Swin 与 ViT 的层次化协同

——中草药图像细粒度分类的动态融合框架

徐世倩　曹昕怡　何淳妍　梁炜盈　郎六琪　林刚 *
珠海科技学院 计算机学院　广东省珠海市　519040

摘　要：精准识别中草药一直是中草药领域的一个挑战。尽管 CNN 和 ViT 等模型在植物识别中占主导地位，但它们在捕捉细节和结构方面存在不足，CNN 难以建模长距离依赖关系，而 ViT 因全局注意力机制计算复杂度且需要大量数据训练导致小规模数据分类受限。本文提出一种结合 Swin-Transformer 和 ViT-Transformer 的双分支融合模型，利用局部窗口注意力和全局自注意力的互补特性，并采用冻结 ViT 浅层参数的优化策略，有效降低计算成本。该模型旨在解决细粒度植物分类问题，为中草药识别提供高效模型。

关键词：中草药图像识别；细粒度分类；Swin-Transformer 模型；ViT-Transformer 模型；特征融合

引言

在中药领域，药材混淆可能引起安全问题，如人参与西洋参误用导致毒性反应。由于外观相似，传统方法难以区分，需要细粒度分类技术。比如人参与西洋参，它们的叶子纹理和整体外观几乎一模一样，传统图像分类技术很难捕捉到这些细微差异。然而，现有的卷积神经网络（CNN）在处理细粒度分类任务时，常因捕捉局部特征能力不足，出现错误分类，或者计算效率不高的情况 [4]。同样，ViT 能捕捉全局信息，但计算复杂且依赖大量数据。

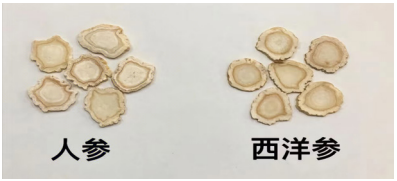

图 1 人参与西洋参对比图

植物识别传统上依赖专家，受限于主观性和局限性。计算机视觉方法，特别是深度学习技术如 CNN 和 ViT 因此受到关注。调查显示 ViT-Transformer 和 Swin Transformer 在图像分类中表现突出。ViT 通过图像分割和全局自注意力提高精度，而 Swin 通过局部窗口自注意力和移位窗口策略提升效率，尤其在处理高分辨率图像时更为灵活。然而，视觉 Transformer（ViT）虽能捕获全局信息，但在处理细粒度分类任务，尤其是数据量有限时，表现也不太理想 [1]。在细粒度分类任务中，Swin Transformer 表现突出，能精确区分相似植物品种。和传统卷积神经网络相比，Swin Transformer 在细粒度分类任务优势显著 [3]。

本研究提出结合 Swin Transformer 和 ViT-Transformer 的深度学习架构，用于中草药图像分类。Swin Transformer 通过局部窗口和多尺度特征技术，减少计算负荷，加快处理速度，有效识别细粒度特征。与传统的卷积神经网络（CNN）和视觉变换器（ViT）相比，Swin Transformer 和 ViT-Transformer 进行特征融合在获取图像细节和全局结构方面能力卓越，对辨别外观相似但类别不同的中草药优势明显 [2]。本研究通过数据增强、算法优化和错误分析，提高了模型泛化和分类精度。改进模型结构和训练方法，比较了 Swin-Transformer 和 ViT-Transformer 在中草药图像上的性能，并融合 ViT 特征于 Swin 中，旨在更精确识别植物种类，解决细粒度植物分类问题。

本文的创新主要在：一是采用双分支并行架构，即 Swin 分支和 ViT 分支协同工作，以提取多尺度特征；二是通过动态特征对齐，利用交叉注意力模块，实现局部与全局信息的加权融合。

1. 实验设计与方法

1.1. 数据集选择

本研究构建了一个中草药图像识别数据集，用于评估 Swin Transformer 与 ViT-Transformer 模型融合在细粒度分类中的效果。数据集包含 60 类植物，共 6001 张图片，每类约

100 张。为增加多样性，我们收集了在不同照明、视角和背景下的中草药照片。

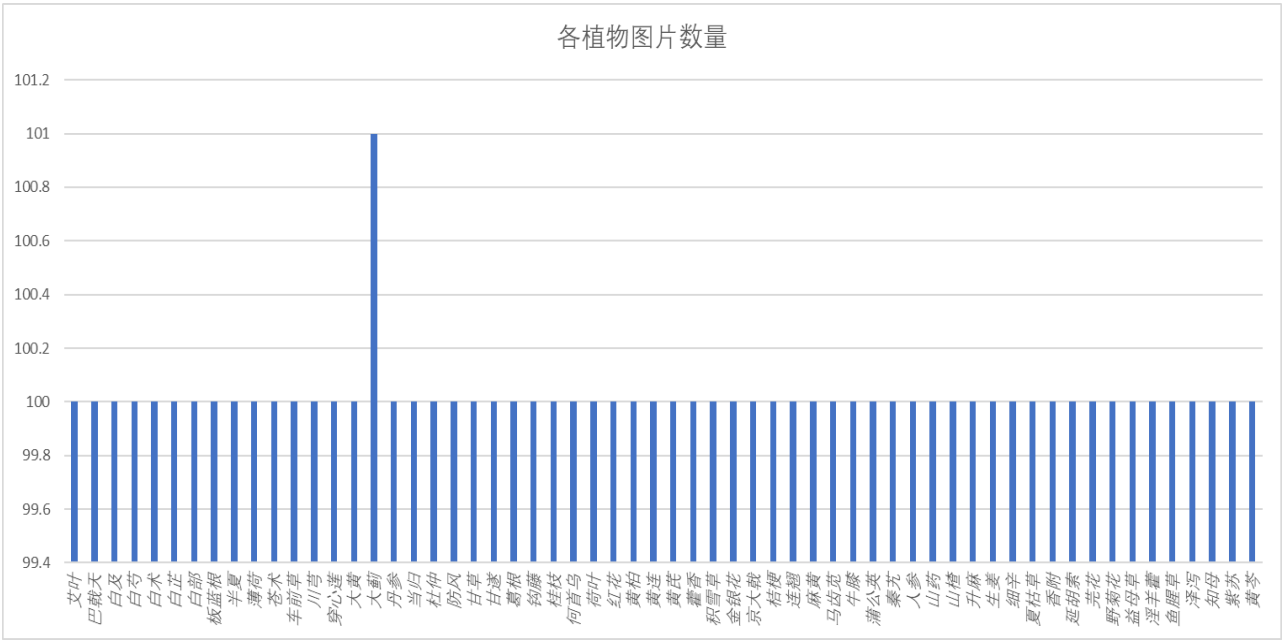

图 2 各植物图片数量条形图

数据集图像经过人工标注，包括植物种类和图像质量，如清晰度、光照等。每种植物有唯一类别标签，确保标注准确一致，提高模型分类精度。数据加载时提供图像路径和标签，检查图像是否为 RGB 格式，确保模型顺利输入。

1.2. 模型架构定义

Swin Transformer 是一种基于 Transformer 的视觉模型，采用层次化设计，将输入图片分割成不重叠的小块。通过局部窗口自注意力机制（W-MSA）和移位窗口自注意力（SW-MSA），逐层提取图片特征 [2]。该模型与 ViT 模型协同工作，ViT 通过全局自注意力机制处理大块图像，而 Swin Transformer 通过局部自注意力机制处理小块图像。图像被分割成“Patches”，每个 Patch 通过“Patch Embedding”转化为特征向量，然后输入模型处理。为了提升分类效果，采用多层 Patch Merging 层合并相邻的 Patch，降低空间分辨率，增加通道数，保留关键特征。模型包含 4 个阶段，每个阶段有多个 Swin Transformer Block，交替使用 W-MSA 和 SW-MSA，增强模型提取图像特征的能力，提高对中草药细微特征的区分能力。具体架构图显示如图 3：

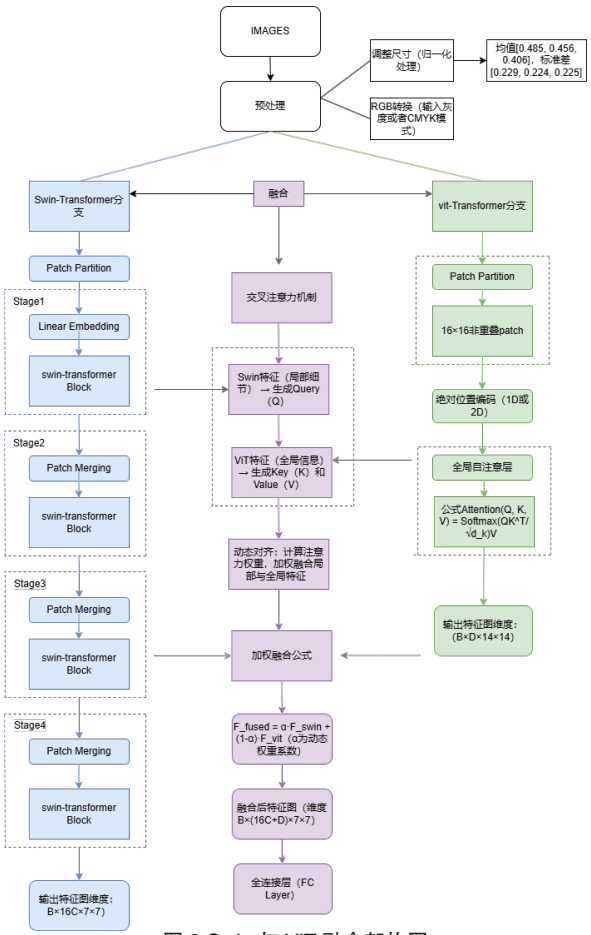

图 3 Swin 与 ViT 融合架构图

1. 图像大小调整与归一化

首先调整图像大小并归一化以适配 Swin Transformer 和 ViT 模型。输入图像可能需从灰度或 CMYK 转换为 RGB 格式。为加快模型收敛和训练稳定性，使用 ImageNet 预训练权重的均值 [0.485, 0.456, 0.406] 和标准差 [0.229, 0.224, 0.225] 进行归一化，帮助模型识别关键特征。

2. 图像增强

为增强模型泛化能力，数据预处理中采用随机裁剪、缩放和水平翻转技术。训练时，随机裁剪和缩放图像，丰富数据集，使模型适应多尺寸和比例的图像，减少过拟合。此外，随机水平翻转图像（概率 50%）增加样本多样性，提高模型对旋转和不同视角的适应性。

3. 批量处理

数据加载时，将预处理好的图像批量堆叠成张量，批量送入模型训练或预测。这提高了加载效率，实现模型并行处理，加速训练过程。

2. 实验运行与结果分析

2.1. 实验流程

实验流程涵盖数据准备、模型构建、训练、评估、预测和部署。

数据准备阶段，我们用 read_split_data 函数将数据集分为训练集和验证集，创建图像路径和标签列表。类别映射保存在 class_indices.json 文件中。训练集经过随机裁剪、水平翻转和归一化处理，验证集则使用中心裁剪和归一化。数据加载通过 MyDataSet 类封装，用 DataLoader 批量加载，支持多线程加速。

构建和训练模型时，采用 Swin-Transformer 和 ViT-Transformer 两种模型。Swin-Transformer 有两种结构，输入尺寸为 224x224 和 384x384。ViT-Transformer 基于 Hugging Face 的 ViTForImageClassification，图像分割为 16x16 Patch。两种模型加载预训练权重，部分层冻结。训练采用交叉熵损失函数，优化器为 AdamW，学习率为 2e-6，批大小为 8，权重衰减为 1e-2。记录每个 epoch 的训练验证损失和准确率，保存模型权重至 ./weights。运用了早停机制，防止模型在训练过程中过拟合同时还优化训练的效率。

模型评估和可视化阶段，通过 evaluate 函数计算验证集的损失和准确率，生成混淆矩阵，输出精度、召回率、特异性。提取预测错误的图像路径及标签，保存到 record.txt 中。训练曲线（损失和准确率）通过 draw 函数绘制并保存为图片，TensorBoard 记录训练过程的关键指标。

预测和部署阶段，采用单张预测方法，加载模型进行图像预处理，输出类别和置信度。批量预测测试集图像，并通过 main 函数构建界面，支持图片上传、预测结果展示及药用功效等详细信息。

表 1 Swin-Transformer 与 ViT-Transformer 对比

特性	Swin-Transformer	ViT-Transformer
注意力机制	局部窗口注意力 + Shifted Window	全局自注意力
计算复杂度	O(4hwC2)（窗口划分降低计算量）	O(N2)（N 为 Patch 数，复杂度高）
数据需求	中等，适合小数据集	需要大规模预训练数据
局部特征提取	通过分层结构和窗口机制保留局部细节	依赖全局注意力，可能忽略局部纹理
位置编码	相对位置偏置（Relative Position BiaS）	绝对位置编码（1D 或 2D）
适用场景	高分辨率图像（如 384x384）	中等分辨率，数据量充足时表现更优

2.2. 结果呈现

实验显示，结合 Swin Transformer 和 ViT 的模型在特征学习和分类性能上优于单独的 Swin Transformer。该模型训练准确率 100%，验证准确率约 90%，泛化能力强。训练和验证损失分别降至 0.001 和 0.373，低于单一模型，表明融合模型分类功能更优。融合模型在提取复杂纹理特征和区分类别差异方面表现突出，有效改善了 Swin Transformer 在处理跨尺度特征关联上的不足，特别是在中草药叶片脉络和花色分布等细粒度特征建模方面。融合模型在训练初期即展现出更强的特征判别能力，验证准确率始终高于单一模型，证明了结构改进对分类性能的正面影响。

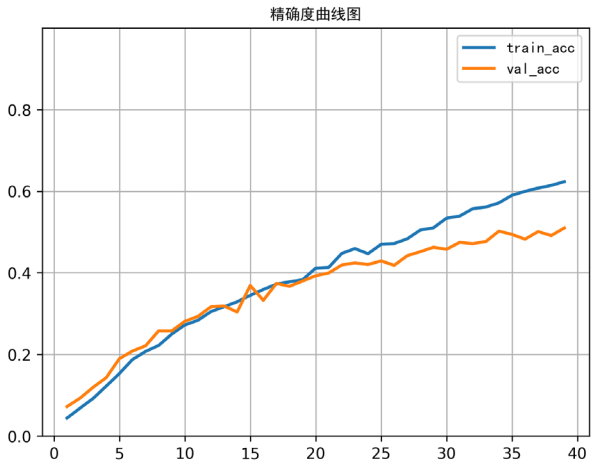

图 4 单一模型精确度曲线

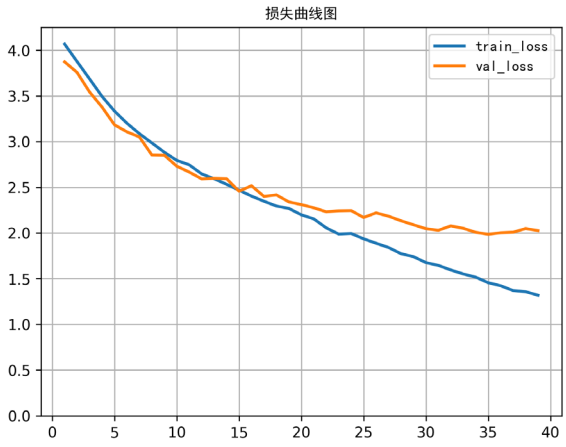

图 5 单一模型损失曲线图

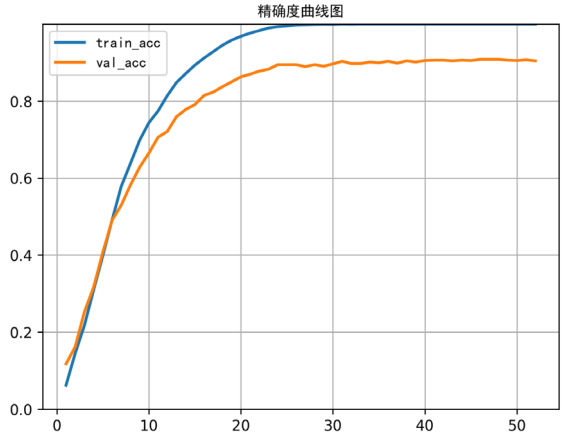

图 6 融合模型精确度曲线图

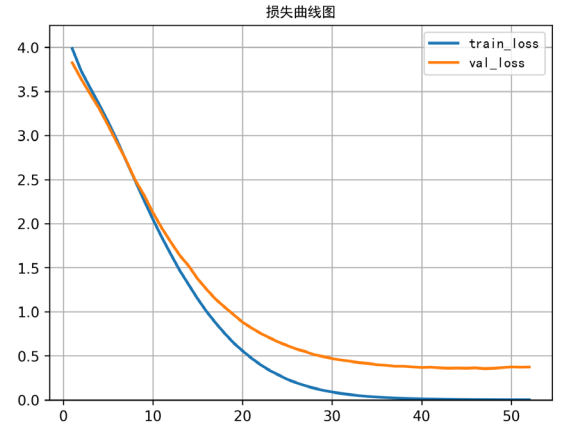

图 7 融合模型损失曲线

3. 结果讨论

3.1. 结果解读

中草药图像分类中，Swin Transformer 与 ViT 融合模型准确率提升至 100%，证明互补策略有效。ViT 的全局注意力和 Swin 的局部窗口机制共同提高分类性能。尽管验证损失较高，但训练损失极低，且验证损失曲线后期稳定，说明模型优化过程稳定。

3.2. 比较分析

Swin Transformer 以局部窗口机制捕捉局部特征，ViT 以全局注意力机制捕捉全局信息。我们结合两者，旨在通过局部与全局特征的互补，来提升模型对图像的理解和处理能力。此外，我们的研究提出了一种创新训练策略，通过冻结 ViT 模型的浅层参数，只微调深层参数。这简化了训练复杂度，深层参数捕捉复杂特征，保持模型全局建模能力。该方法能节约计算资源，维持性能，结合局部与全局特征，提升模型性能，为未来研究和应用开辟新可能。

然而因为 Swin 与 ViT 融合特征需额外模块，所以双分支架构增加了参数量，尽管冻结 ViT 浅层参数减少计算量，但训练时间仍长。且融合模型虽在小数据集表现好，但训练准确率达 100% 可能过拟合。数据增强策略虽然缓解了样本不足，但对相似类别区分需更精细标注，而若类别增至 100 类，模型可能面临长尾分布挑战。

3.3. 实际应用建议

构建数据集时，需确保数据多样性和标注规范。目前有 60 类中草药图像，建议扩大数据规模，涵盖更多类别，每类至少 150 张图像。可与相关机构合作，采集不同生长阶段的图像，确保场景多样性。使用背景分离技术减少干扰，对稀缺类别用 GAN 合成图像补充样本。在标注标准化方面，除了标注类别标签外，还应增加关键部位的标注，以便模型聚焦细粒度特征。建议采用通用格式的标注文件，便于模型迁移与共享。然而，当前模型存在局限，如对不规则中草药建模能力有限，以及数据集类别不平衡问题。下一步需改进模型和算法，提高分类精度。可使用加权损失函数和过采样技术，减少样本分布不均引起的模型偏差。

参考文献：

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[J]. Advances in Neural Information Processing Systems, 2020, 33: 1–22.

[2] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on

Computer Vision. 2021: 10012–10022.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

[4] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 1–20.

[5] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[J/OL]. arXiv preprint arXiv:2010.11929, 2020.

[6] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[J/OL]. arXiv preprint arXiv:2103.14030, 2021.

基金项目：

珠海科技学院 2024 年校级大学生创新训练计划（药膳同源—区域链技术下的药膳产品项目）（项目编号：DC2024030）