基于眼底图像的多标签深度学习分类算法研究

Research on Multi-Label Deep Learning Classification Methods for Fundus Image Analysis

姓名 赵天泽

摘　要

眼底图像分类是眼科疾病自动化筛查的核心任务，但其面临高分辨率图像处理复杂、个体差异显著、多标签病症共存等挑战。传统方法因忽略双眼图像的关联性及缺乏对相似病症的特征区分能力，易导致漏诊与误诊。针对上述问题，本文提出一种基于自注意力机制与多标签监督对比学习的自注意力融合分类网络，旨在通过挖掘双眼底图像的互补信息与全局语义关联，提升多标签分类的准确性与鲁棒性。

首先，设计眼间自注意力模块，通过建模双眼图像的空间与通道相关性，动态调整特征权重，增强模型对病灶区域的定位能力。在此基础上，构建自注意力融合网络，结合深度理解卷积核与空洞空间金字塔模块，实现多尺度特征的高效融合，并引入三分支输入策略，融合浅层局部特征与深层语义信息。进一步提出标签水平嵌入策略，利用多标签监督对比学习优化类别间的特征可分性，缓解相似疾病特征混淆问题。此外，设计基于判别受限玻尔兹曼机与多层感知机的标签嵌入分类器，引入软门控机制，强化正常类别与其他病理类别的区分性，突出正常类别作为互斥类别的独特性，提升多标签预测的稳定性。

实验基于 ODIR-5K 数据集验证方法有效性，所提模型在精准度、召回率、F1 分数与 AUC 指标上分别达到 90.39%、91.61%、90.99%与 90.76%，较次优模型提升 1.75%、2.26%、2%与 3.18%。消融实验表明，眼间自注意力模块与多标签对比学习分别贡献了 20.76%和 11.91%的 F1 分数提升。实验结果显示，模型在各类别数据上均表现良好。

本文方法为眼底疾病的多标签分类提供了新的解决方案，其高精度与强泛化能力可辅助临床大规模筛查，尤其在医疗资源匮乏地区具有重要应用价值。

关键词：眼底图像分类；自注意力机制；多标签对比学习；眼间特征融合；软门控机制

ABSTRACT

Fundus image classification is a core task in automated screening for ophthalmic diseases, yet it faces challenges such as complex high-resolution image processing, significant individual variability, and the coexistence of multiple pathological labels. Traditional methods, which overlook interocular correlations and lack discriminative capability for similar pathologies, often lead to missed diagnoses and misclassifications. To address these issues, this study proposes a Self-Attention Fusion Network based on self-attention mechanisms and multi-label supervised contrastive learning, aiming to enhance the accuracy and robustness of multi-label classification by leveraging complementary information and global semantic correlations from bilateral fundus images.

First, an Interocular Self-Attention Block is designed to dynamically adjust feature weights by modeling spatial and channel correlations between bilateral fundus images, thereby improving the model's ability to localize lesion regions. Building on this, a self-attention fusion network is constructed, integrating Deep Understanding Convolutional Kernels and Atrous Spatial Pyramid Pooling modules to achieve efficient multi-scale feature fusion. A three-stream input strategy is introduced to combine shallow local features with deep semantic information. Furthermore, a label-level embedding strategy is proposed, utilizing multi-label supervised contrastive learning to optimize feature discriminability across categories and mitigate feature confusion among similar diseases. In addition, a label embedding classifier based on a Discriminative Restricted Boltzmann Machine and a Multilayer Perceptron is designed. A soft gating mechanism is introduced to enhance the discriminability between the normal class and other pathological classes, highlighting the exclusiveness of the normal class as a mutually exclusive category and improving the stability of multi-label prediction.

Experiments conducted on the ODIR-5K dataset validate the effectiveness of the proposed method. The model achieves precision, recall, F1-score, and AUC values of 90.39%, 91.61%, 90.99%, and 90.76%, respectively, outperforming the suboptimal model by 1.75%, 2.26%, 2%, and 3.18% in these metrics. Ablation studies demonstrate that the Interocular Self-Attention Block and multi-label contrastive learning contribute to F1-score improvements of 20.76% and 11.91%, respectively. Experimental results demonstrate that the model exhibits strong performance across all data categories.

This method provides a novel solution for multi-label classification of fundus diseases. Its high accuracy and strong generalization capability can assist in large-scale clinical screening, particularly offering significant application value in regions with limited medical resources.

KEY WORDS: Fundus image classification; Self-attention mechanism; Multi-label contrastive learning; Interocular feature fusion; Soft gating mechanism

第 3 章　眼底疾病多标签分类自注意力融合网络

3.2 算法设计与实现

3.2.1 模型总体架构

针对眼底图像特点，及分类任务目前面临的挑战，本文提出了一种创新的眼底疾病多标签分类自注意力融合网络（Self-Attention Fusion Net），该网络包含自注意力特征提取框架（Self-Attention Feature Extraction Framework）与标签嵌入分类器（Label Embedding Classification）两个部分。自注意力特征提取框架由自注意力编码模块（Self-Attention Encoding Block，SAEB）和自注意力融合模块（Self-Attention Merge Block，SAMB）两大核心模块组成，两种核心模块均基于本文新提出的眼间自注意力模块（Interocular Self-Attention Block，ISAB）构建。标签嵌入分类器则结合了多标签监督对比学习，采用标签水平嵌入模块（Label Level Embedding Block，LLEB），计算各个类别的标签水平嵌入（label-level embeddings），最终使用基于判别受限玻尔兹曼机（Discriminative Restricted Boltzmann Machine，DRBM）的深层标签判别受限玻尔兹曼机（Deep Label DRBM，DLDRBM）完成分类任务。其整体网络结构如图 2 所示。

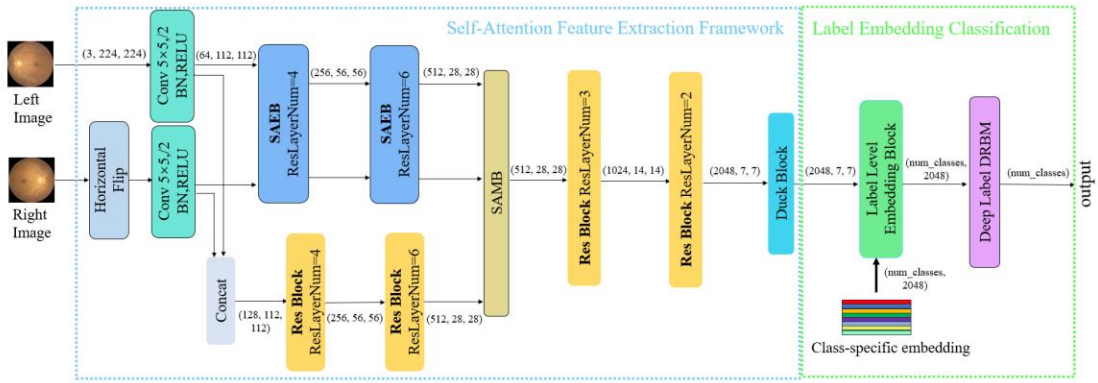

图 1 眼底疾病多标签分类自注意力融合网络

自注意力特征提取框架会同时接受来自同一个人的左眼底图像与水平翻转后的右眼底图像，分别称为左眼特征提取分支与右眼特征提取分支，并在分别进行一个卷积层的预编码后拼接，作为双眼融合特征分支。双眼融合特征分支会在双眼图像特征停留在浅层信息时进行融合，并基于融合后的语义信息继续挖掘深层特征，与深层时再进行融合的双眼特征形成一定程度上的互补。左眼特征提取分支与右眼特征提取分支的输入会在经过两个自注意力编码模块后，与经历两个残差编码模块的双眼融合特征于自注意力融合模块合并融合。

其中，SAEB 同时接受左眼特征提取分支与右眼特征提取分支的输入，并在其中使用不同的残差编码模块为各分支的特征图进行编码与下采样。在两个 SAEB 之后为 SAMB。SAMB 承担了将多个分支的输入融合的作用，三个分支的特征在经过编码后一同输入至自注意力融合模块，并输出融合后的特征图继续进行下采样编码。

融合特征再经过两个多层的残差模块进行特征提炼与增强以及一个深度理解卷积核模块核(deep understanding convolutional kernel, DUCK)[21]模块进行不同尺度感受野的特征整合后，进入标签水平嵌入模块。该模块将得到代表各个类别的 label-level embeddings，其将用于多标签监督对比损失的计算，同时将输入 DLDRBM 进行分类。

最终由 DLDRBM 对标签水平嵌入进行处理，得到最终各个类别的预测值，基于预测值与设置的阈值进行多标签类别判断。

该方法利用自注意力机制有效地建模了左右眼图像之间的潜在关联性，同时结合多标签对比学习策略，进一步提升了模型在多标签分类任务中的性能，尤其对存在特征不明显病症的诊断具有显著优势。

3.2.2 自注意力编码模块

自注意力编码模块接收左眼特征提取分支与右眼特征提取分支的特征图作为输入，由左残差模块（Left Res Block）与右残差模块（Right Res Block）两个残差编码模块与四个眼间自注意力模块构成。其结构如图 3 所示。

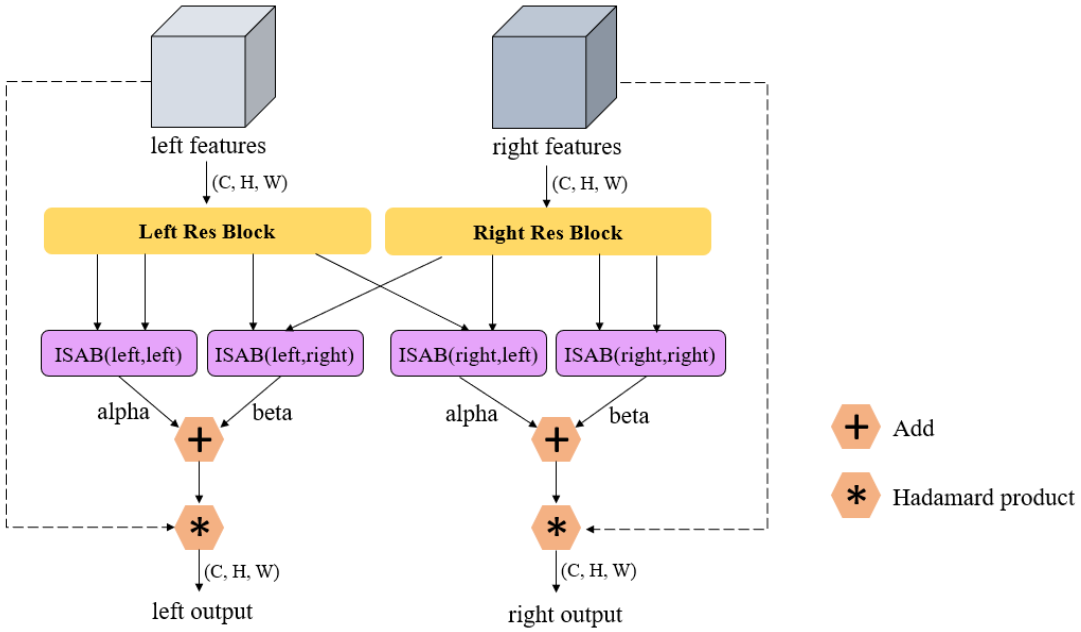

图 2 自注意力编码模块

左残差模块与右残差模块均使用多层残差模块作为编码器，分别处理左眼特征图与右眼特征图。在两个残差模块之后，是四个眼间自注意力模块，每个眼间自注意力模块接受两个特征图输入，分别作为 query 特征图与 key 特征图，四个

眼间自注意力模块则分别处理左眼底图像自身、右眼底图像自身、左眼底图像于右眼底图像以及右眼底图像于左眼底图像的自注意力机制。其中第一个特征图作为 query 特征图，而第二个特征图作为 key 特征图。眼间自注意力模块的结构如图 4 所示。

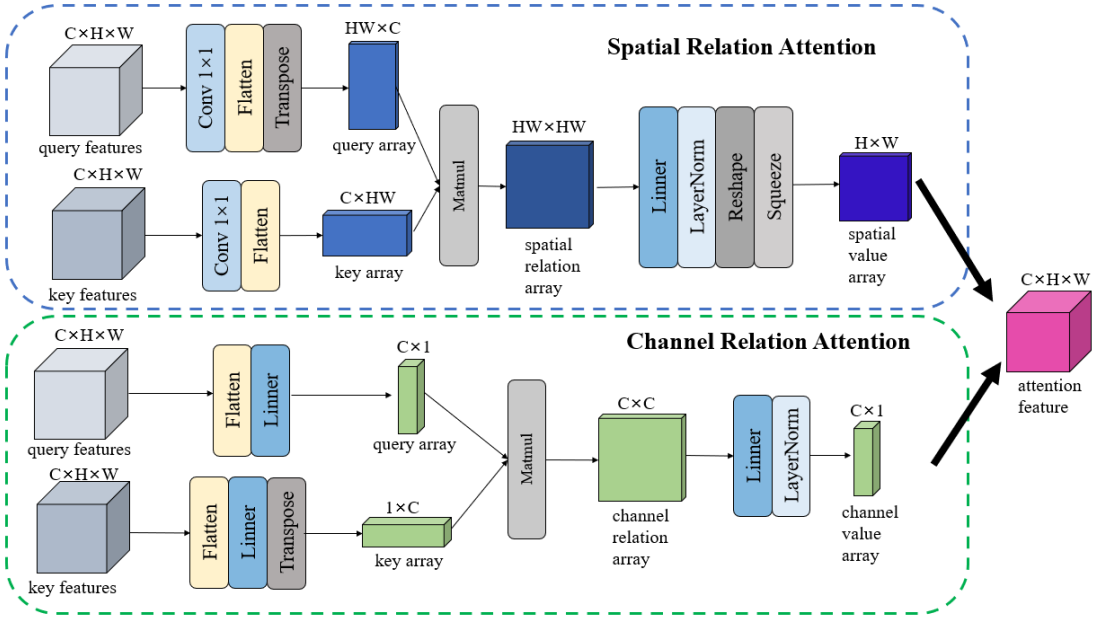

图 3 眼间自注意力模块

ISAB 的计算所得为 query 特征图对于 key 特征图的全局相关性，根据图像之间的相关性，调整全局特征在平面空间维度以及通道方向上的权重，其包含空间相关注意力（Spatial Relation Attention，SRA）与通道相关注意力（Channel Relation Attention，CRA）两部分。

SRA 根据 query 和 key 特征图在空间与通道上的相关性，调整 value 特征图在空间与通道上的权重分配。在空间上的处理，先调整 query 和 key 特征图，将分别通过两个不同的 1×1 卷积核, 得到 query 矩阵与 key 矩阵, 在对两个矩阵进行展平处理后, 再将 query 矩阵进行转置，将转置后的 query 矩阵与 key 矩阵进行矩阵乘法, 得到 relation 矩阵, 该矩阵包含了二维空间中各像素位置之间的相关性. 对 relation 矩阵进行以行为单位的线性运算, 再在运算结束后进行标准化, 并将结果的尺寸调整为宽高与输入特征图相同, 最终得到 spatial value 矩阵, 其代表了每个像素位置与其他像素位置的相关性加权和。

CRA 在通道上的处理，则是将 query 和 key 特征图将分别通过全连接层, 得到每个通道上二维空间上元素的加权和，并对其中之一进行转置, 再进行矩阵乘法。得到的通道相关性矩阵中, 每一个元素均代表了两个通道之间的相关性。再次通过全连接层并进行标准化, 得到每个通道与其他通道的相关性加权之和。

SAEB 中，使用四种 ISAB，不仅计算了左眼特征与右眼特征各自的自注意力，还计算了不同输入顺序下左眼特征与右眼特征之间的注意力。由于双眼特征

源自于同一个人的左右双眼的眼底图像，所以具有高度的相关性与一定的一致性，所以称该注意力为眼间自注意力。第一个输入特征相同的两个 ISAB 所得到的注意力特征分别代表了其自身的全局相关性与对于另一输入特征的全局相关性，两个注意力特征在分别与一个可学习的权重参数 alpha 或 beta 相乘后再相加，得到的最终注意力特征与该模块的第一个输入特征相乘，所得即为经过注意力调整后的输出特征。

该模块在实现了输入特征的全局自注意力机制以外，还实现强相关双图像之间的自注意力机制，根据双眼之间的相似程度对全局特征进行调节，并且，该模块在传统自注意力机制上额外增加了通道方向上的自注意力机制。

3.2.3 自注意力融合模块

在自注意力融合模块 SAMB 中，除了使用本文中提出的 ISAB 以外，也使用了 DUCK 模块与空洞空间池化金字塔(Atrous Spatial Pyramid Pooling，ASPP)[22]模块。

SAMB 会在左眼特征提取分支与右眼特征提取分支之外，额外接受双眼融合特征分支的输入，并在模块内对三个分支的特征进行调整与融合。模型图如图 5 所示。

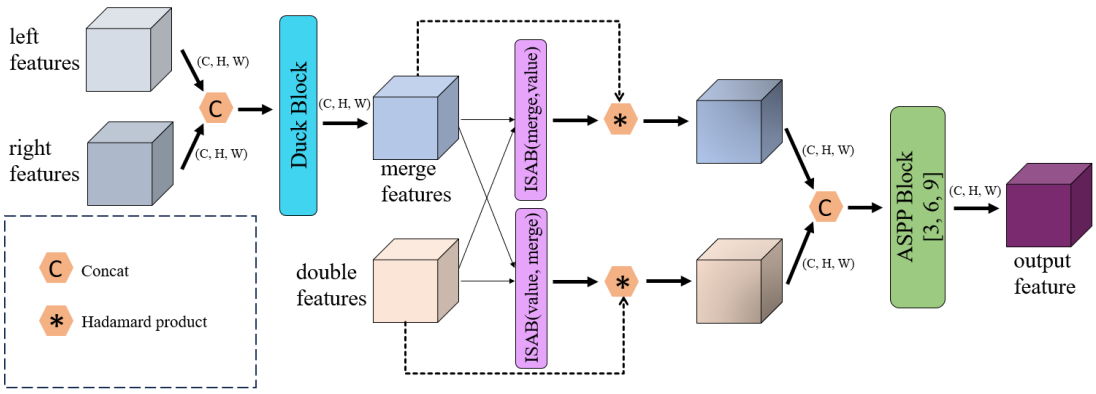

图 4 注意力融合模块

SAMB 模块会首先通过 DUCK 模块对左眼特征与右眼特征的拼接特征进行融合，得到 merge features，并使用两个 ISAB 模块分别计算该特征对于双眼特征提取分支的 double features 的自注意力权重以及 double features 对于 merge features 的自注意力权重。在经过自注意力的权重调整后，再次进行拼接，并通过 ASPP 模块进行语义整合，得到 merge features 与 double features 的融合特征，从而形成更丰富的联合表示。

ASPP 模块可以在编码器进行特征提取之后对信息进行多尺度的整合，其包含了四条卷积分支以及一条池化上采样分支。四条卷积分支为一个 1×1 卷积层以及三个空洞率不同的 3×3 空洞卷积层，能够在不同尺度以不同的信息密度提取特征。池化上采样分支则补充了在全局尺度上的上下文信息，其对输入特征图

的全局平均池化结果进行 1×1 卷积运算，并通过双线性插值进行上采样，还原至与其他分支的输出特征图相同的尺寸。将一共五条分支的输出特征图进行拼接，再通过 1×1 卷积进行通道调整，得到整个模块的输出特征图。

这种策略避免了直接将特征图强硬地融合在一起, 而是结合特征图矩阵之间的相关性, 在注意力上进行调整，对同样是双眼融合特征，但语义信息层次不同的两个特征进行动态调整后再融合，有效地提升了模型对双眼底图像的特征融合能力。

3.2.4 深度理解卷积核模块核

深度理解卷积核模块 DUCK 是一种新型的卷积块, 其设计了 6 种并行的卷积块, 使得网络在训练时采用最合适的卷积块。本文所采用的 DUCK 块如图 6 所示, 其中第 1 个分支和第 2 个分支分别模拟 15×15 和 7×7 的大卷积核, 使用扩展卷积来减少模拟大卷积核所需要的参数, 同时使得网络更好地理解高级别的特征, 有助于保持分割的整体结构完整性。最后一个分支利用通过组合空间可分离卷积, 组合 1×N 卷积核和 N×1 卷积核来模拟 N×N 大小的卷积核, 这样可以使用更少的参数量来实现相同的感受野, 但由于受到卷积核方向的限制, 会难以捕捉感受野内对角线方向的信息。中间 3 个分支分别组合了不同层数的残差模块来获得不同的感受野, 模拟 5×5、9×9 与 13×13 的卷积核大小, 这对于捕获图像中不同尺寸区域的细微特征至关重要。

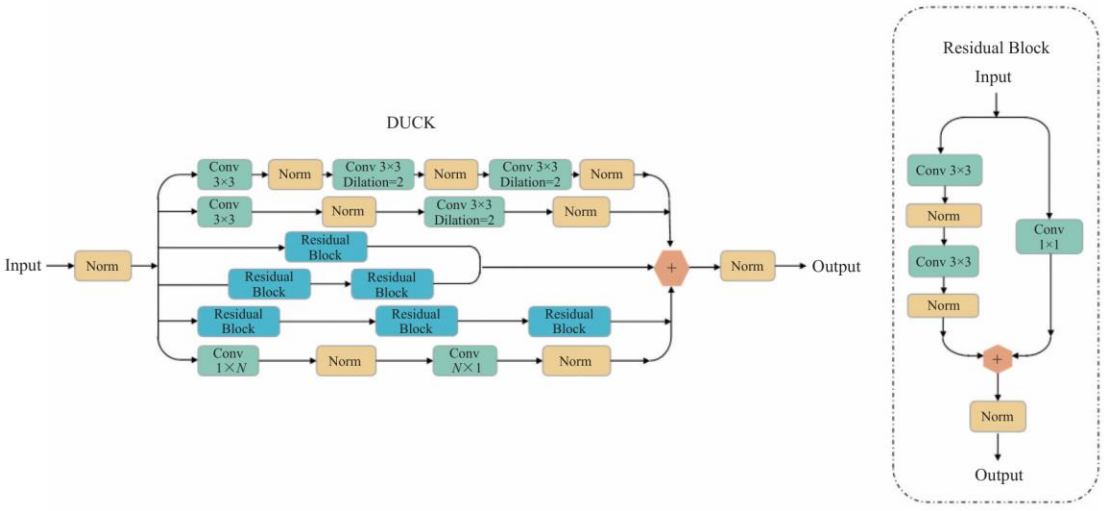

图 5 深度理解卷积核模块

DUCK 模块通过并行操作 6 种不同的卷积块, 使得网络能够在学习过程中选择最适合的特征提取策略. 这个模块的扩展感受野有助于在特征图尺寸逐步扩大的插值过程中, 有效地推理和调整放大区域的细节。

3.2.5 标签嵌入分类器

本文使用的标签嵌入分类器包含两部分，分别是标签水平嵌入模块 LLEB 与深层标签判别受限玻尔兹曼机 DLDRBM。

LLEB 使用多头自注意力机制，捕捉自注意力融合网络的输出与类别专属向量（Class-specific embedding）之间的特征关联，生成 label-level embedding，用来计算多标签监督对比损失，并通过 DLDRBM 进行最终分类。label-level embedding 为一个二维张量，其行数与分类任务的类别数相同，列数则与输入 LLEB 的特征通道数相同，每一行的 embedding 代表了对应类别的深层特征。

类别专属向量为一个随机矩阵 $U \in R^{L\times C}$，其中每一行都是一个类特定的嵌入向量。U在训练初期随机初始化，并在训练过程中作为可学习参数进行优化。

使用r_i表示图像的最终输出特征，为了生成图像的类别水平向量g_i，利用多头注意力机制来捕捉r_i和U之间的交互关系，如公式 14 所示：

$$g_i = \mathrm{MultiAttBlock}(U, r_i, r_i) \qquad \text{（公式 14）}$$

其中，U作为查询（Query），r_i作为键（Key）和值（Value），多头注意力机制能够学习标签对图像不同区域的关注权重。

类别水平向量g_i将用来计算多标签监督对比损失，并接着输入 DLDRBM，计算最终类别的预测概率。

DLDRBM 由多个 DRBM 与两个多层感知机（Multilayer Perceptron，MLP）组成，其中 DRBM 负责对类别水平向量的降维与特征提取，两个 MLP 则分别进行二分类与全类别分类。

DLDRBM 中会为每个类别水平向量配备一个对应的类别 DRBM，类别 DRBM 会处理对应类别的类别水平向量，对其进行降维与特征提取。在类别 DRBM 之后，会再次使用一个异常 DRBM，对所有非正常类别的类别水平向量进行合并与降维，得到一个与正常类别水平向量形状相同的异常类别水平向量。对各个类别的水平向量分别设置一个 DRBM 进行降维，可以独立提取各个类别的语义信息，不会混淆不同类别的特征。在展开后再次经过 DRBM，可以建模所有非正常类别间的交互，建立相似类别之间的联系。

为了更好地建模正常类别与其他病理类别之间的互斥关系，本文引入了一种基于软门控机制（Soft Gating Mechanism）的分类策略。首先使用一个二分类 MLP 和 Softmax 层对正常类别水平向量与异常类别水平向量进行分类，以预测该样本属于正常或异常的概率。该二分类概率作为软门控信号，引导全类别分类过程。在全分类阶段，正常类别的水平向量将与预测得到的正常概率相乘，异常类别的水平向量将与异常概率相乘。经门控调制后的向量随后被输入全类别 MLP 中，并通过 Sigmoid 激活函数输出多标签预测结果。

该软门控机制能够有效强化正常类别与其他病理类别在语义特征空间中的区分性，从而突出正常类别作为互斥类别的独特性。同时，相较于硬门控方法，该机制具备可微性，有利于模型的端到端优化。

该分类器中的 DRBM 均使用对比散度进行独立训练，对比散度公式如公式 15 所示：

$$L_{\text{CD}} = \left(-\sum_i v_i \cdot h_i\right)_{\text{data}} + \left(\sum_i v_i' \cdot h_i'\right)_{\text{reconstruction}} \quad \text{（公式 15）}$$

本文采用 Straight-Through Estimator (STE) 策略对离散操作进行梯度近似。相较于基于采样的方法可能导致的梯度中断问题，STE 能够在保留离散决策行为的同时实现梯度的稳定传递，从而支持更稳定的端到端训练过程。

3.2.6 多标签监督对比学习

在对比学习中，通常的目标是将相似样本（正样本）在嵌入空间中拉近，并将不同样本（负样本）拉远。对于单标签分类任务，监督对比学习的方法通常将具有相同类别标签的样本视为正样本，而其他类别的样本视为负样本，从而提升分类器的判别能力。

然而，在多标签分类任务中，每个图像可能同时属于多个类别，这使得传统的监督对比学习难以直接应用。一个图像的多个标签可能与另一图像的部分标签匹配，因此定义正样本和负样本的过程变得复杂。为了克服上述问题，本文采用了一种标签级别的监督对比学习方法。其核心思想是为每个图像学习多个 label-level embeddings，然后在这些嵌入的基础上定义对比学习的正样本和负样本。

在本文中，某个图像的标签级嵌入z_{ij}（图像i在标签j语境下的嵌入）被视为一个锚点（anchor）。该锚点的正样本（positive samples）定义为其他图像中相同标签j的嵌入，即公式 16 所示：

$$P(i,j) = \{z_{kj} \in A(i,j) \mid y_{kj} = y_{ij} = 1\} \quad \text{（公式 16）}$$

负样本（negative samples）定义为其他图像中不同标签的嵌入，即公式 17 所示：

$$A(i,j) = I \setminus z_{ij} \quad \text{（公式 17）}$$

传统的监督对比损失在单标签任务中的形式如公式 18 所示：

$$L_{\text{sup}} = \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \quad \text{（公式 18）}$$

本文中，针对每个标签级嵌入z_{ij}计算损失，具体计算方式见公式 19 所示：

$$L_{\text{con}}^{ij} = -\frac{1}{|P(i,j)|}\sum_{z_p \in P(i,j)} \log \frac{\exp\left(z_{ij} \cdot z_p/\tau\right)}{\sum_{z_a \in A(i,j)} \exp\left(z_{ij} \cdot z_a/\tau\right)} \quad \text{（公式 19）}$$

最终的多标签监督对比损失为公式 20 所示：

$$L_{\text{con}} = \sum_{z_{ij} \in I} L_{\text{con}}^{ij} \quad \text{（公式 20）}$$

3.2.7 损失函数

本文的最终损失函数由焦点损失（Focal Loss）和多标签监督对比损失组成，如公式 21 所示：

$$L = L_{\text{FL}} + \mathbf{w} * L_{\text{con}} \quad \text{（公式 21）}$$

w为多标签监督对比损失的损失权重，本文中对其进行动态调整，保证两种损失在训练过程中维持在同一个数量级。

其中焦点损失的公式见公式 22 所示：

$$\mathcal{L}_{\mathcal{FL}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \alpha\left(1-\widehat{y_{ij}}\right)^{\gamma} y_{ij} \log\left(\widehat{y_{ij}}\right) + (1-\alpha)\widehat{y_{ij}}^{\gamma}\left(1-y_{ij}\right)\log\left(1-\widehat{y_{ij}}\right)$$

（公式 22）

其中：α是平衡因子；γ是调节因子，控制难分类样本的权重；$\widehat{y_{ij}}$是模型的预测概率，y_{ij}是真实标签，N 是样本总数，C 是类别数。

本文采用这样的损失函数，可以通过拉近相同标签的嵌入向量，增强相同类别样本的聚合性，从而提高分类精度。避免了不同类别之间的特征混叠，使得模型能够更好地区分不同类别的语义信息。

3.3 本章小结

本章提出了一种基于对比学习的眼底疾病多标签分类网络，并详细阐述了模型的设计思路与实现方式。首先，我们分析了眼底图像的特点及分类过程中面临的挑战，包括图像质量的变化、疾病的共现性以及双眼之间的潜在联系。针对这些问题，本文提出了眼间自注意力模块，并以该模块为基础构建了自注意力融合网络的两大核心模块，以增强模型对关键区域的关注能力，并结合多标签监督对比学习优化分类性能。

自注意力融合网络的两大核心模块：自注意力编码模块（Self-Attention

Encoding Block, SAEB）和自注意力融合模块（Self-Attention Merge Block, SAMB），分别用于提取双眼图像特征与进行跨眼融合。在此基础上，本文使用标签嵌入分类器，其包含标签水平嵌入模块（Label Level Embedding Block，LLEB），用于捕捉类别级别的特征表示，并使用深层标签判别受限玻尔兹曼机（Deep Label DRBM，DLDRBM）完成分类任务。损失值的计算，则采用焦点损失（Focal Loss）与多标签监督对比损失相结合的损失函数，以增强类别间的区分度。

该方法能够有效利用双眼图像的互补信息，提高多标签分类的准确性，特别是在易混淆类别的识别上具有显著优势。通过自注意力机制与对比学习相结合，模型在学习特征的同时优化了样本间的表示，使其更加适应复杂的眼底图像分类任务。

第 4 章　实验结果

4.3 对比实验及性能评估

本文的评价标准采用精准度(Precision)、召回率 (Recall)、F1-score、Kappa Score 和 AUC 这 5 个评价指标对比客观结果, 反映每种模型在该数据集上多标签分类任务中的表现。各个指标的计算如公式 23-29 所示。

$$Precision = \frac{TP}{FP + TP} \quad （公式 23）$$

$$Recall = \frac{TP}{TP + FN} \quad （公式 24）$$

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{Precision + Recall} \quad （公式 25）$$

$$KappaScore = \frac{P_o - P_e}{1 - P_e} \quad （公式 26）$$

$$P_o = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r} (TP_i + FN_i)}, \qquad P_e = \frac{\sum_{i=1}^{r} [TP_i + (TP_i + FN_i)]}{N^2} \quad （公式 27）$$

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right) dx \quad （公式 28）$$

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \quad （公式 29）$$

公式中，TP、TN、FP、FN 分别是真正例(True Positive)、真负例（True Negative）、假正例（False Positive）和假负例（False Negative）。

为了验证本文提出的自注意力融合网络有效性, 将本文方法与经典模型 ResNet50、VGG16[24]、HRNet[25]与同样基于多标签对比学习损失的 MulCon[7], 以及同为针对双眼底图像多标签分类任务所提出的 BFENet[26]进行了实验对比。其中只有 BFENet 实现了双眼底图像的融合，所以本文使用逐元素乘法（Element-wise Multiplication）对对比实验中其他方法的左右眼分类结果进行融

合。

根据表 2 中的数据，通过对比可见，所提方法（Ours）在各个指标上均达到最优水平，相较于其他对比模型，所提方法的综合性能显著优于现有方法。

表 1 不同方法在 ODIR-5K 公开数据上的分类表现

Method	Precision	Recall	F1	AUC	Kappa Score
ResNet50	63.74	65.84	64.77	76.46	65.69
VGG16	50.73	52.86	51.77	68.47	55.21
HRNet	66.89	67.93	67.41	79.48	67.78
MulCon	78.94	80.67	79.79	78.43	79.86
BFENet	88.15	88.97	88.56	84.69	84.68
Ours	**90.39**	**91.61**	**90.99**	**90.76**	**88.42**

具体而言，ResNet50 和 VGG16 等经典模型的性能相对较低，其 F1 分数分别为 64.77 和 51.77，表明其在复杂任务中可能存在特征提取能力不足的问题。HRNet 通过改进网络结构、MulCon 通过引入多标签对比学习损失，性能有所提升，但 F1 分数仍低于 80。BFENet 通过设计更复杂的双流交互架构，性能进一步提升，F1 分数分别达到 88.56。本文所提方法通过引入眼间自注意力模块以及在其基础上的自注意力编码模块和自注意力融合模块，进一步挖掘任务相关的细粒度特征，最终在精准度、召回率、F1 分数与 AUC 指标上分别达到了 90.39%、91.61%、90.99%、90.76%，较次优模型分别提升 1.75%、2.26%、2%、3.18%，验证了其在全局特征建模与眼间关联性处理上的有效性。

为了验证所提网络使用残差模块作为编码模块的优势，本文通过实验对比不同 CNN 网络所使用的编码模块对多标签分类任务的影响。本实验使用不同 CNN 网络的编码模块取代本文中的残差模块，并采用迁移学习策略，利用在 ImageNet 数据集上预训练的模型作为初始特征提取器。表 3 中展示了使用不同网络编码模块实现的自注意力融合网络的性能对比。

表 2 本文网络使用不同网络编码模块的表现

Method	Precision	Recall	F1	AUC	Kappa Score
VGG16	66.53	69.81	68.13	74.98	69.61
DenseNet-169[27]	74.23	75.82	75.02	76.08	72.11
ResNet50	86.55	87.86	87.20	87.41	85.69.
MobileNetV3[28]	78.48	79.39	78.93	81.92	76.97
EfficientNetV2[29]	83.12	84.68	83.89	84.11	83.58

实验结果表明，使用残差网络 ResNet 时，各项指标表现最好，在精准度、召回率、F1 分数、AUC 指标与 Kappa 系数上分别达到了 86.55%、87.86%、87.20%、87.41%、85.69%。所以本文使用残差模块作为网络中的主要编码模块，并且根据训练过程中的实际性能表现，调整残差模块的层数，使得本文网络的性能进一

步提升。

4.4 消融实验

为了验证模型中各种结构对于模型性能的影响,进行了消融实验。实验包括:不进行数据处理，不使用对比学习损失（分类器也进行相应调整），SAEB 中不使用 ISAB（仅保留残差模块），SAMB 替换为逐元素乘法 4 种设置。并给出了不同结构的评价指标，如表 4 所示。

表 3 不同结构的评价指标

Method	Precision	Recall	F1	AUC	Kappa Score
不进行数据处理	83.56	85.47	84.50	85.83	85.42
不使用对比学习损失	78.63	79.54	79.08	83.56	80.46
SAEB 中不使用 ISAB	69.45	71.04	70.23	79.78	74.57
SAMB 替换为逐元素乘法	73.85	76.08.	74.95	78.49	75.39
Ours	**90.39**	**91.61**	**90.99**	**90.76**	**88.42**

移除自动裁剪等预处理操作后,模型的精准度、召回率、F1 分数、Kappa Score 与 AUC 分别下降至 83.56%、85.47%、84.50%、85.42%和 85.83%。这表明，眼底图像的黑色边界与不规则形状会引入噪声干扰特征提取,导致模型对有效区域的关注度降低，验证了数据预处理对提升分类性能的必要性。

禁用多标签监督对比损失后，模型性能显著下降。对比学习的缺失导致类别间特征区分度不足，尤其在多标签场景下，相似疾病的特征容易混淆，进一步证明了对比学习在增强类别语义可分性方面的关键作用。

移除自注意力编码模块 SAEB 中的眼间自注意力 ISAB 后，模型性能大幅降低。实验证明，ISAB 模块通过捕捉双眼图像的全局空间与通道关联性，显著提升了模型对关键区域的定位能力，其缺失会导致特征融合效果退化。

将自注意力融合模块 SAMB 替换为简单的逐元素乘法后，F1 分数降至 74.95%，AUC 仅为 78.49%。逐元素乘法无法动态调整特征权重，难以建模复杂语义关联，而 SAMB 通过 ISAB 模块建立眼间联系，并使用 DUCK 模块与 ASPP 模块的多尺度特征融合，有效解决了这一问题。

实验结果表明，数据预处理、多标签对比学习、SAEB 模块及 SAMB 模块共同构成了模型性能提升的核心支撑。

4.5 本章小结

本章基于 ODIR-5K 数据集的公开部分对提出的自注意力融合网络进行了全面验证。首先，通过对比实验表明，所提方法在精准度、召回率、F1 分数与 AUC 指标上均优于现有方法，较次优模型分别提升 1.75%、2.26%、2%与 3.18%。

其次，消融实验进一步揭示了各模块的独立贡献，包括：眼间自注意力模块 ISAB 通过建模双眼图像的全局关联性,显著提升了模型对病灶区域的定位能力；

自注意力融合模块 SAMB 结合多尺度特征融合策略，动态调整不同深度的语义信息并进行融合，增强了复杂语义信息的表达能力；多标签监督对比学习通过拉近同类标签的嵌入表示，有效缓解了多标签分类中的特征混淆问题。

实验结果充分验证了本文方法在双眼底图像多标签分类任务中的有效性与鲁棒性，为临床眼底疾病的自动化筛查提供了可靠的技术支持。

第 5 章　总结与展望

本文针对单独对单眼底图像进行分析处理时会忽略同一患者双眼之间的潜在特征联系，从而导致特征不明显的病症出现漏诊、特征易混淆的病症出现错诊这一挑战，提出了一种基于自注意力机制与多标签监督对比学习的的自注意力融合网络。主要贡献包括：

（1）提出眼间自注意力模块 ISAB，通过联合建模双眼图像的全局空间与通道相关性，增强了模型对病灶区域的关注能力；

（2）构建自注意力融合模块，基于本文提出的 ISAB，结合 DUCK 模块与 ASPP 模块、以及网络设计中的三分支输入，实现了多尺度特征、多深度语义信息的高效融合；

（3）设计多标签监督对比学习策略，通过标签水平嵌入向量优化分类任务，同时构建适用于标签水平嵌入向量的特殊分类器，提升了模型对相似疾病的区分度。

实验表明，所提方法在 ODIR-5K 数据集上综合性能显著优于现有方法，在精准度、召回率、F1 分数与 AUC 指标上分别达到了 90.39%、91.61%、90.99%、90.76%，较次优模型分别提升 1.75%、2.26%、2%、3.18%，验证了其在复杂多标签场景下的实用性。

尽管本文方法取得了显著效果，但仍存在以下改进空间：

1）数据多样性不足：当前实验依赖单一公开数据集，未来可引入更多跨设备、跨人种的眼底图像以增强泛化性；

2）计算效率优化：模型参数量较大，需探索轻量化设计（如知识蒸馏）以适配临床实时诊断需求；

3）可解释性增强：进一步结合病理学先验知识，设计可解释性更强的注意力机制，辅助医生理解模型决策过程；

4）多模态融合：融合眼底图像与患者临床数据（如病史、基因组信息），构建更全面的诊断系统。

未来工作将围绕上述方向展开，推动深度学习在眼科医学中的深度应用。

参考文献

[21] Dumitru RG, Peteleaza D, Craciun C. Using DUCK-Net for polyp image segmentation. Scientific Reports, 2023, 13(1): 9803.

[22] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[25] Sun K, Zhao Y, Jiang B, et al. High-resolution representations for labeling pixels and regions[J]. arXiv preprint arXiv:1904.04514, 2019.

[26] X. Ou, L. Gao, X. Quan, H. Zhang, J. Yang, W. Li, BFENet: a two-stream interaction CNN method for multi-label ophthalmic diseases classification with bilateral fundus images, Comput. Methods Programs Biomed. 219 (2022), 106739

[27] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[28] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.

[29] Tan M, Le Q. Efficientnetv2: Smaller models and faster training[C]//International conference on machine learning. PMLR, 2021: 10096-10106.