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

1 Details of MM-Retinal Dataset Construction

To advance multi-modal fundus foundation model research and foster expert knowledge integration in learning fundus image analysis models, we build MM-Retinal, a high-quality image-text paired fundus dataset that comprises CFP, FFA, and OCT modalities. We design a semi-automatic collection procedure to improve construction efficiency, which consists of four steps: 1) image-text pair collection; 2) image-text alignment; 3) modality classification; 4) text cleaning and bilingual translation.

Step 1: Image-Text Pair Collection. Textual reports for fundus diagnosis are typically not accompanied with images in clinical process. Therefore, unlike X-ray and CT images, it is challenging to directly obtain image-text pairs from fundus clinical reports. To address this, we collect image-text pairs from four diagram books illustrating ocular fundus diseases with high-quality expert captions. First, image-text pairs are captured from the books. In cases that one figure corresponds to one caption, we simply capture them in one screenshot. For the cases that multiple sub-figures correspond to a piece of caption, both the text and its corresponding sub-figures are captured together, as shown in Fig. 1(a) in the main body. Note that we keep the resolution of each image in the pair no less than 800×800. Afterwards, these screenshots are filled into the program we designed to parse images by Adobe and extract texts by OCR technique. Additionally, the color of a certain book suffer a dark tone. We implemented a dehazing operation based on Gamma transformation to correct the color of this book to match the color distribution of other books.

Step 2: Image-Text Alignment. As mentioned earlier, not all the initially extracted images and texts by the program are very well aligned, as there are some sub-figures correspond to one caption within a screenshot. Thus, for those failure cases, we use regular expression matching to split them. Specifically, if the text matches "Figure No.", it indicates the beginning of the text of a new image-text pair. If it matches "Letter.", it indicates the beginning of the text of a new subfigure-text pair. As for the separation of sub-figures, we apply Adobe to automatically implement subgraph segmentation.

Step 3: Modality Classification. Since the fundus images in the books include multiple modalities, we classified them into CFP, FFA, and OCT, separately, which are the mainstream modalities in fundus domain, and exclude others due to their limited samples. Specifically, we first employ K-means to categorize images into three categories based on their color histograms. It effectively separates the CFP modality which characterized by the distinct color pattern. The other two categories are the mixture of FFA, OCT and images from other modalities. Subsequently, we select a reference image from both OCT and FFA respectively. The classification of remaining images is based on the distance between their color histograms and the references' color histograms. This process results in a precise separation of FFA and OCT modality images.

Step 4: Text Cleaning and Bilingual Translation. The text extraction may have some OCR recognition errors or incomplete issues, so we correct the texts manually. Moreover, to enhance the relevance within an image-text pair, we remove irrelevant information from the text, such as the index of the corresponding image. We also make modifications to sentence inaccuracies. For example, the original text describes multi subgraphs (e.g. 'both eyes show'), but after separating them, the text description should be appropriately adjusted to reflect details about each individual subgraph (e.g. 'left eye shows' and 'right eye shows')". In addition, since the diagram books we selected include both Chinese and English, we provide bilingual reports in English and Chinese version to standardize the language and make MM-Retinal more influential. DeepL Translator is applied to translate Chinese to English, and Tencent Translator is used to translate English to Chinese.

2 Comparison of Transferability on Unseen Downstream Datasets

We present the supplementary results of transferability on unseen downstream datasets. The metric for REFUGE, ODIR200×3 and FIVES is ACA, and for AMD is AUC. From Fig. 1, KeepFit performs better with smaller training splits, like 20%-60%. As the training data increases, KeepFit's performance slightly declines, yet it consistently outperforms most baselines. This demonstrates its superior transferability, generalization capabilities and robustness.

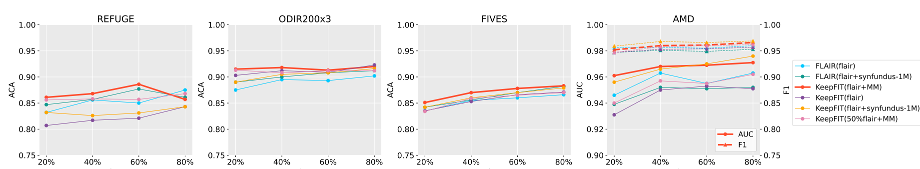

Fig. 1: Comparison of transferability on unseen downstream datasets.

3 Disease and Abnormal Changes Categories in MM-Retinal

This section provides the major categories of fundus diseases and abnormal changes in MM-Retinal with several example images in CFP, FFA and OCT modalities. These categories are summarized from the contents of four diagram books we used.

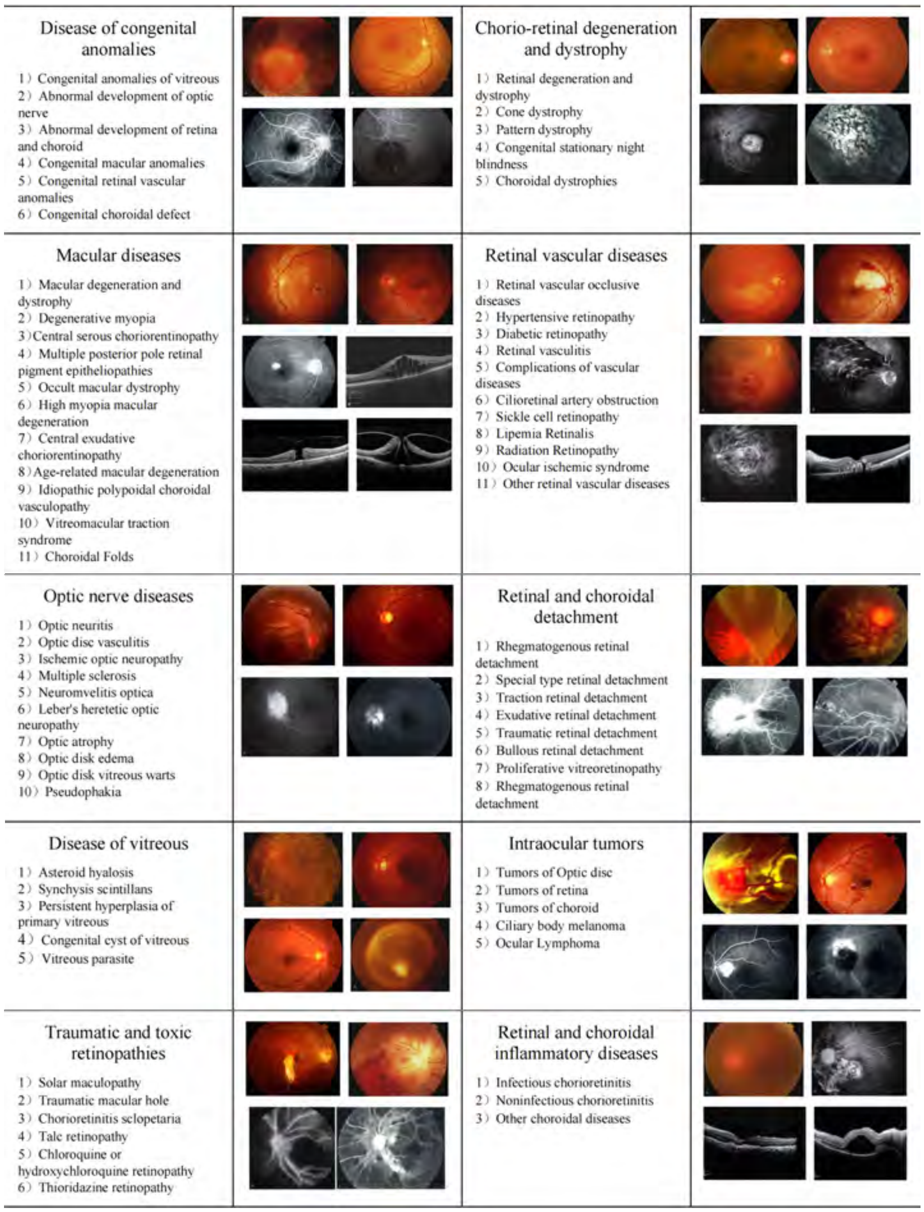

Fig. 2: Major fundus diseases in MM-Retinal

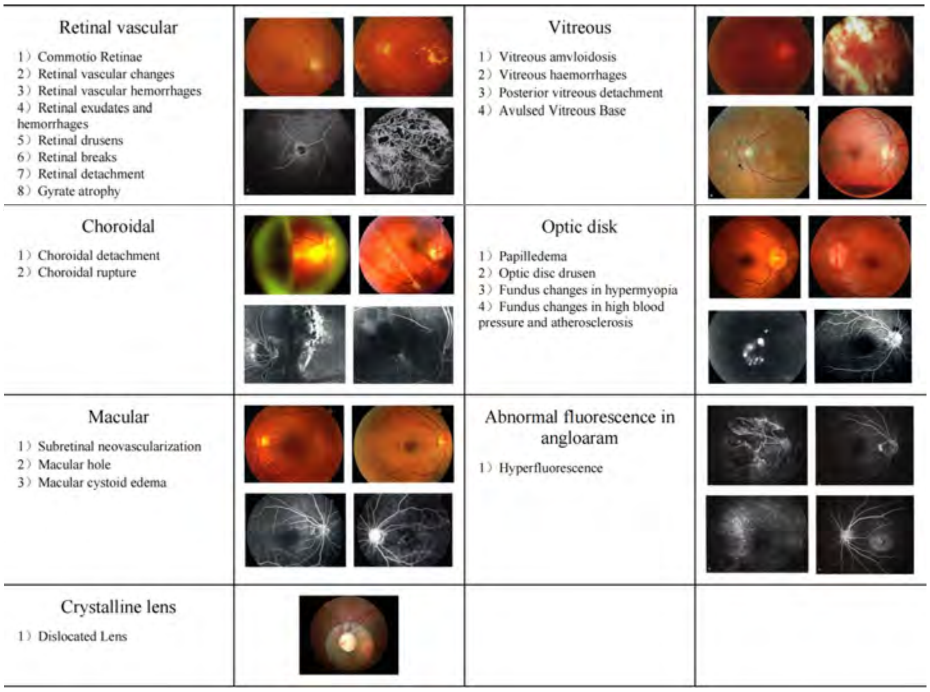

Fig. 3: Major abnormal fundus changes in MM-Retinal