doi: 10.13413/j.cnki.jdxblxb.2023529

基于多层次多尺度注意力融合网络的多模态眼底疾病诊断模型

郭晓新[1,2], 杨 梅[1,2], 杨广奇[1,2], 董洪良[1], 徐海啸[1]
(1. 吉林大学 计算机科学与技术学院, 长春 130012;
2. 吉林大学 符号计算与知识工程教育部重点实验室, 长春 130012)

摘要: 针对单模态眼底图像提取眼底特征的局限性, 提出一个基于多层次多尺度注意力融合网络的多模态眼底疾病诊断模型. 首先, 分别针对彩色眼底图像和视网膜光学相干断层成像设计多层次注意力网络和多尺度注意力网络, 并在特征层进行融合得到融合特征; 其次, 将两种模态的损失函数加权, 与融合特征的损失函数相加, 提取模态的独特和互补信息, 以提高眼底疾病诊断的准确率. 在数据集 MMC-AMD 和 GAMMA 上进行评估的实验结果表明, 该模型优于当前主流模型, 诊断效果优越.

关键词: 医学图像分类; 眼底疾病诊断模型; 多模态分类; 注意力机制
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1671-5489(2025)03-0783-12

Multimodal Retinal Disease Diagnosis Model Based on Multi-level and Multi-scale Attention Fusion Network

GUO Xiaoxin[1,2], YANG Mei[1,2], YANG Guangqi[1,2], DONG Hongliang[1], XU Haixiao[1]
(1. *College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun* 130012, *China*)

Abstract: Aiming at the limitations of extracting retinal features from single-mode retinal images, we proposed a multi-modal retinal disease diagnosis model based on multi-level and multi-scale attention fusion network. Firstly, the multi-level attention network and multi-scale attention network were designed for color retinal images and retinal optical coherence tomography respectively, and the fusion features were obtained by merging at the feature layer. Secondly, the weighted loss function of the two modes and the loss function of the fusion features were added to extract the unique and complementary information of the two modes in order to improve the accuracy of retinal disease diagnosis. The results of evaluation experiments on the MMC-AMD dataset and GAMMA dataset show that the proposed model outperforms the current mainstream models and has superior diagnostic effect.

Keywords: medical image classification; retinal disease diagnosis model; multi-modal classification; attention mechanism

收稿日期: 2023-12-13.

第一作者简介: 郭晓新(1974—), 男, 汉族, 博士, 教授, 从事机器视觉和医疗影像学的研究, E-mail: guoxx@jlu.edu.cn.

基金项目: 国家自然科学基金(批准号: 82071995)和吉林省科技发展计划项目(批准号: 20220201141GX).

目前，常见的眼底疾病包括糖尿病性视网膜病变、青光眼及年龄相关性黄斑变性(age-related macular degeneration，AMD)等. 其中，年龄相关性黄斑变性和青光眼发病率较高. 常见的眼底疾病图像包括彩色眼底图像(color fundus photography，CFP)、视网膜光学相干断层成像(optical coherence tomography，OCT)和眼底荧光素血管造影(fundus fluorescein angiography，FFA)等.

目前大部分眼底疾病的计算机辅助诊断研究都是针对彩色眼底图像. Huang 等[1]提出了一种显著性引导自监督图像 Transformer(SSiT)，用于糖尿病性视网膜病变图像的分级. 视网膜 OCT 是眼底疾病诊断的重要辅助图像，是帮助医生准确诊断的主要依据. Chen 等[2]提出了一种新型的特征交互 Transformer 网络(FIT-Net)，使用视网膜 OCT 诊断病理性近视(PM). Elsharkawy 等[3]提出了一种基于视网膜 OCT 的计算机辅助诊断方法，通过结构三维视网膜成像检测早期糖尿病性视网膜病变.

由于眼底结构较复杂，因此只使用一种模态诊断眼底疾病信息有局限性. 在实际临床的应用中，医生进行眼底疾病诊断的过程中，通常需参考多种不同的眼底图像模态，综合分析判断. 目前已有很多基于多模态的眼底疾病诊断模型[4]. Li 等[5]提出了一种联合使用 CFP 和 FFA 的方法，基于患者特征的归一化指数函数(Softmax)嵌入可通过学习模态不变特征和患者相似特征，实现对眼底疾病的分类. Li 等[6]提出了一种新型跨疾病关注网络(CANet)，通过探索疾病之间的内在联系联合分级糖尿病性视网膜病变和糖尿病性黄斑水肿.

CFP 和视网膜 OCT 是临床上常用的两种眼底图像[7]. Hua 等[8]使用 CFP 结合扫频源光学相干断层扫描血管成像(SS-OCTA)识别糖尿病性视网膜病变的严重程度. Liu 等[9]提出了 CFP 和视网膜 OCT 结合使用，提高了眼科医生和人工智能筛查糖尿病性视网膜病变的准确性. Han 等[10]使用 CFP 和视网膜 OCT 对 AMD 进行分类，建立了一个深度学习模型，利用光谱域光学相干断层成像区分新生血管性年龄相关性黄斑病变的亚型. Wang 等[11]提出了端到端多模态卷积神经网络(MM-CNN)，通过空间不变融合结合 CFP 和视网膜 OCT 流的信息. He 等[12]提出了一种新的多模态视网膜图像分类的模态特异性注意网络(modality-specific attention network，MSAN)，有效利用了 CFP 和视网膜 OCT 图像的模态特异性诊断特征. Yu 等[13]提出了一种基于 Transformer 的跨模态多对比网络，用于有效融合 CFP 和视网膜 OCT 诊断眼科疾病.

多模态眼底疾病诊断模型优于仅使用单模态的模型，但不同模态成像存在模态特异性信息，需针对具体模态，关注重要的特征，对不重要的特征进行抑制，从而提高眼底疾病的诊断效果. 针对上述问题，本文提出一个基于多层次多尺度注意力融合网络(multi-level and multi-scale attention fusion network，MLMSAFN)模型用于多模态眼底疾病诊断. 彩色眼底图像和视网膜 OCT 提供了关于视网膜的独特和互补信息，针对这两个模态，设计了两个独立的分支网络，在特征层进一步融合，以提高眼底疾病诊断的准确率. 本文主要贡献如下：

1) 针对彩色眼底图像设计多层次注意力网络(multi-level attention network，MLAN). 在骨干网络层间加入通道注意力(channel attention，CA)和空间注意力(spatial attention，SA)，通道注意力利用卷积通道之间的相关性增强特征表达，空间注意力捕获图像信息. 多层次注意力网络提取多层特征，充分挖掘垂直扫描方向图像特征，关注重要特征，突出代表性特征，抑制次要特征，从而提高彩色眼底图像的分类性能.

2) 针对视网膜 OCT 设计多尺度注意力网络(multi-scale attention network，MSAN). 为更好地挖掘多尺度特征中的有效信息，将不同尺度的图像特征划分为一系列特征块序列，从不同尺度获得图像全局 class token，与通过骨干网络得到的局部 class token 连接，将卷积神经网络(CNN)的局部特征与 Transformer 的全局特征相结合，从而提高视网膜 OCT 的分类性能.

3) 进一步融合多层次注意力网络的特征和多尺度注意力网络的特征，得到融合特征. 将彩色眼底图像损失函数和视网膜 OCT 损失函数加权，与融合特征的损失函数相加，获得最终的损失函数. 通过协同优化方案，MLMSAFN 可以捕获单模态信息并且学习多模态特征表示，从而提高多模态眼底疾病诊断的效果.

1　算法设计

多层次多尺度注意力融合网络结构包括彩色眼底图像的 MLAN 和视网膜 OCT 的 MSAN，以及松散配对、损失函数和衡量标准.

1.1　网络结构

本文提出的基于多层次多尺度注意力融合网络的多模态眼底疾病诊断模型如图 1 所示. MLMSAFN 模型主要基于 ResNet18[14]，针对彩色眼底图像设计了多层次注意力网络，并针对视网膜 OCT 设计了多尺度注意力网络，在特征层进行融合，得到融合特征，再将彩色眼底图像和视网膜 OCT 图像的损失函数进行加权，与融合特征的损失函数相加，得到总体损失函数.

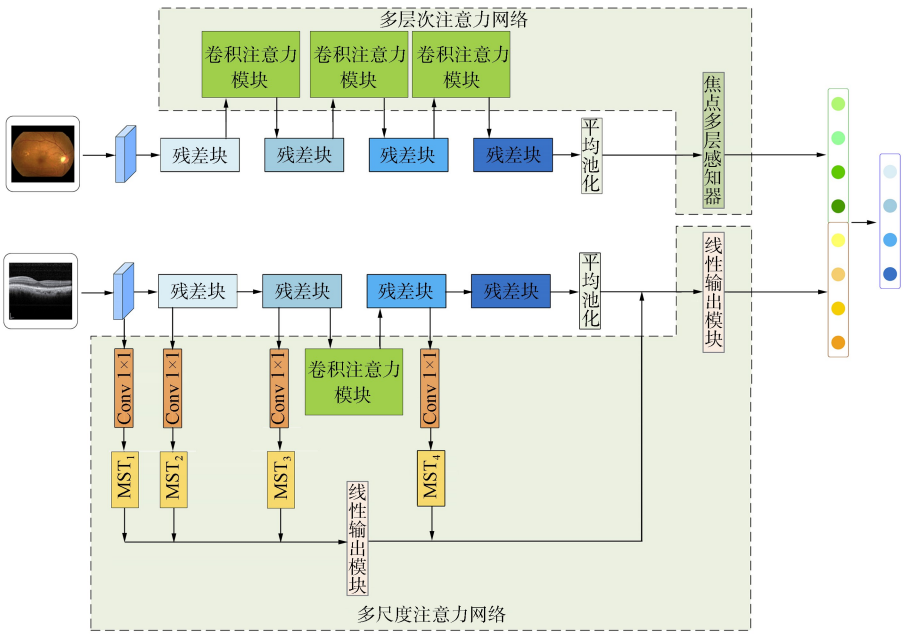


图 1　多模态眼底疾病诊断模型的多层次多尺度注意力融合网络框架

Fig. 1　Framework of multi-level and multi-scale attention fusion network for multi-modal retinal disease diagnosis model

1.1.1　彩色眼底图像分支

医疗图像辅助诊断的研究中，在模型中引入注意力机制，有助于提升模型的特征提取能力[15-19]. 应用在彩色眼底图像的多层次注意力网络主要包括卷积注意力模块(convolutional block attention module，CBAM)[20]和焦点多层感知器模块(Focus MLP)，分别如图 2 和图 3 所示.

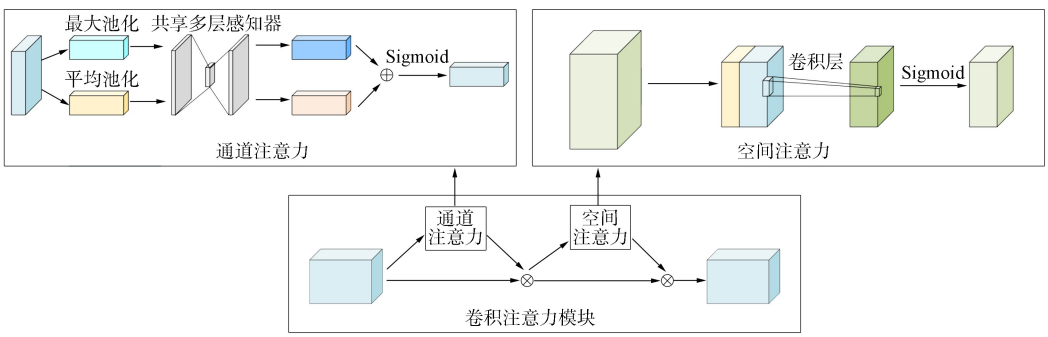


图 2　卷积注意力模块

Fig. 2　Convolutional attention module

由图 2 可见，CBAM 模块包括通道注意力和空间注意力. 首先利用特征的通道关系，生成通道注意力图；然后同时使用平均池化(AvgPool)特征和最大池化(MaxPool)特征，将两种特征分别传入共享网络；最后使用元素求和合并特征向量，再使用激活函数. 该网络是由一层隐藏层构成的多层感知

器(MLP). 通过注意力关注重要特征，抑制不必要的特征增加代表性. 通道注意力模块定义为

$$\mathrm{CA}(x)=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(x))+\mathrm{MLP}(\mathrm{MaxPool}(x))), \tag{1}$$

其中 σ 表示 Sigmoid 激活函数，x 表示输入.

首先，沿通道轴使用平均池化和最大池化操作将其连接，聚合特征图的通道信息. 其次，通过标准卷积层进行连接和卷积，生成二维空间注意力图. 空间注意力模块定义为

$$\mathrm{SA}(x)=\sigma(f^{7\times7}(\mathrm{Concat}[\mathrm{AvgPool}(x),\mathrm{MaxPool}(x)])), \tag{2}$$

其中 σ 表示 Sigmoid 激活函数，$f^{7\times7}$ 表示滤波器大小为 7×7 的卷积运算，Concat 表示连接操作.

卷积注意力模块先给定一个中间特征映射，得到通道注意力和空间注意力，然后将注意力映射乘以特征映射以适应自适应特征细化. 在层间添加卷积注意力模块，用于学习具有代表性、丰富和独特的眼底图像特征，逐层自适应特征细化，得到彩色眼底图像的自适应特征.

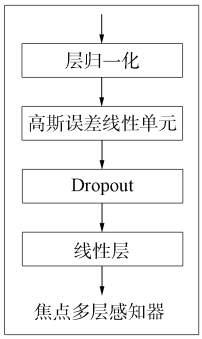


图 3 焦点多层感知器模块

Fig. 3 Focus MLP module

在骨干网络输出特征后添加焦点多层感知器模块. 由图 3 可见，焦点多层感知器先由一个层归一化(layer normalization，LN)接一个高斯误差线性单元(GELU)[21]激活函数层，然后是一层 Dropout 层，最后通过一层线性层进行输出，得到最终的彩色眼底图像特异性特征.

1.1.2 视网膜 OCT 分支

应用在视网膜 OCT 的多尺度注意力网络主要包含多尺度 Transformer(multi-scale transformer，MST)模块和线性输出(LD)模块. 多尺度 Transformer 模块包括实例嵌入、实例嵌入的线性投影、实例位置编码、多头注意力(multi head attention，MHA)、自注意力(self attention，SA)和多层感知器.

将主干卷积网络提取的 4 组多尺度特征分别经过 1×1 卷积核(Conv1×1)对特征进行降维，作为 MST 模块的输入. 表 1 列出了 MST 模块的 4 层(MST_i，$i=1,2,3,4$，i 表示具体层数)参数信息，包括特征尺寸、块维度、多层感知器维度、卷积核尺寸 1×1 以及块大小.

表 1 多尺度 Transformer 参数信息

Table 1 Parameter information of multi-scale Transformer

层	特征尺寸	块维度	多层感知器维度	卷积核尺寸	块大小
MST_1	224×224×64	1 024	2 048	224×224×4	16
MST_2	112×112×64	512	1 024	112×112×8	8
MST_3	56×56×128	256	512	56×56×16	4
MST_4	28×28×256	128	256	28×28×32	2

多尺度 Transformer 模块如图 4 所示. Transformer[22]使用图像的块作为输入 class token，将位置编码信息加入结构中. 将图像 $x\in\mathbb{R}^{H\times W\times C}$ 分成一系列的块 $x_p\in\mathbb{R}^{N\times(P^2\times C)}$，其中 H 和 W 分别为输入图像的高度和宽度，C 为通道数，P 为每个图像块的维度，$N=HW/P^2$ 为块个数. 将这些块作为 Transformer 的实例嵌入输入[13].

Transformer 在其所有层中使用恒定的潜在向量大小为 D，通过可学习的 $\boldsymbol{E}$ 线性投影，将实例嵌入映射到 D 维，$x_p\boldsymbol{E}$，$\boldsymbol{E}\in\mathbb{R}^{(P^2\times C)\times D}$，得到实例嵌入的线性投影. 由于 Transformer 体系结构排列不变，因此可将位置编码添加到实例嵌入中以保留位置信息，再加上位置编码 $\boldsymbol{E}_{\mathrm{pos}}$ 得到嵌入特征序列 $\boldsymbol{z}_0$：

$$\boldsymbol{z}_0=(x_p^1\boldsymbol{E},x_p^2\boldsymbol{E},\cdots,x_p^N\boldsymbol{E})+\boldsymbol{E}_{\mathrm{pos}},\quad \boldsymbol{E}\in\mathbb{R}^{(P^2\times C)\times D},\quad \boldsymbol{E}_{\mathrm{pos}}\in\mathbb{R}^{N\times D}. \tag{3}$$

多头注意力是查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)和值($\boldsymbol{V}$)分别用不同学习的线性投影进行 M 次线性投影. 然后对这些查询、键和值的每个投影版本并行执行多头注意力. 将这些数据进行连接并再次投影，生成最终值. 通过多头注意力提取特征，从不同头中学习到信息，通过自注意力层，最后将输出连接在一起，更好地传播上下文和语义信息，建立远程依赖关系. 该过程可表示为

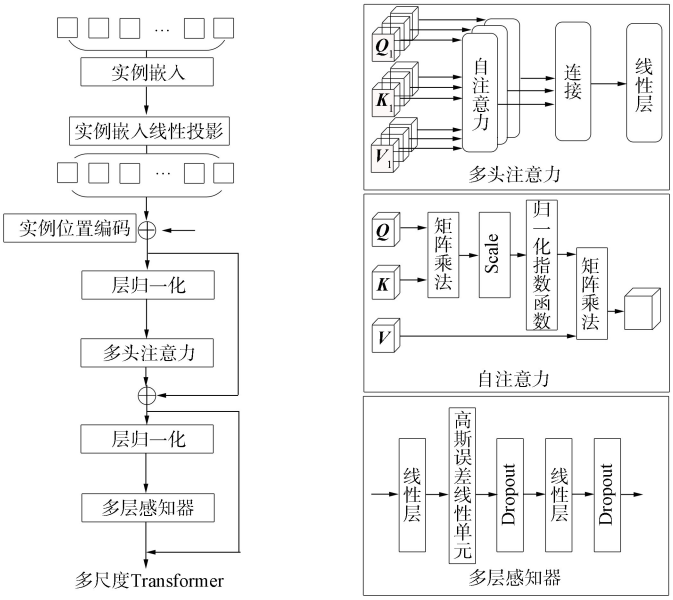


图 4　多尺度 Transformer 模块

Fig. 4　Multi-scale Transformer module

$$\mathrm{MHA}(z_{h-1})=\mathrm{Concat}(\boldsymbol{A}^1,\cdots,\boldsymbol{A}^m,\cdots,\boldsymbol{A}^M)\boldsymbol{W}^o, \tag{4}$$

$$\boldsymbol{A}^m=\mathrm{SelfAttention2}(\boldsymbol{Q}^m,\boldsymbol{K}^m,\boldsymbol{V}^m), \tag{5}$$

其中：Concat 表示连接操作；$m\in M$ 表示多头注意力的索引；$\boldsymbol{W}^o\in\mathbb{R}^{D\times D}$ 表示线性函数的参数矩阵；$D_m=D/M$ 为特征维数；$\boldsymbol{Q}^m=z_{h-1}\boldsymbol{W}_{\boldsymbol{Q}}^m$，$\boldsymbol{K}^m=z_{h-1}\boldsymbol{W}_{\boldsymbol{K}}^m$，$\boldsymbol{V}^m=z_{h-1}\boldsymbol{W}_{\boldsymbol{V}}^m$，参数矩阵 $\boldsymbol{W}_{\boldsymbol{Q}}^m\in\mathbb{R}^{D\times D_m}$，$\boldsymbol{W}_{\boldsymbol{K}}^m\in\mathbb{R}^{D\times D_m}$，$\boldsymbol{W}_{\boldsymbol{V}}^m\in\mathbb{R}^{D\times D_m}$，查询 $\boldsymbol{Q}^m\in\mathbb{R}^{D_m}$，键 $\boldsymbol{K}^m\in\mathbb{R}^{D_m}$，值 $\boldsymbol{V}^m\in\mathbb{R}^{D_m}$．在自注意力中，查询的输出表示为 $\boldsymbol{V}^m$ 加权注意力得分，D_m 用作归一化标量，进行 Scale 操作：

$$\mathrm{SelfAttention}(\boldsymbol{Q}^m,\boldsymbol{K}^m,\boldsymbol{V}^m)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}^m(\boldsymbol{K}^m)^{\mathrm{T}}}{\sqrt{D_m}}\right)\boldsymbol{V}^m. \tag{6}$$

多层感知器紧随层归一化，在每个多头注意力后．该模块由两个线性层组成，用一个高斯误差线性单元[21]激活函数分隔．第一个线性层扩展输入维数，第二个线性层将该维数缩减为原始输入维数．

多尺度 Transformer 的实例位置编码后面部分采用层归一化和残差连接．在多层感知器和多头注意力前应用层归一化操作，在每个多层感知器和多头注意力后应用残差连接．多头注意力和多层感知器叠加在一起，复制 H 次．该过程表示为

$$z_h'=\mathrm{MHA}(\mathrm{LN}(z_{h-1}))+z_{h-1},\qquad h=1,2,\cdots,H, \tag{7}$$

$$z_h=\mathrm{MLP}(\mathrm{LN}(z_h'))+z_h',\qquad h=1,2,\cdots,H. \tag{8}$$

线性输出模块如图 5 所示．由图 5 可见，LD 模块由一层线性层和一层 Dropout 层构成．每层多尺度 Transformer 模块的特征提取为 class token，前三层连接后经过 LD 模块，与第四层和最终层的输出一起连接，通过 LD 模块进行输出，对不同尺度的特征序列学习到的特征进行更有效融合．该过程表示为

$$\mathrm{Out}=\mathrm{LD}(\mathrm{Concat}(\mathrm{LD}(\mathrm{Concat}(\mathrm{MST}_1,\mathrm{MST}_2,\mathrm{MST}_3)),\mathrm{MST}_4,\mathrm{AvgPool}(x))), \tag{9}$$

其中：x 表示由骨干网络输出的特征，经过一层平均池化(AvgPool)；Out 表示输出特征；Concat 表示连接操作．

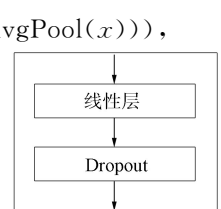


图 5　线性输出模块

Fig. 5　Linear output module

1.1.3　松散配对

多模态的眼底疾病诊断模型根据严格的眼睛身份配对．但彩色眼底图像和视网膜 OCT 的配对眼底图像较难获得，因此本文引入松散配对[23]，实现数据增强，以获得更好的眼底疾病诊断效果．松散配对使用标签而不是眼睛构建输入对．分类标签相同，彩色眼底图像可与视网膜 OCT 进行

匹配. 例如，在训练集中有两只干性年龄相关性黄斑变性眼睛 a 和 b，标签为 $\mathrm{Set}^{\mathrm{a}}$ 和 $\mathrm{Set}^{\mathrm{b}}$. 每张彩色眼底图像关联 1～5 张不等视网膜 OCT 图像，$\mathrm{Set}^{\mathrm{a}}=\{\mathrm{C}^{\mathrm{a}},\mathrm{O}_1^{\mathrm{a}},\mathrm{O}_2^{\mathrm{a}}\}$，$\mathrm{Set}^{\mathrm{b}}=\{\mathrm{C}^{\mathrm{b}},\mathrm{O}_1^{\mathrm{b}},\mathrm{O}_2^{\mathrm{b}},\mathrm{O}_3^{\mathrm{b}},\mathrm{O}_4^{\mathrm{b}}\}$，C 和 O 分别表示彩色眼底图像和视网膜 OCT. 对于$\{(\mathrm{C}^{\mathrm{a}},\mathrm{O}_i^{\mathrm{a}}),(\mathrm{C}^{\mathrm{b}},\mathrm{O}_j^{\mathrm{b}})\}(i=1,2$ 且 $j=1,2,3,4)$，有 6 对严格配对，根据松散配对原理，有 12 个实例配对，可以增加用于多模态训练实例的数量.

1.1.4 多分支损失函数

本文的损失函数采用交叉熵损失函数(CrossEntropyLoss)，定义为

$$\mathrm{CE}=\mathrm{CrossEntropyLoss}=-\frac{1}{n}\sum_{i=1}^{n}y_i\log\ y_i',\tag{10}$$

其中 y_i 表示编号 i 眼底图像分类的真实值，y_i' 表示编号 i 眼底图像分类的预测值，n 为眼底图像个数. 先计算出每张眼底图像的交叉熵后，再将整批眼底图像的交叉熵计算算数平均，得到最终的交叉熵损失.

总体损失函数共包含三部分，分别是彩色眼底图像损失函数 L_{cfp}、视网膜 OCT 损失函数 L_{oct} 和融合特征损失函数 L_{fusion}，定义为

$$L_{\mathrm{cfp}}=\mathrm{CE}(P_i^{\mathrm{cfp}},y^i),\tag{11}$$

$$L_{\mathrm{oct}}=\mathrm{CE}(P_i^{\mathrm{oct}},y^i),\tag{12}$$

$$L_{\mathrm{fusion}}=\mathrm{CE}(P_i^{\mathrm{fusion}},y^i),\tag{13}$$

总体损失函数 L_{overall} 定义为

$$L_{\mathrm{overall}}=\alpha\times L_{\mathrm{cfp}}+(1-\alpha)\times L_{\mathrm{oct}}+L_{\mathrm{fusion}},\tag{14}$$

其中 α 表示用来调节彩色眼底图像和视网膜 OCT 损失函数的比重，y^i 表示编号 i 对应眼底图像分类的真实值，P_i^{cfp} 表示编号 i 彩色眼底图像分类的预测值，P_i^{oct} 表示编号 i 视网膜 OCT 分类的预测值，P_i^{fusion} 表示编号 i 融合特征分类的预测值. L_{overall} 共同训练彩色眼底分支、视网膜 OCT 分支和融合分支. 通过协同优化方案，MLMSAFN 可以捕获单模态信息并且学习多模态特征表示.

1.2 衡量标准

本文使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F_1-Score、灵敏性(Sensitivity)和特异性(Specificity)作为眼底疾病诊断模型的衡量指标，分别定义为

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}},\tag{15}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\tag{16}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\tag{17}$$

$$\mathrm{Sensitivity}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\tag{18}$$

$$\mathrm{Specificity}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\tag{19}$$

$$F_1\text{-}\mathrm{Score}=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},\tag{20}$$

其中 TP,TN,FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性.

2 实验结果与分析

2.1 实验设置

2.1.1 数据集

实验采用两个跨模态的眼底数据集 MMC-AMD 和 GAMMA. 数据集 MMC-AMD 收集于北京协和医院眼科门诊[11]，用于研究多模态 AMD 分类. 该数据集共包含 829 名受试者的 1 093 只眼睛，其中：彩色眼底图像 1 094 张，由 Topcon 眼底相机获得；817 只眼睛的每张图像与 1～5 张视网膜 OCT

有关，OCT 图像共 1 289 张，由 Topcon OCT 相机和 Heidelberg OCT 相机获得. 本文对 AMD 进行 4 分类诊断，将该数据集分为训练集、验证集和测试集，其中训练集中的彩色眼底图像和视网膜 OCT 分别有 934 张和 1 009 张，验证集分别有 80 张和137 张，测试集分别有 80 张和 143 张. 重复 5 次实验，取平均值为最终实验结果.

数据集 GAMMA 用于研究多模态青光眼分类，其由 100 个配对的 3D 视网膜 OCT 和 2D 彩色眼底图像数据组成，青光眼分为 3 类(50 个正常，26 个早期，24 个中晚期). 其中每个 3D 的 OCT 包含 256 个 2D 的 b 扫描切片，每个切片大小为 512×992. 有两种不同大小的彩色眼底图像：2 992×2 000 和 1 956×1 934. 对该数据集进行五折交叉验证.

2.1.2　实验环境及参数设置

实验在 Ubuntu 系统上进行操作，使用 PyTorch 深度学习框架，采用 PyCharm 实验平台，网络训练所用的 GPU 类型为 NVIDIA GTX4090s. 使用随机梯度下降优化(stochastic gradient descent，SGD)算法对网络进行优化，实验的批处理大小设为 16，初始学习率设为 0.005，权重衰退率设为 0.000 1，最小学习率设为 1×10^{-7}. 用 Resnet 在 ImageNet 上进行预训练.

2.2　实验结果分析

2.2.1　不同分类在单模态和多模态的效果对比

在数据集 MMC-AMD 上，表 2～表 5 分别列出了正常(Normal)、干性年龄相关性黄斑变性(DryAMD)、息肉状脉络膜血管病变(PCV)、湿性年龄相关性黄斑变性(WetAMD)分类在彩色眼底图像(CFP)、视网膜光学相干断层成像(OCT)以及多模态下的分类结果.

表 2　正常分类在彩色眼底图像、视网膜光学相干断层成像和多模态下的结果对比

Table 2　Comparison results of normal classification in CFP, OCT and multi-modal　%

模态	灵敏性	F_1-Score	特异性	精确率
彩色眼底图像	99.0	99.0	99.7	99.0
视网膜光学相干断层成像	100.0	100.0	100.0	100.0
彩色眼底图像＋视网膜光学相干断层成像	100.0	100.0	100.0	100.0

表 3　干性年龄相关性黄斑变性分类在彩色眼底图像、视网膜光学相干断层成像和多模态下的结果对比

Table 3　Comparison results of DryAMD classification in CFP, OCT and multi-modal　%

模态	灵敏性	F_1-Score	特异性	精确率
彩色眼底图像	77.0	73.9	89.0	71.0
视网膜光学相干断层成像	86.8	92.7	99.8	99.5
彩色眼底图像＋视网膜光学相干断层成像	85.8	92.3	100.0	100.0

表 4　息肉状脉络膜血管病变分类在彩色眼底图像、视网膜光学相干断层成像和多模态下的结果对比

Table 4　Comparison results of PCV classification in CFP, OCT and multi-modal　%

模态	灵敏性	F_1-Score	特异性	精确率
彩色眼底图像	62.0	65.8	90.7	70.1
视网膜光学相干断层成像	80.4	80.9	91.1	81.4
彩色眼底图像＋视网膜光学相干断层成像	82.1	84.9	94.2	87.8

表 5　湿性年龄相关性黄斑变性分类在彩色眼底图像、视网膜光学相干断层成像和多模态下的的结果对比

Table 5　Comparison results of WetAMD classification in CFP, OCT and multi-modal　%

模态	灵敏性	F_1-Score	特异性	精确率
彩色眼底图像	60.0	61.8	86.7	63.7
视网膜光学相干断层成像	79.0	74.0	87.2	69.6
彩色眼底图像＋视网膜光学相干断层成像	85.3	77.3	86.9	70.6

由表 2～表 5 可见，AMD 的多模态分类结果优于单模态分类结果. 在表 2 的正常分类中，所有指标在多模态中均达 100%. 在表 3 干性 AMD 分类中，多模态分类结果与单模态分类结果较相近，但总体上多模态分类效果更好. 在表 4 的 PCV 分类中，多模态眼底图像的灵敏性达 82.1%，F_1-Score 达 84.9%. 在表 5 的湿性 AMD 分类中，多模态眼底图像的灵敏性达 85.3%，F_1-Score 达 77.3%.

2.2.2 消融实验

表6列出了数据集MMC-AMD上的消融实验结果. 由表6可见,彩色眼底图像中,在基线上添加了卷积注意力模块和焦点多层感知器模块. 在骨干网络层间添加了卷积注意力模块学习彩色眼底图像的有用特征,使分类效果明显提升,F_1-Score达74.5%,比基线提高了1.8个百分点,准确率达74.0%,比基线提高了2个百分点. 添加焦点多层感知器模块,与卷积注意力模块相协作,得到MLAN特异性网络,F_1-Score达到75.1%. 视网膜OCT中,在基线上添加多尺度Transformer模块,F_1-Score提高了1.7个百分点,准确率提高了2.1个百分点. 添加卷积注意力模块,与多尺度Transformer模块协作,得到MSAN特异性网络,准确率达84.5%. 融合网络MLMSAFN的准确率达86.4%,F_1-Score达88.6%,相比于基线模型,性能均有较大提升. 为比较公平,只对网络进行了改动,其他部分均保持一致.

表6 消融实验结果

Table 6 Results of ablation experiments %

模态	方法	准确率	F_1-Score
彩色眼底图像	基线	72.0	72.7
	基线+卷积注意力模块	74.0	74.5
	基线+卷积注意力模块+焦点多层感知器	74.5	75.1
视网膜光学相干断层成像	基线	81.5	84.6
	基线+多尺度 Transformer	83.6	86.3
	基线+多尺度 Transformer+卷积注意力模块	84.5	86.9
彩色眼底图像+视网膜光学相干断层成像	基线	83.4	86.1
	基线+卷积注意力模块	84.2	86.7
	基线+卷积注意力模块+焦点多层感知器	84.9	87.3
	基线+卷积注意力模块+焦点多层感知器+多尺度 Transformer	86.4	88.6

2.2.3 权重因子对诊断结果的影响

表7列出了在数据集MMC-AMD上α对多模态眼底疾病诊断模型的影响. 由表7可见,当$\alpha=0.5$时,MLMSAFN模型效果更好.

表7 α对多模态眼底疾病诊断模型的影响

Table 7 Effect of α on multi-modal retinal disease diagnosis model %

α	准确率	F_1-Score	α	准确率	F_1-Score
0.1	83.6	86.4	0.6	83.5	86.6
0.2	84.9	87.4	0.7	82.0	85.3
0.3	85.6	88.0	0.8	84.5	87.3
0.4	85.9	88.0	0.9	85.2	87.4
0.5	86.4	88.6			

2.2.4 骨干网络的对比

下面讨论融合网络使用的骨干网络选择问题. 表8列出了使用Resnet18,Resnet34,Resnet50,Resnet101骨干网络在数据集MMC-AMD上的分类结果. 这4种卷积网络模型的深度逐渐增加. 由表8可见,在数据集MMC-AMD上,本文提出的MLAN和MSAN网络使用Resnet18表现出更好的AMD诊断效果,随着网络深度的增加,诊断效果提升并不明显.

表8 在数据集MMC-AMD上不同骨干网络的性能对比

Table 8 Performance comparison of different backbone networks on MMC-AMD dataset %

网络	准确率	F_1-Score	网络	准确率	F_1-Score
Resnet18	86.4	88.6	Resnet50	83.8	86.6
Resnet34	85.7	88.0	Resnet101	84.9	87.5

2.2.5 使用松散配对的影响

下面讨论在多模态使用中,在数据集MMC-AMD上使用松散配对对多层次多尺度注意力融合网

络模型的影响. 表 9 列出了不使用松散配与使用对松散配对的对比结果. 松散配对中彩色眼底图像和视网膜 OCT 使用类标签而不是用对眼睛构建数据对，生成更多的多模态训练实例，进一步提高诊断效果. 由表 9 可见，使用松散配对，AMD 诊断效果明显提升，准确率提高了 8.5 个百分点，F_1-Score 提高了 7.3 个百分点.

表 9　不使用松散配对与使用松散配对的对比结果

Table 9　Comparison results between not using loose pairing and using loose pairing　%

网络	准确率	F_1-Score
多层次多尺度注意力融合网络(不使用松散配对)	77.9	81.3
多层次多尺度注意力融合网络	86.4	88.6

2.2.6　分类方法性能对比

表 10 列出了不同方法在数据集 MMC-AMD 上的分类结果对比，按 Accuracy 的升序进行排列. 比较方法包括：文献[24]提出的基于 ImageNet 预训练提取 CFP 和 OCT 的特征诊断方法，为公平，该方法中的 VGG19 替换成 ResNet18；文献[25]提出的多模态网络(COMNet)方法；文献[11]提出的端到端多模态卷积神经网络非数据增强(MM-CNN)方法，以及早期融合方法和晚期融合方法. 文献[24]中模型使用的并不是一种端到端的训练. COMNet 方法侧重于通过融合分支迭代融合每个阶段从两个模式中提取的特征. MM-CNN 方法通过端到端的两个卷积神经网络实例化，缺少对彩色眼底图像和视网膜 OCT 两个模态的特异性特征提取. 本文多层次多尺度注意力融合网络 MLMSAFN 针对彩色眼底图像和视网膜 OCT 分别设计了特征提取网络，通过特征层融合进一步结合彩色眼底图像和视网膜 OCT 的信息，在多模态图像分类中得到的准确率为 86.4%，平均 F_1-Score 为 88.6%. 由表 10 可见，本文方法在实验数据集 MMC-AMD 上优于其他方法.

表 10　不同方法在数据集 MMC-AMD 上的分类结果对比

Table 10　Comparison of classification results of different methods on MMC-AMD dataset　%

模态	方法	准确率	F_1-Score
彩色眼底图像	CFP-CNN[11]	71.7	77.4
	COMNet-CFP [25]	73.1	75.0
	MLAN	74.5	75.1
视网膜光学相干断层成像	CFP-CNN[11]	81.8	88.6
	COMNet-OCT [25]	83.8	86.5
	MSAN	84.5	86.9
彩色眼底图像＋视网膜光学相干断层成像	文献[24]	69.0	79.2
	EarlyFusion	77.9	85.6
	LateFusion	79.2	86.9
	MM-CNN[1]	80.4	87.2
	COMNet[25]	84.9	86.2
	MLMSAFN	86.4	88.6

表 11 列出了不同方法在数据集 GAMMA 上的分类结果对比，按 F_1-Score 的升序进行排列. MBSaNet[26]方法和 DKCNet[27]方法使用单模态眼底图像作为输入. MBSaNet[26]提出了一种结合卷积神经网络和自注意力机制的多阶段眼底图像分类模型. DKCNet[27]是由一个注意力块、一个挤压块和激励块组成的判别卷积核网络. MM-CNN[11]方法和 COROLLA[28]方法使用多模态眼底图像作为输入. COROLLA 使用一种多模态监督对比学习框架. 相对于单模态，多模态模型充分利用了彩色眼底图像和视网膜 OCT 图像的互补信息. 由表 11 可见，本文的多模态诊断模型挖掘了每个模态的特异性信息，表现出更好的青光眼诊断效果，平均 F_1-Score 为 82.3%.

2.2.7　AUC 可视化结果

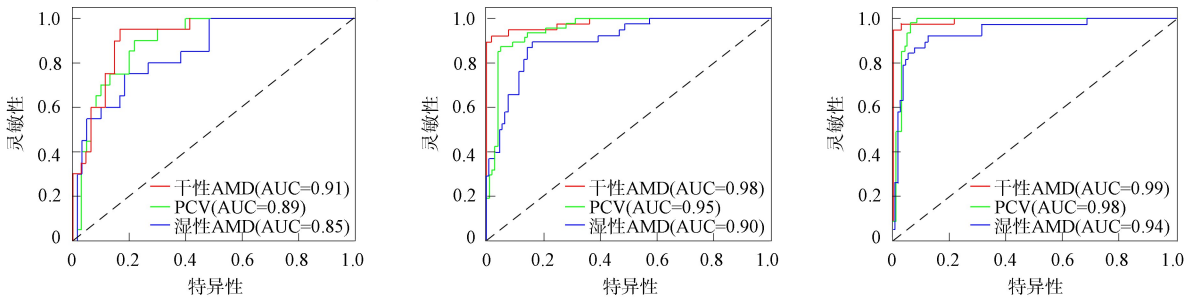
图 6 为不同模型在数据集 MMC-AMD 上的 AUC(area under the curve)曲线. 由图 6 可见，多模态融合模型 MLMSAFN 在 AMD 的分类效果更好，干性 AMD 的 AUC 达 99%，PCV 的 AUC 达

98%，湿性AMD的AUC达94%．并且本文模型在单模态数据上也有良好的性能．

表11 不同方法在数据集GAMMA上的分类结果对比

Table 11 Comparison of classification results of different methods on GAMMA dataset %

方法	灵敏性	精确率	特异性	F_1-Score
MBSaNet[26]	73.0	77.4	88.6	73.9
DKCNet[27]	79.6	80.2	91.3	79.6
MM-CNN[1]	78.4	83.5	91.1	79.8
COROLLA[28]	80.1	82.6	92.3	80.9
MLMSAFN	81.0	81.5	92.7	82.3



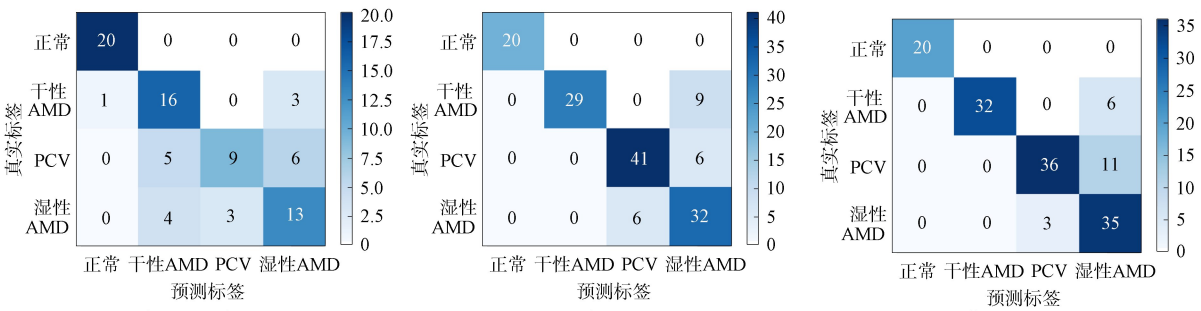
(A) 彩色眼底图像的MLAN (B) 视网膜OCT的MSAN (C) 模型MLMSAFN

图6 不同模型在数据集MMC-AMD上分类的AUC曲线

Fig. 6 AUC curves classified by different models on MMC-AMD dataset

2.2.8 混淆矩阵可视化结果

图7为不同模型在数据集MMC-AMD上的混淆矩阵．由图7可见，多模态融合网络模型在AMD的分类效果更好，并且本文模型在单模态数据上也有良好的性能．



(A) 彩色眼底图像的MLAN (B) 视网膜OCT的MSAN (C) 模型MLMSAFN

图7 不同模型在数据集MMC-AMD上分类的混淆矩阵

Fig. 7 Confusion matrix classified by different models on MMC-AMD dataset

2.2.9 t-SNE可视化结果

图8和图9分别为彩色眼底图像和视网膜OCT图像在数据集MMC-AMD上的t-SNE(t-distributed stochastic neighbor embedding)可视化[29]结果．由图8和图9可见，MLAN和MSAN将彩色眼底图像和视网膜OCT较好地分成了正常、干性AMD、PCV和湿性AMD 4类．正常类别与其他类别距离更远，更易分辨．干性AMD总体可分性较好．PCV是新生血管发生在湿性视网膜色素上皮下方，故PCV和湿性AMD距离较近，分辨更困难．

综上所述，针对单模态眼底图像提取眼底特征的局限性，本文提出了一个基于ResNet的多层次多尺度注意力融合网络模型，用于多模态眼底疾病诊断．通过对彩色眼底图像的多层次注意力特异性特征的提取和视网膜OCT多尺度注意力特异性特征的提取，在特征层进一步融合，得到多模态的眼底疾病诊断模型．针对彩色眼底图像设计的MLAN，通过通道注意力和空间注意力，沿两个独立的维度通道和空间依次推断注意力映射，然后将注意力映射乘以特征映射以适应自适应特征细化，获得骨干网络的层级特征，提取更有效的特征．针对视网膜OCT设计的MSAN，从不同尺度获得图像全局

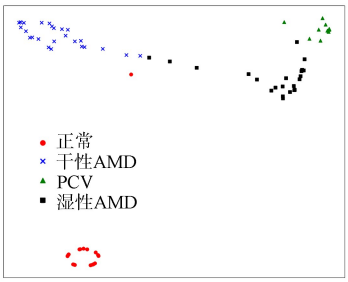


图 8　彩色眼底图像的 t-SNE 可视化结果

Fig. 8　t-SNE visualization results of CFP

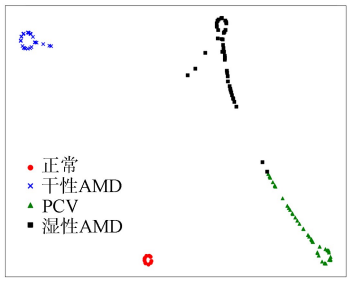


图 9　视网膜光学相干断层成像的 t-SNE 可视化结果

Fig. 9　t-SNE visualization results of OCT

class token，与通过骨干网络得到的局部 class token 连接，专注有意义病灶区域，得到有效的 OCT 特征. 联合损失函数进一步提高多模态眼底疾病诊断效果. 实验结果表明，本文模型能有效提高多模态眼底疾病诊断，有助于眼底疾病的早期检测.

参　考　文　献

[1] HUANG Y J, LYU J Y, CHENG P J, et al. SSIT: Saliency-Guided Self-supervised Image Transformer for Diabetic Retinopathy Grading [EB/OL]. (2022-10-20)[2023-11-10]. https://arxiv.org/abs/2210.10969.

[2] CHEN S B, WU Z Q, LI M Z, et al. FIT-Net: Feature Interaction Transformer Network for Pathologic Myopia Diagnosis [J]. IEEE Transactions on Medical Imaging, 2023, 42(9): 2524-2538.

[3] ELSHARKAWY M, SHARAFELDEEN A, SOLIMAN A, et al. A Novel Computer-Aided Diagnostic System for Early Detection of Diabetic Retinopathy Using 3D-OCT Higher-Order Spatial Appearance Model [J]. Diagnostics, 2022, 12(2): 461-475.

[4] 李锡荣. 多模态深度学习及其在眼科人工智能的应用展望 [J]. 协和医学杂志, 2021, 12(5): 602-607. (LI X R. Multimodal Deep Learning and Its Application Prospect in Ophthalmic Artificial Intelligence [J]. Medical Journal of Peking Union Medical College, 2021, 12(5): 602-607.)

[5] LI X M, JIA M Y, ISLAM M T, et al. Self-supervised Feature Learning via Exploiting Multi-modal Data for Retinal Disease Diagnosis [J]. IEEE Transactions on Medical Imaging, 2020, 39(12): 4023-4033.

[6] LI X M, HU X W, YU L Q, et al. CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading [J]. IEEE Transactions on Medical Imaging, 2019, 39(5): 1483-1493.

[7] WU J D, FANG H H, LI F, et al. Gamma Challenge: Glaucoma Grading from Multi-modality Images [EB/OL]. (2022-12-26)[2023-11-02]. https://arxiv.org/abs/2202.06511.

[8] HUA C H, KIM K, HUYNH-THE T, et al. Convolutional Network with Twofold Feature Augmentation for Diabetic Retinopathy Recognition from Multi-modal Images [J]. IEEE Journal of Biomedical and Health Informatics, 2020, 25(7): 2686-2697.

[9] LIU R, LI Q C, XU F P, et al. Application of Artificial Intelligence-Based Dual-Modality Analysis Combining Fundus Photography and Optical Coherence Tomography in Diabetic Retinopathy Screening in a Community Hospital [J]. BioMedical Engineering OnLine, 2022, 21(1): 47-1-47-11.

[10] HAN J Y, CHOI S, PARK J I, et al. Classifying Neovascular Age-Related Macular Degeneration with a Deep Convolutional Neural Network Based on Optical Coherence Tomography Images [J]. Scientific Reports, 2022, 12(1): 2232-2242.

[11] WANG W S, LI X R, XU Z Y, et al. Learning Two-Stream CNN for Multi-modal Age-Related Macular Degeneration Categorization [J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(8): 4111-4122.

[12] HE X X, DENG Y, FANG L Y, et al. Multi-modal Retinal Image Classification with Modality-Specific Attention Network [J]. IEEE Transactions on Medical Imaging, 2021, 40(6): 1591-1602.

[13] YU Y, ZHU H Q. Transformer-Based Cross-Modal Multi-contrast Network for Ophthalmic Diseases Diagnosis [J]. Biocybernetics and Biomedical Engineering, 2023, 43(3): 507-527.

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition [C]//Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.

[15] 褚张晴晴, 钟志强, 颜子夜, 等. 基于特征融合与注意力机制的脑肿瘤分割算法 [J]. 计算机工程, 2023, 49(10): 154-161. (ZHU Z Q Q, ZHONG Z Q, YAN Z Y, et al. Brain Tumor Segmentation Algorithm Based on Feature Fusion and Attention Mechanism [J]. Computer Engineering, 2023, 49(10): 154-161.)

[16] 蒲秋梅, 田景龙, 邢容畅, 等. 基于改进 Inception-v3 网络的肺炎检测方法 [J]. 东北师大学报(自然科学版), 2023, 55(4): 67-76. (PU Q M, TIAN J L, XING R C, et al. Pneumonia Detection Method Based on Improved Inception-v3 Network [J]. Journal of Northeast Normal University (Natural Science Edition), 2023, 55(4): 67-76.)

[17] 马国祥, 严传波, 杨凌菲, 等. 基于改进的多尺度深度残差网络肝包虫超声影像诊断方法 [J]. 东北师大学报(自然科学版), 2023, 55(1): 80-87. (MA G X, YAN C B, YANG L F, et al. Ultrasonic Image Diagnosis of Hepatic Hydatid Based on Improved Multiscale Depth Residual Network [J]. Journal of Northeast Normal University (Natural Science Edition), 2023, 55(1): 80-87.)

[18] 曹广硕, 黄瑞章, 陈艳平, 等. 基于多模态学习的乳腺癌生存预测研究 [J]. 计算机工程, 2024, 50(1): 296-305. (CAO G S, HUANG R Z, CHEN Y P, et al. Research on Breast Cancer Survival Prediction Based on Multi-modal Learning [J]. Computer Engineering, 2024, 50(1): 296-305.)

[19] 刘兆伟, 方艳红, 郑明宇, 等. 基于注意力机制与多任务的肺部疾病诊断方法 [J]. 计算机工程, 2025, 51(1): 332-342. (LIU Z W, FANG Y H, ZHENG M Y, et al. Lung Disease Diagnosis Method Based on Attention Mechanism and Multi-tasking [J]. Computer Engineering, 2025, 51(1): 332-342.)

[20] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional Block Attention Module [C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.

[21] HENDRYCKS D, GIMPEL K. Gaussian Error Linear Units (Gelus) [EB/OL]. (2023-06-06)[2023-11-01]. https://arxiv.org/abs/1606.08415.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Berlin: Springer, 2017: 6000-6010.

[23] WANG W S, XU Z Y, YU W H, et al. Two-Stream CNN with Loose Pair Training for Multi-modal AMD Categorization [C]//22nd International Conference on Medical Image Computing and Computer Assisted Intervention. Berlin: Springer International Publishing, 2019: 156-164.

[24] YOO T K, CHOI J Y, SEO J G, et al. The Possibility of the Combination of OCT and Fundus Images for Improving the Diagnostic Accuracy of Deep Learning for Age-Related Macular Degeneration: A Preliminary Experiment [J]. Medical & Biological Engineering & Computing, 2019, 57: 677-687.

[25] WANG Q S, GUO Q, LIU X C, et al. Tri-Branch CNN for Age-Related Macular Degeneration Categorization with Incomplete Multi-modality Ophthalmology Images [C]//2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA). Piscataway, NJ: IEEE, 2023: 436-442.

[26] WANG K K, XU C Y, LI G, et al. Combining Convolutional Neural Networks and Self-attention for Fundus Diseases Identification [J]. Scientific Reports, 2023, 13(1): 76-91.

[27] BHATI A, GOUR N, KHANNA P, et al. Discriminative Kernel Convolution Network for Multi-label Ophthalmic Disease Detection on Imbalanced Fundus Image Dataset [EB/OL]. (2022-07-16)[2023-10-30]. https://arxiv.org/abs/2207.07918.

[28] CAI Z Y, LIN L, HE H Q, et al. Corolla: An Efficient Multi-modality Fusion Framework with Supervised Contrastive Learning for Glaucoma Grading [C]//2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). Piscataway, NJ: IEEE, 2022: 1-4.

[29] LAURENS V D M, HINTON G. Visualizing Data Using t-SNE [J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

(责任编辑: 韩 啸)