第 49 卷 第 20 期/2022 年 10 月/中国激光 研究论文

中国激光

# 基于集成卷积神经网络和 Vit 的眼底图像分类研究

袁媛，陈明惠*，柯舒婷，王腾，何龙喜，吕林杰，孙好，刘健南

上海理工大学健康科学与工程学院，上海介入医疗器械工程技术研究中心，教育部医学光学工程中心，上海 200093

**摘要** 在眼底图像的分类任务中，卷积神经网络(CNN)的应用较为普遍，但随着 Transformer 应用的推进，Vit(Vision Transformer)模型在医学图像的领域上展现了更高的性能。然而 Vit 模型通常需要在大型数据集上预训练，受医学图像获取成本较高的限制。因此，本文提出一种基于 EfficientNet-Vit 集成模型的眼底图像分类方法，此方法将卷积神经网络模型 EfficientNetV2-S 和 Vit 模型相结合，分别使用两种完全不同的方法提取眼底图像的特征，通过自适应加权融合算法计算得到最优加权因子 0.6 和 0.4，利用加权软投票法进行模型集成，从而获得更好的分类结果。实验证明，相比于集成前，集成后的模型分类准确率分别提高了 0.5%和 1.6%。

**关键词** 生物光学；眼科学；眼底疾病；图像分类；集成模型；加权融合

**中图分类号** TP391 **文献标志码** A **DOI**：10.3788/CJL202249.2007205

## 1 引 言

眼底成像系统可用于捕捉人眼解剖结构和异常视网膜图像，是观察和检测各种眼科疾病的主要工具。彩色眼底图像是诊断眼病的最基本方法[1]。眼底图像中的血管、视盘、中央凹和黄斑等解剖结构及其周围组织的改变表明患者可能存在糖尿病视网膜病变、黄斑病变、病理近视和白内障等疾病，这有助于进行及时有效的治疗，而早期诊断和及时治疗可有效降低患病率[2]。然而，国内罹患眼病的患者数量较为庞大，眼科医师等相关医疗资源相对有限，使得眼底疾病检查诊断需求过剩。因此，迫切需要探索其他方法来弥补需求漏洞。计算机辅助诊断可以通过对眼底图像的正确处理和分析得出对医生临床诊断有参考价值的信息，它不仅可以减轻医生的工作量，而且具有客观、快速、准确的优点[3]。

近几年来，越来越多的人们致力于将人工智能技术和深度学习方法应用到各种图像处理研究中，例如：Wan 等[4]提出一种基于深度卷积神经网络的高度近视眼底图像分类方法；Imran 等[5]通过卷积和递归神经网络完成了眼底图像的白内障分类；连超铭等[6]通过优化神经网络完成了对视网膜黄斑病变图像自动分类；孙雨琛等[7]和郑雯等[8]分别提出深度神经网络来完成对糖尿病视网膜病变的诊断。上述研究均是对某一疾病的分类，核心技术是具有强大学习能力的卷积神经网络(CNN)，如高度近视眼底的诊断和分类，或白内障等疾病的检测等。虽然这些研究在眼底图像领域中取得了显著的效果，但是在某些情况下，上述算法或方法的结果过于单一，无法提供特定的视网膜病变图像。眼底图像的内容复杂，不同类别的图像可能有很高的相似性[9]，而且眼底图像的检测和识别容易受到血管变形、图像亮度、对比度和其他受损区域的影响[10]。如果使用传统的 CNN 进行分类检测，则需要大量的训练样本，容易陷入局部最优，训练效率较低，难以比较好地解决类内差异大和类间相似性高的问题。为了解决这个问题，Yu 等[11]首次尝试将 Vit(Vision Transformer)应用于医学图像分类任务，探索了 Transformer 模型在眼底图像视网膜疾病分类中的适用性，并在相同的预训练设置下取得了优于 CNN 的性能，但是 Transformer 模型的预训练通常需要大量的数据，对于医学图像分类任务，由于数据收集和注释的成本较高，通常图像数量有限，远少于自然图像的数量。

本研究提出一种将改进的 EfficientNetV2[12]和 Vit 集成的分类模型 EfficientNet-Vit，并用于眼底图像分类，该模型可以用两种完全不同的方式来提取眼底图像的特征，从而达到更好的分类效果。集成 EfficientNetV2-S[13]模型和 Vit 模型，通过自适应加权融合算法计算最优权重，并利用加权软投票法集成了 EfficientNet-Vit 模型；在 OIA-ODIR (南开大学)提供的数据集[14]上进行训练和测试，对眼底图像中 5 种眼底图像进行识别分类，并将该模型与 EfficientNetV2-S 等其他模型进行比较，发现所提模型在分类准确率上有着更好的表现。实验结果表明，所提模型可以对

**收稿日期**：2022-03-04；**修回日期**：2022-04-12；**录用日期**：2022-06-06

**基金项目**：上海市科委产学研医项目(15DZ1940400)

**通信作者**：*cmhui.43@163.com

较少的数据集进行训练，并取得较好的效果，为医学诊断提供可靠的帮助。

## 2 实验方法及原理

本研究的眼底图像分类主要包括 3 个步骤：1）数据增强及预处理，即对采集到的眼底图像进行初步筛选和预处理，划分数据集，并对训练集进行自动数据增强和扩充；2）训练改进的 EfficientNetV2-S 模型和 Vit 模型，通过自适应加权算法计算最优加权因子，集成 EfficientNet-Vit 模型；3）利用训练好的模型进行眼底图像分类，并测试模型的性能。

在本研究中，对眼底图像进行预处理后，首先通过迁移学习微调的方式训练改进后的 EfficientNetV2-S 模型和 Vit 模型；然后，通过自使用加权算法计算最优加权因子加权后集成 EfficientNet-Vit 模型；最后，比较集成后的模型和其他模型的性能指标。

### 2.1 EfficientNetV2 模型和 Vit 模型

本研究使用的 CNN 基础模型为 EfficientNetV2-S，EfficientNetV2 是在 EfficientNetV1 的基础之上进行改进的模型。EfficientNetV1 架构使用带有大量深度卷积[15]的 MBConv 层，虽然拥有比常规卷积更少的参数和浮点运算数（FLOPs），但通常不能充分利用现代 GPU 加速器，这就意味着减少 FLOPs 并不一定会提高训练速度。因此，EfficientNetV2 中使用 Fused-MBConv，它将 MBConv 中的深度 3×3 卷积和扩展 1×1 卷积替换为常规的 3×3 卷积。EfficientNetV2 架构在早期层中广泛利用了 MBConv 和新添加的 Fused-MBConv。EfficientNetV2 使用了更多的 3×3 卷积核，而不是 EfficientNetV1 中的 5×5 卷积，并且移除 EfficientNetV1 中的最后一个 stride-1 阶段，提出一种可根据图像大小调整正则化参数的渐进式学习方法，并在最终实验结果验证中发现该模型性能明显提高。EfficientNetV2 模型有 S、M、L 3 个不同的大小，本研究选用 EfficientNetV2-S 模型，其架构如表 1 所示。

表 1 EfficientNetV2-S 架构

Table 1 EfficientNetV2-S architecture

| Stage | Operator | Stride | Number of channels | Number of layers |
|---|---|---|---|---|
| 0 | Conv 3×3 | 2 | 24 | 1 |
| 1 | Fused-MBConv1, k 3×3 | 1 | 24 | 2 |
| 2 | Fused-MBConv4, k 3×3 | 2 | 48 | 4 |
| 3 | Fused-MBConv4, k 3×3 | 2 | 64 | 4 |
| 4 | MBConv4, k 3×3, SimAM | 2 | 128 | 6 |
| 5 | MBConv6, k 3×3, SimAM | 1 | 160 | 9 |
| 6 | MBConv6, k 3×3, SimAM | 2 | 272 | 15 |
| 7 | Conv 1×1 & Pooling & FC | – | 1792 | 1 |

本研究使用的 Vit 是一种完全基于 Transformer 的架构实现的模型[16]，并在 ImageNet 训练集上进行预训练。经典的 Transformer 结构输出采用的是嵌入的字符或单词，因此为了能够对二维图像进行处理，本研究在准备兼容的眼底图像数据时，将输入的图像 $\boldsymbol{x}\in\mathbb{R}^{H\times W\times C}$ 重塑为一连串扁平化的二维图像块序列 $\boldsymbol{x}_P\in\mathbb{R}^{N\times(P^2\times C)}$，其中，$C$、$H$ 和 $W$ 分别表示图像的通道数、高度和宽度，$P$ 为每个图像块的分辨率，分割后图像块的大小固定为 $P\times P$，$N=HW/P^2$ 是产生的图像块数量，这也是 Transformer 的有效输入序列长度。Transformer 在其所有层中使用恒定的潜在向量大小 $D$，因此本研究将图像块进一步扁平化为一维格式，并通过可训练的线性投影映射到 $D$ 维，这个投影的输出称为图像块嵌入。但是，图像块嵌入不仅需要单个图像块的位置信息，还需要对位置进行编码，添加位置嵌入，将位置嵌入编码为一个可学习的一维向量加入到图像块嵌入中，得到单个图像块的最终嵌入 $\boldsymbol{z}_0$ 为

$$\boldsymbol{z}_0=\left[\boldsymbol{x}_{\text{class}}:\boldsymbol{x}_P^{(1)}\boldsymbol{E}:\boldsymbol{x}_P^{(2)}\boldsymbol{E}:\cdots:\boldsymbol{x}_P^{(N)}\boldsymbol{E}\right]+\boldsymbol{E}_{\text{pos}},\ \boldsymbol{E}\in\mathbb{R}^{(P^2\times C)\times D},\ \boldsymbol{E}_{\text{pos}}\in\mathbb{R}^{(N+1)\times D},\tag{1}$$

式中：$\boldsymbol{x}_{\text{class}}$ 为可学习的类标记，将其作为 class token 部分的向量，在最后分类时，不使用其他维度的向量，只使用 class token 部分的向量；$\boldsymbol{x}_P^{(i)}$ 为由眼底图像块重塑而成的第 $i$ 个图像块；$\boldsymbol{E}$ 为训练好的嵌入矩阵；$\boldsymbol{E}_{\text{pos}}$ 为可训练的位置嵌入。

将得到的向量序列 $\boldsymbol{z}_0$ 作为组合送入标准的 transformer 编码器（encoder）。Vit 由 $L$ 个相同的 transformer 编码器组成。编码器经过层归一化（LayerNorm）、多头注意力机制（MSA）和多层感知器（MLP）处理，其中层归一化应用在每个编码器模

块之前，对输入的图像块进行层归一化标准化处理，即

$$LN(x)=\frac{x-\mu}{\delta}\gamma+\beta, \tag{2}$$

式中：$x$ 为神经网络的净输入；$\mu$ 和 $\delta$ 分别为均值和标准差；$\gamma$ 和 $\beta$ 分别为缩放和平移的参数向量，用于加快网络训练和稳定收敛。多头注意力机制和多层感知器的输出为

$$z'_{\ell}=\zeta_{\mathrm{MSA}}[LN(z_{\ell-1})]+z_{\ell-1},\ \ell=1,\cdots,L, \tag{3}$$

$$z_{\ell}=\zeta_{\mathrm{MLP}}[LN(z'_{\ell})]+z'_{\ell},\ \ell=1,\cdots,L, \tag{4}$$

式中：$\zeta_{\mathrm{MSA}}$ 和 $\zeta_{\mathrm{MLP}}$ 表示 MSA 和 MLP 的输出；$z_{\ell-1}$ 为第 $\ell$ 层的输入。

多头注意力可以表示为

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(H_{\mathrm{head1}},H_{\mathrm{head2}},\cdots,H_{\mathrm{headk}})\boldsymbol{W}^{o}, \tag{5}$$

式中：$\boldsymbol{Q}$ 为查询矩阵；$\boldsymbol{K}$ 为关键矩阵；$\boldsymbol{V}$ 为值矩阵；$\boldsymbol{W}^{o}$ 为可学习的参数。$H_{\mathrm{head}i}(i=1,\cdots,k)$可以表示为

$$H_{\mathrm{head}i}=\mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right)\boldsymbol{V},i=1,\cdots,k, \tag{6}$$

式中：$d_{\boldsymbol{K}}$ 为 $\boldsymbol{K}$ 的维度。

多层感知器由两个全连接层组成，中间有一个高斯误差线性单元（GeLU）激活层。最后，使用 class token 接一层全连接层进行分类任务。将眼底图像块分为以下 5 类：正常（normal）、糖尿病视网膜病变（DR）、老年黄斑变性（ARMD）、近视（myopia）、白内障（cataract）。Vit 模型概述如图 1 所示。

$$y=\mathrm{classifier}[LN(z_L^{(0)})], \tag{7}$$

式中：$z_L^{(0)}$ 为 transformer 编码器的输出；$y$ 为分类器最终输出。

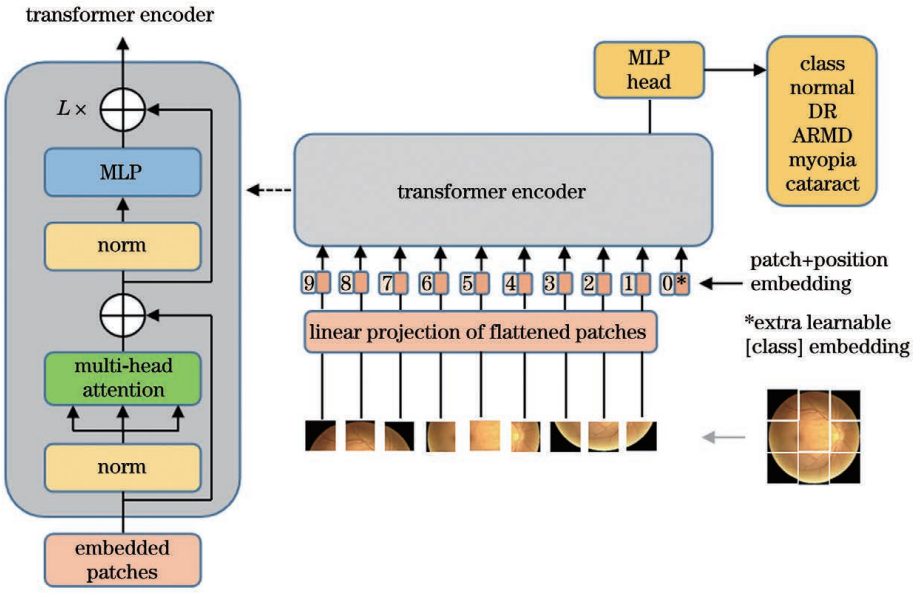


图 1　Vit 模型概述

Fig. 1　Overview of Vit model

## 2.2　SimAM 注意力模块

本研究使用一种无参数注意力模块 SimAM[17] 替换 Fused-MBConv 的 SE 模块，可以在不增加参数的同时提高神经网络的特征提取能力，MBConv 和 Fused-MBConv 的结构如图 2 所示，注意力机制模块的具体细节如图 3 所示。

现有的注意力模块集中在通道维度和空间维度[18]。然而，现实中这两种机制应该共同促进视觉过程中的信息选择。因此，本研究提出一种 3D 注意力模块用于为每个神经元分配一个唯一的权重。在视觉神经科学中，信息量最大的神经元通常表现出与周围神经元不同的放电模式，因此为了构建 3D 注意力模块，需要对每个神经元的重要性进行评估。此外，活跃的神经元也可能抑制周围的神经元活动。因此，在视觉处理中应该给予表现出明显空间抑制效果的神经元更高的重要性。通过测量神经元之间的线性可分离

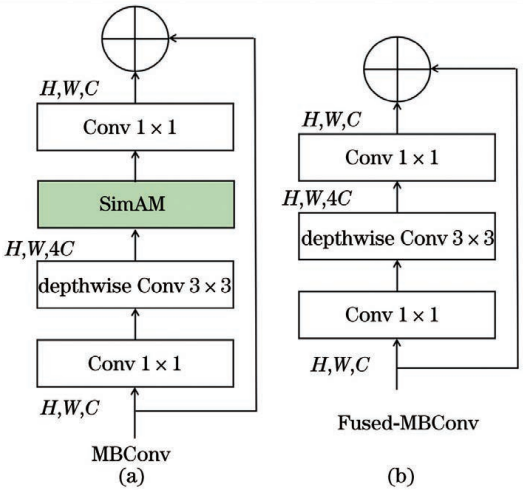


图 2　MBConv 和 Fused-MBConv 的结构。（a）MBConv；（b）Fused-MBConv

Fig. 2　Structure of MBConv and Fused-MBConv. (a) MBConv; (b) Fused-MBConv

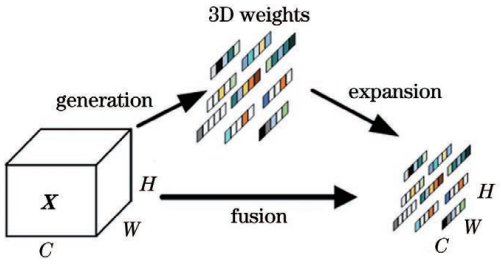


图 3 SimAM 3D 注意力权重

Fig. 3 SimAM 3D attention weights

性，便能计算出神经元的重要性。基于上述发现，本研究定义每个神经元能量函数 $e_t$ 为

$$e_t(w_t,b_t,\boldsymbol{y},x_j)=(y_t-\hat{t})^2+\frac{1}{M-1}\sum_{j=1}^{M-1}(y_o-\hat{x}_j)^2,\tag{8}$$

式中：$t$ 和 $x_j$ 分别为输入单通道上的目标神经元和其他神经元；$w_t$ 和 $b_t$ 分别为线性变换的权重和偏置；$j$ 为空间维度上的索引；$M$ 为该个通道上神经元的数量；$\boldsymbol{y}$ 包括激活值 $y_t$ 和抑制值 $y_o$；$\hat{t}=w_t t+b_t$ 和 $\hat{x}_j=w_t x_j+b_t$ 分别为 $t$ 和 $x_j$ 的线性变换。通过最小化式(8)即可找出目标神经元与其他神经元之间的线性可分行。为简单起见，对 $y_t$ 和 $y_o$ 应用二值标签 $y_t=1$ 和 $y_o=-1$ 并添加正则项 $\lambda w_t^2$，$\lambda$ 可依实验改动。最终的能量函数为

$$e_t(w_t,b_t,\boldsymbol{y},x_j)=\frac{1}{M-1}\sum_{j=1}^{M-1}[-1-(w_t x_j+b_t)]^2+[1-(w_t t+b_t)]^2+\lambda w_t^2。\tag{9}$$

通过迭代优化器来求解这些公式非常耗时，但式(9)有一个快速闭式解，可以由以下方式获得：

$$\begin{cases}w_t=-\dfrac{2(t-\mu_t)}{(t-\mu_t)^2+2\sigma_t^2+2\lambda}\\ b_t=-\dfrac{1}{2}(t+\mu_t)w_t\end{cases}。\tag{10}$$

计算除 $t$ 外其他所有神经元在通道维度上的均值 $\mu_t$ 和方差 $\sigma_t^2$，$\hat{\mu}$ 和 $\hat{\sigma}^2$ 为用来代替 $\mu_t$ 和 $\sigma_t^2$ 的所有神经元的均值和方差，这种方式可以显著降低计算成本，计算最小能量的方法为

$$e_t^*=\frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda}。\tag{11}$$

能量 $e_t^*$ 越低，神经元 $t$ 与周围神经元的差异越大，表示 $t$ 越重要。因此，每个神经元的重要性可以通过 $1/e_t^*$ 得到。至此，推导出一个能量函数并找到了计算神经元重要性的方法。此外，根据注意力调节机制，将缩放运算用于输入 $\boldsymbol{X}$ 增强特征，$\boldsymbol{E}$ 将所有 $e_t^*$ 跨通道和空间维度进行分组，即

$$\tilde{\boldsymbol{X}}=\mathrm{sigmoid}\left(\frac{1}{\boldsymbol{E}}\right)\odot\boldsymbol{X},\tag{12}$$

式中：运算符号⊙表示矩阵对应位置元素相乘。除了均值和方差是按通道(channel-wise)计算外，本研究模块中的所有计算都是按元素（pixel-wise）进行操作。

### 2.3 EfficientNet-Vit 集成模型

本研究采用集成模型的方法来提高分类准确率。模型之间的差异越大，集成后的模型越可能表现出更好的性能[19]。Vit 虽然既能挖掘长距离的依赖关系又能并行计算，但缺少像 CNN 固有的归纳偏差，比如平移不变性和局部相关性。而 EfficientNetV2 核心的卷积操作缺乏对图像本身的全局理解，无法构建特征之间的依赖关系，从而不能充分地利用上下文信息，卷积固定的权重，也不能动态地适应输入的变化[20]。

将 Vit 和 EfficientNetV2-S 两种模型融合在一起，通过两种不同的特征提取方式对眼底图像进行特征提取，可以更全面地区分图像之间的差异，获得更好的分类结果。本研究通过自适应加权算法，计算得到的最优加权因子分别为 0.6 和 0.4，并使用改进的软投票法[21]集成 EfficientNet-Vit 模型。

在分类过程中，各个类别的分类准确率不同，导致其置信度也各不相同。本研究在融合过程中采用自适应加权融合算法，其核心思想是：为了获取最优融合结果，基于所有分类的准确率自适应地查找总体方差最小情况下各分类模型所对应的最优加权因子[22]。

本研究有两种分类模型，则设两种分类模型在分类时的方差分别为 $\sigma_1^2$、$\sigma_2^2$，所要估计的真值为 $S$；各分类模型分类的准确率分别为 $X_1$、$X_2$，它们都是 $S$ 的无偏估计，且相互独立；各分类模型的加权因子分别为 $W_1$、$W_2$，则融合后的 $\hat{X}$ 值和各加权因子满足以下条件

$$\begin{cases}\hat{X}=W_1X_1+W_2X_2\\ W_1+W_2=1\end{cases},\tag{13}$$

则总体方差为

$$\sigma^2=E[W_1(S-X_1)^2+W_2(S-X_2)^2]=W_1^2\sigma_1^2+W_2^2\sigma_2^2。\tag{14}$$

由(14)式知，总体方差 $\sigma^2$ 是关于分类模型各加权因子 $W_1$、$W_2$ 的多元二次函数，一定存在最小值，其最小总体方差为

$$\sigma_{\min}^2=\frac{1}{\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}},\tag{15}$$

对应的最优加权因子为

$$W_1=\frac{1}{\sigma_1^2\left(\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}\right)},\tag{16}$$

$$W_2=\frac{1}{\sigma_2^2\left(\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}\right)}。\tag{17}$$

利用式(14)～(17)计算出各分类模型的方差和自

适应最优加权因子后，再对各分类模型的数据进行自适应加权融合，则融合后的计算值为

$$\hat{\bar{X}}=\sum_{p=1}^{2}W_p\bar{X}_p(k),\tag{18}$$

式中：$W_p$ 为第 $p$ 个模型对应的加权因子；$\bar{X}_p$ 为第 $p$ 个模型多次预测结果的均值；$k$ 为预测次数。

$$\begin{cases}\bar{\sigma}^2 & \sigma^2=\dfrac{1}{k}(W_1^2\sigma_1^2+W_2^2\sigma_2^2)\\ \sigma_{\min}^2=\dfrac{1}{k\left(\dfrac{1}{\sigma_1^2}+\dfrac{1}{\sigma_2^2}\right)}\end{cases},\tag{19}$$

可利用式(19)计算出融合后的总体方差和最小总体方差，评估融合后的准确性。

计算得到加权因子为 0.4 和 0.6，对每一个样本 $x_{mn}$ 在每个类别标签下进行二分类软投票。每个样本 $x_{mn}$ 在每个类别标签两个模型下的分类概率为

$$p_{mn}=(p_{0mn},p_{1mn},p_{2mn},p_{3mn},p_{4mn}),\tag{20}$$

式中：$p_{0mn}$ 表示第 $m$ 个样本在第 $n$ 个模型的第 0 个类别标签下被判定为正例类的概率，$p_{1mn}$，$p_{2mn}$，$p_{3mn}$，$p_{4mn}$ 以此类推。求出 $p_{mn}$ 后，Vit 模型的输出为该标签的概率乘以加权因子 $W_1$，EfficientNetV2-S 模型的输出结果乘以加权因子 $W_2$，两者相加后得到该样本的预测输出为此标签的概率，即

$$P_m=\sum_{n=1}^{2}p_{mn}\cdot W_n。\tag{21}$$

求出样本 $x_{mn}$ 在 5 个类别标签下的最终分类概率 $P_m$，输出该样本为 $P_m$ 下最高概率的标签。

## 3 实验结果与分析

本研究将 EfficientNet-Vit 集成模型应用于眼底图像的分类识别，将收集的数据集划分为训练集、验证集和测试集后，对训练集进行图像预处理和数据增强，在保持训练集、验证集和测试集不变的情况下，各进行 3 次实验，并将测试平均值作为最终结果。本实验还将所提模型与 EfficientNetV2-S、Vit 等其他分类模型进行比较，验证其对眼底图像自动分类的性能。

### 3.1 数据集和实验设置

使用所提分类模型对眼底图像进行分类实验，所使用的数据集为南开大学提供的 OIA-ODIR 数据集，其取样人群年龄涵盖全年龄段，该数据主要针对眼部疾病的诊断。剔除其中多标签的眼底图像后，再参考图片质量、图片数量等因素，将数据集整理为正常、糖尿病视网膜病变、老年黄斑变性、近视、白内障 5 类。其中：糖尿病视网膜病变是世界范围内工作组人群的首要致盲性眼病[23]；老年黄斑变性是老年人失明的主要原因之一；高度近视眼底病变是成人常见的致盲原因之一；白内障是全球第一致盲性眼病，老年性白内障在全世界致盲眼病中所占的比例最高。数据集的具体细节如表 2 所示。

表 2 眼底数据集

Table 2 Fundus dataset

| Degree of illness | Number of training images | Number of testing images | Total number of images |
|---|---|---|---|
| Normal | 2818 | 409 | 3227 |
| DR | 1389 | 203 | 1592 |
| ARMD | 147 | 49 | 196 |
| Myopia | 234 | 35 | 269 |
| Cataract | 262 | 43 | 305 |

本实验在具有 16 GB 显存的 Nvidia Tesla P100 GPU[24]服务器上进行训练，使用 PyTorch[25]作为深度学习开发框架，因为数据集样本比较少，所以训练集使用的数据预处理方式是随机裁剪图片大小为 384 pixel×384 pixel、随机水平翻转和自动数据增强 TrivialAugment[26]。使用混合精度进行训练，Batch_size 设置为 16。训练时的学习策略为预热和余弦退火学习率衰减[27]，训练的最大周期为 60，学习率预热可以使模型更容易进行训练，学习率预热的周期数为 10，最大的学习率设置为 $6\times10^{-4}$。模型训练使用的损失函数是标签平滑损失函数，使用 AdamW 优化器[28]进行参数更新，优化器的 L2 正则化参数 weight_decay 设置为 0.05，其他模型的对比实验都使用相同的训练参数。

### 3.2 性能指标

将测试集划分为 5 类，可以将它们定义为 $T=(T_1,T_2,T_3,T_4,T_5)$，在测试集中，得到预测类 $P=(P_1,P_2,P_3,P_4,P_5)$。为了与其他模型进行性能比较，本研究使用准确率($A$)、精确率($P$)和特异性($S$) 3 个指标作为评价标准。

首先，定义了真正例(TP)、假正例(FP)、真负例(TN)、假负例(FN)。TP 的定义为预测标签与目标标签匹配的图片数量，即 $N_{\mathrm{TP},i}=|P_i\cap T_i|$，$i=1,2,3,4,5$；FP 的定义为预测标签与实际目标标签不匹配的图片数量，即 $N_{\mathrm{FP},i}=|P_i\backslash T_i|$，$i=1,2,3,4,5$；FN 的定义为不属于目标标签但被错误预测的图片数量，即 $N_{\mathrm{FN},i}=|T_i\backslash P_i|$，$i=1,2,3,4,5$；TN 的定义为不属于目标标签且被预测正确的图片数量，不进行分类，即 $N_{\mathrm{TN},i}=\sum_{d\in C,d\neq c}|T_d|$。由此准确率、精确率和特异性的计算公式为

$$A=\frac{\sum_{i=1}^{5}N_{\mathrm{TP},i}}{\sum_{i=1}^{5}(N_{\mathrm{TP},i}+N_{\mathrm{FP},i}+N_{\mathrm{FN},i}+N_{\mathrm{TN},i})},\tag{22}$$

$$P=\frac{N_{\mathrm{TP},i}}{N_{\mathrm{TP},i}+N_{\mathrm{FP},i}},i=1,2,3,4,5,\tag{23}$$

$$S=\frac{N_{\mathrm{TN},i}}{N_{\mathrm{TN},i}+N_{\mathrm{FP},i}},i=1,2,3,4,5,\tag{24}$$

式中：$i$ 为样本标签类别。准确率、精确率越接近 1，模型越可靠。

### 3.3 分类结果和模型可视化解释

图 4 所示为训练好的模型在测试集上运行后绘制的混淆矩阵，通过混淆矩阵计算得到的 EfficientNetV2-S 模型、Vit 模型和 EfficientNet-Vit 集成模型对测试集的准确率、精确率和特异性如表 3 所示。从表 3 可以看到，EfficientNetV2-S 模型的准确率比 Vit 模型高 1.1%，EfficientNetV2-S 模型的精确率比 Vit 模型高 1.2%，说明预训练后的改进 EfficientNetV2-S 模型的性能要优于 Vit 模型。将它们集成到 EfficientNet-Vit 模型中，准确率达到 92.7%，精确率达到 88.3%，特异性达到了 98.1%，说明模型性能得到进一步提高。

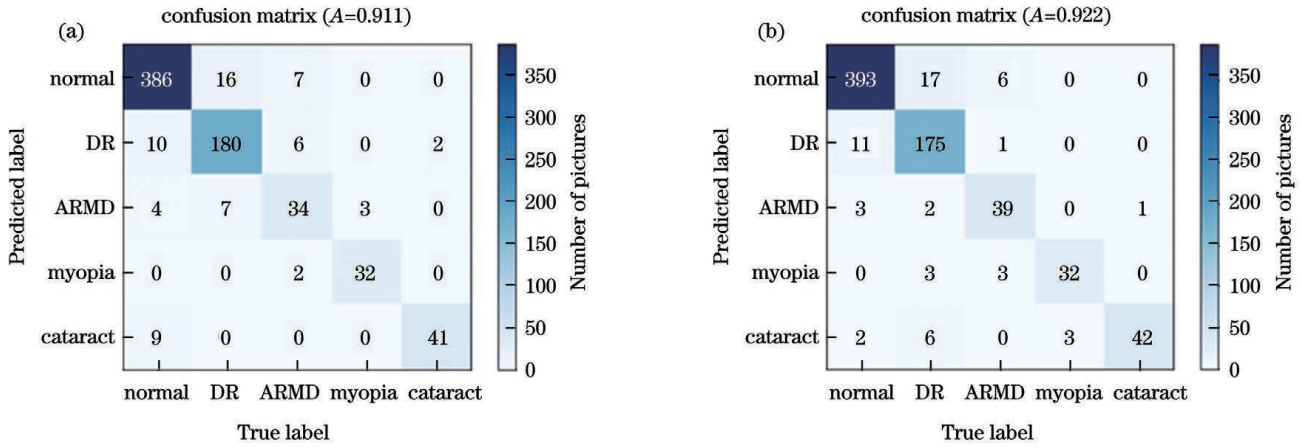


图 4 Vit、EfficientNetV2-S 模型的混淆矩阵。(a) Vit；(b) EfficientNetV2-S

Fig. 4 Confusion matrix of Vit and EfficientNetV2-S models. (a) Vit; (b) EfficientNetV2-S

表 3 Vit、EfficientNetV2-S、EfficientNet-Vit 模型的准确率、精确率和特异性

Table 3 Accuracy, precision, and specificity of Vit, EfficientNetV2-S, and EfficientNet-Vit models

| Model | Accuracy /% | Precision /% | Specificity /% | Training time /h |
|---|---|---|---|---|
| Vit | 91.1 | 86.4 | 97.2 | 11.0 |
| EfficientNetV2-S | 92.2 | 87.6 | 97.5 | 9.2 |
| EfficientNet-Vit | 92.7 | 88.3 | 98.1 | – |

为了证明该方法的有效性，将提出的 EfficientNet-Vit 模型与以下其他模型进行比较，结果如表 4 所示。与 Resnet50[29]、Densenet121[30]、ResNeSt-101[31]和 EfficientNet-B0 相比，EfficientNet-Vit 集成模型的准确率分别提高了 5.4%、3.2%、2.0%、1.4%；与 TNT-B[32] 相比，EfficientNet-Vit 集成模型的准确率提高了 1.6%。结果表明，EfficientNet-Vit集成模型有着更高的准确率，显示了其在眼底图像分类任务中的优越性。不同加权因子集成模型的准确率如表 5 所示。

表 4 不同模型的准确率指标对比

Table 4 Comparison of accuracy indexes of different models

| Model | Accuracy /% |
|---|---|
| Resnet50 | 87.3 |
| Densenet121 | 89.5 |
| ResNeSt-101 | 90.7 |
| EfficientNet-B0 | 91.3 |
| TNT-B | 91.1 |
| EfficientNet-Vit | 92.7 |

表 5 不同加权因子模型的准确率指标对比

Table 5 Comparison of accuracy indexes of models with different weighted factors

| Weighted factor | Accuracy /% |
|---|---|
| 0.3, 0.7 | 92.0 |
| 0.4, 0.6 | 92.7 |
| 0.5, 0.5 | 91.6 |

对于医学图像来说，可解释性非常重要，因此本研究采用加权梯度类激活热力图(Grad-CAM)方法对眼底图像进行特征重要性分析，模型可解释性可以辅助医生定位病变区域，做出可靠的判断。图 5 为 EfficientNetV2-S 模型的激活热力图，2 张眼底图片的原图均为异常眼底，图 5(a)所示为视盘异常，图 5(b)所示为黄斑区异常，每张图的左侧是原图，右侧是可解释性激活热力图，激活热力图的颜色越深，眼底图像的病变越严重，与专业医生的读片结果一致。

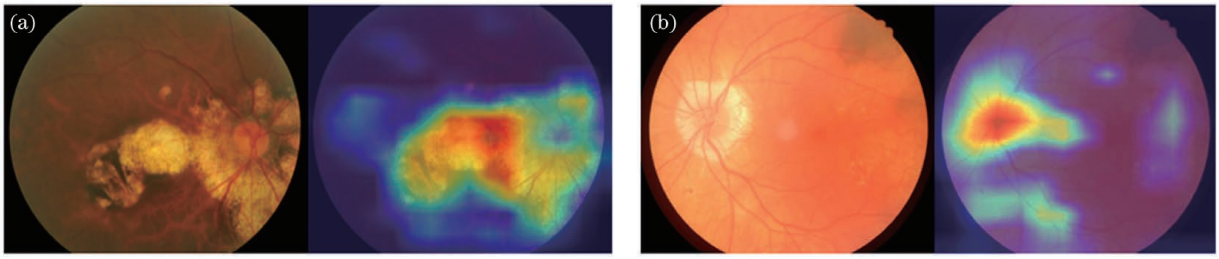


图 5 数据集中异常眼底图片的热力图。(a)视盘异常;(b)黄斑区异常

Fig. 5 Heatmap of abnormal fundus images in dataset. (a) Abnormalities in optic disc; (b) abnormalities in macular

## 4 结 论

所提出的 EfficientNet-Vit 集成模型是将 CNN 和 Transformer 结合在一起优化后的网络模型。CNN 的核心是卷积核,具有平移不变性和局部敏感性等归纳偏置特性,可以捕捉局部的时空信息,但缺乏对图像本身的全局理解。与 CNN 相比,Transformer 的自注意力机制不受局部相互作用的限制,既能挖掘长距离的依赖关系,又能并行计算。在本研究中,CNN 模型选用优化的 EfficientNetV2-S 模型,Transformer 模型选用 Vit 模型,通过自适应加权融合法计算得到最优加权因子,并利用加权软投票方法集成模型。EfficientNet-Vit 可以用两种完全不同的方式提取图像特征。实验证明,将两个模型集成后应用到眼底图像的分类中,集成后的模型精确率分别提高了 0.5% 和 1.6%,准确率分别提高了 0.7% 和 1.9%,特异性分别提高了 0.6% 和 0.9%。若将该模型应用到医学的辅助诊断过程中,可以提高眼科医生的工作效率,有效缓解就医等待时间长、偏远地区就医难等问题。未来投入更多的数据集对模型进行训练,也许可以进一步提高自动分类的准确度、精确度和灵敏度,在临床上获得更好的效果。

## 参 考 文 献


[1] Walter T, Massin P, Erginay A, et al. Automatic detection of microaneurysms in color fundus images[J]. Medical Image Analysis, 2007, 11(6): 555-566.

[2] 朱承璋, 向遥, 邹北骥, 等. 基于分类回归树和 AdaBoost 的眼底图像视网膜血管分割[J]. 计算机辅助设计与图形学学报, 2014, 26(3): 445-451.
Zhu C Z, Xiang Y, Zou B J, et al. Retinal vessel segmentation in fundus images using CART and AdaBoost[J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(3): 445-451.

[3] Anwar S M, Majid M, Qayyum A, et al. Medical image analysis using convolutional neural networks: a review[J]. Journal of Medical Systems, 2018, 42(11): 226.

[4] Wan C, Li H, Cao G F, et al. An artificial intelligent risk classification method of high myopia based on fundus images[J]. Journal of Clinical Medicine, 2021, 10(19): 4488.

[5] Imran A, Li J Q, Pei Y, et al. Fundus image-based cataract classification using a hybrid convolutional and recurrent neural network[J]. The Visual Computer, 2021, 37(8): 2407-2417.

[6] 连超铭, 钟舜聪, 张添福, 等. 光学相干断层扫描视网膜图像的迁移学习分类[J]. 激光与光电子学进展, 2021, 58(1): 0117002.
Lian C M, Zhong S C, Zhang T F, et al. Transfer learning-based classification of optical coherence tomography retinal images[J]. Laser & Optoelectronics Progress, 2021, 58(1): 0117002.

[7] 孙雨琛, 刘宇红, 张达峰, 等. 基于深度学习的糖尿病视网膜病变诊断方法[J]. 激光与光电子学进展, 2020, 57(24): 241701.
Sun Y C, Liu Y H, Zhang D F, et al. Diagnosis method of diabetic retinopathy based on deep learning[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241701.

[8] 郑雯, 沈琪浩, 任佳. 基于 Improved DR-Net 算法的糖尿病视网膜病变识别与分级[J]. 光学学报, 2021, 41(22): 2210002.
Zheng W, Shen Q H, Ren J. Recognition and classification of diabetic retinopathy based on Improved DR-Net algorithm[J]. Acta Optica Sinica, 2021, 41(22): 2210002.

[9] 任龙杰, 孙颖, 丁卫平, 等. 基于单种群蛙跳优化 CNN 的眼底图像多病变检测[J]. 计算机科学与探索, 2021, 15(9): 1762-1772.
Ren L J, Sun Y, Ding W P, et al. Multiple lesions detection of fundus images based on CNN algorithm optimized by single population frog-leaping algorithm[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(9): 1762-1772.

[10] 钟志权, 袁进, 唐晓颖. 基于卷积神经网络的左右眼识别[J]. 计算机研究与发展, 2018, 55(8): 1667-1673.
Zhong Z Q, Yuan J, Tang X Y. Left-vs-right eye discrimination based on convolutional neural network[J]. Journal of Computer Research and Development, 2018, 55(8): 1667-1673.

[11] Yu S, Ma K, Bi Q, et al. MIL-VT: multiple instance learning enhanced vision transformer for fundus image classification[M]// de Bruijne M, Cattin P C, Cotin S, et al. Medical image computing and computer assisted intervention-MICCAI 2021. Lecture notes in computer science. Cham: Springer, 2021, 129008: 45-54.

[12] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. (2019-05-28)[2021-04-05]. https://arxiv.org/abs/1905.11946.

[13] Tan M X, Le Q V. EfficientNetV2: smaller models and faster training[EB/OL]. (2021-04-01)[2021-05-04]. https://arxiv.org/abs/2104.00298.

[14] Li T, Gao Y Q, Wang K, et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening[J]. Information Sciences, 2019, 501: 511-522.

[15] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1): 1-17.
Lu H T, Zhang Q C. Applications of deep convolutional neural network in computer vision[J]. Journal of Data Acquisition & Processing, 2016, 31(1): 1-17.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2021-05-06]. https://arxiv.org/abs/2010.11929.

[17] Yang L, Zhang R Y, Li L, et al. Simam: a simple, parameter-free attention module for convolutional neural networks[C]// International Conference on Machine Learning, July 18-24, 2021, Virtual Event. Cambridge: PMLR, 2021, 139: 11863-11874.

[18] 徐爱生, 唐丽娟, 陈冠楠. 注意力残差网络的单图像去雨方法研究[J]. 小型微型计算机系统, 2020, 41(6): 1281-1285.
Xu A S, Tang L J, Chen G N. Single image rain removal method based on attention mechanism[J]. Journal of Chinese Computer Systems, 2020, 41(6): 1281-1285.
[19] Jiang Z C, Dong Z X, Wang L Y, et al. Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model[J]. Computational Intelligence and Neuroscience, 2021, 2021: 7529893.
[20] 刘文婷, 卢新明. 基于计算机视觉的 Transformer 研究进展[J]. 计算机工程与应用, 2022, 58(6): 1-16.
Liu W T, Lu X M. Research progress of transformer based on computer vision[J]. Computer Engineering and Applications, 2022, 58(6): 1-16.
[21] Zhou Z H. Ensemble methods: foundations and algorithms[M]. London: Taylor& Francis, 2012.
[22] Sung W T, Chen J H, Hsiao C L. Data fusion for PT100 temperature sensing system heating control model[J]. Measurement, 2014, 52: 94-101.
[23] 张炜. 眼底病变智能诊断研究[D]. 成都: 四川大学, 2021.
Zhang W. Research on intelligent diagnosis of fundus lesions [D]. Chengdu: Sichuan University, 2021.
[24] Lindholm E, Nickolls J, Oberman S, et al. NVIDIA tesla: a unified graphics and computing architecture[J]. IEEE Micro, 2008, 28(2): 39-55.
[25] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems 32, December 8-14, 2019, Vancouver, BC, Canada. [S.l.: s.n.], 2019.
[26] Müller S G, Hutter F. TrivialAugment: tuning-free yet state-of-the-art data augmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2021: 754-762.
[27] Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts[EB/OL]. (2016-08-13)[2021-04-05]. https://arxiv.org/abs/1608.03983.
[28] Loshchilov I, Hutter F. Decoupled weight decay regularization [EB/OL]. (2017-11-14)[2021-05-07]. https://arxiv.org/abs/1711.05101.
[29] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
[30] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.
[31] Zhang H, Wu C R, Zhang Z Y, et al. ResNeSt: split-attention networks[EB/OL]. (2020-04-19)[2021-04-08]. https://arxiv.org/abs/2004.08955.
[32] Han K, Xiao A, Wu E H, et al. Transformer in transformer [EB/OL]. (2021-02-27)[2022-01-05]. https://arxiv.org/abs/2103.00112.

# Fundus Image Classification Research Based on Ensemble Convolutional Neural Network and Vision Transformer

Yuan Yuan, Chen Minghui*, Ke Shuting, Wang Teng, He Longxi, Lü Linjie, Sun Hao, Liu Jiannan

*Shanghai Engineering Research Center of Interventional Medical, Ministry of Education of Medical Optical Engineering Center, School of Health Sciences and Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*

## Abstract

**Objective** With the increasing prevalence and blindness rate of fundus diseases, the lack of ophthalmologist resources is increasingly unable to meet the demand for medical examination. Given the shortage of ophthalmic medical staff, long waiting process for medical treatment, and challenges in remote areas, there is an irresistible trend to reduce the workload of medical staff via artificial intelligence. Several studies have applied convolutional neural network (CNN) in the classification task of fundus diseases; however with the advancement of Transformer model application, Vision Transformer (ViT) model has shown higher performance in the field of medical images. ViT models require pretraining on large datasets and are limited by the high cost of medical image acquisition. Thus, this study proposes an ensemble model. The ensemble model combines CNN (EfficientNetV2-S) and Transformer models (ViT). Compared with the existing advanced model, the proposed model can extract the features of fundus images in two completely different ways to achieve better classification results, which not only have high accuracy but also have precision and sensitivity. Specifically, it can be used to diagnose fundus diseases. This model can improve the work efficiency of the fundamental doctor if applied to the medical secondary diagnosis process, thus effectively alleviating the difficulties in diagnosis of fundus diseases caused by the shortage of ophthalmologist staff, long medical treatment process, and difficult medical treatment in remote areas.

**Methods** We propose the EfficientNet-ViT ensemble model for the classification of fundus images. This model integrates the CNN and Transformer models, which adopt the EfficientNetV2-S and ViT models, respectively. First, train the EfficientNetV2-S and ViT models. Then, apply adaptive weighting data fusion technology to accomplish the complementation of the function of the two types of models. The optimal weighting factors of the EfficientNetV2-S and ViT models are calculated using the adaptive weighting algorithm and then the new model (EfficientNet-ViT) is integrated with them. After calculating the weighting factors 0.4 and 0.6, multiply the output of the ViT model by a weighting

factor of 0.4, multiply the output of the EfficientNetV2-S model by a weighting factor of 0.6, and then weigh the two to obtain the final prediction result. According to clinical statistics, the current common fundamental disease in my country includes the following diseases: diabetic retinopathy (DR), age-related macular degeneration (ARMD), cataract, and myopia. These fundus diseases are the main factors that cause irreversible blindness in my country. Thus, we classify fundus images into the following five categories: normal, DR, ARMD, myopia, and cataract. Furthermore, we use three indicators, such as accuracy, precision, and specificity. The EfficientNet-ViT ensemble model can extract the features of fundus images in two completely different ways to achieve better classification results and higher accuracy. Finally, we compare the performance indicators of this model and other models. The superiority of the integrated model in the fundus classification is verified.

**Results and Discussions** The accuracy of EfficientNet-ViT ensemble model in fundus image classification reaches 92.7%, the precision is 88.3%, and the specificity reaches 98.1%. Compared with EfficientNetV2-S and ViT models, the precision of EfficientNet-ViT ensemble model improves by 0.5% and 1.6%, accuracy improves by 0.7% and 1.9%, and specificity increases by 0.6% and 0.9%, respectively (Table 3). Compared with Resnet50, Densenet121, ResNeSt-101, and EfficientNet-B0, the accuracy of the EfficientNet-ViT ensemble model increases by 5.4%, 3.2%, 2.0%, 1.4%, respectively (Table 4), showing its superiority in the fundus image classification task.

**Conclusions** The EfficientNet-ViT ensemble model proposed in this study is a network model combining a CNN and a transformer. The core of the CNN is the convolution kernel, which has inductive biases, such as translation invariance and local sensitivity, and can capture local spatio-temporal information but lacks a global understanding of the image itself. Compared with the CNN, the self-attention mechanism of the transformer is not limited by local interactions and can not only mine long-distance dependencies but also perform parallel computation. This study uses the EfficientNetV2-S and ViT models to calculate the most weighted factors for the CNN and Transformer models through the adaptive weighted fusion method. The EfficientNet-ViT can extract image features in two completely different ways. Our experimental results show that the accuracy and precision of fundus image classification can be improved by integrating the two models. If applied in the process of medical auxiliary diagnosis, this model can improve the work efficiency of fundus doctors and effectively alleviate the difficulties in diagnosis of fundus diseases caused by the shortage of ophthalmic medical staff, long waiting process for medical treatment, and difficult medical treatment in remote areas in China. When more datasets are used to train the model in the future, the accuracy, precision, and sensitivity of automatic classification may be further improved to achieve better clinical results.

**Key words** bio-optics; ophthalmology; fundus disease; image classification; ensemble model; weighted fusion