Computer Science and Application 计算机科学与应用, 2026, 16(1), 154-168
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161013

Hans汉斯

NexusNet：一种面向面部颜色识别的双路径层次化融合混合网络

孙千帅，冯　跃，林卓胜，梁洁欣，赵　雪，刘子豪

五邑大学电子与信息工程学院，广东　江门

收稿日期：2025年12月6日；录用日期：2026年1月7日；发布日期：2026年1月15日

摘　要

针对当前卷积神经网络(CNN)在建模长距离依赖上的局限，以及视觉Transformer因自注意力机制导致的参数量庞大问题，本文提出一种双路径混合模型——NexusNet。该模型通过深度融合CNN的局部表示分支与Transformer的全局建模分支，实现了局部细节特征与全局语义信息的协同编码，在显著提升特征表征能力的同时保持了精简的参数量。在CNN分支中，我们引入了融合动态权重分配与上下文增强机制的新型模块，以增强对判别性局部结构的捕捉能力；在Transformer分支中，采用分层建模与线性复杂度设计，大幅降低了长距离依赖建模的资源开销。此外，设计了一种自适应多层次特征融合模块，通过通道与空间注意力引导的多尺度特征整合，实现跨架构信息的高效聚合与参数优化。在两个面部颜色识别数据集上的实验表明，NexusNet在保持模型轻量化的前提下，分类准确率分别达到88.99%和79.25%，并在多项评价指标上优于现有主流方法，验证了其在局部-全局特征融合与模型轻量化方面的有效性与泛化能力。

关键词

图像分类，面部颜色识别，双路径混合架构，特征融合

NexusNet: A Dual-Path Hierarchical Fusion Hybrid Network for Facial Color Recognition

Qianshuai Sun, Yue Feng, Zhuosheng Lin, Jiexin Liang, Xue Zhao, Zihao Liu

School of Electronics and Information Engineering, Wuyi University, Jiangmen Guangdong

Received: December 6, 2025; accepted: January 7, 2026; published: January 15, 2026

Abstract

To address the limitations of Convolutional Neural Networks (CNNs) in modeling long-range

文章引用: 孙千帅, 冯跃, 林卓胜, 梁洁欣, 赵雪, 刘子豪. NexusNet: 一种面向面部颜色识别的双路径层次化融合混合网络[J]. 计算机科学与应用, 2026, 16(1): 154-168. DOI: 10.12677/csa.2026.161013

dependencies and the high parameter complexity of Vision Transformers, this paper proposes a dual-path hybrid model, NexusNet. The model integrates a CNN-based pathway for local feature extraction with a Transformer-based pathway for global context modeling, enabling effective fusion of fine-grained details and semantic information while maintaining model compactness. In the CNN pathway, we introduce a novel module that combines dynamic weight allocation with a context enhancement mechanism to improve discriminative local feature capture. The Transformer pathway employs a hierarchical structure with linear complexity to efficiently model long-range dependencies. Furthermore, we design an adaptive multi-level feature fusion module that leverages both channel and spatial attention to guide the integration of multi-scale features from both architectures, promoting efficient information aggregation. Experimental results on two facial color recognition datasets demonstrate that NexusNet achieves classification accuracies of 88.99% and 79.25%, respectively, and outperforms existing methods across multiple metrics. This validates the model's strong performance and generalization ability in joint local-global representation learning and efficient model design.

Keywords

Image Classification, Facial Color Recognition, Dual-Path Hybrid Architecture, Feature Fusion

1. 引言

在在中医诊断体系中，面色分析历来是望诊的核心环节。临床诊断将面部色泽系统划分为正常色泽与病理色泽两大类别，其中病理色泽又可细分为青、赤、黄、黑、白五种基本类型[1]。医师通过综合观察这些色泽特征，结合其他诊法信息，完成对患者健康状态的评估。然而，传统面色诊断方法长期依赖医师的主观经验，面临着客观化不足的瓶颈。

随着人工智能技术的突破性进展，计算机视觉在医学图像分类领域展现出显著优势。特别是深度学习技术的引入，为面色识别提供了新的技术路径。林怡等人[2]将 AlexNet、VGGNet、ResNet 等经典卷积神经网络架构应用于中医面色分类，取得了 83.96%的识别准确率，证实了深度学习在该领域的应用潜力。赵康辉等人[3]基于 MobileViT 网络构建的面色分类器更是将准确率提升至 94.1243%，展现了轻量级视觉 Transformer 在移动医疗场景中的独特优势。此外，Yang 等人[4]通过 Transformer 架构实现皮肤病变的精准识别，为面色诊断提供了重要的技术借鉴。这些研究表明，将现代人工智能技术与传统中医理论相结合，不仅能有效提升面色识别的客观性和准确性，也为中医诊断的标准化和智能化发展开辟了新的技术路径。

尽管 CNN 与 Transformer 在面色分类中已取得显著成果，但仍存在以下挑战。

1) 传统 CNN 架构在全局上下文建模方面存在局限。由于卷积操作的局部性特征，这类网络难以有效建立长距离依赖关系，同时对输入数据的空间变换缺乏足够的适应能力[5]。

2) Transformer 架构严重依赖位置编码来构建空间关系，其将图像分割为序列的处理方式破坏了原有的空间连续性。此外，自注意力机制参数量随图像分辨率提升呈二次方增长，导致高昂的计算资源需求。

因此，本文旨在设计一种较为轻量化的双路径混合网络 NexusNet，通过实现局部细节与全局语义的协同建模，为面向真实复杂场景的面色识别任务提供一种结构紧凑且性能可靠的解决方案。

2. NexusNet 模型架构

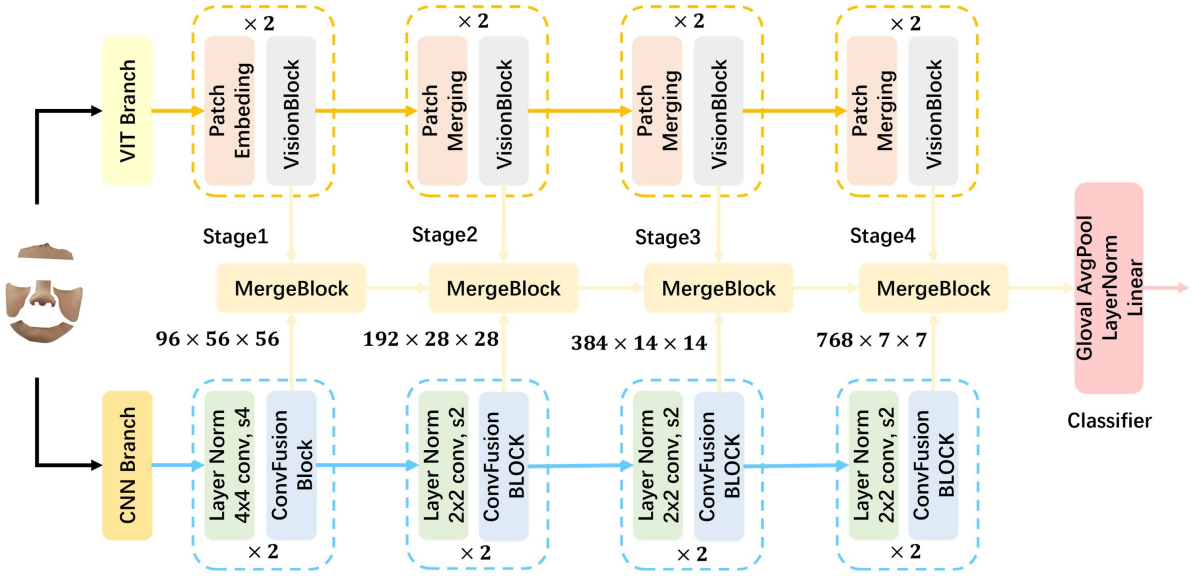

Figure 1. NexusNet network architecture
图 1. NexusNet 网络架构

针对现有技术瓶颈，本文构建了基于 ConvNeXt [6]的局部特征分支与 Swin Transformer [7]的全局特征分支并行架构模型 NexusNet，如图 1 所示。该设计使模型能够同时捕获图像细节特征与长程依赖关系，其中局部分支通过深度可分离卷积与 CTRGC 图卷积[9]增强特征表示能力，全局分支则利用改进的 Swin Transformer V2 [8]模块架构实现跨窗口语义交互。双分支结构在不显著增加参数量的前提下，有效协调了局部感知与全局建模的互补优势。本研究主要具备以下技术优势：

1. 为解决传统 CNN 架构在全局上下文建模和空间变换适应性方面的局限，我们在 CNN 分支中创新设计了卷积融合模块(Convolutional Fusion Module, ConvFusion)。该模块采用深度可分离卷积与 CTRGC 图卷积的并行双路径架构，其中 CTRGC 图卷积专门用于建立长距离依赖关系，有效扩展了传统卷积的感受野范围。同时，模块引入的可学习动态权重分配机制能够自适应调整各路径贡献度，使网络根据不同输入特征自动优化特征提取策略，显著增强了对空间变换的适应能力。无参数 SimAM 注意力机制[10]引入进一步提升了特征表征的判别性，在保持计算效率的同时实现了局部特征与全局上下文的高效融合。这一设计有效克服了传统 CNN 因参数冗余而易出现的过拟合问题。

2. 为解决 Transformer 架构在空间连续性保持方面的固有局限，我们在 Transformer 分支中创新设计了视觉模块(Vision Transformer Block, VisionBlock)。该模块基于改进的 Swin V2 模块架构，采用分层窗口注意力机制替代传统的序列化处理，有效维护了图像的空间连续性。通过引入连续相对位置偏置，该模块在不依赖显式位置编码的情况下保持了精确的空间结构信息，显著降低了模型对位置编码的依赖性。在参数量方面，模块集成的基于对数间隔缩放因子的余弦注意力机制，将计算复杂度从二次方降低至线性级别，大幅减少了对计算资源的需求。同时，归一化层 RMSNorm [11]的采用不仅提升了训练稳定性，还进一步优化了参数效率。结合 SwiGLU 激活的 MLP 感知层增强非线性表征能力，该模块在有限样本条件下实现了长程依赖与局部特征的精准建模，为高分辨率图像处理提供了高效的解决方案。

3. 在 CNN 与 Transformer 之间我们设计了特征融合模块(Feature Merge Block, MergeBlock)，构建了多层次特征融合通路，通过高效多尺度通道注意力与轻量级空间注意力的协同作用，实现了跨模态特征

的精细整合。该模块采用 GhostConv [12]减少参数冗余，集成特征重校准机制动态调整各路径贡献权重，并通过 GhostIRMLP 结构强化特征变换能力。该设计显著提升了模型在复杂场景下数据稀缺时的性能衰减问题。

2.1. 卷积融合模块(Convolutional Fusion Module)

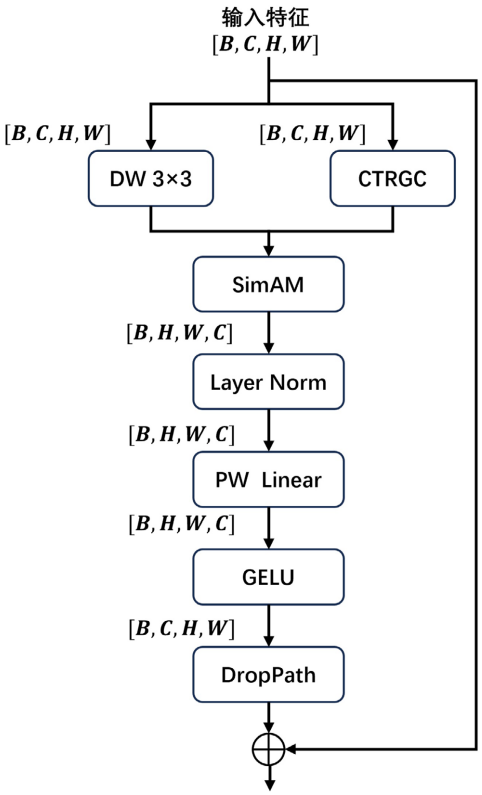

Figure 2. ConvFusion module architecture
图 2. ConvFusion 模块架构

ConvFusion 模块结构见图 2，首先接受输入特征 $x \in \mathbb{R}^{B\times C\times H\times W}$，然后通过的并行双分支架构进行多层次特征提取，其中深度卷积分支利用 3 × 3 深度可分离卷积(DW)以参数高效的方式捕捉局部空间模式和细节特征，同时 CTRGC 图卷积分支通过将特征图重塑为图结构数据并应用通道－空间关系图卷积来建模全局上下文关系和长程依赖，这种双路并进的设计使得模块能够同时从局部细节和全局结构中汲取互补信息，随后两个分支的输出通过基于可学习权重的自适应融合机制进行智能整合，公式如下：

$$x = \text{soft max}(w)[0]\cdot DWConv(x) + \text{soft max}(w)[1]\cdot CTRGC(\text{reshape}(x)) \tag{1}$$

其中 w 为可学习的融合权重向量。该机制通过 Softmax 归一化的融合权重动态平衡局部特征与全局关系的相对重要性，实现了一种自适应的特征选择与增强策略，融合后的特征紧接着通过无参数 SimAM 注意力模块，该注意力机制通过计算每个神经元相对于整个特征图的显著性和能量值，并以 sigmoid 函数激活后作为注意力权重，能够自主地增强信息丰富的特征通道并抑制冗余或噪声响应，公式如下：

$$x_{att} = x \odot \sigma\left(\frac{(x-\mu_x)^2}{4(\sigma_x^2+\epsilon)} + 0.5\right) \tag{2}$$

其中 μ_x 与 σ_x^2 表示特征 x 在空间维度中的均值与方差，σ 表示 Sigmoid 函数。经过注意力训练后的特征随后进入特征精炼阶段，通过通道维度重排、层归一化(LayerNorm)稳定训练过程、线性变换(Linear)进行特征投影以及 GELU 激活函数引入非线性变换，这一系列操作共同完成了特征的深度加工与维度适配，公式如下：

$$x_{out} = permute^{-1}\left(GELU\left(W \cdot LayerNorm\left(permute\left(x_{att}\right)\right)\right)\right) \tag{3}$$

最终，处理后的特征通过带有随机深度正则化(DropPath)的残差连接与原始输入相加，这种设计不仅缓解了深度网络中的梯度消失问题，确保了训练稳定性，还通过随机路径丢弃提供了类似模型集成的正则化效果，使整个模块在保持强大表征能力的同时兼具优秀的泛化性能和训练效率。

2.2. 视觉模块(Vision Transformer Block)

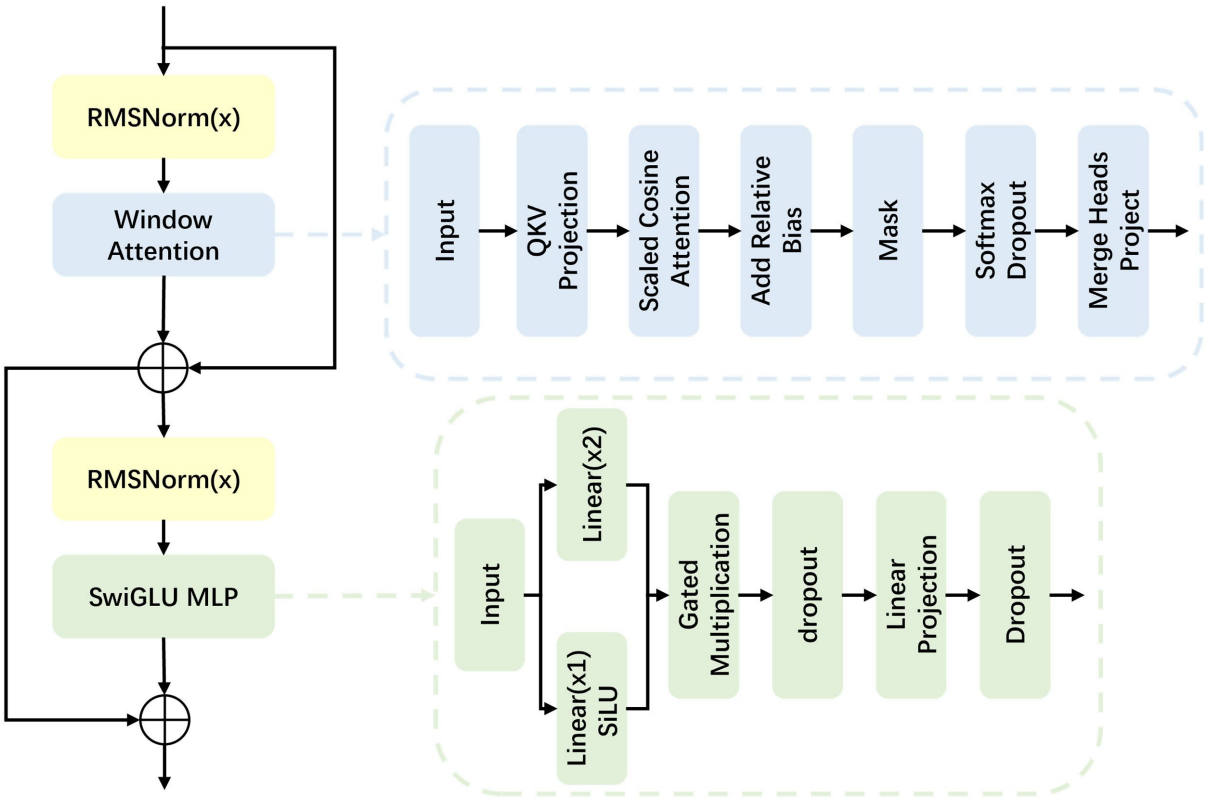

Figure 3. VisionBlock module architecture
图 3. VisionBlock 模块架构

VisionBlock 模块具体结构见图 3，其工作流程起始于输入特征通过 RMSNorm 层处理，计算公式如下：

$$RMSNorm\left(x\right) = \frac{x}{\sqrt{\frac{1}{C}\sum_{i=1}^{C} x_i^2 + \epsilon}} \odot \gamma \tag{4}$$

其中 C 为通道数，γ 为可学习缩放参数，ϵ 为稳定常数。随后归一化后的特征进入窗口注意力机制(Window Attention)进行处理，该机制首先将特征划分为局部窗口并应用基于缩放余弦注意力(Scaled Cosine

Attention)的多头自注意力计算，计算公式如下：

$$Q_{norm}=\frac{Q}{\|Q\|_2},K_{norm}=\frac{K}{\|K\|_2}$$
$$x=\text{soft}\max\left(\min\left(\exp(s),\frac{1}{0.01}\right)\cdot\left(Q_{norm}K_{norm}^{T}\right)+B_{rel}\right) \tag{5}$$

其中，s 为对数缩放因子，B_{rel} 为连续相对位置偏置。该设计是通过查询(Q)和键(K)向量进行 L2 归一化，再通过可学习的对数间隔缩放因子来稳定训练过程，同时融入连续相对位置偏置(Add Relative Bias)以编码空间结构信息，并可选地使用移位窗口策略实现跨窗口连接，随后注意力输出通过残差连接与原始输入相加以保留底层特征信息。接着特征输入到归一化后的增强型 SwiGLU 多层感知机，计算公式如下：

$$SwinGLU(x)=SiLU(W_1x)\odot(W_2x) \tag{6}$$

该结构通过双线性投影生成两个并行特征流并采用 SiLU 激活的门控相乘机制(Gated Multiplication)实现动态特征选择与融合，从而增强非线性表达能力；最终再次通过残差连接整合输出，整个模块融合了局部注意力建模与全局特征变换的优势，结合 RMSNorm 归一化、随机深度丢弃和移位窗口等技术，在保持卓越特征提取能力的同时确保了训练稳定性。

2.3. 特征融合模块(Feature Merge Block)

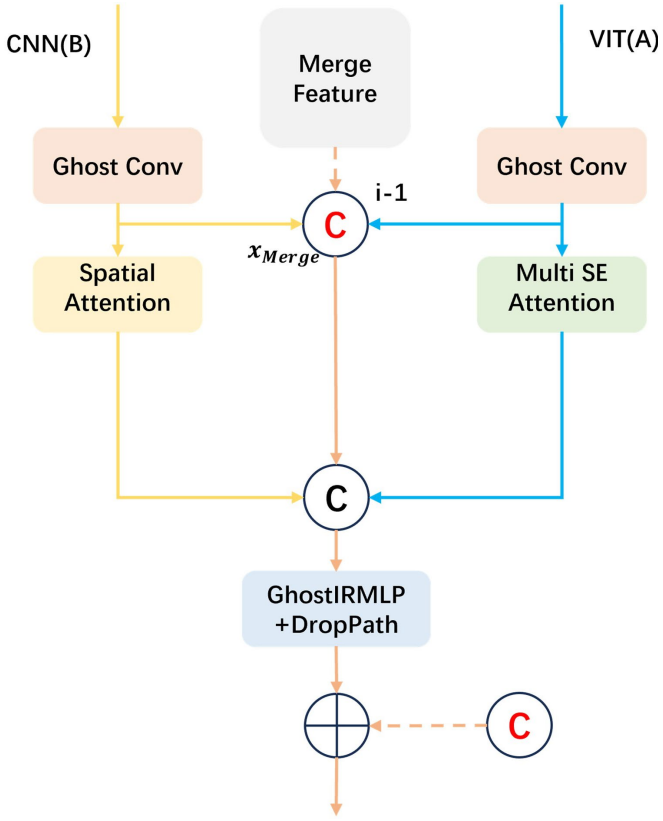

Figure 4. MergeBlock module architecture
图 4. MergeBlock 模块架构

MergeBlock 模块的核心创新在于其层级化融合架构，如图 4 所示，它系统性地集成了 GhostConv、

多尺度通道注意力(MultiSE Attention)、轻量级空间注意力(Spatial Attention)、GhostIRMLP 以及条件残差连接等关键模块。该模块通过三级融合流程进行工作：在早期投影融合阶段，首先接受来自 CNN 分支和 VIT 分支的特征，利用 GhostConv 将不同源的特征映射到统一维度并进行初步交互，计算公式如下：

$$\text{Ghost}Conv(x)=\text{Concat}\left(\text{Primary}Conv(x),\text{Cheap}Conv\left(\text{Primary}Conv(x)\right)\right) \tag{7}$$

其中 Primary*Conv* 生成少量高质量特征，Cheap*Conv* 通过深度可分离卷积等生成大量廉价特征，实现参数量和计算量的显著降低(见图 5(a))。随后进行中期加权融合首先对输入特征分别进行注意力增强，VIT 分支特征通过 MultiSEAttention 处理，其注意力权重生成公式为：

$$A=\sigma\left(Conv_{C/r\to C}\left(\text{Re}LU\left(Conv_{2C/r\to C/r}\left(f_{3\times3}(x)+f_{5\times5}(x)\right)\right)\right)\right) \tag{8}$$

该机制引入 3 × 3 与 5 × 5 两组不同尺度的深度可分离卷积并行提取特征，相对于传统需要高昂资源的通道注意力机制，大幅减少了计算开销。CNN 分支特征通过 SpatialAttention 处理，同样采用多尺度深度可分离卷积。随后，通过可学习的动态权重对这三个特征(原始 A 经通道注意力、原始 B 经空间注意力、以及早期融合结果 x_{Merge} 进行自适应整合，实现特征重校。接着，融合后的特征由 GhostIRMLP 进行非线性变换与精炼，计算公式为：

$$\text{Ghost}IRMLP(x)=Conv_3\left(GELU\left(Conv_2\left(BN\left(Conv_1(x)+x\right)\right)\right)\right)+\text{Proj}(x) \tag{9}$$

Ghost*IRMLP* (见图 5(b))融合了倒残差结构与 Ghost*Conv* 的优势，以较低的计算成本增强了模型的表征能力。最后，在晚期融合阶段，通过条件残差连接将精炼后的特征与来自上层的早期融合结果(当存在上层特征 x_{Merge} 时)，并结合 DropPath 进行正则化，实现了信息的高效保留与梯度稳定传播。该模块的整体设计体现了模块化与自适应处理的核心思想，能够灵活、高效地实现多源、跨尺度的特征融合，并显著降低了计算与存储资源的开销。

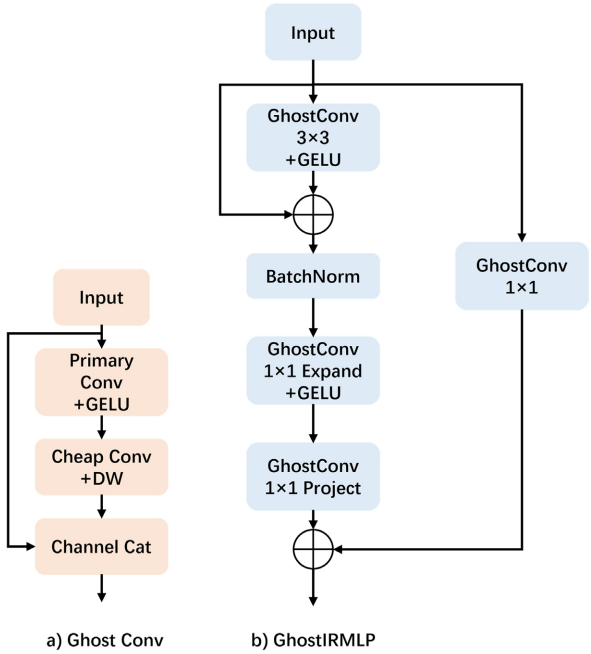

Figure 5. GhostConv and GhostIPMLP module architectures
图 5. GhostConv 与 GhostIPMLP 模块结构

3. 实验

3.1. 数据集

为系统评估 NexusNet 在面色分类任务中的性能，本研究在两个具有不同采集条件的中医面色数据集—Face5c 与 Face3c 上进行了对比实验。所有数据均在专业中医师指导下完成标注，统一遵循红、黄、青、白、黑五类面色分类标准。两个数据集的样本分布情况分别如图 6 与图 7 所示。在实验划分上，采用分层抽样方法将各数据集按 8:1:1 的比例划分为训练集、验证集与测试集，以保证数据分布的均衡性与实验结果的可靠性。

(1) Face5c 数据集

该数据集基于标准化的中医面舌诊仪采集构建。图像采集于严格控制的光照环境：色温维持在 5000~6000 K，显色指数高于 95，照度稳定在 3600 lx 左右。样本来源包括在校学生群体及江门市中心医院的临床志愿者，涵盖红、黄、青、白、黑五类面色数据。具体分布为：面色红样本 238 例，面色黄样本 223 例，面色白样本 54 例，面色黑样本 18 例，面色青样本 10 例，共计 543 例有效数据。

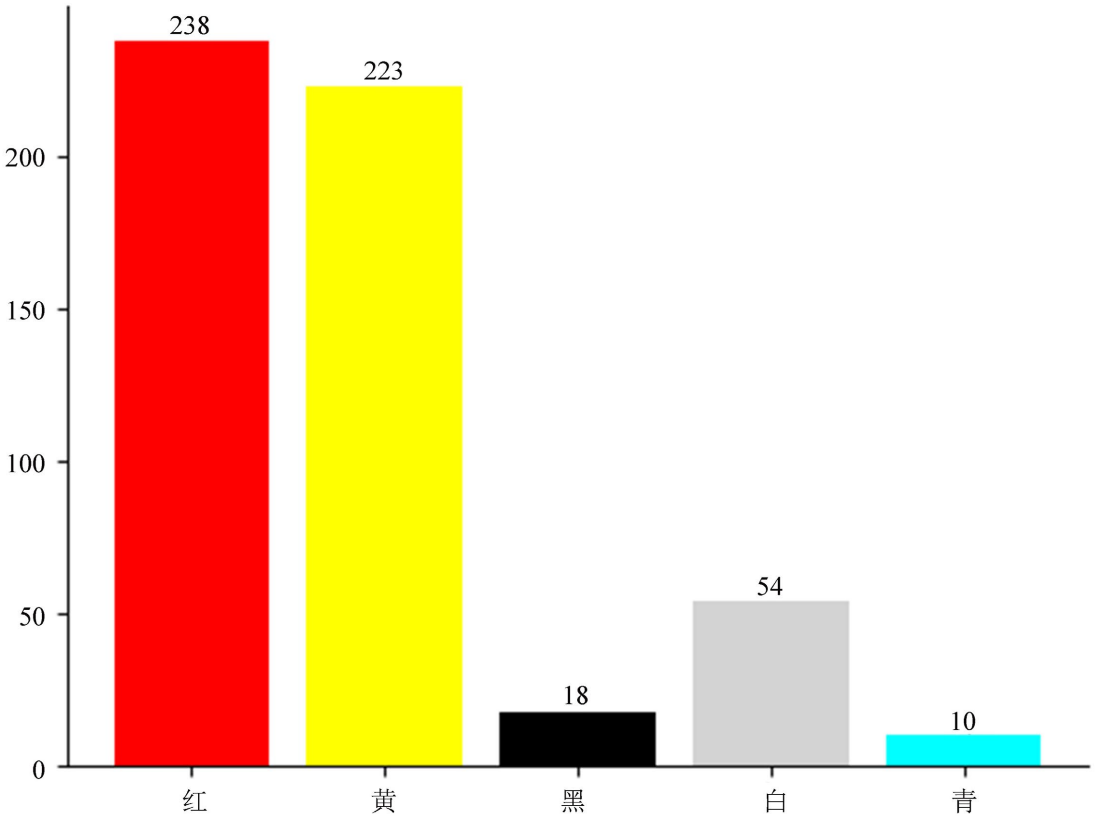

Figure 6. Face5c dataset sample distribution
图 6. Face5c 数据集样本分布

(2) Face3c

Face3c 数据集由上海中医药大学采集构建。与 Face5c 数据集相比，该数据集在面色类别上更为集中，仅包含红、黄、白三类面色样本，未涵盖黑与青两种面色类别。其数据构成如下：面色黄样本 383 例，面色红样本 308 例，面色白样本 116 例，样本总量为 807 例。

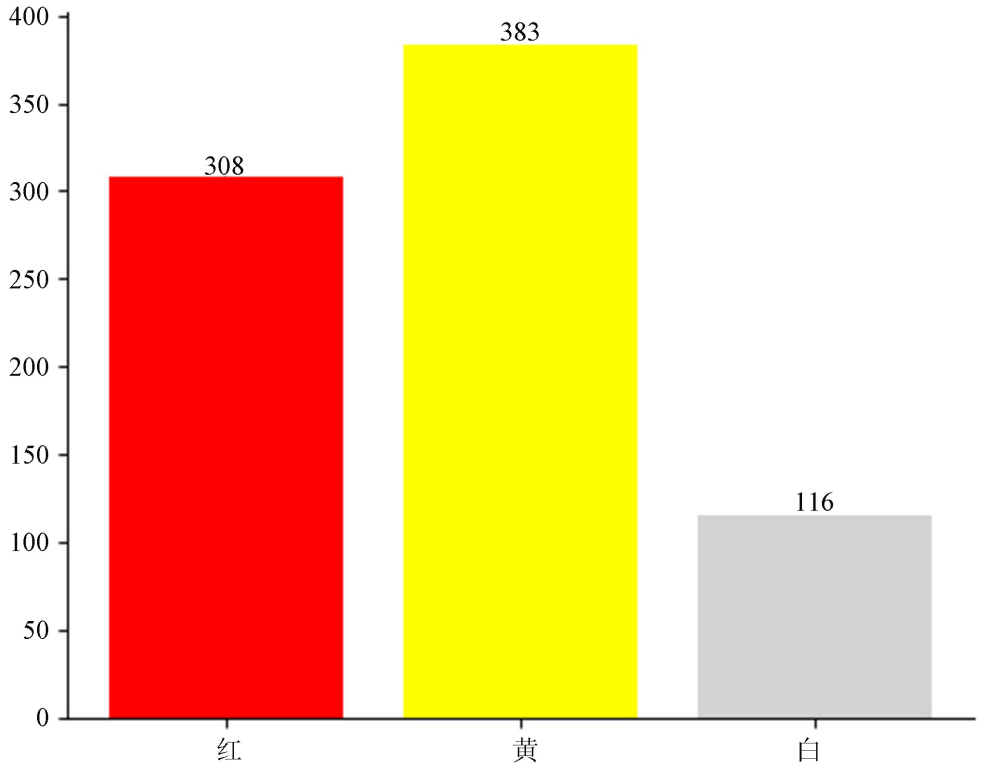

Figure 7. Face3c dataset sample distribution
图 7. Face3c 数据集样本分布

3.2. 实验设置

本实验在 Windows 11 系统下基于 PyTorch 框架进行，使用 NVIDIA RTX A5000 GPU 进行模型训练。输入图像统一预处理为 224 × 224 像素，训练阶段使用随机裁剪和水平翻转进行数据增强。使用 AdamW 优化器进行 300 个 epoch 的训练，初始学习率为 1e-4 并配合余弦退火学习率调整策略，同时通过 TensorBoard 对训练过程中的损失、准确率及学习率变化进行实时监控与记录。

3.3. 评估指标

为量化评估模型在分类任务中的性能，本文选用准确率(Accuracy，Acc)、精确率(Precision，AP)、召回率(Recall，AR)以及特异度(Specificity，AS)四项指标。各指标基于混淆矩阵中的真阳性(TP)、真阴性(TN)、假阳性(FP)与假阴性(FN)进行计算(见公式 33)。准确率反映模型整体分类的正确比例，但在类别分布不均衡的任务中，该指标容易因多数类样本主导而虚高，导致对少数类识别性能的判断失准，因此仅依靠准确率评价模型性能具有较大局限性。

$$\begin{aligned}\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}\\ \text{Precision} &= \frac{TP}{TP+FP}\\ \text{Recall} &= \frac{TP}{TP+FN}\\ \text{Specificity} &= \frac{TN}{TN+FP}\end{aligned} \tag{10}$$

3.4. 损失函数

本实验采用了交叉熵损失函数(Cross-Entropy Loss)，该函数直接优化预测概率与真实标签的分布差异，为分类任务提供了清晰且稳定的梯度信号，能够有效驱动模型快速收敛并提升分类准确率。

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{k} y_{n,i}\log\left(\hat{y}_{n,i}\right) \tag{11}$$

N 为批次中样本总数，k 为类别数量，$y_{n,i}$ 为第 n 个样本在类别 i 上真实标签，$\hat{y}_{n,i}$ 为模型预测第 n 个样本属于类别 i 的概率。

4. 实验结果与分析

4.1. 定量分析

为全面评估模型在面部颜色识别任务中的性能，本研究将所提出的 NexusNet 与当前主流分类模型进行系统对比，涵盖 VGG [13]、ResNet [14]、DenseNet121 [15]、ConvNeXt [6]、Vision Transformer [16]、Swin Transformer [7] [8]、MobileViT [3]以及 HiFuse [17]等方法。实验结果表明，凭借其创新的双路径混合架构设计，NexusNet 在 Face3c 与 Face5c 两个数据集上均实现了全面领先的性能表现，详细量化结果如表 1 与表 2 所示。

为进一步分析各模型的分类依据与特征响应模式，我们针对两个数据集绘制了类别激活热力图。为清晰呈现对比，相同架构系列中选取性能最优的模型进行可视化，不同模型间的热力图对比如图 8 所示。该可视化结果有助于直观理解模型在面色识别任务中的注意力分布与决策机制。

与传统 CNN 相比，其依赖局部卷积操作、难以建模长距离依赖的固有限制，在面对需要全局上下文理解的面色识别任务时表现出了明显短板。例如，VGG 系列虽结构规整，但其参数量巨大且性能平庸，在 Face5c 数据集上准确率(Acc)仅约 60%~65%；ResNet 与 DenseNet 通过残差或密集连接缓解了梯度问题，在 Face3c 上取得了 80%以上的 Acc，但其纯卷积架构在全局语义整合上仍存在瓶颈，导致在更复杂的 Face5c 上性能提升有限，且热力图显示其注意力区域常存在背景干扰。ConvNeXt 在 Face5c 上的精确率(AP)更是大幅降至 39.46%，凸显了传统卷积范式在复杂场景下的适应不足。NexusNet 通过其局部分支的 ConvFusion 模块，创造性地融合了深度卷积与 CTRGC 图卷积，在提取局部细节的同时显式构建了跨区域的上下文关联，从而有效克服了传统 CNN 的感受野局限。这使得 NexusNet 在 Face3c 和 Face5c 上均取得了最高的准确率(分别为 88.99%和 79.25%)，其热力图也展现出对面部核心区域更精准、集中的聚焦能力，背景抑制效果显著。

与纯视觉 Transformer 模型相比，其依赖大规模预训练且在处理小样本数据时泛化能力不稳定的问题在本任务中十分明显。标准的 ViT-B/16 模型在从 Face3c 迁移到 Face5c 时，精确率(AP)从 78.78%急剧下降至 42.14%；采用层次化设计的 Swin Transformer 系列虽有所改善，但 Swin V2 在 Face5c 上的召回率(AR)也仅为 48.55%。热力图显示，这些模型的注意力响应常呈零散、碎片化分布，难以形成对目标整体连贯的理解。NexusNet 的全局分支 VisionBlock 模块，基于改进的 Swin V2 架构，采用线性复杂度的缩放余弦注意力和连续相对位置偏置，在高效捕获长程依赖的同时，大幅降低了对海量数据和超高算力的依赖。因此，NexusNet 在数据分布更具挑战性的 Face5c 上仍能保持 74.55%的高精确率和 69.69%的召回率，其热力图也表现出对目标区域更稳定、完整的关注。

与现有的先进混合架构相比，如轻量级的 MobileViT 和特征融合网络 HiFuse，它们在试图结合 CNN 与 Transformer 优势的同时，往往在参数效率、性能均衡性或泛化稳健性上做出了新的妥协。MobileViT 虽专为移动端设计，但在本任务的数据集上表现不佳，在 Face5c 上的 AP 仅为 44.98%，其热力图虽然红

色显著区域显著，但常分散于目标边缘或目标与背景交界之处，定位精度不足。HiFuse 在 Face3c 上取得了最高的单一精确率指标(AP 89.62%)，但其参数量高达 123.26M，且当面对类别更复杂的 Face5c 时，AP 指标大幅下降 25.28%，显示出明显的过拟合倾向和泛化短板，并且其热力图中所示显著响应区域多出现在图像边缘。与之形成鲜明对比的是。NexusNet 通过其自适应多层次特征融合模块(MergeBlock)，以多尺度注意力机制和 Ghost 卷积等轻量化技术，实现了双路径特征的高效、自适应融合。在将总参数量显著控制在 73.37M(远低于 HiFuse)的前提下，NexusNet 不仅在 Face3c 上获得了最优的综合准确率(Acc)、召回率(AR)和特异度(AS)，更在 Face5c 上展现了卓越的泛化稳定性，各项性能指标下降幅度最小，热力图始终能准确锁定核心特征并有效排除干扰。

综上所述，定量指标与可视化证据共同验证了 NexusNet 架构的有效性。它通过 ConvFusion 模块增强了传统 CNN 的上下文建模能力，通过 VisionBlock 模块降低了纯 Transformer 的数据与计算依赖，并通过 MergeBlock 模块以更高参数效率实现了比现有混合模型更优的性能均衡与泛化鲁棒性。这使其为数据规模有限、应用场景复杂的真实世界面色识别任务，提供了一个高效且可靠的解决方案。

Table 1. Comparison of experimental results on the Face3c dataset
表 1. Face3c 数据集上的实验结果对比

Network	Acc (%)	AP (%)	AR (%)	AS (%)	Parameters (M)
VGG11	80.37	73.89	74.95	88.36	128.78
VGG13	79.36	74.39	77.69	89.95	134.27
VGG16	82.65	76.25	78.96	88.24	139.58
ResNet34	85.66	81.67	80.63	89.87	21.56
ResNet50	85.78	82.29	79.56	89.90	22.65
DenseNet121	83.26	81.95	80.36	89.25	8.95
ConvNeXt	77.22	73.99	74.19	87.72	87.57
ViT-B/16	81.01	78.78	81.6	90.16	85.80
ViT-B/32	82.41	76.32	77.76	90.35	86.75
Swin V1	82.28	79.41	78.64	90.52	86.45
Swin V2	83.36	80.21	77.95	90.45	87.86
MobileViT	86.08	81.39	78.88	91.42	5.35
HiFuse	86.08	**89.62**	82.96	91.72	123.26
NexusNet	**88.99**	87.39	**83.31**	**92.04**	73.37

Table 2. Comparison of experimental results on the Face5c dataset
表 2. Face5c 数据集上的实验结果对比

Network	Acc (%)	AP (%)	AR (%)	AS (%)	Parameters (M)
VGG11	60.48	49.89	49.59	79.48	127.74
VGG13	62.45	51.49	50.68	80.46	133.17
VGG16	64.69	57.95	52.36	81.25	135.28
ResNet34	73.83	74.06	**73.83**	91.74	21.34
ResNet50	76.42	74.35	62.09	92.29	23.52

续表

DenseNet121	77.36	54.06	56.39	92.76	6.96
ConvNeXt	73.58	39.46	35.72	90.89	88.25
ViT-B/16	73.45	42.14	41.93	91.1	84.95
ViT-B/32	72.64	39.55	38.36	90.74	87.46
Swin V1	75.47	42.13	42.81	92.05	86.75
Swin V2	76.95	45.28	48.55	91.37	88.26
MobileViT	78.30	44.98	50.39	91.99	6.21
HiFuse	78.26	64.34	59.51	92.72	123.27
NexusNet	**79.25**	**74.55**	69.69	**92.77**	73.37

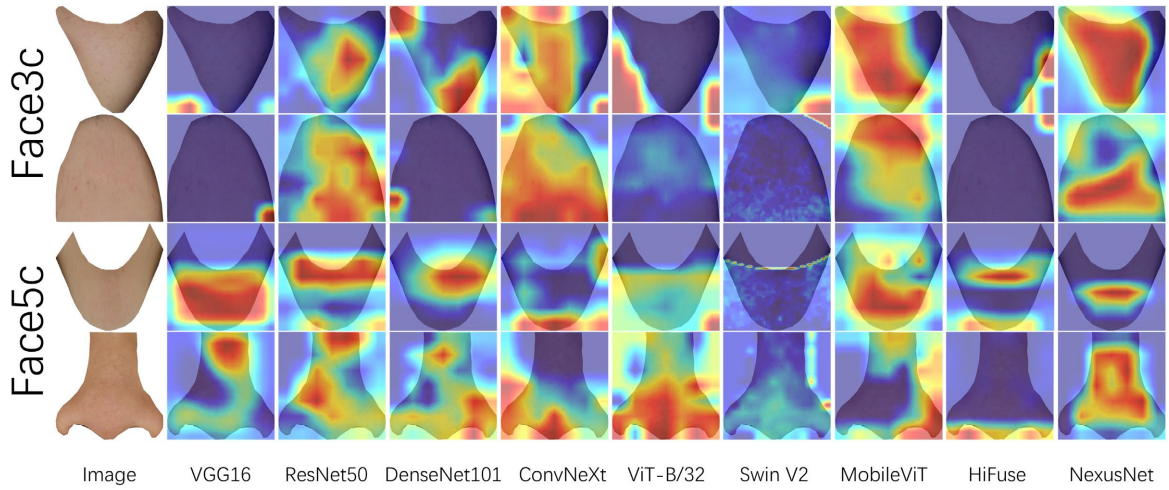

Figure 8. Comparison of heatmaps for different model classifications
图 8. 不同模型分类热力图对比

4.2. 消融实验

为了系统验证 NexusNet 各模块如何协同解决模型设计的核心挑战——即传统 CNN 的远程建模不足与 Transformer 的高计算依赖问题，我们在 Face5c 与 Face3c 数据集上进行了循序渐进的消融研究。结果如表 3 与表 4 所示。

实验始于一个仅具备基础特征融合能力(MergeBlock)的基准模型。在更复杂、类别不均衡的 Face5c 数据集上，该模型取得了 73.90%的准确率，但其较低的精确率(50.26%)与召回率(46.95%)表明，未增强的融合框架对复杂特征的建模能力有限。而在 Face3c 数据集上，其基础性能(Acc 77.95%, AP 77.59%)相对更好，印证了模型具备初始有效性，同时也凸显了性能的上升空间。引入 GhostIRMLP 模块后，Face5c 的准确率升至 74.25%，精确率与召回率分别提升 1.42%与 2.52%，Face3c 的准确率也提高至 79.19%。这表明该模块通过高效的 Ghost 卷积与残差结构，以较低的参数成本增强了特征的非线性表达能力，为模型后续的复杂特征处理提供了更有效的基础。随后加入的 MultiSEAttention 与 SpatialAttention 机制带来了更显著的提升。在 Face5c 上，精确率从 51.68%大幅跃升至 60.59%，召回率也同步增长；Face3c 的召回率则从 77.58%提升至 80.99%。这一变化证明，双重注意力机制通过自适应地聚焦于关键通道与空间区域，有效增强了模型在复杂场景中筛选判别性特征、抑制背景干扰的能力，显著提升了特征选择的鲁棒性。

之后 CNN 分支嵌入的 ConvFusion 模块产生了关键性突破。该模块在 Face5c 上推动准确率大幅增长至 78.95%，精确率更是从 60.59%跃升至 73.65%；在 Face3c 上，准确率也从 80.34%提升至 84.49%。这一跨越式进步直接证实，通过融合深度卷积与 CTRGC 图卷积并引入动态权重分配，ConvFusion 有效建立了长距离特征依赖，从根本上增强了 CNN 路径的上下文建模能力，解决了传统卷积网络在复杂识别任务中的核心短板。最后，Vision Transform 分支引入改进的 VisionBlock 模块进一步优化了全局建模的稳定性与效率。在 Face3c 数据集上，其贡献最为突出，将准确率从 84.49%最终提升至 88.99%，实现了最优性能；在 Face5c 上，模型性能也得到进一步巩固，准确率达到 79.25%，精确率提升至 74.55%。这表明该模块基于层次化窗口注意力与线性复杂度的设计，在显著降低计算开销的同时，有效保障并完善了模型对全局语义的理解与整合。

消融实验清晰地揭示，NexusNet 的卓越性能源于其各组件针对性的协同设计：ConvFusion 解决了 CNN 的远程建模局限，VisionBlock 实现了 Transformer 的高效化，双重注意力与 GhostIRMLP 则分别强化了特征选择与变换的效率。所有这些创新最终通过 MergeBlock 有机整合，使得模型在严格的控制参数量下，实现了局部细节与全局语义的高效平衡与协同。这不仅验证了双路径混合架构思想的正确性，也完整展示了 NexusNet 如何系统性、递进式地攻克了当前视觉识别模型面临的关键挑战。

Table 3. Ablation experiments on the Face5c dataset
表 3. Face5c 数据集上消融实验

	Acc (%)	AP (%)	AR (%)	AS (%)
MergeBlock	73.90	50.26	46.95	89.26
+ GhostIRMLP	74.25	51.68	49.47	90.01
+MultiSEAttention and SpatialAttention	75.85	60.59	52.85	90.58
+ ConvFusion	78.95	73.65	69.36	92.69
+ VisionBlock	**79.25**	**74.55**	**69.69**	**92.77**

Table 4. Ablation experiments on the Face3c dataset
表 4. Face3c 数据集上消融实验

	Acc (%)	AP (%)	AR (%)	AS (%)
MergeBlock	77.95	77.59	76.86	88.65
+ GhostIRMLP	79.19	79.68	77.58	89.17
+MultiSEAttention and SpatialAttention	80.34	80.25	80.99	90.24
+ ConvFusion	84.49	83.26	82.21	91.05
+ VisionBlock	**88.99**	**87.39**	**83.31**	**92.04**

5. 结语

本研究提出了一种融合局部感知与全局建模优势的双分支视觉架构 NexusNet，旨在协同解决卷积神经网络在长距离依赖建模上的局限性与视觉 Transformer 因自注意力机制导致的参数量过大的问题。该模型通过引入层次化特征融合机制，将 CNN 的局部细节提取能力与基于窗口的 Transformer 的全局依赖建模能力有机结合。具体而言，其局部分支通过 ConvFusion 模块集成 CTRGC 图卷积与 SimAM 注意力，增强了对多尺度局部特征的表达能力与空间适应力；全局分支借助 VisionBlock 模块引入 Swin V2 的缩放

余弦注意力、连续相对位置偏置及 SwiGLU MLP，在显著降参数量的同时保持了对长程上下文的高效建模；最后，通过分层级联的 MergeBlock 模块对双分支特征进行自适应加权融合，实现了从底层细节到高层语义的渐进式信息整合。

实验结果表明，NexusNet 在图像分类任务上展现出优越性能。该模型通过并行双路结构与三级融合策略，在显著减少模型参数量的同时，保持了优异的表征能力，从而有效提升了资源利用率；其可学习的动态融合权重与空间－通道协同注意力机制，进一步增强了特征选择的适应性、鲁棒性以及对关键信息的聚焦能力。NexusNet 的设计不仅为异构视觉特征融合提供了一种高效且可扩展的解决方案，也为面向实际部署的、资源受限的视觉模型研究提供了有价值的架构参考。

基金项目

广东省普通高校重点领域专项项目(2021ZDZX1032)；广东省国际及港澳台高端人才交流专项(2020A1313030021)；五邑大学科研项目(2018GR003)。

参考文献

[1] Liu, C., Zhao, C., Li, G., Li, F. and Wang, Z. (2013) Computerized Color Analysis for Facial Diagnosis in Traditional Chinese Medicine. 2013 *IEEE International Conference on Bioinformatics and Biomedicine*, Shanghai, 18-21 December 2013, 613-614. https://doi.org/10.1109/bibm.2013.6732569

[2] 林怡, 王斌, 许家佗, 等. 基于面部图像特征融合的中医望诊面色分类研究[J]. 实用临床医药杂志, 2020, 24(14): 1-5.

[3] Zhao, K., Ma, X., Kuang, H. and Liu, X. (2024) Facial Complexion Classification of Traditional Chinese Medicine Based on Statistical Features and MobileViT. 2024 *IEEE 7th Information Technology*, *Networking*, *Electronic and Automation Control Conference* (*ITNEC*), Chongqing, 20-22 September 2024, 50-54. https://doi.org/10.1109/itnec60942.2024.10733073

[4] Yang, G., Luo, S. and Greer, P. (2023) A Novel Vision Transformer Model for Skin Cancer Classification. *Neural Processing Letters*, **55**, 9335-9351. https://doi.org/10.1007/s11063-023-11204-5

[5] Rangel, G., Cuevas-Tello, J.C., Nunez-Varela, J., Puente, C. and Silva-Trujillo, A.G. (2024) A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks. *Journal of Sensors*, **2024**, Article ID: 2797320. https://doi.org/10.1155/2024/2797320

[6] Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. and Xie, S. (2022) A ConvNet for the 2020s. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11966-11976. https://doi.org/10.1109/cvpr52688.2022.01167

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

[8] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., *et al.* (2022) Swin Transformer V2: Scaling up Capacity and Resolution. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 11999-12009. https://doi.org/10.1109/cvpr52688.2022.01170

[9] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y. and Hu, W. (2021) Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 13339-13348. https://doi.org/10.1109/iccv48922.2021.01311

[10] Yang, L., Zhang, R.Y., Li, L., *et al.* (2021) SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Network. *International Conference on Machine Learning PMLR*, 18-24 July 2021, 11863-11874.

[11] Zhang, B. and Sennrich, R. (2019) Root Mean Square Layer Normalization. arXiv: 1910.07467.

[12] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[13] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[14] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. 2017

IEEE Conference on Computer Vision and Pattern Recognition (*CVPR*), Honolulu, 21-26 July 2017, 2261-2269. https://doi.org/10.1109/cvpr.2017.243

[15] Dosovitskiy, A. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[16] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C. and Xu, C. (2020) GhostNet: More Features from Cheap Operations. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 1577-1586. https://doi.org/10.1109/cvpr42600.2020.00165

[17] Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., *et al.* (2024) HiFuse: Hierarchical Multi-Scale Feature Fusion Network for Medical Image Classification. *Biomedical Signal Processing and Control*, **87**, Article ID: 105534. https://doi.org/10.1016/j.bspc.2023.105534