# UrFound: Towards Universal Retinal Foundation Models via Knowledge-Guided Masked Modeling

Kai Yu[1], Yang Zhou[1(✉)], Yang Bai[1], Zhi Da Soh[2], Xinxing Xu[1], Rick Siow Mong Goh[1], Ching-Yu Cheng[2,3], and Yong Liu[1]

[1] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, 138632, Singapore, Republic of Singapore
zhou_yang@ihpc.a-star.edu.sg
[2] Singapore Eye Research Institute, Singapore National Eye Centre
[3] Centre for Innovation and Precision Eye Health, National University of Singapore, Singapore

**Abstract.** Retinal foundation models aim to learn generalizable representations from diverse retinal images, facilitating label-efficient model adaptation across various ophthalmic tasks. Despite their success, current retinal foundation models are generally restricted to a single imaging modality, such as Color Fundus Photography (CFP) or Optical Coherence Tomography (OCT), limiting their versatility. Moreover, these models may struggle to fully leverage expert annotations and overlook the valuable domain knowledge essential for domain-specific representation learning. To overcome these limitations, we introduce UrFound, a retinal foundation model designed to learn universal representations from both multimodal retinal images and domain knowledge. UrFound is equipped with a modality-agnostic image encoder and accepts either CFP or OCT images as inputs. To integrate domain knowledge into representation learning, we encode expert annotation in text supervision and propose a knowledge-guided masked modeling strategy for model pre-training. It involves reconstructing randomly masked patches of retinal images while predicting masked text tokens conditioned on the corresponding image. This approach aligns multimodal images and textual expert annotations within a unified latent space, facilitating generalizable and domain-specific representation learning. Experimental results demonstrate that UrFound exhibits strong generalization ability and data efficiency when adapting to various tasks in retinal image analysis. By training on ∼180k retinal images, UrFound significantly outperforms the state-of-the-art retinal foundation model trained on up to 1.6 million unlabelled images across 8 public retinal datasets. Our code and data are available at https://github.com/yukkai/UrFound.

**Keywords:** Domain expert knowledge · Masked modeling · Retinal image understanding · Multimodal foundation model.

## 1 Introduction

Foundation models (FMs) are large, powerful artificial intelligence (AI) models pre-trained on vast amounts of unlabeled data. By learning fundamental patterns and relationships within diverse data, FMs gain the ability to adapt to diverse downstream tasks with minimal additional training [20]. Notable examples of FMs, such as CLIP [14], SAM [8], and GPT4 [12], have demonstrated impressive generalization capabilities in various real-world scenarios.

Medical FMs are a specialized type of FM designed for the medical domain [11,19], representing one of the most notable advancements in medical AI. Among these, Medical Vision-Language pre-training stands out as a specific solution that improves medical image analysis by learning domain-specific features from medical images paired with corresponding clinical descriptions or reports [10,18]. Recent medical FMs have focused heavily on radiology, particularly chest X-rays [16,17]. For retinal FMs, RETFound [20] has been proposed, which is pre-trained on 1.6 million retinal images using Masked Autoencoders (MAE). Another notable example is FLAIR [15], a vision-language model that leverages the CLIP architecture to enhance performance in retinal image analysis, supporting zero-shot and few-shot inference through text supervision. Unlike task-specific models that may yield sub-optimal results in the presence of domain shifts, retinal FMs demonstrate robust generalization capabilities across different retinal datasets and tasks. This presents an attractive solution to enhance model efficacy and reduce the annotation burden on experts, thereby enabling widespread clinical AI applications in retinal imaging.

Albeit impressive, existing retinal FMs are restricted to processing a single imaging modality, such as Colour Fundus Photography (CFP) and Optical Coherence Tomography (OCT). In clinical ophthalmology, diagnosis often involves multiple modalities, including CFP, OCT, and Fundus Fluorescence Angiography (FFA) images. This requires training separate FMs for each modality, resulting in higher maintenance costs and hindering the acquisition of complementary information across modalities. The question arises: Can a retinal FM be developed to process multiple modalities? Moreover, expert domain knowledge, often in the form of labels or medical reports, is crucial for effective retinal image analysis. It guides models in capturing clinically relevant information, ensuring clinical significance in real-world healthcare scenarios. However, current retinal FMs struggle to fully leverage expert annotations, potentially hindering specialized representation learning. Another question arises: Can domain knowledge be incorporated into a retinal FM for better generalization ability?

To address the research problems mentioned above, we introduce UrFound, a universal retinal FM designed to learn versatile representations from both multimodal retinal images and domain knowledge. UrFound employs a modality-agnostic image encoder for processing CFP or OCT images and integrates domain knowledge from categorical labels and clinical descriptions through text supervision. To achieve this, we convert expert annotations into detailed clinical descriptions and propose a knowledge-guided masked modeling strategy for UrFound pre-training. This strategy includes a masked image modeling branch

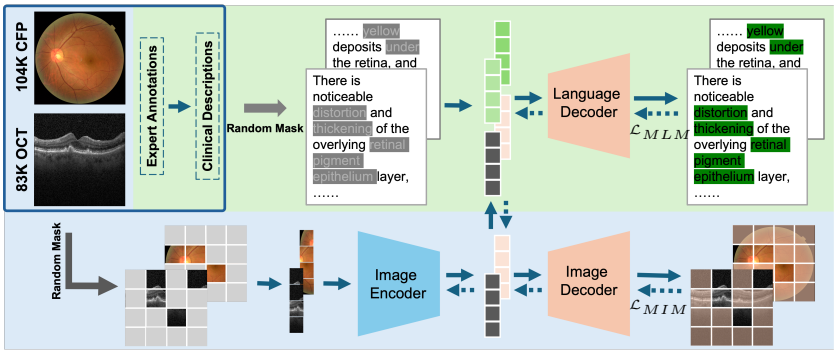



Fig. 1: Knowledge-guided masked modeling framework for UrFound pre-training. Solid arrows represent data flow, while dashed arrows indicate gradient flow.

to reconstruct randomly masked patches of retinal images, and a conditional masked language modeling branch to predict masked word tokens based on the corresponding retinal image. This approach aligns multimodal images and textual expert annotations within a unified latent space, facilitating domain-specific representation learning.

Empirically, we find that incorporating domain knowledge into the retinal FM through text supervision enhances generalization ability. Furthermore, UrFound captures information from CFP and OCT images and performs well with both modalities. Despite being pre-trained on a relatively small dataset of 180k retinal images with expert annotations, UrFound significantly outperforms state-of-the-art (SOTA) retinal FMs trained on up to 1.6 million unlabeled images across eight public retinal datasets. This demonstrates the effectiveness of multimodal images and domain knowledge in training powerful retinal FMs.

Our contribution is threefold: 1. We propose UrFound, a universal retinal foundation model capable of processing CFP and OCT images while incorporating domain knowledge from expert annotations. 2. We introduce a knowledge-guided masked modeling strategy that unifies the pre-training from multimodal images and clinical descriptions, effectively integrating domain knowledge. 3. We provide comprehensive evaluations, comparing UrFound with SOTA retinal FMs across eight public retinal datasets.

## 2 The UrFound Model

In this section, we propose UrFound, a retinal FM designed for CFP and OCT images, as the initial step toward developing universal retinal FMs. UrFound is trained with guidance from expert annotations, which can take the form of categorical labels, clinical descriptions, or any other formats that can be encoded in text supervision. UrFound aims to learn domain-specific representations by reconstructing masked patches of a retinal image while predicting masked word tokens of textual domain knowledge conditioned on the unmasked image patches.

Fig. 1 provides an overview of the UrFound model. UrFound has an image encoder that learns the latent representation of retinal images as well as two decoders that reconstruct the original retinal image and predict the word tokens of the associated clinical descriptions from the latent representation, respectively. The input of the image encoder can be either a CFP or OCT image. During pre-training, we apply masked image modeling to randomly mask certain patches of the input image. Then, the rest unmasked image patches are fed into the image encoder to obtain their embeddings. These embeddings are then forwarded to the image decoder to reconstruct the masked image patches, aiding the model in capturing versatile and informative visual features.

Similar to masked image modeling, we apply conditional masked language modeling to replace certain portions of word tokens of the clinical descriptions with the mask token. The language decoder is then trained to predict the original identity of the masked tokens based on both the unmasked words and the latent image representation from the image encoder. This approach encourages the model to recognize and comprehend the relationships between the retinal image and fine-grained medical knowledge. It serves to bridge the gap between visual features and textual information, integrating domain knowledge from the descriptions into the latent image representation.

### 2.1 Knowledge-guided Masked Modeling

Formally, given a retinal image $X$, it is first reshaped into $n$ patches with the patch size $s$ (e.g., $16 \times 16$ in ViT [5]). A random mask $M \in \{0,1\}^n$ is generated with the mask ratio $\rho$ where 1 indicates a masked patch and 0 indicates an unmasked patch. The masked image $\tilde{X}$ is obtained as: $\tilde{X}_i = X_i \cdot (1 - M_i) + X_0 \cdot M_i, \forall i \in \{1, \cdots, n\}$, where $X_0$ represents the image [MASK] token. Let $f(\cdot)$ be the image encoder that maps each image patch to a latent representation $\mathbf{z}_i = f(\tilde{X}_i)$, and $g^v(\cdot)$ be the image decoder that reconstructs the original image patch $X_i$ from the latent representation. Then the mask imaging modeling (MIM) can be achieved by minimizing the following mean square error (MSE) loss:

$$\mathcal{L}_{MIM} = \sum_{i=1}^{n} M_i \cdot ||X_i - g^v(\mathbf{z}_i)||_2^2, \tag{1}$$

which measures the differences between the reconstructed and original image patches. We adopt the high-resolution trick in [19] to let the model reconstruct high-resolution patches at $2\times$ the input resolution, which allows the model to learn more detailed local features.

For conditional masked language modeling (MLM), the input text is transformed into a sequence of tokens $W = [w_1, \cdots, w_L]$, where $L$ is the sequence length. Then, a certain percentage of tokens in the sequence are randomly replaced with a special [MASK] token, leading to a masked set $\mathcal{W}_\mathcal{M}$ and an unmasked set $\mathcal{W}_\mathcal{N}$. Let $\mathbf{z}$ be the average pooling of the unmasked image patch representations, and $h(\cdot)$ be the text decoder to restore the masked text tokens.

The objective of MLM is to minimize the negative log-likelihood function as follows:

$$\mathcal{L}_{MLM} = -\sum_{w_i \in \mathcal{W}_{\mathcal{M}}} \log P(h(w_i)|\{h(w_j), w_j \in \mathcal{W}_{\mathcal{N}}\}, \mathbf{z}), \tag{2}$$

which predicts the original identities of those masked tokens based on the surrounding context and the latent image representation. The total pre-training objective function of the UrFound model is $\mathcal{L} = \mathcal{L}_{MIM} + \mathcal{L}_{MLM}$. After pre-training, the decoders are discarded and the encoder $f(\cdot)$ can be fine-tuned with a small number of data for specific downstream tasks for retinal image analysis.

### 2.2 Text Preparation

For retinal images, the majority of publicly available expert annotations come in the form of categorical labels rather than text. To maximize the utilization of domain knowledge for pre-training, we follow FLAIR [15] to enhance categorical image labels by augmenting relevant medical findings sourced from established knowledge bases and clinical literature. For instance, the label "drusens" might be described as "yellow deposits under the retina" or "numerous uniform round yellow-white lesions". Each label may have a varying number of descriptions. During pre-training, we randomly select one of these descriptions for samples in each batch, enhancing the diversity and robustness of the text supervision.

### 2.3 Multimodal Image Processing

UrFound directly learns representations for CFP and OCT images using a modality-agnostic encoder. We have observed that this straightforward implementation performs well and achieves superior generalization, particularly when training is guided by domain knowledge through masked modeling. We also explore variants that use separate patch embedding layers, encoders, and decoders for CFP and OCT imaging, respectively, while such modifications do not lead to better results in our experiments.

## 3 Experiments

In this section, we assess the performance of UrFound compared to the SOTA retinal FMs and conduct comprehensive experiments to address the following key questions: **Q1.** Can the imaging modalities of CFP and OCT be encoded in a universal FM? **Q2.** Does domain knowledge improve the generalization ability of FMs? **Q3.** Do CFP and OCT images contain supplementary information that helps representation learning? **Q4.** How well do retinal FMs perform in terms of data efficiency? **Q5.** How effective are retinal FMs in adapting to downstream tasks compared with task-specific models?

Table 1: Performance of retinal FMs on different datasets (**best**, second best).

| Dataset | MAE | | FLAIR | | RETFd-CFP | | RETFd-OCT | | UrFound | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ROC | PRC | ROC | PRC | ROC | PRC | ROC | PRC | ROC | PRC |
| **APTOS** | 94.06 | 67.59 | 92.68 | 62.20 | 94.26 | **71.87** | 87.56 | 53.76 | **94.86** | 71.64 |
| **IDRID** | 79.24 | 43.65 | 80.88 | 49.32 | 83.33 | 51.13 | 59.29 | 28.66 | **85.22** | **57.73** |
| **Messidor** | 84.21 | 48.76 | 81.88 | 48.32 | 86.40 | 58.59 | 65.89 | 28.59 | **88.22** | **60.78** |
| **PAPILA** | 62.85 | 47.48 | 74.80 | 59.30 | 74.36 | 57.27 | 51.67 | 35.03 | **78.32** | **62.54** |
| **GF** | 93.09 | 83.17 | 78.87 | 59.60 | 95.68 | **88.18** | 87.61 | 70.85 | **95.75** | 88.01 |
| **JSIEC** | 98.46 | 81.78 | 93.53 | 52.65 | 99.39 | 86.95 | 88.44 | 41.09 | **99.51** | **92.34** |
| **Retina** | 74.22 | 53.70 | 77.75 | 55.33 | 86.22 | 71.59 | 75.43 | 53.76 | **90.09** | **79.30** |
| **OCTID** | 98.67 | 95.35 | 84.52 | 60.20 | 93.85 | 82.09 | 99.39 | 97.58 | **99.55** | **97.97** |

### 3.1 Experimental Setup

We assess the capabilities of UrFound in adapting to diagnostic classification tasks with minimal additional training. In line with common practice, we add a linear classifier head on top of the learned image encoder and then fine-tune both the encoder and classifier with task-specific labels. We compare the proposed UrFound against the MAE model pre-trained on natural images as well as SOTA retinal FMs including RETFound [20] and FLAIR [15]. For these compared models, we use official checkpoints for fine-tuning. We report the area under the receiver operating curve (ROC) and the area under the precision-recall curve (PRC) as evaluation metrics. We choose the best checkpoints with the highest ROC scores on the validation set for final evaluation.

**Datasets.** For pre-training, we construct a training set by collecting 25 CFP datasets and one large OCT dataset, which include 103,786 CFP images and 83,484 OCT images with expert annotations, covering a wide range of ophthalmic diseases. More details can be found in the supplementary material. We follow [15] to augment domain knowledge and transform categorical labels into textual inputs. For evaluation of fine-tuning performance, we test 8 publicly available datasets across three diagnostic classification tasks including diabetic retinopathy grading (IDRID [13], MESSIDOR [3], APTOS [7]), glaucoma detection (PAPILA [9], GF [1]), and multi-disease diagnosis (JSIEC [2], Retina, OCTID [6]).

**Implementation details.** We implement UrFound by using PyTorch on a single NVIDIA A100 GPU. We employ a Vision Transformer (ViT-base) with 12 Transformer blocks and a patch embedding layer as the retinal image encoder. We utilize 8 and 6 Transformer blocks as the image and text decoders, respectively. In the pre-training stage, we initialize UrFound with the MAE model and use the tokenizer of BERT-Base [4] to convert clinical descriptions into word tokens. We use a mask ratio of 0.75 for image modeling and 0.5 for language modeling. We resize the input image to 224×224 both in the pre-training stage and fine-tuning stage. Random horizontal flip and random crop are implemented in the pre-training stage. And random horizontal flip, and color jitter for data augmentation in the fine-tuning stage, each with a probability of 0.5. The total training epoch is set to 200 with a warm-up period of 40 epochs. The learning

Table 2: Performance of UrFound and its variants (**best**, second best).

| Dataset | W/O Text Supervision | | | | | | W/ Text Supervision | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CFP | | OCT | | CFP+OCT | | CFP | | OCT | | CFP+OCT | |
| | ROC | PRC | ROC | PRC | ROC | PRC | ROC | PRC | ROC | PRC | ROC | PRC |
| **APTOS** | 93.81 | 65.93 | 89.92 | 56.11 | 94.01 | 67.58 | 94.36 | 68.68 | 90.40 | 56.38 | **94.86** | **71.64** |
| **IDRID** | 79.47 | 45.44 | 69.10 | 35.32 | 79.51 | 44.59 | 84.64 | 55.46 | 66.28 | 31.09 | **85.22** | **57.73** |
| **Messidor** | 84.76 | 52.86 | 69.10 | 30.73 | 84.28 | 50.47 | 86.24 | 58.28 | 71.21 | 32.34 | **88.22** | **60.78** |
| **PAPILA** | 69.13 | 52.36 | 47.21 | 35.47 | 69.65 | 53.68 | 73.45 | 54.92 | 56.59 | 38.19 | **78.32** | **62.54** |
| **GF** | 93.84 | 84.50 | 89.01 | 73.30 | 93.48 | 83.70 | 95.16 | 87.17 | 89.61 | 73.62 | **95.75** | **88.01** |
| **JSIEC** | 98.72 | 88.72 | 91.71 | 50.60 | 99.08 | 85.44 | 99.48 | **92.84** | 92.35 | 51.33 | **99.51** | 92.34 |
| **Rtina** | 88.17 | 76.01 | 70.29 | 48.78 | 87.08 | 74.34 | 88.50 | 75.77 | 81.40 | 61.81 | **90.09** | **79.30** |
| **OCTID** | 98.40 | 95.60 | 99.37 | 97.35 | **99.59** | 97.88 | 98.05 | 94.56 | 99.28 | 95.33 | 99.55 | **97.97** |

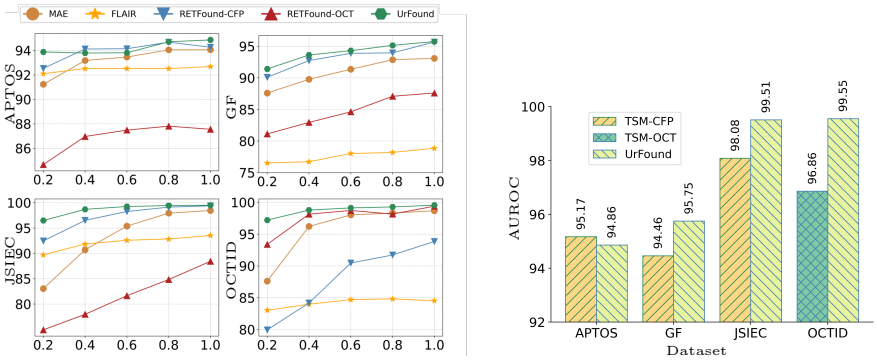
rate is set to 1.5e-4, and the batch size is set to 128. In the fine-tuning stage, the learning rate is adjusted to 1e-4, the batch size is reduced to 16, and the training epoch is set to 50 with a warm-up period of 10 epochs.

### 3.2 Main Results

Table 1 shows the classification results of the compared retinal FMs fine-tuned for various retinal disease diagnosis tasks, and statistical significance analysis is available in the supplementary material. It can be observed that retinal FMs such as RETFound-CFP and UrFound achieve better results than MAE in all the cases, which demonstrates the effectiveness of retinal FMs in learning generalizable representations for retinal imaging analysis. UrFound consistently outperforms the second-best method, RETFound. This superiority can be attributed to the integrated domain knowledge in UrFound through text supervision. In contrast, although FLAIR also leverages domain knowledge, it does not perform well and lags behind MAE in some cases. This is possibly because FLAIR focuses on image-text alignment rather than capturing the visual features of retinal images. It results in a sub-optimal image encoder for image understanding in the pretrain-finetune setting.

RETFound-CFP and FLAIR are designed specifically for CFP images, exhibiting subpar performance when applied to OCT images. Similarly, RETFound-OCT yields the poorest results on CFP datasets. In contrast, UrFound showcases its superiority in processing both CFP and OCT modalities. It achieves this by learning universal and comprehensive representations that span across modalities, demonstrating its capability to effectively handle diverse imaging types.

**Impact of multimodal imaging and domain knowledge.** To investigate how multimodal data and domain knowledge affect the performance of UrFound, we compared UrFound against its single-modality variants, either with or without domain knowledge. As shown in Table 2, without text supervision, UrFound trained from CFP+OCT images achieves reasonably good results on both CFP and OCT datasets. This indicates that it is promising to learn universal FMs for multiple retinal imaging modalities (**Q1**). Furthermore, the inclusion of text supervision significantly enhances the performance of UrFound,




(a) (b)

Fig. 2: (a) Data efficiency of retinal FMs, Axes X and Y are the percentage of training data used and the corresponding AUROC, respectively. (b) Comparison of UrFound with tasks-specific models in AUROC with different datasets.

which demonstrates the effectiveness of domain knowledge in learning domain-specific and generalizable representations (**Q2**). With text supervision, UrFound trained from CFP+OCT images outperforms its single-modality counterparts, which suggests that CFP and OCT images contain supplementary information beneficial for improved representation learning (**Q3**).

**Data efficiency.** Fig. 2a shows the classification results of the compared FMs at different percentages of training data on the APTOS, GF, JSIEC, and OCTID datasets. UrFound outperforms other retinal FMs in most settings and demonstrates a more significant advantage when fewer data are used for training (**Q4**). It is noteworthy that UrFound is pre-trained on ∼180k retinal images, a significantly smaller dataset compared to existing retinal FMs such as RETFound-CFP, which is trained with over 900k CFP images. These demonstrate the superior data efficiency of UrFound, making it well-suited for retinal imaging analysis with limited annotations.

**Comparison with task-specific models.** To verify the advantages of transforming expert annotations into text for pre-training, we compare UrFound with task-specific models (TSMs) that are first trained with the class labels of specific classification tasks (as a supervised way of pre-training) and then adapted to test datasets for evaluation. Specifically, we test two TSMs: one for diabetic retinopathy grading (TSM-CFP) and the other for OCT disease classification (TSM-OCT). TSM-CFP is trained on all the CFP pre-training datasets for diabetic retinopathy grading, comprising 51,556 CFP images and labels of five classes. TSM-OCT is trained on all the OCT pre-training datasets, which include 83,484 OCT images and labels of four classes. In total, the data used for training TSMs account for 72% of those used for pre-training UrFound.

Fig. 2b presents the ROC scores of TSMs and UrFound on the APTOS, GF, JSIEC, and OCTID datasets, where each dataset corresponds to different

downstream tasks. UrFound and TSM-CFP obtain similar results on the ATPOS dataset. This is expected because the task of APTOS aligns with the training of TSM-CPT. UrFound consistently outperforms TSMs on other datasets. This suggests that TSMs lack the flexibility to learn generalizable representations for various tasks. In contrast, UrFound benefits from expert annotations via text supervision, offering a more effective approach to integrating valuable domain knowledge in representation learning (**Q5**).

## 4 Conclusion

We proposed UrFound, a **U**niversal **r**etinal **Found**ation model, which features a modality-agnostic image encoder and utilizes knowledge-guided mask modeling as a pre-training objective, allowing it to learn generalizable representations from both multimodal images and expert annotations. Through comprehensive experiments on 8 public retinal datasets, we demonstrated its strong generalization ability and data efficiency in adapting to various downstream tasks. Nevertheless, UrFound has two limitations: 1. UrFound is designed to process CFP and OCT images while there exist other retinal imaging modalities such as FFA. 2. UrFound is pre-trained on a relatively small dataset with disease labels as expert annotations. In practice, many unlabeled data are available for pre-training.

**Acknowledgments.** This work was supported in part by the National Research Foundation of Singapore under its AI Singapore Programme (AISG) under Award AISG2-TC-2021-003, and in part by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme under Project A20H4b0141, the RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP) (grant no. H20C6a0032), the 2022 Horizontal Technology Coordinating Office Seed Fund (Biomedical Engineering Programme – BEP RUN 3, grant no. C221318005) and partially supported by A*STAR Central Research Fund "A Secure and Privacy-Preserving AI Platform for Digital Health".

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

## References

1. 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)
2. Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature communications **12**(1), 4828 (2021)
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. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL. pp. 4171–4186 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of ICLR (2021)
6. Gholami, P., Roy, P., Parthasarathy, M.K., Lakshminarayanan, V.: OCTID: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (2020)
7. Karthik, Maggie, S.D.: Aptos 2019 blindness detection (2019), `https://kaggle.com/competitions/aptos2019-blindness-detection`
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of ICCV. pp. 4015–4026 (2023)
9. 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)
10. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
11. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
12. OpenAI: Chatgpt-4. `https://openai.com/chatgpt` (2023), accessed: 2024-02-01
13. 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)
14. 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: Proceddings of ICML. pp. 8748–8763 (2021)
15. 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)
16. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
17. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
18. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
19. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: Proceddings of ICLR. pp. 1–16 (2023)
20. 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)