

张善文, 许新华. 基于改进视觉 Mamba-UNet 模型的作物害虫检测方法[J]. 江苏农业学报, 2026, 42(6): 1191-1201.
doi:10.3969/j.issn.1000-4440.2026.06.011

基于改进视觉 Mamba-UNet 模型的作物害虫检测方法

张善文^{1,2}, 许新华¹

(1. 郑州西亚斯学院工学部, 河南 新郑 451150; 2. 河南省农业信息数智化工程研究中心, 河南 新郑 451150)

摘要: 针对田间自然环境下作物害虫种类多样、体积微小、姿态多变及背景复杂等因素导致的检测难题, 以及传统 U-Net 模型全局特征提取能力不足、Transformer 模型计算复杂度高等问题, 本研究提出一种融合空间与通道注意力视觉的 Mamba-UNet 模型的作物害虫检测模型——SCAFVM-UNet。该模型以 U-Mamba 模型架构为基础, 采用编码器-解码器结构, 该模型整合了主干起始模块(Stem)、残差视觉 Mamba 模块(RVM)和视觉状态空间模块(VSS)、空间通道注意力视觉状态空间模块(SCAVSS)、多尺度自适应特征聚合模块(MSAFA)以及交叉二维扫描机制(CS2D)。在 IP102 数据集上的试验结果表明, SCAFAVM-UNet 模型的检测精度与 Dice 相似系数(DSC)分别达到 76.12% 和 73.11%, 并且其检测精度与训练效率整体优于 U-Net 模型、DMSU-Net 模型、MSRAL-UNet 模型、TransFCL 模型、Swin-UNet 模型、Mamba-UNet 模型以及 MSV-Mamba 模型。在害虫目标微小、触角及足部细节丰富且与背景对比度较低的复杂场景下, 该模型能够更精准地定位害虫, 并保持轮廓完整性, 显著减少漏检与目标残缺现象。本研究结果为智慧农业中田间害虫的精准监测与防治决策提供了理论依据。

关键词: 作物害虫; Mamba-UNet 模型; 视觉状态空间模块(VSS); 空间与通道注意力

中图分类号: S433 **文献标识码:** A **文章编号:** 1000-4440(2026)06-1191-11

Crop pest detection method based on improved visual Mamba-UNet model

ZHANG Shanwen^{1,2}, XU Xinhua¹

(1. SIAS Faculty of Engineering, Xinzheng 451150, China; 2. Henan Engineering Research Center for Agricultural Information Digital & Intelligent Technology, Xinzheng 451150, China)

Abstract: Aiming at the detection challenges caused by diverse species, small size, variable postures, and complex backgrounds of crop pests in field natural environments, as well as the insufficient global feature extraction ability of the traditional U-Net model and high computational complexity of the Transformer model, this study proposed a crop pest detection model named SCAFAVM-UNet, which is a visual Mamba-UNet model integrating spatial and channel attention. Based on the U-Mamba architecture, the model adopted an encoder-decoder structure and integrated a Stem module, residual visual Mamba module (RVM), visual state space module (VSS), spatial and channel attention visual state space module (SCAVSS), multi-scale adaptive feature aggregation module (MSAFA), and a crisscross 2D scanning mechanism (CS2D). Experimental results on the IP102 dataset showed that the detection accuracy and Dice similarity coefficient

(DSC) of the SCAFAVM-UNet model reached 76.12% and 73.11%, respectively, and its overall detection accuracy and training efficiency were superior to those of the U-Net, DMSU-Net, MSRAL-UNet, TransFCL, Swin-UNet, Mamba-UNet, and MSV-Mamba models. In complex scenes with tiny pest targets, rich details of antennae

收稿日期: 2025-06-05

基金项目: 河南省科技攻关项目(252102110343, 242102210021); 河南省高等学校重点科研项目(24B510016)

作者简介: 张善文(1965-), 男, 陕西西安人, 博士, 教授, 研究方向为深度学习及其应用。(E-mail) wjd716@163.com

and feet, and low contrast with the background, the model can locate pests more accurately and maintain contour integrity, significantly reducing missed detection and target incompleteness. The results of this study provide a theoretical basis for the accurate monitoring and control decision-making of field pests in smart agriculture.

Key words: crop pests; Mamba-UNet model; visual state space module (VSS); spatial and channel attention

田间害虫严重威胁作物的产量与品质,及时、准确地开展害虫检测已成为实现有效防治的关键前提^[1-2]。传统方式依赖人工检测效率低下,且易受主观因素影响。基于深度学习的检测方法,特别是 YOLO 模型、U-Net 模型及 Transformer 系列模型近年来在作物害虫检测领域得到了广泛应用^[3-5]。为提升对多尺度和小尺度害虫的检测能力,研究人员在经典 YOLO 模型中引入注意力机制、多尺度卷积及多尺度预测损失等策略,提出了一系列改进 YOLO 模型^[6-7]。U-Net 模型凭借其独特的编码器-解码器结构与跳跃连接机制及其持续优化的改进模型,在害虫图像分割与检测任务中展现出强大的潜力^[8]。针对传统方法在农作物分类中存在遥感图像特征提取不充分和识别精度低等问题,李安琦等^[9]提出了一种改进的 U-Net 模型,通过引入多尺度特征融合与结构深化,有效提升了分类准确率。然而,尽管 YOLO 模型与 U-Net 模型及其改进模型在局部特征提取方面表现优异,但在捕捉长距离依赖关系方面仍存在局限^[10]。

随着深度学习技术的发展,Transformer 模型通过其自注意力机制,动态聚焦于输入数据的重要部分,克服了传统深度学习模型难以并行处理的问题,有效解决了 YOLO 模型和 U-Net 模型在捕捉远程依赖关系方面的局限性^[11]。为提升图像处理性能,研究人员提出视觉变换器(ViTs),将图像视为一系列图块,相比 YOLO 模型和 U-Net 模型,其在全局依赖建模方面更具优势^[12]。Liu 等^[13]提出一种基于变换器的特征表示补偿与注意力机制相结合的端到端害虫检测方法,通过减少特征信息损失,在不增加计算量的前提下显著提升了检测性能。Cao 等^[14]将 U-Net 模型与变换器结合,提出混合深度学习模型 Swin-UNet,在图像分割任务中取得了良好效果。尽管变换器与 ViTs 能有效捕获全局依赖,并在多个视觉任务中表现优异,但其二次计算复杂度限制了其在高分辨率和实时图像检测中的应用^[15-16]。

近期研究结果表明,以 Mamba 为代表的状态空间模型(SSM)不仅能够高效建模长距离依赖,还具

备线性计算复杂度,克服了 YOLO 模型与 U-Net 模型在远程依赖建模方面的不足,同时规避了变换器的二次复杂度问题^[17]。视觉 Mamba (VMamba) 通过将状态空间模型与注意力机制相结合,衍生出 Mamba-UNet 和 VM-UNet 等改进模型^[18-19]。Swin-UMamba 模型和 LightM-UNet 模型作为 Mamba-UNet 模型的改进模型,在医学图像分割中表现突出^[20-21]。InsectMamba 模型是一种基于 Mamba 架构的害虫分类模型,融合了状态空间模型(SSM)、卷积神经网络(CNN)、多头自注意力机制(MSA)与多层感知器(MLP),在多个害虫图像数据集上取得了优异性能,但其仍依赖于大规模标注数据,且计算成本较高^[22]。MSV-Mamba 模型则结合了 Swin-Mamba 模块与层次化特征融合机制,显著提升了复杂图像的分割性能^[23]。为此,本研究拟提出一种融合空间与通道注意力的视觉 Mamba-UNet 模型,用于大田作物害虫检测,并在作物害虫数据集上验证其有效性与可行性,以期为复杂背景下的作物害虫智能检测提供一种高效解决方案。

1 材料与方法

1.1 数据集构建

如图 1A 所示,本研究选用 IP102 害虫图像数据集(<https://github.com/xpwu95/IP102>)中玉米、水稻、芒果、小麦、柑橘、葡萄、甜菜、苜蓿 8 种作物的 102 种害虫图像,共 75 000 幅。

该数据集中的害虫图像分布不平衡,水稻蓟马害虫图像和纵卷叶螟害虫图像分别为 173 幅和 1 115 幅。如图 1B 所示,为了克服害虫图像类别的不平衡和提高模型的泛化能力,通过随机移动、缩放、旋转、裁剪、倒转、镜像等变换扩充图像子集。

1.2 SCAFAVM-UNet 模型构建

1.2.1 VM-UNet 模型构建 VM-UNet 模型结合了 U-Net 模型、VMamba 模型和 Mamba-UNet 模型的设计理念,其架构如图 2 所示。对于输入大小为 $H \times W \times 3$ (H : 图像高度, W : 图像宽度) 的二维图像,首先将其划分为 16 个 $H/4 \times W/4$ 的图像块(Patch),并重



图1 IP102 数据集中害虫图像

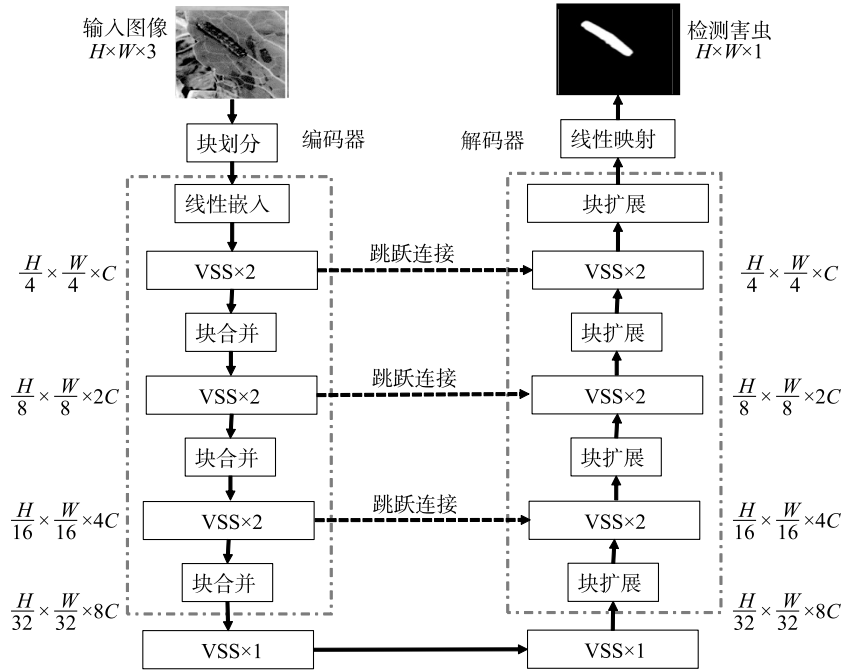
Fig.1 Pest images in the IP102 dataset

塑为 16 个一维序列,随后通过线性嵌入层将特征维度调整为 C (通道数),作为编码器的初始输入。编码器由多个视觉状态空间模块(VSS module)和块合并层模块交替构成,其中块合并层模块负责降采样和维度扩展,视觉状态空间模块则用于学习分层特征表示。经过 4 个阶段的编码后,各层输出特征图的分辨率依次为 $H/4 \times W/4 \times C$ 、 $H/8 \times W/8 \times 2C$ 、 $H/16 \times W/16 \times 4C$ 和 $H/32 \times W/32 \times 8C$ 。解码器采用对称结构,由视觉状态空间模块和块扩展层组成,通过逐步上采样恢复特征图分辨率,并通过跳跃连接融合编码器对应层的特征。最终,经线性映射层输出检测到的害虫。

1.2.2 SCAFAVM-UNet 模型构建 害虫目标尺度小、姿态多变,且常伴随光照变化、遮挡等,这些因素会干扰模型检测。并且害虫的细小部位(如触角、腿)及特定纹理特征,极易被传统卷积神经网络丢失。为了提升模型复杂田间环境下多尺度害虫的检

测性能,本研究在 VM-UNet 模型基础上进行改进,得到 SCAFAVM-UNet 模型。如图 3 所示,SCAFAVM-UNet 模型融合了残差网络、空洞卷积、多尺度卷积与 VMamba 模型的优势,其整体架构包括主干起始模块(Stem)、4 个块合并层模块、4 个残差视觉 Mamba 模块(RVM)、4 个空间通道注意力视觉状态空间模块(SCAVSS)、1 个多尺度自适应特征聚合模块(MSAFA)及上采样模块。

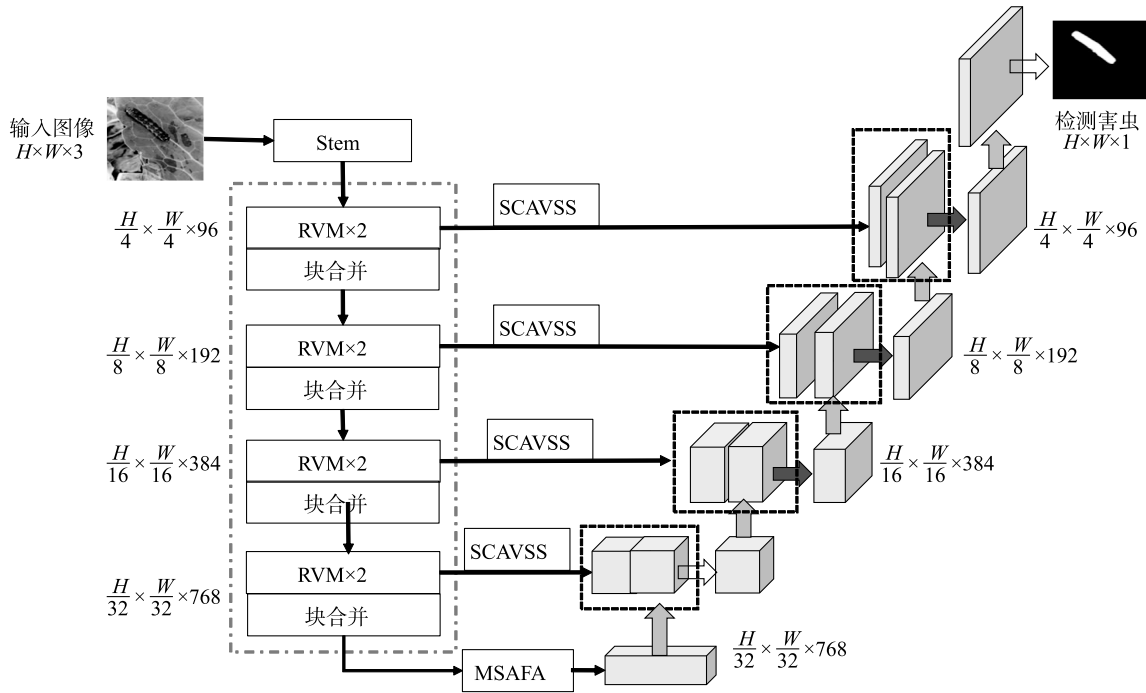
1.2.2.1 主干起始模块(Stem) 在 SCAFAVM-UNet 模型中,对于大小为 $H \times W \times 3$ 的作物害虫输入图像,首先通过主干起始模块(Stem)进行特征嵌入,该模块通过一系列卷积运算逐步降低空间分辨率并增加通道维度,实现图像的嵌入,从而有效保留重要的空间信息。Stem 模块由 5 个卷积层、残差连接、归一化层和激活函数构成,其目的是将原始输入图像转换为适合 Mamba 模块处理的特征,具体过程如下:



VSS:视觉状态空间模块; W :图像宽度; H :图像高度; C :通道数。

图 2 VM-UNet 模型架构

Fig.2 VM-UNet model architecture



↑上采样; ➡全连接; ⇨预测; ⊕加法; □特征拼接。Conv:卷积层; Layer Norm:层归一化操作; RVM:残差视觉Mamba模块; Stem:主干起始模块; SCAVSS:空间通道注意力视觉状态空间模块; MSAFA:多尺度自适应特征聚合模块; W :图像宽度; H :图像高度; C :通道数。

图 3 SCAFAVM-UNet 模型架构

Fig.3 SCAFAVM-UNet model architecture

$$x_0 = F_{\text{Conv3}} \{ F_{\text{Conv2}} [F_{\text{Conv1}} (I_{\text{in}}) \oplus F_{\text{Conv1}} (I_{\text{in}})] \} \quad (1)$$

式中, I_{in} 为输入图像; F_{Conv1} 、 F_{Conv2} 、 F_{Conv3} 为卷积。

1.2.2.2 残差视觉 Mamba 模块 (RVM) 为实现下采样, 采用 4 个合并模块进行特征学习。各层级输出的特征图分辨率依次变为 $H/4 \times W/4 \times 96$ 、 $H/8 \times W/8 \times 192$ 、 $H/16 \times W/16 \times 384$ 和 $H/32 \times W/32 \times 768$ 。这些模块通过整合相邻 2×2 图像块的特征, 并经过线性投影实现下采样, 从而将空间分辨率从 $H \times W$ 降低至 $H/2 \times W/2$, 同时通道数由 C 增至 $2C$ 。

解码器与编码器对称, 采用全连接层和上采样操作逐步恢复特征图分辨率, 最终通过预测分类器完成害虫检测任务。相较于传统卷积和下采样方法, 合并残差视觉 Mamba 模块 (RVM) 不仅提升了作物害虫检测的精度, 也显著降低了计算复杂度。RVM 模块结构如图 4(A) 所示, 其内部集成了 VSS 模块。

在 RVM 模块中, SSM 模块通过交叉扫描模块 (CSM) 实现从输入 x_t 到输出 y_t 的映射, 其过程如下:

$$x_{t+1} = A \cdot x_t + B \cdot u_t + w_t \quad (2)$$

式中, x_t 为 t 时刻的模型状态; u_t 为控制输入; w_t 为过程噪声和状态转移的不确定性; A 为状态矩阵; B 为投影参数。

t 时刻的观测值 y_t 计算公式如下:

$$y_t = D \cdot x_t + E \cdot u_t + v_t \quad (3)$$

式中, v_t 为观测噪声, 用于表征模型预测的观测值与实际观测值之间的差异。

公式(2)和(3)中的矩阵 A 、 D 和 E 共同建立了状态变量与观测值之间的映射关系。交叉扫描模块 (CSM) 通过从右下、左上、右上和左下 4 个方向对像素进行扫描与集成, 实现了对整个空间域的遍历, 其计算复杂度为线性阶 $O(N)$ 。通过这种遍历机制, 任意非因果视觉图像被转换为有序的块序列, 过程如下:

$$\text{CSM}(V) = \text{Order}[\text{Traverse}(V)] \quad (4)$$

式中, V 为视觉特征输入; Traverse 为 CSM 模块所执行的遍历操作。

为进一步提升 SSM 模块对图像深层语义特征的提取能力, 引入了 RVM 模块。RVM 模块通过引入残差连接与可调因子, 在不增加额外参数和计算负担的前提下, 增强了 SSM 模块在远程空间依赖关系建模方面的性能。对于输入的深度特征

F_{in}^l , RVM 模块首先执行层归一化操作, 随后利用 VSS 模块捕获图像中的远程空间依赖关系。最后, 通过残差连接引入一个调整因子 s 以优化模型性能, 过程如下:

$$F^l = \text{VSSM}[\text{LayerNorm}(F_{\text{in}}^l)] + sF_{\text{in}}^l \quad (5)$$

式中, LayerNorm 为层归一化操作, s 为调整因子, 默认值为 0.001。在完成上述特征增强后, 通过映射层将所得特征进一步转化为深层次特征, 过程如下:

$$F_{\text{out}}^l = \text{Proj}[\text{LayerNorm}(F^l)] \quad (6)$$

式中, Proj 为线性映射操作。

VSS 模块用于对远程空间依赖关系进行建模, 其处理流程如下: 给定输入特征 G_{in}^l , VSS 模块将其分别引入两个并行的处理分支。在第 1 个分支中, 特征首先通过 1 个线性层进行通道扩展, 随后依次经过深度可分离卷积 (DWConv) 模块、SiLU 激活函数、SSM 模块以及层归一化 (LayerNorm) 操作。在第 2 个分支中, 特征先通过一个线性层进行通道扩展, 然后应用 SiLU 函数激活。最后, VSS 模块通过 Hadamard 乘积操作融合两个分支的输出特征。融合后的特征经过一个线性层将通道数投影回原始维度 C , 最终得到与输入特征具有相同空间尺寸和通道数的输出。过程如下:

$$\begin{aligned} G_1 &= \text{LayerNorm}[\text{SSM}[\text{SiLU}(\text{DWConv}(\text{linear}(G^l)))] \\ G_2 &= \text{SiLU}[\text{Linear}(G^l)] \end{aligned} \quad (7)$$

$$G_{\text{out}} = \text{Linear}(G^l \otimes G_2)$$

式中, $\otimes$ 为 Hadamard 乘操作。

VSS 模块中的缩放因子主要用于特征归一化, 通过对 SSM 模块输出特征进行动态缩放, 有效缓解梯度爆炸或消失等问题, 提升特征稳定性。该机制具备通道或空间维度的自适应调整能力, 有助于抑制无关区域响应, 增强关键特征表达。

1.2.2.3 空间通道注意力视觉状态空间模块 (SCAVSS) 空间通道注意力视觉状态空间模块 (SCAVSS) 是一种嵌入空间与通道注意力机制并引入跳跃连接的改进 VSS 模块, 其可以在融合编码器的多尺度特征与解码器的上采样特征, 减少空间信息丢失。如图 4(B) 所示, SCAVSS 模块结构由 1 个 VSS 模块、1 个层归一化 (LayerNorm)、1 个卷积层、2 个并行的全局平均池化与全局最大池化操作以及 2 个残差连接组成。该模块的处理流程如下:

$$\begin{aligned}
F^{VSS} &= VSS(F_i^e) \\
F_{LN}^{VSS} &= \text{Conv}[\text{LayerNorm}(F^{VSS})] \\
F^{CAVSS} &= \text{Sigmoid}[\text{Avg}(F_{LN}^{VSS}), \text{Max}(F_{LN}^{VSS})] \otimes F_{LN}^{VSS} + F^{VSS}
\end{aligned} \quad (8)$$

式中, F_i^e 和 F^{CAVSS} 分别为模型的输入和输出, $VSS(\cdot)$ 为 VSS 操作, Sigmoid 为激活函数。

与标准跳跃连接不同, SCAVSS 能够自适应地选择信息通道表示, 有效缩小多尺度编码器特征与解码器特征之间的语义差距。这种选择性注意机制对于保留在降采样过程中易丢失的精细空间细节至关重要。其中, VSS 模块负责增强远程依赖关系建模与全局上下文信息提取, 而全局平均池化与全局最大池化操作则共同捕获丰富的语义表示。

1.2.2.4 多尺度自适应特征聚合模块(MSAFA)

为了增强模型在不同空间尺度下提取各类缺陷特征的能力, 在瓶颈层引入了多尺度自适应特征聚合模块(MSAFA)。其结构如图 4(C) 所示, 由通道注意力(CA)分支、空间注意力(SA)分支及残差连接 3 个并行支路组成。在空间注意力分支中, 依次经过 1×1 卷积、空洞多尺度卷积 Inception 模块(空洞率分别设置为 1、2 和 3)、最大池化、残差连接、 7×7 卷积及 Sigmoid 激活函数。在通道注意力分支中, 输入特征图依次经过全局平均池化、两个 1×1 卷积及 ReLU 激活函数。两分支的输出特征先通过逐元素相乘进行融合, 再与残差连接相加, 最终得到输出特征图。MSAFA 的前向传播过程如下:

$$\begin{aligned}
F_{out} &= \text{Conv}_{1 \times 1}(F_{Spatial} \oplus F_{Channel}) \oplus F_{Dil} \\
F_{MSSA} &= \text{Sigmoid}\{\text{Conv}_{7 \times 7}[\text{Maxpool}(F_{Dil})] \otimes F_{Dil}\} \\
F_{Channel} &= \text{Conv}_{1 \times 1}\{\text{ReLU}[\text{Conv}_{1 \times 1}(\text{Avg}(F_{Dil}))]\} \\
F_{Dil} &= \text{Conv}_{r=1}^{3 \times 3}(F_{in}) \oplus \text{Conv}_{r=2}^{5 \times 5}(F_{in}) \oplus \text{Conv}_{r=3}^{7 \times 7}(F_{in})
\end{aligned} \quad (9)$$

式中, F_{in} 和 F_{out} 分别为模型的输入和输出; F_{Dil} 为空洞多尺度卷积 Inception 模块的输出; $\oplus$ 和 $\otimes$ 分别为逐元素加法和 Hadamard 乘法; $\text{Avg}(\cdot)$ 和 $\text{Maxpool}(\cdot)$ 分别为全局平均池化与最大池化操作; Sigmoid($\cdot$) 和 ReLU($\cdot$) 为激活函数, 用于生成选择性强调信息区域的注意力图。

在 MSAFA 模块处理流程中, 通道注意力分支首先通过全局平均池化将特征压缩为 $C_1 \times 1 \times 1$, 随后经 1×1 卷积与 ReLU 激活生成通道注意力特征图, 并将其扩展至与输入特征相同的空间维度; 该注意

力特征图与空间注意力分支输出的特征图进行逐元素相乘融合, 最后通过 1×1 卷积将融合后的特征图投影回原始通道维度, 生成多尺度双注意力输出特征 F_{out} 。

1.2.2.5 交叉二维扫描机制(CS2D) RVM 模块引入交叉二位扫描机制(CS2D)替代 VSS 模块中的二位选择性扫描机制(SS2D), 其结构如图 4(D) 所示。输入特征经扫描扩展后, 在水平、垂直、对角线及反对角线 4 个方向上进行交叉扫描与平铺, 将二维特征图展开为多方向的一维序列; 各序列通过 S6 模块进行选择性状态空间建模($h_t = \bar{A}h_{t-1} + \bar{B}x_t, y_t = Ch_t + Dx_t$), 经扫描合并后重新映射回二维空间, 输出增强特征。该机制通过多维度扫描, 有效弥补了 SS2D 十字扫描在对角方向上信息覆盖不足的缺陷。

1.3 模型训练

在训练阶段, 采用二值交叉熵作为损失函数。 Pre_i 表示第 i 个模型预测的图像, GT_i 表示第 i 个图像的标注信息, 则二值交叉熵损失函数可定义为:

$$\text{Loss}(Pre_i, GT_i) = - \sum_{i=1}^W \sum_{j=1}^H (GT_{i,j} \cdot \lg Pre_{i,j}) + (1 - GT_{i,j}) \lg (1 - Pre_{i,j}) \quad (10)$$

式中, $GT_{i,j}$ 为坐标 (i, j) 处的真实像素值; $Pre_{i,j}$ 为模型预测图像在坐标 (i, j) 处的预测像素值; W 和 H 分别为图像的宽度和高度。

1.4 评价指标

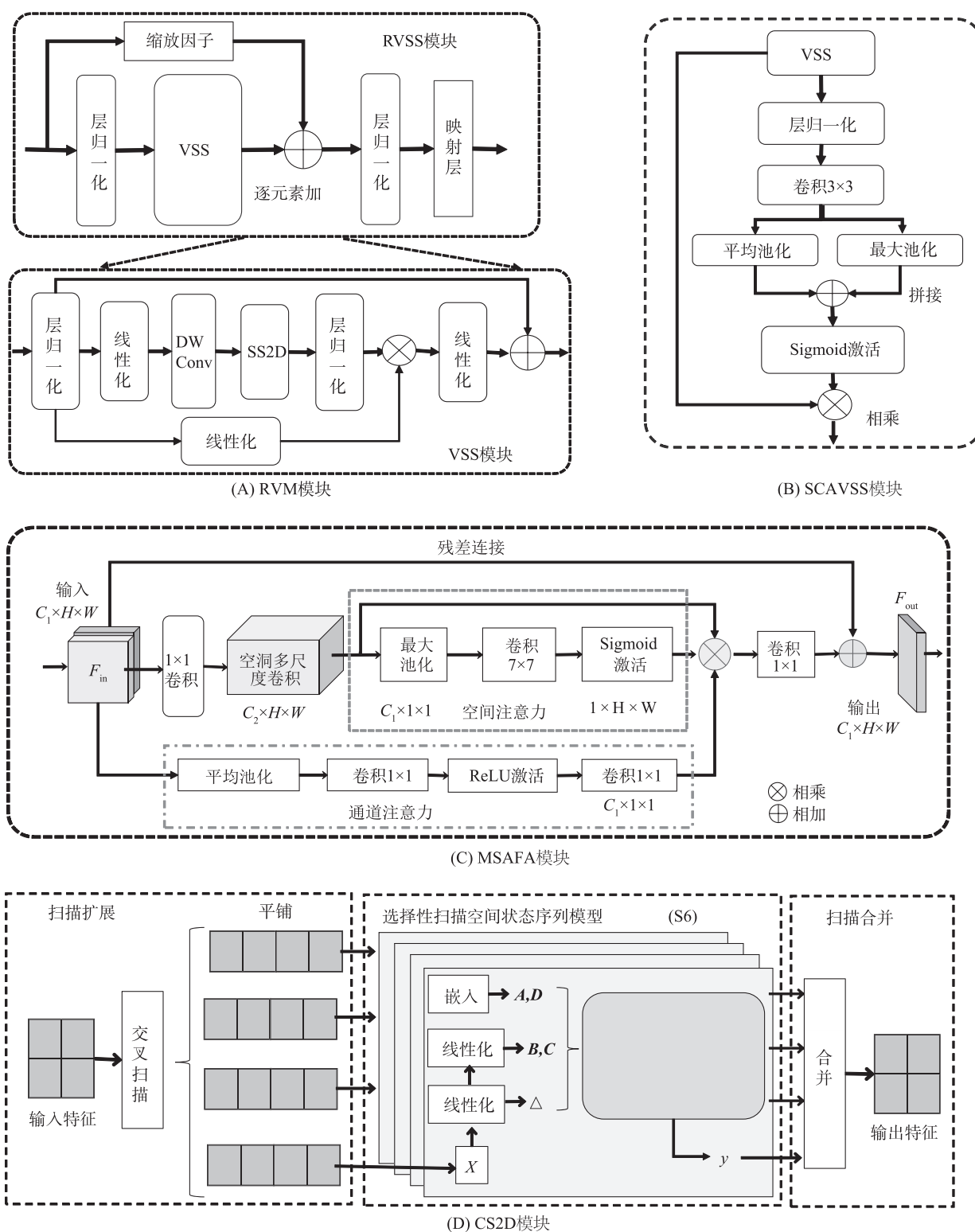
在 IP102 数据集上, 将 SCAFAVM-UNet 模型与基线模型 U-Net 以及 6 种当前主流模型进行对比, 包括 DMAU-Net 模型^[3]、MSRAL-UNet 模型^[8]、TransFCL 模型^[13]、Swin-UNet 模型^[14]、Mamba-UNet 模型^[19] 和 MSV-Mamba 模型^[23]。

Dice 相似系数(DSC)在评估不平衡数据集上的模型性能方面具有显著优势, 因此本研究同时采用精度和 Dice 相似系数(DSC)作为作物害虫检测(CPD)任务的评价指标。精度和 Dice 相似系数(DSC)计算公式如下:

$$\text{Precision} = \frac{T_p}{T_p + F_p} \times 100\% \quad (11)$$

$$\text{DSC} = \frac{2 \times T_p}{2 \times T_p + F_p + F_N} \times 100\% \quad (12)$$

式中, T_p 为检测到的害虫区域与真实害虫区域重叠的部分, F_N 为检测结果中被误判为正常的害虫区域, F_p 为检测结果中被误判的害虫正常区域。



DWConv:深度可分离卷积; F_{in} 和 F_{out} 分别为模型的输入和输出; A 、 B 、 C 、 D 为权值矩阵; W :图像宽度; H :图像高度; C :通道数。SCAVSS:空间通道注意力视觉状态空间模块;RVM:残差视觉 Mamba 模块;VSS:视觉状态空间模块;MSAFA:多尺度自适应特征聚合模块;CS2D:交叉二维扫描机制;SS2D:二位选择性扫描机制。

图4 SCAFAVM-UNet 模型主要模块的结构

Fig.4 Structure of the main modules of the SCAFAVM-UNet model

1.5 试验环境和参数设置

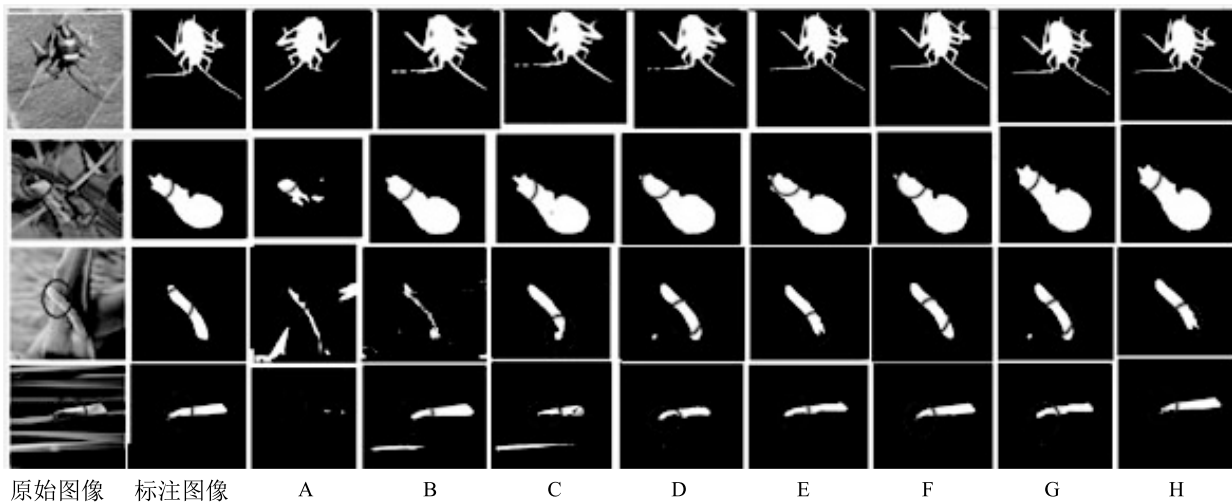
随机设置 SCAFAVM-UNet 的超参数值。试验环境为: Windows 10 操作系统, Intel Xeon E5-2667 v3 @ 3.20 GHz CPU, NVIDIA Quadro M4000 GPU, Python 3.8.8, PyTorch 1.10, CUDA 11.3。

鉴于害虫目标在图像中通常仅占很小区域, 模型训练采用交叉熵损失函数结合 SGD 优化器, 直至损失函数收敛。统一超参数设置如下: 初始学习率 0.001, 批量大小 15, 动量 0.5, 权重衰减 0.000 1, 迭代次数 3 000。所有对比模型均采用相同策略进行训练, 在 IP102 数据集的作物害虫子集上, 利用 5 折交叉验证对模型超参数进行微调。

2 结果与分析

2.1 SCAFAVM-UNet 模型害虫检测性能

采用 10 折交叉验证策略对 SCAFAVM-UNet 模型、U-Net 模型、DMSU-Net 模型^[3]、MSRAL-UNet 模型^[8]、TransFCL 模型^[13]、Swin-UNet 模型^[14]、Mamba-UNet 模型^[19]和 MSV-Mamba 模型^[23]进行训练, 并选取第 3 000 次迭代的模型进行害虫检测。如图 5 所示, 简单场景下, 所有模型均能有效定位害虫, 但 SCAFAVM-UNet 的检测效果最佳, 不仅完整保留了害虫的主体边缘, 其触角、头部、足部等细微结构也清晰可辨, MSV-Mamba 模型的检测效果次之。



A: U-Net 模型; B: DMSU-Net 模型; C: MSRAL-UNet 模型; D: TransFCL 模型; E: Swin-UNet 模型; F: Mamba-UNet 模型; G: MSV-Mamba 模型; H: SCAFAVM-UNet 模型。

图 5 简单场景下不同模型的害虫检测效果

Fig.5 Pest detection performance of different models in simple scenarios

为进一步验证 SCAFAVM-UNet 模型的有效性 with 泛化能力, 选取 4 幅复杂场景下的害虫图像进行测试, 如图 6 所示。原始图像目标尺度不一、图像模糊、背景杂乱、害虫与背景对比度低。尤其后两幅图像中存在大量微小害虫, 传统分割与检测算法难以有效处理此类复杂场景。对比结果表明, SCAFAVM-UNet 模型的检测精度最高, 不仅能准确维持害虫的边界完整性, 还能清晰呈现触角及微小害虫足部等复杂结构。相比之下, U-Net 模型和 DMSU-Net 模型检测到的小害虫形态残缺, 部分细节特征丢失。

如表 1 所示, SCAFAVM-UNet 模型的检测精度和 DSC 最高, 分别为 76.12% 和 73.11%, 但其训练

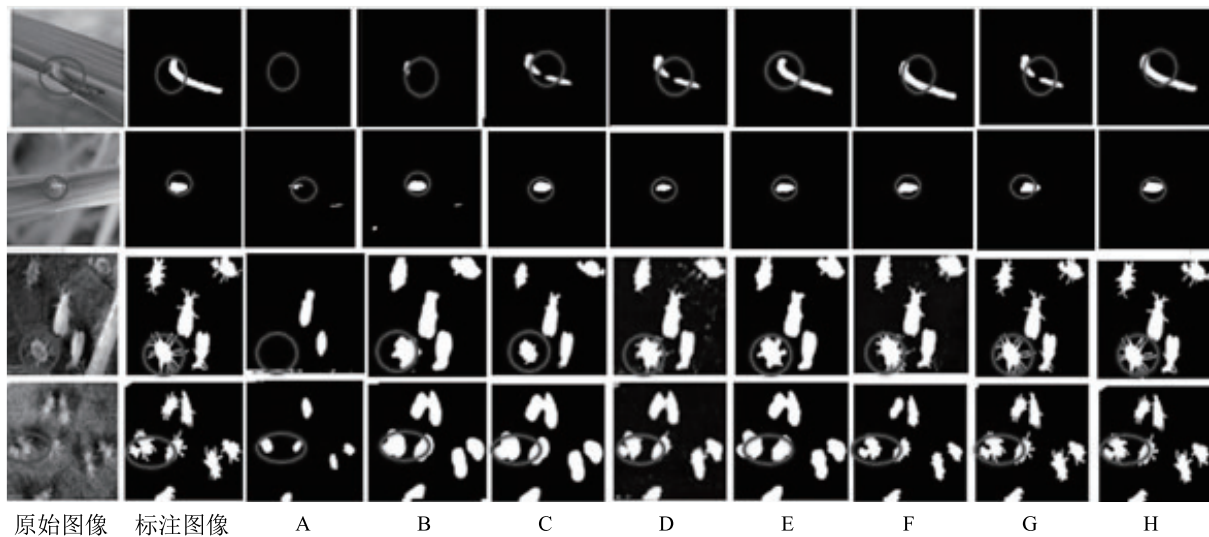
时间略长于 Mamba-UNet 模型和 MSV-Mamba 模型。U-Net 模型的检测精度和 DSC 最低, 主要原因是 U-Net 模型存在最大池化, 导致该模型对害虫的检测和定位不准确。TransFCL 模型和 Swin-UNet 模型的训练时间最长, 原因是 TransFCL 模型和 Swin-UNet 模型中的骨干网络 Transformer 模型的计算复杂度为二次增长。

2.2 消融试验结果

SCAFAVM-UNet 模型与 Mamba-UNet 模型相比, SCAFAVM-UNet 的主要改进包括: 引入空洞多尺度卷积 Inception 模块和 RVM 模块, 采用 CS2D 扫描方式, 使用 SCAVSS 模块替代跳跃连接, 并在瓶颈层

增加 MSAFA 模块,而 Mamba-UNet 模型的瓶颈层由两个 VSS 模块组成,扫描方式为 SS2D。图 7 和表 2

所示,MSAFA 模块和 SCAVSS 模块能够极大增加模型的精度,但训练时间略有增加。



A: U-Net 模型; B: DMSU-Net 模型; C: MSRAL-UNet 模型; D: TransFCL 模型; E: Swin-UNet 模型; F: Mamba-UNet 模型; G: MSV-Mamba 模型; H: SCAFAVM-UNet 模型。

图 6 复杂场景下不同模型的害虫检测效果

Fig.6 Pest detection performance of different models in complex scenarios

表 1 不同模型的性能

Table 1 Performance of different models

模型	精度 (%)	DSC (%)	训练时间 (h)
U-Net	62.47	60.15	11.07
DMSU-Net	67.16	62.42	11.16
MSRAL-UNet	69.44	63.30	10.28
TransFCL	71.08	63.11	12.10
Swin-UNet	71.83	67.53	12.25
Mamba-UNet	73.06	71.44	6.85
MSV-Mamba	74.13	70.26	7.11
SCAFAVM-UNet	76.12	73.11	7.24

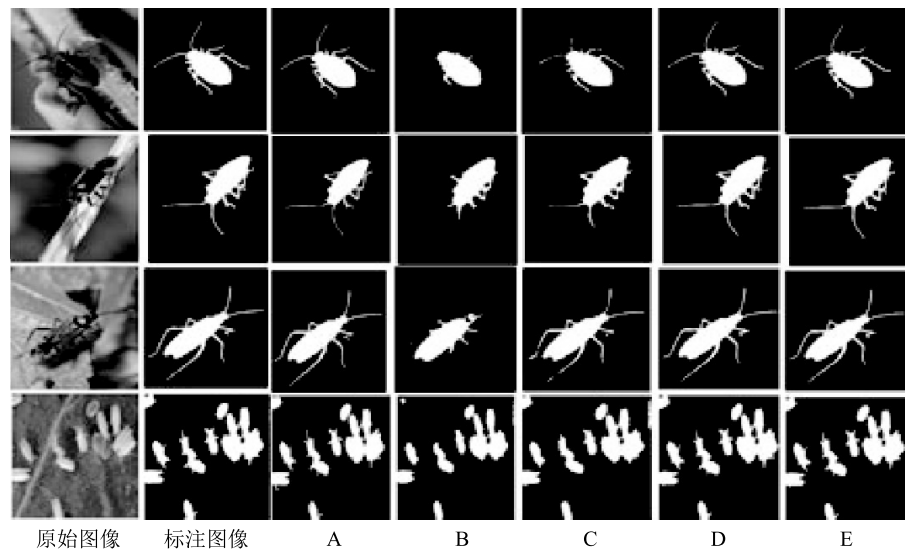
DSC; Dice 相似系数。

3 讨论与结论

SCAFAVM-UNet 模型是一种基于 U-Mamba 模型架构的改进型害虫检测模型。U-Mamba 模型结合了 U-Net 模型的编码器-解码器结构与 Mamba 的状态空间模型,在生物医学影像分割中表现出色,能够有效克服传统 U-Net 模型局部感受野有限及 Transformer 模型计算复杂度过高的问题,同时建立

长距离依赖关系。本研究将其引入农业害虫检测领域,构建了 SCAFAVM-UNet 模型,该模型整合了主干起始模块(Stem)、残差视觉 Mamba 模块(RVM)、空间通道注意力视觉状态空间模块(SCAVSS)、多尺度自适应特征聚合模块(MSAFA)以及交叉二维扫描机制(CS2D)。

SCAFAVM-UNet 模型中多个模块的协同作用显著提升了整体性能:空洞多尺度卷积 Inception 模块的多尺度特征提取策略能够有效识别害虫尺寸的多样性,其优于 Mamba-UNet 模型的简单图像块划分策略;在原始 VSS 模块的基础上,RVM 模块增强了深度语义特征的提取能力;Stem 模块与交叉二维扫描机制(CS2D)能够有效提升模型检测精度;瓶颈层的 MSAFA 模块与 RVM 模块进一步优化了分割精度。在 IP102 数据集上进行试验,结果表明,SCAFAVM-UNet 模型在检测精度与训练效率上整体优于 U-Net 模型、DMSU-Net 模型、Swin-UNet 模型及 Mamba-UNet 模型等。特别是在害虫目标微小、触角及足部等细节丰富、且与背景对比度较低的复杂场景下,SCAFAVM-UNet 模型能够更精准地定位害虫,保持轮廓完整性,显著减少了漏检和目标残缺现象。



A;SCAFAVM-UNet 模型-主干起始模块(Stem)+块划分和线性模块;B;SCAFAVM-UNet 模型-空洞多尺度卷积 Inception 模块;C;SCAFAVM-UNet 模型-残差视觉 Mamba 模块(RVM)+视觉状态空间模块(VSS);D;SCAFAVM-UNet 模型+2 个残差视觉 Mamba 模块(RVM);E;SCAFAVM-UNet 模型-交叉二维扫描机制(CS2D)+二位选择性扫描机制(SS2D)。

图 7 不同模块对 SCAFAVM-UNet 模型害虫检测效果的影响
Fig.7 Effects of different modules on the pest detection performance of SCAFAVM-UNet model

表 2 不同模块对 SCAFAVM-UNet 模型性能的影响
Table 2 Effects of different modules on the performance of SCAFAVM-UNet models

模型	精度 (%)	DSC (%)	训练时间 (h)
SCAFAVM-UNet	76.12	73.11	7.24
SCAFAVM-UNet-主干起始模块和线性模块	75.93	73.03	6.32
SCAFAVM-UNet-空洞多尺度卷积 Inception 模块	67.13	68.07	6.58
SCAFAVM-UNet-RVM+VSS	75.81	73.24	7.65
SCAFAVM-UNet+2×RVM	76.04	69.12	7.03
SCAFAVM-UNet-CS2D+SS2D	76.04	73.25	6.32

-表示去除相应模块;+表示增加相应模块。Stem:主干起始模块;RVM:残差视觉 Mamba 模块;VSS:视觉状态空间模块;CS2D:交叉二维扫描机制;SS2D:二位选择性扫描机制。DSC:Dice 相似系数。

本研究仍存在一定局限。首先,模型的复杂性导致训练耗时较长,未来可探索轻量化设计以提升推理效率。其次,当前试验基于公开数据集,其在真实大田环境中的鲁棒性仍需进一步验证。后续工作将重点研究模型与无人机视觉系统的集成方案,构建适用于实际生产场景的智能害虫监测系统,以实现早期预警和精准施药,最终减少粮食损失并保护生态环境。

参考文献:

[1] CHEN H R, WEN C J, ZHANG L, et al. Pest-PVT A model for multi-class and dense pest detection and counting in field-scale environments[J]. Computers and Electronics in Agriculture,2025, 230:109864.

[2] JIAO L, XIE C J, CHEN P, et al. Adaptive feature fusion pyramid network for multi-classes agricultural pest detection[J]. Computers and Electronics in Agriculture,2022,195:106827.

[3] WANG X Q, ZHANG S W, ZHANG T. Crop insect pest detection based on dilated multi-scale attention U-Net[J]. Plant Methods, 2024,20(1):34-42.

[4] 熊 干,陈慈发,张 上. QMDF-YOLO11:复杂场景下水稻害虫检测算法[J]. 计算机工程与应用,2025,61(13):113-123.

[5] 崔晓辰,雷一东. 基于深度学习的林业害虫智能化检测方法研究进展[J]. 世界林业研究,2024,37(4):53-57.

[6] 杨 锋,姚晓通. 基于改进 YOLOv8 的小麦叶片病虫害检测轻量化模型[J]. 智慧农业(中英文),2024,6(1):147-157.

[7] 张小艾,乔红波,席 磊,等. 基于改进 YOLOv8 的大田农作物害虫小目标检测方法[J]. 农业工程学报,2025,41(13):225-233.

[8] 李 萍,刘 裕,师晓丽,等. 基于多尺度残差空间注意力轻量化 U-Net 的农业害虫检测方法[J]. 江苏农业科学,2023,51(3):187-196.

[9] 李安琦,马 丽,于合龙,等. 改进的 U-Net 算法在遥感图像典型农作物分类研究[J]. 红外与激光工程,2022,51(9):428-

- 434
- [10] MAYYA A M, ALKAYEM N F. Triple-stage crack detection in stone masonry using YOLO-ensemble, MobileNetV2U-net, and spectral clustering [J]. *Automation in Construction*, 2025, 172: 106045.
- [11] LIU Y J, YU J Z, XU Y, et al. A survey on transformer context extension: approaches and evaluation [EB/OL]. (2025-07-08) [2025-06-06]. <https://doi.org/10.48550/arXiv.2503.13299>.
- [12] SARANYA T, DEISY C, SRIDEVI S, et al. Efficient agricultural pest classification using vision Transformer with hybrid pooled multi-head attention [J]. *Computers in Biology and Medicine*, 2024, 177:108584.
- [13] LIU H L, ZHAN Y Z, SUN J, et al. A Transformer-based model with feature compensation and local information enhancement for end-to-end pest detection [J]. *Computers and Electronics in Agriculture*, 2025, 231:109920.
- [14] CAO H, WANG Y Y, CHEN J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation [EB/OL]. (2021-05-12) [2021-05-12]. <https://doi.org/10.48550/arXiv.2105.05537>.
- [15] MAURÍCIO J, DOMINGUES I, BERNARDINO J, et al. Comparing vision transformers and convolutional neural networks for image classification: a literature review [J]. *Applied Sciences*, 2023, 13:5521-5537.
- [16] LIEBER O, LENZ B, BATA H, et al. Jamba: a hybrid Transformer-Mamba language model [EB/OL]. (2024-07-03) [2025-05-06]. <https://doi.org/10.48550/arXiv.2403.19887>.
- [17] LIU Y, TIAN Y J, ZHAO Y Z, et al. Vmamba: visual state space model [EB/OL]. (2024-01-18) [2024-12-29]. <https://doi.org/10.48550/arXiv.2401.10166>.
- [18] RAHMAN M, TUTUL A, NATH A, et al. Mamba in vision: a comprehensive survey of techniques and applications [EB/OL]. (2024-10-04) [2025-05-08]. <https://doi.org/10.48550/arXiv.2410.03105>.
- [19] WANG Z, ZHENG J Q, ZHANG Y, et al. Mamba-UNet: UNet-like pure visual mamba for medical image segmentation [EB/OL]. (2024-03-30) [2025-05-06]. <https://doi.org/10.48550/arXiv.2402.05079>.
- [20] LIU J, YANG H, ZHOU H Y, et al. Swin-UMamba: Mamba-based U-Net with ImageNet-based pretraining [EB/OL]. (2024-03-06) [2025-06-01]. <https://doi.org/10.48550/arXiv.2402.03302>.
- [21] LIAO W B, ZHU Y H, WANG X Y, et al. LightM-UNet: Mamba assists in lightweight u-net for medical image segmentation [EB/OL]. (2024-03-11) [2025-06-01]. <https://doi.org/10.48550/arXiv.2403.05246>.
- [22] WANG Q N, WANG C L, LAI Z X, et al. InsectMamba: insect pest classification with state space model [EB/OL]. (2024-04-04) [2025-06-02]. <https://doi.org/10.48550/arXiv.2404.03611>.
- [23] YANG X X, WANG Q, ZHANG K Q, et al. MSV-Mamba: a multiscale vision mamba network for echocardiography segmentation [EB/OL]. (2025-01-13) [2025-06-05]. <https://doi.org/10.48550/arXiv.2501.07120>.

(责任编辑:成纾寒)