Towards Interpretable Foundation Models for Retinal Fundus Images

Samuel Ofosu Mensah[1,2⋆], Camila Roa[1,2⋆], Kerol Djoumessi[1], and Philipp Berens[1,2]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
[2] Tübingen AI Center, University of Tübingen, Germany
{samuel.ofosu-mensah, maria-camila.roa-carvajal, philipp.berens}@uni-tuebingen.de

Abstract. Foundation models are used to extract transferable representations from large amounts of unlabeled data, typically via self-supervised learning (SSL). However, many of these models rely on architectures that offer limited interpretability, a critical issue in high-stakes domains such as medical imaging. We propose Dual-IFM, a foundation model that is interpretable-by-design via a BagNet backbone whose small receptive fields generate class evidence maps that are faithful to the model's decision-making process. Additionally, Dual-IFM incorporates a $2D$ projection layer during pretraining that enables direct visualization of the representation space, providing a dataset-level view of the learned structure including meaningful clinical clusters as well as potential spurious correlations. We trained Dual-IFM on over 800,000 color fundus photographs from various sources to learn generalizable representations for different downstream tasks. Our model achieves performance comparable to RETFound, which has 16× more parameters, while providing interpretable predictions on out-of-distribution data. These results suggest that large-scale SSL pretraining paired with inherent interpretability can lead to robust representations for retinal imaging. Code and pretrained models are available at github.com/berenslab/interpretable_FM.

Keywords: Foundation Model · Interpretable Model · Self-Supervised Learning · Color Fundus Photography.

1 Introduction

Foundation models are large-scale deep learning models trained on massive, diverse unlabeled datasets, typically via self-supervised learning (SSL), to learn general-purpose representations that can be adapted to downstream tasks [6]. In ophthalmology, a widely used foundation model called RETFound [29] has inspired many studies across ophthalmic imaging tasks [13,24,25]. While these models typically achieve strong performance, they are not inherently interpretable,

⋆ These authors contributed equally.

offering no built-in explanation for each prediction [16,18]. Thus, several studies on foundation models rely on post-hoc attribution methods to explain the model's decisions [25,24,29]. However, these local explanations are often unfaithful, since they do not align with the model's decision-making process [2,23]. This lack of faithfulness limits the adoption of foundation models, especially in high-stakes domains such as medical imaging.

Understanding the model's decision-making process is important, so as the semantic structure of the representations it learns from the data [4]. Foundation models yield high-dimensional representations of input data, embedding semantic relationships in the geometric properties of the latent space. Although attribution methods aim to provide local explanations for individual predictions, they do not reveal whether the underlying representation space accurately captures semantic relationships, that is, whether semantically similar concepts such as disease grades are proximally embedded and whether the geometric structure reflects meaningful patterns rather than data artifacts or spurious correlations [1]. Visualizing representation spaces offers a complementary dataset-level view of the learned global structure, typically done via neighbor embedding methods such as t-SNE [19,25]. However, t-SNE is non-parametric: it learns no explicit high-to-low-dimensional mapping and must be re-fitted whenever new samples are added, which can be costly for large datasets.

Motivated by these considerations, we introduce Dual-IFM, an interpretable foundation model for color fundus photography (CFP) that combines local explanations with a direct visualization of the representation space. We trained our model using the BagNet [7] architecture with the t-SimCNE algorithm [5]. Unlike post-hoc attribution methods, BagNet is interpretable-by-design: by restricting the receptive field to small patches, it generates class evidence maps that indicate which regions of the image contributed to a specific prediction. t-SimCNE trains a parametric projection layer that maps images into a $2D$ approximation of the representation space, enabling Dual-IFM to embed new images directly without re-fitting non-parametric methods such as t-SNE. Together, these components constitute a framework for local interpretability and dataset-level representation inspection of the learned structure, supporting more transparent use of foundation models in high-stakes domains like ophthalmology.

2 Methods

We trained the inherently-interpretable foundation model Dual-IFM on a large dataset of CFP. The model has a BagNet architecture as the backbone [7] and is trained to learn generalizable representations using the self-supervised contrastive t-SimCNE algorithm [5].

2.1 Architecture

Fig. 1 shows the architecture of Dual-IFM, where the upper branch is learned during pretraining. We decompose it into an encoder $f : \mathcal{X} \to \mathcal{H} \subset \mathbb{R}^D$, with

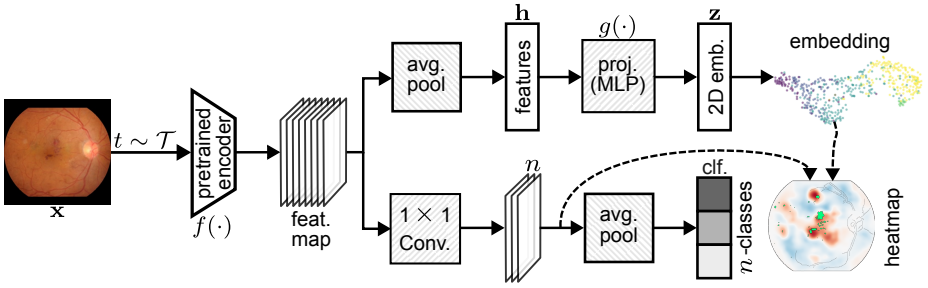

Fig. 1. Architecture of Dual-IFM. The BagNet encoder is pretrained on a large CFP dataset with t-SimCNE, learning generalizable representations and a $2D$ projection of the embedding space. A classification branch can be added for fine-tuning on downstream tasks while preserving the $2D$ projection.

D denoting the encoder feature dimension, and a projection layer $g : \mathcal{H} \to \mathcal{Z} \subset \mathbb{R}^2$. The encoder produces a spatial feature map which is aggregated by global average pooling to obtain a D-dimensional feature vector $\mathbf{h}$, and the projection head maps $\mathbf{h}$ to a $2D$ embedding $\mathbf{z} = g \circ f(\mathbf{x})$, at which the contrastive loss is computed. Following SimCLR [9], the high-dimensional representations $\mathbf{h}$ of the encoder are used for downstream tasks, while the projection head is retained to embed images into a $2D$ visualization of the representation space.

After pretraining, we introduce a classification head to form the lower branch (Fig. 1), operating on the encoder's feature map in parallel with the projection layer. Following [12], we attach a 1×1 convolution to the final feature map to produce n class-wise activation maps that directly highlight the regions of the image that contributed to the model's predictions. Global average pooling is then applied over each activation map to yield a corresponding class score, making the prediction an explicit spatial aggregation of local evidence. This classification head can be used for fine-tuning on downstream tasks.

2.2 Pretraining

We trained Dual-IFM using the t-SimCNE algorithm [5], a variant of SimCLR [9], that adapts the contrastive objective to obtain a $2D$ visualization of the embedding space through a three-stage training procedure. It replaces SimCLR's $128D$ projection, on which the contrastive loss is computed, wit a $2D$ projection, directly embedding the input images in a $2D$ space. Additionally, t-SimCNE replaces the cosine similarity, used by SimCLR's NT-Xent loss with Euclidean similarity to improve visualization quality. In stage one, the full model is trained computing the loss at the $128D$ output of the projector. In stage two, the last projector layer is replaced with a $2D$ output. This layer is then aligned with the encoder's representations, while the other layers remain frozen. In stage three, the full model is unfrozen and trained end-to-end at a lower learning rate.

The authors of t-SimCNE found that using Euclidean similarity throughout all stages yields the best visualizations, but also reported that cosine similarity

during stage one improves linear probe performance at a slight cost to visualization quality. Evaluating both, we observed the opposite effect on visualization quality (see Section 3.2) and therefore adopted cosine similarity in stage one.

2.3 Fine-tuning

During fine-tuning, the encoder and classification head are trained end-to-end while the projector is kept frozen. As the encoder representations are updated, the $2D$ projector's mapping may become suboptimal, with the degree of degradation depending on how much the representations drift from the pretraining ones. When the representations remain close to pretraining, the projector remains a reliable approximation; otherwise, it can be realigned to the fine-tuned representations in a computationally inexpensive way, since only the projector is retrained. This alignment can be performed during or after fine-tuning.

The faithfulness of the activation maps produced by the lower branch stems from the BagNet backbone [7], which restricts the receptive field of each spatial location to a small patch of the input image. This means each entry in the feature map depends only on a local region of the image, ensuring that the class evidence maps directly reflect the local image content that contributed to the prediction. Unlike post-hoc attribution methods, which approximate explanations after inference, this interpretability is built into the architecture itself. The evidence maps are not an explanation layered on top of the decision process, but an intrinsic part of it, eliminating the need for any post-hoc attribution methods.

2.4 Experimental details

We pretrained Dual-IFM on three CFP datasets that were preprocessed and filtered for quality [14]: EyePACS (567,384 images from 44,063 participants) [10], AREDS (110,690 images from 4,432 participants) [26], and UKBiobank (132,010 images from 72,711 participants) [27], yielding 802,360 images. We evaluated the model on out-of-distribution datasets: APTOS [15] (3,662 images), IDRiD [22] (516 images), DeepDRiD [21] (2,000 images) and Messidor-2 [11] (1,718 images) for DR grading, Glaucoma fundus [3] (1,544) and PAPILA [20] (488) for glaucoma grading, and FIVES [17] for multi-disease classification. We used BagNet-33 (variant with a 33×33-pixel receptive field) as Dual-IFM's backbone and compared it with a BagNet-33 initialized with ImageNet weights, a BagNet-33 trained with SimCLR on our pretraining dataset, and RETFound (ViT-Large).

For pretraining, we used the LARS optimizer [28] (batch size 1024, weight decay 10^{-6}, base learning rate 0.075×32). The learning rate was reduced by a factor of 1000 in the last stage, following a cosine annealing schedule with a warm-up of 10 epochs for stages one and three, and kept constant for stage two. Both models were trained with mixed precision. While SimCLR is compatible with float16, we found that t-SimCNE requires bfloat16 on large datasets, as the unconstrained Euclidean distances in the loss can otherwise lead to overflow.

We trained a SimCLR baseline for 1000 epochs and used these weights to initialize the t-SimCNE variant with cosine similarity in stage one, which was

then trained for 25 and 200 epochs in stages two and three, respectively. t-SimCNE with Euclidean similarity was trained from scratch for 775, 25, and 200 epochs. As we show in Section 3.2, the cosine variant yielded better visualizations for our CFP dataset and was therefore adopted as Dual-IFM. Pretraining took up to 6 days using 8 NVIDIA H100 GPUs (see Table 1).

Linear probing was performed by fitting a logistic regression classifier with elastic net penalty and class-balanced weights on the model's representations for each dataset, and evaluating performance on a held-out test set. For fine-tuning, we trained the linear classification head for 10 epochs before unfreezing the full model, continuing for a maximum of 50 epochs with early stopping. We used class-weighted cross entropy and five-fold cross-validation and report average performance across folds. For all datasets, we ensured that splits are performed at participant-level with stratification. Hyperparameters were selected via a small search for each dataset and full configurations are provided in our GitHub repository. We conducted the fine-tuning on one NVIDIA V100 GPU.

3 Results

3.1 Performance on eye disease classification

On held-out test sets of the pretraining datasets, Dual-IFM performed comparably to RETFound on in-distribution DR detection and AMD classification. On out-of-distribution datasets, Dual-IFM matched or outperformed a task-trained BagNet-33 across most tasks (Tables 2 and 3), indicating that self-supervised domain-specific pretraining improves performance on downstream tasks. Additionally, we found that Dual-IFM performed similarly to SimCLR and RETFound, suggesting that the added interpretability and $2D$ projection do not

Table 1. Model complexity and computational cost. Pretraining figures for RETFound are as reported in [29] (*described as "developing time" rather than training wall-clock); all other figures were measured in this work.

	Dual-IFM	RETFound
Architecture	BagNet-33	ViT-Large
# Parameters	18.3M	303.3M
Pretraining		
Data (# images)	802,360	1.6M (900k private)
Hardware	8×H100 (80GB)	8×A100 (40GB)
Epochs	1225 (1000+25+200)	800
Time (effective batch size)	6d 8h (1024)	2 weeks* (1,792)
Inference (batch size 16, V100, fp32)		
Time per batch	81 ms	140 ms
Peak memory	1.12 GB	1.31 GB

impair representation quality. Linear probing results (Table 2) showed improved representations for 4/7 and 6/7 datasets compared to the task-trained BagNet baseline (ImageNet initialized) and RETFound, respectively. Fine-tuning further improved performance for most datasets and models (Table 3), with the exception of PAPILA and FIVES, where performance occasionally decreased, likely due to overfitting given the limited hyperparameter search. Despite this limitation, Dual-IFM performs on par with or above RETFound, suggesting that it learns general representations from retinal images.

3.2 Global structure inspection through 2D projection

Using the learned 2D projection, Dual-IFM enables dataset-level inspection of the representation space by embedding entire datasets into a low-dimensional approximation. Fig. 2 shows two datasets as an example, illustrating how the representation space captures disease severity structure. For the APTOS dataset, samples arranged along a curved manifold following DR severity: images of lower grades clustered in one region, while progressively higher grades occupied increasingly distinct regions (Fig. 2A, yellow to blue). The embedding for Glau-

Table 2. Linear-probe performance across various CFP tasks with model properties.

	ImageNet	SimCLR	RETFound	t-SimCNE	Dual-IFM
Image-level interpretability	+	+	−	+	+
Dataset-level structure	−	−	−	+	+
In-Domain training	−	+	+	+	+
EyePACS	0.727	0.647	0.794	0.648	**0.849**
AREDS	0.877	0.925	0.871	**0.936**	0.919
APTOS	**0.931**	0.929	0.917	0.929	0.925
DeepDRiD	0.854	0.857	**0.887**	0.861	0.829
IDRiD	0.702	0.762	0.668	**0.785**	0.758
Messidor	0.801	**0.868**	0.797	**0.868**	0.820
Glaucoma	0.908	**0.921**	0.913	0.912	0.915
PAPILA	0.655	0.740	0.577	0.682	**0.757**
FIVES	0.895	0.912	0.837	**0.917**	0.881

Table 3. Fine-tuning performance across 5-folds on various CFP tasks.

	ImageNet	SimCLR	RETFound	t-SimCNE	Dual-IFM
EyePACS	0.758 ± 0.073	0.804 ± 0.038	0.767 ± 0.040	0.759 ± 0.037	$\mathbf{0.821 \pm 0.036}$
AREDS	0.908 ± 0.020	0.924 ± 0.008	0.912 ± 0.008	$\mathbf{0.936 \pm 0.007}$	0.923 ± 0.009
APTOS	0.938 ± 0.003	$\mathbf{0.941 \pm 0.009}$	0.940 ± 0.004	0.935 ± 0.010	0.940 ± 0.003
DeepDRiD	0.889 ± 0.008	0.901 ± 0.014	$\mathbf{0.903 \pm 0.007}$	0.883 ± 0.006	0.891 ± 0.012
IDRiD	0.807 ± 0.033	$\mathbf{0.814 \pm 0.029}$	0.764 ± 0.032	0.796 ± 0.018	0.766 ± 0.030
Messidor	0.865 ± 0.020	$\mathbf{0.874 \pm 0.006}$	0.845 ± 0.008	0.860 ± 0.012	0.851 ± 0.038
Glaucoma	0.912 ± 0.013	0.930 ± 0.007	$\mathbf{0.947 \pm 0.007}$	0.933 ± 0.009	0.925 ± 0.012
PAPILA	0.648 ± 0.056	0.664 ± 0.061	$\mathbf{0.689 \pm 0.058}$	0.676 ± 0.051	0.680 ± 0.027
FIVES	0.881 ± 0.021	0.891 ± 0.015	$\mathbf{0.936 \pm 0.012}$	0.910 ± 0.024	0.898 ± 0.030

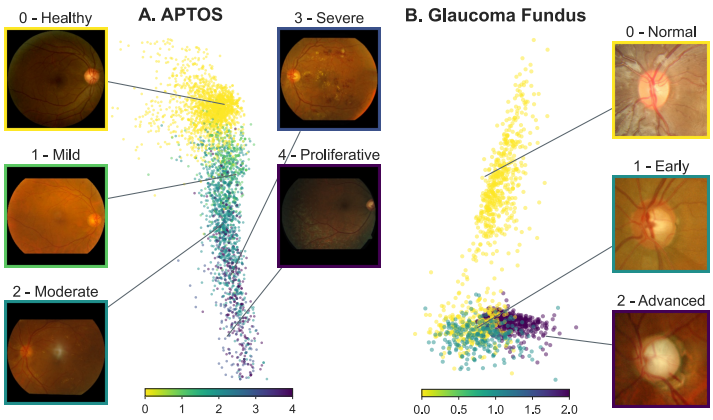

Fig. 2. Representation space visualization. $2D$ embeddings of (A) APTOS (k-NN AUROC: 0.881) and (B) Glaucoma Fundus (k-NN AUROC: 0.883), revealing disease progression structure and potential borderline cases.

coma Fundus exhibited a similar structure (Fig. 2B). For both datasets, we observed overlap among disease grades, highlighting potential borderline cases.

To quantify the quality of the $2D$ embeddings, we measured the k-NN AUROC across all datasets, comparing Dual-IFM to t-SNE and $2D$ PCA applied post-hoc to the high-dimensional representations of the baselines (Table 3). Contrary to the findings of the t-SimCNE, using cosine similarity in stage one achieved better $2D$ embedding quality than Euclidean similarity throughout (0.821 vs 0.745), leading us to adopt this variant as Dual-IFM. The $2D$ visualization can be further improved by aligning the projector to the fine-tuned encoder representations. However, this alignment is not always necessary, as illustrated by the meaningful structure already visible in Fig. 2, where the projector was used directly from pretraining without alignment.

Dual-IFM achieved an average k-NN AUROC of 0.821, compared to 0.851 and 0.839 for $2D$ PCA applied to ImageNet and SimCLR representations, respectively. While this reflects the expected cost of constraining the projection to $2D$ during pretraining, the learned representations nonetheless capture meaningful disease severity structure across datasets, without requiring any post-hoc dimensionality reduction. This cost is expected, since t-SNE and PCA are applied post-hoc to frozen high-dimensional representations, where the model can encode multiple factors of variation without a 2D constraint. In contrast, t-SimCNE's projector must compress these factors into 2D during the later stages of pretraining, while the encoder's representations are still shifting, making them a moving rather than fixed target. Additionally, rather than explicitly preserving the local geometry of the high-dimensional representations, t-SimCNE's objective maximizes augmentation-invariant information.

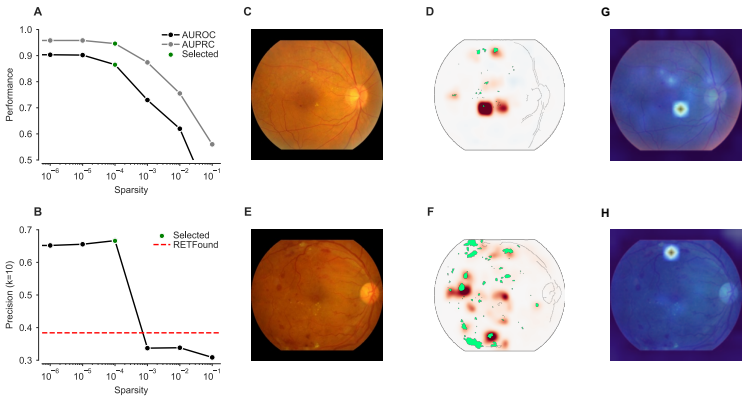

Fig. 3. Local interpretability. (A) Fine-tuning performance and (B) localization precision vs. sparsity. (C, E) Example fundus images with (D, F) class evidence maps of Dual-IFM and (G, H) LRP heatmaps from RETFound. Lesions overlaid in green.

3.3 Local interpretability through class evidence maps

We evaluated the class evidence maps provided by Dual-IFM for individual CFPs through the BagNet architecture (Fig. 3). We incorporated a sparsity constraint in the fine-tuning loss, as proposed by [12], which improves the localization of the evidence maps but affects the classification performance at higher values. We selected the sparsity value that maximized localization before performance deteriorated (Fig. 3 A). To assess the localization precision of the evidence maps, we measured the overlap of the $k = 10$ most activated patches with DR-related lesions on the IDRiD dataset, using Dual-IFM fine-tuned for binary DR classification (healthy vs. diseased). Dual-IFM achieved a precision of 0.674, indicating that the sparsity constraint produced well-localized evidence maps, compared to 0.384 for RETFound using post-hoc LRP [8]. Qualitatively, the class evidence maps showed high activations in CFP regions with disease-related lesions (Fig. 3 C-F), with substantially higher overlap than LRP heatmaps from RETFound (Fig. 3 G,H). Together, these results demonstrate the ability of Dual-IFM to provide faithful and clinically meaningful local explanations without the need for post-hoc attribution methods.

Table 4. k-NN AUROC of $2D$ embeddings. For baselines, t-SNE and PCA are applied post-hoc to high-dimensional representations. † Aligned projector.

Method	ImageNet	SimCLR	t-SimCNE	t-SimCNE†	Dual-IFM	Dual-IFM†
t-SNE	0.904	0.914	0.745	0.773	0.821	0.832
PCA $2D$	0.851	0.839				

4 Discussion

We presented Dual-IFM, an inherently interpretable foundation model for CFPs, that offers local interpretability and a direct visualization of the representation space for dataset-level exploration of learned structure. It achieved competitive performance across CFP benchmarks, demonstrating high-quality representations (linear probing) and adaptability to downstream tasks (fine-tuning). In some tasks, Dual-IFM with 18.3M parameters surpassed RETFound, even though RETFound is a much larger model with over 303M parameters trained on 1.6M retinal images (2× our training data), resulting in 1.7× the inference time (RETFound: 140 ms vs. Dual-IFM: 81 ms). However, due to the BagNet architecture's small receptive field and the resulting large spatial feature maps, Dual-IFM has a high memory footprint relative to parameter count (1.12 GB vs. RETFound's 1.31 GB), as detailed in Table 1.

Crucially, Dual-IFM's BagNet-based architecture provides inherent interpretability through class evidence maps that highlight lesion-relevant regions, unlike models like RETFound which rely on post-hoc attribution methods. These properties make Dual-IFM well-suited for clinical deployment, where both accuracy and transparency are required. Additionally, training a BagNet is memory-intensive, requiring a relatively small batch size despite its low parameter count (Table 1). Therefore, the availability of pretrained weights such as Dual-IFM that improve downstream performance is advantageous.

Our work was limited to CFP and a restricted hyperparameter search during fine-tuning. Future work could extend to additional modalities, interpretable-by-design backbones and SSL objectives that could improve performance from pretraining. The inherent interpretability would make Dual-IFM uniquely suited for a multimodal setting, while enabling dataset-level exploration across modalities. In summary, our results suggest that local interpretability and visualizations of the representation space can be obtained simultaneously without compromising representation quality in foundation models for medical images.

Acknowledgments. This project was supported by the Hertie Foundation and by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy with the Excellence Cluster 2064 "Machine Learning – New Perspectives for Science", project number 390727645. PB is a member of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research". The authors thank the AREDS Research Group for their contribution to this research with funding support provided by the National Eye Institute (N01-EY-0-2127). This research has been conducted using the UK Biobank Resource under Application Number (121699).

Disclosure of Interests. The authors declare no competing interests.

References

1. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations

through concept relevance propagation. Nature Machine Intelligence **5**(9), 1006–1019 (2023)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
3. Ahn, J.M., Kim, S., Ahn, K.S., Cho, S.H., Lee, K.B., Kim, U.S.: A deep learning model for the detection of both advanced and early glaucoma using fundus photography. PloS one **13**(11), e0207982 (2018)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
5. Böhm, J.N., Berens, P., Kobak, D.: Unsupervised visualization of image datasets using contrastive learning. In: International Conference on Learning Representations (2023)
6. Bommasani, R., et al.: On the opportunities and risks of foundation models (2022)
7. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. International Conference on Learning Representations (2019)
8. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 782–791 (June 2021)
9. 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)
10. Cuadros, J., Bresnick, G.: EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology **3**(3), 509–516 (2009)
11. 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)
12. Donteu, K.R.D., Ilanchezian, I., Kühlewein, L., Faber, H., Baumgartner, C.F., Bah, B., Berens, P., Koch, L.M.: Sparse activations for interpretable disease grading. In: Medical Imaging with Deep Learning (2023)
13. Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., Wang, N.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: International conference on medical image computing and computer-assisted intervention. pp. 709–719. Springer (2024)
14. Gervelmeyer, J., Müller, S., Huang, Z., Berens, P.: Fundus image toolbox: A python package for fundus image processing. Journal of Open Source Software **10**(108), 7101 (Apr 2025)
15. Goha, E.F., Chen, Z., Lima, W.X.: APTOS 2019 blindness detection competition dataset (Dec 2024)
16. Grote, T.: The allure of simplicity: On interpretable machine learning models in healthcare. Philosophy of Medicine **4**(1) (2023)
17. 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)
18. Kazmierczak, R., Berthier, E., Frehse, G., Franchi, G.: Explainability and vision foundation models: A survey. Information Fusion **122**, 103184 (2025)

19. Kobak, D., Berens, P.: The art of using t-SNE for single-cell transcriptomics. Nature communications **10**(1), 5416 (2019)
20. 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)
21. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
22. 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)
23. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
24. Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., et al.: A multimodal visual–language foundation model for computational ophthalmology. NPJ digital medicine **8**(1), 381 (2025)
25. Sun, Y., Tan, W., Gu, Z., He, R., Chen, S., Pang, M., Yan, B.: A data-efficient strategy for building high-performing medical foundation models. Nature Biomedical Engineering pp. 1–13 (2025)
26. The Age-Related Eye Disease Study Research Group: The age-related eye disease study (AREDS): design implications AREDS report no. 1. Controlled clinical trials **20**(6), 573 (1999)
27. Warwick, A.N., Curran, K., Hamill, B., Stuart, K., Khawaja, A.P., Foster, P.J., Lotery, A.J., Quinn, M., Madhusudhan, S., Balaskas, K., et al.: UK Biobank retinal imaging grading: methodology, baseline characteristics and findings for common ocular diseases. Eye **37**(10), 2109–2116 (2023)
28. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks (2017)
29. 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)