MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise

Ruiqi Wu[1,2], Chenran Zhang[1,2], Jianle Zhang[1,2], Yi Zhou[1,2⋆], Tao Zhou[3], and Huazhu Fu[4]

[1] School of Computer Science and Engineering, Southeast University, China
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, China
[3] Nanjing University of Science and Technology, Nanjing, China
[4] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
{ruiqiwu@seu.edu.cn, yizhou.szcn@gmail.com}

Abstract. Current fundus image analysis models are predominantly built for specific tasks relying on individual datasets. The learning process is usually based on data-driven paradigm without prior knowledge. To address this issue, we propose **MM-Retinal**, a multi-modal dataset that encompasses high-quality image-text pairs collected from professional fundus diagram books. Moreover, enabled by MM-Retinal, we present a novel **K**nowledg**e**-**e**nhanced foundational **p**retraining model which incorporates **F**undus **I**mage-**T**ext expertise, called **KeepFIT**. It is designed with image similarity-guided text revision and mixed training strategy to infuse expert knowledge. Our proposed fundus foundation model achieves state-of-the-art performance across six unseen downstream tasks and holds excellent generalization ability in zero-shot and few-shot scenarios. MM-Retinal and KeepFIT are available at here.

Keywords: Fundus image · Foundational pretraining · Knowledge-enhanced

1 Introduction

Deep learning has achieved great progress in fundus image analysis. However, most previous works [4,11,22,26] usually utilize individual datasets to train task-specific models. This fashion results in three major model weaknesses: 1) poor generalization ability and robustness across varying scenarios; 2) lack of professional fundus domain-knowledge guidance in learning phase; 3) a huge demand for annotated training data. These challenges underscore the need for developing a general-purpose foundation model which is able to analyze comprehensive ocular diseases in fundus image area. Moreover, learning such a model with less training data but more prior knowledge is preferred.

In fundus image field, there are many specific image-only public datasets for diverse ocular diseases, such as glaucoma [9, 13], diabetic retinopathy [2, 3,

⋆ Corresponding author: Yi Zhou

25], age-related macular degeneration [6], and pathological myopia [5]. Despite remarkable progress of foundation models in many fields, such as natural images [21,28], medical images like chest X-rays [14,18] and MRI [8,12], it lags far behind in fundus image area. Thus, it is of considerable value to explore foundational pretraining in this area. RETFound [27] and FLAIR [17] made two preliminary attempts but suffer from certain limitations. RETFound only relies on large-scale image data and adopts a masked image modeling manner. FLAIR exploits both vision and language modalities through contrastive pretraining objective, yet it simply maps category label names to fixed texts. Both of them lack integrating fundus expertise with rich and profound image-text descriptions.

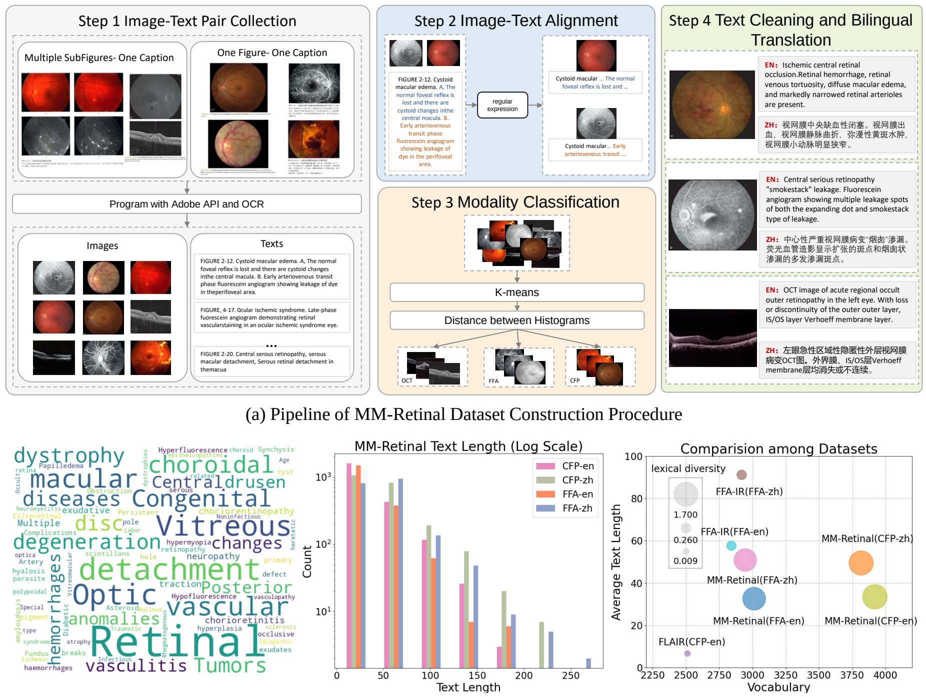
To accomplish our intention, a high-quality vision-language fundus dataset with expert knowledge is required. Such a dataset aims to not only promote the development of foundational fundus models, but also advance research in incorporating knowledge into learning more interpretable models. Additionally, it should propel research of multimodal fundus image analysis and beyond. Therefore, we built **MM-Retinal**, a multi-modal dataset which comprises image-text paired data of color fundus photography (CFP), fundus fluorescein angiography (FFA), and optical coherence tomography (OCT). All these data are collected from fundus diagram books with comprehensive ocular knowledge and accurate image-text descriptions from ophthalmologists. Moreover, we also developed **KeepFIT**, a knowledge-enhanced foundation model, enabled by MM-Retinal. Specifically, image similarity-guided text revision method and mixed training strategy are proposed to inject knowledge from MM-Retinal during training.

We highlight main contributions: 1) We construct a multi-modal MM-Retinal with over 4.3K high-quality image-text pairs in CFP, FFA and OCT modalities. The pairs are accurately matched with long texts, extensive vocabulary and comprehensive ocular disease and abnormalities. **2)** A knowledge-enhanced foundation model, KeepFIT, is proposed including a vision-language pre-training framework and knowledge integration methods. **3)** KeepFIT achieves SOTA on six representative downstream tasks, especially in zero-shot and few-shot scenarios, demonstrating robustness, generalizability and transferability.

2 The MM-Retinal Dataset

2.1 Dataset Construction

To construct MM-retinal with high-quality image-text pairs covering CFP, FFA, and OCT modalities, as shown in Fig. 1(a), our designed semi-automatic dataset construction pipeline contains four steps.1) Collect image-text pairs from diagram books on ocular diseases, which uses Adobe and OCR techniques for raw image and text extraction. 2) Align image-text pairs, and use regular expression matching for sub-figure caption separation. 3) Categorize images into CFP, FFA, and OCT by K-means and color histogram analysis, and exclude unusual modalities with scarce samples. 4) Manually correct OCR errors and irrelevant texts as well as translate texts to offer bilingual (English and Chinese) versions.



(a) Pipeline of MM-Retinal Dataset Construction Procedure

(b) Frequently Used Words

(c) Distribution of Text Length

(d) Dataset Comparison

Fig. 1: Construction workflow and statistical overview of MM-Retinal.

A six-person team took four weeks to get MM-Retinal completed. For further details, please refer to the supplementary material.

2.2 Dataset Statistics

Current version of MM-Retinal dataset includes 2,169 CFP cases, 1,947 FFA cases and 233 OCT cases. Each case is provided with an image and texts in both English and Chinese. Due to the small scale of OCT modality, we do not explore it for now and will extend this part in the future. Detailed statistical analysis of the CFP and FFA data is provided in the Fig. 1 from aspects of frequently used words, text length, vocabulary size, and comparison with other datasets.

Frequently Used Words: Since MM-Retinal dataset is built based on comprehensive ocular fundus diagram books, it covers a wide range of disease categories, such as macular, retinal vascular, and optic nerve diseases. In Fig. 1(b), we only show a small part of the words frequently appeared in CFP and FFA modalities. More details can be found in the supplementary material.

Text Length: Fig. 1(c) illustrates the text length distribution of our dataset. About 75% of English texts and 45% of Chinese texts range from 1 to 40 words, and 19% of English texts and 43% of Chinese texts contain between 41 and 80 words. Since our dataset is sourced from ophthalmologists' diagram books, it features much longer texts compared to FLAIR which simply maps class names of public datasets into fixed texts.

Vocabulary Size: MM-Retinal dataset contains diverse textual descriptions, such as disease diagnosis, lesion characteristics (e.g. color, shape, appearance), clinical manifestations, and post-treatment efficacy information. Fig. 1(d) showcases the average vocabulary size and lexical diversity across different modalities and languages, where average vocabulary size represents the total number of unique words contained in the whole texts within the dataset and lexical diversity refers to the vocabulary size averaged over each text.

Discussion: Compared to public fundus image datasets, MM-Retinal stands out from four aspects: **1) Multi-modality:** MM-Retinal is the pioneering fundus dataset that includes high-quality image-text expertise data for CFP, FFA, and OCT. **2) Data Quality.** In contrast to other medical datasets with low-quality web or academic paper-derived images and captions, MM-Retinal provides high-quality images with a resolution over 800×800, and accurate and pertinent text descriptions. The images closely match clinical data with minimal domain shift. **3) Category Variety.** MM-Retinal covers a broad range of categories with over 96 abnormalities and diseases. **4) Text Diversity.** MM-Retinal encompasses diverse vocabulary and long texts, containing extensive expert knowledge, and can be explored to enhance data-driven fundus image analysis models.

3 Knowledge-Enhanced Foundational Pretraining

3.1 Vision-Language Pre-training Framework

As shown in Fig. 2, we propose **KeepFIT**, a knowledge-enhanced foundational model pretrained on public and MM-Retinal datasets. We define public datasets with category-level labels as unimodal datasets and follow FLAIR [17] to fill labels into a template. We utilize ResNet50 as image encoder E_v and BioClinicalBert [1] as text encoder E_t, followed by projection heads P_v and P_t to match feature dimensions d. The extracted image features v_i and text features t_j are:

$$v_i = P_v \circ E_v(x_i) \in \mathbb{R}^d, \quad t_j = P_t \circ E_t(y_j) \in \mathbb{R}^d. \tag{1}$$

Given that texts detail the associated fundus images, our goal is to maximize similarity between paired image-text and minimize similarity for unpaired ones in a multimodal space. For MM-retinal, featuring genuine texts rather than textual prompts, we follow CLIP [15] to implement the contrastive loss as shown in Eq. 2, and the matching labels G^m_{v2t} and G^m_{t2v} are two $\mid \mathcal{B} \mid \times \mid \mathcal{B} \mid$ identity matrices:

$$\mathcal{L}_m = \mathbb{E}_{(x,y)\sim\mathcal{B}}[CE(G^m_{v2t}, S_{v2t}) + CE(G^m_{t2v}, S_{t2v})], G^m_{v2t}, G^m_{t2v} \in I_{|\mathcal{B}|}, \tag{2}$$

where m represents MM-Retinal, $\mathcal{B}$ is the batchsize, CE denotes InfoNCE loss, $S_{v2t} = \lambda(v_i {t_j}^\top)$ and $S_{t2v} = \lambda(v_i^\top t_j)$ are the cross-modality similarity, λ is a learnable scaling factor, $v2t$ and $t2v$ denote image-to-text and text-to-image.

For public datasets that only have prompt texts mapped from category-level labels, we follow FLAIR to calculate the category co-occurrence relationships

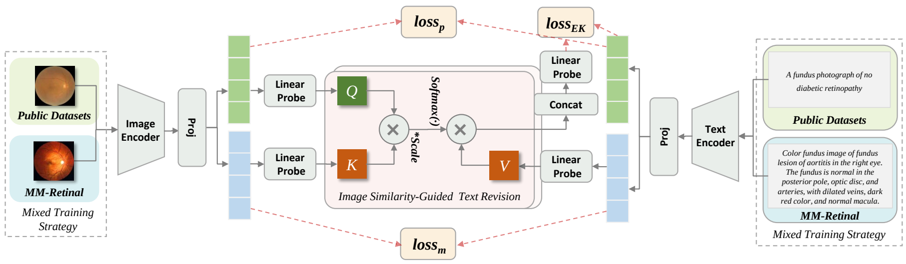


Fig. 2: **KeepFIT**: A vision-language pretraining framework using image-guided text revision and mixed training strategy to infuse expert knowledge.

among samples within a batch and construct a target matrix to encourage the pairs of the same category closer. Objective function Eq. 2 is converted into:

$$\mathcal{L}_p = \mathbb{E}_{(x,y)\sim\mathcal{B}}[CE(G^p_{v2t}, S_{v2t}) + CE(G^p_{t2v}, S_{t2v})], \tag{3}$$

$$G^p_{v2t} = G^p_{t2v} = \begin{cases} 1, & \text{if category}_v = \text{category}_t \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where p represents public datasets that only have category-level labels, the matching labels G^p_{v2t}, G^p_{t2v} are $\mid \mathcal{B} \mid \times \mid \mathcal{B} \mid$ symmetric matrices.

3.2 Expert Knowledge Integration Methods

As our proposed MM-Retinal features high-quality image-text pairs, long text description, rich vocabulary, and comprehensive ocular diseases categories, it encapsulates extensive fundus expert knowledge. Inspired by TipAdapter [23], we propose a lightweight expert knowledge integration method, called **Image Similarity-Guided Text Revision** along with **Mixed Training Strategy**, to infuse the expert knowledge from MM-Retinal to public datasets.

Image Similarity-Guided Text Revision: The descriptions within MM-Retinal are more comprehensive and professional than the simple texts mapped from class names in public datasets. Despite this, the images from both sources share a notable similarity. Hence, we start from identifying visual features in the public datasets that resemble those in MM-Retinal. Visual similarities are used as guidance to extract relevant prior knowledge from MM-Retinal's text features to refine and enhance the textual prompts of the public datasets.

Specifically, given an input image-text pair $[x_p, y_p]$ from public datasets and $[x_m, y_m]$ from MM-Retinal, the extracted features are (v_p, t_p) and (v_m, t_m). A multi-head cross-attention [19] is applied to exploit prior expert knowledge as:

$$EK = \text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O, \tag{5}$$

$$\text{head}_i = \text{ATTN}(v_p W^Q_i, v_m W^K_i, t_m W^V_i), \tag{6}$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter matrics for projection, v_p, v_m, t_m refer to image features of public datasets, image features of MM-Retinal and text features of MM-Retinal, respectively. Then, we establish an expert knowledge revision loss based on Mean Squared Error (MSE) to refine the text features of public datasets by incorporating expert knowledge, as formulated in Eq. 7:

$$\mathcal{L}_{EK} = \text{MSELoss}(EK, t_p). \tag{7}$$

Mixed Training Strategy: Since the public datasets has relatively homogeneous text prompts and almost no expert knowledge, which is vastly different from the texts of our dataset, we propose a mixed training strategy to avoid model optimization bias during the training process. To detail, the samples from public datasets and our dataset are in a 1:1 ratio in each batch.

Overall Training Objective: The overall training objective comprises three parts, which are public datasets contrastive loss $\mathcal{L}_p$, MM-Retinal contrastive loss $\mathcal{L}_m$ and expert knowledge revision loss $\mathcal{L}_{EK}$:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_m + \alpha\mathcal{L}_{EK}, \tag{8}$$

where α is to weight for $\mathcal{L}_{EK}$, empirically set as 100 showing best performance.

4 Experiment and Results

4.1 Datasets and Baselines

Pre-training Data: **1) flair [17] (CFP)** compiles 37 open-access fundus image datasets covering 96 categories with up to 284,660 images. These datasets provide category-level labels for classification. **2) SYNFUNDUS-1M [16] (CFP)** is a synthetic dataset with 1 million images for 14 diseases, created by a diffusion model trained on 1.3 million private fundus images. **3) FFA-IR [10] (FFA)** provides 10,790 reports along with 1,048,584 images from clinical practice. It includes a schema of 46 categories of lesion and bilingual reports. **4) MM-Retinal (CFP+FFA+OCT)** contains over 4.3K high-quality image-text pairs from professional fundus diagram books with over 96 abnormalities and diseases.

Foundation Model Baselines: **1) MoE [20] (MTL)** uses a multi-task approach with mixture-of-experts for fundus, macula, and optic disc images. **2) RETFound [27] (MIM)** is a masked autoencoder with Transformers, trained for retinal image reconstruction. **3) FLAIR [17] (CLIP)** utilizes a CLIP model and brief textual prompts that mapped from category labels for pre-training, but the texts are so limited and brief that fail to fully describe the images.

Evaluation Data and Metrics: For CFP, we evaluate on five datasets across finetuning, few-shot, and zero-shot settings, using classification accuracy (ACA) [24] for REFUGE, FIVES, ODIR200×3 and TAOP, receiving-operative-curve(AUC) for REFUGE and AMD, F1-score for AMD. For FFA, we evaluate on image captioning task using the FFA-IR test split, applying BLUE 1-4, Meter, Rouge, and Cider metrics for consistency with FFA-IR.

Table 1: Comparison of generalization ability on Few-Shot and Zero-Shot tasks. FLAIR and flair denote the model and datasets. syn represents SynFuduns-1M.

Method	Data	Few-Shot									Zero-Shot			Avg
		ODIR200×3									REFUGE	FIVES	ODIR200×3	
		ClipAdapter			TipAdapter			TipAdapter-f						
		1	5	10	1	5	10	1	5	10	AUC	ACA	ACA	
		ACA			ACA			ACA						
FLAIR	flair	0.720	0.823	0.863	0.403	0.413	0.422	0.417	0.462	0.535	0.926	0.670	0.403	0.588
	flair+syn	0.735	0.827	0.852	0.603	0.622	0.632	0.580	0.647	0.672	0.880	0.617	0.520	0.682
KeepFIT	flair	0.763	0.848	0.847	0.780	0.782	0.795	0.775	0.785	0.803	0.931	0.666	0.768	0.795
	flair+syn	0.795	0.858	0.862	0.751	0.777	0.783	0.760	0.795	0.807	0.856	0.696	0.777	0.793
	50%flair+MM	0.832	0.873	0.887	0.862	0.870	0.872	0.870	0.883	0.873	0.934	0.654	0.862	0.856
	flair+MM	0.848	0.878	0.893	0.823	0.843	0.842	0.820	0.847	0.853	0.941	0.731	0.812	0.844

Implementation Summary: In CFP and FFA, image and text encoders are initialized from ImageNet-1K and BioClinicalBERT. Attention mechanisms are trained from scratch, with 512 feature dimensions. Images are adjusted to 512×512 size. All the texts used are in English. Text tokens length is set at 256. Evaluation uses five-fold cross-validation averaging. We employ AdamW (lr=1e-4, decay=1e-5) optimizer and cosine scheduler with initial warm-up for the first epoch. Training is conducted on 4 RTX 3090 GPUs with batches of 24.

4.2 Comparison of Generalization Ability on Zero-Shot and Few-Shot Tasks

To compare the foundational performance of different models, we conducted experiments in zero-shot and few-shot scenarios with unseen categories. We adopted 1, 5, 10 shots and adapter tuning scheme in few-shot, including ClipAdapter [7], TipAdapter [23] that adds a few task-specific parameters.

In Tab 1, KeepFIT trained by MM-Retinal and 50% flair achieves competitive performance across the board. Key insights include: 1) Large datasets may introduce noise and diminish transfer effectiveness. 2) KeepFIT performs better when trained by MM-Retinal and flair than by large datasets like synfundus-1M and fliar, underscoring MM-Retinal's superior expert knowledge for model generalization ability and transferability.

4.3 Comparison of Transferability on Unseen Downstream Datasets

To assess KeepFIT's transferability, we fine-tuned it on six unseen datasets. For CFP, we added and fine-tuned a fully connected layer to the image encoder and keep the other parts frozen. We conducted five fine-tuning settings, including 20%, 40%, 60%, and 80% of the data for training, the rest 20% for testing or followed the official dataset partition. For FFA, due to dataset scarcity, we utilized image captioning for assessment following FFA-IR [10], using ResNet to extract images features and Transformer-based decoder for caption generation.

Table 2: Comparison of transferability on unseen downstream datasets and ablation study. FT refers to finetune.

(a) Comparison of transferability on unseen downstream datasets(CFP)

Type	Method	Data	Data Size	FT(20%-20%)			FT(train-val)	FT(80%-20%)		Avg
				REFUGE	FIVES	ODIR200×3	TAOP	AMD		
				ACA	ACA	ACA	ACA	AUC	F1	
MTL	MoE	flair	278,348	0.543	0.364	0.609	0.239	0.539	0.405	0.450
MIM	RETFound	RETFound	904,170	0.809	0.765	0.907	0.697	0.945	0.805	0.821
CLIP	FLAIR	flair	278,348	0.831	0.835	0.875	0.468	0.963	0.960	0.822
		flair+syn	1,278,366	0.847	0.842	0.890	0.549	0.952	0.953	0.839
	KeepFIT	flair	278,348	0.837	0.838	0.903	0.556	0.951	0.957	0.840
		flair+syn	1,278,366	0.832	0.842	0.890	0.579	0.976	0.969	0.848
		50%flair+MM	141,343	0.856	0.834	0.913	0.700	0.962	0.962	0.871
		flair+MM	280,517	0.861	0.851	0.915	0.684	0.971	0.966	0.875

(b) Comparison of transferability on unseen downstream datasets(FFA)

Model	Data	Data Size	B1	B2	B3	B4	Meter	Rouge	Cider	Avg
CNN + T	FFA-IR	1,048,584	0.321	0.211	0.154	0.122	0.198	0.268	0.283	0.222
CNN + T	FFA-IR+MM	1,050,531	0.363	0.244	0.171	0.127	0.149	0.302	0.314	0.239

(c) Ablation Study(MHCA denotes multi-head cross attention)

Model	Revision	Mixed	Fusion	FT(20%-20%)			FT(train-val)	FT(80%-20%)		Avg
	MHCA	/	MHCA	REFUGE	FIVES	ODIR200×3	TAOP	AMD		
				ACA	ACA	ACA	ACA	AUC	F1	
KeepFIT	✓			0.803	0.841	0.903	0.640	0.964	0.884	0.839
		✓		0.832	0.851	0.907	0.675	0.965	0.905	0.856
		✓	✓	0.832	0.852	0.900	0.673	0.963	0.863	0.847
	✓	✓	✓	0.822	0.851	0.905	0.694	0.964	0.966	0.867
	(Ours) ✓	✓		0.861	0.851	0.915	0.684	0.971	0.966	0.875

As shown in Tab 2(a), KeepFIT trained on MM-Retinal and flair achieves SOTA in almost all the unseen downstream datasets. Similarly in Tab 2(b), the performance of FFA modality is improved when trained on both MM-Retinal and FFA-IR. More results are provided in the supplementary material.

4.4 Ablation Study

We ablated KeepFIT on all flair and MM-Retinal from three aspects. 1) Image similarity-guided text revision: assess the necessity of revising the text features of public datasets by the guidance of image similarity. 2) Mixed training strategy: test the necessity of including data from two sources in one batch. 3) Text fusion module: we substituted text revision with a text fusion module. It integrates knowledge extracted from MM-Retinal by multi-head cross attention into public dataset text features via residual connections.

Table 2(c) shows that image similarity-guided text revision and mixed training strategy are essential for the performance improvement. However, the text fusion module makes minimal contribution, suggesting that text revision is more effective at injecting knowledge than text fusion.

5 Conclusion

In this work, we built a multi-modal MM-Retinal dataset, with high-quality fundus image-text expertise. We also proposed KeepFIT, a vision-language pretraining framework enhancing expert knowledge infusion. Experimental results highlight its transferability to unseen datasets and generalization ability on few-shot and zero-shot scenarios. We expect this work will open up unexplored topics in fundus research, such as building multimodal knowledge graphs of fundus images, and high-quality text-to-image generation.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Grants No 62106043, 62172228), and the Natural Science Foundation of Jiangsu Province (Grants No BK20210225). We're also deeply grateful to Yu-Ang Yao, Minqi Gao, Junkai Chen, Jiaqi Li, Zimeng Zhu, and Jiaqi Xu, who have been instrumental in the construction of the MM-Retinal dataset.

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

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology **3**(3), 509–516 (2009)
3. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology **33**(3), 231–234 (2014)
4. Diao, S., Su, J., Yang, C., Zhu, W., Xiang, D., Chen, X., Peng, Q., Shi, F.: Classification and segmentation of oct images for age-related macular degeneration based on dual guidance networks. Biomedical Signal Processing and Control **84**, 104810 (2023)
5. Fu, H., Li, F., Orlando, J.I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X.: Palm: Pathologic myopia challenge (2019). `https://doi.org/10.21227/55pk-8z03`, `https://dx.doi.org/10.21227/55pk-8z03`
6. Fu, H., Li, F., Orlando, J.I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X.: Adam: Automatic detection challenge on age-related macular degeneration (2020). `https://doi.org/10.21227/dt4f-rt59`, `https://dx.doi.org/10.21227/dt4f-rt59`

7. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
8. Lei, J., Dai, L., Jiang, H., Wu, C., Zhang, X., Zhang, Y., Yao, J., Xie, W., Zhang, Y., Li, Y., et al.: Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. arXiv preprint arXiv:2309.06828 (2023)
9. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10571–10580 (2019)
10. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al.: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
11. Li, X., Zhou, Y., Wang, J., Lin, H., Zhao, J., Ding, D., Yu, W., Chen, Y.: Multi-modal multi-instance learning for retinal disease recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2474–2482 (2021)
12. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
13. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
14. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 (2023)
15. 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)
16. Shang, F., Fu, J., Yang, Y., Huang, H., Liu, J., Ma, L.: Synfundus: A synthetic fundus images dataset with millions of samples and multi-disease annotations. arXiv preprint arXiv:2312.00377 (2023)
17. 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. arXiv preprint arXiv:2308.07898 (2023)
18. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Wang, X., Ju, L., Zhao, X., Ge, Z.: Retinal abnormalities recognition using regional multitask learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 30–38. Springer (2019)
21. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)

22. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
23. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European Conference on Computer Vision. pp. 493–510. Springer (2022)
24. Zhao, Z., Zhang, K., Hao, X., Tian, J., Chua, M.C.H., Chen, L., Xu, X.: Bira-net: Bilinear attention net for diabetic retinopathy grading. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1385–1389. IEEE (2019)
25. Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2079–2088 (2019)
26. Zhou, Y., Yang, G., Zhou, Y., Ding, D., Zhao, J.: Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 704–713. Springer (2023)
27. 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)
28. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)