第 14 卷 第 1 期
2025 年 1 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 14 No. 1
Jan. 2025

引文格式：
熊绍奎, 陈世峰. 基于文本增强的眼底图像多病种识别方法 [J]. 集成技术, 2025, 14(1): 78-90.
Xiong SK, Chen SF. Multi-disease recognition method for fundus images based on text enhancement [J]. Journal of Integration Technology, 2025, 14(1): 78-90.

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 该研究在眼科图像疾病识别中引入了视觉语言模型，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，作者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；其次，作者利用生成式人工智能 GPT-4（Generative Pre-trained Transformer 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；最后，作者计算了平均精度、*F*1 评分和受试者工作特征曲线下面积，并以三者的均值作为最终的性能评价指标。实验结果表明，与传统的卷积神经网络和 Transformer 网络相比，作者提出的方法在性能上分别高出 4.8% 和 3.2%。同时，作者还进行了各模块的消融实验，验证了该方法的有效性，表明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；对比语言图像预训练；专家知识
中图分类号 TP391.7；R770.41 **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*)
[*]*Corresponding Author*: *shifeng.chen@siat.ac.cn*

Abstract In this work, a visual language model is introduced in ophthalmic image disease recognition. And a multi-disease recognition algorithm based on a pre-trained contrasting language-images model is proposed. First, a 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 (Generative Pre-trained

收稿日期：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。

Transformer 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 paper calculates the average precision (AP), *F*1 score, and area under the receiver operating characteristic curve (AUC), and takes the mean value of the three as the final performance evaluation index. The experimental results showed that, the method proposed in this paper 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, demonstrating the potential application of visual language modeling in the field of auxiliary diagnosis of ophthalmic diseases.

Keywords　fundus images; multi-disease; constrastive language-image pretraining; expert knowledge
Funding　This work is supported by Shenzhen Science and Technology Innovation Commission (JSGG20220831105002004)

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、白内障和近视眼等)在早期往往没有明显症状，难以察觉。然而，这些疾病可能随着时间逐渐恶化，不仅严重影响患者的正常生活，甚至可能导致失明。通过定期的眼科检查，可以在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，提升患者的生活质量。

眼部疾病的及时诊断取决于眼科成像和专家检查，这种检查是劳动密集型的，容易出错且十分耗时。基于深度学习的自动分析和诊断技术在眼科疾病的分析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，深度学习模型能自动识别和分类各种眼部疾病。通过学习成千上万的病例，这些模型能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聚焦于卷积神经网络(convolutional neural network，CNN)和基于注意力机制的模型。通过卷积层，CNN 能自动学习和提取图像中的关键特征，并能提取视网膜成像中异常的特征，从而准确区分不同的眼科疾病[1-5]。例如，Sahlsten 等[3]利用 CNN 识别糖尿病视网膜病变和黄斑水肿分级，结果表明：深度学习系统可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不仅能从图像中提取信息，还能揭示视网膜图像中远距离像素之间的全局关联性，进一步提高了眼科疾病诊断任务的准确率[6-7]。例如，Hisham 等[6]采用预训练好的 Swin Transformer V2[8]和 DeiT[9]模型并在眼科疾病数据集上进行了微调，在多标签眼科疾病分类方面取得了较优的实验结果。但是，这些方法只使用了图像模态的信息，忽略了其他模态，如文本模态。此外，这些方法的性能依赖于大量的训练数据，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，如对比语言图像预训练(contrastive language-image pre-training，CLIP)等[10]，通过在大量的图像-文本对上进行预训练，可掌握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。即使在缺乏大量标注数据的情况下，CLIP 模型也能准确识别和分类眼科疾病特征。Silva-Rodriguez 等[11]提出的视网膜基础语言-图

像模型通过文本监督集成专家知识，增强了视网膜图像的解释性。该模型在 37 个开放访问的眼底成像数据集上进行训练，结果表明，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。视觉语言模型的迁移学习能力较好，在眼科疾病识别中潜力巨大。

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[12]将私有数据集合成一个高质量的图像数据集，并引入了一个经过认证的人工智能(AI)辅助诊断系统，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[11]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，利用 CLIP 模型对多标签眼底图像分类问题进行研究。首先，本文从公开可用的多病种眼底图像数据集中筛选图片，并进行标签对齐，构建了一个含有 8 个类别的全新多标签眼底图像数据集；其次，本文利用先进的生成式人工智能 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标签对应的专家知识进行随机采样带来的问题，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和消融实验，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，本文提出方法的性能更好。

2 材料与方法

2.1 数据集收集

由于专业壁垒、隐私保护和标记数据成本过高等，大量的医学数据通常很难被获得。因此，为得到高质量的眼底彩照多疾病数据集，本文收集了 4 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 4 个数据集分别是 ARIAS[13]、RFMiD[14]、RFMiD2.0[15]和 ODIR-5K[16]。其中，ARIAS 包含 143 张图像和 3 种标签；RFMiD 包含 3 200 张图像和 46 种标签，分别为正常和 45 种病理；RFMiD2.0 包含 860 张图像和 50 种标签，分别为正常和 49 种病理；ODIR-5K 包含 5 000 张图像和 8 种标签，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，数据集中的图像质量过关；第二，数据集中的图像数量足够大，且每一类眼科疾病的数量足够用来训练。初步合成的数据集由多个公开数据集构建，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和数量满足要求。

对于第一点而言，有些图像因设备误用、环境条件等造成眼底图像质量不佳，这些不利因素引入噪声、模糊和伪影等，降低了图像的质量，增加了不确定性和错误分类的风险。因此，先剔除标签中带有“低质量”“镜头灰尘”“图像偏移”等注释的眼底图像，避免因拍摄时光照条件差、镜头被灰尘遮蔽或抖动等外界因素影响，导致图像质量不佳。为检测低质量图像，本文基于 Kanjar 等[17]提出的模糊测度(blur measure，*BM*)计算图像质量分数，计算模糊测度的具体步骤为使用 Canny 边缘检测器检测眼底图像 I 的边缘集合 E，表示如下：

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

其中，$I(x,y)$为图像 I 的第(x,y)个像素值；$I(x',y')$为图像 I 的第(x',y')个像素值；E 为图像 I 的边缘像素点集合；N_{xy} 为 $I(x,y)$的八邻域；

$|N_{xy}|$为 N_{xy} 中元素的个数。

选取一个合适的模糊测度阈值，若大于或等于这个阈值，则图像具有较高的清晰度；若小于这个阈值，则图像较模糊，需去除。Rodriguez 等[18]比较视觉上评价得出的低质量图像和根据模糊测度(阈值设为 0.058)自动筛选得出的低质量图像，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 为最终阈值，采用自动计算模糊度的方法去除低质量图像。不同模糊度下的图像示例如图 1 所示。

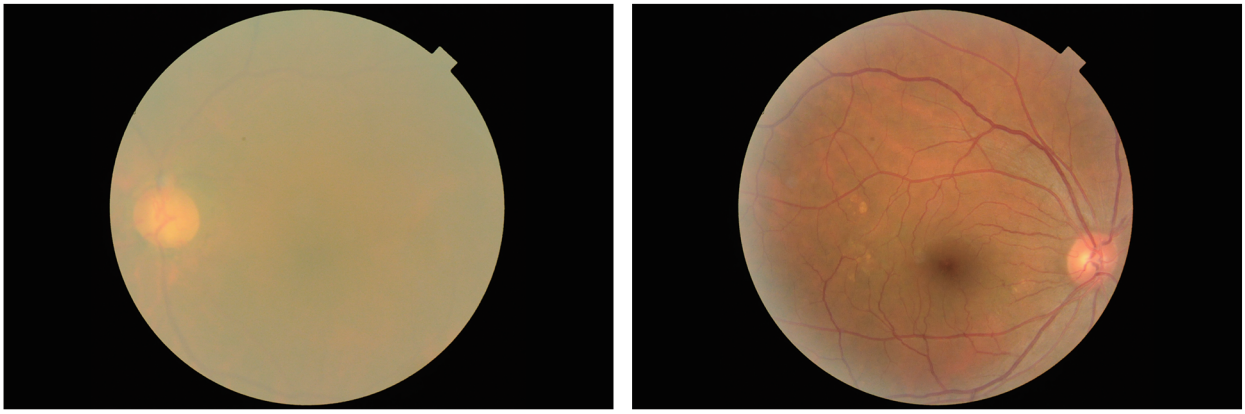
对于第二点而言，数据集之间不仅疾病的种类不一样，标签的形式也有显著差异，例如，ODIR-5K 数据集的疾病标签是基于患者标注的，而其他数据集的疾病标签是基于眼睛标注的。该数据集除了疾病标签外，还有关于患者的年龄、性别和病灶的描述。因此，需要对齐各个数据集的标签，以便于模型训练。对于数据集中样本较少的病理，数据增强技术并不会提高这些病理的识别性能，反而可能会对模型的整体性能产生负面影响。因此，本文把这样的病理标签并入“其他”类。由于两个不同数据集之间的病理标签有部分重叠，且剔除样本数量较少的病理标签，本文最终得到的新数据集共包含 8 种眼底图像，记为 MDFCD8，分别是正常、糖尿病视网膜病变、青光眼、白内障、年龄相关性黄斑变性、高血压视网膜病变、近视和其他。8 种眼底图像如图 2 所示。MDFCD8 数据集中 8 种眼底图像的数量分布如图 3 所示。

3　研究方法

3.1　网络架构

本文模型的整体网络结构如图 4 所示，在训练阶段，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[19]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特征，如图 4(a)所示。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，作为监督信号监督网络的训练。在生成文本时，输入眼底图像和合适的提示词，使用 GPT-4 生成眼底图像对应的包含专家知识的文本，如图 4(b)所示。

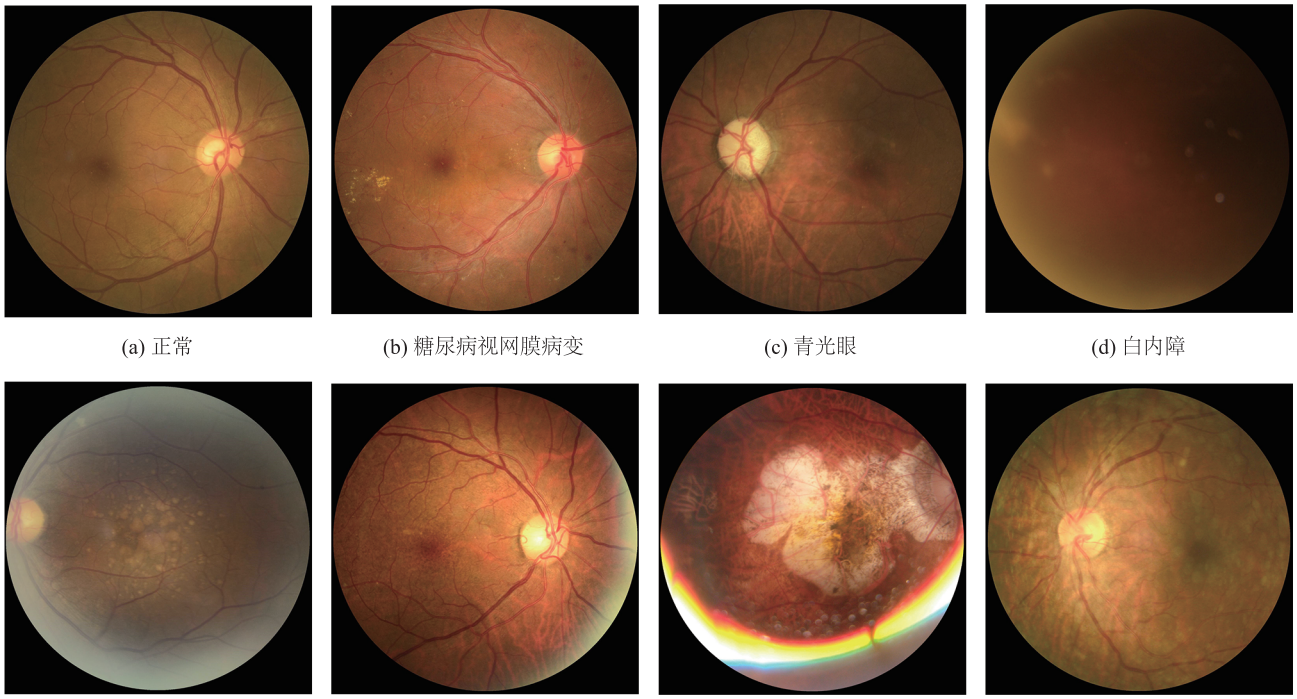
在推理阶段，待预测图像经过固定尺寸和归一化之后，送入图像编码器进行编码。同时，待预测的各类标签模板化后，通过文本编码器编码，如图 4(c)所示。编码后的图像嵌入分别与各类文本计算相似度，超过阈值则可以判定图像



(a) $BM=0.045$　　(b) $BM=0.090$

图 1　不同模糊度的眼底图像

Fig. 1　Fundus images with different degree of fuzziness



(a) 正常 (b) 糖尿病视网膜病变 (c) 青光眼 (d) 白内障

(e) 年龄相关性黄斑变性 (f) 高血压视网膜病变 (g) 近视 (h) 其他

图 2 8 种眼底图像

Fig. 2 Fundus images of 8 types

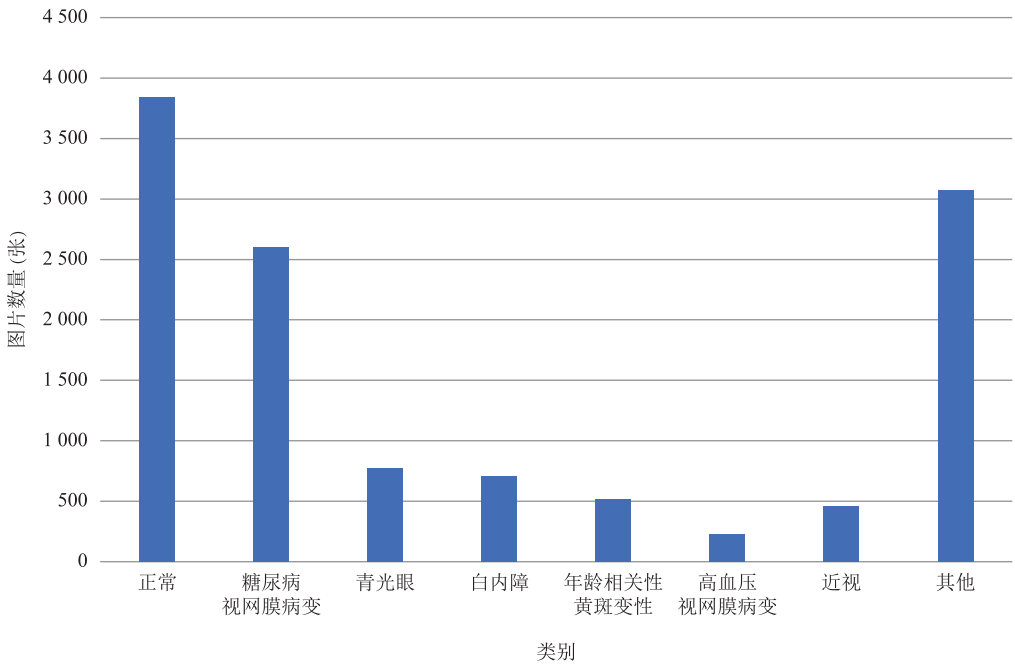


图 3 8 种眼底图像的数量分布

Fig. 3 Number distribution of 8 fundus images

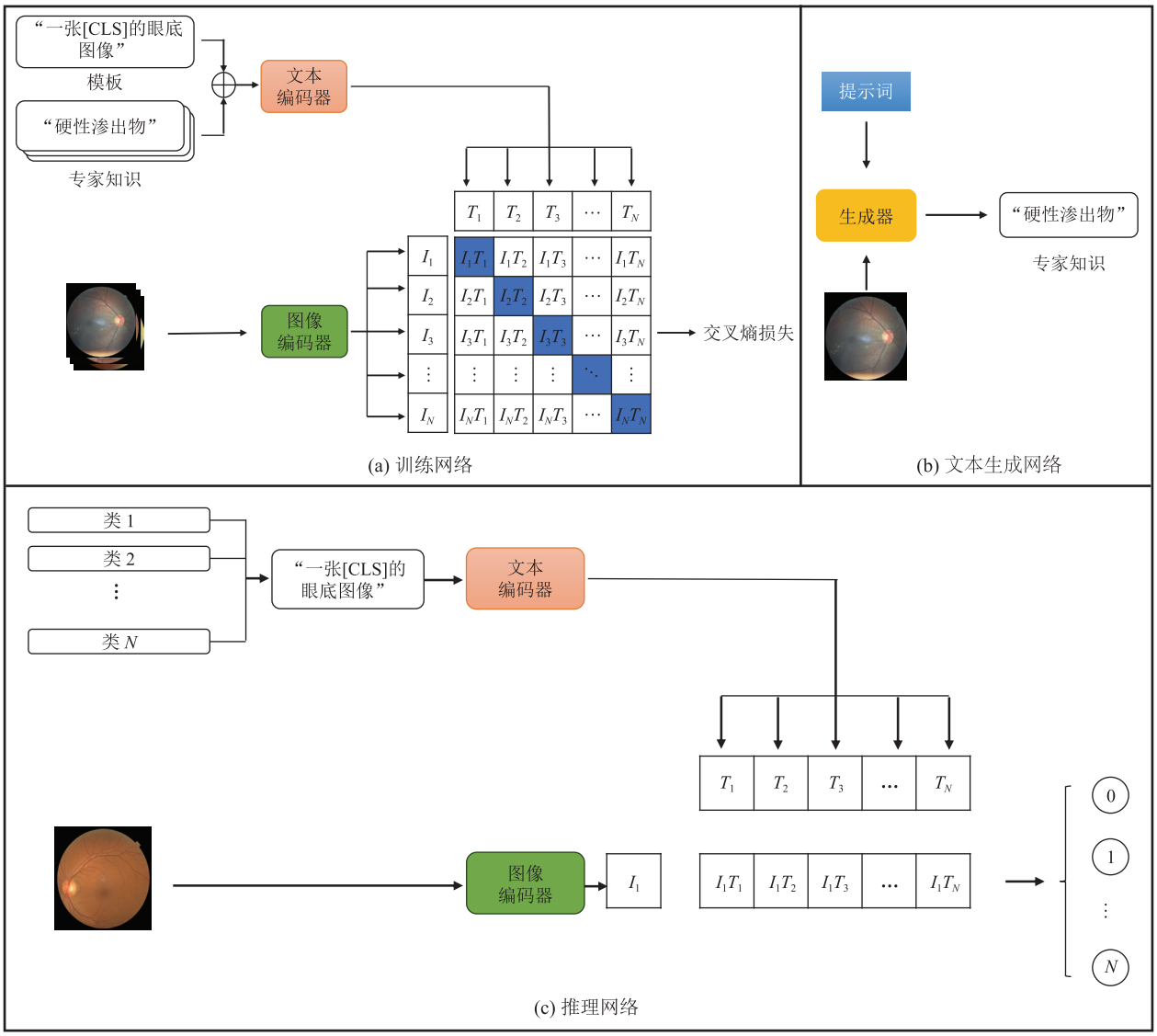


图 4　网络结构

Fig. 4　Network structure

属于该类。

3.2　数据对扩增

本文的研究对象为眼底图像，其中，一张眼底图像可能存在多种疾病，此时，这张眼底图像和其所含有的疾病类别共同组成一个数据样本。若一张眼底图像同时包含 $n(n<8)$ 种疾病，则这个数据样本可以表示为 $\{I,L_1,L_2,\cdots,L_n\}$。一些眼部疾病之间虽然可能会相互影响或协同进展，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：$\{I,L_1\},\{I,L_2\},\cdots,\{I,L_n\}$。

3.3　文本增强

3.3.1　文本生成策略

目前主要采用两种方法生成眼底图像的文本。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。这种方法虽然能保持类别名称的语义一致性，但仅包含病理名称，未反映具体的病理结构，信息量较少。第二种方法[12]

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构建文本描述。这样虽然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，有时甚至可能造成对模型的误导。

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，不仅需要专业的眼科疾病知识，而且耗时、劳动力密集。大型生成式人工智能提供了一种解决思路。GPT-4（Generative Pre-trained Transformer 4）是由 OpenAI 提出的一种多模态对话模型，能理解和生成文本、图像等多种数据类型，可深入理解复杂信息。已有研究表明[20]，GPT-4 具备大量的医疗知识，拥有生成较精准的医学图像描述的能力，能为医疗图像提供详细的、较为准确的文本解读，因而能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，从而以较低的成本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步骤如下：

（1）设定不同病理特征对应的专家知识集。本文定义的数据集包含 8 种病理特征，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[21]和社区标准[22]，本文总结了不同的病理及其对应的专家知识描述，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，利用 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设定，首先查询眼底图像标签对应的专家知识，以眼底图像标签的描述作为引导词输入 GPT-4，然后询问 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。

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 6 种特征，这 6 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、视神经盘变形及脉络膜新生血管。最终输入的引导词提示 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 6 种特征，如图 5 所示。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、视神经盘变形这 4 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、视网膜裂孔和脱离及视神经盘变形。”

3.3.2 文本生成的有效性

为验证本文病理特征文本生成方法的有效性，本研究团队与一名资深的高级眼科医生和两名眼

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”

图 5　细粒度病理特征描述

Fig. 5　Description of fine-grained pathology characterization

科专业的学生进行合作，并进行了定量评估。首先，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 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。其次，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，若不同，即有争议时，则由高级眼科医生裁决。最后，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。

最终的实验结果表明，这 100 张眼底图像的病理特征总数为 360，在 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 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

本文旨在学习特征表示，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。

本文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数为 N，则这 N 个样本中，图像表示为 $\{I_i\}_{i\in X_{\mathrm{B}}}$，其中，$X_{\mathrm{B}}\subset\{1,2,\cdots,N\}$，$X_{\mathrm{B}}$ 为该批量图像的索引；文本表示为 $\{T_j\}_{j\in T_{\mathrm{B}}}$，其中，$T_{\mathrm{B}}\subset\{1,2,\cdots,N\}$，$T_{\mathrm{B}}$ 为该批量文本的索引。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 $\boldsymbol{u}'$ 和文本特征嵌入 $\boldsymbol{v}'$，两者为同一空间的向量。

之后，对图像特征 $\boldsymbol{u}'$ 和文本特征 $\boldsymbol{v}'$ 进行 $L2$ 范数归一化，表示如下：

$$\boldsymbol{u}_i=\frac{\boldsymbol{u}_i'}{\left\|\boldsymbol{u}_i'\right\|_2} \tag{2}$$

$$\boldsymbol{v}_j=\frac{\boldsymbol{v}_j'}{\left\|\boldsymbol{v}_j'\right\|_2} \tag{3}$$

其中，$\boldsymbol{u}_i$ 为第 i 张图像通过图像编码器编码后的特征嵌入；$\boldsymbol{v}_j$ 为第 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，表示如下：

$$\mathrm{sim}\left(\boldsymbol{u}_i,\boldsymbol{v}_j\right)=\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j \tag{4}$$

其中，$\boldsymbol{u}_i^{\mathrm{T}}$ 为 $\boldsymbol{u}_i$ 的转置。本文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，表示如下：

$$L_{i2t}=-\sum_{i\in X_{\mathrm{B}}}\frac{1}{\left|K_{T_{\mathrm{B}}}(i)\right|}\sum_{i'\in K_{T_{\mathrm{B}}}(i)}\log\left(\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{u}_i,\boldsymbol{v}_{j'}\right)/\tau\right)}{\sum_{j\in T_{\mathrm{B}}}\exp\left(\mathrm{sim}\left(\boldsymbol{u}_i,\boldsymbol{v}_j\right)/\tau\right)}\right) \tag{5}$$

$$L_{t2i}=-\sum_{j\in T_{\mathrm{B}}}\frac{1}{\left|K_{X_{\mathrm{B}}}(j)\right|}\sum_{j'\in K_{X_{\mathrm{B}}}(j)}\log\left(\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{u}_{j'},\boldsymbol{v}_j\right)/\tau\right)}{\sum_{j\in T_{\mathrm{B}}}\exp\left(\mathrm{sim}\left(\boldsymbol{u}_i,\boldsymbol{v}_j\right)/\tau\right)}\right) \tag{6}$$

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

其中，L_{i2t} 指图像到文本之间的交叉熵损失；L_{t2i} 指文本到图像之间的交叉熵损失；τ 为可训练的温度系数；$\mathrm{sim}(*,*)$为余弦相似度；$K_{T_{\mathrm{B}}}(i)=$

$\left\{i'\middle|i'\in T_{\mathrm{B}},y_{i'}=y_i\right\}$；$K_{X_{\mathrm{B}}}(j)=\left\{j'\middle|j'\in X_{\mathrm{B}},y_{j'}=y_j\right\}$。

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，相关的指标有精确率(*precision*)、召回率(*recall*，*TPR*)、假正率(false positive rate，*FPR*)、平均精度(average precision，*AP*)、*F*1 和曲线下面积(area under the curve，*AUC*)等，其中，*F*1 与精确率和召回率呈正相关，表示如下：

$$precision=\frac{TP}{TP+FP} \tag{8}$$

$$recall=TPR=\frac{TP}{TP+FN} \tag{9}$$

$$FPR=\frac{FP}{FP+TN} \tag{10}$$

$$AP=\int_{x=0}^{1}precision\left(recall^{-1}(x)\right)\mathrm{d}x \tag{11}$$

$$F1=\frac{2\times precision\times recall}{precision+recall} \tag{12}$$

$$AUC=\int_{x=0}^{1}TPR\left(FPR^{-1}(x)\right)\mathrm{d}x \tag{13}$$

其中，*TP* 为属于该种眼科疾病且被正确预测的数量；*FP* 为该种眼科疾病被误检的数量；*FN* 为属于该种疾病但是被漏检的数量；*TN* 为不属于该种疾病且预测正确的数量。

为评估模型的性能，先计算所有标签的 *AP*、*F*1 和 *AUC* 分数，然后计算各标签对应指标的平均值，表示如下：

$$\overline{AP}=\frac{1}{|T|}\sum_{i=0}^{|T|-1}AP_i \tag{14}$$

$$\overline{F1}=\frac{1}{|T|}\sum_{i=0}^{|T|-1}F1_i \tag{15}$$

$$\overline{AUC}=\frac{1}{|T|}\sum_{i=0}^{|T|-1}AUC_i \tag{16}$$

其中，*T* 为疾病标签；AP_i、$F1_i$ 和 AUC_i 分别为第 *i* 个标签的 *AP*、*F*1 和 *AUC* 分数。最后，将各个指标的平均值作为模型的最终分数，表示如下：

$$Final=\frac{\overline{AP}+\overline{F1}+\overline{AUC}}{3} \tag{17}$$

4.2 实验细节

在本文采用的 CLIP 模型中，图像编码器为 ViT-B/32，文本编码器为 12 层的 Transformer，CLIP 框架的超参数如表 2 所示。

在表 2 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。

本研究的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 型号为 NVIDIA GeForce RTX 2080 Ti，GPU 内存为 11 GB。本文采用 8∶2 的比例划分数据集，即 20% 的样本用来验证，80% 的样本用来训练。本研究训练时采用的超参数如表 2 所示。

训练时，图像预处理尺寸为 224×224，使用 Adam[23]优化器，学习率为 1×10^{-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。

4.3 对比实验

为评估本文提出模型的性能，本实验在自定

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

义的 MDFCD8 数据集上与几种现有方法进行了比较。在实验过程中面临的一个主要挑战是，该数据集缺乏可用于比较的相关工作。为快速比较本文提出的模型和其他先进方法的性能，本文从已有工作中选择已经公开源码的方法，并分别从基于 CNN 的方法和基于 Transformer 的方法中选择具有代表性的算法进行比较，结果如表 3 所示。

表 3　与其他相关工作的比较

Table 3　Comparison with other related work

模型	$\overline{AP}$	$\overline{F1}$	$\overline{AUC}$	*Final*
MCG-Net[24]	0.672	0.615	0.902	0.730
InceptionV3[7]	0.693	0.627	0.913	0.744
C-tran[17]	0.702	0.636	0.941	0.760
文本增强＋CLIP	0.737	0.682	0.958	0.792

与其他方法相比，本文方法的 $\overline{AP}$、$\overline{F1}$、$\overline{AUC}$ 和 *Final* 均较高，表明 CLIP 模型在眼底疾病识别方面具有较大潜力。本文方法在 8 种类别上的性能指标如表 4 所示。

由表 4 可知，大部分类别的 *AUC* 大于 90%。由于 *AUC* 可反映一个模型的分类性能，因此，总体上该模型的性能较好。只有最后一个类“其他”的 *AUC* 小于 90%，原因是“其他”类包含的疾病种类较多，没有准确的文本专家知识来给模型提供有效的信息。在精确率和召回率两个指标上，“高血压视网膜病变”类的表现较差，原因是这种类别的正样本数量较小，从而造成精确率和召回率较低。

4.4　消融实验

4.4.1　文本嵌入方式的影响

眼底图像数据缺乏文本描述，而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，如果眼底图像显示白内障这种疾病，那么这张图像的文本为“白内障”。模板嵌入将 3.3.1 节中的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”。两种嵌入方式的实验结果如表 5 所示。由表 5 可知，模板嵌入的实验效果更好，表明 CLIP 对文本的敏感性较高，原因是 CLIP 通过对大量的图像文本对进行预训练，可学习到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 4　8 种眼底类别的性能指标

Table 4　Performance indicators for 8 fundus categories

类别	精确率	召回率	*F*1	*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

表 5　词嵌入与模板嵌入的实验结果

Table 5　Experimental results for word embedding and template embedding

嵌入方式	$\overline{AP}$	$\overline{F1}$	$\overline{AUC}$	*Final*
词嵌入	0.374	0.342	0.823	0.513
模板嵌入	0.515	0.474	0.890	0.626

4.4.2 专家知识的作用

与模板嵌入相比，有专家知识的文本包含病理特征，文本信息更丰富。模板嵌入与专家知识嵌入的实验结果对比如表 6 所示。由表 6 可知，有专家知识的文本嵌入使得分类效果大幅提升，表明了领域专家知识的重要性。

4.4.3 专家知识多样性的作用

由表 6 可知，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，专家知识是对眼底图像病灶特征的描述，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。本文将研究专家知识多样性对模型性能的影响，即实验组的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；而对照组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。将这两组实验的结果与 4.4.1 节的实验结果进行对比，如表 7 所示。其中，“单调专家知识”表示病理文本只含有病理名称，“多样专家知识”表示病理文本含有多样性的病理特征。由表 7 可知，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地丰富了语义信息，更充分地利用了 CLIP 模型对文本数据提取信息的能力。

4.4.4 数据对增加的作用

在 3.2 节，本文将多标签分类问题转化为单标签分类问题，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 种病理，分别是糖尿病视网膜病变和老年黄斑变形，即{I，“糖尿病视网膜病变”，“年龄相关性黄斑变性”}。原数据可拆分成两个数据对，即{I，“糖尿病视网膜病变”}和{I，“年龄相关性黄斑变性”}。为证明这种数据处理方法的有效性，本文对这两种方法的实验结果进行对比，结果如表 8 所示。结果表明，本文采用的这

表 6 模板嵌入与专家知识嵌入的实验结果对比

Table 6 Comparison of experimental results for template embedding and expert knowledge embedding

嵌入方式	$\overline{AP}$	$\overline{F1}$	$\overline{AUC}$	*Final*
模板嵌入	0.515	0.474	0.890	0.626
专家知识嵌入	0.733	0.680	0.958	0.790

表 7 多样专家知识和单调专家知识的对比实验

Table 7 Comparison experimental results for diverse expert knowledge and harmonized expert knowledge

嵌入方式	$\overline{AP}$	$\overline{F1}$	$\overline{AUC}$	*Final*
模板嵌入	0.515	0.474	0.890	0.626
单调专家知识嵌入	0.603	0.546	0.933	0.694
多样专家知识嵌入	0.733	0.680	0.958	0.790

表 8 多标签数据对与单标签数据对的实验对比

Table 8 Experimental comparison of multi-label data pairs and singls-label data pairs

方法	$\overline{AP}$	$\overline{F1}$	$\overline{AUC}$	*Final*
单标签数据对	0.533	0.510	0.878	0.640
多标签数据对	0.733	0.680	0.958	0.790

种数据增强方法对任务起正向作用。

5 结　论

本文通过数据收集、预处理、模型训练和测试等关键步骤，深入讨论了数据清洗过程、数据对扩增的方法、GPT-4 增强数据的方法和损失函数的计算。本研究旨在通过这一系列的技术和策略提高眼科疾病识别的准确性和效率，提供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，与其他传统方法进行了对比实验，结果表明：视觉语言模型在眼科疾病识别领域的潜力较大。此外，本文还重点进行了消融实验，研究了 CLIP 模型中各模块在眼科疾病识别中的有效性，具体内容如下：比较不同文本嵌入方式的专家知识对模型的影响以及单标签数据对与多标签数据对对模型的影响。在未来的研究中，本研究团队将考虑进一步研究更多种类的疾病，以更好地处理种类不平衡问题。

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: 4828.

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

[3] 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: 10750.

[4] 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.

[5] 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.

[6] Hisham I, Khalil MI, Abbas H. Multi-label ophthalmological disease classification using vision transformers [C] // Proceedings of the 2023 5th Novel Intelligent and Leading Emerging Sciences Conference, 2023: 279-284.

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

[8] Liu Z, Hu H, Lin YT, 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.

[9] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 10347-10357.

[10] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proceedings of the 38th International Conference on Machine Learning, 2021: 8748-8763.

[11] 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.

[12] 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.

[13] Bendary NE, Hassanien AE, Corchado E, et al. ARIAS: automated retinal image analysis system [C] // Proceedings of the Soft Computing Models

in Industrial and Environmental Applications, 6th International Conference SOCO 2011, 2011: 67-76.

[14] 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): 14.

[15] 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): 29.

[16] Grand Challenge. 北京大学“智慧之眼”国际眼底图像智能识别竞赛 [EB/OL]. (2019-07-13)[2024-05-14]. https://odir2019.grand-challenge.org.
Grand Challenge. Peking University International Competition on Ocular Disease Intelligent Recognition [EB/OL]. (2019-07-13)[2024-05-14]. https://odir2019.grand-challenge.org.

[17] Kanjar D, 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.

[18] Rodriguez MA, Almarzouqi H, Liatsis P. Multi-label retinal disease classification using transformers [J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(6): 2739-2750.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [Z/OL]. arXiv Preprint, arXiv: 2010.11929, 2020.

[20] Nazir A, Wang Z. A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges [J]. Meta-Radiology, 2023, 1(2): 100022.

[21] Garner A, Ashton N. Pathogenesis of hypertensive retinopathy: a review [J]. Journal of the Royal Society of Medicine, 1979, 72(5): 362-365.

[22] 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.

[23] Kingma DP, Ba J. Adam: a method for stochastic optimization [Z/OL]. arXiv Preprint, arXiv: 1412.6980, 2014.

[24] Lin JK, Cai QL, Lin MY. Multi-label classification of fundus images with graph convolutional network and self-supervised learning [J]. IEEE Signal Processing Letters, 2021, 28: 454-458.