轻量化 MobileNet－YOLOv7 融合模型在山区分散农田病害识别中的适配研究

万小雨[1]，张信得[2]，李绍稳[3]

（1.合肥师范学院 数学与统计学院，安徽 合肥 230601；2.安徽省农业科学院 经济作物研究所，安徽 合肥 230001；
3.安徽农业大学 信息科学与技术学院，安徽 合肥 230001）

摘 要：针对山区分散农田病害识别中环境复杂、目标尺度多变及边缘设备算力受限等导致的检测精度低与部署困难等问题，提出一种轻量化 MobileNet－YOLOv7 融合模型。其核心创新在于：将 YOLOv7 的 ELAN 主干替换为改进型 MobileNet 结构，并引入基于 L1 范数的动态通道剪枝机制以实现参数压缩；在第 3～5 阶段中嵌入轻量多尺度注意力模块，结合并行深度可分离卷积与通道激励生成联合注意力权重；同时重构检测头并采用跨阶段部分特征融合策略，实现浅层细节与高层语义信息的逐级融合。模型通过两阶段训练与 TensorRT INT8 量化，适配 Jetson Nano 等边缘设备部署。实验结果显示，该方法在多尺度病害目标检测中的 AP@0.5 均值为 0.764～0.891，AP@0.75 均值为 0.621～0.842；在复杂背景下 mAP@0.5∶0.95 达 71.8%～73.2%，Precision@0.5 为 84.9%～86.1%；在 Jetson Nano 上平均功耗为（4.82±0.11）W，温升控制在（18.3±0.7）℃。与现有轻量化模型相比，本方法在精度、效率与稳定性方面均具有明显优势，为山区农业智能感知提供了更具适应性的技术路径。

关键词：分散农田；病害识别；YOLOv7；轻量化模型；多尺度注意力

中图分类号：TP31 **文献标志码**：A **文章编号**：1009－1734(2026)02－0045－14

山区农田生态系统具有高度的空间异质性和环境不确定性，作物病害多呈碎片化分布，早期症状隐蔽，给精准识别带来了严峻挑战[1]。实现病害区域的准确检测，对保障作物产量、减少植保投入等都具有重要意义[2]。在边缘计算资源受限的条件下，开发兼具低推理开销与高检测灵敏度的智能视觉系统，已成为支撑田间实时决策的关键技术路径[3]。然而，现有检测系统在实际部署中面临诸多结构性瓶颈[4]。骨干网络普遍存在参数冗余，导致边缘设备推理延迟过高[5]。轻量型网络虽可降低计算负荷，但对小目标病害的识别能力弱，导致假阴性率上升[6]；此外，训练过程中的梯度冲突与泛化能力不足，也制约了模型在真实农田环境中的稳定应用[7]。本研究基于改进的 MobileNet 与 YOLOv7 架构，提出一种轻量级病害识别模型，并围绕轻量化网络设计与硬件感知优化展开研究，旨在推动通用人工智能模型向边缘端迁移，为复杂农村场景提供高效、实用的农业智能解决方案。

1 模型设计与算法改进

针对山区农田病害识别中的挑战，本文提出了一种轻量化 MobileNet－YOLOv7 融合模型，并详细阐述了其创新性设计与应用效果。具体而言，该模型基于改进的 MobileNet 主干网络[8]，通过引入 L1 范数动态通道剪枝机制实现参数压缩；同时嵌入多尺度注意力模块，增强特征响应能力；并采用跨阶段部分特征融合策略，有效结合浅层细节与高层语义信息。经两阶段训练与 TensorRT INT8 量化[9]，该模型成功

收稿日期：2025－11－24
基金项目：合肥师范学院校级科研项目（2024KY63）
通信作者：张信得，助理研究员，从事农业经济，农业病害识别、计算机视觉研究

适配 Jetson Nano 等边缘设备，实现山区农田病害的实时精准检测。轻量化 MobileNet－YOLOv7 融合架构的具体设计见图 1。

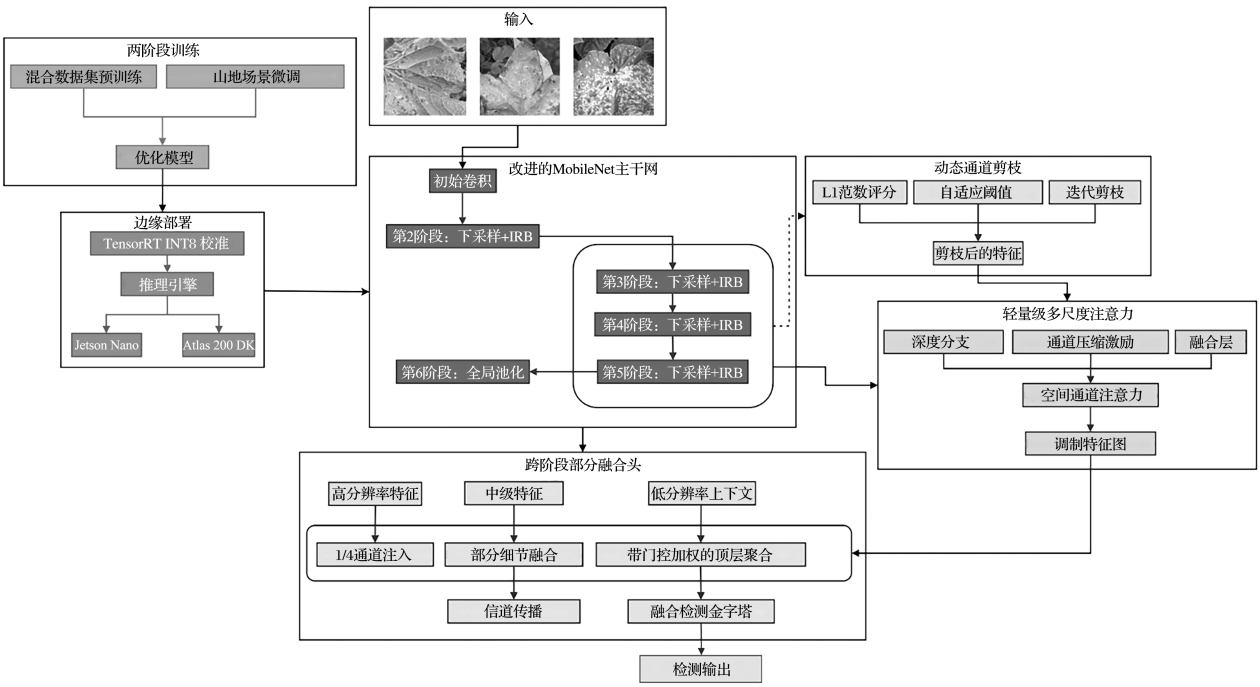


图 1 轻量化 MobileNet－YOLOv7 融合架构

1.1 主干网络重构与动态通道剪枝实施

原始 YOLOv7 的 ELAN 主干架构存在参数冗余、计算复杂度高等不足，不适合边缘部署[10]。本文针对这些问题，重新设计骨干网络架构，采用基于深度可分离卷积和反向残差块的增强型 MobileNet 结构，显著降低参数冗余和浮点运算次数。该设计通过将空间混合与通道混合分离，有效提高了计算资源的利用效率。

1.1.1 主干网络轻量化重构与结构适配

原始 YOLOv7 的 ELAN 架构因密集连接和通道堆叠，易导致 ELAN 模块计算复杂，难以在边缘设备上部署。为克服这些不足，本文重新设计骨干网络，并采用基于反向残差块的增强型 MobileNet 架构。该架构基于深度可分离卷积，使其在大部分卷积运算中实现空间混合与通道混合的分离，从而显著降低参数冗余和浮点运算次数。深度可分离卷积的输出可表示为

$$\boldsymbol{Y}=\sigma(\boldsymbol{W}_{\mathrm{d}} * \mathrm{DepthConv}(\boldsymbol{W}_{\mathrm{p}} * \boldsymbol{X})), \tag{1}$$

式中：$\boldsymbol{X}$ 为输入特征图；$\boldsymbol{W}_{\mathrm{d}}$ 与 $\boldsymbol{W}_{\mathrm{p}}$ 分别为深度卷积与逐点卷积的权重张量；$*$ 表示卷积运算；$\sigma(\cdot)$ 为 H－Swish 激活函数。

整个网络包含 6 个递进阶段（第 1～6 阶段）。第 1 阶段使用标准卷积进行初始特征编码，后续各阶段则通过步长大于 1 的深度可分离卷积实现空间下采样，以提升特征语义抽象程度。网络通道数通过宽度乘数进行缩放，以获取不同尺寸的模型。其中，第 3～5 阶段的输出被定义为用于多尺度检测的关键特征，其维度经专门设计，以有效捕捉山区农田中不同大小与距离的病害目标。此外，所有非线性激活均采用硬件友好的 H－Swish 函数，并结合批量归一化层来稳定数据分布，从而确保训练过程具备更强的收敛性和鲁棒性。其定义为

$$\mathrm{H-Swish}(x)=x \cdot \frac{\mathrm{ReLU6}(x+3)}{6}。 \tag{2}$$

轻量级多尺度注意力机制——清晰的增强过程见图 2。在图 2(a)的原始特征映射中，病害疑似区域

与背景噪声混杂。注意力权重热图[图 2(b)]显示，模块能精准聚焦于病灶空间位置，生成有效掩码。调制后的特征图[图 2(c)]显著增强了病灶响应，抑制了背景干扰。这得益于本设计的双分支结构：小卷积捕捉局部纹理，大卷积捕获跨区域上下文；通道激活机制通过压缩激活生成自适应权重，使模型能自动聚焦于显著特征。特征调制通过逐元素乘法，在保留空间信息的同时提升了特征判别力。

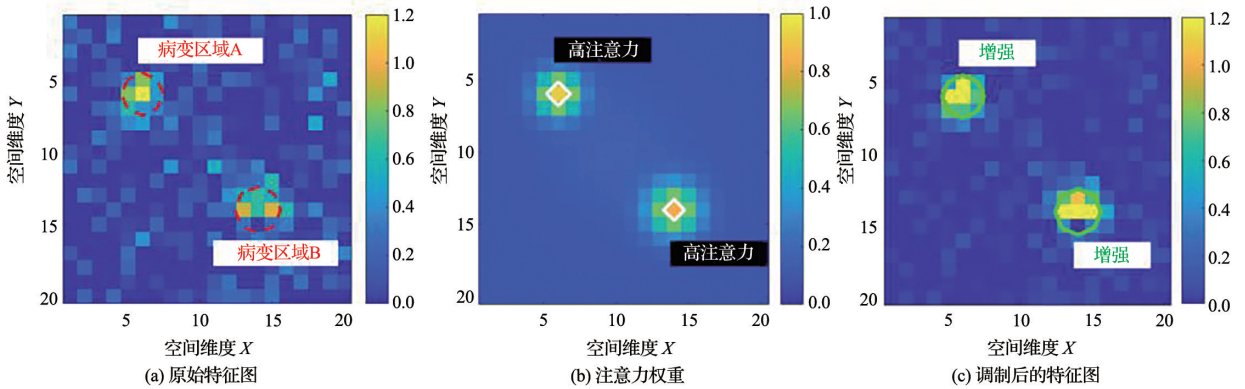


图 2　轻量级多尺度注意力机制——清晰的增强过程

1.1.2　动态通道剪枝机制的执行及精度恢复

在骨干网络的重构训练中，引入基于 L1 范数的通道重要性评估机制，以识别并剪枝响应微弱的冗余滤波器。该方法通过计算卷积层各输出通道权重的 L1 范数绝对值之和来衡量其贡献度。剪枝强度随训练进程动态调整，根据权重稀疏度逐渐增加的趋势，在迭代中逐步提高剪枝阈值，从而避免过早剪除尚有潜力的特征通道。第 l 层第 c 个通道的重要性评分定义为

$$S_{l,c}=\sum_{k}\left|w_{l,c,k}\right|, \tag{3}$$

式中，$w_{l,c,k}$ 表示该通道对应卷积核中第 k 个权重元素。

所有权重低于阈值的通道将被置零并停止梯度更新，以实现结构化稀疏。为防止精度骤降，每次剪枝量受到严格限制。剪枝后立即进入局部微调阶段，对未冻结参数进行一定轮次的反向传播，损失函数保持与原始设计一致(边界框回归、目标置信度与分类)。学习率采用余弦退火策略调度，初始值设定为较低水平以防止权重震荡；优化器选择随机梯度下降算法，以保持更新方向一致。通过重复“评估－剪枝－微调”闭环过程，网络逐步学习选择对病变识别最有效的特征路径，从而实现渐进式压缩与优化。训练轻量级骨干网络的主要超参数设置见表 1。

表 1　主干网络训练超参数配置

参数	值	描述
扩展比率	6.00	反向残差块中的通道扩展因子
批归一化动量	0.03	批归一化的移动平均动量
初始学习率	0.01	优化器的起始学习率
SGD 动量	0.94	随机梯度下降中的动量系数
马赛克增强概率	0.75	应用四图像马赛克数据增强的概率
MixUp 概率	0.20	通过线性插值生成混合样本的比例
水平翻转概率	0.50	应用随机水平翻转的概率

1.2　轻量多尺度注意力模块嵌入

在第 3～5 阶段中嵌入轻量多尺度注意力模块，采用双分支并行深度可分离卷积(3×3 与 5×5 核)提取局部与跨区域特征，结合通道激励机制生成联合注意力权重，以增强对病斑区域的选择性响应。

1.2.1　多尺度空间特征提取路径构建

在 MobileNet 骨干网络第 3～5 阶段的输出端引入注意力扩展模块，以构建多粒度注意力机制。该

注意力扩展模块中3×3与5×5卷积核的组合选择，是基于对目标尺度分布、感受野需求和计算复杂度的综合权衡。首先，通过对训练集中的28 460张标注图像进行统计分析，发现约65%的病斑目标尺寸介于16×16至64×64像素之间。3×3卷积核能高效捕捉此类中小尺度目标的局部纹理与边缘细节，而5×5卷积核则提供了更大的感受野，有助于识别散布性病斑或跨区域的上下文关联模式。同时，对比了不同卷积核组合的浮点运算量，发现3×3+5×5组合在保持多尺度表达能力的同时，其计算成本相较于3×3+7×7组合降低了约18%，更符合边缘设备的部署约束。消融实验进一步表明，该组合在mAP@0.5与Recall@0.5指标上均优于其他对称或更大核尺寸的组合，尤其在小目标病斑(<32×32像素)的检测上表现更佳。基于上述分析，每个注意力扩展模块采用双分支并行结构：一支使用3×3深度可分离卷积，专注于捕捉病斑边缘、纹理等局部细节；另一支使用5×5卷积扩大感受野，以识别散布或聚集的病斑分布模式。两组卷积均采用通道级处理，在增大感受野的同时保持计算复杂度呈线性增长。各分支经批量归一化与H-Swish激活函数处理后，通过带可学习尺度因子的加权融合模块进行通道维融合，实现自适应多尺度特征整合。设第1阶段输出的多尺度融合特征为

$$F_l=\alpha_l\cdot D_{3\times3}(X_l)+(1-\alpha_l)\cdot D_{5\times5}(X_l), \tag{4}$$

式中：X_l为Stage (l)的原始输出特征；$D_{k\times k}(\cdot)$表示核尺寸为$k\times k$的深度可分离卷积操作；$\alpha_l\in[0,1]$为可学习融合权重。融合后的特征图融合了多粒度空间响应，既能保留高分辨率细节，又可承载长程语义关联。整个提取路径采用纯深度可分离卷积进行空间建模，在避免使用标准卷积的同时，显著降低计算复杂度，从而有效抑制浮点运算量的增长，满足边缘设备对实时性的严苛要求。

1.2.2 联合注意力权重生成与特征调制机制

融合的多尺度特征首先输入通道压缩激活子模块，通过全局平均池化生成通道统计描述符；其次，对一个通道数压缩至1/16的瓶颈式全连接层进行非线性变换，并使用ReLU6激活函数确保输出稳定；最后，线性层将通道数恢复，并通过Sigmoid函数生成[0,1]区间内的通道注意力权重序列。设通道压缩-激励输出的权重向量为$\boldsymbol{W}_l$，其计算过程为

$$\boldsymbol{W}_l=\sigma(\boldsymbol{W}_2\mathrm{ReLU6}(\boldsymbol{W}_1\mathrm{GAP}(F_l))), \tag{5}$$

式中：GAP(·)表示全局平均池化；$\boldsymbol{W}_1\in R^{\frac{C}{16}\times C}$，$\boldsymbol{W}_2\in R^{C\times\frac{C}{16}}$为可学习权重矩阵；$\sigma(\cdot)$为Sigmoid函数。该权重序列与上一阶段的多尺度融合输出逐元素相乘，从而耦合空间和通道特异性敏感性。由此得到的张量是一个联合注意力掩码，它与原始主干网络的输出特征图相乘，实现逐样本、逐通道、逐位置地动态调整增益。设原始特征为X_l，最终调制后的特征$\hat{X}_l$为

$$\hat{X}_l=X_l\odot(\boldsymbol{W}_l\otimes 1_{H\times W}), \tag{6}$$

式中：$\odot$表示逐通道相乘；$\otimes$表示通道权重在空间维度上的广播扩展；$1_{H\times W}$为与特征图空间尺寸一致的全一张量。在病灶中心、颜色异常区以及轮廓快速变化等典型区域的强响应会被显著增强，而背景干扰或均匀纹理则会被弱化。这种调制无需改变网络拓扑结构，特征重加权仅通过张量代数计算实现，因而能很好地适配当前最先进的推理引擎的优化机制。

1.3 跨阶段部分特征融合路径设计

采用跨阶段部分特征融合策略重构检测头结构，按通道比例逐级融合浅层细节特征(第2阶段)与高层语义特征(第4、6阶段)，避免全通道融合带来的内存开销与梯度混淆问题。在融合节点引入动态门控机制，以自适应调节不同阶段的特征贡献，从而提升对小目标检测的一致性。

1.3.1 跨阶段部分连接拓扑构建

本文重新设计检测头，并采用通道拼接的方法进行特征融合。选取骨干网络的第2、4、6阶段分别提供高、中、低分辨率特征。设其输出特征图的通道数分别为C_2、C_4和C_6。第2阶段保留细节以定位微小病灶，第4阶段具备语义抽象能力以识别病斑形态，第6阶段编码全局上下文信息辅助病灶区分。为避免直接全融合带来的内存与梯度问题，引入跨阶段部分连接：将第2阶段特征经1×1卷积压缩至1/4通道

后，与第 4 阶段特征拼接，实现中层特征增强。设压缩后的特征通道数为 $C_{2\to4}=C_4/2$。融合后的第 4 阶段（设其通道数为 C_4'，且 $C_4'=C_4+C_{2\to4}$）输出再经 1×1 卷积选择一半通道，与第 6 阶段特征在顶层聚合，最终形成兼具细节与语义的检测输入。上述通道压缩比例（1/4 与 1/2）是基于对不同比例的消融实验（详见本文 2.9.2 节）结果确定的，该比例在内存占用与检测精度间取得了最佳平衡。

设第 2 阶段输出特征为 $\boldsymbol{F}_2$，经降维后进入第 4 阶段的特征表示为

$$\boldsymbol{F}_{2\to4}=\sigma(\boldsymbol{W}_{2\to4}*\boldsymbol{F}_2+\boldsymbol{b}_{2\to4}),\tag{7}$$

式中：$\boldsymbol{W}_{2\to4}\in R^{C_{2\to4}\times C_2\times1\times1}=R^{\frac{C_4}{2}\times C_2\times1\times1}$ 为 1×1 卷积核；$\sigma(\cdot)$ 表示非线性激活函数。第 4 阶段经融合后的输出 $\boldsymbol{F}'_4$ 再以半通道比例参与第 6 阶段 6 融合，其压缩形式为

$$\boldsymbol{F}_{4\to6}=\sigma(\boldsymbol{W}_{4\to6}*F'_4+\boldsymbol{b}_{4\to6}),\tag{8}$$

式中：$\boldsymbol{W}_{4\to6}\in R^{C_{4\to6}\times C'_4\times1\times1}=R^{\frac{C_6}{2}\times C'_4\times1\times1}$，且 $C_{4\to6}=C_6/2$。该半量特征进入第 6 阶段对应的融合节点，与高层语义特征共同构成最终检测输入。整个连接路径呈现逐渐收敛的形式，信息流严格遵循由低到高、由细到粗的层级演化规律。每个阶段都通过仅传递所需信息来抑制无关或重复特征的级联放大效应。

1.3.2　动态门控融合机制与特征调制策略

每个融合节点都设置轻量级门控模块，以动态调节融合特征。该模块将来自底层的注入特征与当前阶段主干特征经 1×1 卷积通道对齐后拼接，并通过全局平均池化获取全局描述，再经两个全连接层（前者压缩通道，后者恢复通道）生成归一化权重向量。此权重与注入特征进行逐通道相乘，实现内容感知的增益控制：当浅层特征包含显著病灶边缘时，自动增强细节注入；对噪声或模糊内容，则抑制传输信号。设拼接后的联合特征为 $\boldsymbol{F}_{\text{cat}}$，其对应的门控权重 $\boldsymbol{G}$ 为

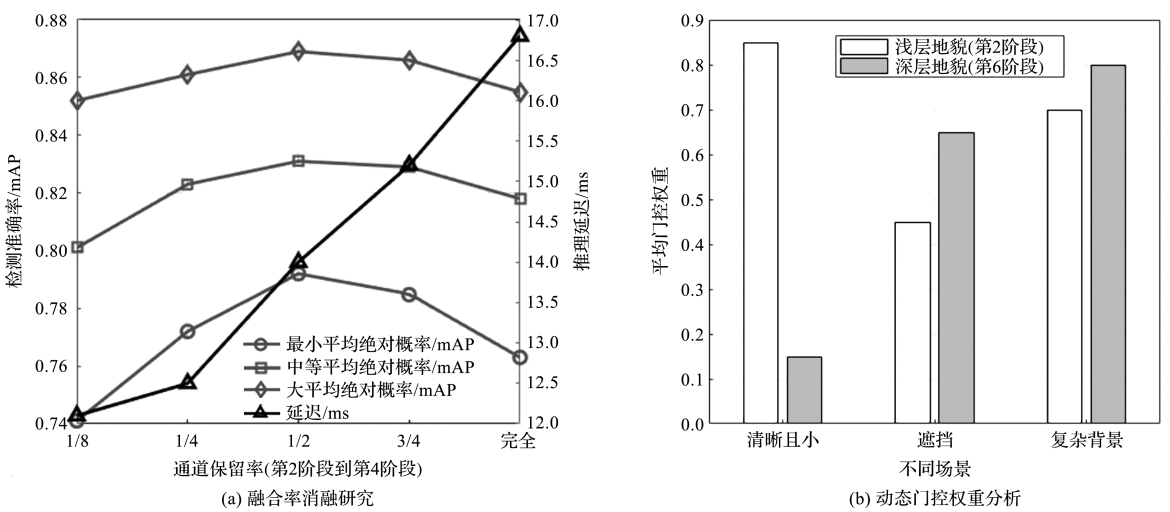
$$\boldsymbol{G}=\text{Sigmoid}(\boldsymbol{W}_2\cdot\text{ReLU}(\boldsymbol{W}_1\cdot\text{GAP}(\boldsymbol{F}_{\text{cat}}))),\tag{9}$$

式中：GAP$(\cdot)$ 表示全局平均池化；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 为可学习权重矩阵。门控权重参与端到端训练，反向传播过程中根据损失梯度持续优化，确保融合策略随任务需求动态演化。融合结果可表示为

$$\boldsymbol{F}_{\text{out}}=\boldsymbol{F}_{\text{local}}+\boldsymbol{G}\odot\boldsymbol{F}_{\text{injected}},\tag{10}$$

式中：$\odot$ 表示逐通道乘法；$\boldsymbol{F}_{\text{local}}$ 为本阶段主干特征；$\boldsymbol{F}_{\text{injected}}$ 为跨阶段注入特征。

融合后的输出直接作为输入特征送入检测头，无需添加任何变换层。该机制显著提高了模型在多尺度下对小目标检测的响应一致性，同时保持了捕捉微小病变的能力，尤其适用于远距离摄影或低分辨率图像输入。特征调制在融合节点内部实现，无需外部指令信号或预设规则，因而具有高度的环境适应性。图像融合和门控机制的细粒度分析结果见图 3。



(a) 融合率消融研究　　(b) 动态门控权重分析

图 3　融合和门控机制的细粒度分析结果

1.4 端到端训练与边缘部署适配

本研究采用两阶段训练策略：第一阶段使用混合数据集训练主干与检测头；第二阶段使用山区农田数据进行微调。最终，通过 TensorRT INT8 量化生成适配 Jetson Nano 的推理引擎[11]，实现低精度高效推理。

1.4.1 两阶段渐进式参数优化流程

模型训练采用两阶段渐进式学习流程。

第一阶段使用整合公开与自采样本的大规模混合数据集，该数据集涵盖多作物、多病害及复杂田间环境。数据预处理通过多尺度随机裁剪（包括旋转和缩放）、光照扰动、颜色空间转换与仿射变换等手段，以增强样本多样性，提升模型对田间变化的适应能力。训练过程中采用动量优化器，并配合梯度裁剪技术防止参数更新失控。损失函数由边界框回归、目标置信度与分类损失 3 部分组成，其权重经经验验证设定了固定的比例，确保多任务目标的均衡收敛。

整体损失函数被表达为三项加权和的形式：

$$\mathcal{L}_{\text{total}}=\lambda_1\mathcal{L}_{\text{cls}}+\lambda_2\mathcal{L}_{\text{obj}}+\lambda_3\mathcal{L}_{\text{box}}, \tag{11}$$

式中：$\mathcal{L}_{\text{cls}}$ 表征类别交叉熵损失；$\mathcal{L}_{\text{obj}}$ 为目标置信度二元交叉熵损失；$\mathcal{L}_{\text{box}}$ 为边界框回归的广义交并比损失；λ_1、λ_2、λ_3 为经过经验验证设定的非负权重系数。学习率调度采用余弦退火策略，初始值设置在稳定下降区间，随后逐步衰减至接近零值，以促进网络在后期进入精细调参状态：

$$\eta_t=\eta_{\min}+\frac{1}{2}(\eta_{\max}-\eta_{\min})\left(1+\cos\left(\frac{t\pi}{T}\right)\right), \tag{12}$$

式中：η_t 为第 t 轮学习率；T 为总训练轮数；$\eta_{\max}$ 与 $\eta_{\min}$ 分别为初始与终态学习率边界值。每轮前向传播结束后，计算整体损失并执行反向传播更新参数。此阶段持续运行预定周期，使主干特征提取能力与检测头判别逻辑建立初步协同关系，完成通用病害模式的建模积累。训练过程中持续监控验证集性能指标，保存性能最优的检查点作为下一阶段的初始条件。

第二阶段则是基于全山地数据集进行局部微调。该数据集包含陡坡、散布地块和泥土阴影背景下的真实射击样本。为重启优化过程，采用较小的学习率，并只允许部分预定义层的权重更新，其余层则保持冻结状态，以维持先前学习获得的有效特征表示。微调过程被严格限制在一个较小的窗口内，这样模型就不会过度拟合小规模数据，从而避免损害其泛化能力。两阶段训练策略的内在动力学机制见图 4。

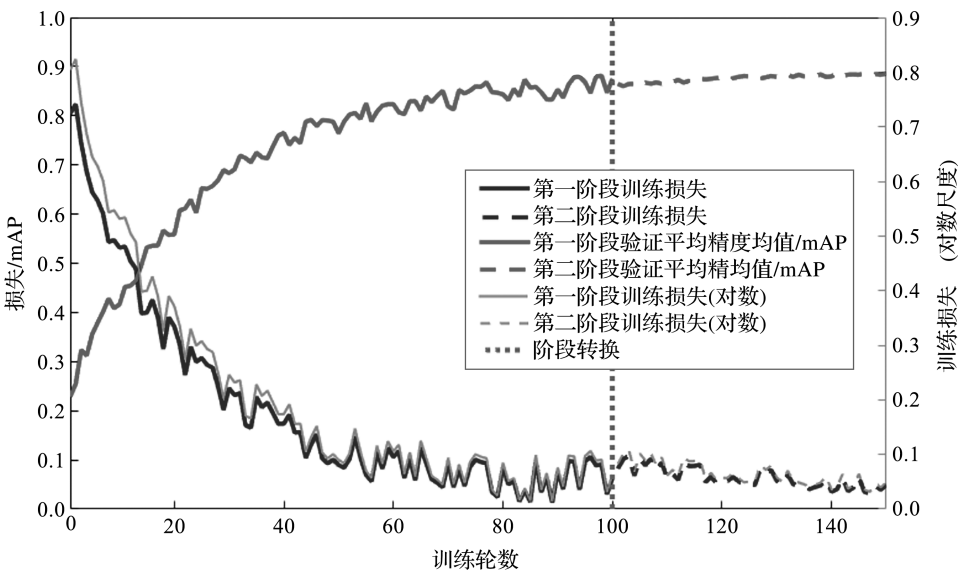


图 4 两阶段训练策略的内在动力学机制

1.4.2　硬件感知量化压缩与推理引擎生成

训练完成后，浮点模型通过 TensorRT 8.6 工具链转换为 INT8 数据类型，以满足边缘设备的部署要求。经 ONNX 格式解析模型，并选取具有代表性的山地图像作为校准集，通过统计激活值动态范围，逐层确定量化缩放因子与零点偏移。为保持关键路径准确性，对检测头及注意力模块等敏感层进行更精细的子范围量化。量化后模型会测试校准集精度，若偏差超预设阈值，则触发局部重校准。最终生成包含优化计算图与量化参数的推理引擎，其可部署于 Jetson Nano 等边缘平台，并利用硬件加速核心实现高效低功耗推理。

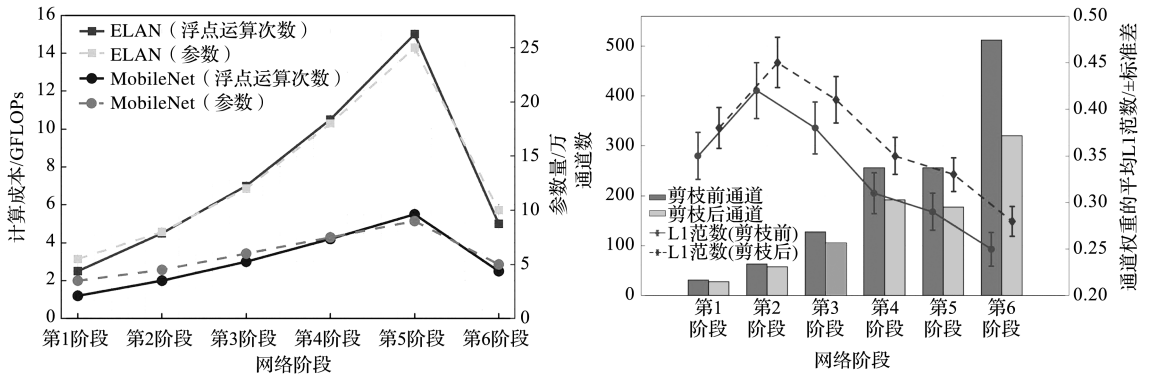
2　实验设计与结果分析

2.1　数据集选择

本实验数据融合公开的 PlantVillage 数据集与自建的山区农业病害图像库[12]。后者聚焦中国西南梯田及坡地环境，收录水稻、玉米、马铃薯等 6 类作物，涵盖叶斑病、锈病、枯萎病等 12 种常见病害。所有图像均通过无人机与手持设备在多光照条件下采集，分辨率统一为 1 920×1 080 像素，总计 28 460 张。数据经人工标注边界框，并由 3 位植保专家交叉验证（Kappa 一致性系数达 0.87），以确保标签可靠性。按 7∶2∶1 的比例将数据划分为训练集、验证集与测试集，测试集进一步细分为小目标、多尺度、复杂干扰（强阴影/逆光/遮挡）及能量评估序列。全部图像保留原始色彩与背景复杂度，未做去噪或增强处理，以真实反映部署场景。

2.2　主干结构效率与通道演化联合分析

本研究对恢复后的主干网络进行联合分析，比较原始 ELAN 与所提 MobileNet 在各阶段的参数量与 GFLOPs，以评估模型复杂度分配变化；在 MobileNet 训练中集成 L1 范数监控，实时追踪剪枝后的通道保留比例与权重平均幅度；通过分析通道数与 L1 统计量在迭代中的相互作用，量化结构稀疏性对特征表达的影响。所有比较均在输入尺寸、批大小与迭代次数相同的公平条件下进行。两个骨干网络在各个阶段的复杂度（成本），以及渠道留存率和权重度演变见图 5。



(a)两种主干网络在各阶段的复杂度分布　　(b)渠道留存率和权重度演变

图 5　两种主干网络的复杂度、渠道留存率和权重度演变

由图 5(a)可见，ELAN 架构在第 5 阶段出现显著计算强度（15 GFLOPs/25 万参数），而 MobileNet 同期仅需 5.5 GFLOPs/9 万参数，体现了深度可分离卷积的结构优势。ELAN 通过通道堆叠保持特征维度导致参数持续增长，MobileNet 采用反向残差块实现参数空间紧凑化。虽然这两种架构在第 6 阶段参数规模趋同，但是 MobileNet 将计算重心重新分配至中间层，在保留特征提取能力的同时显著降低了计

算负载，更契合边缘设备对实时病害检测的资源约束需求。

图5(b)显示，剪枝后各阶段通道数与L1范数分布发生显著变化。第1阶段因需保留基础纹理编码能力，剪枝幅度最小；第3～5阶段因嵌入注意力模块存在大量冗余，剪枝比例最高；第6阶段为保障高层语义的完整性，保留了较多通道。剪枝后L1范数均值系统性上升，标准差收窄，表明低响应通道被有效剔除，模型判别力集中至高贡献通路，实现了效率与检测性能的再平衡。

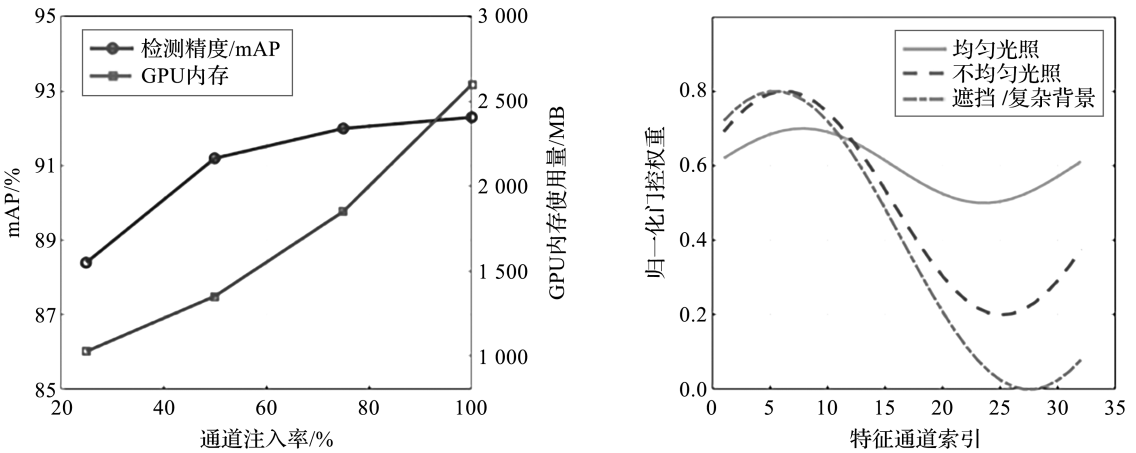
为定量评估动态通道剪枝对模型性能的影响，对比剪枝前后模型在测试集上的精度与参数量，结果见表2。剪枝后模型参数量从剪枝前的1.53 M减少至0.89 M，下降了约42%。与此同时，模型在测试集上的mAP@0.5仅从87.1%下降至85.9%，下降了1.2个百分点。这一结果表明，本文所采用的基于L1范数的动态通道剪枝机制能够高效地识别并移除冗余参数，在实现显著模型压缩的同时，对模型的整体检测精度影响甚微，达成了轻量化与高精度之间的有效平衡。

表2 动态通道剪枝前后的性能对比表

模型状态	参数量/M	mAP@0.5/%
剪枝前	1.53	87.1
剪枝后	0.89	85.9

2.3 跨阶段部分特征融合与动态门控机制分析

本研究选取第2、4、6阶段的多尺度特征构建多阶段特征融合网络，通过调整通道注入比例优化各阶段特征的贡献，并记录不同比例下的精度与GPU内存占用情况的相关性。动态门控模块在多种环境下统计通道激活并生成权重向量，用于指导通道选择和特征分配。训练采用两阶段策略：先冻结主干微调检测头，再进行端到端精调，以确保通道剪枝与门控权重同步更新。所有实验均重复多次取均值，以消除随机因素影响。跨阶段部分特征融合与动态门控机制分析结果见图6。



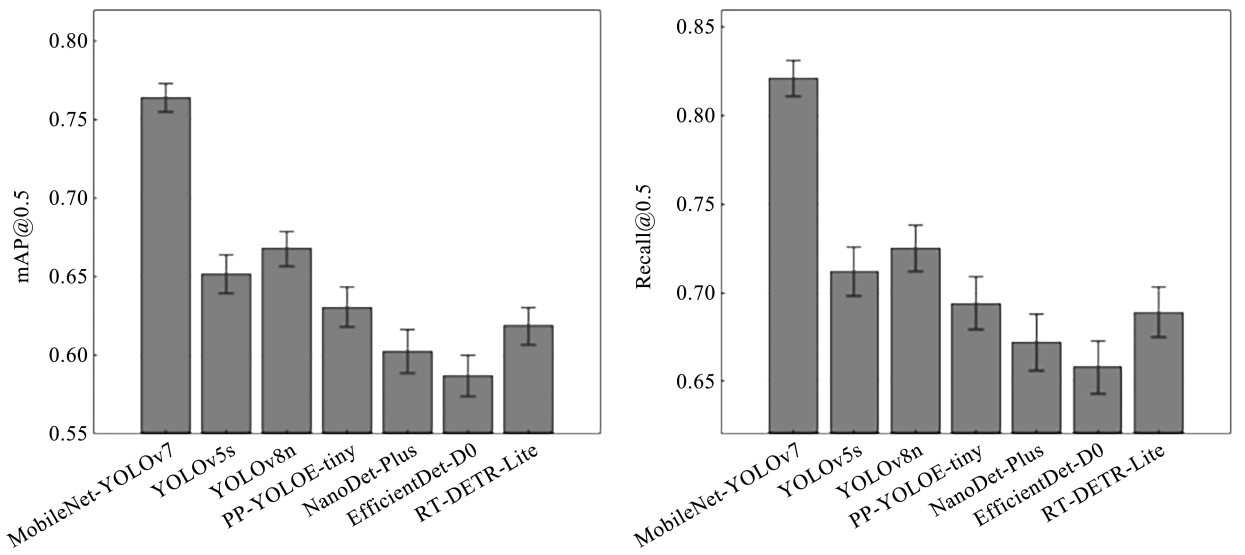
(a)跨阶段部分融合效率 (b)不同场条件下的动态门控响应

图6 跨阶段部分特征融合与动态门控机制分析结果

图6(a)显示，随着通道注入比例从25%增至50%，检测精度(mAP)快速提升，表明适量浅层特征注入能有效增强小目标识别能力。当比例继续增至75%及以上时，精度增长趋缓，而GPU内存占用显著上升，反映出全通道传递会引入冗余信息。图6(b)中，不同环境下的门控权重分布呈现差异化：光照均匀时响应稳定；光照不均时波动加剧，以捕捉局部纹理；复杂背景下则通过抑制低响应通道、增强高频特征来维持判别力，体现了通道调节机制对环境的自适应能力。

2.4　小目标检测精度对比分析

在针对病斑面积小于 32×32 像素的小目标子集上的检测任务时，采用 mAP@0.5 和 Recall@0.5 两个指标评估检测精度。本研究将 MobileNet－YOLOv7 模型与 YOLOv5s、YOLOv8n、PP－YOLOE－tiny、NanoDet－Plus、EfficientDet－D0 及 RT－DETR－Lite 等主流轻量模型进行了对比。所有对比实验均在相同的输入尺寸、训练策略及硬件平台上进行，以确保评估的公平性。图 7 分别展示了各模型在小于 32×32 像素病斑子集上的 mAP@0.5 与 Recall@0.5 表现。



(a)小目标检测：mAP@0.5　　(b)小目标检测：Recall@0.5

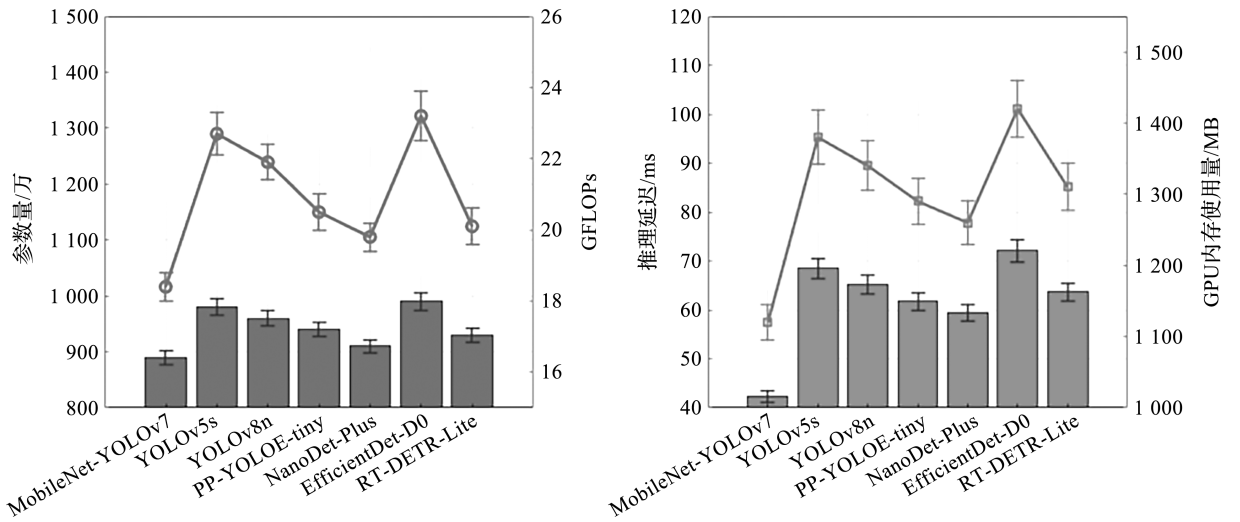
图 7　各模型在小于 32×32 像素病斑子集上的 mAP@0.5 与 Recall@0.5 表现

图 7 显示，在针对小目标病斑（＜32×32 像素）的检测任务中，本文提出的 MobileNet－YOLOv7 模型在 mAP@0.5（0.764±0.009）和 Recall@0.5（0.821±0.010）两项指标上均优于其他 6 种主流轻量检测器。这一优势源于以下 3 种关键技术：一是主干网络的轻量化设计保留了高频细节通路；二是多尺度注意力机制增强了对微小病斑的响应能力；三是跨阶段部分融合策略有效整合了浅层定位线索与高层语义特征。相比之下，其他对比模型因结构限制，导致小目标特征在传递过程中的信息衰减严重。误差分析进一步验证了本方法在推理稳定性方面的优势。研究表明，要解决山区农田环境下的小病斑检测难题，必须在模型轻量化的同时，协同优化特征提取、注意力机制和跨尺度融合路径，从而在边缘设备上实现精度与鲁棒性的统一。

2.5　模型复杂度与计算开销对比分析

统计参数量与浮点运算量（GFLOPs），并在 Jetson Nano 上测量平均推理延迟与 GPU 内存占用。对比对象同上 6 种模型，所有测试在 Ubuntu 20.04、CUDA 11.4 环境下运行，输入图像批量大小设为 1。模型复杂度与边缘端性能的对比分析结果见图 8。

图 8(a)显示，所有模型参数量均符合轻量化要求，且 GFLOPs 均大于 15，说明计算强度得以保持。本研究方法的参数量仅为（89±1.2）万，GFLOPs 达 18.4±0.4，在参数最少的同时维持较高计算密度，得益于动态剪枝保留了关键通路，避免了表达能力下降。图 8(b)显示，在 Jetson Nano 上，本研究的推理延迟（42.3±1.2）ms 与 GPU 内存占用（1 120±25）MB 均显著低于对比模型，这归因于跨阶段部分融合与轻量注意力机制的优化，有效降低了特征输送与计算冗余。部分对比模型虽参数略少，但因结构冗余导致内存占用与推理延迟上升。本研究的结果表明，山区农田部署需实现四维度（参数、计算、内存与延迟）的协同优化，单一策略压缩无法兼顾所有性能指标。



(a)模型复杂度:参数量和 GFLOPS (b)Jetson Nano 上运行时的开销

图 8 边缘部署约束下的复杂性和计算效率对比分析结果

2.6 复杂环境干扰下的稳定性对比分析

从总测试集中选取包含强阴影、逆光与植被遮挡等典型干扰因素的样本构成干扰子集(占比 37.2%)。为精确界定各类干扰样本的特征,首先筛选强阴影样本,即通过测量图像内非阴影区域与阴影区域的平均光照强度进行判别,并限定阴影区域光照强度≤500 lx;其次筛选逆光样本,即通过计算图像中心主体区域与背景天空或亮部的平均像素值之比进行筛选,限定光照对比度≥10∶1;最后筛选植被遮挡样本,即要求病斑目标被其他植物器官(如叶片、茎秆)遮挡面积超过其自身面积的 30%。所有样本均保留原始采集状态,未进行增强或校正处理。在模型微调方面,各模型基于统一训练流程进行优化,未针对特定干扰场景进行专门优化。评估采用 mAP@0.5∶0.95 和 Precision@0.5 两项指标,其分别反映模型在多种 IoU 阈值下的鲁棒性和精确度。通过 3 次独立推理的均值与标准差的对比分析(结果见表 3),确保统计结果的可靠,并为实际部署提供参考。

表 3 复杂环境干扰下的稳定性对比分析结果

模型	干扰类型	mAP@0.5∶0.95/%	Precision@0.5/%	P (vs. MobileNet-YOLOv7)
MobileNet-YOLOv7	强阴影	71.8±0.9	84.9±1.2	—
	逆光	73.2±0.8	86.1±1.0	—
	植被遮挡	72.1±0.8	86.0±1.1	—
YOLOv5s	强阴影	52.3±1.3	77.5±1.6	0.003 2
	逆光	53.8±1.2	78.9±1.5	0.002 8
	植被遮挡	53.2±1.3	78.5±1.6	0.003 5
YOLOv8n	强阴影	53.9±1.2	78.7±1.5	0.002 1
	逆光	55.1±1.1	80.2±1.4	0.001 9
	植被遮挡	54.8±1.2	79.9±1.4	0.002 3
PP-YOLOE-tiny	强阴影	50.9±1.4	76.1±1.7	0.004 1
	逆光	52.4±1.3	77.5±1.6	0.003 7
	植被遮挡	52.1±1.4	77.2±1.7	0.004 3

表 3(续)

模型	干扰类型	mAP@0.5∶0.95/%	Precision@0.5/%	P（vs. MobileNet－YOLOv7）
NanoDet－Plus	强阴影	48.3±1.5	73.2±1.8	0.000 9
	逆光	49.8±1.4	74.8±1.7	0.000 7
	植被遮挡	49.5±1.5	74.5±1.8	0.000 8
EfficientDet－D0	强阴影	46.7±1.6	71.9±1.9	0.000 5
	逆光	48.1±1.5	73.4±1.8	0.000 4
	植被遮挡	47.7±1.6	73.1±1.9	0.000 6
RT－DETR－Lite	强阴影	49.4±1.4	74.7±1.7	0.001 2
	逆光	51.0±1.3	76.3±1.6	0.00 1
	植被遮挡	50.5±1.4	76.0±1.7	0.001 3

注：P，基于配对样本 t 检验(双尾)，比较各模型与本文方法在相同干扰类型下的 mAP@0.5∶0.95 差异；所有模型使用相同微调策略，未进行干扰场景专项增强

表 3 展示了各模型在强阴影、逆光和植被遮挡三类典型干扰下的检测稳定性。MobileNet－YOLOv7 在各项场景中均取得最优结果，mAP@0.5∶0.95 达 71.8%～73.2%，Precision@0.5 为 84.9%～86.1%，且所有对比均具备统计显著性（$P<0.01$）。其优势源于多尺度注意力对低对比区域的增强，以及跨阶段融合对遮挡细节的补偿。其他模型在阴影与遮挡下性能下降明显，主因特征提取依赖全局统计，缺乏局部响应调制能力；RT－DETR－Lite 虽引入了注意力，但在低分辨率下边界收敛能力有限。本研究通过硬件感知剪枝保留关键特征通路，在不增加训练负担的前提下显著提升鲁棒性，更好地满足了边缘端部署对稳定性的要求。

此外，我们在 Jetson Nano 上进一步测试了各模型在干扰子集上的平均推理延迟，以评估其计算效率的稳定性，结果见表 4。

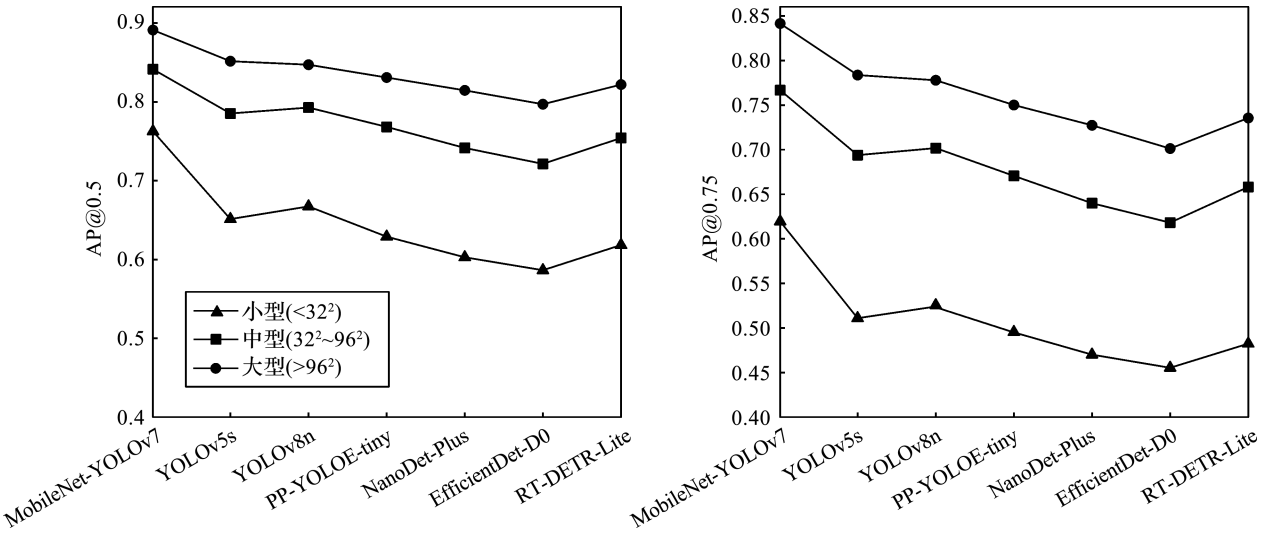
表 4　复杂环境下各模型在 Jetson Nano 上的推理延迟　　**ms**

模型	强阴影	逆光	植被遮挡
MobileNet－YOLOv7	43.1±1.3	42.8±1.2	43.5±1.4
YOLOv5s	58.9±2.1	57.7±1.9	59.8±2.3
YOLOv8n	55.3±1.8	54.6±1.7	56.1±2.0
PP－YOLOE－tiny	61.4±2.4	60.5±2.2	62.2±2.5
NanoDet－Plus	49.2±1.5	48.7±1.4	49.9±1.6
EfficientDet－D0	65.1±2.7	64.2±2.5	66.3±2.8
RT－DETR－Lite	53.6±1.7	52.9±1.6	54.4±1.9

由表 4 可知，本研究提出的 MobileNet－YOLOv7 模型在上述 3 类干扰环境下均保持了最低且最稳定的推理延迟（约 43 ms），波动范围显著小于其他对比模型。这得益于其轻量化的主干网络与高效的特征融合路径，使其计算负荷对复杂背景变化不敏感。而其他对比模型在阴影和遮挡环境下延迟有所增加，表明其计算图可能因环境干扰而触发了更低效的硬件执行路径。这进一步印证了本模型在边缘计算资源受限的复杂山区环境中，兼具精度鲁棒性与效率稳定性的双重优势。

2.7　多尺度目标鲁棒性表现对比分析

根据病害目标面积，测试集划分为 3 个尺度：小（$<32^2$ 像素）、中（32^2～96^2 像素）和大（$>96^2$ 像素）。所有模型在统一标注、输入分辨率及 INT8 量化条件下，在 Jetson Nano 环境中评估，确保公平性。分别计算各尺度在 IoU 阈值为 0.5 和 0.75 下的平均精度（AP），后者对定位精度要求更高。结果基于多次推理取均值与标准差，以分析模型在多尺度病害上的识别鲁棒性与定位敏感性。分析结果见图 9。



(a)多尺度性能:AP@0.5 (b)多尺度性能:AP@0.75

图9 不同IoU阈值下目标尺度的稳健性对比分析结果

由图9可知,不同IoU阈值下各模型对多尺度病害的检测能力不同。小尺度目标因像素少、纹理模糊,AP普遍较低,尤其在AP@0.75时出现显著下降,凸显其定位敏感性不足。中尺度目标特征平衡,多数模型在此达到性能峰值。大尺度目标易被检测,但在IoU@0.75的高精度要求下,部分模型因边界回归精度不足而表现较弱。本研究通过动态剪枝保留关键通路、注意力机制强化病斑区域,并结合分比例融合策略,在3种尺度上实现了更均衡的检测性能:AP@0.5均值为0.764~0.891,AP@0.75均值为0.621~0.842,在严苛定位条件下仍保持较高鲁棒性。

2.8 边缘设备持续运行能效对比分析

Jetson Nano连续运行各模型1小时,记录平均功耗(W,通过USB功率计采样,USB全称为Universal Serial Bus)与温度上升值(℃,通过tegrastats工具获取),同时统计累计处理帧数以计算有效吞吐率(Frames Per Second,FPS)。测试平台为Jetson Nano Developer Kit (4 GB版本),配置为:Ubuntu 20.04 LTS操作系统,JetPack 4.6.1 SDK,CUDA 10.2,cuDNN 8.2,TensorRT 8.2.1。所有测试的输入图像批量大小(Batch Size)固定为1,以模拟边缘设备单张串行处理的典型部署场景。6种对比模型均部署为INT8量化版本,运行环境温度控制在(25±2) ℃,避免散热差异干扰结果。对比分析结果见表5。

表5 边缘设备持续运行能效的对比分析结果

模型	平均功耗/W	温升/℃	FPS/(帧/s)	P (vs. MobileNet-YOLOv7)
MobileNet-YOLOv7	4.82±0.11	18.3±0.7	21.4±0.5	—
YOLOv5s	5.67±0.13	23.1±0.9	14.2±0.4	0.001 8
YOLOv8n	5.53±0.12	22.4±0.8	14.9±0.4	0.002 3
PP-YOLOE-tiny	5.41±0.12	21.7±0.8	15.6±0.4	0.003 1
NanoDet-Plus	5.28±0.12	20.9±0.8	16.3±0.4	0.004 7
EfficientDet-D0	5.89±0.14	24.5±1.0	13.1±0.3	0.000 9
RT-DETR-Lite	5.36±0.12	21.2±0.8	15.9±0.4	0.003 9

由表5数据可知,模型的功耗与温升呈现强正相关性。本研究构建的MobileNet-YOLOv7模型凭借其深度可分离卷积、动态通道剪枝以及跨阶段部分融合等轻量化设计,实现了最低的平均功耗(4.82 W)与温升(18.3 ℃)。在能效权衡方面,本模型在实现最优能耗表现的同时,仍保持了最高的FPS(21.4帧/s)。充分体现了其在计算效率上的优势。通过消除冗余计算与数据搬运,该模型将能量更集中地用于有效特征提取,而非硬件调度开销,从而达成了低功耗、低热耗与高实时性的统一。相比之下,部分对比模型(如

EfficientDet－D0）虽在结构上追求轻量，但因计算图复杂度或内存访问模式欠佳，导致其在 Jetson Nano 平台上的能效比较低，表现为高功耗、高温升和低 FPS。

2.9　消融实验与分析

2.9.1　多尺度卷积核组合选择分析

为验证 1.2.1 节中 3×3 与 5×5 卷积核组合的合理性，对比多种卷积核组合方案的性能，结果见表 6。

表 6　不同卷积核组合在多尺度特征提取路径中的性能对比分析结果

卷积核组合	参数量/万	GFLOPs	mAP@0.5	Recall@0.5	mAP@0.5（小目标）
3×3＋3×3	87.2	17.8	0.841	0.802	0.749
3×3＋5×5	89.0	18.4	0.856	0.821	0.764
5×5＋5×5	91.5	20.9	0.848	0.812	0.752
3×3＋7×7	92.8	22.4	0.850	0.815	0.758

由表 6 可知，本研究采用的 3×3＋5×5 组合，在参数量和计算量仅轻微增加的情况下，取得了全面的精度领先，尤其是在对小目标病害的检测上（mAP@0.5 达 0.764）优势明显。这证实了该非对称多尺度设计能够更有效地平衡细节捕捉与上下文信息获取，是精度与效率权衡下的最优选择。

2.9.2　跨阶段融合通道比例选择分析

为确定 1.3.1 节跨阶段部分融合中最优的通道注入比例，本研究对第 2～4 阶段的通道压缩比例 $\boldsymbol{F}_{2\to4}$ 进行了消融实验。在控制其他变量一致的条件下，对比了 $\boldsymbol{F}_{2\to4}$ 为 1/8、1/4、1/2 及 3/4 时的性能，评估指标包括小目标检测精度（mAP@0.5）与 Jetson Nano 上的 GPU 内存占用，结果见表 7。

表 7　不同通道压缩比例对性能与内存占用的影响对比分析结果

第 2～4 阶段的压缩比例 $\boldsymbol{F}_{2\to4}$	mAP@0.5	mAP@0.5（小目标）	GPU 内存占用/MB
1/8	0.843	0.741	1 052
1/4	0.856	0.764	1 120
1/2	0.855	0.762	1 258
3/4	0.854	0.760	1 435

由表 7 可知，当压缩比例从 1/8 放宽至 1/4 时，本模型能够融合更多来自浅层网络的高分辨率细节特征，小目标检测精度（mAP@0.5）从 0.741 显著提升至 0.764，整体 mAP@0.5 也达到最高的 0.856，而此时内存占用（1 120 MB）仍处于可接受范围。当比例继续增大至 1/2 和 3/4 时，精度提升已趋于饱和，甚至因引入过多冗余信息而略有下降，但内存占用却呈线性显著增长。因此，选择 1/4 作为第 2 阶段的通道压缩比例，是在小目标检测精度（mAP@0.5＝0.764）与边缘设备内存开销（1 120 MB）之间达成的最佳平衡点。基于此结论，第 4～6 阶段的通道压缩比例经验性地设定为 1/2，以在高层语义融合中维持类似的计算与内存效率。

3　结 语

本研究构建的轻量化 MobileNet－YOLOv7 融合架构，面向山区分散农田病害识别场景实现精度与效率的协同优化。主干网络以改进型 MobileNet 替代 ELAN 结构，结合 L1 范数驱动的动态通道剪枝，在压缩参数的同时保留高频病斑通路。轻量多尺度注意力模块嵌入中高层阶段，以增强微小目标响应；跨阶段部分特征融合策略以通道比例逐级聚合浅层细节与高层语义，降低内存访问开销。两阶段训练与 TensorRT INT8 量化保障边缘部署可行性。实验表明，该方法在小目标检测、多尺度鲁棒性、复杂干扰稳定性及能效表现上均优于主流轻量模型，为资源受限山地农业场景提供了可落地的智能感知方案。

参考文献：

[1] TIAN Q，ZHAO G，YAN C Q，et al. Enhancing practicality of deep learning for crop disease identification under field

conditions:insights from model evaluation and crop-specific approaches[J]. Pest Management Science,2024,80(11):5864-5875.
[2] 侯海敏,沈梦姣,张艳.作物病害检测关键技术分析[J].中国农机化学报,2024,45(6):90-97.
[3] 薛飞跃,周玉玲,李俊凯,等.智慧养殖农业物联网与边缘计算中大模型技术应用综述[J].农业机械学报,2025,56(9):291-311.
[4] 王荣赫,尚文周,王启月.智慧农业技术应用的现实困境与优化路径[J].国际会计前沿,2025(3):597-602.
[5] 于萍.面向边缘设备的智慧农业病虫害图像识别关键技术研究[D].长春:吉林大学,2025:12-18.
[6] 胡根生,谢一帆,鲍文霞,等.基于轻量型网络的无人机遥感图像中茶叶枯病检测方法[J].农业机械学报,2024,55(4):165-175.
[7] 邱云桥,何婷婷,王春霞,等.农业生产中机器视觉技术应用现状[J].农业工程,2025,15(6):19-25.
[8] QIN D F,LEICHNER C,DELAKIS M,et al. MobileNetV4:universal models for the Mobile ecosystem[C]//Computer Vision-ECCV 2024. Cham:Springer,2025:78-96.
[9] HIROSE S,WADA N,KATTO J,et al. Research and examination on implementation of super-resolution models using deep learning with INT8 precision[C]//2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Jeju Island,Republic of Korea IEEE,2022:133-137.
[10] 张鹏程,矫桂娥,毕卓.基于YOLOv7的轻量化农田害虫检测算法[J].湖南农业大学学报(自然科学版),2025,51(2):103-112.
[11] 荆旭君,郭永刚,李峰.基于Jetson Nano的农业监测系统的设计与实现[J].现代计算机,2023,29(21):71-76.
[12] 刘晓彬,彭俊桂,黄有章.基于MobileNetV3模型的农作物病害识别研究[J].信息与电脑(理论版),2022,34(19):61-63.

Adaptation Research of a Lightweight MobileNet-YOLOv7 Fusion Model for Disease Detection in Scattered Mountainous Farmlands

WAN Xiaoyu[1], ZHANG Xinde[2], LI Shaowen[3]
(1. School of Mathematics and Statistics, Hefei Normal University, Hefei 230601, China;
2. Institute of Industrial Crops, Anhui Academy of Agricultural Sciences, Hefei 230001, China;
3. School of Information Science and Technology, Anhui Agricultural University, Hefei 230001, China)

Abstract: To address the challenges of low detection accuracy and difficult deployment in identifying diseases in scattered mountainous farmlands caused by complex environments, variable target scales, and limited computing power on edge devices:this study proposes a lightweight MobileNet-YOLOv7 fusion model. The core innovations of this research include: replacing the ELAN backbone of YOLOv7 with an improved MobileNet architecture and introducing a dynamic channel pruning mechanism based on the L1 norm to achieve parameter compression; embedding a lightweight multi-scale attention module from stage3 to stage5, which combines parallel depthwise separable convolutions with channel excitation to generate joint attention weights; and reconstructing the detection head while adopting a cross-stage partial feature fusion strategy to progressively integrate low-level details with high-level semantic information. The model undergoes two-stage training and TensorRT INT8 quantization to enable deployment on edge devices such as the Jetson Nano. Experimental results show that the proposed method achieves an AP@0.5 range of 0.764-0.891 and an AP@0.75 range of 0.621-0.842 for multi-scale disease target detection. Under complex backgrounds, the mAP@0.5∶0.95 reaches 71.8%-73.2%, with a Precision@0.5 of 84.9%-86.1%. On the Jetson Nano, the average power consumption is (4.82±0.11) W, and the temperature rise is controlled within (18.3±0.7) ℃. Compared to existing lightweight models, this approach demonstrates significant advantages in accuracy, efficiency, and stability, offering a more adaptable technical pathway for intelligent agricultural perception in mountainous regions.

Keywords: scattered farmland; disease identification; YOLOv7; lightweight model; multiscale attention