基于文本增强的眼底图像多病种识别方法

熊绍奎[1]，陈世峰[2]

[1]（南方科技大学，深圳 518055）

[2]（中国科学院深圳先进技术研究院，深圳 518055）

摘要：本研究在眼科图像疾病识别中引入了视觉语言模型这一新的范式，提出了一种基于对比语言图像预训练模型的多疾病识别算法。首先，基于多个公开可用的眼底图像数据集构建了一个含有 8 个类别的多标签眼底图像数据集 MDFCD8；然后利用生成式人工智能 GPT-4 生成描述眼底图像细粒度病理特征的专家知识，解决了眼底图像数据集文本标签缺乏的问题。实验结果表明，与传统的卷积神经网络和 Transformer 网络相比，本文提出的方法在性能上分别高出 4.8%和 3.2%。同时，本文还进行了各模块的消融实验，验证了该方法的有效性，显示了视觉语言模型在眼科疾病辅助诊断领域的应用潜力。

关键词 眼底图像；多病种；对比语言预训练；专家知识

中图分类号：TP391；R77 文献标志码 A

doi: 10.12146/j.issn.2095-3135.20240422001

Multi-disease Recognition Method for Fundus Images Based On Text Enhancement

XIONG Shaokui[1], CHEN Shifeng[2]

[1] (South University of Science and Technology, Shenzhen, 518055, China)

[2](Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China)

Abstract: In this work, a new paradigm of visual language modeling is introduced in ophthalmic image disease recognition. And a multi-disease recognition algorithm based on a pre-trained model of contrasting language images is proposed. First, a new multi-labeled fundus image dataset MDFCD8 containing 8 categories is constructed based on several publicly available fundus image datasets. Then, the generative artificial intelligence GPT-4 is utilized to generate expert knowledge describing the fine-grained pathological features of fundus images, which solves the problem of the lack of text labels in fundus image datasets. The experimental results showed that, the proposed method outperforms the traditional convolutional neural network and Transformer network by 4.8% and 3.2%, respectively. This study also conducted ablation experiments on each module to validate the effectiveness of the method, and also demonstrated the potential of visual language modeling in ophthalmic disease research.

Key words: fundus images; multimorbidity; contrastive language-pretraining; expert knowledge

来稿日期：2024-04-22 修回日期：2024-05-14

基金项目：深圳市技术攻关项目（JSGG20220831105002004）

作者简介：熊绍奎，硕士研究生，研究方向为计算机视觉；陈世峰（通讯作者），副研究员，研究方向为人工智能，计算机视觉，图像视频处理，模式识别，机器学习等，E-mail：shifeng.chen@siat.ac.cn

1 引 言

眼科疾病的及时诊断至关重要。许多眼科疾病如糖尿病视网膜病变、青光眼、白内障和近视等，在早期阶段往往没有明显的症状，不易察觉，但这些疾病可能会逐渐恶化，严重影响患者的正常生活，甚至导致失明。通过定期的眼科检查，可以早期发现并诊断这些疾病，从而采取适当的治疗措施，有效控制病情的发展，保护视力，提高生活质量。

眼部疾病的及时诊断取决于眼科成像和专家检查，是劳动密集型的，容易出错并且十分耗时。基于深度学习的自动分析和诊断技术在眼科疾病分析问题中展现出卓越的性能和巨大的潜力。通过学习大量的眼科图像数据，深度学习模型能够自动识别和分类各种眼部疾病。这些模型通过学习成千上万的病例，能够捕捉到疾病特征的细微差别，从而提供比传统方法更快、更准确的诊断。现阶段对多病种眼底图像的研究，主要集中在卷积神经网络（Convolutional Neural Network, CNN）以及基于注意力机制的模型。CNN 通过其卷积层能够自动学习和提取图像中的关键特征，可以提取视网膜成像中异常的特征，从而最终准确区分出不同的眼科疾病[1-6]。例如，Sahlsten 等人[4]使用 CNN 来识别糖尿病视网膜病变和黄斑水肿分级，证明了深度学习系统可以提高筛查和诊断的成本效益。与基于 CNN 的模型相比，基于注意力机制的模型除了从图像中提取信息外，还可以揭示视网膜图像中远程像素之间的全局关联，进一步提升了眼科疾病任务的准确率[7-8]。例如，Hisham 等人[7]在他们的方法中使用预训练好的 Swin Transformer V2[9]和 DeiT[10]在眼科疾病数据集上微调，在多标签眼科疾病分类方面取得了最优的实验结果。但是，这些方法只使用了图像模态的信息，忽略了其他模态，例如文本模态。另一方面，这些方法的性能依赖于大量的训练数据，然而获取眼科疾病数据的成本较高，这限制了这些方法的性能。

视觉语言模型，例如对比语言预训练（Contrastive Language-Image Pre-training, CLIP）等[11]，通过在大量的图像−文本对上进行预训练，掌握了理解图像内容与其自然语言描述之间复杂关系的能力，为眼科疾病诊断提供了全新的研究思路。CLIP 模型即使在缺乏大量标注数据的情况下也能准确识别和分类眼科疾病特征。Silva-Rodriguez 等人[12]提出的 FLAIR 模型，通过文本监督集成专家知识以增强视网膜图像的解释性，在 37 个开放访问的分类眼底成像数据集上进行训练，表明嵌入了专家知识的模型具有更强大的泛化能力和可迁移能力。视觉语言模型显示出了令人印象深刻的迁移学习能力，在眼科疾病识别中有着巨大的潜力。

挖掘图像中的文本信息，并集成到图像分析中的思想并不新鲜。例如，Shang 等人[13]利用私有数据集合成一个高质量的图像数据集，并引入一个经过认证的 AI 辅助诊断系统来为图像注释文本，其文本内容只是简单的类别描述，不包含图像中具体的病理特征。Silva-Rodriguez 等人[12]则是从疾病对应的专家知识集中随机采样，并将采样获得的病理特征与模板结合组成文本，忽略了眼底图像真实的病理特征。本文利用生成式人工智能来生成眼底图像对应的专家知识文本，既丰富了文本的内容，又保证了文本的真实性。

本文在眼底图像多病种识别领域引入视觉语言模型这一新的范式，使用 CLIP 模型对多标签眼底图像分类问题进行研究。首先，利用公开可用的多病种眼底图像数据集，并对这些图片进行筛选和标签对齐，最终构建了一个新的含有 8 个类别的多标签眼底图像数据集。然后利用生成式人工智能 GPT-4 生成描述眼底图像细粒度病理特征的专家知识，既避免了对眼底图像专家知识随机采样存在的问题，又解决了眼底图像数据集文本标签缺乏的问题。最后，本文进行了对比实验与消融实验，证明了这种文本增强方法的有效性。与传统的卷积神经网络和基于注意力机制的网络相比，本文提出的方法具有更高的性能。

2 材料与方法

2.1 数据集收集

由于专业壁垒、隐私保护以及标记数据成本过高等原因，通常很难获得大量的医学数据。为了得到高质量的眼底彩照多疾病数据集，本文收集了四个权威机构发布的公共数据集，并将这些数据集合并成一个新的数据集。这四个数据集分别是 ARIA[14]、RFMID[15]、RFMID2.0[16]、ODIR-5K[17]。其中，ARIA 包含 143 张图像和 3 种类别的标签；RFMID 包含 3200 张图像，数据集中有 46 种标签，分别为正常和 45 种病理；RFMID2.0 包含 860 张图像，数据集中有 50 种标签，分别为正常和 49 种病理；ODIR-5K 包含 5000 张照片，数据集中有 8 种类别的标签，分别为正常、6 种常见病理和其他异常。

2.2 数据集清洗

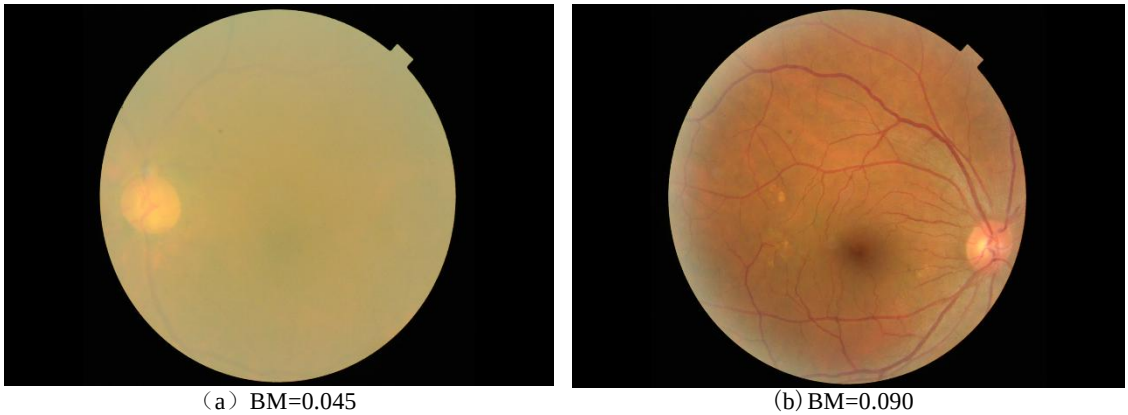
一个高质量的数据集具有以下特点：第一，数据集中的图像具有一定的质量。第二，数据集数量足够大，且每一类眼科疾病的数量足够用来训练。由于初步合成的数据集是由多个公开数据集构建的，具有不同的图像质量和标签类型。因此，需要对初步合并后的数据进行清洗，以确保整个眼底图像数据集的图像质量以及数量满足要求。

对于第一点，首先，观察原始数据集中的标签，有些图像由于设备误用、环境条件等造成眼底图像质量不佳，这些因素都会插入噪声、模糊和伪影等来降低图像的质量，从而增加不确定性和错误分类的风险。因此，先去除标签中带有“低质量”、“镜头灰尘”和“图像偏移”等注释的眼底图像，这些图像因为拍摄时光照条件差、镜头被灰尘遮蔽或抖动等外界因素影响而质量不佳。为了检测低质量图像，我们基于 Kanjar 和 Masilamani[18] 提出的模糊度量来计算图像质量分数，模糊分数计算的具体步骤是先使用 Canny 边缘检测器检测眼底图像 I 的边缘集合 E，则模糊度指标由公式（1）得到：

$$BM=\frac{\sum_{I(x,y)\in E}\sqrt{\sum_{I(x',y')\in N_{xy}}\frac{\{I(x,y)-I(x',y')\}^2}{|N_{xy}|}}}{\sum_{I(x,y)\in E}I(x,y)} \quad (1)$$

其中，N_{xy} 是 $I(x,y)$ 的八邻域，$|N_{xy}|$ 为 N_{xy} 中元素的个数。

选取一个合适的模糊分数阈值，大于或等于这个阈值则意味着图像具有较高的清晰度，而低于这个阈值意味着图像较模糊，需要去除。Rodrguez[19]等人比较视觉上评价得出的低质量图像和根据模糊度指标以 0.058 作为阈值自动确认得出的低质量图像，发现两种方法得出的低质量图像有 90%是相同的。因此，本研究以 0.058 作为最终阈值，采用自动计算模糊度指标的方法去除低质量图像。图 1 为模糊度指标不同的图像示例。



（a）BM=0.045　　（b）BM=0.090

图 1　不同模糊度

Fig. 1 Different degrees of fuzziness

对于第二点，数据集之间不仅疾病的种类不一样，标签的形式有显著差异，例如，ODIR-5K 数据集的疾病标签是基于患者标注的，而其他数据集的疾病标签是基于眼睛标注的。该数据集除了疾病标签外，还有关于患者的年龄、性别以及病灶的描述。因此，需要对齐各个数据集的标签，以便于模型训练。对于数据集中样本较少的病理，它们不会从数据增强技术中受益，而且模型的性能也会受到影响，因此，我们把这样的病理标签并入“其他”类，考虑到两个不同数据集之间的病理标签有部分重叠，以及去除样本较少的病理标签，我们最终得到的新数据集使用 8 个类用于预测，记为 MDFCD8，分别是正常、糖尿病视网膜病变、青光眼、白内障、年龄相关性黄斑变性、高血压视网膜病变、近视以及其他，图 2 展示了 8 种类别的眼底图像。MDFCD8 数据集中 8 种类别的数据分布如图 3 所示。

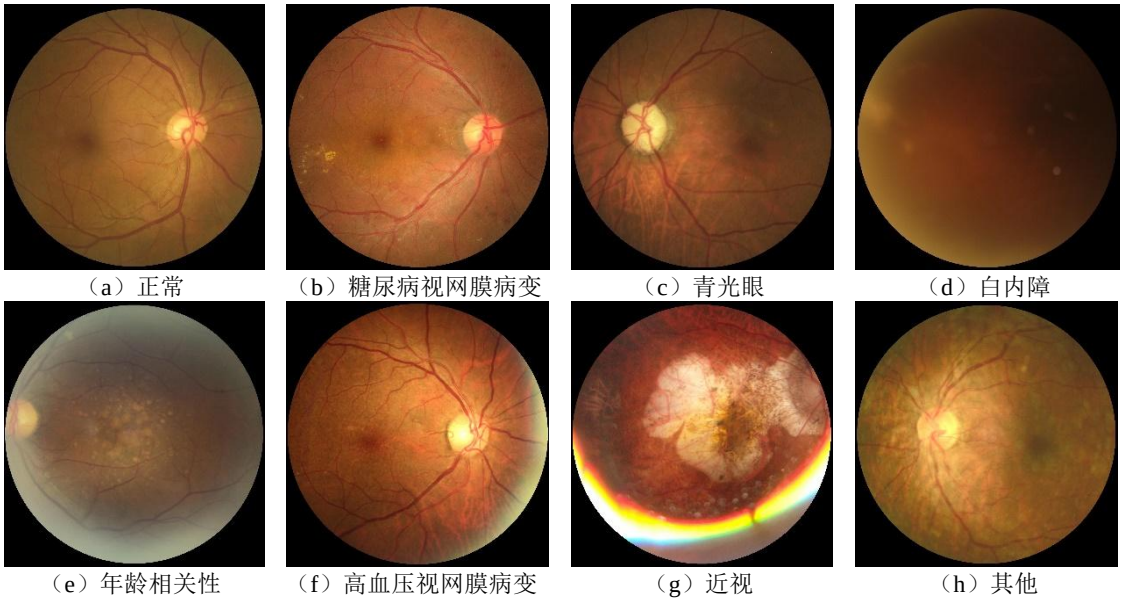


图 2　8 种类别眼底图像

Fig. 2 8 types of fundus images

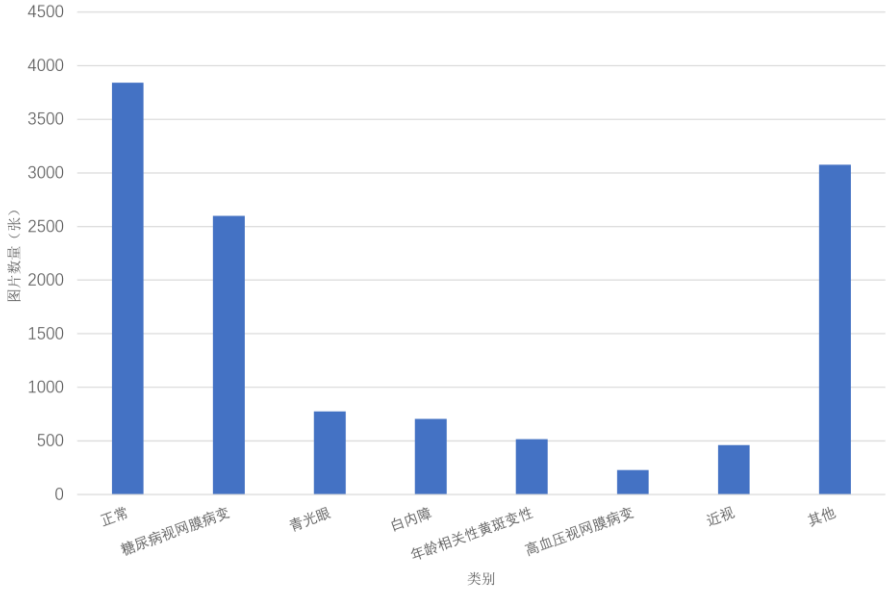


图 3　不同类别数据分布

Fig. 3 Distribution of data in different categories

3 研究方法

3.1 网络架构

本文模型整体网络结构如图 4 所示，在训练阶段，如图 4（a）所示，网络结构包括一个图像编码器和一个文本编码器，图像编码器采用 ViT[20]结构，用于提取眼底图像的特征，文本编码器采用 Transformer 结构，用于提取包含专家知识的文本的特征。将图像特征和文本特征映射到同一向量空间并求出图像与文本之间的相似度表，这张表作为监督信号来监督网络的训练。在生成文本时，如图 4（b）所示，输入眼底图像和合适的提示词，使用 GPT-4 来生成眼底图像对应的包含专家知识的文本。

在推理阶段，如图 4（c）所示，待预测图像经过固定尺寸和归一化之后，送入图像编码器进行编码，同时，待预测的各类别标签模板化后，通过文本编码器编码。编码后的图像嵌入分别与各类别的文本计算相似度，超过阈值则可以判定为图像属于该类。

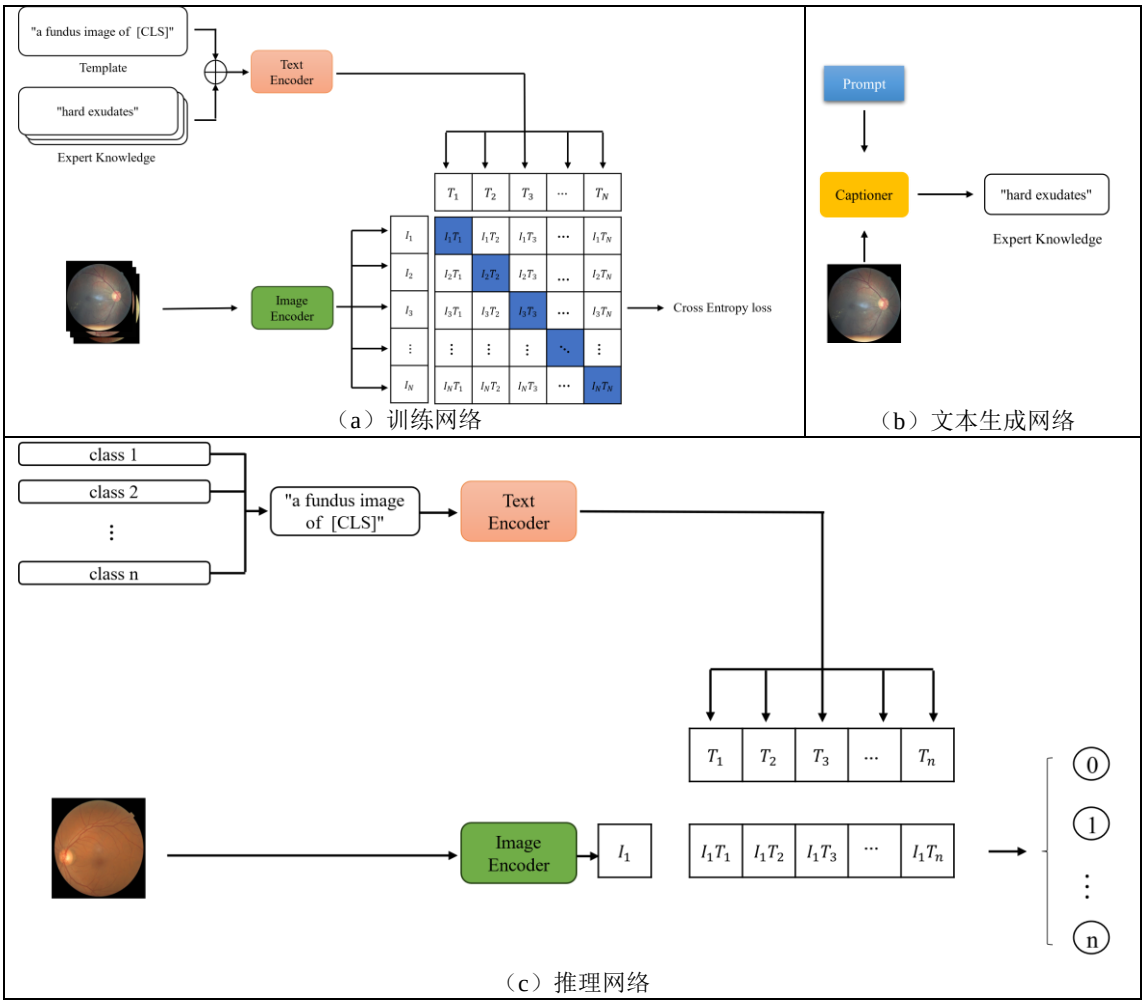


图 4　网络结构

Fig. 4　network structure

3.2 数据对扩增

本文的研究对象是眼底图像，其中，一张眼底图像可能存在多种疾病。若一张眼底图像存在 8 种类别中的 n 种类别，则这个数据样本可以表示为$\{I, L_1, L_2 ... L_n\}$。尽管一些眼部疾病之间可能会相互影响或共同进展，但是，在表现层面上，不同的眼底疾病可能有着独特的表现特征。因此，可以把一个具有多种疾病的数据对扩增为多个具有单种疾病的数据

对：$\{I, L_1\}$、$\{I, L_2\}\ldots\{I, L_n\}$。

3.3 文本增强

3.3.1 文本生成策略

在眼底图像的文本生成中，目前主要采用两种方法。第一种方法使用固定的文本模板，例如，“一张患有[类别]的眼底图像”，直接将图像的类别标签转换成文本。这种方法虽然能够保持类别名称的语义一致性，但仅仅包含病理名称，未能反映出具体的病理结构，信息量较少。第二种方法[12]是使用简单的模板引入细粒度病理特征，通过从专家知识集中随机选取特征来构建文本描述。尽管这样可以增加描述的细节，但由于文本选取的随机性，这种描述往往无法准确反映图像内容，有时甚至可能导致信息的误导。

本文对图像的文本描述增加眼底图像中存在的细粒度病理特征。然而，采用人工的方式对每一张眼底图像进行病理特征的标注成本较高，这不仅需要专业的眼科疾病相关知识，而且耗时且劳动力密集。大型生成式人工智能提供了一种新的解决思路。GPT-4 是由 OpenAI 提出的一种先进的多模态对话模型，能够理解和生成包括文本、图像在内的多种数据类型，展现出对复杂信息的深入理解能力。已有研究指出[21]，GPT-4 具备大量的医疗知识，拥有生成较为精准医学图像描述的能力，能够为医疗图像提供详细的、较为准确的文本解读，从而辅助医生和研究人员更好地理解和利用医学图像数据。因此，本文利用完善的领域专家知识，采用 GPT-4 为眼底图像生成包含病理特征的专家知识文本，从而在较低成本下构建眼部疾病多模态数据集。

采用 GPT-4 为眼底图像生成包含病理特征的文本的具体步骤如下：

（1）设定不同类别对应的专家知识集。本文自定义的数据集包含八种不同的类别，不同类别对应的专家知识描述的数量和内容各异。通过查阅相关的临床文献[22]和社区标准[23]，本文总结出不同的病理和对应专家知识描述，如表 1 所示。

（2）根据眼底图像的标签，查询专家知识集并使用合适的引导词使得 GPT-4 生成细粒度病理特征描述。引导词根据眼底图像的标签来设定，首先查询眼底图像标签对应的专家知识，引导词内容则为告诉 GPT-4 眼底图像的标签并且询问 GPT-4 目标眼底图像是否含有对应标签的专家知识集中的病理特征。若结果为是，则将该病理特征作为文本描述的一部分。

表 1　不同病理的专家知识

Table 1 Expert knowledge of different pathologies

类别	专家知识
正常	“健康”，“无病变”
糖尿病视网膜病变	“硬性渗出”，“微动脉瘤”，“出血”，“微血管异常”
青光眼	“视神经异常”，“视盘异常大小”，“视杯异常大小”
白内障	“晶状体轻度混浊”，“晶状体严重混浊”
年龄相关性黄斑变性	“小玻璃膜疣”，“大玻璃膜疣”
高血压视网膜病变	“硬性渗出”，“血管壁变化”，“视盘水肿”，“动脉狭窄”，“棉絮斑”
近视	“眼底扩张和薄化”，“脉络膜萎缩”，“Fuches 斑”，“视网膜裂孔和脱离”，“视神经盘变形”，“脉络膜新生血管”
其他	“不健康”，”“病变征兆”

以一张具有病理性近视的眼底图像为例，说明生成病理知识文本的过程。首先，根据表 1 查询病理性近视对应的专家知识集，一共有六种特征，这六种特征分别是眼底扩张和薄化、脉络膜萎缩、Fuches 斑、视网膜裂孔和脱离、视神经盘变形和脉络膜新生血管。最终输入的引导词告诉 GPT-4 眼底图像的标签为病理性近视，并询问眼底图像是否含有这六种特征，如图 5 所示。根据结果显示，这张眼底图像含有眼底扩张和薄化、脉络膜萎缩、视网膜裂孔和脱离和视神经盘变形这四种特征，其最终的文本描述为：“一张患有近视的眼底图像，伴有眼底扩张和变薄、脉络膜萎缩、视网膜裂孔和脱离以及视神经盘变形”。

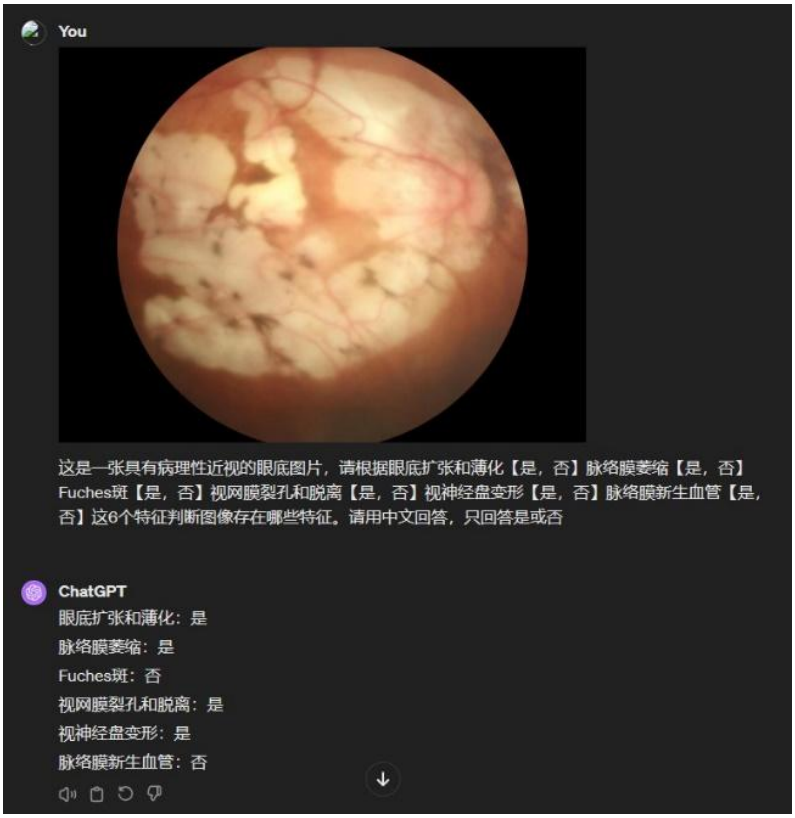


图 5　生成细粒度病理特征描述

Fig. 5 Generating fine-grained pathology characterization

3.3.2 文本生成的有效性

为了验证这种病理特征文本生成方法的有效性，本文与一名资深的高级眼科医生和两名眼科专业的医学生进行合作，并进行了定量的评估。本文从训练集中选择 100 张眼底图像作为实验对象。一方面，本文采用 GPT-4 来生成对应的细粒度病理特征。另一方面，由一名高级眼科医生和两名眼科专业的学生标注眼底图像的真实标签。在标注眼底图像的真实标签时，若两名眼科专业学生的结论相同，则直接采纳为眼底图像的真实病理特征，否则，有争议的病理特征由高级眼科医生裁决。将 GPT-4 生成的病理特征和眼底图像的真实病理特征进行比对。

最终实验结果表明，这 100 张眼底图像的病理特征总数为 360，其中，GPT-4 生成的结果中正确的个数为 295，正确率为 81.9%。这一结果符合专业眼科医生对眼底图像标注的要求，因此，本文使用 GPT-4 来为眼底图像生成对应的细粒度病理特征文本描述的方法是有效的。

3.4 损失函数

我们的目标是学习特征表示，使配对图像和文本描述之间的距离最小化，同时使未配对样本之间的距离最大化。

我们根据可用的分类标签信息建立图像-文本对，从而促使属于同一类别的样本在图像和文本领域中都具有接近的特征表征。假设一个批量的数据样本个数为 N，这 N 个样本中，图像表示为 $\{I_i\}_{i\in X_B}$，其中，$X_B \subset \{1,..N\}$，X_B 表示这一个批量图像的索引，文本表示为 $\{T_j\}_{j\in T_B}$，其中，$T_B \subset \{1,..N\}$，T_B 表示这一个批量文本的索引。将这一批图像 I 和文本 T 分别通过图像编码器和文本编码器，从而产生图像特征嵌入 u' 和文本特征嵌入 v'，两者为同一空间的向量。

之后，对图像特征 u' 和文本特征 v' 进行 L2 范数归一化，计算过程如下式：

$$u_i = \frac{u_i^{'}}{\| u_i^{'} \|_2} \quad (2)$$

$$v_j = \frac{v_j^{'}}{\| v_j^{'} \|_2} \quad (3)$$

其中，u_i 表示第 i 张图像通过图像编码器编码后的特征嵌入，v_j 表示第 j 个文本通过文本编码器编码后的特征嵌入。得到归一化的特征后，进行矩阵点积运算得到余弦相似度，公式如下式：

$$sim(u_i, v_j) = u_i^T v_j \quad (4)$$

其中，u_i^T 表示 u_i 的转置。我们考虑文本到图像和图像到文本之间的双向学习，采用对称交叉熵来最大化匹配的图像-文本对之间的相似性，最小化非匹配的图像-文本对之间的相似性，公式如下：

$$L_{i2t} = -\sum_{i \in X_B} \frac{1}{| K_{T_B}(i) |} \sum_{i^{'} \in K_{T_B}(i)} \log\left(\frac{\exp(sim(u_i, v_{j^{'}}) / \tau)}{\sum_{j \in T_B} \exp(sim(u_i, v_j) / \tau)}\right) \quad (5)$$

$$L_{t2i} = -\sum_{j \in T_B} \frac{1}{| K_{X_B}(j) |} \sum_{j^{'} \in K_{X_B}(j)} \log\left(\frac{\exp(sim(u_{j^{'}}, v_j) / \tau)}{\sum_{j \in T_B} \exp(sim(u_i, v_j) / \tau)}\right) \quad (6)$$

$$L = \frac{1}{2}(L_{i2t} + L_{t2i}) \quad (7)$$

其中，τ 为可训练的温度系数，$sim(*,*)$ 为余弦相似度，$K_{T_B}(i) = \{i^{'} \mid i^{'} \in T_B, y_{i^{'}} = y_i\}$，$K_{X_B}(j) = \{j^{'} \mid j^{'} \in X_B, y_{j^{'}} = y_j\}$。

4 实验设计及结果

4.1 评价指标

数据集中每个标签都有两种情况，即存在和不存在。因此，多标签问题可以看成是多个二分类的单标签问题。对于二分类单标签，相关的指标有 *precision*、*recall*（*TPR*）、*FPR*、*AP*、*F1* 和 *AUC* 等，*precision* 表示精确率，*recall* 表示召回率，也叫真正率，*FPR* 表示假正率，*AP* 表示平均精度、*F1* 与精确率和召回率呈正相关，相关计算公式如下：

$$precison = \frac{TP}{TP + FP} \quad (8)$$

$$recall = TPR = \frac{TP}{TP + FN} \quad (9)$$

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

$$AP = \int_{x=0}^{1} precision(recall^{-1}(x))dx \quad (11)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (12)$$

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \quad (13)$$

式中 *TP* 表示属于该种眼科疾病且被正确预测的数量，*FP* 表示该种眼科疾病被误检的数量，*FN* 表示属于该种疾病但是被漏检的数量，*TN* 表示不属于该种疾病且预测正确的数量。

为了评估模型的性能，先计算所有标签的 *AP*、*F1* 和 *AUC* 分数，然后计算各标签对应

指标的平均值，计算公式如下：

$$ml\_AP=\frac{1}{|T|}\sum_{i=0}^{|T|-1}AP_i \quad (14)$$

$$ml\_F1=\frac{1}{|T|}\sum_{i=0}^{|T|-1}F1_i \quad (15)$$

$$ml\_AUC=\frac{1}{|T|}\sum_{i=0}^{|T|-1}AUC_i \quad (16)$$

式中，T 表示疾病标签，AP_i、$F1_i$、AUC_i 分别表示第 i 个标签的 *AP*、*F1* 和 *AUC* 分数。最后，取各个指标的平均值作为模型的最终分数，计算方法如公式（17）所示。

$$Final=\frac{ml\_AP+ml\_F1+ml\_AUC}{3} \quad (17)$$

4.1 实验细节

本文采用的 CLIP 模型中，图像编码器为 ViT-B/32，文本编码器为 12 层的 Transformer，CLIP 的框架超参数详细信息如下表 2 所示。

表 2　CLIP 框架的超参数

Table 2 Hyperparameters of the CLIP framework

模型	图像编码器			文本编码器		
	层数	宽度	头数	层数	宽度	头数
ViT-B/32	12	768	12	12	512	8

表中，层数表示 Transformer 子层的层数，宽度表示嵌入向量的长度，头数表示子层多头的个数。

本研究所有实验均在专用服务器上进行，CPU 型号为 Intel(R) Xeon(R) CPU E5-2690，GPU 的型号是 NVIDIA GeForce RTX 2080 Ti，GPU 内存为 11GB。。本文采用 8：2 的比例来划分数据集，即 20%的样本用来验证，其余 80%的样本用来训练。本研究训练时采用的超参数如下表 3 所示。

表 3　训练超参数

Table 3 Hyperparameters for training

超参数	值
批次大小	8
迭代次数	50
优化器	Adam
学习率	1e-5
衰减系数	0.2
早停耐心	10
输入尺寸	224

训练时，图像预处理尺寸为 224×224，使用 Adam[24]优化器，学习率为 1e-5，衰减系数为 0.2，损失函数为 BCE，批次大小为 8。训练迭代总次数为 50，使用早期停止策略监控验证损失，耐心为 10。

4.2 对比实验

为了评估我们提出模型的性能，我们在自定义的 MDFCD8 数据集上与几种现有方法进行了比较。面临的主要挑战之一是该数据集缺乏用于比较的相关工作。为了快速比较我们提出的模型和其他先进方法之间的性能，我们从已有工作中选择已经公开源码的方法，并且分别从基于 CNN 的方法和基于 Transformer 的方法中选择代表性的算法来作为比较，比较的结果如下表 4 所示。

表 4　与其他相关工作的比较

Table 4 Comparison with other related work

模型	*ml_AP*	*ml_F1*	*ml_AUC*	*Final*
MCG-Net[25]	0.672	0.615	0.902	0.730
InceptionV3[26]	0.693	0.627	0.913	0.744
C-tran[18]	0.702	0.636	0.941	0.760
Our CLIP	**0.737**	**0.682**	**0.958**	**0.792**

我们的方法在 ml_AP、ml_F1、ml_AUC 以及模型的最终分数 $Final$ 上比其他方法的性能都要高，由此也说明了 CLIP 模型在眼底疾病识别方面的巨大潜力。我们的方法在 8 种类别中具体的性能指标如下表 5 所示。

表 5　8 种类别的性能指标

Table 5 Performance indicators for 8 categories

类别	精确率	召回率	*F1*	*AUC*
正常	0.775	0.789	0.782	0.973
糖尿病视网膜病变	0.845	0.754	0.797	0.987
青光眼	0.661	0.581	0.618	0.948
白内障	0.722	0.773	0.747	0.972
年龄相关性黄斑变性	0.744	0.621	0.677	0.963
高血压视网膜病变	0.387	0.323	0.352	0.942
近视	0.737	0.793	0.764	0.980
其他	0.769	0.682	0.723	0.898

从每个类别的分类结果可以看出，大部分类别的 *AUC* 分数都在 90%以上，*AUC* 反映了一个模型的分类性能，该模型总体上具有良好的性能。只有最后一个类“其他”低于 90%，这是因为“其他”类包含的疾病种类太多，没有明显的文本专家知识来给模型提供有效的信息。在精确率和召回率这两个指标上，“高血压视网膜病变”类表现很差，这是由于这种类别的正样本数量较小，从而造成精确率和召回率太小。

4.2 消融实验

4.2.1 文本嵌入方式的影响

眼底图像数据缺乏文本描述，词嵌入和模板嵌入是为眼底图像生成文本的较为简单的方式。词嵌入以其类别名作为文本，例如，如果眼底图像存在白内障这种疾病，那么这张图像的文本为“白内障”。模板嵌入采用 3.3.1 节中句子模板作为图像的文本，其文本只是简单的模板形式，即“一张患有[类别]的眼底图像”。两种不同文本嵌入方式的实验结果如下表 6 所示。从实验结果可以看出，使用模板嵌入取得更好的实验效果，表明 CLIP 对于文本的具有很高的敏感性，这是由于 CLIP 通过对大量的图像文本对进行预训练，学习了如何将图像内容与自然语言描述匹配起来。因此，即便是微小的文本变化也可能导致模型对图像的理解和响应发生改变。

表 6 词嵌入和模板嵌入的实验结果

Table 6 Comparison of experimental results for word embedding and template embedding

方法	*ml_AP*	*ml_F1*	*ml_AUC*	*Final*
单词文本	0.374	0.342	0.823	0.513
模板文本	**0.515**	**0.474**	**0.890**	**0.626**

4. 2. 2 专家知识的作用

对于有专家知识的文本来说，其文本包含病理特征，具有比模板嵌入更为丰富的文本信息。我们采用第一组消融实验中模板嵌入的实验结果作为本组的实验结果之一，实验结果对比如下表 7 所示。从表中的实验看出，有专家知识的文本嵌入使得分类效果大幅度提升，也说明了领域专家知识的重要性。

表 7　模板嵌入和专家知识嵌入的实验结果对比

Table 7 Comparison of experimental results for template embedding and expert knowledge embedding

方法	*ml_AP*	*ml_F1*	*ml_AUC*	*Final*
模板文本	0.515	0.474	0.890	0.626
专家知识文本	**0.733**	**0.680**	**0.958**	**0.790**

4.2.3 专家知识多样性的作用

从已有实验结果可以看出，眼科疾病的领域专家知识使得 CLIP 模型的性能大幅提升。然而，领域专家知识是每张眼底图像对应的病灶特征的描述，对于同一种病理来说，由于其轻重程度不一，其病灶特征也不一致。因此，每一种病理的专家知识库里有多个不同的专家知识描述。我们将研究专家知识多样性对模型性能的影响，即一组实验中，眼底图像的病理文本包含所有图像中存在的病理特征，是多样性的。而另一组对照实验中，从图像存在的病理特征中随机挑选一个作为图像的病理文本。将两组实验结果结合之前的简单模板嵌入实验结果进行对比，对比结果如下表 8 所示。表中，“单调专家知识文本”表示病理文本只含有一个病理，“多样专家知识文本”表示病理文本含有多样性的病理特征。从实验结果可以看出，专家知识提升了模型的性能，而多样性的专家知识描述给模型提供了更加多样化的数据，扩展了模型学习的范围和深度。多样化的专家知识极大的丰富了语义信息，更充分地利用 CLIP 模型对文本数据提取信息的能力。

表 8　专家知识多样和专家知识单调的对比实验

Table 8 Comparative experiments of diverse expert knowledge and harmonized expert knowledge

方法	*ml_AP*	*ml_F1*	*ml_AUC*	*Final*
模板文本	0.515	0.474	0.890	0.626
单调专家知识文本	0.603	0.546	0.933	0.694
多样专家知识文本	**0.733**	**0.680**	**0.958**	**0.790**

4.2.4 数据对增加的作用

在前文中，我们将多标签分类问题转化为一组单标签分类问题，并对数据对进行了扩增，将一个多标签数据对扩增为多个单标签数据对。例如，原始数据是一张含有多种疾病的眼底图像，假设有 2 种，分别是糖尿病视网膜病变和老年黄斑变形这两种疾病，即{I，“糖尿病视网膜病变”，“年龄相关性黄斑变性”}。原数据可以拆分成两个数据对，即{I，“糖尿病视网膜病变”}和{I，“年龄相关性黄斑变性”}。为了证明这种数据处理方法的有效性，我们对这两种方法的实验结果进行对比，结果如下表 9 所示。由此可见，我们采用这种数据增强的方法，是对任务起正向作用的。

表 9　多标签数据对和单标签数据对的实验对比

Table 9 Experimental comparison of multi-label data pairs and single tab data pairs

方法	*ml_AP*	*ml_F1*	*ml_AUC*	*Final*
单标签数据	0.533	0.510	0.878	0.640
多标签数据	**0.733**	**0.680**	**0.958**	**0.790**

5 结 论

本文从数据收集、预处理、模型训练到测试等关键步骤，并深入讨论了数据清洗过程、数据对扩增的方法、GPT-4 增强数据的方法，以及损失函数的计算。通过这一系列的技术和策略，研究旨在提高眼科疾病识别的准确性和效率，提供了一种综合利用深度学习技术和自然语言处理工具的有效框架。

本文在新构建的数据集上，与其他传统的方法进行对比实验，实验结果表明了视觉语言模型在眼科疾病识别领域也有巨大的潜力。此外，本文还重点进行了消融研究，研究了 CLIP 模型中各模块在眼科疾病识别中的有效性，具体包括比较了不同文本嵌入方式、专家知识的作用、数据对增加以及图像编码器的影响。在未来的研究中，将进一步考虑研究更多种类的疾病，以及更好地处理类不平衡问题。

参考文献

[1] Cen LP, Ji J, Lin JW, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks [J]. Nature Communications, 2021, 12(1): 4828-4828.

[2] Ju L, Wang X, Yu Z, et al. Long-tailed multi-label retinal diseases recognition using hierarchical information and hybrid knowledge distillation [Z/OL]. arXiv Preprint arXiv:2111.08913, 2021.

[3] Prawira R, Bustamam A, Anki P. Multi label classification of retinal disease on fundus images using AlexNet and VGG16 architectures [C] // 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2021: 464-468.

[4] Sahlsten J, Jaskari J, Kivinen J, et al. Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading [J]. Scientific Reports, 2019, 9(1): 10750-10750.

[5] Sengar N, Joshi RC, Dutta MK, et al. EyeDeep-Net: A multi-class diagnosis of retinal diseases using deep neural network [J]. Neural Computing and Applications, 2023, 35(14): 10551-10571.

[6] Nawaz A, Ali T, Mustafa G, et al. Multi-Class Retinal Diseases Detection Using Deep CNN With Minimal Memory Consumption [J]. IEEE Access, 2023, 11, 56170-56180.

[7] Hisham I, Khalil MI, Abbas H. Multi-label Ophthalmological Disease Classification Using Vision Transformers [C] // 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES). IEEE, 2023: 279-284.

[8] Wang XL, Lu YJ, Wang YJ, et al. Diabetic retinopathy stage classification using convolutional neural networks [C] // 2018 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, 2018: 465-471.

[9] Liu Z, Hu H, Lin Y, et al. Swin transformer v2: Scaling up capacity and resolution [C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12009-12019.

[10] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [C] // International conference on machine learning. PMLR, 2021: 10347-10357.

[11] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // International conference on machine learning. PMLR, 2021: 8748-8763.

[12] Silva-Rodriguez J, Chakor H, Kobbi R, et al. A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision [Z/OL]. arXiv Preprint arXiv:2308.07898, 2023.

[13] Shang FX, Fu J, Yang YH, et al. SynFundus: A synthetic fundus images dataset with millions of samples and multi-disease annotations [Z/OL]. arXiv Preprint arXiv:2312.00377, 2023.

[14] Bendary NE, Hassanien AE, Corchado E, et al. ARIAS: Automated retinal image analysis system [C] // Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011. Springer Berlin Heidelberg, 2011: 67-76.

[15] Pachade S, Porwal P, Thulkar D, et al. Retinal Fundus Multi-Disease Image Dataset

(RFMiD): A Dataset for Multi-Disease Detection Research [J]. Data, 2021, 6(2): 1-14.
[16] Panchal S, Naik A, Kokare M, et al. Retinal Fundus Multi-Disease Image Dataset (RFMiD) 2.0: A Dataset of Frequently and Rarely Identified Diseases [J]. Data, 2023, 8(2): 1-16.
[17] Grand Challenge. 智慧之眼 预见未来[EB/OL]. [2024-03-28]. https://odir2019.grand-challenge.org.
[18] De K, Masilamani V. A new no-reference image quality measure for blurred images in spatial domain [J]. Journal of Image and Graphics, 2013, 1(1): 39-42.
[19] Rodriguez M A, Almarzouqi H, Liatsis P. Multi-Label Retinal Disease Classification Using Transformers [J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(6): 2739-2750.
[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [C] // ICLR. 2021.
[21] Nazir A, Wang Z. A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges [J]. Meta-radiology, 2023, 1(2): 100022-100022.
[22] Garner A, Ashton N. Pathogenesis of hypertensive retinopathy: a review [J]. Journal of the Royal Society of Medicine, 1979, 72(5): 362-365.
[23] Wilkinson CP, Ferris IIIFL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales [J]. Ophthalmology, 2003, 110(9): 1677-1682.
[24] Kingma DP, Ba J. Adam: A method for stochastic optimization [Z/OL]. arXiv Preprint arXiv:1412.6980, 2014.
[25] Lin J, Cai Q, Lin M. Multi-label classification of fundus images with graph convolutional network and self-supervised learning [J]. IEEE Signal Processing Letters, 2021, 28: 454-458.
[26] Wang X, Lu Y, Wang Y, et al. Diabetic retinopathy stage classification using convolutional neural networks [C] // 2018 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, 2018: 465-471.