstream tasks: (1) type 2 diabetes mellitus (T2DM) detection using AI-READI (OCT and OCTA) and (2) finding vs. no-finding using VinDr-CXR chest x-rays. To obtain finding vs. no-finding labels, we convert VinDR-CXR's original multi-part labels into binary labels based on majority vote across radiologists.

4.1 CoMoGe Prevents Catastrophic Forgetting by Aligning Multimodal Representations

Figure 2 reports downstream performance after continually pre-training RETFound on each new modality. Figure 2 illustrates that continual MAE pre-training causes forgetting, reflected by the decreasing AUC, accuracy, and loss across each added modality. CoMoGe mitigates this degradation by regularizing changes in representations from the original modality and thereby achieving higher AUC and accuracy and lower loss compared to MAE pre-training. While CoMoGe is designed to preserve multi-scale relational representations, RKD is also designed to preserve single-scale ones, possibly explaining their comparable performance. We expect CoMoGe's performance to improve with further hyperparameter optimization. We further applied Centered Kernel Alignment (CKA) [11] to measure the similarity of all other pretrained representations vs. CoMoGe's representations in Figure 3. CoMoGe's representations maintain higher similarity to those of fundus-image pre-trained RETFound across most layers, indicating better preservation of pretrained fundus representations (limited forgetting) compared to continual MAE. The remaining decrease in similarity is

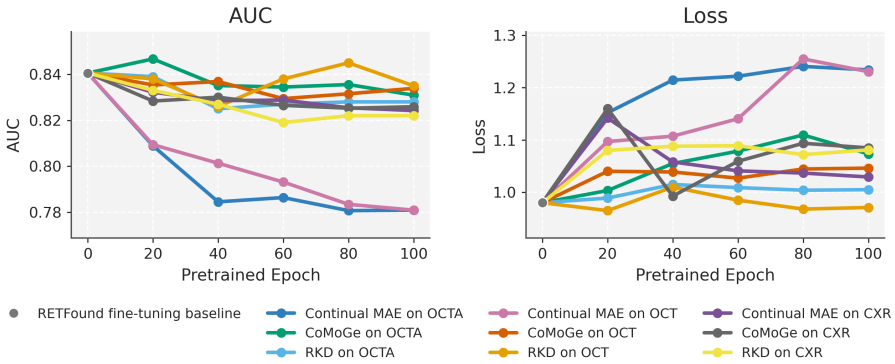

Fig. 2. Downstream DR task evaluation after continual pre-training on OCTA, OCT, or CXR and fine-tuning on fundus IDRiD images. Epoch 0 finetunes the unmodified RETFound encoder. Each subsequent point represents a downstream model independently fine-tuned at the indicated continual-pre-training epoch. We report test AUC and loss. Continual MAE exhibits greater AUC degradation after OCT and OCTA pre-training than after CXR pre-training. Loss shows modalitiy-specific trends. CoMoGe generally reduces AUC degradation compared to continual MAE and performs comparably to RKD.

Method	OCT	OCTA	CXR
RETFound	0.636	0.663	0.827
Continual MAE	**0.666**	0.671	0.904
RKD	0.625	0.613	**0.907**
CoMoGe (ours)	0.660	**0.694**	0.889

Table 1. Downstream performance of the given methods after learning each new modality. AUC is reported on each dataset's test set. T2DM prediction is reported using OCT and OCTA, and findings vs. no-findings prediction is reported using chest x-rays (CXRs).

likely due to applying the regularizer to the final layer, so this decrease could potentially be mitigated by varying layer-positioning for the regularizer. [6].

4.2 Learning New Data Modalities

Table 1 reports CoMoGe's performance on downstream tasks. T2DM performance using AI-READI is reported for both OCT and OCTA. For OCT, CoMoGe performs comparably to continual MAE. For OCTA, CoMoGe achieves higher AUC than both original RETFound and continual MAE. While CoMoGe outperforms RKD on downstream tasks, the performance gain can be attributed to relatively more forgetting compared to RKD. Further hyperparameter tuning for CoMoGe may improve the balance between retention (not forgetting) and plasticity (new-modality learning).

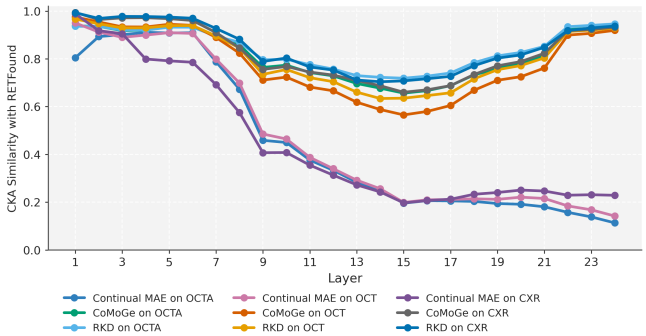

Fig. 3. CKA similarity with fundus RETFound across ViT layers. measured on AI-READI's fundus images testset. Continual MAE exhibits substantially greater representation drift, particularly in deeper layers, while CoMoGe and RKD better preserve the pretrained representations.

5 Conclusion and Future Work

CoMoGe establishes a method for mitigating forgetting by preserving learned representations, enabling learning new data modalities. Across OCT, OCTA and chest X-rays (CXRs), CoMoGe retains more of the original fundus representations than continual MAE while performing better on downstream tasks with new data modalities on T2DM prediction with OCT and OCTA, and on findings/no-findings prediction with CXRs.

Our experiments consider the addition of a single new modality and evaluate forgetting on a limited set of downstream tasks. Future work will study continual learning over longer sequences of modalities, as in RetCoP [21]. In its current form, CoMoGe preserves representations of the original pretrained modality but does not explicitly preserve representations learned from intermediate modalities during subsequent continual training. Additionally, while we proposed a regularization-based method on changes in representation, there are other methods for continual learning such as LoRA [8]. Future work will explore the efficacy of CoMoGe independently compared to these other continual learning methods.

Our method is also dependent on the dataset from the original data modality that is accessible during continual learning. At the same time, this dataset's coverage might cause the model to retain knowledge for only some downstream tasks, for example a dataset with only diseased images. Lastly, future work will also explore opportunities to both prevent forgetting and further improve multimodal representations, as regularizing the model's representations not to change much from the pretrained representations might inhibit the model's ability to improve performance on new downstream tasks.

Acknowledgements

Left blank for double-blind submission.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical AI. Nat. Med. **28**(9), 1773–1784 (Sep 2022)
2. AI-READI Consortium: AI-READI: Artificial intelligence ready and equitable atlas for diabetes insights. https://aireadi.org (2024), dataset
3. AI-READI Consortium: Flagship dataset of type 2 diabetes from the AI-READI project (2025)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020), https://arxiv.org/abs/2002.05709
5. Fini, E., da Costa, V.G.T., Alameda-Pineda, X., et al.: Self-supervised models are continual learners. In: CVPR (2022)
6. Gröger, F., Wen, S., Le, H., Brbić, M.: With limited data for multimodal alignment, let the structure guide you (2025), https://arxiv.org/abs/2506.16895
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021), https://arxiv.org/abs/2111.06377
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
9. Huh, M., Cheung, B., Wang, T., Isola, P.: The platonic representation hypothesis (2024), https://arxiv.org/abs/2405.07987
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. PNAS **114**(13), 3521–3526 (2017)
11. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited (2019), https://arxiv.org/abs/1905.00414
12. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language (2019), https://arxiv.org/abs/1908.03557
13. Li, Z., Hoiem, D.: Learning without forgetting (2017), https://arxiv.org/abs/1606.09282
14. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T.T., Dinh, D.H., Do, C.D., Doan, L.T., Nguyen, C.N., Nguyen, B.T., Nguyen, Q.V., Hoang, A.D., Phan, H.N., Nguyen, A.T., Ho, P.H., Ngo, D.T., Nguyen, N.T., Nguyen, N.T., Dao, M., Vu, V.: VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. Sci. Data **9**(1), 429 (Jul 2022)
15. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation (2019), https://arxiv.org/abs/1904.05068
16. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y.H., Lee, Y.C., Jung, S.H., Li, Z., Sui, X., Wu, J., Li, X., Zhou, T., Toth, J., Baran, A., Kori, A., Chennamsetty, S.S., Safwan, M., Alex, V., Lyu, X., Cheng, L., Chu, Q., Li, P., Ji, X., Zhang, S., Shen, Y., Dai, L., Saha, O., Sathish, R., Melo, T., Araújo, T., Harangi, B., Sheng, B., Fang, R., Sheet, D.,

Hajdu, A., Zheng, Y., Mendonça, A.M., Zhang, S., Campilho, A., Zheng, B., Shen, D., Giancardo, L., Quellec, G., Mériaudeau, F.: IDRiD: Diabetic retinopathy - segmentation and grading challenge. Med. Image Anal. **59**(101561), 101561 (Jan 2020)

17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), `https://arxiv.org/abs/2103.00020`
18. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
19. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE TPAMI (2024)
20. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application (2024), `https://arxiv.org/abs/2302.00487`
21. Yao, Y., Wu, R., Zhou, Y., Zhou, T.: Continual retinal vision-language pre-training upon incremental imaging modalities (2025), `https://arxiv.org/abs/2506.19320`
22. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning (2023), `https://arxiv.org/abs/2311.17597`
23. Zhang, D., et al.: Merge then realign: Simple and effective modality-incremental continual learning for multimodal llms. arXiv preprint arXiv:2503.07663 (2025)
24. 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., Kihara, Y., UK Biobank Eye & Vision Consortium, Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., Keane, P.A.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (Oct 2023)
25. Zhu, Z., et al.: Oculomics: Current concepts and evidence. Prog. Retin. Eye Res. **106**(101350), 101350 (May 2025)