基于RGB与深度图像融合的生菜表型特征估算方法

陆声链[1,2]　李沂杨[1,2]　李　帼[1,2]　贾小泽[1,2]　鞠青青[3,4]　钱婷婷[3,4]

(1.广西师范大学计算机科学与工程学院, 桂林 541004; 2.广西多源信息挖掘与安全重点实验室, 桂林 541004;
3.上海市农业科学院农业科技信息研究所, 上海 201403; 4.上海数字农业工程技术研究中心, 上海 201403)

摘要: 采用自动化手段对植物生长过程中的表型特征进行精准测量对于育种和栽培等应用具有重要意义。本文围绕工厂化生菜种植中的表型特征无损精准检测需求,通过融合深度相机采集的RGB图像和深度图像,利用改进的DeepLabv3+模型进行图像分割,并通过双模态回归网络对生菜表型特征进行估算。本文改进的分割模型的骨干网络由Xception替换为MobileViTv2,以增强其全局感知能力和性能;在回归网络中,提出了卷积双模态特征融合模块CMMCM,用于估算生菜的表型特征。在包含4个生菜品种的公开数据集上的实验结果表明,本文方法可对鲜质量、干质量、冠幅、叶面积和株高共5种生菜表型特征进行估算,决定系数分别达到0.922 2、0.931 4、0.862 0、0.935 9和0.887 5。相较于未添加CMMCM和SE模块的RGB和深度图的表型参数估计基准ResNet-10(双模态),本文改进的模型决定系数分别提高2.54%、2.54%、1.48%、2.99%和4.88%,单幅图像检测耗时为44.8 ms,说明该方法对于双模态图像融合的生菜表型特征无损提取具有较高的准确性和实时性。

关键词: 生菜; 表型估算; 模态融合; 分割模型; RGB图像; 深度图像

中图分类号: TP391.4　**文献标识码**: A　**文章编号**: 1000-1298(2025)01-0084-08　**OSID**:

Lettuce Phenotype Estimation Using Integrated RGB-Depth Image Synergy

LU Shenglian[1,2]　LI Yiyang[1,2]　LI Guo[1,2]　JIA Xiaoze[1,2]　JU Qingqing[3,4]　QIAN Tingting[3,4]

(1. *School of Computer Science and Engineering*, *Guangxi Normal University*, *Guilin* 541004, *China*
2. *Guangxi Key Laboratory of Multisource Information Mining and Security*, *Guilin* 541004, *China*
3. *Institute of Agricultural Science and Information*, *Shanghai Academy of Agricultural Sciences*, *Shanghai* 201403, *China*
4. *Shanghai Digital Agriculture Engineering Technology Research Center*, *Shanghai* 201403, *China*)

Abstract: Accurate measurement of phenotypic traits in plant growth using automated methods is crucial for applications such as breeding and cultivation. Aiming to address the need for non-destructive, precise detection of phenotypic traits in factory-grown lettuce, by integrating RGB images and depth images collected by depth cameras, an improved DeepLabv3+ model was used for image segmentation, and a dual-modal regression network estimated the phenotypic traits of lettuce. The backbone of the improved segmentation model was replaced from Xception to MobileViTv2 to enhance its global perception capabilities and performance. In the regression network, a convolutional multi-modal feature fusion module (CMMCM) was proposed to estimate the phenotypic traits of lettuce. Experimental results on a public dataset containing four lettuce varieties showed that the method estimated five phenotypic traits—fresh weight, dry weight, canopy diameter, leaf area, and plant height—with determination coefficients of 0.922 2, 0.931 4, 0.862 0, 0.935 9, and 0.887 5, respectively. Compared with the RGB and depth image-based phenotypic parameter estimation benchmark ResNet-10 (Dual) without CMMCM and SE modules, the improved model increased the determination coefficients by 2.54%, 2.54%, 1.48%, 2.99%, and 4.88%, respectively, with an image detection time of 44.8 ms per image. This demonstrated that the method achieved high accuracy and real-time performance for non-destructive detection of lettuce phenotypic traits through dual-modal image fusion.

Key words: lettuce; phenotypic estimation; modality fusion; segmentation model; RGB images; depth images

收稿日期: 2024-10-31　修回日期: 2024-11-20

基金项目: 国家自然科学基金项目(61762013)、上海市农业科技创新项目(2023-02-08-00-12-F04621)和农业农村部长三角智慧农业技术重点实验室开放课题(KSAT-YRD2023011)

作者简介: 陆声链(1979—),男,教授,博士,主要从事机器视觉和智能农业研究,E-mail: lsl@gxnu.edu.cn

通信作者: 钱婷婷(1983—),女,副研究员,博士,主要从事设施园艺和数字农业研究,E-mail: qiantingting@saas.sh.cn

0 引言

在设施栽培环境中,作物形态参数及生物量等表型相关参数的动态变化是衡量作物生长状况的重要指标之一[1]。生菜富含多种营养成分,是一种重要的叶类经济作物[2]。通过测量生菜的表型特征参数,可以获取关于其生长发育的重要信息,帮助评估其健康状况、生长速度及生物量积累情况。通过监测鲜质量、干质量和叶面积等关键参数,种植户和研究人员能够及时调整栽培管理措施,优化作物生长速度、提高产量和品质。此外,鲜质量也是评估农产品品质与市场价值的关键指标之一。

精准的环境控制对于温室作物的生长至关重要,温室环境-作物生长机理模型则为模型预测控制或最优控制提供了基础[3]。然而,缺乏有效的作物生长模型及适当的植物表型特征在线测量方法,已成为阻碍作物最优控制算法实施的重要因素。例如,传统测量生菜鲜质量的方法通常是破坏性采样,即将生菜从栽培板上拔出,抖落叶片上的多余水分后进行称量[4]。同时,在作物群体表型特征估算中,主观选择样本和取平均值的过程也容易产生误差。因此,开发非破坏性的间接测量方法对实现作物生长全过程检测监测具有实际意义[5]。

通过计算机技术实现作物表型特征的精准、自动提取始于 20 世纪 90 年代。早期研究者通常采用传统浅层机器学习方法,通过颜色空间将作物图像与背景分割。如通过人工经验提取与作物长势相关的表型特征,并将这些特征与标注的作物生长数据拟合,进行表型特征估计[6]。然而,这些方法不但需要人工调参,而且模型的泛化能力较差,缺乏通用性。

近年来,随着深度神经网络技术不断发展,基于深度学习的图像目标检测技术为植物表型特征测量提供了强大支撑[7]。例如,GANG 等[8]通过将 RGB 与深度图像输入特征提取网络,再通过多个估计网络完成特定场景下生菜表型参数的估计。然而,由于背景也被纳入特征提取,影响了表型特征估计结果的泛化能力,限制了这些模型在实际栽培场景中的应用。目前基于深度和 RGB 图像估算绿叶菜表型参数的报道较少,大多数研究是关于果实的质量预测[9-10]。

本研究提出一种双阶段生菜表型参数估算方法。首先,基于深度摄像头捕获的 RGB 和深度图像,训练改进的 DeepLabv3 +[11]模型进行生菜分割,以去除无关背景干扰;然后采用双模态表型参数回归模型对分割后图像进行特征提取,并经双模态融合后输入回归模型,对生菜关键表型特征进行估算。

1 材料与方法

1.1 数据集

采用 HEMMING 等[12]制作的温室生菜数据集。该数据集包含 4 个生菜品种的 RGB 图像、深度图像及其对应的干质量、湿质量、叶面积、株高和冠幅等表型特征实测数据。其中,图像数据使用 Intel Realsense D415 型相机在距离植株 0.9 m 处拍摄。深度图像为单通道 16 位 PNG 格式,RGB 图像为标准 3 通道 RGB 的 PNG 格式,分辨率均为 1 080 像素 × 1 920 像素。拍摄过程中对每株植株均保持一致的拍摄高度和参数设置。表型特征数据则通过破坏性方法获得。湿质量通过称量每株植株的地上部分获取,干质量通过叶片干燥后的质量测定。冠幅表示生菜顶部最长两端间的距离,叶面积为将全部叶片展平后测算的面积,株高为植株基部第一片叶的附着点到植物最高处的距离。

数据集包含 4 个生菜品种为 Aphylion、Lugano、Satine 和 Slanova。所有植株均种植于荷兰一水培控制温室中,每周采集一次生长数据,直至 49 d 后收获。每个品种对应的样本数量分别为 92、96、97、102。

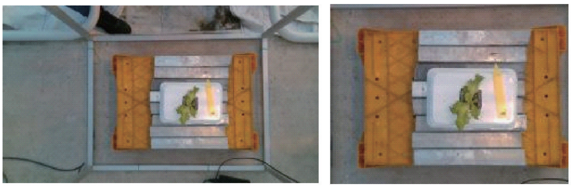
1.2 图像处理

本研究的图像处理过程分为两个阶段:首先,构建用于提取生菜部分与背景的分割模型;其次,基于分割后的图像和深度信息,构建 RGB 和深度融合表型特征估计模型。

针对分割模型,为去除架子旁的无关背景,使用 OpenCV 对数据集进行预处理。从拍摄架的左上角金属支架位置(x, y) = (550 像素, 200 像素)开始,裁切出 1 000 像素 ×750 像素的区域。这一预处理步骤有助于减少拍摄背景对模型的干扰。随后,采用 Labelme 对预处理图像进行标注,以区分生菜与背景,处理前后图像示例如图 1 所示。在标注过程中,针对 Aphylion、Lugano、Satine 和 Slanova 不同类型生菜,将 56 株样本均匀划分为测试集,其余 331 株作为训练集。

在模型训练阶段,使用 MMsegmentation 提供的默认数据集配置文件,设置 batch_size 为 4,尺度因子分别为 0.5、0.75、1.0、1.25、1.5、1.75,不采用水平翻转,以确保数据的均一化和标准化。

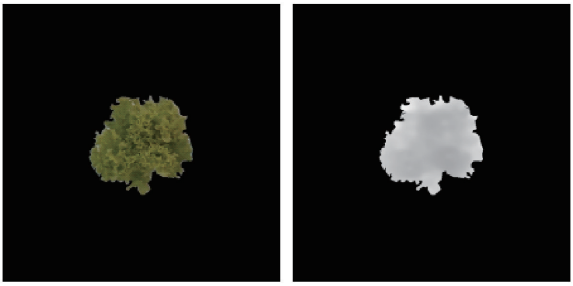
在表型特征参数估计阶段,首先使用获得的生菜分割掩膜提取生菜部分。RGB 图像和深度图像采用一致的提取方式。如图 2a 所示,具体的提



(a) 原始图像　(b) 预处理后图像

图 1　原始图像和预处理后图像示例

Fig. 1　Original image and preprocessed image



(a) RGB图像　(b) 深度图像

图 2　处理后 RGB 图像和深度图像

Fig. 2　Processed RGB image and depth image

取方法为：在掩膜图像中识别生菜部分后，保留 RGB 和深度图像中相应像素点的原始值；对于掩膜图中的背景部分，则将 RGB 和深度图像的 3 个通道值置为 0，并统一裁切为 720 像素 ×720 像素的图像。

随后，对深度图进行后处理。通过 OpenCV 中的 convertScaleAbs 函数将深度图像中生菜区域的深度值(650 ~905)转换为 8 位色深的 RGB 图像。结果如图 2b 所示，这一处理步骤在保留图像深度信息的同时，确保深度模型能够直接对深度图像进行有效的特征提取。

1.3　损失函数及评价指标

1.3.1　轻量化生菜分割模型损失函数及评价指标

采用交叉熵损失函数，该损失函数能够有效衡量分割模型性能。采用平均 Dice 系数(mDice)、平均准确率(mACC)和平均交并比(mIoU)作为评价指标。

1.3.2　模态融合并行表型特征回归模型损失函数及评价指标

在回归模型中使用 MSE(均方误差损失)作为模型计算过程的损失函数。为评估模型性能，采用回归决定系数(R^2)和相对均方根误差(RRMSE)作为评价指标。

2　模型建立与训练

表型特征回归任务中，干质量和湿质量主要依赖于 RGB 图像中的高级语义信息，而冠幅、株高和叶面积则更依赖于植株的形态信息。这些指标反映了生菜的生长过程，对其经济价值的评估具有重要意义。采用一种两阶段的方法进行模型训练。

在第 1 阶段，使用 DeepLabv3 + MobileViTv2[13] 模型进行分割，旨在增强全局特征提取能力，从而生成精确的生菜区域掩膜，并有效去除无关的背景信息。这一过程有助于提高后续表型特征估计的准确性。

在第 2 阶段，利用注意力机制选择性地加强不同模态的空间和通道特征信息，同时结合特征融合策略，实现对生菜 RGB 图像和深度图像的模态融合精准表型特征回归。

2.1　训练环境

软件环境为 Windows 11 64 位系统，使用 Python 3.10 环境运行，深度学习框架为 Pytorch 2.1，硬件环境为 Intel ® Core(R) CPU i9 - 10900K 处理器和 NVIDIA TITAN RTX GPU，使用 NVIDIA CUDA12.1 进行 GPU 加速。

2.2　轻量化生菜分割模型建立

语义分割任务旨在输出生菜的掩膜[14]。第 1 阶段生菜分割模型采用基于卷积神经网络(CNN)的 DeepLabv3 + 模型。模型的输入图像为 1 000 像素 ×750 像素，使用 MobileViTv2 作为骨干网络。DeepLabv3 + 模型结构如图 3 所示。在模型训练过程中使用 MMsegmentation 的默认优化器和学习率调度器，配置初始学习率为 0.01，动量为 0.9，权重衰减系数为 0.000 5，衰减指数为 0.9，最小学习率为 0.000 1，并进行 40 000 次迭代，采用旋转、垂直翻转等数据增强方法与初始图像共同训练。

MobileViTv2 是一种结合卷积与 Transformer 架构的轻量化骨干网络，具备卷积局部特征提取能力与视觉 Transformer[15](ViT)全局特征建模能力，因而在轻量化骨干网络中具有竞争优势。为解决多头自注意力(MHA[16])模型的高计算复杂度问题，MobileViTv2 引入了可分离自注意力机制。不同于多头注意力(MHA)，可分离自注意力机制仅计算与潜在 token L 相关的上下文得分，从而将计算复杂度从 $O(k^2)$ 降低至 $O(k)$。

在特征提取阶段，将 MobileViTv2 模型提取的低维特征作为解码器的输入，将最终提取的高维特征输入到 DeepLabv3 + 的编码器部分，以生成后续掩膜。

DeepLabv3 + 结合了编码器-解码器结构和膨胀卷积。编码器模块对输入大小为 $H \times W \times C$ 的特征图进行处理，通过一个 3 ×3 的卷积层对局部空间信息进行编码。通过 1 ×1 卷积层提升特征通道数，学习输入通道的线性组合。随后，逐层级联将数据投影高维空间 d，此时特征图 x 的大小为 $H \times W \times d$。

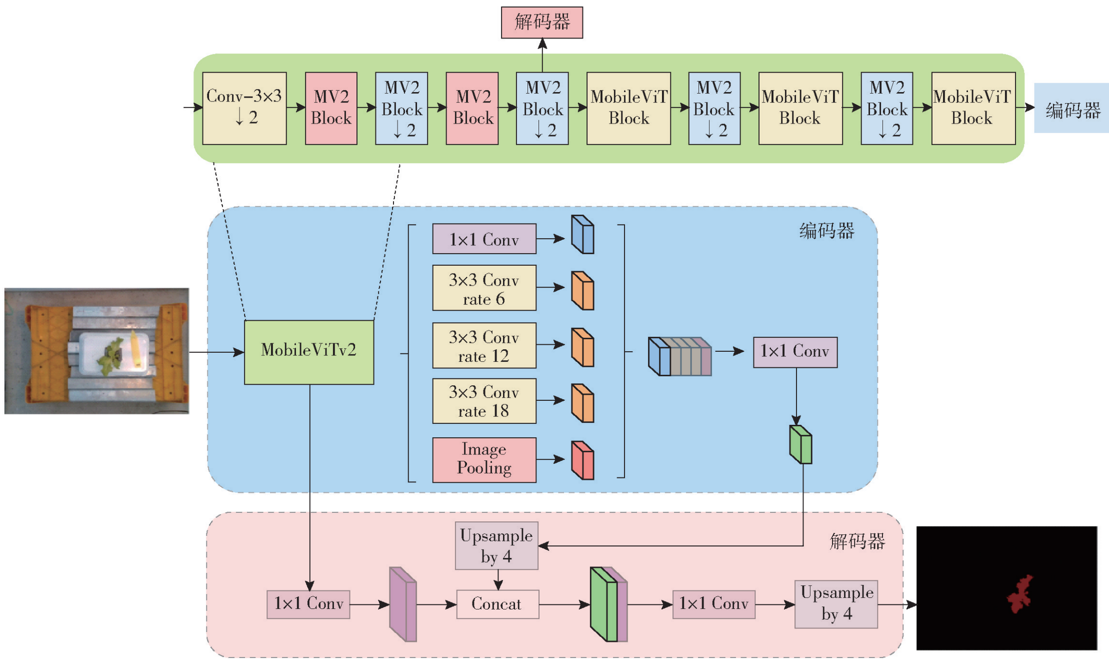


图 3　DeepLabv3 + MobileViTv2 分割模型框图

Fig. 3　DeepLabv3 + MobileViTv2 segmentation model

为了使模型能够学习具有空间分离卷积归纳偏差的全局表示，将特征图 x 拆分成 N 个不重叠的图像块，每个图像块大小为 $h \times w$。此时，图像块序列 X 大小为 $P \times N \times d$，其中 $P = h \times w$，$N = \frac{H \times W}{P}$。

通过膨胀卷积编码图像块之间的关系，实现全局信息关注。为避免信息丢失，需将编码后的图像块还原至编码前的维度。解码器模块会将编码器特征双线性上采样 4 倍，并与具有相同空间分辨率的低级特征进行连接。低级特征通常包含大量通道（如 256 或 512），在连接之前，使用 1 × 1 卷积减少通道数。随后，应用多个 3 × 3 卷积进一步精炼特征，最终通过一个简单的双线性上采样 4 倍生成高分辨率特征图。

膨胀卷积可以显式控制深度卷积神经网络计算的特征分辨率，并调整滤波器的视野以捕获多尺度信息。对于二维信号，膨胀卷积在输入特征图 x 上的应用如下：$y[i] = k\sum x[i + dot(r,k)]w[k]$。其中，膨胀率 r 决定了采样输入信号的步幅。标准卷积是膨胀率 $r = 1$ 的特例。这种设计的 DeepLabv3 + 在保持高效计算的同时，能够进行多尺度信息的提取和细粒度特征的恢复，从而高效生成高质量的分割掩码。

2.3　双模态表型参数估计模型

通过计算得出的掩膜去除了生菜无关部分后，将得到的深度图像和 RGB 图像输入到基于 ResNet - 10[17-18] 改进的双模态分支并行回归模型中。输入回归模型后，将输出 5 种表型特征的估计值。

在 PyTorch[19] 平台上进行模型训练。为增强模型的性能，提出了一种卷积双模态融合模块（CMMCM），该模块针对包含丰富图像空间信息的深度特征图进行空间信息增强，同时对 RGB 图像提取的高级语义信息进行通道增强，随后利用卷积操作实现双模态融合。CMMCM 模块由 CBAM[20]（Convolutional block attention module）及其子模块组成。CBAM 通过顺序推断出两个维度（通道和空间）上的注意力图，然后将这些注意力图乘以输入特征图，以进行自适应特征修正。CBAM 包含 2 个子模块：通道注意力模块（CAM）和空间注意力模块（SAM）。

通道注意力模块（CAM）通过对特征图进行最大池化和平均池化，生成两个特征图。接着，这两个特征图经过全连接层和 ReLU 激活函数处理，最终通过 sigmoid 函数生成通道注意力图，从而使模型关注有意义的通道信息。

空间注意力模块（SAM）则通过对特征图进行最大值和平均值操作，生成两个特征图，将这两个特征图拼接，随后通过卷积层和 sigmoid 函数生成空间注意力图，以从空间层面关注信息分布。

卷积注意力机制模块（CBAM）按照特定顺序对原有特征图 F 应用通道注意力和空间注意力。首先，CBAM 生成通道注意力图 $M_c(F)$，并将其与输入特征图 F 逐通道相乘得 $F_c = M_c(F)F$。

CBAM 生成空间注意力图 $M_s(F_c)$，并与前一步的输出特征图逐元素相乘得 $F_s = M_s(F_c)F_c$。通过这种方式，模型能够在通道和空间维度上有效地捕

捉并强调重要信息。

如图 4 所示，CMMCM 通过 CAM 增强 RGB 通道，使得模型能够关注到有意义的三通道特征信息；同时通过 SAM 增强深度信息，使得模型能够捕捉到空间维度的特征信息。随后，通过一个卷积层进一步对两种模态进行融合，融合后的结果将输入完整的空间通道注意力模块，以进一步整合双模态信息。

利用该模块对 ResNet－10 进行了改进，在基准模型特征提取的中间层中加入 CMMCM 以增强双模态特征融合，在特征图链接后添加 SE 注意力模块自适应校准通道信息，随后将校准后1 000 维张量输入包含两层全连接层的多层感知机（MLP），公式为

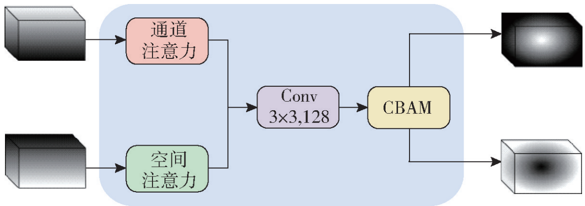


图 4　卷积双模态融合模块（CMMCM）

Fig. 4　Convolutional multi-modal fusion module (CMMCM)

$$y=\phi_2(\boldsymbol{W}_2(\phi_1(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1))+\boldsymbol{b}_2)$$

其中 $\boldsymbol{x}$ 为输入张量，乘以权重矩阵 $\boldsymbol{W}_1$ 加上偏置 $\boldsymbol{b}_1$ 经过修正线性单元（ReLU）激活函数 ϕ_1，输出 128 维张量，第 2 个全连接层将 128 维张量降维到 5 维，输出预测表型特征，以此构建了卷积双模态融合回归模型（CMMCRN），如图 5 所示。

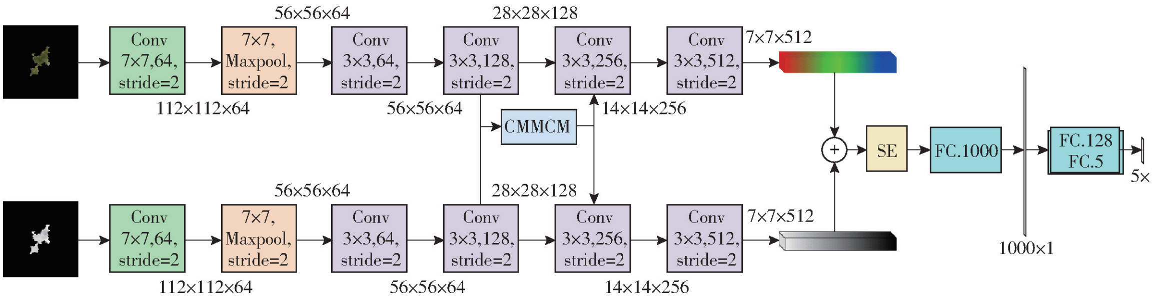


图 5　卷积双模态融合回归模型（CMMCRN）

Fig. 5　Convolutional multi-modal fusion regression model (CMMCRN)

实验在 Windows 系统环境下使用 TITAN RTX GPU 训练，输入图像统一处理成 224 像素 × 224 像素。学习率固定为 0.001，每轮训练 25 幅图像，总共训练 200 轮，比较结果并选择最优模型作为后续的回归模型。

3　实验结果和分析

3.1　生菜分割模型性能分析

3.1.1　结果分析

对提出的 DeepLabv3 + MobileViTv2 改进模型在 4 个生菜分割数据集上进行了训练，训练的验证结果如图 6 所示。模型经过 40 000 次迭代，每 100 次迭代称为 1 个轮次，并在每 500 次迭代后进行一轮模型测试，共计 480 轮次。由图可见，随着迭代次数增加，模型分割准确性逐渐提升并趋于稳定。最终，模型训练曲线趋于平稳，表明训练结果表现良好。

在使用最终训练得到的模型权重后，改进后的模型在 Aphylion、Lugano、Salanova 和 Satine 数据集上的 mDice 评价指标分别达到 0.994 2、0.993 1、0.994 5 和 0.993 6，表明模型具有出色的分割性能，测试结果如图 7 所示。

3.1.2　对比实验

为验证 DeepLabv3 + 与 MobileViTv2 组合的模型性能及其泛化能力，将其与多种常见分割模型进

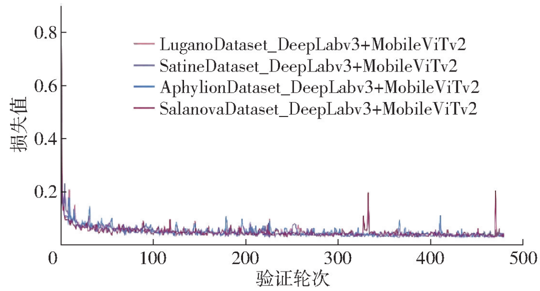


图 6　4 类生菜训练损失曲线

Fig. 6　Training loss curves for four lettuce varieties

图 7　分割模型样例

Fig. 7　Examples of segmentation models

行性能对比，包括 KNet[21]、FastRCNN[22]、PSPNet[23]，以及替换 MobileNetv3[24]、ResNet－18[17]和 MobileNetv4[25]骨干网络的 DeepLabv3＋。其中，本文将 DeepLabv3＋MobileViTv2 模型与具有竞争力的 KNet－tiny 模型分割结果进行可视化对比，如图 8 所示。结果显示，在 Satine 生菜的右下方空隙部分，DeepLabv3＋MobileViTv2 模型对生菜叶片间隙的分割表现出色。

(a) DeepLabv3+MobileViTv2 分割示例　(b) 原始图像　(c) KNet分割示例

图 8　分割模型性能对比

Fig. 8　Performance comparison of segmentation models

和其它常见目标分割模型性能对比如表 1～3 所示。本文改进的模型在 Satine、Salanova、Lugano 生菜品种上取得了更好的性能，而在 Aphylion 生菜品种上的 mDice 评分为第 3，mIoU 和 mACC 的性能次优，说明了模型具有较好的泛化性，适合于多种生菜品种图像分割任务，MobileViTv2 的全局关注注意力增强了模型对高级语义的理解。

表 1　分割模型 mDice 实验结果

Tab. 1　Performance metrics for segmentation models（mDice）

模型	数据集			
	Aphylion	Satine	Lugano	Salanova
DeepLabv3＋MobileNetv4	0.9941	0.9036	0.9923	0.9831
DeepLabv3＋MobileNetv3	0.9941	0.9924	0.9926	0.9938
DeepLabv3＋R18	0.9947	0.9934	0.9921	0.9941
PSPNet	0.9944	0.9926	0.9930	0.9935
FastRCNN	0.9917	0.9902	0.9907	0.9901
KNet－swin－t	0.9932	0.9929	0.9922	0.9935
DeepLabv3＋MobileViTv2	0.9942	0.9936	0.9931	0.9945

表 2　分割模型 mIoU 实验结果

Tab. 2　Performance metrics for segmentation models（mIoU）　%

模型	数据集			
	Aphylion	Satine	Lugano	Salanova
DeepLabv3＋MobileNetv4	98.84	83.28	98.49	96.71
DeepLabv3＋MobileNetv3	98.82	98.50	98.55	98.78
DeepLabv3＋R18	98.94	98.70	98.45	98.83
PSPNet	98.90	98.54	98.61	98.72
FastRCNN	98.63	98.07	98.16	98.05
KNet－swin－t	98.65	98.60	98.47	98.72
DeepLabv3＋MobileViTv2	98.86	98.73	98.64	98.92

从图 8 也可以看出本文改进模型对边缘分割更为精细。

表 3　分割模型 mACC 性能指标

Tab. 3　Performance metrics for segmentation models（mACC）　%

模型	数据集			
	Aphylion	Satine	Lugano	Salanova
DeepLabv3＋MobileNetv4	99.52	96.34	99.45	97.48
DeepLabv3＋MobileNetv3	99.50	99.27	99.42	99.52
DeepLabv3＋R18	99.57	99.30	99.52	99.44
PSPNet	99.51	99.01	99.29	99.31
FastRCNN	99.42	98.93	99.11	99.00
KNet－swin－t	99.27	99.17	99.17	99.35
DeepLabv3＋MobileViTv2	99.52	99.35	99.41	99.43

3.2　表型特征估算结果分析

3.2.1　表型特征回归基准模型比较分析

为选择适用于 5 种表型特征回归的基准模型，分别选用 MobileNetv4、MobileViTv2 和 ResNet－18 作为对比模型，针对 RGB 图像和深度图像分别进行特征提取，并输入多层感知机回归头。对比实验结果如表 4、5 所示。

表 4　RGB 图像回归性能对比

Tab. 4　Comparison of regression performance for RGB images

模型		鲜质量	干质量	冠幅	叶面积	株高
R^2	ResNet－10	0.9168	0.9148	0.7923	0.8950	0.8319
	ResNet－18	0.8576	0.8609	0.6488	0.8614	0.8020
	MobileNetv4	0.8847	0.8842	0.7750	0.8692	0.8173
	MobileViTv2	0.8871	0.8966	0.7935	0.8852	0.8203
RRMSE	ResNet－10	0.2864	0.2957	0.4598	0.3273	0.5125
	ResNet－18	0.3463	0.3421	0.4737	0.3374	0.4209
	MobileNetv4	0.3083	0.3120	0.4136	0.3263	0.4021
	MobileViTv2	0.3056	0.2946	0.4084	0.3136	0.4008

表 5　深度图像回归性能对比

Tab. 5　Comparison of regression performance for depth images

模型		鲜质量	干质量	冠幅	叶面积	株高
R^2	ResNet－10	0.8917	0.8952	0.7431	0.8918	0.8333
	ResNet－18	0.8784	0.8669	0.7545	0.8517	0.8043
	MobileNetv4	0.8862	0.8806	0.7831	0.9010	0.8437
	MobileViTv2	0.8569	0.8680	0.7605	0.8745	0.8217
RRMSE	ResNet－10	0.3112	0.4258	0.4651	0.3316	0.4708
	ResNet－18	0.3213	0.3336	0.4309	0.3518	0.4140
	MobileNetv4	0.3103	0.3189	0.4169	0.2925	0.3717
	MobileViTv2	0.3518	0.3405	0.4548	0.3321	0.3949

实验结果表明，ResNet－10 和 MobileViTv2 在使用 RGB 图像进行 5 种表型特征回归时表现较佳，而 MobileNetv4 在纯深度图像回归中表现出色。然而，由于 MobileNetv4 在 RGB 图像回归性能相对较低，最终选择在 RGB 和深度图像上均具竞争力的

ResNet－10 作为基准模型。

3.2.2 双模态回归模型有效性验证分析

基于 ResNet－10 基准模型，构建了一个并行分支，分别计算 RGB 和深度图像的特征图，并在最后将两种模态的特征图拼接后输入多层感知机进行回归。单模态与双模态回归性能对比如表 6 所示。

表 6 RGB、深度图像单模态与双模态模型性能对比

Tab. 6 Comparison of performance between RGB, depth single-modal and dual-modal models

模态类型	表型参数	R^2	RRMSE
双模态	鲜质量	0.899 275	0.283 050
	干质量	0.908 275	0.240 150
	冠幅	0.849 375	0.098 525
	叶面积	0.908 700	0.229 200
	株高	0.846 200	0.134 625
单模态（RGB 图像）	鲜质量	0.916 775	0.286 350
	干质量	0.914 750	0.295 675
	冠幅	0.792 250	0.459 800
	叶面积	0.895 025	0.327 275
	株高	0.831 900	0.512 525
单模态（深度图像）	鲜质量	0.891 650	0.311 225
	干质量	0.895 200	0.425 750
	冠幅	0.743 125	0.465 100
	叶面积	0.891 800	0.331 600
	株高	0.833 250	0.470 775

在并行结构下，得益于深度图像提供的植物外观形态与植株表面各点到相机的视距差[26]，模型冠幅、株高和叶面积的回归性能获得明显提升，在 RRMSE 这一性能指标上体现的尤为明显。与此同时，在鲜质量和干质量的回归性能上出现一定下降，推测是由于 ResNet－10 双模态模型仅在最终特征融合时进行模态合并，缺乏对双模态低分辨率特征的融合，影响了模型对双模态信息的理解和利用。

3.2.3 消融实验

为验证 CMMCM 模块与 SE 模块[27]对 ResNet－10 性能的改进效果，在验证集上进行了消融实验。分别去除 CMMCM＋SE（ResNet－10）中的 CMMCM 和 SE 模块，其余部分不变。4 种模型在 R^2、RRMSE 性能指标上对比如表 7 所示。

结果可视化见图 9、10。实验结果表明，在仅保留 CMMCM 并去除 SE 模块后，本文改进模型在叶面积、株高和冠幅等依赖深度模态的表型特征预测任务上表现出了较好的性能，分别提升 1.239 1%、0.429 2%、1.760 8%，表明 CMMCM 模块在双模态特征理解上发挥了作用。

同时，SE 模块的引入则有助于增强模型对高级语义信息的捕捉，从而提升了鲜质量（2.944 0%）和干质量（2.193 7%）的回归性能。综合来看，在添加 CMMCM 和 SE 模块后，模型在不同数据集的整体性能均有大幅提升，特别是在回归系数指标 R^2 上，相较于基准 ResNet－10（双模态）模型获得了显著提升。

表 7 消融实验性能对比结果

Tab. 7 Comparison of performance results from ablation experiments

模型	表型参数	R^2	RRMSE
ResNet－10（双模态）	鲜质量	0.899 275	0.283 050
	干质量	0.908 275	0.240 150
	冠幅	0.849 375	0.098 525
	叶面积	0.908 700	0.229 200
	株高	0.846 200	0.134 625
ResNet－10＋SE	鲜质量	0.925 750	0.239 425
	干质量	0.928 200	0.212 350
	冠幅	0.829 250	0.105 225
	叶面积	0.912 975	0.225 025
	株高	0.854 750	0.125 750
ResNet－10＋CMMCM	鲜质量	0.896 600	0.285 500
	干质量	0.909 750	0.236 325
	冠幅	0.859 900	0.095 550
	叶面积	0.912 600	0.226 150
	株高	0.861 100	0.129 775
ResNet－10＋SE＋CMMCM	鲜质量	0.922 15	0.246 125
	干质量	0.931 35	0.208 025
	冠幅	0.861 950	0.094 900
	叶面积	0.935 900	0.189 150
	株高	0.887 475	0.114 900

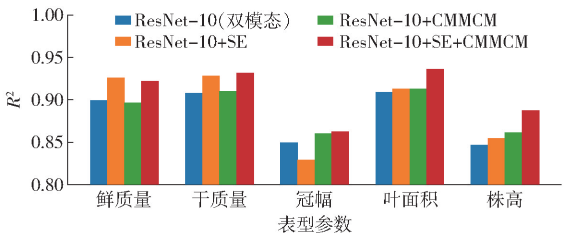


图 9 消融实验 R^2 性能对比

Fig. 9 Comparison of R^2 performance from ablation experiments

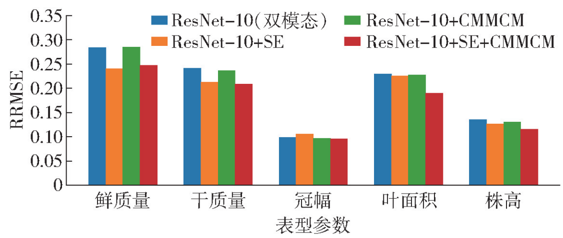


图 10 消融实验 RRMSE 性能对比

Fig. 10 Comparison of RRMSE performance from ablation experiments

3.2.4 不同设备上推理时间测试

同时在服务器和便携式计算机上部署了本文提出的表型特征回归模型，以测试模型推理速度。在配置了 NVIDIA TITAN RTX 显卡服务器上，估算单

幅生菜图像 5 种表型特征的耗时约为 21.3 ms，而在普通计算机上单幅图像推理时间为 44.8 ms，说明该模型具有实时检测性能。

4 结论

（1）通过在 DeepLabv3+ 模型中替换为 MobileViTv2 骨干网络，提升了模型全局特征注意能力，从而提高了生菜分割精度，在 4 个生菜数据集上的 mDice 分别达到 0.994 2、0.993 1、0.994 5 和 0.993 6，对不同生菜品种分割性能优良，去除了背景干扰。

（2）提出的 CMMCM 和 SE 模块有效促进了模型低级特征的双模态融合，使得模型能够更加高效地利用双模态信息。在实验中，模型在 4 个生菜品种鲜质量、干质量、冠幅、叶面积和株高等表型特征的估算任务上的决定系数分别达到 0.922 2、0.931 4、0.862 0、0.935 9 和 0.887 5。与未添加卷积双模态回归模块和 SE 模块的原始模态模型 ResNet-10（双模态）相比，本文改进模型在 5 个表型特征参数指标预测上决定系数分别提高 2.54%、2.54%、1.48%、2.99% 和 4.88%，相对均方根误差分别减少 3.69%、3.21%、0.36%、4.01% 和 1.97%。在与 ResNet-10、MobileViTv2、MobileNetv4 和 ResNet-18 等单模态 RGB 图像或深度图像回归模型相比，所提方法在整体性能上均表现出显著优势，能够更加准确地针对生菜进行更精准的表型特征预测。

（3）在模型推理性能方面，在普通计算机上单幅生菜图像表型特征估算耗时约为 44.8 ms，可以满足嵌入式设备低成本部署和实时检测需求。

参 考 文 献

[1] ZHANG X, HE D, NIU G, et al. Effects of environment lighting on the growth, photosynthesis, and quality of hydroponic lettuce in a plant factory[J]. International Journal of Agricultural and Biological Engineering, 2018, 11(2): 33-40.

[2] SUBLETT W L, BARICKMAN T C, SAMS C E. The effect of environment and nutrients on hydroponic lettuce yield, quality, and phytonutrients[J]. Horticulturae, 2018, 4(4): 48.

[3] 黄林生，邵松，卢宪菊，等．基于卷积神经网络的生菜多光谱图像分割与配准[J]．农业机械学报，2021，52(9)：186-194.
HUANG Linsheng, SHAO Song, LU Xianju, et al. Multispectral image segmentation and registration of lettuce based on convolutional neural networks[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(9): 186-194. (in Chinese)

[4] 董默．基于深度学习的生菜表型精准鉴定与应用研究[D]．长春：吉林大学，2024.
DONG Mo. Precise identification and application research of lettuce phenotypes based on deep learning[D]. Changchun: Jilin University, 2024. (in Chinese)

[5] 马义东，胡鹏展，金鑫，等．水培生菜低损柔性采收装置设计与试验[J]．农业机械学报，2022，53(10)：175-183，210.
MA Yidong, HU Pengzhan, JIN Xin, et al. Design and experiment of low damage flexible harvesting device for hydroponic lettuce[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(10): 175-183, 210. (in Chinese)

[6] 刘林，苑进，张岩，等．日光温室基质培生菜鲜质量无损估算方法[J]．农业机械学报，2021，52(9)：230-240.
LIU Lin, YUAN Jin, ZHANG Yan, et al. Non-destructive estimation method of fresh weight of substrate cultured lettuce in solar greenhouse[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(9): 230-240. (in Chinese)

[7] 蒋心璐，陈天恩，王聪，等．农业害虫检测的深度学习算法综述[J]．计算机工程与应用，2022，59(6)：30-44.
JIANG Xinlu, CHEN Tian'en, WANG Cong, et al. Survey of deep learning algorithms for agricultural pest detection[J]. Computer Engineering and Applications, 2022, 59(6): 30-44. (in Chinese)

[8] GANG M S, KIM H J, KIM D W. Estimation of greenhouse lettuce growth indices based on a two-stage CNN using RGB-D images[J]. Sensors, 2022, 22(15): 5499.

[9] 李杰，王俊，李波，等．基于 CBAM 的 CNN-Bi-Coupled LSTM 的苹果产量预测[J]．信息技术与信息化，2023(10)：4-7.
LI Jie, WANG Jun, LI Bo, et al. Apple yield prediction based on CBAM's CNN-Bi-Coupled LSTM[J]. Information Technology and Informatization, 2023(10): 4-7. (in Chinese)

[10] 张润芝，张晓，吴刚．基于 Kinect 相机的香梨重量预测方法[J]．食品与机械，2023，39(9)：77-82，88.
ZHANG Runzhi, ZHANG Xiao, WU Gang. A method for predicting the weight of fragrant pears based on Kinect camera[J]. Food and Machinery, 2023, 39(9): 77-82, 88. (in Chinese)

[11] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Computer Vision-ECCV 2018, Lecture Notes in Computer Science, 2018: 833-851.

[12] HEMMING S, DE ZWART F, ELINGS A, et al. 3rd autonomous greenhouse challenge: online challenge lettuce images [DS/OL]. (2011-08-18)[2024-06-05]. https://doi.org/10.4121/15023088.v1.

[13] MEHTA S, APPLE M. Separable self-attention for mobile vision transformers[J]. arXiv Preprint, arXiv:2206.02680, 2022.

（下转第 101 页）

[36] BENMOUNA B, GARCÍA-MATEOS G, SABZI S, et al. Convolutional neural networks for estimating the ripening state of fuji apples using visible and near-infrared spectroscopy[J]. Food and Bioprocess Technology, 2022, 15(10): 2226 –2236.

[37] ZENG J, GUO Y, HAN Y, et al. A review of the discriminant analysis methods for food quality based on near-infrared spectroscopy and pattern recognition[J]. Molecules, 2021, 26(3): 749.

[38] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700 –4708.

[39] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. PMLR, 2019: 6105 –6114.

[40] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314 –1324.

[41] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132 –7141.

[42] WOO S, PARK J, LEE J, et al. Cbam: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3 –19.

[43] LIU Y, SHAO Z, HOFFMANN N. Global attention mechanism: retain information to enhance channel-spatial interactions[J]. arXiv Preprint, arXiv:211205561, 2021.

[44] WANG Q, WU B, ZHU P, et al. ECA – Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534 –11542.

[45] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13713 –13722.

(上接第 91 页)

[14] 蒋易宇, 王硕, 张丽娜, 等. 基于水培生菜力学特征的成熟度分类方法[J]. 农业工程学报, 2023, 39(1): 179 –187.
JIANG Yiyu, WANG Shuo, ZHANG Li'na, et al. Maturity classification method based on the mechanical characteristics of hydroponic lettuce[J]. Transactions of the CSAE, 2023, 39(1): 179 –187. (in Chinese)

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv Preprint, arXiv: 2010.11929, 2020.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017,30:5998 –6008.

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

[18] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] ANSEL J, YANG E, HE H, et al. PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation[C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024, 5(2):929 –947.

[20] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Computer Vision-ECCV 2018,Lecture Notes in Computer Science, 2018:3 –19.

[21] ZHANG W, PANG J, CHEN K, et al. K – Net: towards unified image segmentation[J]. Advances in Neural Information Processing Systems, 2021, 34: 10326 –10338.

[22] GIRSHICK R. Fast R – CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[23] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[24] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[25] QIN D, LEICHNER C, DELAKIS M, et al. MobileNetV4 – universal models for the mobile ecosystem[J]. arXiv Preprint, arXiv:2404.10518, 2024.

[26] 胡松涛,翟瑞芳,王应华,等.基于多源数据的马铃薯植株表型参数提取[J].智慧农业(中英文),2023,5(1):132 –145.
HU Songtao, ZHAI Ruifang, WANG Yinghua, et al. Extraction of potato plant phenotypic parameters based on multi-source data[J]. Smart Agriculture, 2023, 5(1): 132 –145. (in Chinese)

[27] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.