山西大学学报(自然科学版)49(2):284－294,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024087

基于CNN与Transformer混合的轻量化皮肤病变分类网络

徐健,赵欣*,李鑫杰
(大连大学 信息工程学院,辽宁 大连 116622)

摘　要:针对皮肤病变图像类别数量分布不均衡,现有的分类算法参数量大,计算复杂且分类性能有待提高的问题,本文提出了轻量级Transformer模块和新的卷积神经网络(Convolutional Neural Network, CNN)与Transformer结合策略以提升网络分类性能,同时采用逆类频率损失函数加权方案解决图像类别数量分布不均带来的训练影响。轻量级Transformer将输入序列进行显著特征提取后进行可分离自注意力计算来捕获皮肤病灶区域的全局特征信息,并解决Transformer计算量大的缺陷;新策略有效结合网络浅层的全局细节特征信息与深层的语义特征信息来提升网络表达能力。在HAM10000数据集上的实验结果表明,所提算法各评价指标高于其他对比算法,同时模型的参数保持在230万,对于推广自动皮肤病变分类工具具有重要意义。

关键词:图像分类;皮肤病变分类;轻量化网络;混合网络;HAM10000
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2026)02-0284-11

A Lightweight Classification Network for Skin Lesion Based on the Hybridization of CNN and Transformer

XU Jian, ZHAO Xin*, LI Xinjie
(*College of Information Engineering, Dalian University, Dalian* 116622*, China*)

Abstract: Existing classification algorithms often suffer from a large number of parameters, high computational complexity, and sub-optimal classification performance in classifying skin lesion images, due to their uneven distribution attibutes. To address this issue, in this paper, we propose a lightweight transformer module and a novel strategy that combines Convolutional Neural Network (CNN) with Transformer to enhance network classification performance. Additionally, we adopt an inverse class loss function weighting scheme to mitigate the impact of imbalanced image category distribution during training. The lightweight transformer extracts essential features from input sequences, and performs separable self-attention computations to capture global feature information from skin lesion regions. This approach addresses the computational limitations of traditional transformer. Furthermore, our new strategy effectively integrates shallow global detail features with deep semantic features, enhancing the network's expressive ability. Experimental results on the HAM10000 dataset demonstrate that our algorithm outperforms other comparative methods in terms of evaluation metrics. Remarkably, we achieve these results while maintaining a model size of only 2.3 million parameters, which holds significant promise for advancing automatic skin lesion classification tools.

Key words: image classification; skin lesions classification; lightweight network; hybrid network; HAM10000

收稿日期:2024-03-04;**修回日期**:2024-05-28
基金项目:国家自然科学基金(61971424)
作者简介:徐健(1999—),男,河北廊坊人,硕士研究生,研究方向为医学图像处理、深度学习模型轻量化。E-mail:xujian1117@126.com
* **通信作者**:赵欣(ZHAO Xin),E-mail:zhaoxin@dlu.edu.cn
引文格式:徐健,赵欣,李鑫杰. 基于CNN与Transformer混合的轻量化皮肤病变分类网络[J]. 山西大学学报(自然科学版),2026,49(2):284-294. DOI:10.13451/j.sxu.ns.2024087.

0 引言

皮肤病变是一种常见的皮肤增生性疾病,包含多种类型。临床上,分为良性与恶性病变。良性皮肤病不会转移或扩散到身体的其他部位,一般不需要治疗。然而,恶性皮肤病变有侵袭周围组织器官或者转移其他身体部位的危险,其危害不容忽视。早期发现和诊断恶性皮肤病变至关重要,如果能够及时识别和治疗,可以显著提高治愈率,减少患者的痛苦和死亡率。然而,由于不同病例的病灶部位具有高强度的相似性和微小的差异,仅靠医生肉眼观察容易发生误诊。随着人工智能(Artificial Intelligence,AI)兴起,以计算机进行皮肤病变的辅助诊断已成为可能,通过AI寻找一种高效、准确地针对早期皮肤病变自动诊断方法具有重要意义。

近年来,基于卷积神经网络(Convolutional Neural Network,CNN)的方法在计算机视觉方面取得了令人印象深刻的进步和发展。卷积层凭借其强大的归纳偏置和先验知识,使得CNN具有出色的泛化能力和快速的收敛速度,这促使越来越多的学者选择使用CNN进行皮肤病变分类。Han等[1]使用CNN对134种疾病进行了分类,结果表明CNN能有效识别出模糊的、难以区分的病变图像。

传统CNN因其庞大的参数量,通常需要花费更多的计算资源和时间来训练,同时增大了部署难度。皮肤病分类的轻量级网络作为皮肤病自动识别的一个重要方向,因为可以在移动端部署,使患者在医生诊断前可获得准确的检测结果,同时为医生提供了诊断参考。Toğaçar等[2]提出了一种基于自编码器、脉冲神经网络和CNN的网络模型,通过使用MobileNetV2[3]轻量级分类网络进行分类,以自编码器和脉冲神经网络弥补MobileNetV2网络分类精度低的缺陷。类似地,Srinivasu等[4]向MobileNetV2中添加长短期记忆(Long Short-term Memory,LSTM)机制后在HAM10000数据集上达到了85%的分类准确率。而Hoang等[5]提出了一种基于新的分割方法和wide-ShuffleNet网络的皮肤病变分类方法,通过计算皮肤图像的熵加权和一阶累积矩(Entropy-based Weighting and First-order Cumulative Moment,EW-FCM)来从背景中分离病灶,将分割结果输入到wide-ShuffleNet中确定皮肤病变类型,实验在HAM10000数据集上达到了84.80%的准确率,同时模型有180万个可学习参数,使算法更加适用于移动医疗保健系统。

上述工作尽管已经取得了很大的成功,但在模型轻量化的前提下算法的分类性能仍有待进一步提高。而在皮肤病变图像类内差异性大和类间相似性强的挑战下,类间差异学习需要完整病变的全局上下文信息,但是现有方法缺少全局上下文信息。考虑到复杂皮肤条件的存在,局部病变定位可能会丢失皮肤病变的全局信息,进而导致类间特征学习的判别信息丢失[6]。

浅层CNN无法学习远程上下文信息,这主要是由于卷积操作的感受野很小[7]。Chen等[8]试图通过使用扩展卷积增加卷积层的感受野来克服这一限制,然而当叠加连续扩张率的卷积核时会损失特征信息的连续性,导致一些特征信息并不会参与运算。视觉Transformer(Vision Transformer,ViT)[9]因为能够捕获图像中的远程依赖关系,使其成功应用于图像识别任务,Cheslerean-Boghiu等[10]通过使用ViT架构捕获皮肤病灶区域的全局上下文信息,并实现单级多模态数据融合,这种方法在图像丰富和患者数据丰富的环境中取得了杰出的成果。但是,相较于CNN,Transformer计算成本更高,对空间位置的不敏感以及对数据集预处理的高要求也影响着它在皮肤病图像领域的实际应用。为探索CNN与Transformer相结合对网络分类性能产生的影响,Peng等[11]提出了Resnet[12]与ViT并行的网络Confomer,并设计特征耦合单元以消除两种不同风格特征的语义差异,在ImageNet上的性能表现超越Resnet或ViT单分支网络。类似地,Mehta等[13]提出了MobileViT网络,具体来说MobileViT通过引入MobileViT块来进行有效的局部与全局信息编码,MobileViT块通过使用Transformer来将卷积中的局部建模替换为全局建模使得MobileViT块同时具备CNN与ViT的优势。因此,设计有效的CNN与Transformer结合算法成为一种提

升网络分类性能的有效途径。

针对上述问题,本文提出了一种新的轻量化的CNN与Transformer混合分类网络Cross-MCViT(Cross-layer Mobile CNN-vit Classification Network),用于实现皮肤病变图像自动分类。

具体而言,本文的贡献如下:

1)为了更高效地提取皮肤病灶的全局信息,提出轻量级Transformer模块。相对于输入数据中序列k的数量,轻量级Transformer进行自注意力计算的时间复杂度为$O(k/4)$,比Transformer计算成本更低,更有利于未来在移动端部署。

2)在Cross-MCViT网络中,为加强对浅层细粒度特征信息的复用,提出了新的CNN与Transformer混合策略,通过在网络浅层阶段末尾放置轻量级Transformer块赋予浅层捕获全局信息的能力,再以跨层的方式拼接到网络深层进行信息融合以丰富特征细节,提升网络对于皮肤病变的判别能力。

3)为解决皮肤病变图像不同类别数量分布不均衡给模型训练带来的影响,采用逆类频率加权方案对损失函数进行加权训练。

1 方法

1.1 Cross-MCViT网络主体结构

Cross-MCViT网络的设计目的是建立一个高效的混合网络利用CNN与Transformer各自优势以提升皮肤病变分类水平。网络模型主体结构如图1所示,主要由Stem块、MV2(mobilenetv2 block)块、轻量级Transformer块组成。首先,输入图像经Stem块进行细粒度特征提取,Stem块由一个常规3×3卷积、批标准化(BatchNorm)和ReLu6激活函数组成,然后分别输入到MV2块与轻量级Transformer块进行表示学习。网络主体结构采用了CNN的分层架构设计,以步长为2的MV2块逐步降低分辨率,分层架构设计有助于提取多尺度特征,以用于类内差异性学习,捕获皮肤病变的局部病变信息。

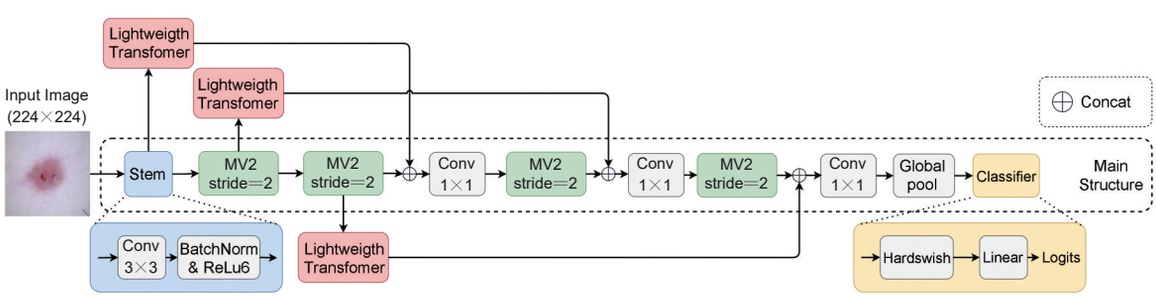

图1 Cross-MCViT网络结构

Fig. 1 Structure of Cross-MCViT network

此外,考虑到类间差异学习需要完整病变的全局上下文信息,而在CNN网络中,浅层特征图相较于深层特征图有更加丰富的空间信息且能够进一步捕获皮肤病灶区域的细粒度特征信息,这些浅层捕获的细粒度特征信息包含图像更多的低级细节与纹理信息,对最后的类区分尤为关键。因此,为了更高效地增强对网络浅层特征信息复用,通过在网络浅层卷积块的基础之上引入并行的轻量级Transformer块,以一种横向方式扩展网络,并在网络深层进行两者的逐步信息融合以提升网络性能。轻量级Transformer块通过将浅层特征图经显著信息提取后送入LinearTransformer进行自注意力计算以获取不同类别病灶区域的全局结构信息,并以跨层的方式拼接到网络深层进行特征信息补偿,从而提高模型对重要特征的关注程度,同时这种类残差结构也避免网络出现退化现象。

另一方面,通过卷积算子收集的局部特征信息与使用全局自注意力机制获取的全局特征信息存在着明显的语义差异[11],因此,在Cross-MCViT网络深层每一个拼接处应用点卷积(Pointwise Convolution)进行信息融合,以填补语义空白。经由信息融合后网络深层特征图结合了皮肤病灶区域低层次的细节特征和高层次的语义特征,丰富的细节保证了预测的准确性,使最终预测效果更好。添加点卷积也增加了网络的非线性特性,同时控制了特征层通道数量,避免网络参数量膨胀。

Cross-MCViT的两个关键模块MV2与轻量级Transformer具体表述如下。

1.2 MV2块

深度可分离卷积(Depthwise Separable Convolution,DWConv)在MobileNetV1[14]网络中被提出,由深度卷积(Depthwise Convolution)和点卷积组成,虽然减少了CNN计算量与参数量,但是不可避免地会丢失部分信息。为了解决MobileNetV1中出现的信息丢失问题,MobileNetV2提出了倒残差结构以进行更高效的特征提取,如图2所示,具体来说,就是在DWConv前加上一个点卷积,目的是将低维空间映射到高维空间,因为DWConv本身并没有改变通道数的能力。本文中MV2块将输入维度c_{in}通过点卷积扩展6倍以丰富特征细节。

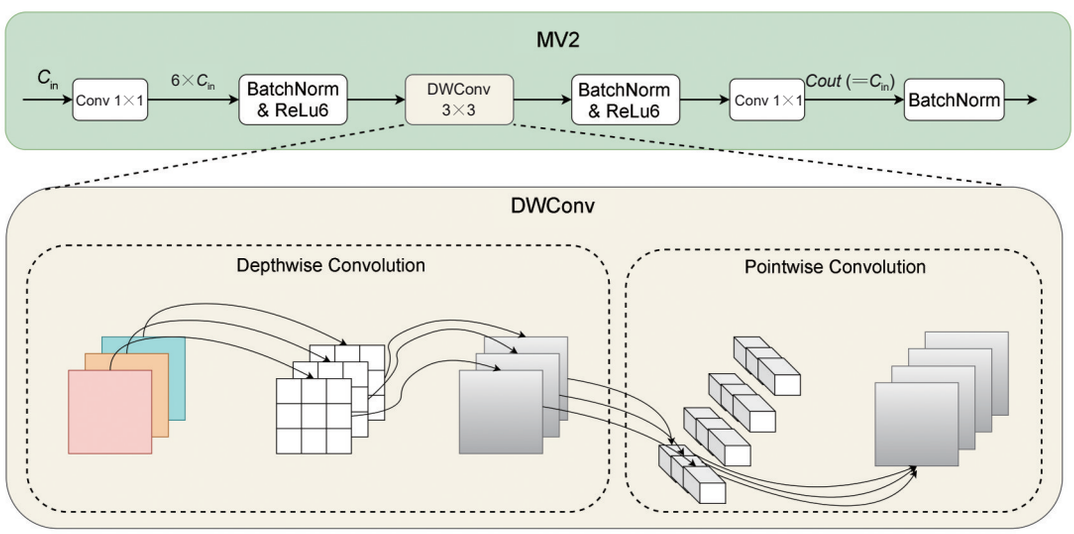

图2 倒残差结构

Fig. 2 Inverted residual block

1.3 轻量级Transformer块

ViT与基于CNN的模型相比,具有较高的计算成本与延迟,主要效率瓶颈是多头自注意力机制(Multi-headed Self-attention,MHA),面对输入数据中序列k的数量,它需要$O(k^2)$的时间复杂度。

为了更高效地提取皮肤病变区域全局上下文信息,本文提出了轻量级Transformer模块,如图3所示,通过引入一种时间复杂度为$O(k)$的可分离自注意力计算方法(Separate Selfattention)[15],并将特征信息在送入LinearTransformer前,经由大小为2×2,步长为2的平均池化与最大值池化进行显著特征信息提取。经显著特征提取后输入到LinearTransformer的序列长度变为原本的1/4,有效减少特征信息的规模,进一步减少计算量。因为输入到CNN与输入到LinearTransformer的特征信息格式不同,需要将特征信息进行二维特征图与一维序列的转换。最终,相较于Transformer,本文提出的方法时间复杂度仅为$O(k/4)$,在计算成本方面更为高效。

与MHA类似,可分离自注意力结构如图4(a)所示。输入$\boldsymbol{x}$使用三个分支进行处理,即L、键$\boldsymbol{K}$和值V。L使用一个权重为$\boldsymbol{W}_I\in\boldsymbol{R}^d$的线性层将$\boldsymbol{x}$中的每个$d$维序列线性映射为一个标量,线性映射是一个内积运算。$\boldsymbol{W}_I$作为图4(b)中的潜在节点L,通过计算L和$\boldsymbol{x}$之间的距离,得到一个k维向量,然后对这个k维向量应用softmax来生成上下文分数$\boldsymbol{c}_s\in\boldsymbol{R}^k$。与计算每个序列相对于所有$k$个标记的上下文分数的Transformer不同,可分离自注意力只计算相对于潜在序列L的上下文分数。这将计算上下文分数的成本从$O(k^2)$降低到$O(k)$。

上下文分数$\boldsymbol{c}_s$用于计算上下文向量$\boldsymbol{c}_v$,它通过使用权值为$\boldsymbol{W}_K\in\boldsymbol{R}^{d\times d}$的键分支$\boldsymbol{K}$将输入$\boldsymbol{x}$线性投影到$d$维空间,产生输出$\boldsymbol{x}_K\in\boldsymbol{R}^{k\times d}$,然后将上下文向量$\boldsymbol{c}_v\in\boldsymbol{R}^d$作为$\boldsymbol{x}_K$的加权和,如图4(c)所示。数学上,$\boldsymbol{c}_v$定义为:

$$\boldsymbol{c}_v=\sum_{i=1}^{k}c_s(i)x_K(i)。\tag{1}$$

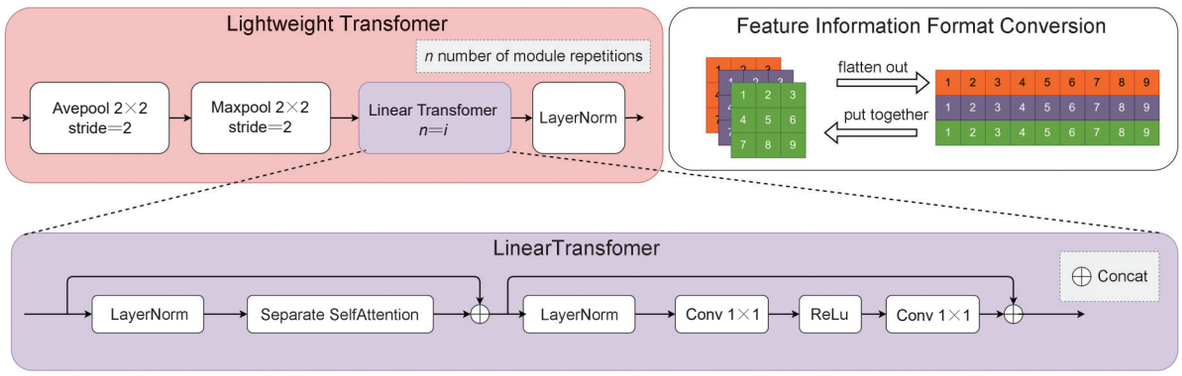

图3 轻量级Transformer模块(其中i的取值在Cross-MCViT网络中为2,3,3)

Fig. 3 Lightweight Transformer block (the value of i is 2, 3, 3 in the Cross-MCViT network)

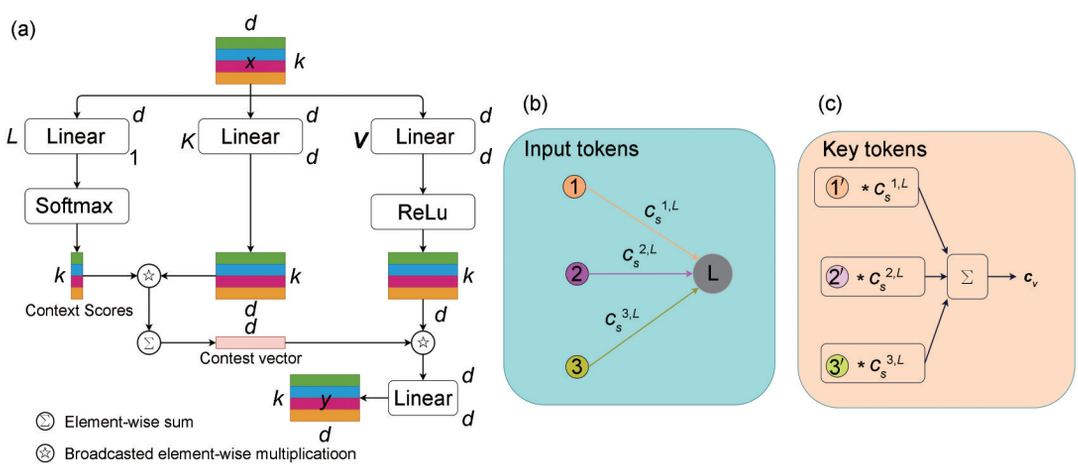

图4 基于可分离自注意力的轻量化Transformer模块

(a)可分离自注意力计算;(b)计算上下文分数c_s;(c)计算上下文向量c_v。

Fig. 4 Lightweight transformer module based on separable self-attention

(a) Separable self-attention calculation; (b) Calculation of contextual score c_s; (c) Calculation of contextual vector c_v.

为了使c_v中编码的上下文信息与x中的所有序列共享,使用权重为$W_V \in R^{d\times d}$的分支V将x投影到d维空间,通过ReLU激活产生输出$x_V \in R^{k\times d}$。然后,c_v中的上下文信息通过广播的元素相乘(broadcasted element-wise multiplication)操作传播到x_V。最后,将结果馈送到另一个权重为$W_O \in R^{d\times d}$的线性层,以产生最终输出$y \in R^{k\times d}$。数学上,可分离自注意力定义为:

$$y=\left(\sum(\sigma(xW_I)*xW_K)*\mathrm{ReLu}(W_V)\right)W_O, \quad (2)$$

其中σ代表softmax函数,$*$为支持广播的逐元素乘法。

2 实验

2.1 HAM10000数据集

HAM10000是皮肤镜数据库[16],由7种不同类型的皮肤病组成:光化性角化病(Actinic Keratoses, AKIEC)、基底细胞癌(Basal Cell Carcinoma, BCC)、良性角化病(Benign Keratosis, BKL)、皮肤纤维瘤(Dermatofibroma, DF)、黑色素瘤(Melanoma, MEL)、黑色素细胞痣(Melanocytic Nevi, NV)和血管病变(Vascular Skin Lesions, VASC)。HAM10000数据库的一个主要问题是类别不平衡,这会导致模型在训练时偏向占比比重大的类。如表1所示,NV类约占总图像数的67%,其他类贡献的图像数量很少,特别是DF类的图像,占总图像的比例不到2%。此外,如图5所示,皮肤病变在大小、形状、颜色和质地等病变特征上有比较大的类内变化,同时类间又存在着高度的视觉相似性。其次,皮肤镜图像经常包含外来的伪影(气泡、尺子、校准图等),以及固有的皮肤特征(毛发、皮肤线

条、血管等),这些特征会遮挡皮肤病变,使病变分类复杂化。

2.2 实验环境设置

实验环境采用深度学习框架Pytorch,在NVIDIA RTX 3090(24 GB)GPU上进行,训练过程中,通过使用图像随机翻转、随机旋转、中心裁剪等数据增强策略,来增强模型的泛化能力,所使用的数据增强参数设置如表2所示。初始学习率设置为0.008,使用余弦退火学习率衰减策略,批次数量(batchsize)设置为16,训练epoch次数为300,优化算法采用随机梯度下降算法(Stochastic Gradient Descent, SGD)。

表1 HAM10000数据集类别分布

Table 1 Distribution of classes in HAM10000 dataset

类别	图像数量
AKIEC	327
BCC	514
BKL	1 099
DF	115
MEL	1 113
NV	6 705
VASC	142

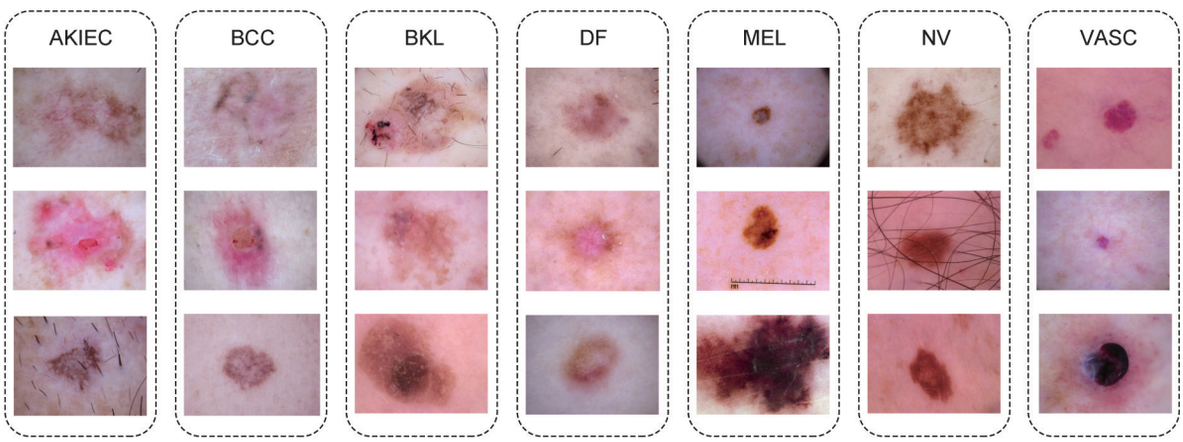

图5 HAM10000中的皮肤病变图像,从上到下展现了类内差异性,从左到右展现了类间相似性

Fig. 5 Images of skin lesions in HAM10000, from top to bottom shows the class difference, from left to right shows the similarity between classes

表2 数据增强参数设置

Table 2 Parameters setting of data enhancement

数据增强类别	参数设置
RandomHorizonalFlip	p=0.4
RandomRotation	degrees=20
Resize	size=(288,288)
CenterCrop	size=(224,224)

同时,为了避免模型在训练过程中偏向比重大的类,损失函数采用加权交叉熵损失函数。通过采用逆类频率加权方案对交叉熵损失函数进行加权,权值计算方法如下式所示:

$$\omega_{\text{class\_weight}\,i}=\frac{\sum_{j=1}^{k}N_j}{k\times N_i},\tag{3}$$

其中N_i是第i类的样本数,$\omega_{\text{class_weight}\,i}$为计算出的$i$类别权重,$k$是类别数量。

2.3 评价指标

采用常用的5种评价指标:准确率(Accuracy, Acc)、灵敏度(Sensitivity, $r_{\text{Sensitivity}}$)、特异性(Specificity, $r_{\text{Specificity}}$),精确度(Precision, P)以及$F1$值($F1$-score),来对所有模型性能进行综合评价。

准确度表示为正确分类的样本数占总样本数的比例,用来衡量分类模型对所有类别的分类能力的整体表现。数学上其计算公式定义如下:

$$\text{Acc}=\frac{TP+TN}{TP+FP+TN+FN},\tag{4}$$

其中TP为真阳性,TN为真阴性,FN为假阴性,FP为假阳性。

灵敏度表示为成功预测为阳性样本的个数占所有阳性样本个数(包含误测)的比例,用来衡量分类模型预测阳性样本的能力。在多分类问题中,如果将其中一类视为阳性,则其他类别视为阴性。数学上其计算公式定义如下:

$$r_{\text{Sensitivity}}=\frac{TP}{TP+FN}。\tag{5}$$

特异性表示为成功预测为阴性样本的个数占所有阴性样本个数(包含误测)的比例,用来

衡量分类模型预测阴性样本的能力。数学上其计算公式定义如下：

$$r_{\text{Specificity}}=\frac{TN}{TN+FP}。\tag{6}$$

精确度表示为预测为阳性的样本中有多少是真正的阳性，用来衡量分类模型成功预测阳性样本的能力。数学上其计算公式定义如下：

$$P=\frac{TP}{TP+FP}。\tag{7}$$

*F*1-Score 通过综合考虑灵敏度与精度，兼顾阳性样本的误报和漏报，可视为一种更为均衡的评价指标，尤其适用于类别不平衡的情况。数学上其计算公式定义如下：

$$F1=\frac{2\times P\times r_{\text{Sensitivity}}}{P+r_{\text{Sensitivity}}}。\tag{8}$$

2.4 经典网络对比实验

Cross-MCViT 网络作为一种新型的 Transformer 与 CNN 混合模型，通过与经典网络 MobileNetV2、ResNet50、DenseNet121[17]、ViT16 与 MobileViT 进行性能比较，以验证 Cross-MCViT 未来作为皮肤病分类基准网络的潜在效果。

数据集划分策略采用 HAM10000 数据集的 80% 用于训练，20% 用于测试。在图像输入网络之前，对图像进行上文提到的数据增强和归一化操作，将图像大小调整为 224×224 像素。各网络在 HAM10000 数据集上的混淆矩阵如图 6 所示，指标值如表 3 所示，由表 3 实验结果可以得出以下两点结论：

1）Cross-MCViT 在准确率、灵敏度、特异性、精确度与 *F*1-Score 上均取得了最优值，在参数量上相较于 MobileNetV2 略有增加。此外，ViT16 的各项指标性能表现并不如 CNN，这也表明 Transformer 架构的网络在面对有限的皮肤病变数据样本以及数据增强策略时，并不能很好地胜任皮肤病变分类任务。

2）MobileViT 网络得益于融合了卷积（空间归纳偏置和对数据增强的低敏感度）和 Transformer（全局处理）的优点，在各项指标上的结果均超过了 CNN 与 ViT16，表明融合架构相较于单一架构在面对皮肤病变分类任务时更具优势，同时模型参数量保持在 430 万也更有利于移动端部署。与 Cross-MCViT 网络不同的是，MobileViT 将 Transformer 块放置在网络深层阶段末，但针对皮肤病变图像特性来说，如何更好地利用网络浅层特征信息尤为关键，表 3 的实验结果也验证了 Cross-MCViT 所使用的混合策略相较于 MobileViT 更适用于皮肤病变分类任务，在保持更高性能同时参数量上也下降了 46.51%。

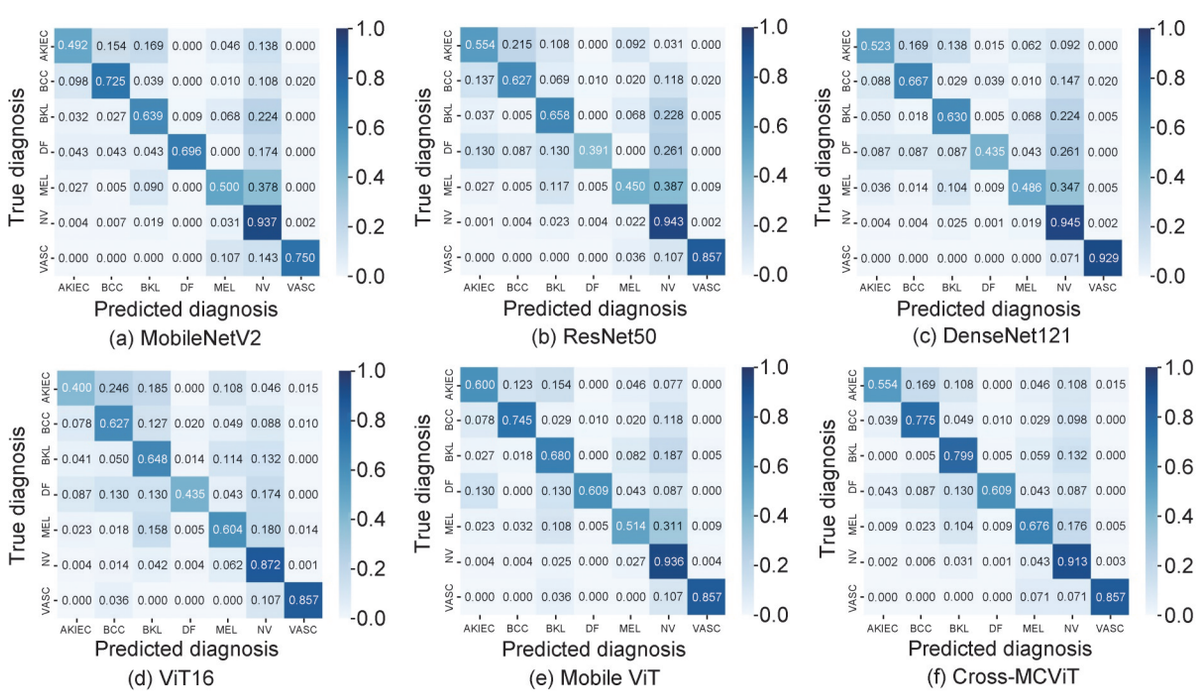

图 6 混淆矩阵

Fig. 6 Confusion matrix

表3 经典网络对比实验

Table 3 Comparative experiments on classical networks

评价指标	算法					
	MobileNetV2	Resnet50	DenseNet121	ViT16	MobileViT	Cross-MCViT
Specificity	95.03%	94.99%	95.13%	95.40%	95.59%	**96.43%**
Sensitivity	66.70%	64.00%	65.91%	63.48%	70.58%	**74.04%**
Precision	73.75%	67.63%	68.27%	60.62%	74.19%	**75.73%**
*F*1-Score	70.60%	65.77%	67.08%	62.02%	72.34%	**74.87%**
Accuracy	82.50%	82.05%	82.55%	78.50%	83.55%	**85.15%**
Parameters	**2.2×10^6**	23.5×10^6	7.0×10^6	85.7×10^6	4.7×10^6	2.3×10^6
Flops	**0.33×10^9**	4.13×10^9	2.90×10^9	16.83×10^9	1.33×10^9	2.37×10^9

注:加粗字体为行最优值,红色字体为行次优值,ViT16代表16层ViT模型。

2.5 改进算法对比实验

为了验证Cross-MCViT网络结构的有效性与先进性,本文与近6年皮肤病变自动分类相关的改进算法进行对比实验,包括:PNASNet[18]、Densenet121 with SVM[19]、ResNet50+gcForest[20]、VGG16+GoogLeNet Ensemble[21]、Bayesian DenseNet169[22]、Shifted MobileNetV2[23]、Shifted GoogLeNet[23]、Shifted 2-Nets[23]、EW-FCM+wide-ShuffleNet与DMN(Dual-path Multi-Branch Network)[24]。

由表4与表5可知,Cross-MCViT相较于其他改进算法在准确度上取得了次优值。表5中对比的5种改进算法给出了HAM10000数据集上的混淆矩阵,相比Shifted MobileNetV2、Shifted GoogLeNet、Shifted 2-Nets、EW-FCM+wide-ShuffleNet与DMN,Cross-MCViT在特异性、灵敏度与*F*1-Score上取得了最优值,分别得到了0.40%~1.73%、3.33%~15.94%与1.39%~14.07%性能提升。DMN通过在MobileNetV2上添加分支和连接来增加网络宽度以提高网络特征提取的能力,在准确度和精确度上取得了85.67%与77.96%的最优值,但对于皮肤病变图像样本不均衡的情况,准确度与精确度的最优值并不能合理反映模型的预测能力,同时分支和连接的增加也使得DMN的参数量超过了500万,运算量超过了190亿,相较于Cross-MCViT分别增长了130%与725%。综合各项指标的结果可以得出,Cross-MCViT通过有效结合病灶的局部与全局判别信息,在皮肤病变自动分类任务上取得了更加全面的性能表现。

表4 改进算法对比实验1

Table 4 Comparison experiments on improved algorithms 1

评价指标	算法					
	PNASNett	ResNet+gcForest	VGG16 + GoogLeNet Ensemble	DenseNet121 with SVM	Bayesian DenseNet169	Cross-MCViT
Accuracy	76.00%	80.40%	81.50%	82.70%	83.59%	**85.15%**

注:加粗字体为行最优值。

2.6 消融实验

Cross-MCViT网络是以MV2块为主体的网络,为验证在Cross-MCViT网络浅层添加轻量级Transformer块的有效性,通过从浅层到深层逐渐去除轻量级Transformer块进行消融实验。本消融实验分为三组:第一组去除首层轻量级Transformer块;第二组去除首层和第二层轻量级Transformer块;第三组去除全部轻量级Transformer块。其次,为验证采用逆类频率损失函数加权策略对模型训练带来的影响,通过与未加权的交叉熵损失函数训练出的网络Cross-MCViT*进行对比,以进行实验验证。

消融实验结果如表6所示,通过逐层去除轻量级Transformer块各项指标值均有下降,其中*F*1-Score值下降最为明显,完全去除轻量级Transformer后下降了7.87%,这也侧面印证了在网络中添加轻量级Transformer块有助于模型

表5 改进算法对比实验2

Table 5 Comparison experiments on improved algorithms 2

评价指标	算法					
	Shifted MobileNet	Shifted GoogLeNet	Shifted 2-Net	EW-FCM + wide-ShuffleNet	DMN	Cross-MCViT
Specificity	95.20%	94.70%	95.30%	96.03%	95.85%	**96.43%**
Sensitivity	65.90%	58.10%	64.40%	70.71%	70.21%	**74.04%**
Precision	71.40%	68.50%	76.10%	75.15%	**77.96%**	75.73%
*F*1-Score	67.00%	60.80%	67.80%	72.61%	73.48%	**74.87%**
Accuracy	81.90%	80.50%	83.20%	84.80%	**85.67%**	85.15%
Parameters	3.4×10^6	7×10^6	10.4×10^6	$\mathbf{1.8\times10^6}$	5.3×10^6	2.3×10^6
Flops	—	—	—	—	19.55×10^9	$\mathbf{2.37\times10^9}$

注:—代表引用论文中并未给出具体数值,且未给出网络模型具体配置,也无法估计。加粗字体为行最优值,红色字体为行次优值。

稳定性的提升。此外,Cross-MCViT*虽然在精度和准确度上取得了最优值,但结合图7混淆矩阵分析,Cross-MCViT*结果偏向图像数量占比大的NV类,而在其他少数类上的预测结果均不如Cross-MCViT,尤其对于恶性病变MEL类的预测较为一般。综合各指标值结果分析,通过采用逆类频率损失函数加权策略在一定程度上避免了模型在训练过程中偏向样本数量占比大的类。

表6 消融实验

Table 6 Ablation experiments

评价指标	算法/%				
	第三组	第二组	第一组	Cross-MCViT*	Cross-MCViT
Specificity	96.14	96.28	96.34	95.89	**96.43**
Sensitivity	68.96	73.69	72.75	69.57	**74.04**
Precision	69.21	70.47	71.87	**80.26**	75.73
*F*1-Score	67.00	72.04	72.31	74.54	**74.87**
Accuracy	81.20	82.30	83.80	**86.15**	85.15

注:加粗字体为行最优值,红色字体为行次优值。

2.7 可解释性实验

深度学习模型的不可解释性影响着其在医疗领域的实际应用,在皮肤科医生和数据驱动系统之间建立信任需要设计透明的模型,解释"它们为何做出相应的预测"。Grad-CAM[25]是一种类激活映射可解释性方法,通过生成热力图的方式可视化模型在作出对应判断时其在图像上的重点关注区域。它利用梯度来计算卷积层中空间位置的重要性,由于梯度是针对唯一类计算的,红色显示的区域表示这些区域更具类区分性。

如图8所示,Cross-MCViT网络在对皮肤病变类型做出正确预测时,热图所显示出的重点区域定位在病灶部位,表明Cross-MCViT网络能够在皮肤病图像中获得代表语义特征的细节。

3 结论

由于皮肤病变图像类别分布不均,图像表

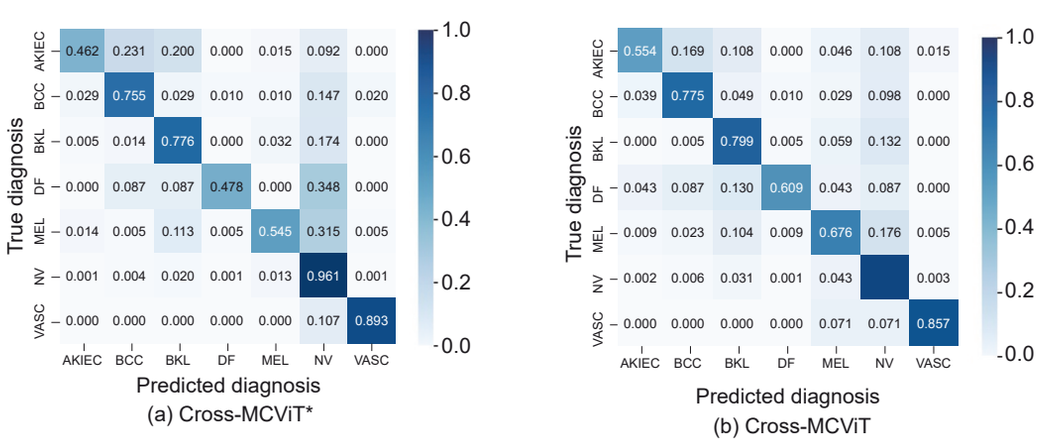

图7 不同损失函数上的混淆矩阵

Fig. 7 Confusion matrix on different loss functions

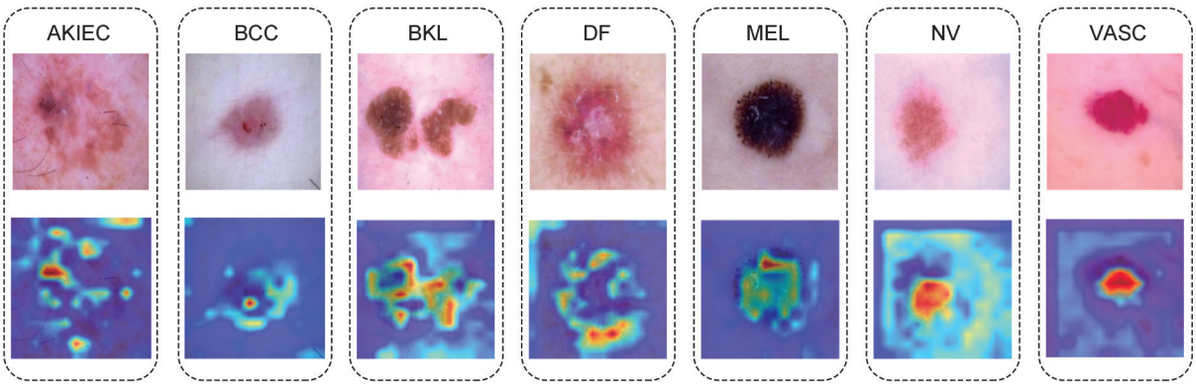

图8 Grad-CAM热力图可视化结果

Fig. 8 Visualization results of Grad-CAM heat map

现出类间相似性高、类内差异大的视觉特性,以及固有的皮肤特征、图像伪影等问题,导致皮肤病变图像分类任务极具挑战性,本文提出了一种新的网络架构Cross-MCViT,它通过高效的卷积模块与本文提出的轻量级Transformer模块提取皮肤病变的局部与全局判别信息,并将病灶的全局信息跨层拼接到网络深层,增强网络深层的表达与判别能力。在HAM10000数据集上的实验表明,Cross-MCViT相较于其他算法在多个评价指标上取得了最优值。其次,Cross-MCViT是基于轻量化设计的网络,该算法在保持更高性能的同时将参数量减少到230万。

在未来的工作中,为进一步验证Cross-MCViT网络的有效性,将与更多的先进算法进行比较,研究将所提出的方法集成到移动医疗保健系统等现实问题中。

参考文献:

[1] HAN S S, PARK I, CHANG S E, *et al*. Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders[J]. *J Invest Dermatol*, 2020, **140**(9): 1753-1761. DOI: 10.1016/j.jid.2020.01.019.

[2] TOĞAÇAR M, CÖMERT Z, ERGEN B. Intelligent Skin Cancer Detection Applying Autoencoder, MobileNetV2 and Spiking Neural Networks[J]. *Chaos Solitons Fractals*, 2021, **144**: 110714. DOI: 10.1016/j.chaos.2021.110714.

[3] SANDLER M, HOWARD A, ZHU M L, *et al*. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 4510-4520. DOI: 10.1109/CVPR.2018.00474.

[4] SRINIVASU P N, SIVASAI J G, IJAZ M F, *et al*. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM[J]. *Sensors*, 2021, **21**(8): 2852. DOI: 10.3390/s21082852.

[5] HOANG L, LEE S H, LEE E J, *et al*. Multiclass Skin Lesion Classification Using a Novel Lightweight Deep Learning Framework for Smart Healthcare[J]. *Appl Sci*, 2022, **12**(5): 2677. DOI: 10.3390/app12052677.

[6] WANG L T, ZHANG L, SHU X, *et al*. Intra-class Consistency and Inter-class Discrimination Feature Learning for Automatic Skin Lesion Classification[J]. *Med Image Anal*, 2023, **85**: 102746. DOI: 10.1016/j.media.2023.102746.

[7] HE X Z, TAN E L, BI H W, *et al*. Fully Transformer Network for Skin Lesion Analysis[J]. *Med Image Anal*, 2022, **77**: 102357. DOI: 10.1016/j.media.2022.102357.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, *et al*. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFS[J]. *IEEE Trans Pattern Anal Mach Intell*, 2018, **40**(4): 834-848. DOI: 10.1109/TPAMI.2017.2699184.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. (2021-06-03)[2024-03-04]. https://arxiv.org/abs/10.48550/arxiv.2010.11929.

[10] CHESLEREAN-BOGHIU T, FLEISCHMANN M E, WILLEM T, *et al*. Transformer-based Interpretable Multi-modal Data Fusion for Skin Lesion Classification [EB/OL]. (2023-08-31)[2024-03-04]. https://arxiv.org/abs/2304.14505.

[11] PENG Z L, HUANG W, GU S Z, *et al*. Conformer: Local Features Coupling Global Representations for Visual Recognition[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 357-366. DOI: 10.1109/ICCV48922.2021.00042.

[12] 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 (CVPR). New York: IEEE, 2016: 770−778. DOI: 10.1109/CVPR.2016.90.

[13] MEHTA S, RASTEGARI M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer[EB/OL]. (2022-03-04) [2024-03-04]. https://arxiv.org/abs/2110.02178.

[14] HOWARD A G, ZHU M L, CHEN B, *et al*. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[EB/OL]. (2017−04−17)[2024−03−04]. https://arxiv.org/abs/1704.04861.

[15] MEHTA S, RASTEGARI M. Separable Self-attention for Mobile Vision Transformers[EB/OL]. (2022−06−06)[2024−03−04].https://arxiv.org/abs/2206.02680.

[16] TSCHANDL P, ROSENDAHL C, KITTLER H. The HAM10000 Dataset, a Large Collection of Multi-source Dermatoscopic Images of Common Pigmented Skin Lesions[J]. *Sci Data*, 2018, **5**: 180161. DOI: 10.1038/sdata.2018.161.

[17] HUANG G, LIU Z, VAN DER MAATEN L, *et al*. Densely Connected Convolutional Networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 2261−2269. DOI: 10.1109/CVPR.2017.243.

[18] MILTON M A A. Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018: Skin Lesion Analysis towards Melanoma Detection Challenge[EB/OL]. (2019−01−30) [2024−03−04]. https://arxiv.org/abs/1901.10802.

[19] GESSERT N, SENTKER T, MADESTA F, *et al*. Skin Lesion Diagnosis Using Ensembles, Unscaled Multi-crop Evaluation and Loss Weighting[EB/OL]. (2018−08−05)[2024−03−04].https://arxiv.org/abs/1808.01694.

[20] RAY S. Disease Classification within Dermascopic Images Using Features Extracted by ResNet50 and Classification through Deep Forest[EB/OL]. (2018−07−25)[2024−03−04]. https://arxiv.org/abs/1807.05711.

[21] PEREZ F, AVILA S, VALLE E. Solo or Ensemble? Choosing a CNN Architecture for Melanoma Classification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New York: IEEE, 2019: 2775−2783. DOI: 10.1109/CVPRW.2019.00336.

[22] MOBINY A, SINGH A, VAN NGUYEN H. Risk-aware Machine Learning Classifier for Skin Lesion Diagnosis [J]. *J Clin Med*, 2019, **8**(8): 1241. DOI: 10.3390/jcm8081241.

[23] THURNHOFER-HEMSI K, LÓPEZ-RUBIO E, DOMÍNGUEZ E, *et al*. Skin Lesion Classification by Ensembles of Deep Convolutional Networks and Regularly Spaced Shifting[J]. *IEEE Access*, 2021, **9**: 112193−112205. DOI: 10.1109/ACCESS.2021.3103410.

[24] WANG H, QI Q Q, SUN W J, *et al*. Classification of Skin Lesions with Generative Adversarial Networks and Improved MobileNetV2[J]. *Int J Imaging Syst Tech*, 2023, **33**(5): 1561−1576. DOI: 10.1002/ima.22880.

[25] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 618−626. DOI: 10.1109/ICCV.2017.74.