DOI:10.16861/j.cnki.zggc.2024.0325

基于深度学习的可食用野菜种类识别

吴玉强[1,2],孙 荀[1],季呈明[2],胡乃娟[3]

(1.南京警察学院信息技术学院 南京 210023; 2.南京农业大学工学院 南京 210095;
3.江苏省农业科学院农业经济与发展研究所 南京 210014)

摘 要:可食用野菜兼具营养价值和药用价值,然而传统采摘可食用野菜的分辨主要依赖人为主观经验,效率低且错误风险高,因此对可食用野菜快速准确的识别对实现野菜产业开发和保障食用安全具有重要意义。以南京地区“七头一脑”共 8 种可食用野菜为研究对象,构建了 8 种野菜的 2400 张图像数据集,采用 3 种具有代表性的卷积神经网络(convolutional neural network,CNN)模型(AlexNet、VGG16 和 ResNet50)和 3 种视觉自注意力(vision transformer,ViT)模型(ViT、CaiT 和 DeiT)共 6 种不同的深度学习模型进行训练和验证,并通过梯度加权类激活映射(gradient-weighted class activation mapping,Grad-CAM)来分析深度学习模型的决策机制。结果表明,ResNet50 在验证集上的准确率达到 94.68%,精确率、召回值和 F1 分数分别为 97.66%、97.74%和 97.70%,在 6 个模型中表现最佳。随后,在最优模型 ResNet50 基础上添加卷积模块的注意力机制(convolutional block attention module,CBAM)和坐标注意力机制(coordinate attention,CA)模块进行模型优化,结果显示,CBAM-ResNet50 准确率达到了 97.67%,CA-ResNet50 准确率达到了 98.34%,分别提高了 2.99 个百分点和 3.66 个百分点。以上研究结果证实了 CNN 模型在数据集上能取得比 ViT 更好的结果,利用深度学习识别可食用野菜种类是可行的,且添加注意力模块能够实现更高的识别准确率。

关键词:可食用野菜;种类识别;卷积神经网络;视觉自注意力;注意力机制模块

中图分类号:S647 **文献标志码**:A **文章编号**:1673-2871(2024)11-057-10

Identification of edible wild vegetable species based on deep learning

WU Yuqiang[1,2], SUN Xun[1], JI Chengming[2], HU Naijuan[3]

(*1.College of Information Technology, Nanjing Police University, Nanjing 210023, Jiangsu, China; 2. College of Engineering, Nanjing Agricultural University, Nanjing 210095, Jiangsu, China; 3. Institute of Agricultural Economy and Development, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, Jiangsu, China*)

Abstract: Edible wild vegetables possess both nutritional and medicinal values. However, the traditional identification of wild edible vegetables mainly relies on subjective human experience, which is inefficient and carries a high risk of error. Therefore, rapid and accurate identification of edible wild vegetables is of great significance for the development of the wild vegetable industry and the assurance of food safety. Eight types of edible wild vegetables known as the "Seven Heads and One Brain" in the Nanjing region were selected as the research subjects and a database of 2400 images were constructed. Training and validation were conducted using 6 different deep learning models, including 3 representative convolutional neural network (CNN) models (AlexNet、VGG16 and ResNet50) and 3 vision transformers (ViT) models (ViT、CaiT and DeiT). Furthermore, the decision-making mechanisms of the deep learning models were analyzed using Gradient-Weighted Class Activation Mapping. The results showed that ResNet50 achieved an accuracy rate of 94.68% on the validation set, with precision, recall value, and *F*1-score of 97.66%, 97.74%, and 97.70%, respectively, and performed the best among the 6 models. Subsequently, the attention mechanism modules, convolutional block attention module and coordinate attention module were added to the optimal ResNet50 model for further optimization. The results showed that the accuracy of CBAM-ResNet50 and CA-ResNet50 models achieved 97.67% and 98.34%, respectively, representing enhancements of 2.99 and 3.66 percent point. The above research results confirmed that the CNN model can achieve

收稿日期:2024-05-13;**修回日期**:2024-09-09

基金项目:江苏省重点研发计划项目(BE2019762);中央高校基本科研业务费专项资金项目(LGZD202408);国家自然科学基金(32201923);“十四五”江苏省重点学科“公安技术”(苏教研函(2022)2 号)

作者简介:吴玉强,副教授,研究方向为计算机视觉与食品安全检测技术。E-mail:wuyq@nfpc.edu.cn

通信作者:胡乃娟,副研究员,研究方向为智慧农业技术。E-mail:20210107@jaas.ac.cn

better results than ViT on the dataset in this paper. It is feasible to use deep learning to identify edible wild vegetable species, and adding attention modules can lead to higher recognition accuracy.

Key words: Edible wild vegetables; Species identification; Convolutional neural networks; Vision transformer; Attention mechanism modules

野菜是一类未经人工培育、在野外环境中生长的植物。很多野菜具有重要的经济价值和一定的药用价值，对改善人类的膳食结构、丰富药用植物资源都具有积极的意义[1]。我国是拥有丰富野菜资源的国家之一，野菜种类有 700 种以上，其中常见的有 100 多种[2]。但是由于野菜种类繁多、形态相似，在日常生活中经常发生误食有毒野菜导致中毒等安全事件。据统计，我国每年发生的食物中毒案件中，误食野菜引起中毒死亡占很大比例，贵州省在 2016—2021 年共报告有毒植物中毒事件 550 起，其中误食有毒野菜的就有 178 起，且中毒事件数量还有不断上升的态势[3]。同时，随着社会生活方式的丰富，人们饮食结构也变得多样化，野菜食用越来越受到现代人的追捧，因此，如何精准识别出可食用野菜对保障食物安全具有重要意义。

传统的可食用野菜识别分类最普遍的方法是人工感官识别，但这种方法依赖个人经验，成本高、效率低且准确率不稳定。近年来，深度学习在图像识别、语音识别、目标检测等多个领域的应用都取得显著的进展和成功[4-6]。基于深度学习的计算机视觉技术应用于植物表型分类识别具有高效、无损、易操作等优点，已逐渐成为农业领域和食品安全领域的重要研究方向之一。

卷积神经网络（convolutional neural network，CNN）和视觉自注意力模型（vision transformer，ViT）是目前广泛应用于图像分类识别和目标检测领域的两大类模型[7-8]。其中，CNN 是一类包含卷积计算且具有深度结构的前馈神经网络，是深度学习（deep learning，DL）的代表算法之一。2012 年 AlexNet 在 ImageNet 挑战赛的成功，重新点燃了人们对深度学习领域研究的兴趣[9]。林伟等[10]以大豆籽粒分类为目标，构建大豆籽粒图像数据集，通过对传统 AlexNet 模型进行改进来对大豆籽粒验证集进行分类；王圆等[11]将番茄叶片病虫害数据集分为 5 类，并采用改进的 ResNet50 网络识别番茄叶片病虫害，取得了不错的效果。然而，当前大多数研究人员都专注于应用 CNN 对研究中的图像进行分类，缺少与 ViT 模型的比较。与 CNN 比，ViT 在最新计算机视觉研究进展中展现出了显著的性能[12]。王杨等[13]将改进的 Vision Transformer 网络应用于一个包含 9 种番茄叶片病害图像、1 种健康叶片图像和 1 种无关背景图像的共 11 种番茄叶片数据集进行病虫害识别，取得了 99.63%的分类准确率；Castellano 等[14]提出了一种基于轻量级 Transformer 的新方法，在不影响推理时间的情况下，实现在多光谱无人机图像中绘制杂草地图，实现可持续和更高效的农业生产。因此，进一步研究 ViT 模型在野菜种类识别领域的应用具有重要意义。深度学习模型通常被认为是一个“黑匣子”，这意味着这些模型的决策机制是不透明的，而模型的透明度可以让研究人员在模型决策过程中更有信心。

笔者通过采集江苏南京地区有名的“七头一脑”（苜蓿头、枸杞头、豌豆头、荠菜头、马兰头、香椿头、小蒜头和菊花脑）共 8 种野菜图像样本构建数据集，并分别选用 3 种经典的 CNN 模型和 3 种 ViT 共 6 种模型对这 8 种野菜进行分类识别，然后利用梯度加权类激活映射（gradient-weighted class activation mapping，Grad-CAM）[15]算法在给定图像中可视化对模型预测贡献最大的像素，并以热力图的方式输出，从而深入了解 CNN 和 ViT 的决策过程，进而为野菜精准识别模型选择和食品安全检测领域提供借鉴。

1 材料与方法

1.1 材料和设备

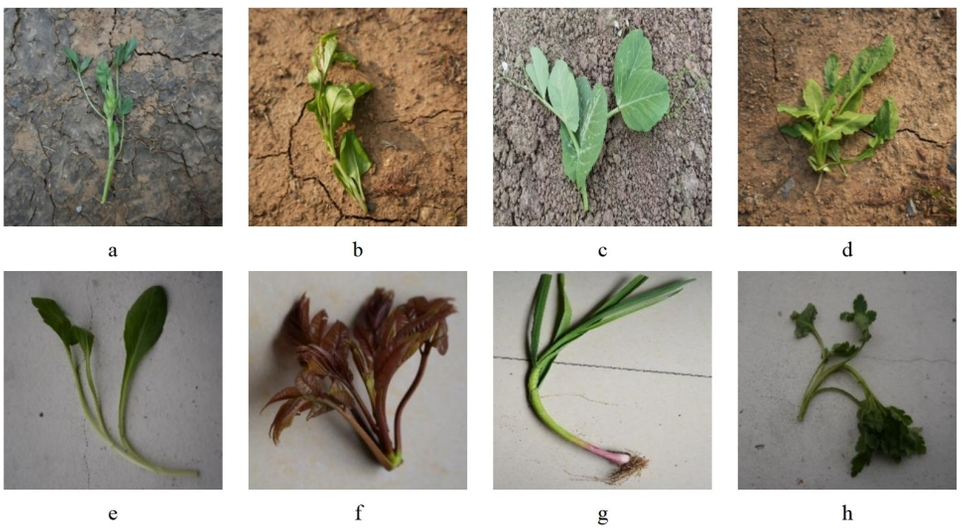
试验材料“七头一脑”8 种野菜均购买于南京市栖霞区仙林街道菜市场。图像拍摄设备为索尼 ILCE-7M4 相机，图像分辨率为 7008 × 4072。图像采集地点位于南京警察学院敏行楼实验室，以及校园内裸露的黑土和黄土地面上，采集时间为 2024 年 3 月 11 日到 2024 年 3 月 27 日，共分 3 个批次采集。

1.2 方法

1.2.1 *数据预处理* 为了更好地贴近采集野菜的野外环境，笔者选用包括黑土、黄土及淡黄色、白色等不同颜色的背景进行拍摄，除了自然光下正常拍摄之外，还采用背光、逆光等不同光线来增加数据集的丰富性，为后期训练增加难度。由于数据集由相机拍摄，原始图片较大，为了更方便进行操作，利

用 Python 3.8 进行编程，批量将图像分辨率调整为 640 × 640。共拍摄“七头一脑”野菜图片 2487 张，后期通过筛选，最终得到图像 2400 张。其中，苜蓿头 300 张、枸杞头 298 张、豌豆头 296 张、荠菜头 302 张、马兰头 298 张、香椿头 303 张、小蒜头 302 张、菊花脑 301 张。8 种野菜的样本实例如图 1 所示。

1.2.2　基于传统 CNN 模型的识别　传统 CNN 模型包括 AlexNet、VGG16 和 ResNet50 等 3 种。



注：a. 苜蓿头；b. 枸杞头；c. 豌豆头；d. 荠菜头，e. 马兰头；f. 香椿头；g. 小蒜头；h. 菊花脑。

Note: a. represents alfalfa sprouts; b. represents goji sprouts; c. represents pea sprouts; d. represents shepherd's purse sprouts; e. represents horseradish; f. represents Chinese toon sprouts; g. represents garlic chives; h. represents chrysanthemum stems.

图 1　南京地区“七头一脑”样本图

Fig. 1　Sample pictures of "seven heads and one brain" in Nanjing

AlexNet 是由 Krizhevsky 等[9]在 2012 年提出的，是深度学习在计算机视觉领域取得突破性进展的标志模型。AlexNet 共包含 8 层神经网络，采用了大量的卷积核和池化层，以及修正线性单元作为激活函数，同时采用了 Dropout 来缓解过拟合问题。

VGG16 由牛津大学提出，该模型于 2014 年在 ImageNet 图像分类挑战赛中取得了很大的成功[16]。它采用了连续的卷积层和池化层，通过增加网络深度来提高性能，其核心思想是通过多个小尺寸的卷积核和池化层来堆叠网络，以增加感受野大小和提高非线性表达能力。

Residual Network（ResNet）是由微软亚洲研究院的研究员 He 等[17]于 2015 年提出的网络模型，并在 ImageNet 图像分类挑战赛上取得了非常出色的成绩。ResNet 引入残差学习机制，从而解决了深度神经网络中的梯度消失和梯度爆炸等问题。ResNet 提出了跳跃连接的概念，即在网络中引入直接连接，将输入信息绕过一些层直接传递给后续层，从而使得网络可以学习到残差的表示，而不是直接学习原始的映射。ResNet50 模型后面之所以有 50，是因为该网络包含了 49 个卷积层、1 个全连接层，除此之外还有 ResNet101 和 ResNet152 等更多层次的模型。本文中的 ResNet50 对输入野菜图片进行识别的整体架构如图 2 所示。

1.2.3　基于新型 ViT 模型的识别　视觉自注意力（ViT）模型包括 ViT、CaiT 和 DeiT 等 3 种。

ViT 将 Transformer 应用在图像分类任务中，是首个将 Transformer 模型应用于计算机视觉任务的模型，因其可扩展性强，使其成为了 Transformer 在计算机视觉应用的里程碑模型[18]。ViT 将输入图像划分为多个 patch，并将每个 patch 作为固定长度的向量投影到 Transformer 中。随后的编码器操作与原始 Transformer 中的操作完全相同。在图像分类任务中，一个特殊的令牌被添加到输入序列中，该令牌的相应输出就是最终的类别预测。ViT 模型的整体架构如图 3 所示。

CaiT 是在 ViT 之后提出的模型，由 Touvron 等[19]于 2021 年提出。CaiT 采用了级联的注意力机制，引入了坐标注意力机制，将图像块的位置信息纳入注意力计算中，以更好地处理图像中的局部结

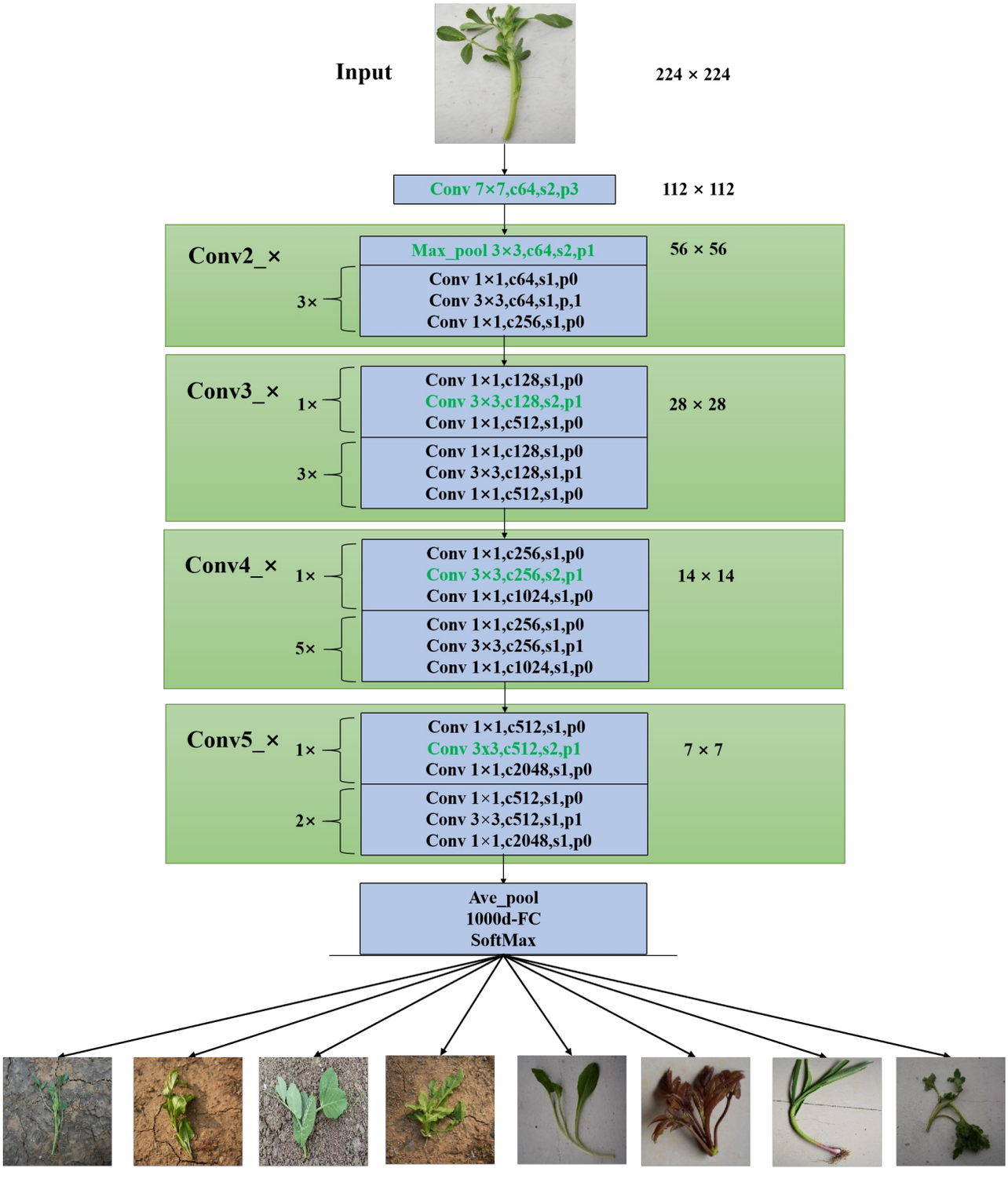


图 2　ResNet50 的网络架构图

Fig. 2　Network architecture diagram of ResNet50

构信息。因此，CaiT 模型在图像局部结构信息上更有优势[20]。

针对数据不足会导致 ViT 性能较差的问题，DeiT 核心共享采用蒸馏策略，通过引入跨模态对齐训练和自注意力机制，以及使用更小的模型参数，实现了对数据更加高效的利用，从而在较小的数据集上取得了很好的性能[21]，作为加入了蒸馏模块的 Transformer 模型也在图像识别中崭露头角[22]。

1.3　模型评价指标

为了评估 6 种模型的野菜识别分类性能，笔者使用了 4 个常用的评价指标，即精确率（precision）、召回率（recall）、*F*1 分数（*F*1-score）和准确率（accu-

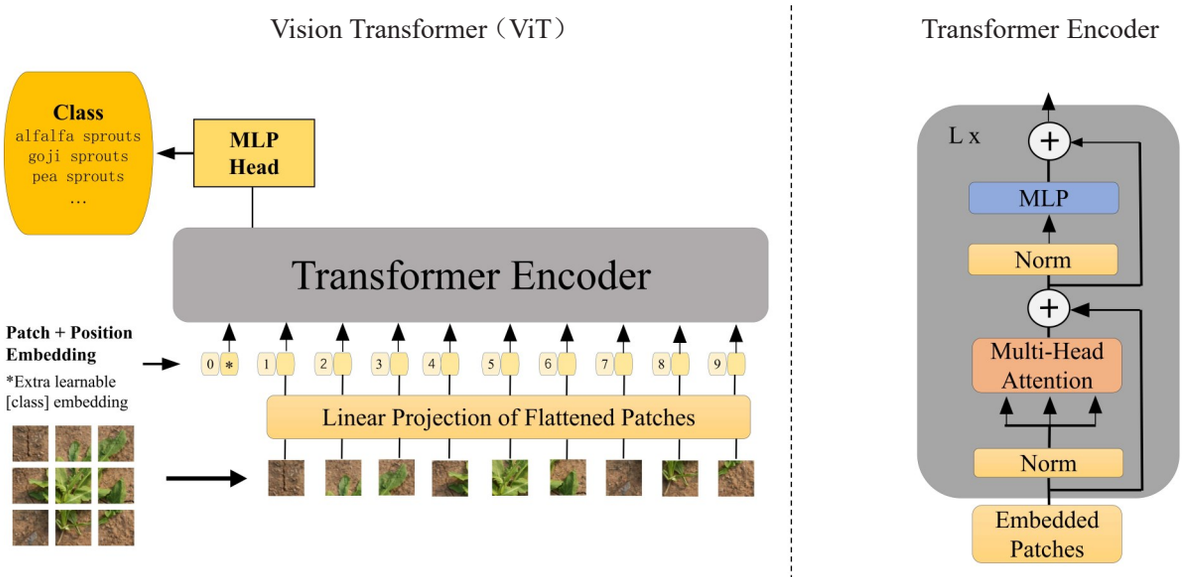


图 3　ViT 模型的网络结构图

Fig. 3　Network structure diagram of ViT

racy）。精确率指的是在所有预测的正样本中，真正的正样本的比例。召回率表示在所有实际正样本中，被正确预测为正样本的比例。*F*1 分数是精确率和召回率的调和平均数，可以评估模型的整体分类性能。准确率表示在总样本数中被正确分类的图像的比例。准确率越高，说明模型在野菜识别分类方面的性能越好。

$$\text{Precision}=\frac{TP}{TP+FP}\ ; \tag{1}$$

$$\text{Recall}=\frac{TP}{TP+FN}\ ; \tag{2}$$

$$F1-\text{score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\ ; \tag{3}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\ 。 \tag{4}$$

1.4　基于 Grad-CAM 的模型可视化

为了便于理解特征学习的过程，对模型进行可解释性分析，笔者使用 Grad-CAM 算法来可视化每个图像中用于种类识别和预测的特征区域，直观地表达算法获得结果的原因。CAM 提取最后一个要素图层和输出之间的全局平均池化层（global average pooling，GAP）。其局限性在于网络模型中必须存在 GAP，而 Grad-CAM 适用于非 GAP 连接的网络结构，其应用范围比 CAM 更广。在许多实际应用中，如医疗诊断[23]、自动驾驶[24]等领域，模型解释性可以帮助医生或工程师更清楚地理解模型的预测依据，增强他们对模型的信任感。此外，模型解释性还有助于检测模型的潜在偏差或错误，从而提高模型的鲁棒性和可靠性。

2　结果与分析

2.1　模型训练

所有试验均在 64 位 Ubuntu18.04 上使用 Pytorch 1.11.0 编程。所用服务器配置为 Intel（R）Xeon（R）Platinum 8255C CPU、64 GB RAM 和 NVIDIA GeForce RTX 3090（24GB）GPU。为了进行试验，笔者将数据集按 8∶2 的比例划分为训练集和验证集。为了确保结果的可靠性和一致性，笔者采用了五折交叉验证方法，并且每个模型进行了 80 个 epoch 的训练，使用相同的超参数设置，将批量大小设置为 64，初始学习率为 0.001，优化器使用 Adam。

2.2　CNN 与 ViT 的对比试验结果

笔者选择 AlexNet、VGG16 和 ResNet50 3 种 CNN 模型与 ViT、CaiT 和 DeiT 3 种 Vision Transformer 模型进行可食用野菜种类识别对比试验。6 种模型在验证数据集上的准确率曲线和损失值曲线如图 4 和图 5 所示，所有模型的损失曲线逐渐下降并达到收敛状态，而准确率曲线逐渐上升并最终稳定。所有模型曲线表现正常，表明模型持续学习具有更准确的特征。在准确率变化曲线方面，所有模型的初始准确值均不高，在前 10 轮中，6 种模型的准确值均快速上升，然后增速变缓，在 50 轮次时逐渐平缓。ViT 和 DeiT 在准确率曲线方面最初表现良好，但后来被 ResNet50 超越，最终 ResNet50 达到所有模型中最优的准确率。在模型损失曲线上，6 种模型初始损失率都很高，且可以观察到 CNN 模

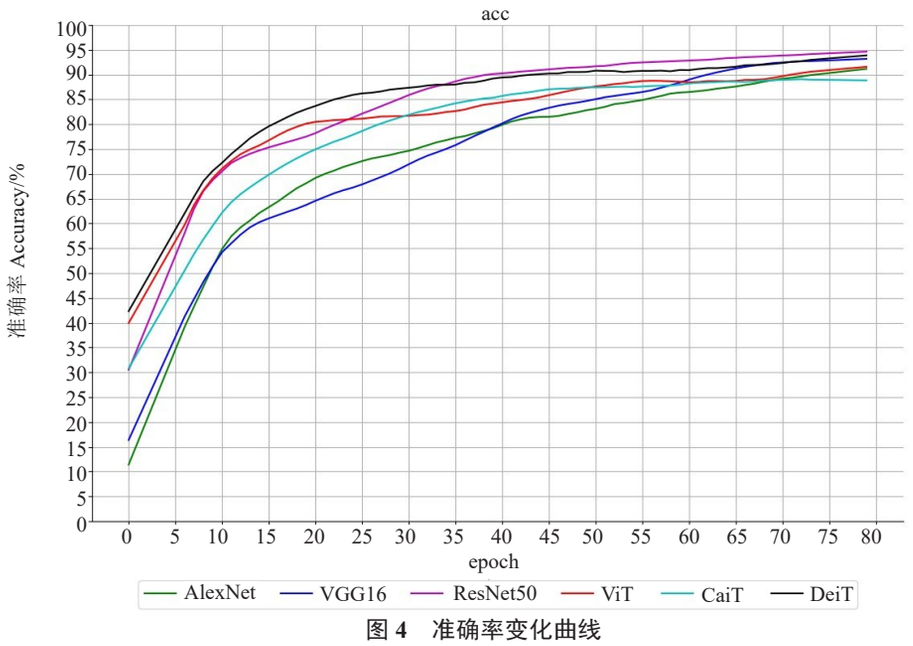


图 4　准确率变化曲线

Fig. 4　The accuracy change curve of six models

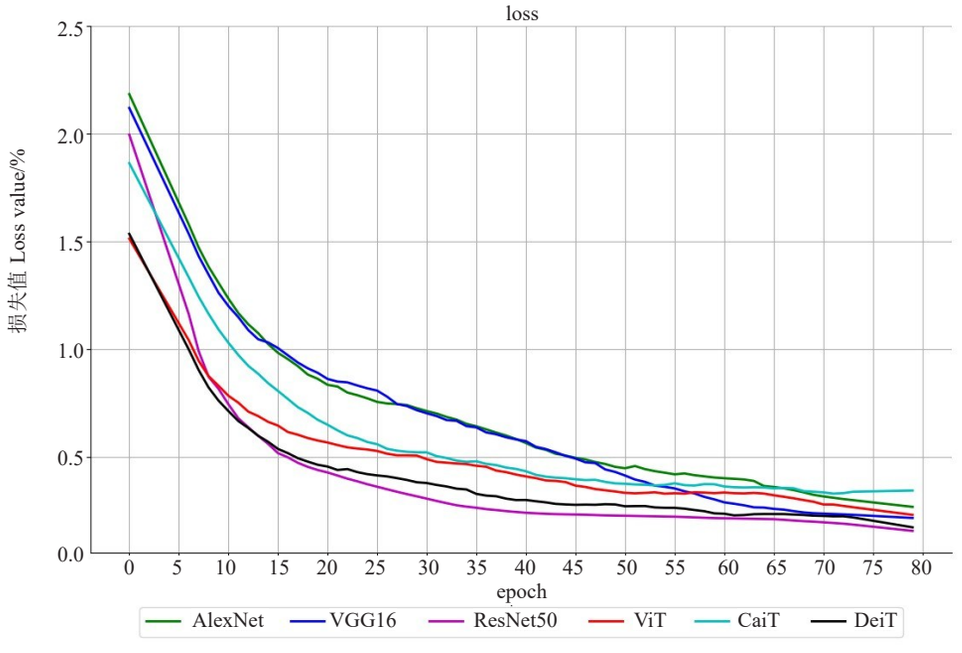


图 5　损失值变化曲线

Fig. 5　The loss change curve of six models

型在初始验证阶段具有较大的损失值，但他们收敛速度更快，而 3 种 ViT 模型的损失值相对较低。当迭代次数达到 60 次后，验证损失值曲线逐渐变平，表明模型接近收敛。与其他训练模型相比，ResNet50 模型的损失曲线收敛更加迅速，表现出更好的训练模型效果，在迭代的初始阶段表现出更好的网络优化性能。因此，可以认为 ResNet50 模型在本研究中更加稳定，表现出良好的鲁棒性。

笔者使用 4 个评价指标数据对 6 种模型的性能进行评估。如表 1 所示，6 种模型在 4 个指标表现上基本都达到了 90%以上，其中 ResNet50 模型在可食用野菜识别中表现最优，其精确率、召回率、*F*1 分数和准确率分别为 97.66%、97.74%、97.70%和 94.68%。在准确率方面，ResNet50 比第二的 DeiT 高出 0.63%，比第三的 VGG16 高出 1.53%。表现最差的模型是 CaiT 模型，准确率只有 89.36%。从整体来看，相较于 Transformer 模型，CNN 模型的效果更好一些。

2.3　ResNet50 模型优化

在 ResNet50 模型中分别加入卷积模块的注意

表 1　6 种模型各项指标比较

Table 1　Comparison of indicators of the six models　%

模型 Model	精确率 Precision	召回率 Recall	*F*1 分数 *F*1-score	准确率 Accuracy
AlexNet	95.81	95.71	95.76	91.36%
VGG16	96.71	95.06	95.88	93.15%
ResNet50	97.66	97.74	97.70	94.68%
ViT	96.46	97.21	96.83	92.52%
CaiT	95.96	93.33	94.62	89.36%
DeiT	96.77	97.13	96.95	94.05%

力机制（convolutional block attention module，CBAM）[25]和坐标注意力机制模块（coordinate attention，CA）[26]，结果如表 2 所示，添加了 CA 注意力模块的 ResNet50 模型的表现最优，其精确率、召回率、*F*1 分数和准确率分别为 98.86%、99.38%、99.12%和 98.34%。两个注意力模块均对 ResNet50 模型效果产生了较大的提升，CBAM 将 ResNet50 模型的准确率提升了 2.99 个百分点，CA 注意力模块将 ResNet50 模型的准确率提升了 3.66 个百分点。

表 2　优化模型的各项指标比较

Table 2　Comparison of various indicators of optimization models　%

模型 Model	精确率 Precision	召回率 Recall	*F*1 分数 *F*1-score	准确率 Accuracy
ResNet50	97.66	97.74	97.70	94.68
CBAM-ResNet50	98.80	99.40	99.10	97.67
CA-ResNet50	98.86	99.38	99.12	98.34

2.4　模型可视化结果

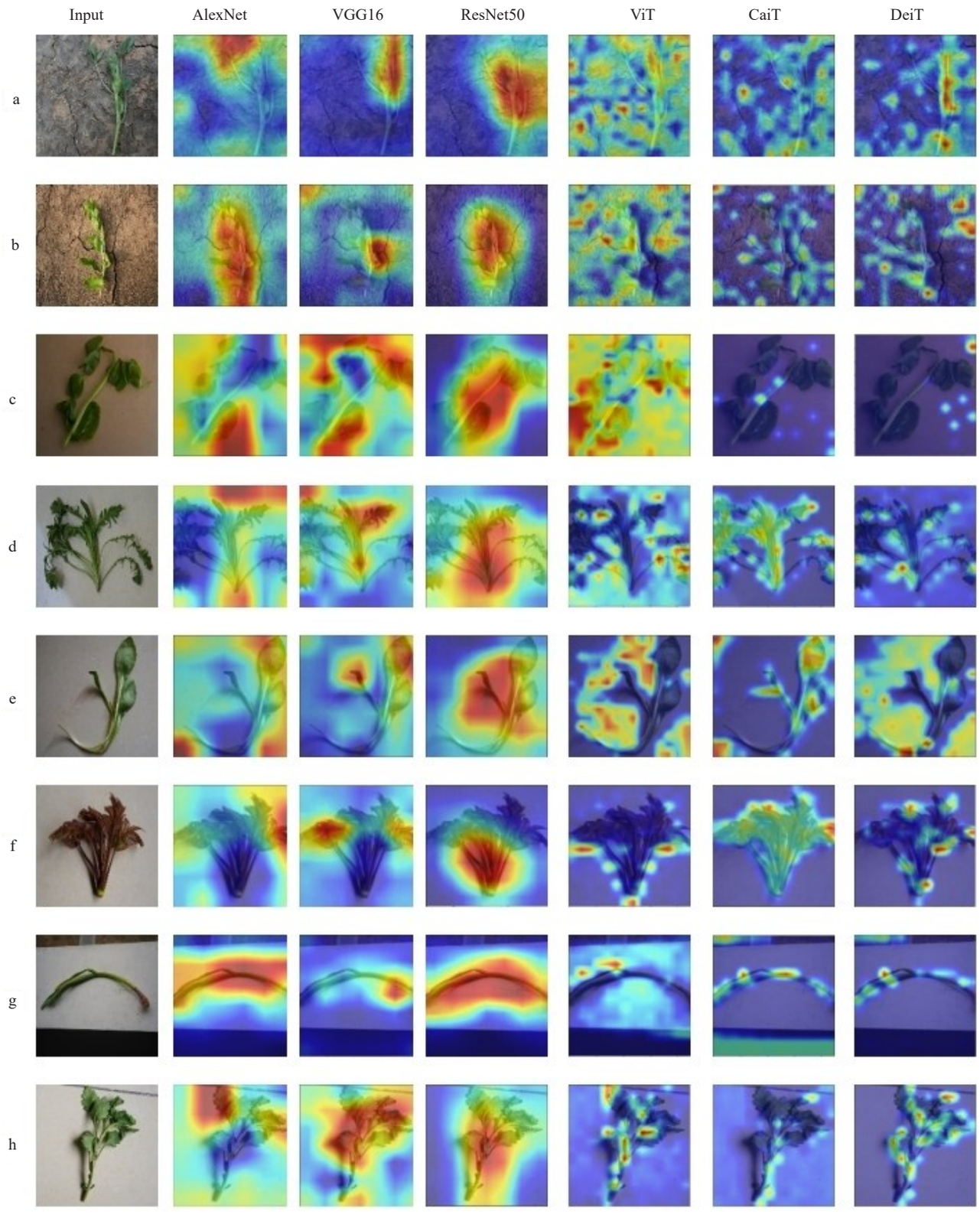
为了确保模型的鲁棒性，笔者使用 Grad-CAM 为 6 种模型的最后一层生成热力图。Grad-CAM 是一种可解释的方法，用于分析深度学习模型的决策机制（即可视化深度学习模型的关注点）。颜色越红，该区域对最终预测结果的贡献越大，而蓝色区域表示贡献较小。对于野菜种类识别分类任务，不同的模型对同一张图片的关注点不同，如图 6 所示。

3　讨论与结论

在模型性能上，本研究中的 3 种 CNN 模型的识别准确率都到了 91%以上，ResNet50 的各项评估指标优于 VGG16 和 AlexNet，且 CNN 总体性能相比 ViT 更好。这是因为 CNN 具有深层架构来对数据进行训练[27]。AlexNet 是一个 8 层的卷积神经网络，而 VGG16 是一个 16 层的卷积神经网络，ResNet50 可以达到 50 层。神经网络的深度越大，学习能力越强，VGG16 使用了更多的数据增强方式，比 AlexNet 有更大的操作空间，因此 VGG16 效果要明显优于 AlexNet，而 AlexNet 最终表现最差。ResNet 提出了残差连接，有效解决了深层网络退化的问题，因此，ResNet50 在 3 种 CNN 中实现了最佳性能。

在 CNN 与 ViT 模型的对比中，除了 AlexNet 外，其他两种 CNN 模型都稍微优于 ViT 和 CaiT。这是因为 CNN 和 Transformer 的架构存在重大差异[28]。CNN 中的卷积核通过局部感受野和共享权重来捕获局部特征。而 Vision Transformer 将给定的图像分成多个 patch，并为每个 patch 分配一个位置嵌入，ViTs 在生成输出时评估每个 patch 在整个图像中的贡献[29]，通过计算图像块之间的关系来捕捉图像全局信息。在这种情况下，Transformer 的性能依赖于训练的数据集的大小，较大的数据集可以取得更好的训练结果[8]，当提供的数据不足时，Transformer 的泛化能力会受到限制。Li 等[30]对野外环境中收集的 2095 份病态和健康甘蔗叶片进行了比较研究，发现 ResNet50 比 ViT 有更显著的效果。Li 等[31]在 7434 个表面人脸识别任务上将 ViT 与 ResNet50 进行比较，结果显示，ResNet50 的性能优于 ViT Base，但低于 Swin Base。ViT 和 CNN 的性能不完全取决于数据集的大小，而是与图像数据集本身的质量、分布、应用场景等密切相关。Li 等[32]根据医学图像和自然图像之间的差距，提出了一种基于视觉 Transformer 骨干的专用医学图像分类模型。而本研究中 DeiT 取得第二名的成绩，相对于 DeiT，ResNet50 在训练和推理阶段通常具有更高的计算效率。ResNet50 使用了卷积层和残差连接等结构，在数据量较小的数据集计算上比 Transformer 的自注意力机制更为高效。和 ViT 相比，DeiT 在模型结构和特征提取方面可能进行了一些改进，更专注于提高数据利用效率，通过一系列的训练策略和技巧，例如知识蒸馏和对比学习等，使得 DeiT 在相对较小的数据集上也能够表现出色[21]。因此，本研究在现有数据集情况下，DeiT 取得了仅次于 ResNet50 的效果。

CBAM 模块将通道和空间注意力添加到 ResNet50 模型当中，能够获取更加全面的图像信息，强化有效特征，进而提高了网络模型的性能[25]。在本研究中的图像数据中，野菜的根茎、叶片占据图片的主要部分，通道注意力和空间注意力机制可以缩小图像中干扰信息的影响，更加注重野菜本身



注：a~h 分别代表苜蓿头、枸杞头、豌豆头、荠菜头、马兰头、香椿头、小蒜头和菊花脑的原始输入图像以及他们在 6 种不同模型下的热力图。

Note: a-h represent the original input images of alfalfa sprouts, goji sprouts, pea sprouts, shepherd's purse sprouts, horseradish, Chinese toon sprouts, garlic chives and chrysanthemum stems, as well as their corresponding heatmaps under six different models.

图 6　6 种模型的热力图

Fig. 6　Heat maps of six models

所在区域，以达到更好的训练效果，所以添加CBAM注意力模块的ResNet50模型要比ResNet50模型自身要好。CA模块通过精确的位置信息对通道关系和长程依赖进行编码，能帮助模型更加精准地定位和识别感兴趣的目标；还能够在移动网络中参与大区域的建模，并避免大量的计算开销[26]。在野菜数据集中，定位好野菜的位置可以让训练数据更加准确，而野菜的方向总是从根茎到枝，坐标注意力利用这一点以实现更高的准确率。所以添加CBAM和CA注意力模块都使ResNet50模型在本研究的野菜数据集上表现更加优秀。

在模型可视化特征上，与AlexNet和VGG相比，ResNet50特征提取能力将大大增强[33]。一方面，浅层可以捕获边缘和纹理等特征，而深层能够提取语义信息，因此ResNet50能更加全面准确地捕捉野菜的关键特征区域；另一方面，ViT模型由于是将图片细分为更小的区域，注意力以点状散开，导致有些时候无法将注意力分散的区域连接到一起，在3种分类状态下显示出更为分散的关注点。ViT的优势在于能够捕捉图像中的远距离依赖关系，无需复杂的卷积操作，而CaiT更关注野菜本身区域。DeiT基于ViT并加入了蒸馏学习的蒸馏标记，更关注野菜周围区域。从图6可以看出，CNN模型抗图像背景干扰更强，即使苜蓿头和枸杞头分别在黑土地和黄土地上拍摄，但CNN模型也能很好地捕捉到叶片特定区域，尤其是ResNet50网络，总能获得最大的感兴区域，从而进行分类决策。而图像背景对ViT分类结果有较大影响，可以看出，ViT类模型在黑土和黄色背景下，关注了大量非野菜区域的嘈杂背景信息，而在白色等简单背景下，关键点像素则聚焦到识别目标本身。这说明，未来在使用ViT系列模型时，可以考虑去除嘈杂背景等数据增强技术来提高模型的识别能力。

笔者基于两种广泛使用的深度学习框架，即CNN模型和Vision Transformer模型，对自建的野菜数据集识别分类任务进行比较评估，总体而言，6种模型中5种的准确率都在90%以上，表明他们能够准确提取不同野菜图像的特征。其中，ResNet50在验证集上的准确率达到94.68%，在所有评估标准下均取得了最佳性能；通过加入CBAM和CA模块进行模型优化，取得了更加显著的效果。最后，采用Grad-CAM算法对先前6种模型的分类效果进行可视化，对模型训练过程进行可视化解释。基于深度学习的野菜识别技术无论在食品安全领域还是在农业生产过程中都具有广阔的应用前景，但同时也面临着许多挑战。首先是数据集场景简单、丰富性不足的问题。由于当前很少有专门针对野菜识别的公开数据集，本文中可食用野菜从菜市场购得，虽然部分图片模拟了田间地头背景进行拍摄，但图像总体背景并不复杂，后续需要深入野菜生长真实场景拍摄更多样本，以进一步扩大数据集的多样性和复杂性，增强模型的泛化性和鲁棒性。其次是模型轻量化问题。尽管具有深层次的深度学习模型有实现智慧农业应用中所期望检测结果的潜力，但他们可能需要更多的训练时间。考虑到在计算能力有限的边缘设备上部署模型变得越来越重要，设计轻量级网络模型的趋势日益增长[34]。如今，将ViT网络与CNN方法相结合进行图像分类的研究逐渐受到关注，这种结合的方法旨在充分利用ViT对全局信息的优势以及CNN在局部特征提取方面的优势，提升图像分类性能，后续可以在本研究的基础上进一步优化模型，达到更优效果。

参考文献

[1] 卢超.长沙地区野菜资源开发利用研究[D].长沙：湖南农业大学，2017.

[2] 查金平.利用野菜资源开展校本研究提升学生核心素养[J].科技风，2019(8)：37.

[3] 何进，刘琳，朱姝，等.贵州省2016—2021年有毒植物及其毒素中毒暴发事件监测情况分析[J].现代预防医学，2022，49(21)：4009-4013.

[4] 刘文斌，庹先国，张贵宇，等.基于卷积神经网络的白酒上甑探汽方法[J].食品研究与开发，2024，45(5)：139-144.

[5] WANG M S，MA H B，WANG Y L，et al. Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion[J]. Applied Acoustics，2024，218：109886.

[6] XU G，YUE Q R，LIU X G. Real-time multi-object detection model for cracks and deformations based on deep learning[J]. Advanced Engineering Informatics，2024，61：102578.

[7] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the Ieee，1998，86(11)：2278-2324.

[8] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16×16 Words：Transformers for image recognition at scale，May 04，2021[C]. Vienna：International Computer on Learning，2021.

[9] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the Acm，2017，60(6)：84-90.

[10] 林伟，仲伟波，袁毓，等.基于改进AlexNet与CUDA的大豆快速三分类方法[J].计算机与数字工程，2023，51(12)：2997-3003.

[11] 王圆，祝俊辉，周贤勇，等.基于改进 ResNet 模型的番茄叶片病虫害识别[J].激光杂志，2024，45(5)：209-214.

[12] ZHOU B，YU X，LIU J，AN D，et al. Effective vision transformer training：A data-centric perspective[J]. Computer Vision and Pattern Recognition，2022，2209：15006.

[13] 王杨，李迎春，许佳炜，等.基于改进 Vision Transformer 网络的农作物病害识别方法[J].小型微型计算机系统，2024，45(4)：887-893.

[14] CASTELLANO G，MARINIS P D，VESSIO G. Weed mapping in multispectral drone imagery using lightweight vision transformers[J]. Neurocomputing，2023，562：126914.

[15] SELVARAJU R R，COGSWELL M，DAS A，et al. Grad-CAM：Visual explanations from deep networks via gradient-based localization[C]. Ieee International Conference on Computer Vision（ICCV），2017：618-626.

[16] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science，2014.

[17] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition（CVPR），2016：770-778.

[18] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Computation and Language，2017，30：6000-6010.

[19] TOUVRON H，CORD M，SABLAYROLLES A，et al. Going deeper with image transformers[C]. IEEE/CVF International Conference on Computer Vision（ICCV），2021：32-42.

[20] LIU Y，ZHANG Y，WANG Y，et al. A survey of visual transformers[J]. IEEE Transactions on Neural Networks and Learning Systems，2024，35：7478-7498.

[21] TOUVRON H，CORD M，DOUZE M，et al. Training data-efficient image transformers & distillation through attention[C]. International Conference on Machine Learning，2021，139：7358-7367.

[22] 赵婷婷，高欢，常玉广，等.基于知识蒸馏与目标区域选取的细粒度图像分类方法[J].计算机应用研究，2023，40(9)：2863-2868.

[23] 曹明亮，尹蜜，王庆彬，等.基于深度学习算法联合 Grad-CAM 的宫腔镜子宫内膜病变诊断模型研究[J].实用妇产科杂志，2024，40(5)：409-413.

[24] 谢瑞麟，崔展齐，陈翔，等.IATG：基于解释分析的自动驾驶软件测试方法[J].软件学报，2024，35(6)：2753-2774.

[25] WOO S H，PARK J，LEE J Y，et al. CBAM：Convolutional block attention module[J]. Computer Vision，2018，11211：3-19.

[26] HOU Q B，ZHOU D Q，FENG J S，et al. Coordinate attention for efficient mobile network design[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：13708-13717.

[27] PRAKASH J A，ASSWIN C R，KUMAR K S D. A，et al. Transfer learning approach for pediatric pneumonia diagnosis using channel attention deep CNN architectures[J]. Engineering Applications of Artificial Intelligence，2023，123：106416.

[28] XIONG B P，CHEN W S，NIU Y X，et al. A Global and Local Feature fused CNN architecture for the sEMG-based hand gesture recognition[J]. Computers in Biology and Medicine，2023，166：107497.

[29] ZHOU D，KANG B，JIN X，et al. DeepViT：Towards deeper vision transformer[J]. Computer Vision and Pattern Recognition，2021.

[30] LI X C，LI X H，ZHANG M Q，et al. SugarcaneGAN：A novel dataset generating approach for sugarcane leaf diseases based on lightweight hybrid CNN-Transformer network[J]. Computers and Electronics in Agriculture，2024，219：108762.

[31] LI X P，XIANG Y Y，LI S Q. Combining convolutional and vision transformer structures for sheep face recognition[J]. Computers and Electronics in Agriculture，2023，205：107651.

[32] LI Y X，HUANG Y W，HE N J，et al. Improving vision transformer for medical image classification via token-wise perturbation[J]. Journal of Visual Communication and Image Representation，2023，98：104022.

[33] LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition（CVPR），2017，106：936-944.

[34] KIM W，JUNG W S，CHOI H K. Lightweight driver monitoring system based on multi-task mobilenets[J]. Sensors，2019，19(14)：3200.