基于改进 ConvNeXt 模型的农作物害虫细粒度识别

韩源涛[1]，张 聪[2※]，詹晓芸[2]，王 正[3]

（1. 武汉轻工大学数学与计算机学院，武汉 430048；2. 武汉轻工大学电气与电子工程学院，武汉 430048；
3. 武汉大学计算机学院，武汉 430072）

摘 要：精确分类不同时期的农业害虫对控制其发生和发展至关重要。针对目前不同生长时期农作物害虫分类不准确的问题，该研究创建了一个关注虫态的害虫数据集并提出了一种基于改进 ConvNeXt 网络的农作物害虫识别模型。通过引入多种虫态共同监督来重构网络主干，以便模型学习不同虫态的特征，引入空间注意力（spatial attention，SA）来改进模型结构，增强对害虫位置信息的提取能力。在大型公开数据集 IP102 上进行试验，与现有的同类最优基于 Vision Transformer 的方法相比，在保持模型参数量基本没有增加的前提下，准确率提高 3.67 个百分点，F1 值提高 2.49 个百分点。试验证明，该研究提出的模型针对不同虫态害虫具备较强的识别准确率，可为精准农业害虫识别提供一定的参考。

关键词：害虫识别；农作物；ConvNeXt；空间注意力机制；多虫态识别

doi：10.11975/j.issn.1002-6819.202408055

中图分类号：TP391.41 **文献标志码**：A **文章编号**：1002-6819(2025)-04-0185-08

韩源涛，张聪，詹晓芸，等. 基于改进 ConvNeXt 模型的农作物害虫细粒度识别[J]. 农业工程学报，2025，41(4)：185-192. doi: 10.11975/j.issn.1002-6819.202408055 http://www.tcsae.org

HAN Yuantao, ZHANG Cong, ZHAN Xiaoyun, et al. Fine-grained identification of crop pests using an enhanced ConvNeXt model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2025, 41(4): 185-192. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202408055 http://www.tcsae.org

0 引 言

在全球农业生产与管理过程中，有效控制农作物害虫已成为确保粮食安全与提升产量的关键议题。在此背景下，精准而高效的害虫识别技术发挥着至关重要的作用。传统的害虫分类严重依赖农技人员的个人经验，不仅费时费力且效率有限。

随着模式识别与图像处理技术的迅速发展，更多的研究者开始探索基于图像的害虫识别方法。早期的方法主要依赖于手工特征提取技术，例如局部二进制模式（local binary pattern，LBP）[1]、梯度方向直方图（histogram of oriented gradient，HoG）[2] 和尺度不变特征转换（scale invariant feature transform，SIFT）[3]。这些方法在捕获低层次的特征，如颜色、边缘和纹理方面表现良好。研究者通过这些手工设计的特征提取器来采集局部特征，并依据局部纹理差异对昆虫进行识别和分类。例如，WEEKS 等[4] 提出了一种基于主成分分析（principal component analysis，PCA）的昆虫标本识别方法，该方法专注于利用翅脉和色素沉积模式对近缘寄生蜂进行分类识别。SAMANTA 等[5] 在一个包含 609 个样本的数据集上，使用基于相关性的特征选择和人工神经网络对 8 种茶树害虫进行诊断。此外，支持向量机（support vector machine，SVM）分类器也被用于识别叶片图像中的烟粉虱、蚜虫和蓟马[6-7]。尽管这些方法在小数据集上能够有效提取几种典型的手工特征并进行评估，但它们在特征表示的丰富度上存在局限。

近年来，深度学习技术已引起了广泛关注[8-10]。与依赖手工特征截然不同的是，深度学习模型能够通过多层神经网络自动从数据中学习并提炼特征，进而显著提升分类精度。深度学习的成功在很大程度上有赖于高质量与大规模的数据集。

为推动害虫识别技术的发展，研究者构建了大量的害虫图像数据集。LIU 等[11] 通过训练一个深度卷积神经网络（deep convolutional neural networks, DCNNs）对稻田害虫进行分类，其数据集包含 12 类约 5 000 个训练样本。WU 等[12] 提出了一个大规模的农业害虫数据集 IP102，包含 75 222 张由 102 个类别组成的图像，用于害虫分类任务。此数据集在应用 AlexNet[13]、GoogLeNet[14]、VGGNet[15] 和 ResNet[16] 进行分类时显示了显著的提升潜力。此外，BOLLIS 等[17] 提供了一个柑橘害虫数据集（citrus pest benchmark，CPB），通过应用弱监督学习方法来改进柑橘害虫的检测和分类。这些数据集在害虫分类研究方面发挥了积极作用，然而它们仍主要聚焦于不同害虫物种间的分类，对害虫在不同时期虫态变化的关注度相对不足。

在丰富的数据集支持下，各种深度学习模型被应用于害虫分类任务。特别是 DCNNs 在图像分类任务中表现出卓越的性能。基于 ImageNet 预训练的 VGG16 网络架构，VALAN 等[18] 在单类别训练样本量不足 100 张的

收稿日期：2024-08-07 修订日期：2024-12-06
基金项目：湖北省技术创新重大项目（2018A01038）
作者简介：韩源涛，研究方向为农业图像处理。Email：hansel_00@163.com
※通信作者：张聪，博士，二级教授，研究方向为多媒体信息处理和网络通信，人工智能与大数据技术。Email：hb_wh_zc@163.com

条件下，成功实现了与专业昆虫分类相媲美的识别精度。同时，Bert[19]模型在自然语言处理领域的成功也激发了将其应用于计算机视觉的尝试，例如 Vision Transformer[20]几乎没有对原有的 Transformer[21]架构进行大的改动，在大型数据集上与 CNN 模型的表现相当，甚至更佳。LIU 等[22]提出的自监督的基于 Transformer 的预训练方法，利用潜在语义遮蔽自编码器（latent semantic masking auto-encoder, LSMAE）在 IP102 数据集上达到了 74.69% 的准确率。在模型轻量化方向的探索中，彭红星等[23]在 ShuffleNet V2 结构中引入多尺度特征融合模块，在一组包含 24 类害虫的数据集上获得了 79.39% 的平均准确率。张佳敏等[24]利用 DeAnchor 算法对 Mask-RCNN 的锚框引导机制进行了改进，使其在处理虫体密集场景的图像时，识别准确率提升至 90.6%。然而，由于害虫在不同生长时期的形态特征存在显著差异，现有的识别算法仍难以充分利用该领域特定的高层语义信息，导致识别性能受到限制。

针对这一难题，本研究提出了多虫态共同监督策略，以同时捕捉害虫在不同虫态下的共性与个性特征，从而强化模型对多样化形态信息的整合与辨识。同时，引入空间注意力机制，引导模型关注图像中害虫目标的关键区域，以更有效地提取小目标和独特形态特征，通过这些改进与优化，在农业害虫识别的准确率与鲁棒性方面取得提升，为农业害虫防治提供精准、高效的技术支持。

1 材料与方法

1.1 农作物害虫数据集

本研究基于最流行的农业害虫数据集 IP102 构建了本数据集 Age AP。IP102 是一个在互联网中收集的大规模数据集，包含 75 222 张照片，102 类害虫。为了各类害虫进一步的虫态细化分类，本研究对数据集中的图片进行了虫态标注。虫态是指昆虫在其生命周期中的不同形态和发育阶段。在昆虫学和农业科学中，虫态通常包括卵、幼虫、蛹和成虫等阶段。每个阶段的形态特征和生理行为可以有显著差异，这对于识别和管理害虫种群特别重要。例如，在农业害虫管理中，不同的虫态可能需要不同的防治策略，因为某些虫态可能对农药更敏感，或者对作物的危害程度更高。因此，精确识别害虫的虫态是实施有效害虫管理措施的关键。数据集的构建过程如下：

1）基于《中国农业害虫图鉴》[25]进行目标害虫初步比对。

2）剔除含多虫态或非目标对象的无效图像。

3）由专业志愿者团队完成标注，经双重校验保留一致性结果，分歧数据予以排除。

对于完全变态的害虫，一般可以按照生长时期，将其分成卵、幼虫、蛹和成虫 4 个虫态。由于有部分害虫属于不完全变态，没有蛹这一生长时期，故只有 3 种虫态。最终收集到的数据集 Age AP 总共由 102 类害虫，369 类不同虫态的 51 670 张图片组成。各类具体害虫的训练集、验证集、测试集包含的图像数量汇总如表 1 所示。

表 1 Age AP 害虫数据集分类体系

Table 1 Classification system for the Age AP pest dataset

害虫类别 Pest category		害虫虫态数 Number of insect stage	图像数量 Image number		
			训练集 Training set	验证集 Validation set	测试集 Test set
农田作物害虫 Field crops pest	水稻害虫	51	3 184	486	1 511
	玉米害虫	49	6 462	1 055	3 168
	小麦害虫	30	1 251	193	688
	甜菜害虫	32	2 112	353	1 022
	苜蓿害虫	46	3 445	529	1 704
经济作物害虫 Economic crops pest	葡萄害虫	55	7 721	1 318	3 910
	柑橘害虫	71	2 603	440	1 339
	芒果害虫	35	4 297	718	2 161
害虫数据集 Pest dataset Age AP		369	31 075	5 092	15 503

本研究进行大规模的数据筛选和虫态信息标注，构建了农业害虫领域内较为全面的害虫分类数据集。与 IP102 数据集相比，本数据集增加了对同类害虫不同生长时期的详细信息标注。

在害虫研究中，同一种类害虫在不同生长阶段可能展现出显著不同的形态，这要求识别并学习它们在各个生长阶段的共性特征。同时，不同种类的害虫处于相同生长阶段，也有可能呈现出形态上的相似性，这就需要辨识出它们之间的差异性特征。如图 1 所示，二化螟在其不同的发育阶段形态差异极大，而与三化螟在相同生长阶段，却表现出相似的形态。这些特点体现了该数据集所带来的研究挑战。

1.2 细粒度害虫分类模型

与体积庞大的 Transformer 模型相比，ConvNeXt V2 在保持较高性能的同时具备更低的计算复杂度，因而非常适合计算资源受限的应用场景。基于这些特性，本研究将 ConvNeXt V2 选定为基准模型。然而，该模型在小目标识别上仍存在一定局限，可能难以充分捕捉此类目标的关键信息。此外，作为通用模型，ConvNeXt V2 并未针对害虫的多虫态形态特征进行专门优化，从而在利用特定领域知识提升模型性能方面仍有不足。

为克服上述问题，本文在基准模型上提出了两项改进策略：1）多虫态监督策略：针对害虫在不同生长阶段呈现的显著形态差异，设计 2 个独立的子网络，分别提取种内共性特征和种间差异特征，以充分挖掘各类虫态信息；2）空间注意力机制：引入空间注意力模块，增强网络对图像中关键区域及小目标的捕捉能力，从而提升特征提取的精准度。

整体架构设计如图 2 所示，各模块协同作用，进一步提升了模型在细粒度害虫识别任务中的精度和鲁棒性。

1.3 特征提取器

特征提取器是本研究中针对害虫多样化形态特征进行提取的重要模块。本研究参考 Swin Transformer 的结构设计思路，采用相近的卷积模块堆叠策略构建特征提取模块，将 4 个阶段的卷积块比例设定为 1:1:3:1。通过

在高层阶段增加卷积块的堆叠深度，模型对图像整体信息的提取能力显著提升，从而有助于进一步提高害虫识别的精度和鲁棒性。

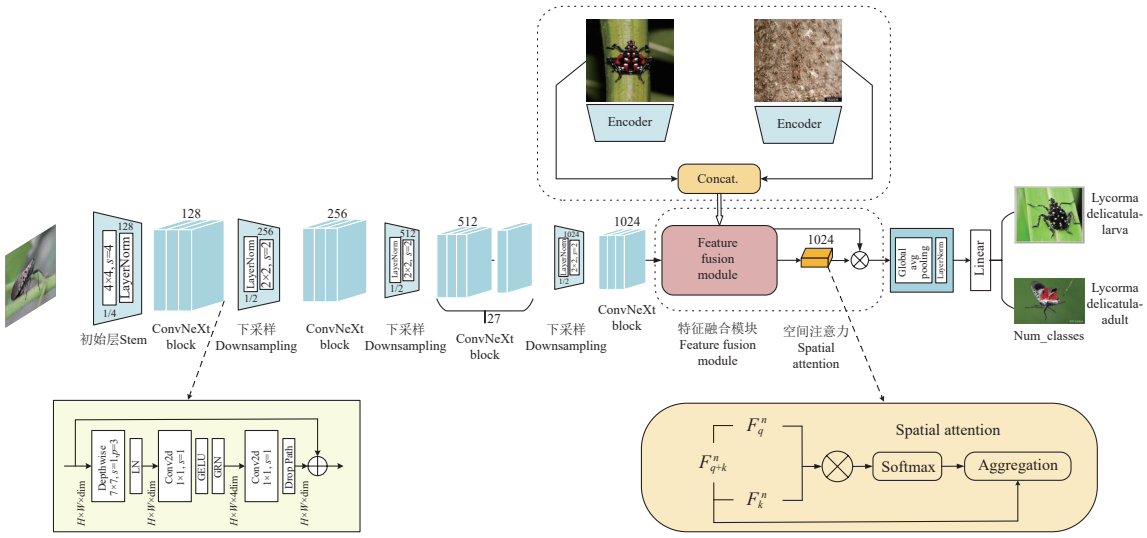
假设原始输入图像表示为P_s，$P_s \in \mathbb{R}^{H\times W\times C}$，其中 H、W、C 分别表示图像的高度、宽度、通道数。用函数$F(\cdot)$来表示特征提取块的操作，因此对输入其中的图片提取到的特征f_s可以表示为

$$f_s = F(P_S) \tag{1}$$

a. 二化螟卵　a. Asiatic rice borer egg
b. 二化螟幼虫　b. Asiatic rice borer larva
c. 二化螟蛹　c. Asiatic rice borer pupa
d. 二化螟成虫　d. Asiatic rice borer adult
e. 三化螟卵　e. Yellow rice borer egg
f. 三化螟幼虫　f. Yellow rice borer larva
g. 三化螟蛹　g. Yellow rice borer pupa
h. 三化螟成虫　h. Yellow rice borer adult

图 1　Age AP 数据集中具有挑战性的图片示例

Fig.1　Examples of challenging images in the Age AP dataset



注：⊕表示通道级别相加，⊗表示通道级相乘。Depthwise 表示深度可分离卷积，Conv2d 表示二维卷积，GELU 表示激活层，GRN 表示全局归一化，Drop Path 表示随机丢失路径，F^n_{q+k}表示第 n 阶段 Query 与 Key 连接的融合特征，F^n_q表示第 n 阶段 Query 特征，F^n_k表示第 n 阶段 Key 特征。p 表示填充的大小，s 表示步幅的大小，h 表示图像高度，w 表示图像的宽度，dim 表示图像的通道数。

Note: ⊕ denotes the channel level summation, ⊗ denotes channel-level multiplication, Depthwise denotes depth-separable convolution, Conv2d denotes 2D convolution, GELU denotes activation function, GRN denotes global normalization, and DropPath denotes random path dropping, F^n_{q+k} denotes the fusion feature of the nth stage Query and Key concatenation, F^n_q denotes the n^{th} stage Query feature, F^n_k denotes the n^{th} stage Key feature. p denotes the padding size; s denotes the stride; h denotes the image height; w denotes the image width; dim denotes the number of channels in the image.

图 2　改进 ConvNeXt 网络模型

Fig.2　Network model of improved ConvNeXt

1.4　多虫态融合

由于害虫在其生命周期内经历了显著的形态变化，这给网络提取虫态信息并进行合理分类带来了巨大的挑战。本研究中采用了 2 个独立的神经网络来分别捕捉同一种害虫的共性特征与不同害虫种间的个性特征。具体而言，个性特征通过 1.3 节中描述的特征提取器来学习，而共性特征则通过 Resnet50 的第一个残差块进行提取，Resnet50 的其他参数则共享。为了有效整合这些共性与个性特征，本研究还设计了一个特征融合模块，通过深度融合技术增强了模型的识别能力。

假定存在多张记录害虫生长阶段的图片，表示为$\{P_1, P_2, ..., P_n\}$，其中$n \geqslant 1$。首先需对每张图片应用特征提取函数$F(\cdot)$生成相应的特征图$\{F(P_1), F(P_2), ...F(P_n)\}$。随后，对获取的特征图进行拼接以便于整合不同级别的特

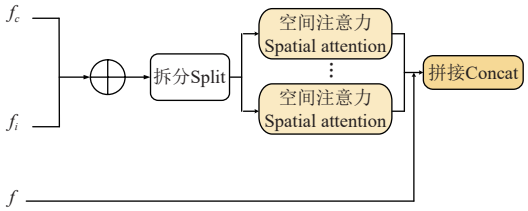
征信息。经过连接操作后，其特征图f可表示为

$$f = Concat(F(P_1), F(P_2), ..., F(P_n)) \tag{2}$$

对于学习到的同一种害虫种内的共性特征采用f_c表示，对于不同害虫种间的个性特征采用f_i表示。随后，将所有学习到的特征输入特征融合模块。其中，采用跳接的方式整合来自f的低层信息和来自f_c和f_i的多虫态特征，从而获得最终的特征表达，⊕表示通道级别相加。因此合并 2 个不同的虫态特征可表示为

$$f_{i+c} = f_i \oplus f_c \tag{3}$$

合并后的特征f_{i+c}分成 N 个区块，N 为注意力区块的总数，随后通过注意力模块得到特征图后进行合并操作。特征融合模块具体结构如图 3 所示。



注：f表示低层次提取的特征，f_i表示个性特征，f_c表示共性特征。
Note: f denotes low-level features, f_i denotes individual-specific features, and f_c denotes shared features.

图 3 特征融合模块结构

Fig.3 Structure of feature fusion module

1.5 空间注意力模块

考虑到图像中关键区域信息对目标识别的核心作用，本文引入空间注意力机制，旨在聚焦关键区域特征、降低背景干扰，从而提升识别精度。合并后的特征f_{i+c}，将其分割成 N 个f'并通过空间注意力模块获取注意力图$A_{spatial}$，可以表示为：

$$A_{spatial} = \sigma(f') = W \otimes \Phi(f') + f' \tag{4}$$

式中$\sigma(\cdot)$表示空间注意力模块，W是一个可学习的参数矩阵，⊗表示通道级相乘。$\Phi(\cdot)$表示空间注意力的具体操作，计算式如下：

$$\Phi(f') = \frac{\exp(f_i')}{\sum_{j=1}^{N} \exp(f_i')},\ i = 1, \cdots, N \tag{5}$$

式中$\exp(f_i')$表示对特征进行指数化，i，j 表示索引从 1 到 N，N 表示空间注意力的个数。

最后将这个注意力图应用于每一个特征图得到第 i 步特征F_i：

$$F_i = A_{spatical} \odot f \tag{6}$$

式中⊙表示逐元素乘法，用于加权。

AdamW 优化器更新策略按照以下表达式执行：

计算第 t 步梯度g_t：

$$g_t = \nabla_\theta L(\theta_t) \tag{7}$$

式中∇_θ为对参数θ进行梯度运算，$L(\theta_t)$表示第 t 步模型参数的损失函数值。

一阶动量均值m_t的递推式如式（8）所示：

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \tag{8}$$

式中m_{t-1}表示第 t-1 步的均值，β_1表示指数衰减因子，默认值为 0.9。

二阶动量未中心化方差v_t的递推式：

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \tag{9}$$

式中v_{t-1}表示第 t-1 步的均值，β_2为指数衰减因子，默认值为 0.999。

由于初始时刻递推值通常设置为 0，导致前期的近似值存在偏差，因此需要进行偏差修正。偏差修正后的均值$\hat{m}_t$计算式为

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{10}$$

式中β_1^t表示经过 t 个时间步后的累积影响程度。

偏差修正后的方差$\hat{v}_t$计算式为

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{11}$$

式中β_2^t表示经过 t 个时间步后的累积影响程度。

模型参数的更新：

$$\theta_{t+1} = \theta_t - lr \times \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} - \lambda \times \theta_t \tag{12}$$

式中θ_t表示第 t 步的模型参数，θ_{t+1}表示第 t+1 步的模型参数，lr 表示学习率，ε表示小常数，用于数值稳定性，λ表示权重衰减项。以上参数默认值为lr=0.000 02，$\varepsilon = 10^{-8}$，$\lambda = 0$。

1.6 评价指标

为了评估所提出的害虫识别模型的分类性能，本研究采用了以下评价指标：识别准确率（accuracy，A）、F1 值（F1 score），并使用模型参数量（parameters）作为模型复杂度的衡量指标。

A 表示正确预测的样本数占测试样本总数的比例，计算式如下：

$$A = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{13}$$

式中 TP 代表真阳性，即正确预测为阳性的样本数；FP 代表假阳性，即错误预测为阳性的样本数；TN 代表真阴性，即正确预测为阴性的样本数；FN 代表假阴性，即错误预测为阴性的样本数。

F1 值是精确率和召回率的调和平均数，计算式如下：

$$F1 = \frac{2P \cdot R}{P + R} \times 100\% \tag{14}$$

式中精确率（precision，P）指的是在所有预测为正样本的害虫图像中实际被正确识别的比例。召回率（recall，R）表示模型正确预测的正样本数量占实际正样本总数的比例。精确率和召回率的计算式如下：

$$\begin{cases} P = \dfrac{TP}{TP + FP} \times 100\% \\ R = \dfrac{TP}{TP + FN} \times 100\% \end{cases} \tag{15}$$

1.7　试验设置

本研究基于先前的图像分类技术，选择在 ImageNet 数据库上预训练的 ConvNeXt V2 网络的卷积模块作为特征提取器。为了增强模型的泛化能力，本研究采用了随机裁剪和水平翻转技术对训练数据进行数据增强。训练过程中，批处理大小设置为 32，最后特征的维度为 1 024。此外，使用了与 ConvNeXt V2 相同配置的 AdamW 优化器进行模型优化，并实施了渐进式热身策略以稳步提升学习效率。所有试验均在 PyTorch 框架下进行端到端训练，确保了试验的可复制性。为保证试验结果的可重复性，所有的随机种子均设定为 0。关于硬件配置，本研究使用 NVIDIA A40 GPU 48GB 显存进行训练和测试，训练阶段一轮约需 30 min，测试阶段一轮约需 5 min。

2　结果与分析

2.1　对比试验

为了全面评估本研究提出的网络在害虫虫态识别性能上的优越性，并考虑到其与 IP102 数据集相关问题的相似性，本研究选择了在 IP102 数据集上表现突出的几种深度网络模型进行性能对比。比较的模型包括 ResNet-50、ResNet-101、MobileNetV2、Vision Transformer、Swin Transformer、Swin Transformer V2、CoAtNet、ConvNeXt 及 ConvNeXt V2 等。在 Age AP 数据集上，研究重点是害虫虫态的细粒度分类性能，因此选择了在细粒度分类任务中表现优异的模型。为公平对比，所有模型均基于 PyTorch 框架实现，并采用一致的网络架构与训练策略（迁移学习及数据增强）。

根据表 2 的数据，与其他经典分类网络相比，本研究提出的模型在害虫识别准确率上表现最优。特别是，与试验性能最佳的 CoAtNet 网络相比，本研究模型的准确率提高了 5.07 个百分点，F1 分数提升了 5.48 个百分点。这证明了该模型设计的高效性和优越性。特别是在参数数量相似的情况下，该模型不仅在准确率上优于 ConvNext 和 CoAtNet，而且 F1 分数也展现了竞争力。此外，与传统的 CNN 模型如 ResNet-50 和 EfficientNet-B0 相比，尽管这些模型参数量较少，但在性能上仍有较大差距，显示了该方法在处理复杂图像数据时的强大能力。

表 2　主流网络模型在 Age AP 数据集上的性能对比

Table 2　Performance comparison of mainstream network models on Age AP dataset

模型 Model	准确率 Accuracy/%	F1 值 F1 score /%	参数量 Parameters/M
ResNet-50[16]	49.40	40.10	23.71
ResNet-101[16]	56.23	48.19	42.71
EfficientNet-B0[28]	63.76	60.32	5.29
Vision Transformer[20]	68.01	59.83	86.00
Swin Transformer[26]	65.12	58.53	88.00
Swin TransformerV2[29]	69.87	61.78	88.00
CoAtNet[30]	77.21	69.35	75.00
ConvNeXt[31]	72.13	61.39	89.00
ConvNeXt-V2[32]	75.60	68.76	89.00
本文	82.28	74.83	89.00

为了验证该方法的泛化能力，本研究不仅进行了针对不同害虫虫态的模型对比试验，还进行了跨数据集的评估，结果如表 3 所示。而在 CPB 和 IP102 数据集的测试中，本研究关注的是通用的害虫分类任务。这些数据集包含了更为广泛的害虫种类和更复杂的背景，因此选择了在 IP102 和 CPB 数据集中表现突出的 7 种模型。在 IP102 数据集上，本方法实现了 78.36% 的准确率和 76.85% 的 F1 分数，明显优于传统的 CNN 方法，如 VGGNet 和 ResNet-101，其中 VGGNet 的准确率为 43.65%，F1 分数为 40.21%，ResNet-101 的准确率为 47.19%，F1 分数为 41.06%。这说明本方法能够更有效地处理害虫图像的多样性和复杂性。

表 3　主流网络模型在 CPB 和 IP102 数据集上的测试结果

Table 3　Evaluation results of mainstream network models on the CPB and IP102 datasets

模型 Models	IP102		CPB		参数量 Parameters/M
	准确率 Accuracy/%	F1 值 F1 score/%	准确率 Accuracy/%	F1 值 F1 score/%	
EfficientNet-B0[28]	60.46	59.21	74.89	72.67	4.1
VGGNet[15]	43.65	40.21	60.15	54.16	
ResNet-101[16]	47.19	41.06	63.21	59.62	42.71
Attention-based MIL-Guided[33]	68.31	68.02	78.10	76.80	–
Vision Transformer[20]	72.47	72.04	75.92	74.08	86.00
ViT-B/16+ FRCF[22]	74.69	74.36	76.99	75.02	–
CA-EfficientNet[34]	69.45	63.06	–	–	5.38
Ensemble[35]	73.46	72.90	–	–	–
本文	78.36	76.85	80.16	77.39	89.00

注：“-”表示该模型无开源代码，或者在相关文献中未提供对应的数据。

Note: “-” indicates that there is no open source code for the model, or the corresponding data are not available in the relevant literature.

在 CPB 数据集中，尽管图像尺寸较小，该方法仍然达到了 80.16% 的准确率和 77.39% 的 F1 分数，表现优于 Attention-based MIL-Guided 方法，后者在调整图像大小后的准确率为 78.10%，F1 分数为 76.80%。这进一步证明了本模型在小图像数据上的鲁棒性和高效性。

从技术角度看，该方法主要依赖于多虫态共同监督和空间注意力模块。多虫态监督使得模型可以更精细地学习并区分不同害虫的生长阶段，而空间注意力模块则提高了模型对细节的捕捉能力，这对于精确识别小尺寸图像中的害虫尤为关键。这些技术的结合不仅提高了准确率和 F1 分数，而且增强了模型对不同数据集和变化条件下的适应能力，使得该方法成为害虫分类领域中一种高效且可靠的技术。

2.2　消融试验

为了验证本研究提出模型的有效性，本研究开展了一系列的消融试验。

农业害虫图像识别面临目标小、背景干扰强及类间差异细微等挑战。针对小目标特征提取，本研究分析了不同卷积核尺寸的影响：较小的卷积核可能更擅长捕捉局部细节特征，而较大的卷积核可能有助于获取更大的感受野，从而捕捉全局信息。基于此，设计了卷积核尺寸消融试验。

表 4 总结了主干网络内深度可分离卷积模块的消融

分析结果。试验结果表明随着卷积核的增大，试验效果也一直在提升，网络学习全局信息的能力也在提高，但是当卷积核大小达到 7，性能出现瓶颈。所以最终选定卷积核大小为 7，试验效果最佳。较小的卷积核（如 3）可能不足以捕捉害虫的关键特征，而过大的卷积核（如 9）可能引入过多的背景噪声。这表明，针对小目标的害虫分类任务，卷积核大小需要进行调整，以平衡局部细节和全局信息的提取。

表 4 不同卷积核大小消融试验结果

Table 4 Results of ablation experiments with different convolutional kernel sizes

卷积核大小 Kernel size	准确率 Accuracy/%	F1 值 F1 score/%
3	81.03	73.56
5	82.01	74.67
7	82.28	74.83
9	82.15	74.69

表 5 为模型的消融试验研究。当不采用多虫态监督和空间注意力时，准确率和 F1 分数分别只有 76.50% 和 67.65%。但是当采用多虫态监督策略后，分类准确率提高了 3.81 个百分点，F1 分数提升了 5.33 个百分点；单独使用空间注意力机制时，准确率和 F1 分数分别提高了 2.76 和 3.18 个百分点；而综合本研究提出的方法后，准确率和 F1 分数分别提升了 5.78 和 7.18 个百分点。这表明了本研究采用的策略的有效性。

表 5 模型消融试验结果

Table 5 Results of model ablation experiments

数据增强 Data augmentation	多虫态监督 Multi-stage co-supervision	空间注意力 Spatial attention	准确率 Accuracy/%	F1 值 F1 score/%
			75.56	66.73
√			76.50	67.65
√	√		80.31	72.98
√		√	79.26	70.83
√	√	√	82.28	74.83

注：“√”表示执行此操作。

Note: "√" indicates that the iperation is performed.

2.3 可视化分析

本研究采用 Grad-CAM 等[27]可视化技术对模型处理的害虫数据进行对比分析，直观展示了模型的性能，如图 4 所示。从图 4b 中可见，该模型能够精确地聚焦于害虫所在位置，对背景区域的关注度较低，有效地将害虫与背景区分开，降低了背景的干扰。图 4c 显示，采用 Grad-CAM 技术将斑衣蜡蝉的幼虫阶段和成虫阶段进行了类激活可视化；在成虫阶段，模型预测类别和生长阶段的准确率为 0.885；而在幼虫阶段，该准确率达到 0.913。尽管同一种害虫在不同生长阶段的形态特征存在明显差异，该模型依然能够识别出各个生长阶段的独特特征，并准确地将其归类。这一结果表明，多虫态共同监督策略能有效促进模型学习不同生长时期的特异性特征及其共性特征。

此外，本研究还将每个特征通道的激活图进行了可视化，如图 5 所示。虽然各通道学习到的特征存在差异，但本模型能够精准地分割出害虫在不同生长阶段的轮廓，并有效地抑制背景噪声，对于高维的特征识别较为准确。而从激活图中可以显示模型主要关注的害虫特征区域，可以看出，该模型关注重点在于害虫各虫态差异部位，这也能解释模型对于虫态的识别能力，从而提升不同虫态归属于同一类害虫的准确率。

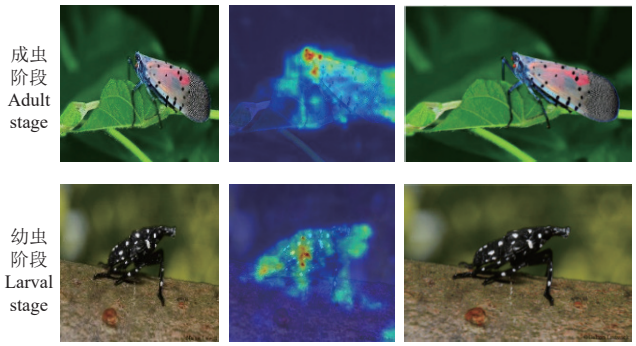
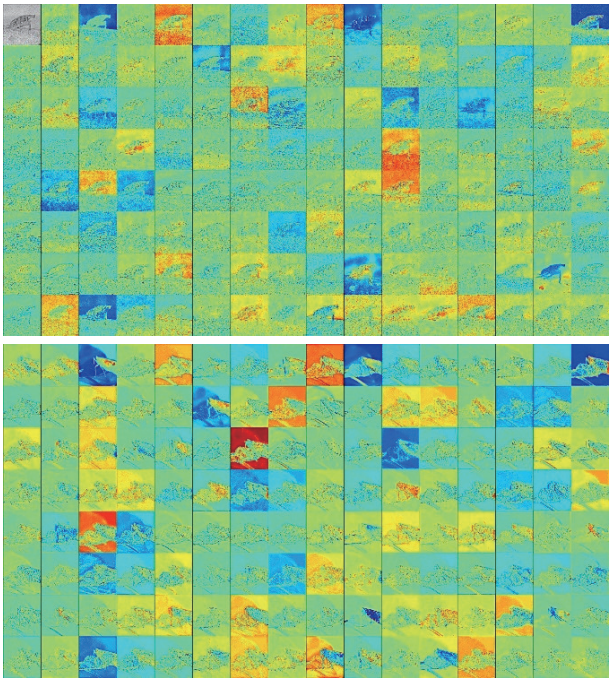


图 4 斑衣蜡蝉的幼虫和成虫阶段类激活图可视化

Fig.4 Visualization of class activation maps for larval and adult stages of spotted lanternfly



注：图中每个特征通道的激活图都用小方块表示。为确保清晰度，每次可视化显示 128 个通道。

Note:Activation maps for each feature channel are depicted using small squares. To ensure clarity, 128 channels were displayed in each visualization.

图 5 斑衣蜡蝉幼虫及成虫图像的特征图可视化

Fig.5 Visualization of feature maps for larval and adult images of spotted lanternfly

3 结 论

本研究以细粒度农作物害虫分类为研究对象，针对害虫在不同生长阶段中形态变化显著、识别精度不足以及目标定位困难等挑战，在 ConvNeXt V2 网络的基础上进行改进，并在该任务上取得更好的分类效果。主要研究结论如下：

1）构建了包含 102 类害虫的 369 种虫态，共计 51 670 张图像的数据集 Age AP，为后续研究提供了针对虫态的害虫数据集；

2）设计了多虫态共同监督机制，对网络结构进行整合与优化，从而分别提取并融合害虫在不同虫态下的共性与个性特征，显著提高了模型对害虫形态的识别能力。试验表明，增加了多虫态共同监督机制后，该模型在 Age AP 测试集上的准确率相比基础模型提升了 3.81 个百分点，在 F1 分数上提高了 5.33 个百分点；

3）引入空间注意力模块，在特征融合完成后，于归一化前加入空间注意力模块，增强模型对小目标位置信息的提取能力。试验结果表明，引入空间注意力机制后，该模型在 Age AP 测试集上的准确率和 F1 分数上分别提高了 2.76 个百分点和 3.18 个百分点；

本研究表明多虫态共同监督机制对害虫细粒度分类任务有显著的提升，从可视化的结果中也直观显示本模型对害虫不同虫态的识别性能有明显的改进。同时，所提出的改进策略并未显著增加模型参数量，较好地平衡了识别精度与计算效率，在农业生产实际应用中具备较好的应用前景。

[参　考　文　献]

[1] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, UA, USA: IEEE, (CVPR'05), 2005, 1: 886-893.

[3] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60: 91-110.

[4] WEEKS P J, O'NEILL M A, GASTON K J, et al. Species–identification of wasps using principal component associative memories[J]. Image and Vision Computing, 1999, 17(12): 861-866.

[5] SAMANTA R K, GHOSH I. Tea insect pests classification based on artificial neural networks[J]. International Journal of Computer Engineering Science (IJCES), 2012, 2(6): 1-13.

[6] MANOJA M, RAJALAKSHMI J. Early detection of pest on leaves using support vector machine[J]. International Journal of Electrical and Electronics Research, 2014, 2(4): 187-194.

[7] RANI R U, AMSINI P. Pest identification in leaf images using SVM classifier[J]. International Journal of Computational Intelligence and Informatics, 2016, 6(1): 248-260.

[8] HINTON G, LECUN Y, BENGIO Y. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[9] CHEN L, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[11] LIU Z, GAO J, YANG G, et al. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network[J]. Scientific Reports, 2016, 6(1): 20410.

[12] WU X, ZHAN C, LAI Y-K, et al. Ip102: A large-scale benchmark dataset for insect pest recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 8787-8796.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[14] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, US, USA: IEEE, 2015: 1-9.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations, San Diego, CA, USA: Computational and Biological Learning Society, 2015: 1-14.

[16] 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,Las Vegas, NV, USA: IEEE, 2016: 770-778.

[17] BOLLIS E, PEDRINI H, Avila S. Weakly supervised learning guided by activation mapping applied to a novel citrus pest benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual Conference: IEEE,2020: 70-71.

[18] VALAN M, MAKONYI K, MAKI A, et al. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks[J]. Systematic Biology, 2019, 68(6): 876-895.

[19] DEVLIN J, CHANG MW, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Minneapolis, MN, USA: Association for Computational Linguistics, 2019: 4171-4186.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//Proceedings of the International Conference on Learning Representations, 2021: 1-22.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010.

[22] LIU H, ZHAN Y, XIA H, et al. Self-supervised transformer-based pre-training method using latent semantic masking auto-encoder for pest and disease classification[J]. Computers and Electronics in Agriculture, 2022, 203: 107448.

[23] 彭红星，徐慧明，刘华鼐. 基于改进 ShuffleNet V2 的轻量化农作物害虫识别模型[J]. 农业工程学报，2022，38(11)：161-170.
PENG Hongxing, XU Huiming, LIU Huanai. Lightweight agricultural crops pest identification model using improved ShuffleNet V2[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(11): 161-170. (in Chinese with English abstract)

[24] 张佳敏，闫科，王一非，等. 基于改进 Mask-RCNN 算法的作物害虫分类识别[J]. 农业工程学报，2024，40(7)：202-209.
ZHANG Jiamin, YAN Ke, WANG Yifei, et al. Classification and identification of crop pests using improved Mask-RCNN algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(7): 202-209. (in Chinese with English abstract)

[25] 国家基础学科公共科学数据中心. 农业病虫害研究图库（IDADP）数据库 [DB/OL]. (2023-02-22). https://www.heywhale.com/mw/dataset/63e50cfea2c1716e14fb9db6，2023.
[26] 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, Virtual Conference: IEEE, 2021: 10012-10022.
[27] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy: IEEE, 2017: 618-626.
[28] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning.Long Beach, CA, USA: PMLR,2019: 6105-6114.
[29] LIU Z, HU H, LIN Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA: IEEE,2022: 12009-12019.
[30] DAI Z, LIU H, LE Q V, et al. Coatnet: Marrying convolution and attention for all data sizes[J]. Advances in Neural Information Processing Systems, 2021, 34: 3965-3977.
[31] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA: IEEE, 2022: 11976-11986.
[32] WOO S, DEBNATH S, HU R, et al. Convnext v2: Co-designing and scaling convnets with masked autoencoders[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada: IEEE, 2023: 16133-16142.
[33] BOLLIS E, MAIA H, PEDRINI H, et al. Weakly supervised attention-based models using activation maps for citrus mite and insect pest classification[J]. Computers and Electronics in Agriculture, 2022, 195: 106839.
[34] 甘雨，郭庆文，王春桃，等. 基于改进 EfficientNet 模型的作物害虫识别[J]. 农业工程学报，2022，38(1)：203-211.
GAN Yu, GUO Qingwen, WANG Chuntao, et al. Recognizing crop pests using an improved EfficientNet model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(1): 203-211. (in Chinese with English abstract)
[35] NANNI L, MANFè A, MAGUOLO G, et al. High performing ensemble of convolutional neural networks for insect pest image detection[J]. Ecological Informatics, 2022, 67: 101515.

Fine-grained identification of crop pests using an enhanced ConvNeXt model

HAN Yuantao[1] , ZHANG Cong[2※] , ZHAN Xiaoyun[2] , WANG Zheng[3]

(1. *School of Mathematics and Computer Science*, *Wuhan Polytechnic University*, *Wuhan* 430048, *China*; 2. *School of Electrical and Electronic Engineering*, *Wuhan Polytechnic University*, *Wuhan* 430048, *China*; 3. *School of Computer Science*, *Wuhan University*, *Wuhan* 430072, *China*)

Abstract: Precise and rapid identification of diverse pest species can greatly contribute to crop disease prevention and control in modern agriculture. However, the accuracy of pest identification has been frequently confined to the varied insect stages during different pest growth. Among them, the same pest can display distinctly different morphological features across various growth stages, while the different pests can exhibit similar morphologies in the same developmental periods. Both manual identification and machine learning approaches can often struggle to fully meet these demands of complex recognition. In this study, fine-grained identification was performed on the crop pests using an enhanced ConvNeXt model. A series of experiments were also carried out on the large-scale pest dataset with the morphological diversity of insects. The large-scale dataset contained 102 pest categories and 51 670 images representing 369 classes of pests at different stages. The largest dataset was focused mainly on the whole stages of insects; Each image was precisely labelled with the pest species and their developmental stages. A robust foundation was provided for the subsequent morphological studies. Furthermore, the ConvNeXt V2 was adopted as the baseline model. A multi-stage co-supervision strategy was then introduced to optimize the structure for better feature variability of the same species across different pest stages, as well as the significant inter-species differences. Two independent streams of neural networks were also constructed during optimization. Specifically, the species-specific features were learned by the feature extraction module within the ConvNeXt Block. While the shared features were derived through the first residual block of ResNet50, and then shared with the subsequent parameters of layers. A feature fusion module was then employed to effectively integrate these shared and species-specific features. A deep feature fusion was also designed to enhance the overall performance of recognition. Moreover, there were pronounced morphological differences among various pest species, leading to the varying spatial location. Therefore, the spatial attention module was further introduced to improve the sensitivity of the model to the spatial distribution of the target. Comparison experiments were conducted on the large public dataset IP102. The results demonstrate that the accuracy and F1 score of the model was improved by 3.67 and 2.49 percentage points, respectively, compared with the state-of-the-art models. Meanwhile, the corresponding metrics were improved by 5.07 and 5.48 percentage points, respectively, on the Age AP dataset. There were increases of 2.06 and 0.59 percentage points, respectively, on the CPB dataset. Ablation experiments show that the accuracy increased by 3.81 percentage points, compared with the original baseline model, when only the multi-stage co-supervision was adopted; The spatial attention module was raised by an additional 2.76 percentage points; These strategies ultimately improved the accuracy by 5.77 percentage points, compared with the original model. Significant improvements were achieved in the feature extraction and spatial information capture across multiple pest stages. Better performance was also achieved, compared with similar models. A reliable scheme was also presented for crop pest identification in smart agriculture.

Keywords: pest recognition; crops; ConvNeXt; spatial attention mechanism; recognition of multiple insect stages