

基于空间门控与边缘感知的苹果枝条多模态分割方法

王 峥¹ 黄德钊² 程泽元² 宋卓颖² 许翔虎³ 杨福增²

(1. 西北农林科技大学林学院, 陕西杨凌 712100; 2. 西北农林科技大学机械与电子工程学院, 陕西杨凌 712100;
3. 哈尔滨工业大学机电工程学院, 哈尔滨 150001)

摘要: 苹果树枝条实例分割既是苹果采摘机器人实现机械臂采摘路径规划与无损避障采摘的重要前提, 又是信息采集机器人逐树逐枝感知果树表型信息的重要基础。本研究针对矮化密植苹果园非结构化环境下, 光照剧变及背景干扰导致的枝条分割边界模糊, 且末端细微枝条在深度学习网络的下采样过程中特征易丢失的难题, 提出一种面向果园的跨模态注意力与边缘感知的苹果树枝条实例分割模型 CAROL - YOLO (Cross-modal attention and refinement for orchard localization YOLO)。首先, 设计了一种可见光图像辅助的“特征仿射-非刚性变换”图像级联配准流程, 构建了宽波段近红外 (Visible - near infrared, Vis - NIR) 图像-深度图像配准的苹果树枝条多模态数据集; 在此基础上, 以 YOLO v8 为基础, 设计了伪孪生双流空间几何门控骨干网络, 通过该骨干网络, 设计了空间几何门控模块 (Spatial-geometry aware gating encoder, S - Gate) 提取深度图像中的前景枝条空间位置, 并自适应增强宽波段近红外图像中前景枝条的空间位置特征; 同时设计了一种边缘感知模块 (Edge-aware module, EAM), 显式增强细微枝条高频边界信息, 以增强深度学习网络下采样后果树枝条的分割精度。实验结果表明: CAROL - YOLO 模型在不同的参数量级 (Nano/Small/Large) 上均取得一致性的性能提升。以 CAROL - YOLO - s 为例, 模型 mAP50 - 95 达 (39.6 ± 0.1) %, 参数量仅为 2.03×10^7 ; 与基线模型 YOLO v8m - seg 相比, 在参数量减少 25.6% 的情况下, mAP50 - 95 仍提升了 1.4 个百分点; 与 YOLO 11m - seg 及 YOLO 26m - seg 模型相比, 在参数量分别减少 8.9% 和 24.5% 的情况下, mAP50 - 95 仍分别提升了 0.8 个百分点和 1.4 个百分点。该方法有效解决了果园非结构化复杂环境下的枝条漏检问题, 同一模型不同大小下的一致性能提升可为不同类型果园机器人精细化作业提供可靠的感知算法支撑。

关键词: 苹果枝条; 实例分割; 多模态融合; YOLO; 农业机器人

中图分类号: TP391.4; S24 文献标识码: A 文章编号: 1000-1298(2026)21-0310-14

OSID:



Multimodal Instance Segmentation Method for Apple Branches
Based on Spatial - Geometric Gating and Edge-awareness

Wang Zheng¹ Huang Dezha² Cheng Zeyuan² Song Zhuoying² Xu Xianghu³ Yang Fuzeng²

(1. College of Forestry, Northwest A&F University, Yangling, Shaanxi 712100, China

2. College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China

3. School of Mechatronics Engineering, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Instance segmentation of apple tree branches is essential for enabling robotic manipulators to plan harvesting paths and perform collision-free, damage-free picking. It also provides a fundamental basis for information-acquisition robots to perceive tree- and branch-level phenotypic traits. However, in unstructured high-density dwarf apple orchards, drastic illumination variations and background interference often result in blurred branch boundaries, while features of fine terminal branches are readily lost during downsampling in deep neural networks. To address these challenges, the cross-modal attention and refinement for orchard localization YOLO (CAROL - YOLO), a cross-modal attention- and edge-aware instance segmentation model was proposed for apple branches in orchard environments. Firstly, a visible-light-assisted cascaded image registration pipeline integrating feature-based affine and nonrigid

收稿日期: 2026 - 02 - 27 修回日期: 2026 - 05 - 22

基金项目: 陕西省“两链”融合重点专项(2023 - LLRH - 01)和国家重点研发计划项目(2023YFD1600800)

作者简介: 王峥(1972—), 女, 讲师, 主要从事农林智能装备研究, E-mail: wangzh925@163.com

通信作者: 杨福增(1966—), 男, 教授, 博士, 主要从事农业机器人关键技术和丘陵山地智能农机装备研究, E-mail: yfz0701@163.com

transformations was developed to construct a multimodal apple branch dataset comprising registered visible – near-infrared (Vis – NIR) and depth images. Building on YOLO v8, a pseudo-Siamese dual-stream backbone with spatial-geometry-aware gating was then designed. Within this backbone, a spatial-geometry-aware gating encoder (S – Gate) extracted the spatial locations of foreground branches from depth images and adaptively enhanced the corresponding spatial features in Vis – NIR images. In addition, an edge-aware module (EAM) was introduced to explicitly enhance the high-frequency boundary information of fine branches, thereby improving branch segmentation accuracy after network downsampling. Experimental results demonstrated that CAROL – YOLO consistently improved segmentation performance across different model scales (Nano, Small, and Large). In particular, CAROL – YOLO – s achieved an mAP50 – 95 of $(39.6 \pm 0.1)\%$ with only 2.03×10^7 parameters. Compared with the baseline YOLO v8m – seg, it improved mAP50 – 95 by 1.4 percentage points while reducing the number of parameters by 25.6%. Compared with YOLO 11m – seg and YOLO 26m – seg, it achieved improvements of 0.8 and 1.4 percentage points, respectively, while using 8.9% and 24.5% fewer parameters. The proposed method effectively mitigated missed detections of apple branches in complex, unstructured orchard environments. Its consistent performance gained across different model scales can provide reliable perceptual support for precision operations by various types of orchard robots.

Key words: apple tree branches; instance segmentation; multimodal fusion; YOLO; agricultural robotics

0 引言

我国是苹果生产大国,随着“精准农业”国家战略的深入推进,研发具备自主修剪、疏花及采摘能力的农业机器人已成为产业升级的核心需求。在矮砧密植的高纺锤形果园中,非结构化环境下的苹果树枝条实例分割是苹果采摘机器人实现机械臂采摘路径规划与无损避障采摘、信息采集机器人逐树逐枝感知果树表型信息的前提^[1-2]。然而,不同于常规果实检测,枝条具有极端的长宽比与复杂的拓扑结构,且往往被枝叶深度遮挡,光照条件较为复杂,这给农业机器人开展相关工作带来极大的困难。

早期的果园视觉感知研究多使用传统的图像处理技术,例如基于颜色空间的阈值分割和基于梯度的边缘检测,但这些方法在光照剧变和背景杂波严重的自然果园中鲁棒性较差。近年来,以深度学习为代表的驱动方法在农业目标检测与分割领域取得了突破性进展。Mask R – CNN、YOLO 系列等通用实例分割模型被广泛迁移应用于果园场景^[3-5]。

已有研究充分证明了实例分割在农业目标定位中的可行性^[6-7]。然而,这些研究大多针对番茄等具有较好几何包容性的离散块状目标,但在面对苹果树枝条这一特定对象时,通用视觉算法仍面临三大困难:①光照干扰与强伪装性。果园中树叶间隙的动态光斑与背光阴影极易破坏图像的纹理特征,且木质化枝条的颜色与树干、老叶高度相近,导致特征难以被卷积网络有效提取^[8-9]。②严重的遮挡与几何混叠。高密度种植模式下,枝条与枝条、枝条与

果实叶片之间存在复杂的遮挡关系^[10]。二维图像上的投影交织使得前景与背景枝条难以区分^[11],现有的分割模型往往难以推断被遮挡部分的连续性,导致分割出的枝条掩膜破碎、断裂,丧失了拓扑结构的完整性^[12-13]。Gené-Mola 等^[14]的研究指出,在严重遮挡下,常规模型无法像人类视觉那样进行非模态补全,严重影响了后续的尺寸估计和路径规划。③细微目标的特征丢失。枝条属于典型的细长结构。在深度卷积神经网络中,为了扩大感受野而进行的连续下采样操作(如步长卷积和池化)会不可避免地滤除高频空间信息^[15-16]。对于末端细小的枝条,其特征在深层特征图中往往会完全消失,导致重构出的掩膜边缘粗糙,甚至漏检整个细枝^[17-18]。

为了突破单模态感知的局限,多模态信息融合成为近年来的研究热点^[19-21]。尽管多模态视觉为果园环境感知提供了新思路,但在非结构化环境中直接应用现有融合策略仍面临 2 个难点:消费级深度相机在复杂光照与重度遮挡下极易产生伪影与数据空洞,常规的对称融合易导致深度噪声污染纹理特征;苹果树枝条具有极端的长宽比,深层网络的连续下采样极易使末端微细枝条的空间细节被平滑丢失。

为解决上述问题,本文提出一种面向果园主动感知的轻量级跨模态实例分割模型 CAROL – YOLO。引入空间几何门控模块 (Spatial geometry aware gating encoder, S – Gate),通过几何引导与残差机制,在增强宽波段近红外图像纹理特征的同时,有效阻断跨模态噪声污染,确保极端工况下的特征鲁棒性;引入边缘感知模块 (Edge-aware module,

EAM),采用频域代数差分策略,在低算力开销下,对于因下采样丢失的细微枝条高频特征进行显式锐化,有效增强边界特征。结合构建的多模态数据集,显著提升强伪装与复杂背景下的枝条实例分割精度,以期为果园自动化巡检与采摘提供高精度的视觉支撑。

1 数据集构建

1.1 图像采集

实验数据采集于陕西省延安市黄陵县苏家险村西北农林科技大学苹果试验站及陕西省延安市黄陵县古路村苹果试验站,研究对象为纺锤形栽培的富士苹果树,两处试验站果树均为3年树龄,果园行距3.5~4.0 m,株距0.9~1.2 m。图像采集设备为海康机器人MV-CH250-90GN型宽波段近红外工业面阵相机和Intel D435if型深度相机,其中宽波段近红外相机搭载12 mm焦距镜头。

针对苹果园非结构化环境的光照剧变与背景干扰,海康机器人MV-CH250-90GN型相机未配置窄波滤光片,直接采集波长400~1 100 nm的宽波段(Visible-near infrared, Vis-NIR)图像。宽波段成像利用植物特有的光谱响应差异,在波长700 nm后的近红外区域^[22],显著增强了暗部枝干与亮部叶片的对比度。同时,通过舍弃滤光片,保证成像时拥有充足的光通量,充分利用CMOS传感器在850 nm处的高量子效率。若采用如850 nm的窄带纯近红外方案,虽可滤除部分非目标波段,但会阻断可见光波段的反射信息,导致枝条表面及边缘的高频纹理细节大量衰减。由于本文模型中的EAM模块依赖此类纹理特征进行边界锐化,因此,宽波段无滤光片设计在算法层面,兼顾了近红外波段的背景对比度与可见光波段的表面纹理特征,为特征提取保留了丰富的高频信息;在硬件层面,增加了光学系统的总进光量,结合CMOS传感器的量子效率,使系统在弱光条件下能够以较短的曝光时间获取高信噪比图像,有效抑制了田间风扰动带来的运动模糊。

为模拟修剪机器人的实际作业视角,相机组安装于移动平台上,距离树干约1.5 m水平拍摄。数据采集覆盖了春季苹果树始花期与秋季落叶期,并在晴天(顺光/逆光)、阴天及傍晚等多种光照条件下获取原始图像551组,拍摄中相机均设置为自动曝光。其中Vis-NIR图像分辨率为5 120像素×5 120像素(bmp格式);深度及可见光图像为1 280像素×720像素(png格式,已使用Intel D435if型相机自带的函数完成配准),用于提供几何与色彩参考,图像采集系统构成如图1所示。

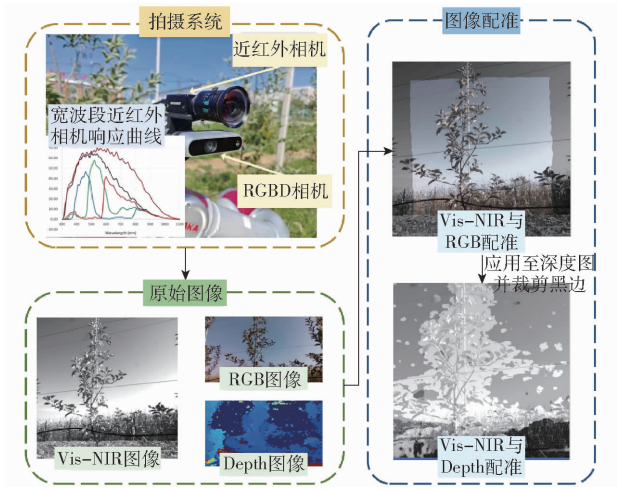


图1 图像采集系统构成

Fig.1 Camera components

1.2 数据预处理与图像配准

由于近红外相机与深度相机存在物理基线、视场角及分辨率的显著差异,且果树枝条在自然环境下受微风影响存在轻微晃动,传统的刚性变换无法实现异构图像的对齐。为此,本文利用SimpleITK开源图像处理库,设计了一种“可见光图像辅助的特征仿射-非刚性变换”级联配准流程,具体流程如图2所示。

1.2.1 可见光与Vis-NIR图像仿射粗配准

为增强本配准方法在不同硬件平台上的泛用性,在仿射粗配准阶段并未依赖相机内参及外参矩阵进行三维空间重投影,而是采用了一种基于二维图像间共有特征点映射的配准策略。

鉴于可见光与Vis-NIR图像均具备丰富的高频纹理特征,首先采用SIFT(Scale-invariant feature transform)方法提取两者在枝条分叉点及叶片边缘的共有特征点,两者成像效果如图3所示。为了抑制图像边缘背景杂波(如远处树木、地面杂草)的干扰,设定了一个以图像中心为焦点的椭圆形掩膜,强制程序仅在树干及主枝区域提取特征,并采用基于仿射变换的几何映射模型。该模型包含平移、旋转、缩放及剪切4类变换,通过最小化特征点对的欧氏距离误差,解算出全局最佳变换参数,将可见光图像初步变换至Vis-NIR视场坐标系下,消除主要的刚性视差。

1.2.2 基于B样条的非刚性精配准

针对枝条非共面及微风引起的局部非刚性形变,在粗配准的基础上引入自由变形模型。采用三次B样条作为变换模型,在图像域内设置均匀的控制点网格。配准过程以归一化互信息作为多模态相似性度量函数,并迭代调整控制点位置。该方法能够模拟图像局部的弹性形

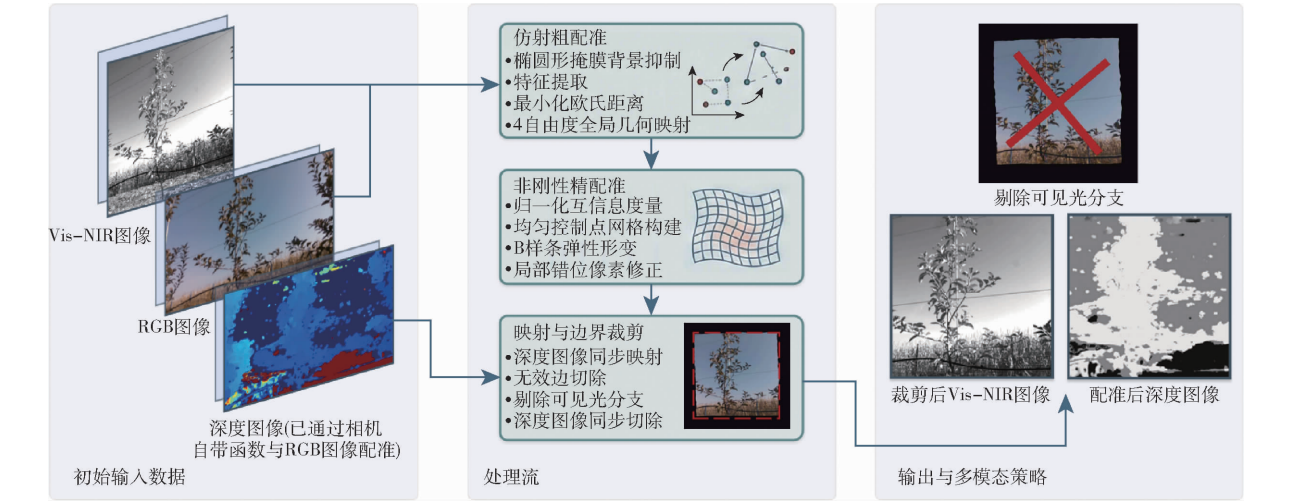


图 2 图像配准流程图

Fig. 2 Image registration process

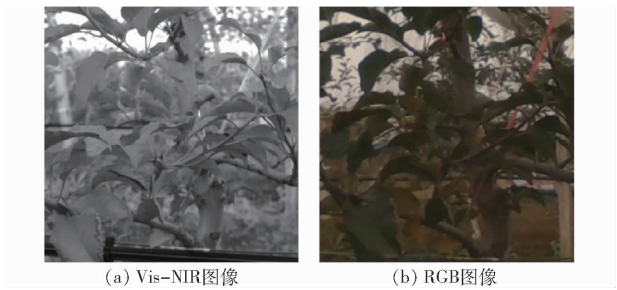


图 3 Vis - NIR 与 RGB 成像对比

变,有效修正了因深度差异和微动产生的局部像素错位。

1.2.3 深度空间映射与多模态输入策略

利用深度相机自带的可见光图像与深度图像配准功能,将经过非刚性校正后的深度图像空间最终投影至 Vis - NIR 图像。虽然可见光数据作为参考源参与了中间配准,但考虑到 Vis - NIR 成像涵盖波段 400 ~ 1 100 nm,兼具高频纹理与高感光特性,信噪比显著优于易受动态光斑干扰的可见光图像,剔除冗余的可见光分支可以避免色度噪声干扰,提升模型的分割效果,更契合农业机器人复杂环境下作业需求,故本文网络训练与推理阶段仅输入对齐后的 Vis - NIR 和深度图像,提高系统全天候的环境适应性。

为客观评价上述“特征仿射-非刚性变换”级联配准流程在复杂果园工况下的鲁棒性,将验证样本按干扰类型划分为常规、强光与阴影(逆光或光线不足)、重度遮挡 3 个子集,每组随机选择 30 组图像进行定量分析,共计 90 组图像。需要特别说明的是,为了保证模态间配准的准确度,本研究数据集均 在无风或微风时采集,故不涉及强风的特殊工况。

由于模态间存在显著的纹理差异且深度图像质量较低,采用人工提取 Vis - NIR 图像和 RGB 图像

的枝条交叉点与显著边缘特征的方式,计算配准后图像对的特征映射误差。结果表明,在原始分辨率下,常规、强光与阴影、重度遮挡组的均方根误差(RMSE)为(17.49 ± 2.535)像素、(20.85 ± 6.384)像素与(25.84 ± 13.754)像素,平均 RMSE 为(21.39 ± 8.698)像素(约为 0.4%)。

实验观察发现,该误差主要集中于受微风干扰的细微枝条以及深度突变的边缘区域,其在全图中的空间占比极小。在果园非结构化环境下,深度相机机会因遮挡和强光导致深度图像存在边缘伪影与数据空洞。0.3% ~ 0.5% 的配准偏差未超过深度图像自身的边缘噪声,并不会从根本上破坏 Vis - NIR 图像与深度图像之间的宏观几何对应关系。同时,由于深度卷积网络会下采样提取特征(P3 至 P5 层),该尺度下的输入端偏差在深层特征图中会被进一步压缩至 0.6 ~ 3.2 像素。为量化配准误差对空间几何门控 S - Gate 模块利用深度先验进行抗噪引导的实际影响,将在后文进行配准偏移情况下的模型性能测试,进一步分析其对模型最终分割精度的实质影响。

1.3 数据集构建方法

为解决原始 5 120 像素 × 5 120 像素高分辨率图像直接缩放导致微小枝条特征丢失的问题,本文采用一种无填充自适应切片策略,对配准后的 Vis - NIR 图像、深度图像及矢量标注进行同步处理。

在切片处理前,利用 Labelme 标注软件对原始 5 120 像素 × 5 120 像素分辨率的 Vis - NIR 图像进行人工标注。考虑到果园环境的复杂性,采用多边形工具沿枝条边缘进行精细描点,生成对应的 JSON 矢量文件。由于 Vis - NIR 与深度图像已完成配准,该标注文件可作为 2 种模态的共享标注文件。

不同于部分研究^[20]将果树细分为主干、主枝与

侧枝,本研究采取了统一语义标注策略,将所有木质化与半木质化茎干统一定义为枝条类别,标注完成图像如图 4 所示。这一策略主要基于面向的作业场景考量,在机器人进行绕障采摘、图像拍摄等场景时,任何直径的枝干均构成实体障碍物,全类别的统一分割更能迫使网络关注枝条的整体连通性与几何边界,从而为后续的果树整体骨架提取或避障规划提供最完备的环境语义图。

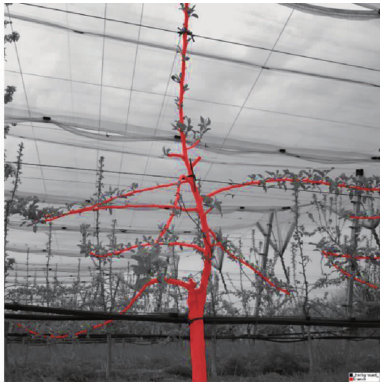


图 4 标注完成的图像
Fig. 4 Annotated image

参照 YOLO 系列标准,分辨率设定图像切片尺寸 S_{slice} 为 640 像素;为了保证跨越切片边界的长枝条及分叉结构的拓扑连通性,防止硬切割造成的上下文语义截断,设定最小重叠率阈值 $\lambda_{\min}$ 为 0.2。

首先计算满足重叠率约束的固定滑动步长 L_{str} , 计算式为

$$L_{\text{str}} = S_{\text{slice}} (1 - \lambda_{\min}) \quad (1)$$

基于该步长,在图像宽和高方向上生成基础采样坐标序列 $\{c_0, c_1, \dots, c_n\}$, 其中 $c_i = iL_{\text{str}}$, 直至满足 $c_n + S_{\text{slice}} > L$ (L 为原始图像在宽或高方向上的绝对边长,像素)。

针对图像边缘无法被完整窗口覆盖的区域,传统方法常采用纯黑背景填充,但这会引入边缘伪影。本研究采用回退补偿机制,当基础滑动无法刚好覆盖边界时,强制增加一个以图像边缘为终点的回退采样窗口。最终的切片起始坐标集合 C_{final} 计算式为

$$C_{\text{final}} = \{c_i \mid c_i + S_{\text{slice}} \leq L\} \cup \{L - S_{\text{slice}}\} \quad (2)$$

式中 c_i ——基础采样序列中第 i 个坐标位置

该策略不仅确保了全图无死角覆盖,且在图像边缘处通过增加重叠率实现了数据的自然增强。

利用生成的坐标集合,同步裁剪 Vis-NIR 与深度图像,并对多边形标注进行重映射。通过计算原始标注 P_{org} 与切片窗口 W_{slice} 的几何交集 $P_{\text{new}} = P_{\text{org}} \cap W_{\text{slice}}$, 将绝对坐标转换为切片系相对坐标,修正被窗口截断的实例边界,生成实例分割标签。

为防止数据泄露,本文严格按照“先划分,后切片”的方式进行数据集划分,首先将原始 551 组高分辨率图像按照比例 8:2 划分为训练集和验证集,确保同一幅大图的所有小型切片仅存在于同一数据子集中,避免了重叠区域导致的训练集与验证集特征同源的问题。

经上述处理,原始 551 组高分辨率图像被扩充为 25 637 组的样本(每组含 Vis-NIR 图像、深度图像及对应标签),切片后部分图像仅包含背景信息,作为负样本包含在数据集中,提升训练后网络的鲁棒性,共 7 648 组(占 29.8%)。最终数据集中训练集图像 20 904 组,包含负样本 6 194 组(占 29.6%);验证集图像 4 733 组,包含负样本 1 454 组(占 30.7%),训练集与验证集的正负样本比例差异的原因为“先划分,后切片”的数据集划分方式。该数据集规模及多样性能够有效支撑深度神经网络在复杂非结构化环境下的特征学习。

2 CAROL-YOLO 网络模型构建

在农业非结构化环境的实时感知任务中,单阶段实例分割网络因其高效的推理速度而被广泛应用,与 Mask R-CNN 等经典二阶段实例分割网络相比,一阶段检测器从一定程度上可以加快检测速度。其中,YOLO v8 网络作为目前主流的一阶段检测器,在 NVIDIA Jetson 等嵌入式平台上展现出优异的实时响应能力与算子兼容性,契合农业机器人对端侧部署与实时响应的需求。

然而,原生的 YOLO v8-seg 网络仅设计用于处理单模态图像的三通道数据,缺乏对多模态异构信息的融合能力;且随着网络深度的增加,多次下采样操作导致细微枝条的高频边缘信息丢失,降低实例分割任务中的边缘分割精度。本文以 YOLO v8-seg 为基准骨干,设计了双流伪孪生特征提取网络,并嵌入空间几何门控(S-Gate)与边缘感知模块(EAM),提出了 CAROL-YOLO 架构,旨在突破传统单模态网络在复杂光照及精细分割任务上的性能瓶颈,图 5 为 CAROL-YOLO 模型总体框架,具体改进包括伪孪生骨干网络设计、多尺度门控融合和边缘高频特征增强。

2.1 双流特征提取与融合机制

为了实现 Vis-NIR 与深度信息的有效互补,设计了基于伪孪生架构的双流特征提取网络,并通过 S-Gate 机制实现跨模态特征的融合。

(1) 伪孪生骨干网络设计

在多模态视觉系统中,骨干网络的设计直接决定了特征编码的质量^[23]。针对 Vis-NIR 图像与深

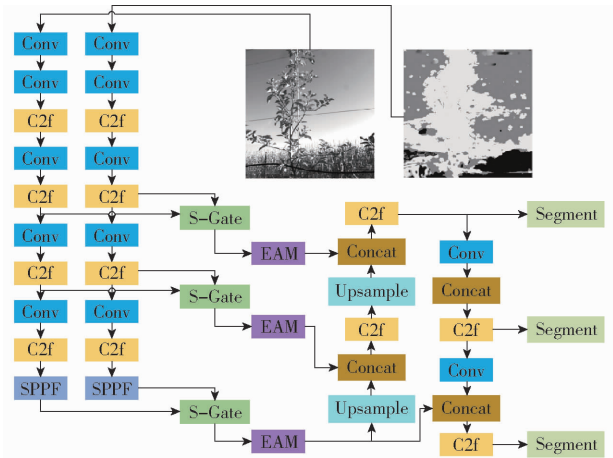


图 5 CAROL-YOLO 架构图

Fig. 5 CAROL-YOLO architecture diagram

度图像 2 种模态,考虑到 Vis-NIR 图像与深度图像的数据物理属性不同,Vis-NIR 图像主要表征物体表面的反射光强与纹理细节,深度图像主要表征物体到相机的距离;且深度图像受限于主动双目/结构光成像原理,原始深度图像常伴随边缘伪影与数据空洞,数据质量较低,直接进行简单的通道拼接会导致网络无法完全学习到 2 个模态的物理特征,并可能导致深度图像的噪声污染 Vis-NIR 图像的特征。

在此基础上,采用常见的孪生网络让 2 个分支共享权重,会导致同一组卷积核同时需要提取纹理边缘与深度信息,这种耦合会导致网络优化方向发散,难以同时达到最优。

因此,本研究并未采用简单的通道拼接或孪生网络,而是构建了伪孪生架构,本文设计的伪孪生网络虽然在 2 个分支上均保持了与 YOLO v8-seg 一致的 CSPDarknet 拓扑结构,但不共享权重参数。这种解耦设计赋予了网络针对特定模态独立学习最优特征表示的能力,例如 Vis-NIR 分支专注于提取叶脉、树皮纹理,而 Depth 分支专注于提取枝干的空间信息。

(2) S-Gate 空间几何门控

在特征融合阶段,传统方法^[24]通常采用通道拼接或逐元素相加的形式进行模态间的融合。然而,这类线性融合方式忽略了模态间的噪声干扰,特别是在深度图像质量较低的情况下,这样的融合方式会导致 Vis-NIR 图像受到深度图像噪声的影响,降低分割质量。

为此,本文提出了 S-Gate 空间几何门控单元,其核心逻辑为利用深度图像几何信息的确定性,给 Vis-NIR 图像提供基于几何线索的空间注意力指导,其结构如图 6 所示。

假设输入特征为 F_{nir} 和 F_{depth} ,S-Gate 的处理过程分为 2 部分:空间几何注意力建模和跨模态门控

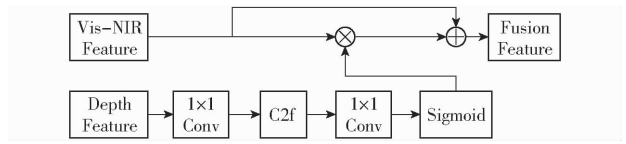


图 6 空间几何门控(S-Gate)结构示意图

Fig. 6 Schematic of spatial geometric gate (S-Gate) structure

与残差连接。

输入的深度特征 F_{depth} 经过第 1 层 1×1 卷积处理。该层仅在通道维度进行线性映射,不具备空间感受野,从而在实现深度值特征编码的同时,确保了单个像素的深度信息互不干扰,有效避免了背景噪声向枝条区域的扩散,编码后的特征输入至 C2f 模块,提取出更鲁棒的高层几何纹理特征 $\hat{F}_{\text{depth}}$,提取过程式为

$$\hat{F}_{\text{depth}} = \text{C2f}(\text{Conv}_{1 \times 1}(F_{\text{depth}})) \quad (3)$$

式中 C2f——YOLO v8 中的特征提取模块

$\text{Conv}_{1 \times 1}$ ——卷积核尺寸为 1×1 的卷积运算

聚合后的特征经过第 2 层 1×1 卷积与 Sigmoid 激活函数,生成与输入特征维度一致的空间几何注意力图 M_{geo} ,计算式为

$$M_{\text{geo}} = \sigma(\text{Conv}_{1 \times 1}(\hat{F}_{\text{depth}})) \quad (4)$$

式中 σ ——Sigmoid 非线性激活函数

生成空间几何注意力图后,模块与 Vis-NIR 特征进行逐元素相乘并引入残差连接,确保了即使深度图像因为噪声在枝条区域产生了空洞,也不会影响到 Vis-NIR 特征流的特征提取,残差连接过程式为

$$F_{\text{fused}} = F_{\text{nir}} + F_{\text{nir}} \odot M_{\text{geo}} \quad (5)$$

式中 F_{fused} ——跨模态门控融合输出的特征图

在此机制下,深度图像的空间几何特征充当了空间滤波器,鉴于切片化处理导致原始图像中苹果树枝条的全局拓扑结构与整体空间几何特征不可避免地丢失,S-Gate 模块利用深度模态中依然保留的局部空间一致性,有效弥补了 Vis-NIR 图像分支在局部受限视场下的语义模糊问题,即便深度图像在非结构化环境中因传感器噪声或强光干扰产生数据空洞,残差连接的方式也可以完整保留原始 Vis-NIR 图像流的特征,显著增强了网络在复杂背景中精准剥离前景枝条的鲁棒性。

2.2 EAM 边缘感知模块设计

由于卷积神经网络的固有特性,随着骨干网络层数的加深(P5 层下采样倍率达 32 倍),图像中的高频空间细节不可避免地在连续的跨步卷积与池化操作中发生衰减。这一问题在处理多尺度目标时尤为突出,这体现在全枝条类别的统一标注策略上,分

割头需在同一特征尺度下同时回归粗壮的主干与末端微细枝条。这使得常规网络在深层特征模糊与多尺度目标耦合的双重挑战下,难以精准捕捉细微枝条的像素级边界。

因此本文并未采用复杂的解码器结构,而是基于频域差异分析原理^[21],设计了一种轻量级的边缘感知模块 EAM,网络结构如图 7 所示。

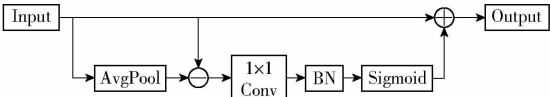


图 7 边缘感知模块 (EAM) 结构示意图
Fig. 7 Schematic of edge awareness module (EAM) structure

(1) 频域差异分析与高频残差提取

不同于传统边缘检测算子 (如 Sobel, Laplacian 等) 依赖固定权重的卷积核计算梯度, EAM 模块通过自适应差分来解耦特征图中的高频边缘分量。假设输入特征图为 $F_{in} \in \mathbf{R}^{C \times H \times W}$, EAM 模块首先构建一个基于平均池化的低通滤波器, 用于提取图像的低频背景信息 F_{low} , 计算式为

$$F_{low} = \text{AvgPool}(F_{in}, k = 3, s = 1, p = 1) \quad (6)$$

式中 k ——池化核大小

s ——滑动步长

p ——边缘填充大小

AvgPool——二维平均池化

由于平均池化具有平滑去噪特性, F_{low} 主要保留了平坦区域的语义强度, 而抑制了边缘突变, 由于该模块主要针对的是末端细微枝条的特征增强, 因此选用 3×3 的池化核足以捕捉边界特征。模块根据信号处理中的非锐化掩膜原理, 通过将原始特征与低频特征进行差分, 解耦出包含物体边界与纹理细节的高频残差 F_{high} , 计算式为

$$F_{high} = F_{in} - F_{low} \quad (7)$$

相较于可学习的卷积操作, 这种基于代数差分的边缘提取方式不引入额外参数, 且能够更纯粹地捕获特征层面的几何突变。

(2) 动态局部对比度增强

为了将提取到的高频边缘信息转换为对分割任务的有效监督, EAM 模块进一步设计了自适应特征重校准机制。模块利用 1×1 卷积与 Sigmoid 激活函数将高频残差 F_{high} 映射为像素级的边缘增强权重图 W_{edge} , 计算式为

$$W_{edge} = \sigma(\text{BN}(\text{Conv}_{1 \times 1}(F_{high}))) \quad (8)$$

式中 BN——批归一化操作

由式(7)代数差分产生的高频残差 F_{high} 内部包含了正负双向的激变值, 表征了枝条边缘的梯度方向与强度。通过引入 1×1 卷积, 网络对这些无约束

的频域响应进行特征通道间的线性组合与投影, 并利用 Sigmoid 激活函数将其非线性映射至 $(0, 1)$ 的概率区间内。这一过程将包含了正负梯度的高频差分信号稳定地转换为了正向的边缘注意力权重 W_{edge} , 有效避免了深层网络前向传播中的数值振荡。

该权重图在边缘纹理区域趋近于 1, 在平坦背景区域趋近于 0。利用该权重图, 模块采用一种乘性增强策略将边缘权重注入原始特征流, 实现局部对比度的动态拉伸, 计算式为

$$F_{out} = F_{in} \odot (1 + W_{edge}) \quad (9)$$

式中 F_{out} ——最终输出特征图

通过该模块, 在枝条内部区域特征值按比例保留, 维持基础语义信息; 在枝条边界区域提供增益系数, 放大边缘的特征响应幅值。从拓扑结构上看, 上述方法构成了一种带有注意力机制的残差连接, 但其本质功能不同于传统残差连接的特征相加, 而是实现了一种基于乘性的信号幅度调制。通过 W_{edge} 项, 网络能够动态地重校准局部特征的响应强度, 从而显著提升边缘区域的信噪比。

3 实验

3.1 实验环境与评价指标

本研究依托西北农林科技大学高性能计算平台进行训练, 运行硬件包括 8 个英特尔至强 Platinum 8358 的核心, 64 GB 运行内存, 1 块英伟达 A800 SXM 80GB 显卡, 使用服务器端的 conda 建立虚拟环境, 选择 Python 3.9 作为开发语言。

考虑到 Vis - NIR 图像和深度图像多模态数据集的特殊性, 所有模型均采用从零训练策略, 不加载预训练权重 (Mask R - CNN 由于网络特性加载预训练权重进行训练)。输入图像尺寸统一调整为 640 像素 $\times$ 640 像素。优化器选用随机梯度下降 (SGD), 初始学习率设为 0.01, 权重衰减系数为 0.001, 训练轮次 (Epochs) 设定为 300。

为了在有限的显存资源下最大化训练效率与梯度稳定性, 本研究实施了动态批次策略: 针对 CAROL - YOLO 参数量较小的 Nano/Small 模型, Batch Size 设为 64 以加快收敛; 针对参数量较大的 Large 模型 Batch Size 调整为 32 以避免显存溢出。

在训练阶段, 启用 Mosaic 马赛克增强, 提升模型鲁棒性。针对单通道的 Vis - NIR 图像与深度图像, 在输入端执行通道复制操作, 将其扩展为 $B \times 3 \times 640 \times 640$ 张量 (其中 B 为批量大小), 以适配骨干网络的底层卷积接口。

本研究的验证集共包含 4 733 组图像, 涵盖了大规模的复杂果园场景与枝条实例。在此大样本基数

下计算所得的 mAP 等评估指标具有充分的统计学稳定性,可以有效避免小样本评估中的偶然性误差。

3.2 CAROL – YOLO 模型在不同规模下的性能评估

为了验证本文提出的 CAROL – YOLO 模型的泛化能力,探究模型在不同场景部署下的收益边界,基于原版 YOLO v8 – seg 的 Nano(n)、Small(s)与 Large(l)3 种不同缩放因子构建了改进模型,并与对应的基准模型及原版 Medium(m)模型进行了性能对比,具体对比结果如表 1 所示。

表 1 改进算法性能

Tab.1 Performance of improved algorithm

模型	参数量	精确度/%	召回率/%	mAP50/%	mAP50 – 95/%
n	3.2×10^6	68.2	56.4	63.0 ± 1.3	33.4 ± 0.9
改进 – n	5.4×10^6	70.4	59.5	67.8 ± 1.7	36.8 ± 0.9
s	1.16×10^7	67.5	63.7	65.0 ± 0.3	35.0 ± 0.1
改进 – s	2.03×10^7	72.0	64.6	70.9 ± 0.2	39.6 ± 0.1
m	2.73×10^7	70.0	64.3	68.8 ± 1.6	38.2 ± 1.2
l	4.57×10^7	72.4	65.3	70.5 ± 0.6	39.9 ± 0.4
改进 – l	7.83×10^7	73.6	67.8	73.8 ± 0.2	42.4 ± 0.2

注:表中的精确度 (Precision) 与召回率 (Recall) 等单点指标,均为 YOLO 模型默认评估策略,取自各模型在验证集上达到最大 F1 分数时的自动搜索置信度阈值,旨在展现各网络架构在最佳平衡点下的理论性能上限。面向实际工程部署的固定置信度 (Conf 为 0.25 与 0.50)性能对比详见表 2。

由于额外的骨干网络和模块的引入,各级别模型的数量均有所提升,但带来的是更显著的性能提升,为了更直观展示改进模型相较于基准模型的效能提升,绘制了参数量与分割精度的曲线,如图 8 所示。

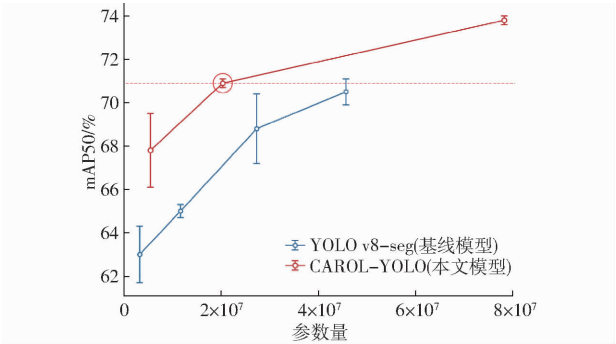


图 8 不同尺寸模型性能对比图

Fig.8 Performance comparison chart for models of different sizes

可以看出,本文提出的 CAROL – YOLO 模型的性能曲线始终位于基准 YOLO v8 – seg 曲线的上方,即改进模型在任意给定参数量级下,均能提供更优秀的分割精度。

在 Nano 量级,改进模型的 mAP50 提升约 4.8

个百分点,略高于参数量是其 2 倍的原版 s 模型的性能,体现了网络对于浅层网络特征丢失情况的修复能力;在 Large 量级,虽然 mAP50 只提升约 3.3 个百分点,但达到了 73.8%,体现了本研究的改进结构不是针对某一尺寸进行的专门优化,而是在不同尺寸下均有性能提升,证明了模型拥有极强的泛化能力。

由图 8 可知,性能表现最好的是改进模型的 Small 版本,在该版本下,改进模型体现出了优越性能,显著超越了更大参数量的原版 m 模型,在展现目标定位与基础分割能力的 mAP50 指标上,与参数量庞大的原版 YOLO v8l 模型基本持平,这体现了通过引入多模态信息和边缘感知机制, CAROL – YOLO – s 成功突破了轻量级模型的性能瓶颈,展现了最佳的效能平衡性能。结合农业机器人对嵌入式部署的模型大小需求,本文选定 CAROL – YOLO – s 作为核心代表。

3.3 性能对比实验

为了验证 CAROL – YOLO 模型的有效性,实验选取了最新的 YOLO 11 和 YOLO 26 中与 CAROL – YOLO – s 参数量相近的 Medium(m)模型,及经典的二阶段模型 Mask R – CNN 进行对比。推理均在同一本地平台 (Intel i5 – 13400F, 32 GB 内存, NVIDIA RTX 4070 12 GB)上进行。推理速度测试中,为消除 GPU 冷启动带来的显存分配延迟误差,正式测定前均对模型进行 10 幅图像的预热。随后选取 100 帧测试集图像进行连续评估,为过滤底层系统环境的随机毛刺扰动,单帧图像均重复推理 5 次取均值作为单次有效采样基准,以评估模型之间的性能,结果如表 2 所示。由于单阶段网络 YOLO 系列与基于 COCO 评估标准的二阶段网络 Mask R – CNN 在单点指标的默认评估策略上存在差异。为保证横向对比的公平性,表 2 所有模型的精确度与召回率均统一使用 YOLO 推理默认置信度阈值 (Conf 为 0.25) 与标准置信度阈值 (Conf 为 0.50)进行统计,故结果与表 1 中基于最大 F1 分数自动搜索得到的结果存在差异。表中列出的推理速度为纯网络前向推理时间,旨在对比不同模型架构的计算效率。考虑到本研究面向采摘和巡检的工程任务,末端微细枝条需较高分辨率与抗干扰能力才能精准分割,消费级深度相机自带的可见光传感分辨率有限,故本系统采用外挂高分辨率工业相机以保障感知精度。在此特定硬件架构下,为获取机械臂所需的深度坐标,相机之间的配准为系统前置步骤。实际工程部署中,系统采用了控制系统多线程处理机制,整体耗时满足实际作业需求。

表 2 主流模型性能对比

Tab. 2 Performance comparison of mainstream models

模型	参数量	推理速度/(f·s ⁻¹)	置信度	精确度/%	召回率/%	mAP50/%	mAP50-95/%
CAROL-YOLO-s	2.03×10 ⁷	103.1±6.3	0.25	70.0	66.9	70.9±0.2	39.6±0.1
			0.50	83.5	52.7		
YOLO v8m-seg	2.73×10 ⁷	101.0±3.1	0.25	70.5	62.4	68.8±1.6	38.2±1.2
			0.50	85.0	35.2		
YOLO 11m-seg	2.23×10 ⁷	90.3±5.9	0.25	74.3	63.5	69.6±0.4	38.8±0.6
			0.50	84.6	33.7		
YOLO 26m-seg	2.69×10 ⁷	84.8±4.1	0.25	76.8	58.3	68.4±0.7	38.2±0.5
			0.50	88.5	26.8		
Mask R-CNN	4.44×10 ⁷	22.7±1.5	0.25	59.5	65.4	58.6±0.4	29.7±0.2
			0.50	64.6	62.9		

实验表明,CAROL-YOLO-s 实现了轻量化与高性能的统一。在其参数量仅为 2.03×10^7 , 相比 YOLO 11m-seg (2.23×10^7) 和 YOLO 26m-seg (2.69×10^7) 分别减少了 8.9% 和 24.5% 的情况下取得了最优的分割精度。对比 YOLO 11m-seg 与 YOLO 26m-seg 可知,YOLO 26m-seg 更大的分割头并未带来显著的边界精度提升,证明单纯堆叠参数已面临边际效应。而本文提出的 CAROL-YOLO-s 得益于空间几何门控(S-Gate)与边缘感知(EAM)机制,mAP50 达到 70.9%,更为严苛的 mAP50-95 指标达到 39.6%,证明了多模态几何信息对枝条边缘分割的显著增益。

为客观评估模型的工程部署潜力,对比了各模型在固定置信度下的精确度与召回率。由表 2 可知,在 YOLO 推理默认置信度(Conf 为 0.25)下,YOLO 系列基线模型虽具有略高的精确度,但漏检相对较多;Mask R-CNN 在置信度较低时召回率保持了较好的水平,但精确度较低;CAROL-YOLO-s 则在保持精确度 70.0% 的同时,取得了 66.9% 的最高召回率,实现了较好的 P-R 平衡。

此外,当置信度阈值提高到 0.5 时,YOLO 系列单模态基线模型受复杂背景干扰,检出能力明显衰退,存在漏检隐患;Mask R-CNN 作为二阶段模型,并未体现明显的精确度提升,这代表了该模型在推理过程中存在大量的假阳性误检。而本文模型得益于 S-Gate 模块的空间深度特征补充,依然维持了精确度 83.5% 与召回率 52.7%,表明其在严苛防碰撞工况下具有更好的鲁棒性与实际应用价值。

为确保对比的绝对公平,实验预先加载了 COCO 预训练权重以最大化其特征初始化能力,且其拥有高达 4.44×10^7 的庞大参数量,但 mAP50-95 仅为 29.7%,远落后于本文模型与其他一阶段基线模型。其主要原因并非是模型训练不充分,更有可能是两阶段架构不适用于苹果树枝条的几何特

征以及本研究面向全局骨架提取任务的标注方式。

在本研究中,为了向机器人避障规划与全局骨架提取提供完备的环境语义,标注阶段将粗壮的主干与极细的末端枝条统一标注为枝条。这一策略虽然贴合本研究针对的实际工程任务,但不可避免地导致了数据集极高的类内方差,需要分割目标的长宽比与尺度跨度极大。Mask R-CNN 网络强依赖区域建议网络生成预设锚框,难以同时拟合大面积的垂直主干与长宽比极端的微细枝条;其次,网络将提取到的特征图强制量化并池化为固定的小尺寸(如常规的 28 像素×28 像素)进行掩膜预测,尽管通过修改锚框比例可以一定程度上缓解该问题,但枝条形态与方向多变,固定修改尺寸无法从根本上解决网格化压缩的问题。这种空间维度上的强制降采样与网格化压缩,会导致末端细微枝条特征丢失。此外,单模态架构缺乏几何深度的显式约束,使其在面对颜色、纹理与前景高度相似的背景枝条时极易陷入特征混淆。

相比之下,本研究提出的 CAROL-YOLO 采用基于 YOLO 系列的一阶段分割网络,无需预设锚框,可以更灵活地自适应类内方差更大的分割任务;同时配合 EAM 边缘感知模块,增强极端长宽比分割目标的高频边缘特征,保障了苹果树从主干到末端枝条的特征提取的完整性。该模型在大幅降低计算负担的同时保持了极具竞争力的精度,更适配果园嵌入式终端。

3.4 消融实验

为了验证 CAROL-YOLO 模型中各个关键组件(双流骨干架构、S-Gate 门控模块、EAM 边缘感知模块)的有效性,实验设计了逐项递进的消融分析。以改进后的 YOLO v8s 为基础架构,在相同数据集和模型大小(s)下进行了对比实验,其中伪孪生架构未使用本文提出的空间几何门控模块 S-Gate 的

实验,而是使用传统的拼接卷积进行融合 (Concat + Conv),结果如表 3 所示。

表 3 模型消融实验
Tab.3 Model ablation experiment

模型 编号	伪孪生 架构	S - Gate	EAM	mAP50/ %	mAP50 - 95/ %	参数量
0	×	×	×	64.8	34.9	1.16×10^7
1	×	×	✓	69.3	38.4	1.21×10^7
2	✓	×	×	68.4	37.6	1.69×10^7
3	✓	✓	×	70.3	39.0	1.99×10^7
4	✓	✓	✓	71.0	39.7	2.03×10^7

注:表中✓表示使用了该模块或架构;×表示未使用该模块或架构

对比模型 2 与模型 3 可以发现,在同为双流架构的前提下,使用本文提出 S - Gate 替代传统的拼接卷积融合,模型的 mAP50 从 68.4% 跃升至 70.3%,mAP50 - 95 提升了 1.4 个百分点。S - Gate 通过“点空间几何编码”与门控机制,成功抑制了无

效深度噪声,实现了真正的有效融合;对比模型 0 与模型 1,在引入 EAM 模块后,模型在参数量变化不大的情况下精度大幅度提高,这表明通过增强枝条边缘高频特征,可以帮助模型显著提升特征提取能力,增加分割精度;对比模型 1、模型 3 和模型 4(完整模型),引入深度信息并配合合理的门控与边缘增强机制后,整体 mAP50 相较于仅添加单模块的模型进一步得到了提升,这证明了 NIR 光谱信息与 Depth 几何信息在果园非结构化环境中具有高度互补性,双流架构能够更稳健地应对光照变化与背景杂波干扰。

3.5 热力图分析

为了探究 CAROL - YOLO 模型内部的特征演化机制,验证多模态融合策略与边缘感知模块在复杂果园环境下的物理有效性,本研究提取了网络推理过程中关键层级的特征响应热力图进行可视化分析,特征演化过程与特征响应差异热力图如图 9 所示。

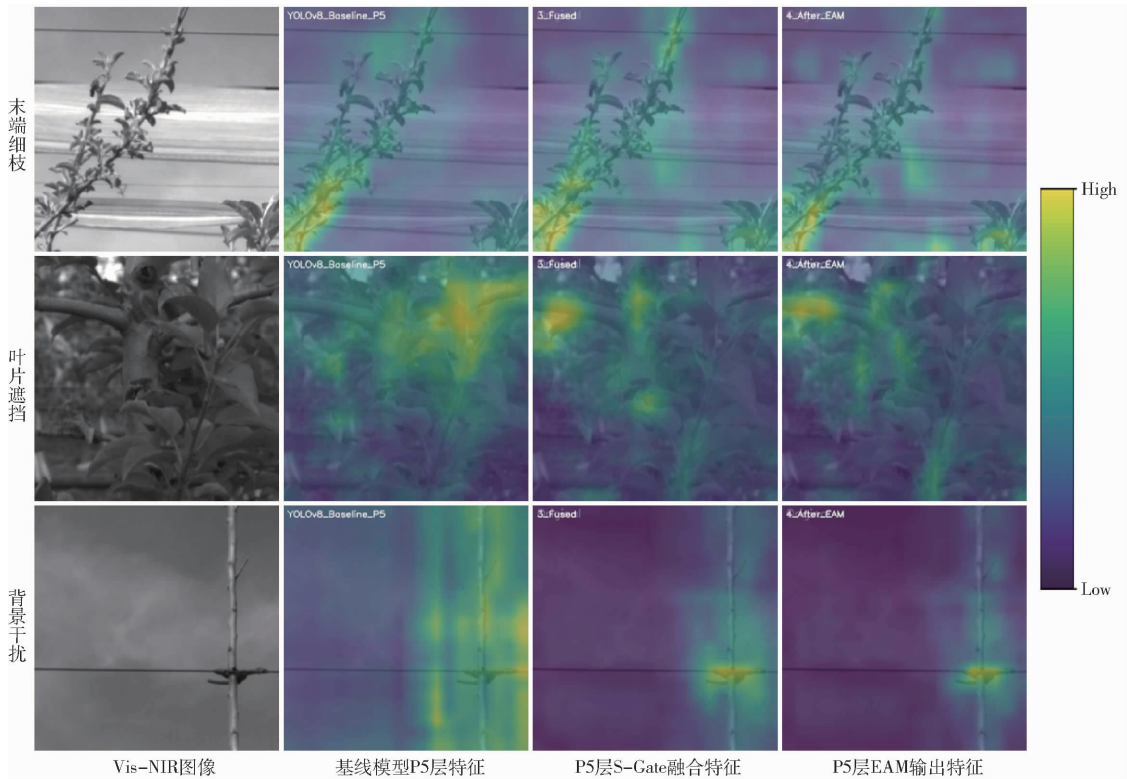


图 9 CAROL - YOLO 热力图对比

Fig.9 CAROL - YOLO heatmap comparison

热力图中高响应区域 (High) 表征网络对特征的高强度激活,低响应区域 (Low) 表征特征响应微弱或被抑制。图 9 展示了在复杂干扰下,本文模型能够有效滤除环境噪声,实现对目标枝条拓扑结构的聚焦;图 9 中各层热力图均采用统一的相对颜色映射尺度。各层激活特征均经局部最大值归一化至 [0,1] 区间,以直观展示网络在不同深度的相对空

间注意力分布。

观察基线模型 (YOLO v8m - seg) 的 P5 层特征可知,在非结构化环境中,常规单模态网络存在特征弥散与注意力偏移的问题。面对枝条颜色相近的遮挡叶片时,基线模型表现出了冗余激活与误判;同时,针对末端细微枝条,多次下采样导致高频空间信息丢失,目标特征较为暗淡。

CAROL-YOLO 在引入 S-Gate 模块后,利用深度信息生成的几何先验提供了更高的语义纯度。得益于深度约束,门控机制成功构建了仅关注前景深度的空间注意力掩膜,滤除网络的背景干扰,强制关注前景信息。融合特征图中的主干响应获得了定向增强,展现了跨模态几何约束在复杂背景剥离中的强大能力。

观察 EAM 模块作用后的输出特征图,在深层网络中略显模糊、发散的枝条边界被重新锐化。通过频域边缘的显式增强,细微枝条区域的像素响应强度显著提升,原本在基线模型中丢失的末端细枝特征被重新关注,提升了网络的分割精度。

这种使用空间几何门控先剔除无效背景、再使用边缘感知增强拓扑边界的特征演化过程,从可视化的角度证明了多模态协同机制的有效性。得益于这种高效的特征提取策略,网络无需堆叠参数,也能在复杂遮挡与背景干扰下实现较高精度的实例分割。

3.6 环境遮挡度分级性能测试

由于实际果园环境中,果树枝条存在严重的遮挡情况,为了验证本文模型在不同遮挡情况下的分割精度,本研究根据枝条的遮挡情况,由两名具备农业视觉研究经验的研究人员交叉评估,将测试集划分为 3 个测试子集,每个场景挑选 100 组典型图像作为测试子集,定义与划分标准如下:

(1)轻度遮挡/无遮挡:目标结构完整且视觉显著性高。枝条具有连续的几何形态,基本无背景果树枝条干扰,目标边界清晰;研究人员可直接无歧义辨识枝条的位置与边界。

(2)中度遮挡:局部截断与中度语义混淆。枝条受叶片或果实局部遮挡导致空间几何特征不连续,或背景中存在果树枝条,但较为模糊;研究人员需要结合上下文枝条走向进行短暂推理后辨识枝条的位置与边界。

(3)重度遮挡:严重结构破碎与极低信噪比。目标发生大面积物理截断,且受强阴影、背光或极度复杂的相似背景影响,研究人员需借助图像处理软件的对比度变化或放大功能,才能完成枝条位置与边界辨识。

为避免性能测试中引入非目标变量(如 YOLO 跨代版本中其他模块的差异),选用基线模型 YOLO v8-seg 的 Medium(m) 版本作为性能对比模型,讨论在不同场景下的分割性能。

如表 4 所示,在轻度与中度遮挡工况下,本文提出的 CAROL-YOLO 模型展现出显著优势,其 mAP50 分别达到 74.3% 与 73.1%,超越参数量更

大、通道更多的 YOLO v8m-seg 基线模型。在重度遮挡工况下,两者 mAP50 基本持平,其原因在于通道容量与特征输入的相互制约。面对重度遮挡的目标,CAROL-YOLO-s 虽提供了丰富的跨模态与高频边界特征,但受限于 YOLO 架构 Small(s) 尺寸的颈部通道数量限制,形成了信息传递瓶颈,制约了特征在深层的充分融合。相比之下,YOLO v8m-seg 依托更宽的通道容量提升了网络表征能力,在重度遮挡环境中保持了特征信息的丰富度。但本研究提出的模型精确度在 3 个场景下保持稳定水平,证明了 EAM 模块在复杂场景下的特征边界提取能力。

表 4 遮挡实验性能对比

Tab.4 Comparison of performance in occlusion tests

测试子集	模型名称	精确度	召回率	%	
				mAP50	mAP50-95
轻度	CAROL-YOLO-s	71.2	67.4	74.3	40.1
	YOLO v8m-seg	68.7	66.7	70.1	36.9
中度	CAROL-YOLO-s	72.6	67.2	73.1	41.9
	YOLO v8m-seg	69.4	71.6	71.7	41.1
重度	CAROL-YOLO-s	70.5	62.0	66.7	36.4
	YOLO v8m-seg	65.8	66.3	66.8	37.1

注:表中性能参数均为 YOLO 在默认参数下得出。

值得讨论的是,在轻度遮挡场景中,基线模型性能相比中度遮挡出现大幅度下降。通过对轻度遮挡的测试子集研究发现,在此类场景中,枝条无叶片遮挡与背景干扰,对人类视觉而言属于毫无干扰的“最易辨识”的情况,但这种情况往往意味着背景是充斥着高频细节的天空、果园防雹网或者土地,由于基线模型的单模态输入缺乏深度物理先验,极易被远景天空等背景干扰,出现严重的感受野特征稀释的情况,这与图 9 中热力图对比的背景干扰情况吻合。而本文模型凭借 S-Gate 的空间几何门控机制,有效强化了近景特征,实现了对枝条更精准的感知,具体分割效果差异如图 10 所示。轻度干扰情况下,基线模型受远处防雹网干扰产生了特征混淆;中度干扰情况下,基线模型由于背景的果树枝条干扰,出现了前景果树枝条的漏检;重度干扰情况下,通道数更多的基线模型检出了更多枝条,但也错误地识别了果树固定用塑料管。中度遮挡工况下,CAROL-YOLO 未识别出的为背景枝条。

3.7 模态配准误差传递分析

为评估多模态图像配准误差对模型分割精度的影响,本文在验证集的深度图像中增加随机方向、固定像素的偏移扰动(方向为 0~360°随机值,使用双线性插值映射像素),并进行性能测试。在极端的 50 像素(约等于重度遮挡平均配准误差加两倍标准

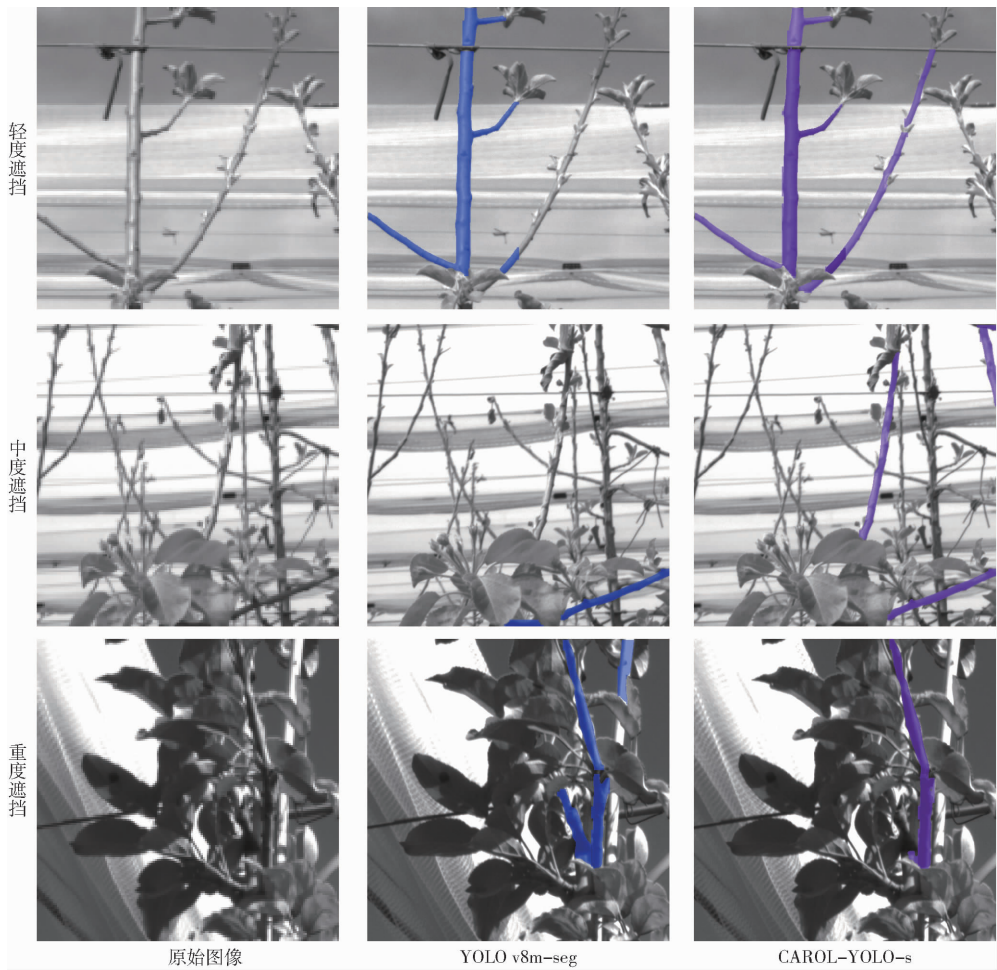


图 10 不同等级复杂场景下模型分割结果对比

Fig. 10 Comparison of model segmentation results in scenes of varying complexity

差) 偏移下, 本文提出的 CAROL - YOLO - s 的 mAP50 仅由 70.9% 降至 70.6% (降幅 0.3 个百分点)。进一步深度模态失效性能测试(深度图的值设置为全零)中, 缺失深度先验后模型 mAP50 降至 68.5% (下降 2.4 个百分点)。这证实了模型针对模态误差的鲁棒性并非因为深度特征权重过低, 结合真实果园深度图像常伴随边缘伪影与空洞的工况, 推测网络在拟合低质量样本时, 自适应降低了对深度高频边界特征的感知权重, 转而将其作为宏观区域的辅助特征。

此外, S - Gate 模块基于通道交互, 对空间严格对齐的依赖相对较低。输入端 50 像素的错位经 32 倍下采样至深层后, 理论偏移仅约 1.56 像素。相对于深层卷积的感受野尺寸, 该偏移未超出模型在训练中建立的特征适应范围。带有偏移的深度特征仍能覆盖目标区域, 从网络结构的角度印证了上述容错表现。

4 讨论

4.1 多模态融合与边缘感知的协同机制

本文提出的 CAROL - YOLO 模型是对原始 YOLO v8 - seg 模型的改进, 通过利用伪孪生骨干网

络和空间几何门控, 让深度图像辅助 Vis - NIR 图像流的特征感知枝条的空间几何信息, 有效抑制了背景果树枝条的干扰, 提升了网络的召回率; 利用边缘感知模块强化网络对高频边界特征的关注, 使得网络下采样的过程中仍然能保留细微枝条目标的边缘细节, 提升网络在背景干扰下的特征提取能力; 两者共同构成了“语义 - 边界”双重空间注意力机制, 通过对无效背景与平滑纹理的特征抑制, 提升了单位参数的有效信息承载量, 实现了轻量级参数量对原版更大参数量模型的性能超越。

4.2 面向机器人自主作业的系统集成与验证

为了验证提出算法在实际工程环境中的部署能力, 本研究将训练好的改进模型集成到了自主研发的苹果园巡检机器人系统中。如图 11a 所示, 该系统硬件包含六自由度机械臂 (JAKAZu7)、工业宽波段近红外相机及深度相机, 软件端通过 PyQt 多线程架构实现了从图像采集、实例分割到后端控制的实时闭环, 如图 11b 所示。

在实际作业流程中, 改进模型输出的高质量二值掩膜被直接输入到骨架提取与路径规划模块。系统能够从分割结果中成功重构出反映枝条拓扑结构

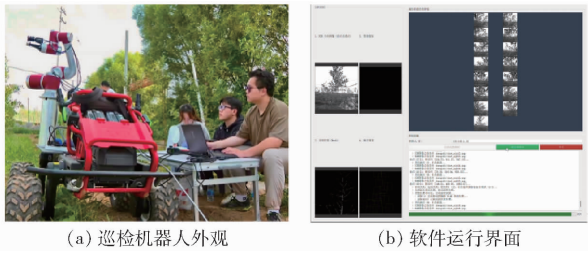


图 11 果园巡检机器人外观及软件界面
Fig. 11 Appearance and software interface of orchard inspection robot

的连通骨架,如图 12 所示。基于该骨架数据,机器人成功生成了覆盖全树的自主拍摄路径。集成测试表明,本文提出的分割网络具备极高的输出质量与环境适应性,能够作为底层感知核心,支撑上层复杂的机器人自主作业任务,具有显著的工程实用价值。

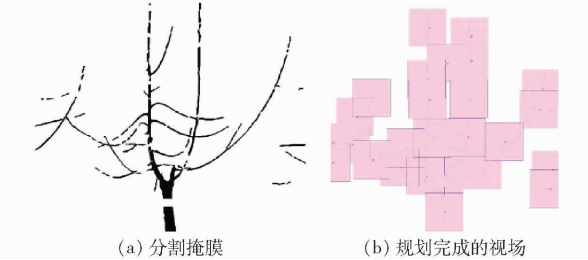


图 12 分割掩膜及规划图像采集用拍摄点
Fig. 12 Segmentation mask and planned image acquisition shooting points

4.3 局限性分析

尽管 CAROL-YOLO 模型展现出优越性能,但在极端条件下仍存在一定局限性亟待解决。鉴于双流架构对传感器对齐的高度依赖,若深度相机在强光直射下出现空洞或失效,深度图像分支可能引入噪声,未来需引入模态缺失训练策略以提高模型在传感器故障下的鲁棒性。

5 结论

(1)针对果园非结构化环境下枝条识别难、背

景干扰大的问题,本研究提出了一种融合光谱与几何信息的分割方法。构建了像素级配准的宽波段近红外与深度图像(Vis-NIR/Depth)数据集,利用近红外光对植物组织的敏感性与深度信息的几何约束,解决了常规可见光图像在光照剧变下特征不稳定的问题。在此基础上提出的 CAROL-YOLO 模型,通过伪孪生骨干网络解耦了纹理与空间特征,并利用空间几何门控(S-Gate)抑制了背景枝条的干扰。同时,引入边缘感知模块(EAM)针对性地增强了末端细微枝条的高频特征,有效改善了深层网络下采样过程中细小目标漏检和边缘模糊的现象。

(2)提出的模型在轻量化与检测精度之间取得了良好的平衡,验证了算法的有效性。实验结果表明,改进模型在不同参数量级上均实现了性能提升。以适于移动端部署的 CAROL-YOLO-s 为例,在参数量(2.03×10⁷)比基线模型 YOLO v8m-seg 减少 25.6% 的情况下,其 mAP50-95 提升 1.4 个百分点,达到 39.6%。与最新的 YOLO 11 及 YOLO 26 同量级模型相比,该方法在减少参数量的同时,依然保持了精度优势。消融实验数据进一步证实,引入的多模态几何约束与边缘增强机制,能够显著提升算法在复杂背景下的鲁棒性,说明了针对农业特定场景进行特征增强的必要性。

(3)验证了算法在果园机器人自主作业系统中的工程适用性。将该算法集成至自研的苹果园巡检机器人系统,实地测试表明,分割生成的二值掩膜质量高、连通性好,能够支持后续的枝条骨架提取与机械臂避障路径规划。尽管目前模型在强光导致深度相机失效的极端场景下仍存在依赖性,且数据集尚未覆盖冬季修剪场景,但本研究为果园机器人的精细化感知提供了一种低算力消耗、高精度的解决方案。

参 考 文 献

[1] Mhamed M, Zhang Z, Yu J, et al. Advances in apple's automated orchard equipment: a comprehensive research[J]. Computers and Electronics in Agriculture, 2024, 221: 108926.

[2] Zhang J, He L, Karkee M, et al. Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN)[J]. Computers and Electronics in Agriculture, 2018, 155: 386-393.

[3] Kang H, Chen C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards[J]. Computers and Electronics in Agriculture, 2020, 171: 105302.

[4] Gongal A, Amatya S, Karkee M, et al. Sensors and systems for fruit detection and localization: a review[J]. Computers and Electronics in Agriculture, 2015, 116: 8-19.

[5] Zeng T, Li S, Song Q, et al. Lightweight tomato real-time detection method based on improved YOLO and mobile deployment[J]. Computers and Electronics in Agriculture, 2023, 205: 107625.

[6] 张勤,庞月生,李彬. 基于实例分割的番茄串视觉定位与采摘姿态估算方法[J]. 农业机械学报, 2023, 54(10): 205-215.

- Zhang Qin, Pang Yuesheng, Li Bin. Visual positioning and picking pose estimation of tomato clusters based on instance segmentation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10): 205–215. (in Chinese)
- [7] 徐文凯,郎平,李道亮. 多目标鱼类实例分割 LIS-YOLO 轻量化模型研究[J]. 农业机械学报, 2025, 56(10): 94–101, 139.
- Xu Wenkai, Lang Ping, Li Daoliang. Lightweight LIS-YOLO model for multi-target fish instance segmentation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(10): 94–101,139. (in Chinese)
- [8] Lin Y, Xia Y, Xia P, et al. YOLO11-ARAF: an accurate and lightweight method for apple detection in real-world complex orchard environments[J]. Agriculture, 2025, 15(10): 1104.
- [9] Tang Y, Chen M, Wang C, et al. Recognition and localization methods for vision-based fruit picking robots: a review[J]. Frontiers in Plant Science, 2020, 11: 510.
- [10] 夏天,谢纯,李琳一,等. 基于改进 YOLO v8 的复杂温室环境黄瓜果实分割方法[J]. 农业机械学报, 2025, 56(5): 433–442.
- Xia Tian, Xie Chun, Li Linyi, et al. Improved YOLO v8 method for cucumber fruit segmentation in complex greenhouse environments[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(5): 433–442. (in Chinese)
- [11] 肖伸平,赵倩颖,曾甲元,等. 基于 YOLO-DCL 的复杂环境油茶果遮挡检测与计数研究[J]. 农业机械学报, 2024, 55(10): 318–326,480.
- Xiao Shenping, Zhao Qianying, Zeng Jiayuan, et al. *Camellia oleifera* fruits occlusion detection and counting in complex environments based on improved YOLO-DCL[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(10): 318–326,480. (in Chinese)
- [12] López-Barrios J D, Escobedo Cabello J A, Gómez-Espinosa A, et al. Green sweet pepper fruit and peduncle detection using Mask R-CNN in greenhouses[J]. Applied Sciences, 2023, 13(10): 6296.
- [13] Chen Z, Ting D, Newbury R, et al. Semantic segmentation for partially occluded apple trees based on deep learning[J]. Computers and Electronics in Agriculture, 2021, 181: 105952.
- [14] Gené-Mola J, Ferrer-Ferrer M, Gregorio E, et al. Looking behind occlusions: a study on amodal segmentation for robust on-tree apple fruit size estimation[J]. Computers and Electronics in Agriculture, 2023, 209: 107854.
- [15] Terven J, Córdova-Esparza D M, Romero-González J A. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS[J]. Machine Learning and Knowledge Extraction, 2023, 5(4): 1680–1716.
- [16] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961–2969.
- [17] Wang D, He D. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning[J]. Biosystems Engineering, 2021, 210: 271–281.
- [18] Shit S, Paetzold J C, Sekuboyina A, et al. cDice-a novel topology-preserving loss function for tubular structure segmentation [C] // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 16555–16564.
- [19] Li Y, Feng Q, Li T, et al. Advance of target visual information acquisition technology for fresh fruit robotic harvesting: a review[J]. Agronomy, 2022, 12(6): 1336.
- [20] Majeed Y, Zhang J, Zhang X, et al. Deep learning based segmentation for automated training of apple trees on trellis wires [J]. Computers and Electronics in Agriculture, 2020, 170: 105277.
- [21] 李道亮,赵晔,杜壮壮. 农业领域多模态融合技术方法与应用研究进展[J]. 农业机械学报,2025, 56(1): 1–15.
- Li Daoliang, Zhao Ye, Du Zhuangzhuang. Advances in multi-modal fusion techniques and applications in agricultural field [J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(1): 1–15. (in Chinese)
- [22] Liu C, Feng Q, Sun Y, et al. YOLACTFusion: an instance segmentation method for RGB-NIR multimodal image fusion based on an attention mechanism[J]. Computers and Electronics in Agriculture, 2023, 213: 108186.
- [23] Tang L, Xiang X, Zhang H, et al. DIVFusion: darkness-free infrared and visible image fusion[J]. Information Fusion, 2023, 91: 477–493.
- [24] Kulić D, Nakamura Y, Khatib O, et al. 2016 International symposium on experimental robotics: vol 1 [M/OL]. Cham: Springer International Publishing, 2017. <http://link.springer.com/10.1007/978-3-319-50115-4>.