DOI:10.3969/j.issn.1005-202X.2025.09.008

医学影像物理

面向糖尿病视网膜病变分级的多层特征关注增强网络

梁冰雪[1],王文婧[1],王皓祺[2],关权[1],秦玉华[1]
1.青岛科技大学信息科学技术学院,山东 青岛 266061;2.华东理工大学商学院,上海 200237

【摘要】为进一步提高对糖尿病视网膜病变严重程度的诊断准确率,提出一种多层特征关注增强网络(MFAE-Net)。针对处理糖尿病视网膜病变图像时全局与局部特征的统一表达方面不足的问题,采用双分支并行的ResNet-50和DeiT-S模型作为骨干架构,并在网络末端位置设计特征融合模块。同时,设计多尺度位置感知增强模块,通过空洞卷积结合位置注意力机制提取多尺度信息,增强眼底图像中病变的特征表示;设计局部特征增强模块,强化对局部信息的提取能力,从而提高模型识别小病变和微小变化的能力。实验结果表明,本研究提出的MFAE-Net达到87.61%的准确率,表现出优异的分类效果,为进一步推动糖尿病视网膜病变检测技术的发展提供有力的支持。

【关键词】糖尿病视网膜病变;图像分类;特征融合;计算机辅助诊断
【中图分类号】R318;TP391.41 【文献标志码】A 【文章编号】1005-202X(2025)09-1174-10

Multi-layer feature attention enhanced network for diabetic retinopathy staging

LIANG Bingxue[1], WANG Wenjing[1], WANG Haoqi[2], GUAN Quan[1], QIN Yuhua[1]
1. School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China; 2. School of Business, East China University of Science and Technology, Shanghai 200237, China

Abstract: A multi-layer feature attention enhanced network is proposed to further improve the diagnostic accuracy of the severity of diabetic retinopathy. To address the inconsistent expression of global and local features when processing diabetic retinopathy images, a dual-branch parallel model combining ResNet-50 and DeiT-S is employed as the backbone architecture, and a feature fusion module is designed at the end of the network. Concurrently, a multi-scale location awareness enhancement module is developed to extract multi-scale information through dilated convolution with positional attention mechanism for enhancing the feature representation of lesions in fundus images, and a local feature enhancement module is constructed to strengthen the model's capability in extracting local information, thus improving model's capability to identify small lesions and minor changes. The experimental results show that the proposed multi-layer feature attention enhanced network achieves an accuracy of 87.61%, exhibiting excellent classification performance. This advancement provides a strong support for further development of diabetic retinopathy detection technology.

Keywords: diabetes retinopathy; image classification; feature fusion; computer-aided diagnosis

前 言

糖尿病视网膜病变(Diabetic Retinopathy, DR)是一种由于糖尿病患者长期血糖及血压偏高所导致的慢性眼部疾病,它是全球视力损失的主要诱因之一[1]。根据国际糖尿病联合会(IDF)的数据,全世界每10名成年人中就有1人患有DR[2]。在糖尿病患者中,超过1/3的人会患上DR,这也是20~65岁成人视力丧失的主要原因。DR的全球患病率正在上升,预计到2030年将有6.43亿人受到影响,到2045年这一数字将增加到7.83亿人[3]。在病理阶段,DR图像根据病变的数量和类型分为3种类型的非增殖性DR和1种类型的增殖性DR(Proliferative DR, PDR)。分级范围从无到最严重,依次为正常(No DR)、轻度DR(Mild DR)、中度DR(Moderate DR)、重度DR(Severe DR)和PDR[4]。非增殖性阶段是早期阶段,表现为视网膜血管的微小损伤,如微血管瘤、硬性渗出、视网膜出血等。这一阶段对患者的视觉影响较轻,且疾病的进展可以通过控制血糖和定期眼科检查进行管理。早期发现和干预非增殖性

【收稿日期】2025-03-11
【基金项目】青岛市科技惠民示范项目(23-2-8-smjk-20-nsh)
【作者简介】梁冰雪,硕士,研究方向:医学图像处理、智慧医疗,E-mail: lbx_1209@163.com
【通信作者】秦玉华,博士,教授,研究方向:人工智能及应用、数据挖掘与分析,E-mail: yqin@qust.edu.cn

病变至关重要，可有效阻止其发展到增殖性阶段，严重时会引发视网膜脱离和视力丧失。因此，早期发现并治疗非增殖性病变能够极大降低DR导致的不可逆性视力丧失风险[5]。然而，在对大规模糖尿病患者进行手动筛查时会面临实际挑战。相同等级的病变部分可能在大小、形状和数量上有所不同，不同等级之间的差异也可能很微小。此外，在某些情况下，病变部分的颜色可能仅与视网膜的一些解剖结构有微小差异。有时，甚至噪声粒子也可能在眼底图像中出现并被误认为是病变部分，这大大增加了筛查的难度。此外，依赖人工诊断需要眼科医生的专业知识，这不仅增加了医疗保健成本，还可能延迟诊断和治疗过程[6]。

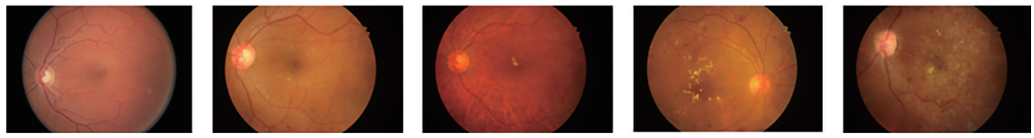
深度学习算法可以从大量数据中学习和提取特征，以实现复杂任务的自动化处理和决策，非常适合用于DR检测等图像的应用[7]。基于卷积神经网络（Convolutional Neural Network, CNN）的方法能够从眼底图像中高效地提取多层次特征，同时具备处理海量数据及对未知数据进行有效泛化的能力，已经成为自动DR检测和阶段分级的理想工具[8]。然而，CNN的感受野有限，故其无法有效捕获像素间的长距离依赖关系，无法充分理解DR图像的复杂结构。为了克服CNN的局限性，近期的研究开始采用基于Transformer的方法处理DR医学图像[9]。这类方法通过自注意力机制能够捕捉长距离依赖关系和全局上下文信息，同时摆脱固定感受野的限制，灵活适应不同分辨率和尺寸的图像，这对于在眼底图像分析中捕获病变信息至关重要[10]。但当前基于Transformer的方法需要在大规模数据集上训练才能发挥其作用，在增强DR图像特征的全局表示方面存在局限性。此外，纯粹的Transformer网络并不非常适合局部特征表示，而局部特征对于表达视网膜图像中的纹理和边缘细节至关重要。通过将CNN和Transformer充分结合，可以有效地综合两者的优势，提升糖尿病视网膜图像的分析性能。

在上述研究的启发下，本文提出一种基于CNN-Transformer的多特征关注增强网络模型（MFAE-Net），有效应用于DR图像分级任务。该模型包括两个分支：第一条是CNN分支，其中设计多尺度位置感知增强（Multi-scale Location Awareness Enhancement, MsLAE）模块，旨在扩展病变特征提取的感受野，更好地理解DR图像的上下文信息；第二条分支则是在Transformer的基础上，引入局部特征增强（Local Feature Enhancement, LFE）模块，以加强对局部信息的提取，提升模型在识别微小病变和细微变化方面的能力。此外，设计特征融合模块（Feature Fusion Module, FFM）来融合两个分支的特征，弥补CNN和Transformer各自只关注单一特征的不足。通过这一模块，可以有效结合CNN提取的局部特征与Transformer捕捉的全局信息，从而提升模型的综合表现和对复杂病变的识别能力。通过在DR数据集上的实验结果表明，本研究提出的MFAE-Net在糖尿病视网膜等级分类任务上能够显著提升分类精度，展现其优越性和竞争力。

1 相关工作

1.1 深度学习在DR分级的应用

近年来，在深度学习算法的支持下，医学图像分析取得重大进展[11]，图1展示了不同程度的DR严重性的眼底彩色图像。Qureshi等[12]提出一种用于自动DR阶段识别的主动深度学习（ADL）系统。ADL-CNN模型使用主动学习来学习视网膜样本的深度视觉特征以进行准确的DR分级，并生成用于预测和分割感兴趣区域的掩模。Wang等[13]提出一种名为DeepMT-DR的联合学习多级DR分级任务的方法，该方法主要关注使用低分辨率的眼底图像，能够同时处理图像超分辨率的低级任务、病变分割的中级任务和疾病严重程度分类的高级任务。赵爽等[14]提出一种基于特征融合网络的DR分类模型，该模型以EfficientNet-B0为基础结合不同空洞率的空洞卷积形成多尺度特征，引入MS-CAM模块融合高低层特征。实验证明，该模型的分类准确率能够达到85.25%。为了解决医学图像分析中标记数据不足的问题，Wang等[15]提出一种基于多通道半监督的生成对抗网络来应对这个问题。该模型从标记和未标记的数据中生成眼底下图像，并将其与分散的DR特征进行匹配。同样，Han等[16]提出一种类别加权网络（CWN）来实现模型层面的数据平衡，在CWN中通过计算类别梯度范数，减少实验开销，为权重设置提供参考。同时，提出使用关系加权标签代替one-hot标签来考察标签之间的距离关系。



a：正常　b：轻度DR　c：中度DR　d：重度DR　e：PDR

图1 不同程度的DR图像示例

Figure 1 Examples of different stages of diabetic retinopathy images

自从Dosovitskiy等[17]开创性地提出VisionTransformer(ViT)并在图像分类任务上取得显著成就后,众多研究开始积极探索将ViT技术引入医学领域的应用潜力。Wu等[18]率先使用ViT进行DR等级识别,证明纯注意力机制在DR分级的有效性,Transformer可以取代传统CNN进行DR分级。然而,由于数据不平衡的问题,需要额外采用图像处理技术进行预处理,以提升模型性能。Gu等[19]利用ViT和残差注意力模块对DR不同阶段进行分类,可以更好地关注图像的细粒度特征,捕捉不同类别的空间分布。Sun等[20]提出一种基于编码器-解码器架构的新型病变感知Transformer(LAT)联合进行DR分级和病变发现。尽管所提出的架构优于最先进的技术,但它有时会错过病变部分,特别是当发生不同病变特异性像素的聚类时。

1.2 基于CNN和Transformer结合的网络模型

在DR分类领域,传统CNN仍然是主流方法,Transformer模型在DR图像分析中的应用相对较新,其分类效果仍有待进一步改进。随着技术的发展,将Transformer与CNN等结构相结合,并进行多模态信息融合,有望在DR分类领域取得更多突破。一种方法是使用CNN模型来提取视觉特征,然后将其馈送到Transformer层进行进一步优化。Pham等[21]提出一种简单而有效的UNet-Transformer模型(seUNet-Trans),用于医学图像分割。seUNet-Trans使用UNet作为特征提取器,随后将特征图通过桥阶层送入Transformer。Jiang等[22]提出一种名为RGTransformer的小样本分类算法,使用CNN进行初步特征提取后连接RGTransformer,学习图像数据的空间上下文感知区域表示。另一种方法是使用多分支融合,将CNN和Transformer的特征图进行融合。该模型不仅保留本地和全局上下文信息的保存,而且还能够捕获输入元素之间的远程关系。庄建军等[23]提出一种多层级深度特征融合的乳腺癌病理图像分类方法,该方法以双分支并行的DeiT-B和ResNet-18模型作为骨干架构,在双分支网络中间层和末端位置分别引入特征融合操作,有效加强乳腺癌病理图像全局与局部深度特征的联合学习。Zang等[24]提出一种Transformer、引导的类别-关系注意网络(CRA-Net),利用Transformer框架进一步处理CNN提取的特征图,捕获病变的长距离特征依赖性和空间相关性,从而获得上级DR分级性能。

2 多层特征关注增强网络

本文提出的模型旨在结合CNN和Transformer来学习更有效的医学图像表示,从而实现DR的精确分类。网络整体架构如图2所示,它由用于学习局部信息的CNN分支、捕捉全局信息的Transformer分支以及FFM组成。首先,将DR图像分别送入CNN分支和Transformer分支,以获得相应的特征图。CNN分支采用以ResNet-50为基础的卷积神经网络,利用残差网络捕获眼底图像的局部特征。在此过程中,设计MsLAE模块,以提升网络对图像的感知能力,并减少对背景干扰的敏感度。Transformer分支则设计LFE模块,以进一步强化对局部信息的提取能力。随后,两个分支的输出特征图将通过FFM,利用多重注意力机制融合局部和全部信息。最后,通过全局平均池化和全连接层生成DR严重性等级。

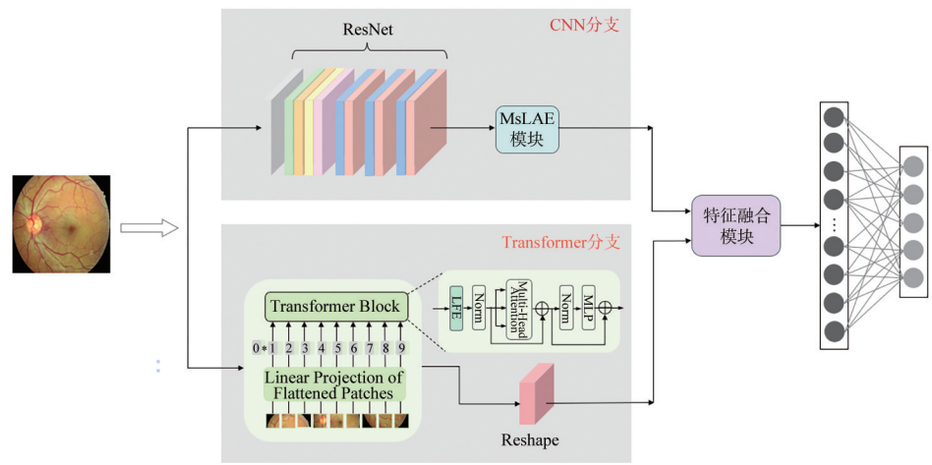


图2 多层特征关注增强网络

Figure 2 Multi-layer feature attention enhanced network

2.1 CNN分支

CNN分支采用ResNet-50网络来提取糖尿病视网膜图像的深层特征。由于眼底图像通常包含多种复杂的细微结构,如血管、黄斑和视盘,这些特征的

差异和复杂性使得特征提取具有挑战性。为了学习更具泛化能力的特征表示，本文移除了ResNet-50网络的最后一层，并添加MsLAE模块。MsLAE模块通过不同膨胀率的空洞卷积层和位置注意力的恰当组合，实现对眼底图像的多尺度特征提取。MsLAE结构图如图3所示。

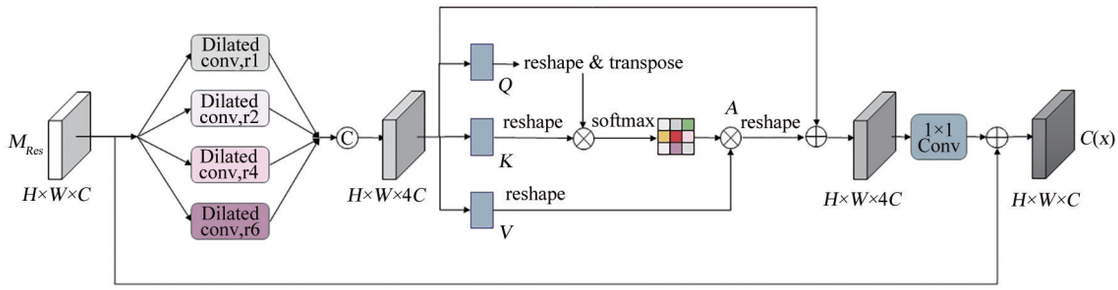


图3 MsLAE结构

Figure 3 MsLAE structure

首先，使用ResNet-50网络得到特征$\boldsymbol{M}_{\text{Res}}\in R^{H\times W\times C}$。为了扩大感受野，采用空洞卷积作为关键手段，以提高对不同尺度下眼底图像的特征提取能力。空洞卷积通过调整膨胀率，控制卷积核在输入特征图上的间隔，从而改变卷积核的感受野大小。空洞卷积的感受野大小可以计算为：

$$R_i=(k_i-1)\times(d_i-1)+k_i \tag{1}$$

其中，k_i是卷积核的大小，d_i是膨胀率。将$\boldsymbol{M}_{\text{Res}}$分别输入到4个并行的空洞卷积，空洞率设置为1、2、4和6，以提取图像的多尺度特征。公式如下：

$$X^i=f_{\text{dil}}^i(\boldsymbol{M}_{\text{Res}}),i=1,2,4,6 \tag{2}$$

其中，$f_{\text{dil}}^i(\cdot)$为空洞率为i的空洞卷积，X^i是空洞率为i的卷积提取到的感受野特征。

通过插值上采样技术，将获取到的多感受野特征调整至统一的尺寸，随后执行通道拼接操作，最终生成一个维度为$\boldsymbol{M}_{\text{dil}}\in R^{H\times W\times 4C}$的特征图。这种处理方式旨在增加网络对于不同尺度信息的感知能力，从而更好地理解图像中不同大小的目标或结构。描述如下：

$$\boldsymbol{M}_{\text{dil}}=\text{Concat}[X^1,X^2,X^4,X^6] \tag{3}$$

在糖尿病眼底图像中，病变通常表现为微小且不易察觉的病变特征，这些细微的病变常常被复杂的背景结构所掩盖。因此，需要加大对病变区域的关注，以确保这些关键特征能够被准确识别和提取。为此，将$\boldsymbol{M}_{\text{dil}}$输入到位置注意力机制（Position Attention Module, PAM）中[25]。该模块能够根据病变的位置自适应地调整网络中的权重，减少噪声干扰，并捕捉细粒度特征。整个过程描述如下：

$$\boldsymbol{Q},K,V=\boldsymbol{f}_c^{1,1}(\boldsymbol{M}_{\text{dil}})\in R^{N\times 4C} \tag{4}$$

$$A=\text{softmax}(K\boldsymbol{Q}^{\text{T}})\in R^{N\times N} \tag{5}$$

$$\boldsymbol{M}_{\text{PA}}=(\boldsymbol{M}_{\text{dil}}+A^{\text{T}}V)\in R^{H\times W\times 4C} \tag{6}$$

其中，$\boldsymbol{f}_c^{1,1}(\cdot)$表示1×1的普通卷积层，$\boldsymbol{Q}^{\text{T}}$为$\boldsymbol{Q}$的转置。

经过位置注意力机制之后，将$\boldsymbol{M}_{\text{PA}}$经过1×1卷积层处理，以降低通道维度，从而减少模型的参数数量和计算量。然后，将降维后的特征图与$\boldsymbol{M}_{\text{Res}}$相加，得到输出特征$\boldsymbol{C}(x)$，公式如下：

$$\boldsymbol{C}(x)=\boldsymbol{M}_{\text{Res}}+f_c^{1,1}(\boldsymbol{M}_{\text{PA}}) \tag{7}$$

MsLAE首先采用不同尺寸的空洞卷积来获得多尺度的特征表示，增加网络对图像的感知能力，从而捕获更广泛的上下文信息。接着，引入位置注意力机制，帮助网络将注意力集中在眼底图像中最相关和关键的病变区域，同时抑制不相关的噪音。通过在ResNet-50网络中添加MsLAE模块，模型在不同层次上都具有较强的感知能力和区分能力，这使得模型能够更好地捕捉图像中的细微特征和结构信息，从而提高病变分级的准确性。

2.2 Transformer分支

CNN模型在处理长距离依赖时可能存在困难，因为信息需要在卷积层之间逐步传递。为此，模型中添加Transformer分支，用于学习糖尿病视网膜图像中的全局信息和位置信息，从而帮助模型更好地理解图像中的病变特征。同时，在Transformer编码器中添加LFE模块，以提高对局部特征的感知能力，并进一步提升Transformer的建模能力。给定输入的眼底图像$I\in R^{H\times W\times C}$，首先将其切分成$N$个相同大小且不重叠的补丁图像$I\in R^{N\times(p^2\cdot C)}$，对每个补丁图像进行线性变换，将其映射到一个较高维度的向量空间中。为了保留位置信息，还添加位置编码，以形成最终的输入。然后，将嵌入的补丁和位置编码输入到Transformer编码器中。

LFE模块的结构如图4所示。给定输入特征序列$\boldsymbol{I}_t=\left[\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_{N-1},\boldsymbol{I}_N\right]$，将其重塑为2维的特征图$\boldsymbol{S}$。紧接着，将特征图$\boldsymbol{S}$送入深度可分离卷积层（Depthwise Separable Convolution, DSC），并进行归一化操作，得到新的特征图$\boldsymbol{S}'$。深度可分离卷积层在处理特征图时，会先针对每个通道独立进行卷积运算，随后再将各通道的输出结果进行合并。这有助于减少参数量并提高计算效率。随后，将原始特征图$\boldsymbol{S}$与处理后的特征图$\boldsymbol{S}'$相加，形成一个残差结构块。这个残差结构块通过引入跳跃连接，有助于网络学习残差并加速训练收敛。最后，将生成的特征图重新映射为特征序列$\boldsymbol{I}_t'$。整个过程可以描述为：

$$\boldsymbol{S}=\text{Reshape}(\boldsymbol{I}_t) \tag{8}$$

$$\boldsymbol{S}'=\boldsymbol{f}_{\text{norm}}(\boldsymbol{f}_{\text{DSC}}^{3,3}(\boldsymbol{S}))+\boldsymbol{S} \tag{9}$$

$$\boldsymbol{I}_t'=\text{Flatten}(\boldsymbol{S}') \tag{10}$$

其中，$\boldsymbol{f}_{\text{DSC}}^{3,3}(\cdot)$为3×3的深度可分离卷积层，$\boldsymbol{f}_{\text{norm}}(\cdot)$为归一化操作，Flatten$(\cdot)$为特征映射函数。LFE模块通过引入包括深度可分离卷积层的残差结构块，能够帮助网络学习到相邻图像块之间的空间信息，使Transformer分支能够更好地捕捉和学习局部特征，从而提升对图像细节和微小病变的识别能力。

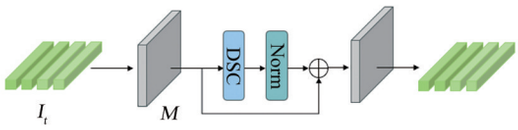


图4 LFE模块

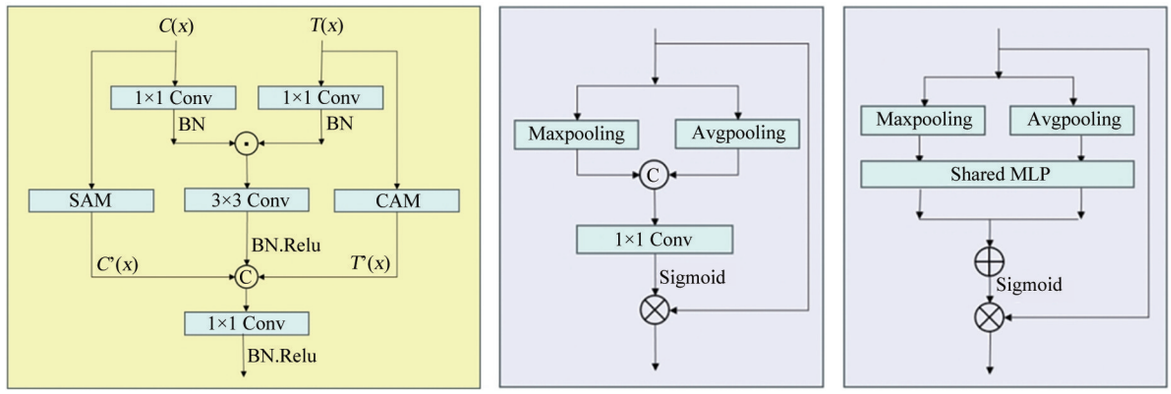
Figure 4 LFE module

Transformer编码器的输出需要与CNN分支的输出具有相同的空间分辨率。为了实现这一点，首先对Transformer分支生成的特征图执行上采样操作，以确保其空间尺度与CNN分支的特征图保持一致。接着，使用一个1×1卷积层对通道数进行调整，并添加BatchNorm进行归一化处理，生成对齐后的特征图$\boldsymbol{T}(x)$。在两条分支的特征图尺寸对齐后，通过FFM模块将局部特征与全局特征进行融合。最终，将融合后的特征图送入到分类器，以进行DR等级分类。

2.3 FFM模块

为了有效结合CNN和Transformer分支的特征向量，本文在网络的最后设计特征融合模块，其结构如图5a所示。首先，将双分支的$\boldsymbol{C}(x)$和$\boldsymbol{T}(x)$分别送入空间注意力（Channel Attention Module, CAM）[26]和通道注意力（Spatial Attention Module, SAM）[26]，以进一步加强特征信息，得到增强后的特征$\boldsymbol{C}'(x)$和$\boldsymbol{T}'(x)$。接着，对$\boldsymbol{C}(x)$和$\boldsymbol{T}(x)$进行3×3卷积和BatchNorm处理，然后通过哈达玛积对位置特征进行加权。这种处理方法旨在保持位置特征的细节，同时有效融合两组特征，避免细节信息的丢失。最后，将加权后的特征图沿通道维度与$\boldsymbol{C}'(x)$、$\boldsymbol{T}'(x)$拼接，再使用1×1卷积进行特征降维，实现局部信息和全局信息的有效融合。这种方法弥补CNN和Transformer各自只能关注单一特征的不足，增强模型在捕捉细粒度特征和全局语义信息方面的能力。

由于CNN分支在不断下采样的过程中会造成信息丢失，因此可以利用空间注意力机制增强局部信息。首先，将$\boldsymbol{C}(x)$在通道维度上执行全局最大池化和全局平均池化，并将两者的结果相结合。接着，利



a：特征融合模块（FFM） b：SAM c：CAM

图5 特征融合模块及其内部结构

Figure 5 Feature fusion module and its internal structure

用一个7×7卷积以及Sigmoid激活函数处理这个组合结果，从而生成一个空间注意力权重矩阵。最终，将这个权重矩阵与原始输入特征$\boldsymbol{C}(x)$相乘，得到增强后的特征$\boldsymbol{C}'(x)$。过程描述为：

$$\boldsymbol{C}'(x)=\sigma(\boldsymbol{f}_c^{7,7}([\mathrm{AvgPool}(\boldsymbol{C}(x));\mathrm{MaxPool}(\boldsymbol{C}(x))])) \quad (11)$$

其中，$\boldsymbol{f}_c^{7,7}(\cdot)$为7×7的普通卷积层，$\sigma(\cdot)$为Sigmoid激活函数，AvgPool为全局最大池化，MaxPool为全局平均池化。Transformer的自注意力机制主要关注特征在空间维度上的关系，可能忽略了特征图各通道之间的重要性。为了弥补这一不足，可以引入通道注意力机制，通过强调不同通道的重要性，提升特征表达的准确性和全面性。首先，将$\boldsymbol{T}(x)$沿在空间维度上分别进行全局最大池化和全局平均池化，随后将这两个池化结果输入到一个共享的多层感知机（MLP）中进行学习。接着，将MLP的输出结果相加，并通过Sigmoid激活函数处理，以生成通道注意力权重矩阵。最后，将该权重矩阵与原始输入特征$\boldsymbol{T}(x)$相乘，得到增强后的$\boldsymbol{T}'(x)$。过程可以描述为：

$$\boldsymbol{T}'(x)=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{T}(x)))+\mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{T}(x)))) \quad (12)$$

3 实验与结果分析

3.1 实验数据

本文采用IDRiD[27]、DDR[28]两个公共数据集与某医院的临床数据组成最终的数据集来评估模型的有效性。IDRiD数据集是根据印度一家眼科诊所的真实临床检查结果创建的，有413张训练图像和103张测试图像。该数据集根据国际临床DR分级标准分成5个等级。DDR数据集包含从中国147家医院收集的13 673张彩色眼底图像。这些图像由9 598名患者提供。该数据集在分级标准的基础上添加不可分级，模糊程度超过70%且没有清晰可见病变的图像会被归入此类别。经医院采集的临床数据共有2 100张彩色眼底图像，均由专业眼底照相机拍摄，图像为JPG格式，分辨率为3 072×2 048。图像均由眼科专家按照国际临床DR分级标准标注。

表1列出原数据集中不同类别的数量，本文在实验中不使用DDR数据集中不可分级的图像，最终数据集为15 138张图像。为保证实验的公平性和科学性，数据集按照8:1:1的比例将数据集随机划分为训练集、验证集和测试集。最终，训练集含有12 110张图像，验证集含有1 514张图像，测试集含有1 514张图像。

表1 不同类别的数量分布

Table 1 Distribution of quantities across categories

类别	IRDIR	DDR	医院数据
正常	168	6 266	1 468
轻度DR	25	630	193
中度DR	168	4 477	326
重度DR	93	236	72
增值性DR	62	913	41
不可分级	-	1 151	-
总计	516	13 673	2 100

预处理的过程如图6所示。(1)裁剪和调整尺寸：由于眼底图像中存在大量无信息的黑色区域，将这些区域裁剪掉有利于减少数据冗余。然后，将所有图像的尺寸统一调整为1 024×1 024。(2)绿色通道处理：眼底图像由红、绿、蓝3通道组成，其中绿色通道的病变特征更加明显，如出血点、微动脉瘤和渗出物，这对于检测DR的等级至关重要。因此，将绿色通道分离出来，并对其执行最小-最大归一化处理，将像素值映射到0~255之间。(3)对比度增强：归一化之后，利用对比度受限自适应直方图均衡化增强图像的颜色和亮度。(4)去噪处理：使用中值滤波器进行去噪处理，通过将每个像素值替换为其相邻像素值的中值来保留图像的锐度。最后，将处理后的绿色通道重新合并回原始图像。

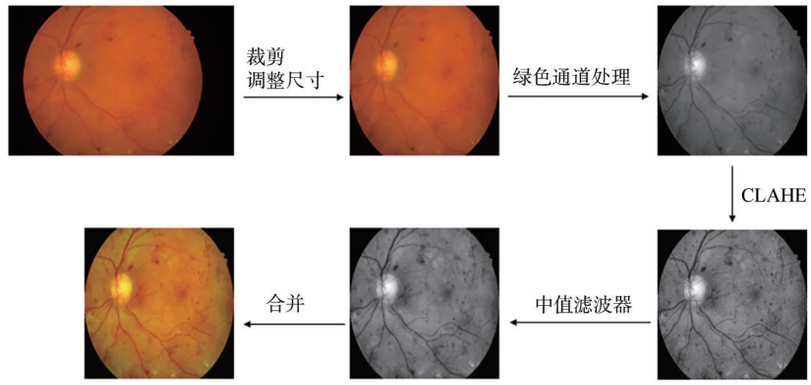


图6 预处理过程

Figure 6 Pre-processing

3.2 实验环境与设置

所有实验均在NVIDIA GeForce RTX2080Ti GPU上的Pytorch环境中运行。在训练阶段，epoch设置为100，batchsize设置为8，学习率为0.0001，使用Adam优化器进行优化。由于数据集中存在明显的样本不均衡问题，采用加权交叉熵损失函数，通过为每个类别分配不同的权重来平衡样本的不均衡情况，如式(13)所示：

$$L_{\mathrm{WCE}} = -\frac{1}{N}\sum_{n=1}^{N}\left(wy_i\log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right) \quad (13)$$

3.3 评价指标

本文选择准确率(Accuracy)、精确率(Precision)、召回率(Recall)作为模型的评价指标。准确率是指正确预测的样本数量占总样本数量的比例；精确率则反映了在预测为正类的样本中，真正为正类的样本所占的比例；召回率衡量了所有实际为正类的样本中，被正确预测为正类的样本所占的比例。以下是指标的计算方法：

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (14)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (15)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (16)$$

其中，TP、TN、FP、FN分别表示正阳性、正阴性、假阳性、假阴性。

3.4 性能分析

图7描绘了训练过程中，训练集与验证集上损失值的变化趋势。训练集的损失值迅速降低并维持在较低水平，而验证集的损失值则整体呈现递减趋势，最终趋于平稳状态。尽管验证集损失在初期波动较大，但随后逐渐下降并趋于稳定，这表明模型在学习过程中逐步提高了泛化能力。总体而言，模型表现出良好的学习和泛化效果。

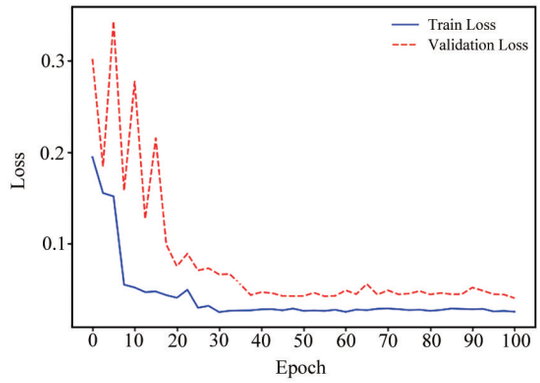


图7 损失曲线

Figure 7 Loss curve

图8显示了测试集的混淆矩阵，其中纵轴上的标签代表真实类别，横轴上的标签代表预测结果。本文模型在区分正常眼底图像和病变的眼底图像方面表现良好。然而，由于病变在早期眼底图像中的症状通常十分轻微，一些正常图像可能会被误分类为轻度DR和中度DR。此外，由于病变眼底图像在相近级别之间的特征差异不大，不同级别的病变都可能发生不同程度的误分类现象。此外，还生成了按类分类的报告，描述本文模型在测试数据集上的按类性能。本文模型在正常图像上表现最好，精确率为99.86%，召回率为87.72%。而在轻度DR图像上表现最差，精确率为33.45%，召回率为87.06%。表2展示本文提出的模型按类性能评估结果。

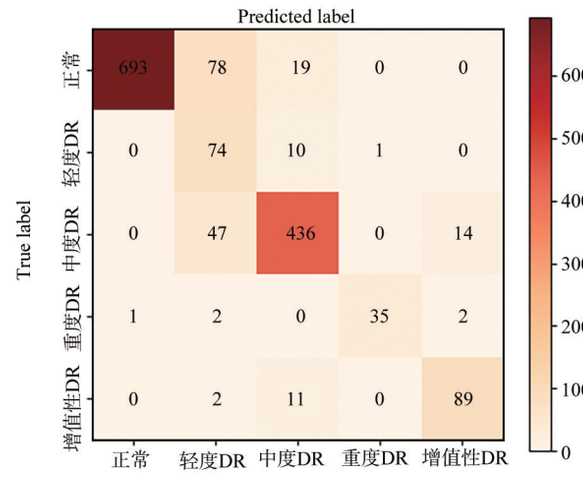


图8 混淆矩阵

Figure 8 Confusion matrix

表2 按类性能评估结果(%)

Table 2 Performance evaluation by category (%)

类别	精确率	召回率
正常	99.86	87.72
轻度DR	33.45	87.06
中度DR	91.60	89.53
重度DR	97.22	87.50
增值性DR	84.76	87.25

3.5 对比实验

为验证本文提出模型的有效性，分别与传统的CNN、Transformer，以及基于CNN和Transformer的混合模型在眼底图像数据集上进行对比实验。表3展示MFAE-Net和其他方法的性能数值比较。结果显示，MFAE-Net在分类精度上取得显著的性能，准确率、精确率和召回率分别达到87.61%、81.58%和87.81%。

表3 与现有方法的比较

Table 3 Comparison with the existing methods

模型	方法	准确率/%	精确率/%	召回率/%	Params/M
基于CNN的分类模型	VGG16[29]	80.75	73.22	81.35	138.3
	AlexNet[30]	81.81	75.46	82.23	61.1
	ResNet50[31]	83.20	77.71	84.85	25.6
	EfficientNet-B0[32]	84.44	77.13	83.83	5.3
基于Transformer的分类模型	ViT-B[17]	83.64	78.61	82.24	86.0
	Swin-T[33]	84.21	78.93	83.98	28.3
基于CNN和Transformer的分类模型	CMT-S[34]	85.38	80.82	85.96	25.3
	Conformer[35]	84.89	78.99	85.23	43.1
	MFAE-Net	87.61	81.58	87.81	83.3

与传统的基于CNN的方法相比，MFAE-Net在各项性能指标上均有明显提升。例如，与VGG16相比，精确率提升了8.36%，充分展现本文方法在特征提取能力上的优势。相较于CMT-S和Conformer等结合了CNN与Transformer的混合模型，虽然性能提升幅度较小，但MFAE-Net在捕捉病变区域的细微特征上表现出更高的潜力，进一步验证多层特征关注的有效性。尽管MFAE-Net的参数量相较于部分对比模型有所增加，但得益于CNN局部特征提取能力与Transformer全局特征建模能力的深度融合，该模型能够更加精确地识别和定位DR中的微小病变与细微变化。这种性能与参数量的权衡在医疗场景中尤为重要，因为医疗任务通常对模型的准确性和鲁棒性要求更高。整体来看，MFAE-Net在性能上的显著提升使其成为一个高效、可靠的DR分类模型，为疾病的早期诊断提供强有力的支持。

3.6 消融实验

为评估不同模块对整体模型性能的贡献，本文设计消融实验，以更好地理解各模块的运作原理，表4提供不同模块的消融实验结果。从表中可知，FFM显著改善了准确率、精确率和召回率。这是因为FFM在CNN和Transformer的特性之间架起桥梁，有效地融合双分支的特征信息，实现互补，从而提升模型的整体性能。一方面，CNN擅长捕捉细粒度局部特征，而FFM的融合机制确保这些局部信息能在全局语义上下文中得以保留和强化。另一方面，Transformer的全局建模能力通过FFM的引导，能够有效地与细粒度局部信息协同作用，从而增强对复杂病变的理解。MsLAE模块有助于丰富对图像多尺度信息的学习，从而获得更准确的病变定位。去除该模块后，准确率、精确率和召回率分别降低了2.52%、0.57%和3.58%。LFE模块则有助于Transformer在捕获远距离关系时关注局部信息，弥补模型在这方面的局限性，进一步提高模型对微小病变的敏感性，最终实现最佳性能。去除MsLAE和LFE模块后，指标降低明显，分别降低6.52%、8.83%和11.17%。这表明局部和全局特征的结合对于复杂病变特征的识别尤为重要。局部与全局特征的深度整合使得MFAE-Net在不同等级的眼底图像中均能表现优异，特别是在复杂病变区域和多变的图像环境下，体现出更强的鲁棒性。

表4 消融实验结果(%)

Table 4 Ablation study results (%)

方法	准确率	精确率	召回率
MFAE-Net	87.61	81.58	87.81
(w/o)MsLAE	85.09	81.01	84.23
(w/o)LFE	84.68	76.87	84.37
(w/o)FFM	80.45	73.34	76.06
(w/o)MsLAE & LFE	81.09	72.75	76.64

3.7 不同注意力机制对模型性能的影响

选取PAM、CAM和SAM进行实验验证，以研究不同注意力机制对MFAE-Net模型精度的影响。如表5所示，PAM表现最佳，能够处理眼底图像中的复杂背景和细微特征。相比之下，虽然CAM和SAM在局部特征的捕捉上有所增强，但在全局信息的整合和多尺度特征的融合方面，它们的能力不及PAM。CAM主要侧重于通道间的关系，而SAM则注重空间位置间的关系，这可能会导致模型在某些情况下对细节或全局特征的关注不够全面。这种局限性导致它们在整体性能上略有不足，无法达到PAM的效果。

表5 不同注意力机制的MFAE-Net精度(%)

Table 5 MFAE-Net accuracies for different attention mechanisms (%)

方法	准确率	精确率	召回率
PAM	87.61	81.58	87.81
CAM	85.86	78.63	86.98
SAM	86.01	80.98	84.32

3.8 卷积核的尺寸对模型性能的影响

为研究不同尺寸卷积核对MFAE-Net精度的影响,本文对LFE模块分别选取卷积核尺寸大小为3×3、5×5和7×7的深度可分离卷积层进行实验验证。如图9所示,随着卷积核尺寸的增大,MFAE-Net的精度呈现下降的趋势。较大的卷积核覆盖了更广的区域,这导致它们捕捉了更多不相关或噪声信息,从而干扰有用信号的提取。这不仅增加模型的计算复杂度,还可能导致模型过拟合于训练数据的背景细节,忽略关键的细微特征。因此,尽管较大的卷积核在理论上可以提供更丰富的上下文信息,但在实际应用中,它们反而可能会削弱模型的精度和识别能力。

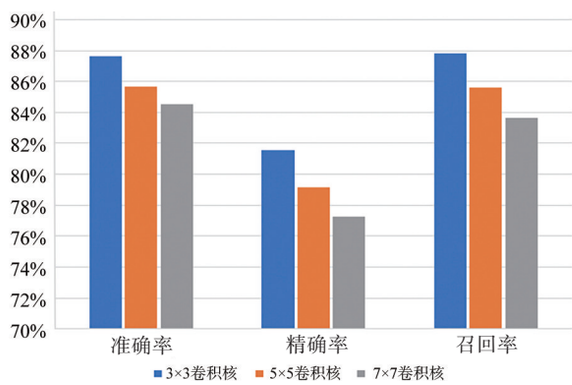


图9 不同卷积核尺寸的MFAE-Net精度

Figure 9 MFAE-Net accuracies for different convolution kernel sizes

4 结 论

本文提出一种DR分级的多层特征关注增强网络(MFAE-Net)。通过构建双分支特征提取网络,分别使用CNN和Transformer结构对DR图像进行特征提取,有效地增强眼底图像局部与全局深度特征的联合学习。在CNN分支中,加入多尺度位置感知增强模块,通过空洞卷积提取多尺度信息,并结合位置注意力机制来增强病变的特征表示。在Transformer分支中,设计局部特征增强模块,以提升对局部细节的捕捉和表达能力。最后,设计FFM模块来整合双分支的特征,以此充分捕捉具有辨识度的病变特征,进而提升分类的准确性。相较于现有的技术,本文提出的方法展现出一定的优势,并证明其高度的可行性和实效性,为临床医生做出准确的DR评估和治疗方案提供可靠依据。然而,由于使用Transformer,本文方法的模型参数量和计算复杂度较大,使得模型训练速度与传统卷积神经网络相比较慢。未来的工作将致力于探索CNN与Transformer的有效融合方式,设计更为精简高效的模型,并将其应用于更广泛的医疗图像任务中。

【参考文献】

[1] Madarapu S, Ari S, Mahapatra KK. A deep integrative approach for diabetic retinopathy classification with synergistic channel-spatial and self-attention mechanism[J]. Expert Syst Appl, 2024, 249(Part A): 123523.

[2] Sun H, Saeedi P, Karuranga S, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045[J]. Diabetes Res Clin Pract, 2022, 183: 109119.

[3] Bhati A, Gour N, Khanna P, et al. An interpretable dual attention network for diabetic retinopathy grading: IDANet[J]. Artif Intell Med, 2024, 149: 102782.

[4] Zhang D, Liu MT, Chen FS, et al. Graph-based multi-level feature fusion network for diabetic retinopathy grading using ultra-wide-field images[J]. Biomed Signal Process Control, 2024, 93: 106134.

[5] 杨佳楠, 姜同连, 朱福彬, 等. miRNA在糖尿病肾病足细胞损伤中作用及其机制的研究进展[J]. 吉林大学学报(医学版), 2023, 49(6): 1677-1682.
Yang JN, Jiang TL, Zhu FB, et al. Research progress in effect of miRNA on podocyte injury in diabetic nephropathy and its mechanism [J]. Journal of Jilin University (Medicine Edition), 2023, 49(6): 1677-1682.

[6] Sebastian A, Elharrouss O, Al-Maadeed S, et al. A survey on deep-learning-based diabetic retinopathy classification[J]. Diagnostics (Basel), 2023, 13(3): 345.

[7] 孙石磊, 李明, 刘静, 等. 深度学习在糖尿病视网膜病变分类领域的研究进展[J]. 计算机工程与应用, 2024, 60(8): 16-30.
Sun SL, Li M, Liu J, et al. Research progress on deep learning in field of diabetic retinopathy classification[J]. Computer Engineering and Applications, 2024, 60(8): 16-30.

[8] Grauslund J. Diabetic retinopathy screening in the emerging era of artificial intelligence[J]. Diabetologia, 2022, 65(9): 1415-1423.

[9] Li J, Chen JY, Tang YC, et al. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives[J]. Med Image Anal, 2023, 85: 102762.

[10] Matsoukas C, Haslum JF, Söderberg M, et al. Is it time to replace CNNs with transformers for medical images?[EB/OL]. (2021-08-20). https://arxiv.org/abs/2108.09038.

[11] Zhang L, Wang XS, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains *via* deep stacked transformation[J]. IEEE Trans Med Imaging, 2020, 39(7): 2531-2540.

[12] Qureshi I, Ma J, Abbas Q. Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning[J]. Multimed Tools Appl, 2021, 80(8): 11691-11721.

[13] Wang XF, Xu M, Zhang JC, et al. Joint learning of multi-level tasks for diabetic retinopathy grading on low-resolution fundus images[J]. IEEE J Biomed Health Inform, 2022, 26(5): 2216-2227.

[14] 赵爽, 穆鸽, 赵文华, 等. 基于特征融合网络的糖尿病视网膜病变分类[J]. 激光与光电子学进展, 2023, 60(14): 300-306.
Zhao S, Mu G, Zhao WH, et al. Classification of diabetic retinopathy with feature fusion network[J]. Laser & Optoelectronics Progress, 2023, 60(14): 300-306.

[15] Wang SQ, Wang XY, Hu Y, et al. Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision[J].

IEEE Trans Autom Sci Eng, 2021, 18(2): 574-585.
[16] Han ZK, Yang B, Deng SG, et al. Category weighted network and relation weighted label for diabetic retinopathy screening[J]. Comput Biol Med, 2023, 152: 106408.
[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2021-06-03). https://arxiv.org/abs/2010.11929.
[18] Wu JF, Hu R, Xiao ZH, et al. Vision transformer-based recognition of diabetic retinopathy grade[J]. Med Phys, 2021, 48(12): 7850-7863.
[19] Gu ZY, Li Y, Wang ZJ, et al. Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention[J]. Comput Intell Neurosci, 2023, 2023: 1305583.
[20] Sun R, Li YH, Zhang TZ, et al. Lesion-aware transformers for diabetic retinopathy grading[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 10933-10942.
[21] Pham TH, Li XQ, Nguyen KD. seUNet-trans: a simple yet effective UNet-transformer model for medical image segmentation[J]. IEEE Access, 2024, 12: 122139-122154.
[22] Jiang B, Zhao KK, Tang J. RGTransformer: region-graph transformer for image representation and few-shot classification[J]. IEEE Signal Process Lett, 2022, 29: 792-796.
[23] 庄建军, 吴晓慧, 景生华, 等. 多尺度特征融合的改进残差网络乳腺癌病理图像分类[J]. 中国生物医学工程学报, 2024, 43(4): 419-428. Zhuang JJ, Wu XH, Jing SH, et al. Improved residual network classification of breast cancer pathological images based on multi-scale feature fusion[J]. Chinese Journal of Biomedical Engineering, 2024, 43(4): 419-428.
[24] Zang F, Ma H. CRA-Net: transformer guided category-relation attention network for diabetic retinopathy grading[J]. Comput Biol Med, 2024, 170: 107993.
[25] Fu J, Liu J, Tian HJ, et al. Dual attention network for scene segmentation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 3141-3149.
[26] Woo S, Park J, Lee JY, et al. CBAM: convolutional block attention module [C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.
[27] Porwal P, Pachade S, Kokare M, et al. IDRiD: diabetic retinopathy-segmentation and grading challenge[J]. Med Image Anal, 2020, 59: 101561.
[28] Li T, Gao YQ, Wang K, et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening[J]. Inf Sci, 2019, 501: 511-522.
[29] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10). https://arxiv.org/abs/1409.1556.
[30] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2012: 1097-1105.
[31] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.
[32] Tan MX, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks [C]//Proceedings of the 36th International Conference on Machine Learning. Chia Laguna Resort, Sardinia, Italy: PMLR, 2019: 6105-6114.
[33] Liu Z, Lin YT, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 9992-10002.
[34] Guo JY, Han K, Wu H, et al. CMT: convolutional neural networks meet vision transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 12165-12175.
[35] Peng ZL, Huang W, Gu SZ, et al. Conformer: local features coupling global representations for visual recognition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 357-366.

（编辑：陈丽霞）