

基于改进 RTMPose3D 模型的番茄三维关键点估计方法

王蓬勃^{1,2} 刘宇^{1,2} 赵胜辉^{1,2} 傅毅凯^{1,2}

(1. 苏州大学机电工程学院, 苏州 215137; 2. 江苏省具身智能机器人技术重点实验室, 苏州 215137)

摘要: 针对温室环境下串番茄枝叶遮挡严重、光照干扰强,导致自动采摘机器人难以稳定获取果实三维位姿的问题,提出了一种改进的串番茄三维关键点估计模型 TomatoPose3D。该模型在训练阶段引入 RGB 图像与三维真值关键点的联合约束,增强结构一致性与泛化能力;在推理阶段,仅输入单幅 RGB 图像即可端到端回归三维关键点坐标,避免了因点云缺失或稀疏导致的定位失败。改进模型以 RTMPose3D 为基线,引入全局结构感知模块 MobileVit Block 与分布感知解码策略 (DARK),在保持轻量化的同时提升了定位精度。温室场景对比实验表明, TomatoPose3D 的 PCK@0.05 指标较 RTMPose3D 和 SimpleBaseline3D 分别提升 5.18、9.98 个百分点;在无深度信息辅助下,其定位精度与基于 RGB-D 投影的方法相当,且鲁棒性更优。此外,模型经 TensorRT 加速部署于工业级嵌入式平台,端到端推理速度达 37 f/s,满足采摘机器人实时空间视觉感知需求。

关键词: 温室番茄; 采摘机器人; 三维关键点估计; 轻量化网络; RTMPose3D

中图分类号: S24

文献标识码: A

文章编号: 1000-1298(2026)05-0149-10

OSID:



Tomato 3D Keypoint Estimation Method Based on Improved RTMPose3D Model

WANG Pengbo^{1,2} LIU Yu^{1,2} ZHAO Shenghui^{1,2} FU Yikai^{1,2}

(1. College of Mechanical and Electrical Engineering, Soochow University, Suzhou 215137, China

2. Jiangsu Key Laboratory of Embodied Intelligent Robot Technology, Suzhou 215137, China)

Abstract: Aiming to address the challenge of reliably acquiring 3D pose information of truss tomatoes for autonomous harvesting robots under conditions of severe occlusion and strong light interference in greenhouses, an improved 3D keypoint estimation model named TomatoPose3D was proposed. During the training phase, the model incorporated joint constraints between RGB images and 3D ground-truth keypoints to enhance structural consistency and generalization capability. In the inference phase, the model can end-to-end regress 3D keypoint coordinates from a single RGB image, thereby avoiding localization failures caused by sparse or missing point clouds. Based on the RTMPose3D baseline, the improved model introduced the global structure-aware MobileVit Block and the distribution-aware coordinate representation of keypoints (DARK) decoding strategy, improving localization accuracy while maintaining a lightweight architecture. Comparative experiments in greenhouse scenarios indicated that TomatoPose3D improved the PCK@0.05 score by 5.18 and 9.98 percentage points compared with RTMPose3D and SimpleBaseline3D, respectively. Without the assistance of depth information, the model achieved localization accuracy comparable to RGB-D projection-based methods while demonstrating superior robustness. Furthermore, the model was deployed on an industrial-grade embedded platform accelerated by TensorRT, achieving an end-to-end inference speed of 37 f/s, which met the real-time spatial visual perception requirements of harvesting robots.

Key words: greenhouse tomato; harvesting robot; 3D keypoint estimation; lightweight network; RTMPose3D

0 引言

为了实现高效、精确的果实采摘作业,自动化采

摘机器人需要准确获取果实的空间位置与姿态信息,然而实际果实的生长环境常伴随着严重遮挡、光照干扰等现象,显著提高了自动化采摘机器人视觉

收稿日期: 2025-12-22 修回日期: 2025-12-24

基金项目: 苏州市科技强农创新项目 (SNG2025009) 和国家重点研发计划项目 (2022YFB4702202)

作者简介: 王蓬勃 (1984—), 男, 副教授, 博士, 主要从事农业机器人和农业人工智能研究, E-mail: pbwang@suda.edu.cn

系统的识别难度^[1]。

近年来,深度学习方法的迅速发展为农业视觉感知问题提供了新的解决思路,凭借其强大的特征识别能力,深度学习技术在农业视觉感知领域得到广泛应用,显著推动了自动化采摘机器人的发展。然而现有大部分工作^[2-8]主要聚焦于获取果实二维信息,如成熟度、果实尺寸和位置等,忽视了包含空间位置与三维姿态在内的果实三维信息。事实上,三维信息是采摘机器人选择抓取策略的重要依据,其精度直接影响采摘作业成功率与效率。

近年来部分研究开始关注果实三维信息的获取,目前相关方法主要分为三维重建^[9-11]与关键点检测两类,前者常利用多视角 RGB 图像融合或改进神经辐射场(NeRF)模型^[12]等方法实现果实的高精度三维重建,尽管这些方法能较为完整地恢复果实的几何结构,但其计算开销大,耗时过长,难以满足在采摘机器人这类资源受限的边缘设备的实时作业需求。

相比之下,关键点检测方法能够以较低的计算成本获取果实三维姿态,因而近年来受到越来越多学者关注。文献[13]通过改进 YOLO v8-GP 模型实现了葡萄串目标及其采摘关键点的同步检测。文献[14]提出了一种两阶段架构的深度学习网络(TPMv2),可在遮挡与劣质点云条件下同时预测番茄串的二维与三维关键点。文献[15]提出了一种创新方法,通过将二维关键点检测结果投影到点云获取了串番茄的三维姿态,在真实温室中实现了毫米级的误差估计。

尽管上述方法取得了一定进展,但普遍采用“2D 检测 + 深度投影”的两阶段策略,存在明显的局限性:①深度依赖性强:现有商业深度相机在温室强光或复杂遮挡下,常出现深度图缺失或噪点,导致细小果梗处的点云稀疏,直接影响投影精度。②缺乏三维约束:训练阶段仅利用 2D 监督信息,模型未能学习果实固有的三维几何结构,限制了其在遮挡下的推断能力。因此,研究一种不依赖实时深度信息、能直接从单目图像感知三维结构的方法,对于提升采摘机器人的环境适应性具有重要意义。

针对上述问题,本研究以 RTMPose3D 模型^[16]为基础提出改进的串番茄三维姿态关键点估计模型 TomatoPose3D,训练阶段通过融合 RGB 图像与三维真值关键点信息引入真实串番茄三维结构约束,抑制不合理预测结果的出现以提高关键点定位精度;推理时,模型由单幅 RGB 图像端到端输出串番茄三维关键点坐标,无需通过点云投影来补全深度信息,避免因温室环境中常见的深度缺失、点云稀疏等现

象带来的投影失败问题。

1 材料与方法

1.1 样本数据采集

番茄图像于 2025 年 4—5 月在安徽桐城现代智慧农业产业园番茄种植温室采集。该温室采取标准化种植流程,在实施采摘作业前已经去除了位于下层的成熟串番茄果实周围的叶片(打叶),以降低病虫害风险以及减少植株不必要的营养损耗。在图像采集过程中出于实验考虑没有对温室场景进行任何修改,保障了数据集的真实性。

使用自行设计的串番茄采摘机器人平台进行数据采集,如图 1 所示,机械臂安装在移动底盘上,相机固定于机械臂末端。相机型号为 Intel RealSense D405,可采集 RGB 图及点云图像,两者分辨率均为 1 280 像素 × 720 像素。本文从采摘机器人作业需求出发,在距离串果实 30 ~ 40 cm 处进行图像获取,仅采集符合采摘条件的串番茄果实对象。通过调节相机姿态,系统可采集到不同生长姿态和光照条件下的串番茄图像。筛选后的数据集包括 1 296 幅 RGB 图像及对应的点云图像,其中 1 064 幅用于模型训练,232 幅用于测试。串番茄样本分布如表 1 所示。图 2 展示了不同生长姿态与光照条件下的串番茄样本。数据集样本的多样性可确保本文模型在非结构化温室场景中的适应性。



图 1 串番茄温室种植环境

Fig. 1 Tomato greenhouse planting environment

表 1 串番茄样本分布

Tab. 1 Distribution of tomato samples

光照	正向数量(百分比/%)	侧向数量(百分比/%)	合计
顺光	206(16)	271(21)	477
逆光	159(12)	237(18)	396
侧光	218(17)	205(16)	423
总计	583(45)	713(55)	1 296

为了更系统地表达数据集中的遮挡特征,并分析 TomatoPose3D 对不同遮挡情况下串番茄对象的适应性,本文依据串番茄关键部位的可视区域比例对样本进行了遮挡等级划分。如图 3 所示,轻度遮挡样本中仅有少量果梗或果实出现局部遮挡;中度

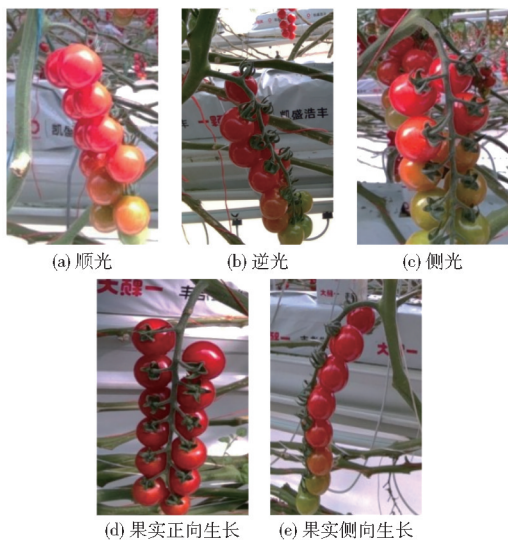


图2 温室场景番茄图像

Fig. 2 Tomato images in greenhouse environment

遮挡样本中主茎或果梗结构存在较明显的重叠与遮挡;重度遮挡样本中相关区域大范围遮挡或重叠,结构信息缺失明显。



图3 不同遮挡程度的串番茄对象

Fig. 3 Tomato objects with different degrees of occlusion

1.2 番茄三维关键点定义与数据集制作

数据集使用 Labelme 工具进行图像标注,标注文件格式为 JavaScript Object Notation (JSON)。为了给采摘机器人提供完整的采摘场景信息,避免采摘过程中末端执行器与主茎或者果实发生碰撞造成损失,本文在串番茄的主茎、果梗与果实上定义了7个关键点(如图4所示)用于描述串番茄的空间姿态。关键点 m 位于果梗中点,被定义为机械臂执行剪切作业的理想作业点,其三维坐标即为采摘控制的目的



图4 关键点定义

Fig. 4 Keypoints definition

标位置。 S_u 与 S_d 分别表示主茎的上、下节点,两者构成的主茎向量 V_{stem} 为

$$V_{stem} = S_u - S_d \quad (1)$$

用于描述主茎的空间生长方向;关键点 n 为主茎与果梗连接点,与关键点 m 构成的向量 $V_{peduncle}$ 为

$$V_{peduncle} = m - n \quad (2)$$

反映了果梗的生长姿态。基于 V_{stem} 与 $V_{peduncle}$ 可确定采摘切割平面 ρ , 公式为

$$\rho = \text{span}(V_{stem}, V_{peduncle}) \quad (3)$$

计算其法向量 N_ρ 为

$$N_\rho = V_{stem} \times V_{peduncle} \quad (4)$$

考虑到关键点 m 是末端执行器执行剪切动作的目标位置,如图5所示,本研究在文献[17]提出的最优采摘姿态平面 (Harvesting optimal posture plane, HOPP) 的基础上,引入“姿态平面需通过点 m ”的几何约束,将模型输出的三维关键点信息映射为采摘作业需要的六自由度位姿指令。具体而言,目标采摘位置由果梗关键点 m 的三维预测坐标确定,末端执行器的进给方向被约束为沿切割平面 ρ 的法向量 N_ρ ,为了约束绕进给方向的自由旋转以降低末端执行器与主茎或果梗干涉的风险,旋转方向被约束为与主茎向量 V_{stem} 平行。通过最优采摘姿态平面的约束,数据集定义的串番茄关键点三维信息被映射为机械臂末端执行器的六自由度空间位姿,为实际采摘作业提供了稳定、无碰撞的采摘姿态参考。

为了提高网络对番茄果梗三维关键点的估计精度,本文构建了包含真值深度信息的三维关键点数据集。将标注得到的二维关键点坐标映射至配准后的 RGB-D 点云空间,以获取其对应的三维坐标。然而使用商业相机获取的现现场景点云通常呈现稀

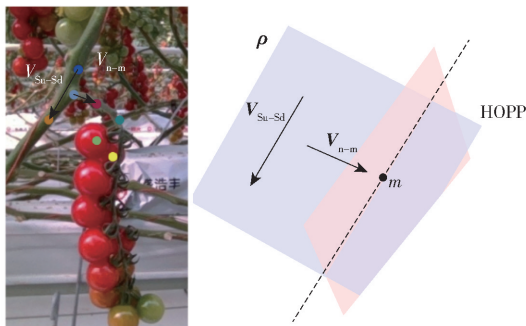


图5 最优采摘姿态平面

Fig. 5 Harvesting optimal posture plane

疏分布,关键点投影可能因深度缺失而失败,参考文献[15]的方法,本文使用关键点邻域深度补全策略生成真值三维关键点,以标注真值二维关键点为中心,在其局部圆形邻域内统计可用深度,其中圆形邻域半径 r 为

$$r = w \sqrt{W^2 + H^2} \quad (5)$$

式中 w ——权重因子

W ——图6中关键点范围矩形框的宽度

H ——矩形框高度

权重因子 w 经实验整定为 0.02,该阈值覆盖了约为果梗直径 1.5 倍的区域,既能保证获取足够的有效深度,又避免了背景噪声的引入。

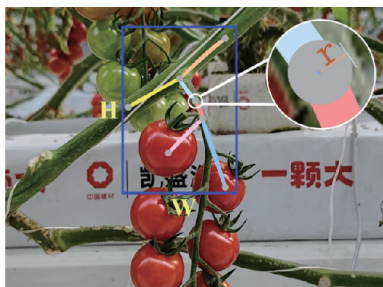


图6 拓展圆形掩码

Fig. 6 Expanded circular mask

2 TomatoPose3D 三维关键点检测模型

RTMPose3D 是由 OpenMMLab 推出的实时高精度三维人体姿态估计项目,在 Backbone 部分采取 CSPNet^[18] (Cross stage partial networks) 架构,通过多层级残差块提取不同层级特征图增强模型感受野,并使用深度可分离卷积 (Depthwise separable convolution)^[19] 降低计算复杂度。Neck 部分使用改进双向特征金字塔结构 (Bidirectional feature pyramid network, BiFPN) 融合多尺度特征,增强遮挡条件下的空间感知能力。Head 部分采用 Simcc (Simulated classification coordinate)^[20] 解码器,将三维关键点回归问题分解为各坐标轴上的区间分类问题,并结合骨骼几何约束损失,实现高精度且快速的

三维关键点预测。

当尝试将 RTMPose3D 模型用于番茄果实三维点检测时,面临温室场景内枝叶遮挡、果实表面高光反射等诸多挑战,现有骨干网络主要关注稀疏关键点的局部纹理信息,难以捕捉串番茄特有的几何结构关系;再加上关键点数量有限、外观相似度高,模型在光照变化或遮挡条件下的判别能力容易下降,从而导致三维定位精度降低。

为此本文在 RTMPose3D 基础上进行两方面改进:在骨干网络引入全局结构感知 MobileViT Block 模块^[21] 提升模型长距离依赖建模能力,同时依靠轻量化设计保持较低计算成本与较高推理速度;对 Simcc 解码头的坐标解码机制进行优化,提高关键点定位精度与稳定性。

2.1 MobileViT Block 模块

传统卷积神经网络 (Convolutional neural network, CNN) 凭借其局部感受野与共享权重机制,能有效捕捉到图像的局部细节特征,然而其有限的感受野导致 CNN 很难建立长距离依赖来捕获特征全局空间的结构联系。相比之下,Transformer^[22] 架构则借助自注意力机制实现全局建模,但其高计算复杂度和对大规模数据的依赖限制了其在边缘设备上的应用。串番茄具有极高的外观相似性,且常发生重叠。针对上述问题 MobileViT Block 将 CNN 与 Transformer 相结合,在保持轻量化的同时兼顾局部细节提取与全局建模能力,该模块首先利用深度可分离卷积处理局部细节特征,再利用标准 Transformer 模块捕获全局依赖关系,在保证计算效率的同时提升模型全局表征能力,有助于模型在严重遮挡下利用上下文信息推断被遮挡关键点的位置,保持整串果实结构的拓扑一致性。

如图 7 所示,MobileViT Block 首先利用一个深度可分离卷积对输入特征图进行局部特征提取。该操作通过将标准卷积分解为逐通道的深度卷积与逐点卷积两个阶段,有效捕捉空间上下文信息的同时显著降低了计算复杂度与参数量。其次,通过一个 1×1 的卷积层对通道维度进行压缩,将特征图从原始通道数 C 映射至一个较小的维度 d ,以适配后续的 Transformer 模块。通道压缩后的特征图被划分为 N 个非重叠的二维特征块,并展平成序列形式,输入至标准的 Transformer 编码器中,以捕获全局依赖。Transformer 输出的序列随后被重构回空间特征图结构,并通过 1×1 卷积层恢复通道数。然后,为了融合局部与全局语义信息,该模块采用跳跃连接的方式将重构特征图与初始局部特征图进行通道拼接。最终,拼接后的特征图经过一个 3×3 卷积融合

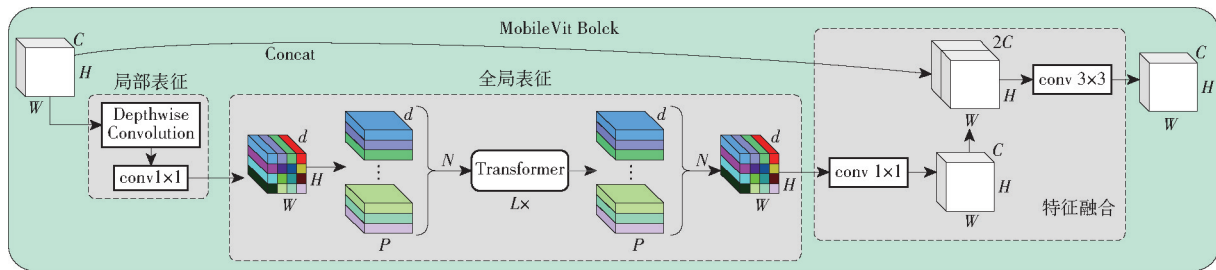


图 7 MobileVit Block 结构图

Fig. 7 MobileVit Block structure diagram

信息并恢复原始空间分辨率,形成最终的模块输出。

2.2 改进 Simcc head

当前三维关键点估计方法主要分为二维坐标提升法与直接回归法两类。前者首先在 RGB 图像上预测二维关键点位置,再通过轻量化回归网络将其映射至三维空间。计算成本低,适合部署于资源受限的边缘设备。然而,该方法在三维回归阶段完全舍弃了图像的纹理与上下文信息,性能对训练数据分布高度敏感,在未见场景下的泛化能力有限,容易出现明显的性能衰减。相比之下,直接回归法通过深度神经网络直接从原始图像中预测三维关键点,充分利用图像特征与结构特征,通常具有更高精度,但模型参数量大、计算开销高,难以满足采摘机器人等资源受限场景对实时性的要求。

RTMPose3D 受 Simcc 方法的启发,将连续三维坐标进行离散化处理,将三维坐标回归问题转化为在各个坐标轴上的分类问题,有效降低回归难度与计算复杂度。推理阶段通过解码模块将分类结果映

射回连续坐标空间,生成最终的三维关键点坐标。该策略在保持高精度的同时兼顾了推理效率与部署友好性,适用于边缘计算平台下的实时三维感知任务。

Simcc head 是 RTMPose3D 的核心结构。如图 8、9 所示,它同时接收来自骨干网络的底层空间细节特征图与高层上下文信息特征图。高层特征图首先经过 3×3 卷积并进行特征展平投影至目标维度,同时高层特征图还经过上采样与 1×1 卷积得到增强的高层特征,增强后的高层特征与底层特征拼接后再经过 1×1 卷积实现特征融合,融合特征图同样经特征展平后投影到统一维度。两分支输出在通道维上拼接得到融合局部细节特征与全局语义特征的融合表征,融合表征随后输入门控注意力单元 (Gated attention unit, GAU)^[23] 进行全局上下文融合。GAU 输出通过 Simcc Regression 回归,经过 3 条并行的独立线性层分别生成 X、Y、Z 轴上的一维离散概率分布,最后对坐标离散概率分布进行解码得到预测连续坐标。

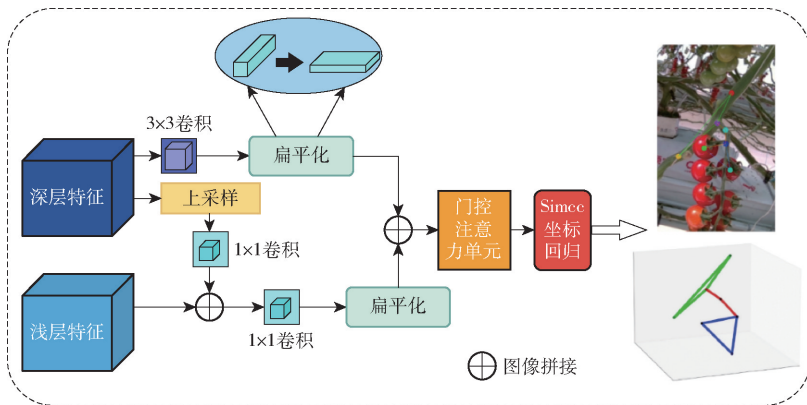


图 8 Simcc head 结构图

Fig. 8 Simcc head structure diagram

传统的 Simcc head 采用 Argmax 函数进行坐标解码,其精度受限于特征图分辨率,存在固有的量化误差。本文引入 DARK (Distribution-aware coordinate representation of keypoints) 解码策略^[24]代替原始 Argmax 硬解码,通过高斯核平滑与泰勒展开进行亚像素级峰值细化,有效克服了离散化带来的精度损失,在不显著增加计算量的前提下提升了微小目标 (如果梗节点) 的定位准确度。

假设离散概率分布区间为 $[x_{-r}, \dots, x_i, \dots, x_r]$, $-r \leq i \leq r$, 高斯核宽度为 σ , 则高斯核为

$$G(x_i - x_j; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right) \quad (6)$$

得到 x_i 处平滑分布概率, 即

$$p'(x_i) = \sum_{j=1}^N p(x_j) G(x_i - x_j; \sigma) \quad (7)$$

其中 $p(x_j)$ 为原始概率分布在 x_j 的值。

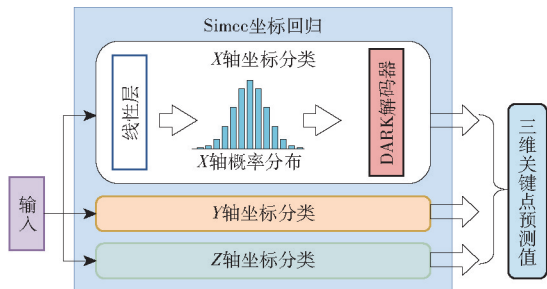


图9 Simcc regress 结构图

Fig.9 Simcc regress structure diagram

假设平滑后的概率峰值位置为 x_k , 为了提升数据稳定性, 对峰值邻域 $(x_{k-1}, p(x_{k-1}))$, $(x_k, p(x_k))$, $(x_{k+1}, p(x_{k+1}))$ 3 点进行抛物线拟合, 取偏移量 δ 为

$$\delta = \frac{p'(x_{k-1}) - p'(x_{k+1})}{2(p'(x_{k-1}) - 2p'(x_k) + p'(x_{k+1}))} \quad (8)$$

则最终连续位置坐标 \hat{x} 为

$$\hat{x} = p'(x_i) + \delta \quad (9)$$

DARK 解码通过在离散概率分布上施加高斯平滑, 有效抑制噪声与相似峰干扰, 并在峰值邻域内进行二次曲线拟合, 从而获得亚像素级的连续坐标估计。相较于 Argmax 硬解码直接选取单一峰值位置, DARK 能在多峰相似或分布平缓的情况下更准确地恢复关键点位置。本文在 RTMPose3D 中引入 DARK 解码函数, 以替代原有的 Argmax 硬解码, 在不显著增加模型计算量的前提下提升三维关键点预测精度与稳定性。

2.3 TomatoPose3D 模型

温室中串番茄复杂的生长环境对自动采摘机器人视觉检测系统的实时性和准确性提出了更高的要求。因此本研究基于 RTMPose3D 改进特征提取网络与回归解码模块, 提出一种轻量化串番茄三维关键点估计网络 TomatoPose3D (图 10)。在骨干网络部分将原始 CspLayer 层中卷积模块替换成具有全局结构感知能力的 MobileVit Block 以增强模型全局依赖建模能力。同时将 Simcc 解码流程中的 Argmax 模块替换为 DARK 解码器。DARK 解码通过高斯平滑与二次曲线拟合实现亚像素级的连续坐标估计, 能够在多峰或平缓分布的情况下提升坐标预测精度。该改进在不增加显著计算开销的前提下, 有效提高了 TomatoPose3D 的三维关键点预测稳定性与精度。

3 实验与结果分析

3.1 实验平台

实验硬件配置为 GeForce RTX 4090D 24G GPU 显卡、Intel(R) Xeon(R) Platinum 8474C CPU 处理

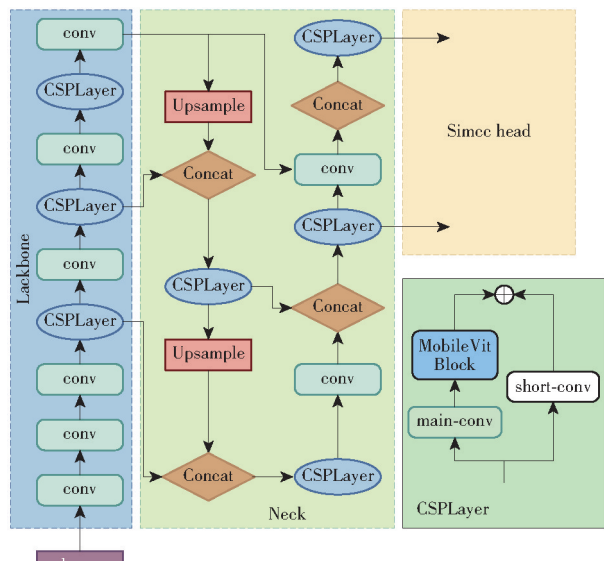


图10 TomatoPose3D 结构图

Fig.10 TomatoPose3D structural diagram

器, 运行内存容量为 80 GB。使用 Ubuntu 20.04 操作系统, 在 Python 3.8 以及 Cuda 11.3 环境下基于 Pytorch 深度学习框架搭建模型进行训练。

3.2 评估指标

为了评估模型性能, 在测试集上采用平均关节位置误差、正确关键点平均百分比 (PCK@0.05)、推理速度和浮点运算量作为评估指标。

PCK@0.05 表示模型在给定阈值为 0.05 时的预测正确关键点百分比。使用浮点运算量和推测速度来评估模型实时推理能力与计算开销。

3.3 消融实验

以 RTMPose3D 为基准模型进行消融实验以分析模型的性能, 并评估每个组件的贡献和功能。在基准模型 RTMPose3D 上逐步引入基于 Transformer 的全局特征提取模块 MobileVit Block 与改进 DARK 解码算法。

消融实验结果如表 2 所示, 首先使用 MobileVit Block 替换原模型特征卷积模块, 结果表明平均关节位置误差从 4.489 mm 降低到 3.711 mm, PCK@0.05 从 81.55% 提升至 84.93%, 模型在关键点定位精度上取得显著提升, 且推理速度未出现明显下降。这主要归功于全局注意力机制增强了模型对串番茄整体骨架的感知能力, 减少了因局部特征混淆导致的误检。

随后将 Argmax 解码替换为 DARK 解码策略, 实验结果显示, 该改进使得平均关节位置误差降低 0.481 mm, 并将 PCK@0.05 从 81.55% 提高到 82.59%, 验证了 DARK 解码进一步修正了回归坐标的微小偏移, 证明了在精细农业操作中亚像素级定位的必要性。进一步, 同时引入 MobileVit Block

表 2 消融实验对比结果

Tab. 2 Comparison results of ablation experiments

模型	平均关节位置误差/mm	PCK@0.05/%	推理速度/(f·s ⁻¹)	浮点运算量
RTMPose3D	4.489	81.55	86	1.771 × 10 ¹⁰
RTMPose3D + MobileVit Block	3.711	84.93	82	1.908 × 10 ¹⁰
RTMPose3D + DARK	4.305	82.59	86	1.771 × 10 ¹⁰
RTMPose3D + MobileVit Block + DARK	3.558	86.73	81	1.908 × 10 ¹⁰

模块与 DARK 解码函数,充分结合前者在特征表达与推理效率上的优势,以及后者在关键点定位精度上的提升能力,两者的优势相互叠加,使得各项性能指标均优于单一改进方案。

3.4 解码策略对比实验

为了进一步验证本文所采用的 DARK 解码策略在精度与实时度方面的优势,在消融实验的基础上开展了专项对比实验。实验选取 3 种典型的坐标解码方法进行比较:Standard Argmax、Soft - Argmax 以及本文采用的 DARK 解码。Standard Argmax 通过检索一维分类向量的极值索引获得坐标,计算高效但受限于特征离散性而存在固有量化误差;Soft - Argmax 通过 Softmax 归一化后计算概率分布的期望值以获得连续坐标,能够在一定程度上缓解量化偏差;DARK 解码通过高斯平滑预处理与泰勒展开计算对分布峰值进行局部修正,从而提升在噪声背景下的定位稳定性。

在保持 RTMPose3D 模型权重不变的条件下,将 3 种解码策略分别应用于测试集评估,各策略精度指标(平均关节位置误差、PCK@0.05)与时间性能指标(解码耗时)如表 3 所示。

表 3 解码策略对比实验结果

Tab. 3 Experimental results of decoding strategy comparison

解码策略	平均关节位置 误差/mm	PCK@0.05/ %	解码耗时/ ms
Standard Argmax	4.711	84.93	0.08
Soft - Argmax	4.543	86.15	0.42
DARK	4.201	87.73	0.25

结果显示,与 Standard Argmax 相比,Soft - Argmax 与 DARK 两种策略均在关键点定位精度上有所提升。其中,Soft - Argmax 通过全局期望计算有效减小了离散分布固有量化误差,但易受农业复杂场景下的多峰值热力图影响从而导致预测位置偏移。相比之下,DARK 通过对峰值邻域进行高斯平滑和局部拟合处理,更有效地抑制了背景噪声对峰值定位的干扰,其在平均关节位置误差上优于 Soft - Argmax 解码 0.342 mm,在 PCK@0.05 上也提升了 1.58 个百分点。从推理性能来看,DARK 的解码速

度(0.25 ms)略低于 Standard Argmax(0.08 ms),但明显快于 Soft - Argmax(0.42 ms)。鉴于 DARK 仅在解码阶段引入较小的计算代价,却带来了明显的定位精度提升,该方法在精度与推理速度之间取得了更加均衡的表现。综合分析可知,DARK 解码更适合作为采摘机器人三维视觉感知任务的关键点解码策略,尤其在复杂的温室环境中表现出更高的稳定性和可靠性。

3.5 对比实验

为了验证 TomatoPose3D 在实际检测任务中的表现,本文选取了文献[15,25]的串番茄三维关键点估计方法以及 RTMPose3D 作为对比模型。上述方法分别代表了当前三维关键点估计的三类典型技术路线。SimpleBaseline3D^[25]属于典型的二维提升三维方法,它接收独立二维关键点检测网络预测得到的二维关键点坐标,并通过轻量回归网络回归得到三维信息,结构简单、计算代价极低。文献[15]的方法则结合了二维关键点和点云深度信息,先通过 Keypoint R - CNN^[26]检测得到串番茄的二维关键点,再利用深度图和点云投影得到对应的三维位置。与前两类方法不同,RTMPose3D 以及本文提出的 TomatoPose3D 均从单幅 RGB 图像中直接回归三维关键点坐标。本文分别采用上述 4 种方法针对串番茄测试样本数据集(共 232 幅)开展验证实验,结果如表 4 所示。

表 4 不同模型的对比实验结果

Tab. 4 Comparative experimental results of different models

模型	平均关节位 置误差/mm	PCK@0.05/ %	推理速度/ (f·s ⁻¹)	浮点运算 量
RTMPose3D	4.48	81.55	86	1.771 × 10 ¹⁰
Keypoint R - CNN	3.12	89.43	105	9.674 × 10 ⁹
SimpleBaseline3D	6.36	76.75	147	8.230 × 10 ⁸
TomatoPose3D	3.55	86.73	81	1.608 × 10 ¹⁰

文献[15]的方法依赖 RGB - D 相机提供的点云来恢复三维关键点,在部分样本中会因点云缺失而无法完成投影,本文在评估该方法时仅在成功获取三维关键点的样本上计算平均关节位置误差。Keypoint R - CNN 在点云数据良好的样本上取得了

最优的平均关节位置误差 (3.12 mm) 和最高的 PCK@0.05 (89.43%), 说明在深度信息充分、点云质量良好的情况下依赖 RGB-D 的关键点回归方式能提供较高的三维精度。但该方法在部分样本中会因点云缺失或噪声而无法输出三维关键点, 在采摘场景中, 这类情况会直接影响采摘机器人对串番茄位置与姿态的判断从而降低整体采摘成功率。

其余 3 种方法在所有样本上均能稳定给出预测结果, 其中 TomatoPose3D 整体平均关节位置误差为 3.55 mm、PCK@0.05 达到 86.73%, 在不依赖点云数据的前提下取得了接近 Keypoint R-CNN 的精度, RTMPose3D 的平均关节位置误差为 4.48 mm, PCK@0.05 为 81.55%, 精度略低于 TomatoPose3D, 这表明通用人体 3D 回归框架在串番茄结构上存在一定适应性限制。SimpleBaseline3D 由于网络结构轻量, 在浮点运算量和推理速度上表现出显著优势, 是所有方法中计算效率最高的。然而, 其仅利用稀

疏的 2D 关键点坐标进行 3D 提升, 丢失了图像中的纹理与光照等视觉上下文信息, 在温室非结构化环境中难以处理复杂的姿态变化, 精度上受到一定限制, 平均关节位置误差和 PCK@0.05 均为 4 种方法中最低。

考虑到自动采摘机器人在执行采摘作业对关键点定位的精确性要求远高于推理速度与计算复杂度, 以避免采摘过程中对果实造成碰撞损伤, 保障果实品质, TomatoPose3D 在保持较高精度的同时仍然具备合理的实时性与计算开销, 体现了在精度、推理速度与计算复杂度之间的良好平衡, 上述结果表明该方法能在满足现实农业采摘机器人实时性要求的前提下, 提供稳定准确的串番茄三维关键点信息。

为了展示不同方法在复杂温室场景下的关键点预测差异, 本文从 232 幅测试集样本中按生长姿态与遮挡情况选取 5 个代表性样本进行可视化对比分析, 结果如图 11 所示。

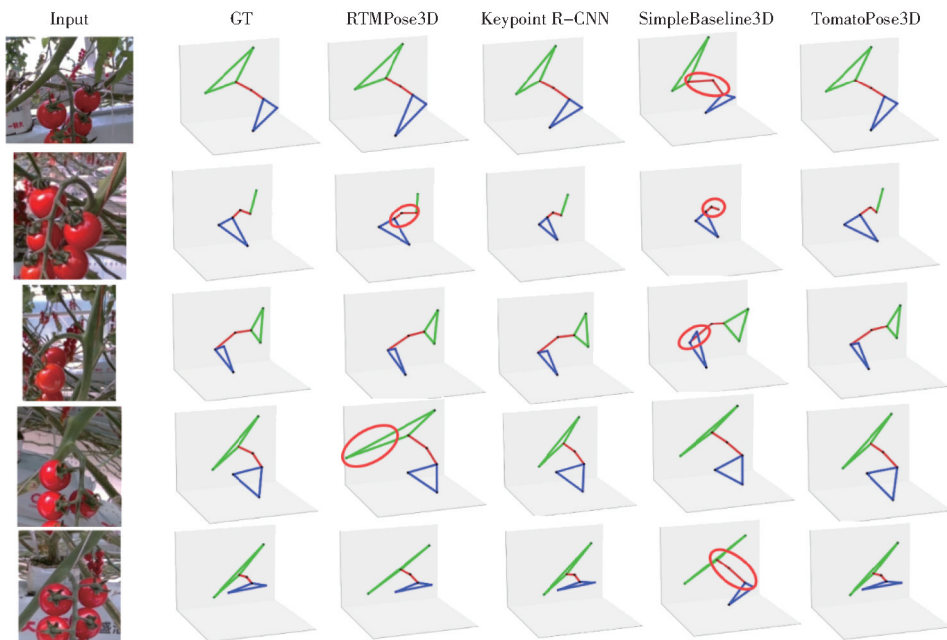


图 11 温室场景代表案例对比图

Fig. 11 Comparison diagrams of representative cases in greenhouse scenarios

在 5 个样本对比实验中, 对比模型出现了不同程度的误检或漏检, 相关位置在图 11 中以红圈标注。样本 1 中目标串番茄正向生长, SimpleBaseline3D 在果梗位置出现误检测, 样本 2 实验中, 选取的目标番茄主茎被果实部分遮挡, RTMPose3D 在主茎处出现误检测, 而 SimpleBaseline3D 未能检测到位于主茎上的关键点。样本 3 实验选取果实侧向生长的番茄对象, 其中果实关键点被部分遮挡, SimpleBaseline3D 对于被遮挡的果实关键点出现误检测。样本 4 实验选取果梗与背景主茎出现重叠的串番茄对象, RTMPose3D 对

主茎下侧关键点的估计出现明显偏移。这些结果表明, 在存在遮挡、重叠或形态特殊的复杂场景下, TomatoPose3D 在三维关键点定位的稳定性与精度方面表现更具优势。

4 模型部署与推理性能评估

为了验证 TomatoPose3D 模型在边缘计算平台中的实时运行能力, 本研究进行了模型部署实验, 模型部署在基于 NVIDIA Jetson Orin 的嵌入式工控机平台并利用 TensorRT 构建 FP16 推理引擎。部署实验旨在衡量模型的实际可用性, 包括 GPU 稳态推理

性能测试以及在真实推理流程中的推理速度测试。结果如表 5 所示,在 GPU 稳态推理情况下, TomatoPose3D 的 FP16 引擎推理速度为 78 f/s,表明模型在工控机平台上具有较高的推理效率。该结果展示了 Tomatopose3D 的良好可加速性,能够充分发挥边缘 GPU 的推理能力,为实时三维姿态估计任务提供了良好的性能保障。

表 5 模型部署实验结果

Tab. 5 Model deployment experiment results

推理方式	平均帧率/ ($f \cdot s^{-1}$)	平均单帧耗时/ ms	是否包含预 处理与后处理
稳态推理	78	12.8	否
端到端推理	37	26.0	是

为了进一步评估模型在实际部署场景中的表现,构建了包含图像读取、预处理、模型推理以及解码函数模块的端到端推理流程,并在 1 000 幅测试图像上进行推理时间统计。实验结果显示,实际工作场景端到端推理速度为 37 f/s,单幅图像平均推理耗时为 26.0 ms。虽然低于稳态推理速度,但仍远高于一般采摘机械臂的控制周期(通常小于 10 Hz),完全满足闭环控制的实时性要求。

综合上述结果可知, TomatoPose3D 在资源受限的边缘平台上仍能实现较高的推理效率,具备良好的轻量化特性与工程可部署性。模型在 GPU 满载条件下具备较高的吞吐能力,而端到端场景下的平均推理速度也能够满足实际采摘任务的实时性需求,为其在温室农业场景中的实际应用奠定了基础。

5 结论

(1)本研究提出了一种面向温室环境的轻量化串番茄三维关键点估计方法 TomatoPose3D,能够从单视角 RGB 图像直接回归预测串番茄的三维关键点位置与空间姿态,为自动采摘机器人提供精确的果实空间位姿信息。该方法以 RTMPose3D 为基准模型,结合全局结构感知 MobileVit Block 模块与亚像素级 DARK 解码策略,有效解决了遮挡与相似目标干扰下的定位难题,平均关节位置误差缩小为 3.55 mm。

(2)在生产型温室串番茄数据集上的实验结果表明, TomatoPose3D 在单目三维关键点估计任务中取得了优于 RTMPose3D 与 SimpleBaseline3D 的性能表现, PCK@0.05 达到 86.73%,平均关节位置误差分别降低 0.93 mm 与 2.81 mm;并在无深度信息的条件下取得与基于 RGB-D 的关键点回归方法相当的精度。考虑到实际温室场景中点云质量常受光照反射、果实表面纹理及相机位置等因素影响而表现不稳定,本文方法在无需依赖深度数据的前提下表现出更高的稳定性与适应性。

(3)在边缘端上进行 TomatoPose3D 模型部署实验,利用 TensorRT 构建轻量化推理引擎。结果显示模型在 GPU 稳态推理条件下推理速度可达到 78 f/s,端到端运行速度为 37 f/s,具备实时性与资源受限环境下的部署能力。综上所述, TomatoPose3D 在精度、速度与模型复杂度之间实现了良好平衡,为智能采摘机器人的视觉系统提供了高可靠性的解决方案。

参 考 文 献

- [1] 赵春江,范贝贝,李瑾,等. 农业机器人技术进展、挑战与趋势[J]. 智慧农业(中英文), 2023, 5(4): 1-15. ZHAO Chunjiang, FAN Beibei, LI Jin, et al. Agricultural robots: technology progress, challenges and trends[J]. Smart Agriculture, 2023, 5(4): 1-15. (in Chinese)
- [2] 于丰华,周传琦,杨鑫,等. 日光温室番茄采摘机器人设计与试验[J]. 农业机械学报, 2022, 53(1): 41-49. YU Fenghua, ZHOU Chuanqi, YANG Xin, et al. Design and experiment of tomato picking robot in solar greenhouse[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(1): 41-49. (in Chinese)
- [3] 刘建航,何鉴恒,陈海华,等. 基于改进 YOLO v4 和 ICNet 的番茄串检测模型[J]. 农业机械学报, 2023, 54(10): 216-224, 254. LIU Jianhang, HE Jianheng, CHEN Haihua, et al. Development of detection model for tomato clusters based on improved YOLO v4 and ICNet[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10): 216-224, 254. (in Chinese)
- [4] 雷志龙,刘畅,王权. 设施番茄采摘机器人识别定位与采摘方法研究[J]. 农业机械学报, 2025, 56(7): 219-226. LEI Zhilong, LIU Chang, WANG Quan. Recognition-localization and picking methods of facility tomato picking robots[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(7): 219-226. (in Chinese)
- [5] 龙洁花,赵春江,林森,等. 改进 Mask R-CNN 的温室环境下不同成熟度番茄果实分割方法[J]. 农业工程学报, 2021, 37(18): 100-108. LONG Jiehua, ZHAO Chunjiang, LIN Sen, et al. Segmentation method of the tomato fruits with different maturities under greenhouse environment based on improved Mask R-CNN[J]. Transactions of the CSAE, 2021, 37(18): 100-108. (in Chinese)
- [6] GUO J, YANG Y, LIN X, et al. Revolutionizing agriculture: real-time ripe tomato detection with the enhanced tomato-YOLO v7 system[J]. IEEE Access, 2023, 11: 133086-133098.

- [7] 毕泽洋, 杨立伟, 吕树盛, 等. 基于 EDH - YOLO 的轻量型温室番茄检测方法[J]. 农业机械学报, 2024, 55(增刊2): 246 - 254.
BI Zeyang, YANG Liwei, LÜ Shusheng, et al. Lightweight greenhouse tomato detection method based on EDH - YOLO[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(Supp. 2): 246 - 254. (in Chinese)
- [8] GAO G, SHUAI C, WANG S, et al. Using improved YOLO V5s to recognize tomatoes in a continuous working environment [J]. Signal, Image and Video Processing, 2024, 18(5): 4019 - 4028.
- [9] ZHENG X, XINYI A I, QIN H, et al. Tomato - nerf: advancing tomato model reconstruction with improved neural radiance fields[J]. IEEE Access, 2024, 12: 184206 - 184215.
- [10] MAGALHÃES S A C, DOS SANTOS F N, MOREIRA A P, et al. MonoVisual3DFilter: 3D tomatoes' localisation with monocular cameras using histogram filters[J]. Robotica, 2024, 42(8): 2528 - 2547.
- [11] SMITT C, HALSTEAD M, ZIMMER P, et al. PAg - NeRF: towards fast and efficient end-to-end panoptic 3D representations for agricultural robotics[J]. IEEE Robotics and Automation Letters, 2023, 9(1): 907 - 914.
- [12] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99 - 106.
- [13] CHEN J, MA A, HUANG L, et al. Efficient and lightweight grape and picking point synchronous detection model based on key point detection[J]. Computers and Electronics in Agriculture, 2024, 217: 108612.
- [14] ZHANG F, GAO J, SONG C, et al. TPMv2: an end-to-end tomato pose method based on 3D key points detection[J]. Computers and Electronics in Agriculture, 2023, 210: 107878.
- [15] CI J, WANG X, RAPADO-RINCÓN D, et al. 3D pose estimation of tomato peduncle nodes using deep keypoint detection and point cloud[J]. Biosystems Engineering, 2024, 243: 57 - 69.
- [16] JIANG T, XIE X, LI Y. RTMW: real-time multi-person 2D and 3D whole-body pose estimation[J]. arXiv Preprint, arXiv: 2407.08634, 2024.
- [17] LI Y, FENG Q, ZHANG Y, et al. Peduncle collision-free grasping based on deep reinforcement learning for tomato harvesting robot[J]. Computers and Electronics in Agriculture, 2024, 216: 108488.
- [18] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390 - 391.
- [19] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251 - 1258.
- [20] LI Y, YANG S, LIU P, et al. Simcc: a simple coordinate classification perspective for human pose estimation[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 89 - 106.
- [21] MEHTA S, RASTEGARI M. MobileVit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv Preprint, arXiv:2110.02178, 2021.
- [22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998 - 6008.
- [23] HUA W, DAI Z, LIU H, et al. Transformer quality in linear time[C]// International Conference on Machine Learning. PMLR, 2022: 9099 - 9117.
- [24] ZHANG F, ZHU X, DAI H, et al. Distribution-aware coordinate representation for human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 7093 - 7102.
- [25] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2640 - 2649.
- [26] DING X, LI Q, CHENG Y, et al. Local keypoint-based Faster R - CNN[J]. Applied Intelligence, 2020, 50(10): 3007 - 3022.

(上接第 126 页)

- [22] ZHENG C, ZHU Q, XU W, et al. FAST-LIVO: fast and tightly-coupled sparse-direct lidar-inertial-visual odometry[C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022: 4003 - 4009.
- [23] ZHENG C, XU W, ZOU Z, et al. FAST-LIVO2: fast, direct lidar-inertial-visual odometry[J]. IEEE Transactions on Robotics, 2025, 41: 326 - 346.
- [24] ZERMAS D, IZZAT I, PAPANIKOLOPOULOS N. Fast segmentation of 3D point clouds: a paradigm on LiDAR data for autonomous vehicle applications[C]// 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017: 5067 - 5073.
- [25] 郑立华, 麦春艳, 廖岁, 等. 基于 Kinect 相机的苹果树三维点云配准[J]. 农业机械学报, 2016, 47(5): 9 - 14.
ZHENG Lihua, MAI Chunyan, LIAO Wei, et al. 3D point cloud registration of apple trees based on Kinect camera[J]. Transactions of the Chinese Society of Agricultural Machinery, 2016, 47(5): 9 - 14. (in Chinese)
- [26] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the KITTI vision benchmark suite[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354 - 3361.
- [27] ZHANG Y, SHI P, LI J. LiDAR-based place recognition for autonomous driving: a survey[J]. ACM Computing Surveys, 2024, 57(4): 1 - 36.