doi:10.6041/j.issn.1000-1298.2025.06.046

温室番茄场景单目 RGB 模态向深度模态转换模型研究

高　旺　邓寒冰　邢志鸿　朱彦强
(沈阳农业大学信息与电气工程学院, 沈阳 110866)

摘要: 在温室场景下,针对番茄的表型解析、自主采摘、多模态联合分割等任务,快速、高精度、低成本地获取场景深度信息对农机视觉系统至关重要。本研究提出了一种嵌入注意力机制的 RGB 模态向深度模态转换的单目深度估计网络(RGB to depth conversion network,RDCN),以解决传统算法无法充分挖掘编码器的特征提取能力、深度估计精度低以及边界模糊问题。首先以 ResNext101 替换原来的 ResNet101 骨干网络,提取各个不同层级的特征图并将其融合到拉普拉斯金字塔分支,强调特征的尺度差异性并强化特征融合的深入与广泛性;同时为了增强模型获取全局信息以及上下文信息交互的能力,引入了置换注意力模块(Shuffle attention module, SAM),以减少下采样过程造成的局部细节信息丢失;其次,为了改善预测深度图的边界模糊问题,嵌入深度细化模块(Depth refinement module, DRM),感知预测特征图物体附近的深度变化;实现了温室场景下番茄植株图像深度信息的精准预测。试验结果表明,RDCN 在测试集上的平均相对误差、均方根误差、对数均方根误差、对数平均误差相比于基准模型分别降低了 20.5%、10.3%、8.3%、21.8%,在 1.25、1.25^2、1.25^3 阈值下的准确率分别提高 3.2%、1.2% 和 1.0%;并且网络生成的深度图像视觉上全局完整清晰且有较多的纹理细节;研究表明,RDCN 在温室场景下能够基于 RGB 信息获得高质量的深度信息,可为基于单目传感器的温室场景农机导航以及深度图像在多模态任务中的应用提供技术支持。

关键词: 温室; 番茄图像; 单目视觉; 深度估计; 拉普拉斯金字塔; 模态转换
中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)06-0499-10　　**OSID**:

Monocular RGB to Depth Conversion Model for Greenhouse Tomato Scene

GAO Wang　DENG Hanbing　XING Zhihong　ZHU Yanqiang
(*College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110866, *China*)

Abstract: In greenhouse environments, fast, high-precision, and low-cost acquisition of scene depth information is crucial for agricultural machine vision systems in tasks such as tomato phenotype analysis, autonomous harvesting and multimodal joint segmentation. An attention-embedded RGB-to-depth conversion network (RGB to depth conversion network, RDCN) for monocular depth estimation was proposed, addressing issues in traditional algorithms such as insufficient feature extraction capability of encoders, low depth estimation accuracy, and blurred boundaries. Firstly, ResNext101 was employed to replace the original ResNet101 backbone network, extracting feature maps from different levels and integrating them into the Laplacian pyramid branches. This approach emphasized the scale differences of features and enhances the depth and breadth of feature fusion. To enhance the model's capacity for capturing global information and contextual interactions, a shuffle attention module (SAM) was introduced. This module also helped minimize the loss of local detail information caused by the down-sampling process. This module also mitigated the loss of local detail information caused by the downsampling process. Secondly, to address the issue of blurred boundaries in the predicted depth maps, a depth refinement module (DRM) was embedded to capture depth variations near object edges in the predicted feature maps. For the study, an RGBD image acquisition platform for tomatoes was constructed in a daylight greenhouse environment using an Azure Kinect DK depth camera. To ensure diversity in the

收稿日期: 2024-09-22　修回日期: 2024-11-21
基金项目: 国家重点研发计划项目(2022YFD2002303-01)和辽宁省教育厅基本科研项目面上项目(JYTM20231303)
作者简介: 高旺(1999—),男,硕士生,主要从事计算机视觉和植物表型检测研究,E-mail: huasan999@126.com
通信作者: 邓寒冰(1984—),男,副教授,博士,主要从事计算机视觉和植物表型检测研究,E-mail: denghanbing@syau.edu.cn

dataset, images were collected at different times of the day based on varying light intensities in the greenhouse environment. The training set was then augmented by using three methods: horizontal mirroring, random rotation, and color jittering, resulting in a total of 8 515 aligned RGBD image sets of tomatoes. Experimental results indicated that by introducing the shuffle attention module and the depth refinement module, the model achieved accurate depth information prediction in greenhouse scenes. Compared with the baseline model, the visualized depth maps generated by the network demonstrated global completeness and clarity, with more texture details, especially in regions with complex geometries and significant depth variations, exhibiting superior visual effects. Experimental results showed that, compared with the baseline model, RDCN reduced the mean relative error, root mean square error, log root mean square error, and log mean error on the test set by 20.5%, 10.3%, 8.3%, and 21.8%, respectively. Additionally, accuracy under the 1.25, 1.25^2, and 1.25^3 thresholds was improved by 3.2%, 1.2%, and 1.0%, respectively. Moreover, the depth images generated by the network were visually complete and clear, with abundant texture details. Studies showed that RDCN can obtain high-quality depth information from RGB data in greenhouse environments, providing technical support for agricultural machine navigation in greenhouse scenarios using monocular sensors, as well as for the application of depth images in multi-modal tasks.

Key words: greenhouse; tomato images; monocular vision; depth estimation; Laplacian pyramid; modal conversion

0 引言

在日光温室环境下,能够实时获取场景的深度信息对于番茄的表型解析、自动采摘、农机导航等应用至关重要。通过获取深度信息,可以获得植物表型信息[1];自动采摘机器人可以利用深度信息来区分果实和叶片,精确定位待采摘的果实位置,从而实现高效、无损的采摘操作[2];室内无人车能够利用深度信息识别并避开各种温室内的障碍物,规划安全的行进路径[3];仇瑞承等[4]利用田间玉米的RGBD图像,实现了单株玉米株高测量;赵永杰等[5]结合深度信息和玉米植株高度测量方法完成了株高测量;徐胜勇等[6]基于RGBD图像实现了柑橘三维重建和表型检测。王纪章等[7]使用Kinect相机获得黄瓜苗的深度图像,实现了其生长监测;计算机视觉系统能够利用深度信息构建详细的三维环境地图,提高系统对场景的识别和认知能力。麦春艳等[8]基于RGBD图像完成果树三维重构以及果实识别定位;研究人员通常采用激光雷达(Light detection and ranging, LiDAR)等3D传感器来获取深度信息,但其价格昂贵且便携性较低,需要联合其他成像设备才能构成完整的深度信息,而且3D点云数据规模也较大,这导致其获取稠密且准确的深度图像的成本过高。

近年来,随着基于图像的深度估计算法的发展[9-10]和硬件计算能力的提升,利用基于RGB图像来获取场景深度信息的方法逐渐成为研究热点。依据输入图像数量的差异,这些方法可分为单目和双目深度估计。其中双目深度估计通常建立在多视点特征匹配上,采用立体视觉匹配方法,使用双目相机获取成对的左右目图像,或者连续图像序列作为模型的输入,但是不仅对输入有较强的限制,还存在着标定过程复杂、数据处理时间过长等问题。传统的单目深度估计算法包括运动恢复结构信息法、光照恢复结构信息法等,这些算法主要从运动和阴影中恢复形状,深度预测精度不高。随着深度学习在图像领域的不断应用,深度学习强大的建模能力和对数据内在相关性深度学习的优势,使得单目图像的精确深度估计成为可能。因此,利用单目深度估计网络学习输入彩色图像的纹理、形状和空间布局特征,编码彩色信息和深度信息之间的映射关系,不需要昂贵的视觉传感器以及专业科研人员,就能够获得准确的深度图像,极大地降低了技术的应用成本。

基于深度学习的单目深度估计方法主要包括监督和自监督两种方式。自监督不需要真实的深度标签,而是利用连续帧之间的时序约束来训练深度模型[11-15]。然而,自监督学习面临的一个主要挑战是缺乏大规模和多样化的数据集,这限制了其提升精度的能力。监督学习方式基于编码器-解码器的结构也取得了极大的进展[16-22]。

传统卷积神经网络在提取特征的过程中,受卷积核大小的限制,模型仅考虑局部信息,缺乏对上下文全局特征的解析能力,导致场景深度估计精度较低;目前的深度估计网络在解析场景中纹理等几何信息时能力有限,这使得生成的深度图缺乏足够的细节表现,边缘模糊的问题较为突出。为解决上述问题,本研究提出一种嵌入注意力的RGB模态向深度模态转换的单目深度估计网络RDCN,实现温室

场景下番茄植株图像深度信息的精确预测。

1　材料和方法

1.1　RGBD 图像数据集构建

1.1.1　数据采集

Azure Kinect 是由微软发布的多功能 RGB－D 深度感应相机，专门为计算机视觉及其相关应用设计，搭载了 100 万像素的基于时间飞行算法（Time of flight，TOF）的深度传感器，能够进行精准的深度感知，捕捉环境中的三维空间结构信息。本研究使用该深度相机搭建日光温室环境下的番茄 RGBD 图像采集平台。深度镜头采用 WFOV 宽视场非装箱模式，镜头分辨率为 1 024 像素 ×1 024 像素，工作范围为 0.25～2.21 m，帧率为 15 f/s，彩色镜头分辨率为 1 920 像素 ×1 080 像素，帧率为 15 f/s。

为了使采集到的番茄 RGBD 图像数据集具有多样性，根据日光温室环境下光照强度的不同，分别在 08:00—11:00、12:00—15:00、16:00—19:00 3 个时间段采集番茄 RGB 图像与深度图像。数据采集过程中，始终保持相机的中心线与番茄植株垂直，并从多个角度、方位和视角对番茄植株进行拍摄，采集到的图像如图 1 所示。由于 RGB 镜头和深度镜头的物理位置和视角不同，导致采集的图像在空间上不能完全对齐，为了找到两者的映射关系，首先，通过官方提供的 Azure Kinect SDK 获取相机的内参和外参，包括深度相机和 RGB 相机之间的平移和旋转矩阵等；使用 SDK 的接口自动生成深度相机和 RGB 相机之间的标定参数；然后再次调用相关接口逐帧读取深度图像进行配准，使其对齐到 RGB 图像的坐标系中。最终获得逐像素一一对齐的 RGB 图像与深度图像共 8 515 组，其中光照强 1 500 组、正常光照 5 515 组、光照弱 1 500 组，分辨率均为 1 920 像素 ×1 080 像素。为了简化模型训练与评估过程，将采集到的所有 RGBD 图像进行中心裁剪，获取对齐后分辨率均为 640 像素 ×480 像素的 RGB 图像和深度图像。

图 1　不同角度的番茄植株图像

Fig. 1　Images of tomato plants from different angles

1.1.2　数据预处理及增强

NYU Depth V2 是最常用的室内场景 RGBD 数据集，提供了 Matlab 工具箱用于数据预处理，其中包括去噪、深度对齐、原始数据集深度补全等多功能工具集，由于光反射等因素，采集的深度图像通常包含一定的噪声，为了填补深度图像中的黑点，本研究采集的原始深度图像通过利用上述工具箱提供的交叉双边滤波算法进行深度补全，平滑小黑点，然后对深度图像进行可视化处理，利用 plasma 颜色方案对深度图像作可视化，处理后的数据集 RGBD 图像如图 2 所示，光照强度自上而下依次为强光照、正常光照、弱光照，与原始深度图像相比，补全后的深度图像深度值的分布更加连续，减少了原始深度图像的黑点和缺失区域，同时降低了噪声干扰，使得深度信息更加平滑。

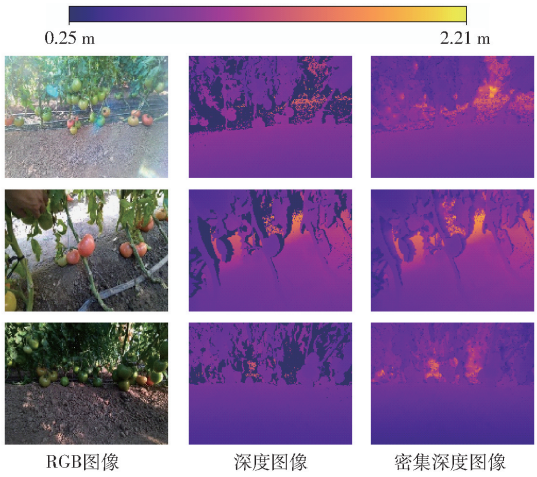

图 2　温室番茄 RGBD 图像

Fig. 2　Greenhouse tomato RGBD images

将数据集按照 8∶2的比例分成训练集 6 812 组、测试集 1 703 组，为了防止训练过程中出现过拟合，并提升模型的泛化能力，采用 3 种方式对训练集中的光照强和光照弱数据进行增强。①随机旋转：将 RGB 图像和深度图像在角度［－3°，3°］范围内同时旋转，概率设置为 50%。②色彩抖动：将 RGB 图像的亮度、对比度和饱和度按照比例因子［0.9，1.1］进行随机缩放。③水平镜像：RGB 图像和深度图像同时水平翻转，概率设置为 50%，将光照强和光照弱同时扩充至 4 200 组，最终共获得训练样本 12 812 组。在模型的训练过程中，从训练集随机选择 20% 的样本用于验证和模型调优。

1.2　RDCN 单目深度估计网络

针对传统的卷积池化操作会引起局部信息丢失，以及忽略深度图像边界特征的提取，本文提出 RDCN 单目深度估计网络，在 Lapdepth 模型框架的基础上，对解码器网络进行了改进与优化，具体结构如图 3 所示。网络首先使用 ResNext101 编码器将输入 RGB 图像通过深度堆叠的卷积块进行深度压缩，从而获得了蕴藏丰富信息的深度特征；解码过程

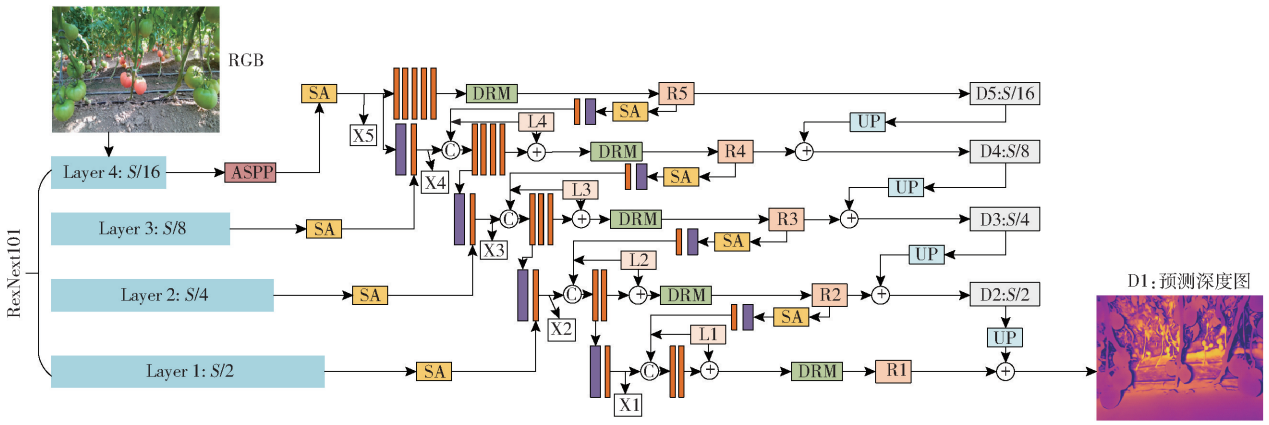

图 3 RDCN 单目深度估计模型

Fig. 3 RDCN monocular depth estimation model

中模型使用 4 层拉普拉斯金字塔,对应 Layer 1 ~ Layer 4,其中最高层用于预测深度图像的全局特征;其余 3 层融合编码特征和拉普拉斯残差。为了增强模型获取全局信息以及上下文信息交互的能力,引入了置换注意力模块,以帮助更好地恢复深度图像中的局部细节;为了更准确地捕捉边界信息,深度残差生成过程中融入了置换注意力模块;同时,在深度预测阶段,通过引入深度细化模块来自适应学习边界特征,强化模型的边界感知能力,最终完成深度图像的准确预测。

1.2.1 ResNext101 主干网络

本研究以 ResNext101[23] 作为模型的主干特征提取网络,相比于传统的 ResNet 系列,引入了组卷积以及模块化设计,不仅加深了网络的深度还保持着更高的性能,同时显著减少了参数量。与 Inception 系列相比有着更强的特征提取能力,且模型在结构上更为简单。在编码器端,输入的 RGB 图像经过深度堆叠的卷积块对特征进行大幅压缩,生成了 4 个不同尺度的特征图,通道数分别为 64、256、512 和 1 024,分辨率对应为原图的 1/2、1/4、1/8 和1/16。这些特征图富含嵌入空间中的颜色与深度信息之间的关系。

1.2.2 残差模块

如何提取边界信息,准确预测出深度图的边界特征非常重要,因此,本研究定义了两种残差:拉普拉斯残差 L_k、融合置换注意力模块的深度残差 R_k,来提取差异特征。在解码过程中,首先计算输入 RGB 图像的拉普拉斯残差 L_k,即

$$L_k = I_k - U_p(I_{k+1}) \quad (k=1,2,3,4) \tag{1}$$

式中 k——拉普拉斯金字塔中的层级索引号

$U_p(\cdot)$——上采样操作

拉普拉斯残差的结构如图 4 所示,利用 encoder 下采样过程中的特征图减去 decoder 上采样得到的特征图来获得深度边界信息,分别对各个层级做相同操作,提取不同尺度的差异,从而得到 L_k 引导解码过程精确恢复不同尺度空间的局部细节,准确解释深度边界,有效避免了预测深度图像的伪影模糊。I_k 通过对原始输入 RGB 图像进行向下采样 $1/2^{k-1}$ 得到。

假设 R_k 为第 k 层深度残差,生成深度残差的过程如下:将潜在特征 X_k、L_k 与上采样后的拉普拉斯金字塔 $(k+1)$ 获得的深度残差融合,随后,将这些拼接的特征输入到堆叠的卷积块中,并将相应的输出再次拼接到 L_k 中,此过程可表述为

$$R_k = B_k([X_k, L_k, R'_{k+1}]) + L_k \quad (k=1,2,3,4) \tag{2}$$

$$R'_{k+1} = \mathrm{Conv}(\mathrm{Up}(R_k), \mathrm{Up}(\mathrm{SA}(R_k))) \quad (k=2,3,4,5) \tag{3}$$

$$X_1 = \mathrm{Conv}(\mathrm{Up}(X_2, L_2, R'_3)) \tag{4}$$

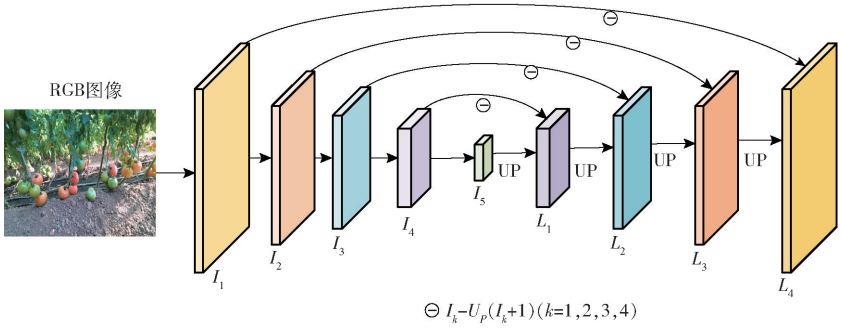

图 4 拉普拉斯残差结构图

Fig. 4 Laplacian residual structure diagram

$$X_2 = \mathrm{Conv}(\mathrm{Up}(X_3, L_3, R_4'), \mathrm{SA}(S/2)) \quad (5)$$

$$X_3 = \mathrm{Conv}(\mathrm{Up}(X_4, L_4, R_5'), \mathrm{SA}(S/4)) \quad (6)$$

$$X_4 = \mathrm{Conv}(\mathrm{Up}(X_5), \mathrm{SA}(S/8)) \quad (7)$$

式中　$[\cdot]$——拼接操作

B_k由多个卷积块组成，输出空间分辨率与 L_k相同的单通道图像。

最后，深度图像从拉普拉斯金字塔的最高层逐步重建，即

$$D_k = R_k + \mathrm{Up}(D_{k+1}) \quad (k = 1, 2, 3, 4) \quad (8)$$

$$D_5 = R_5 \quad (9)$$

式中　D_k——各层恢复的深度图像

D_5——金字塔最高首先恢复的深度全局图像

D_1——最终预测深度图像

1.2.3　置换注意力模块

日光温室环境下的番茄多成群出现，易出现重叠与遮挡；枝叶连结，且颜色相近，视觉较相近，不易区分；在估计深度的时候容易出现边界误差以及精度较低。为了解决这些问题，本文引入了置换注意力机制 SA[24]，其结构如图 5 所示，通过增强编码器提取的特征间位置关系，强化通道间表达，减少噪声干扰，提升网络对视觉相似的深度估计效果。

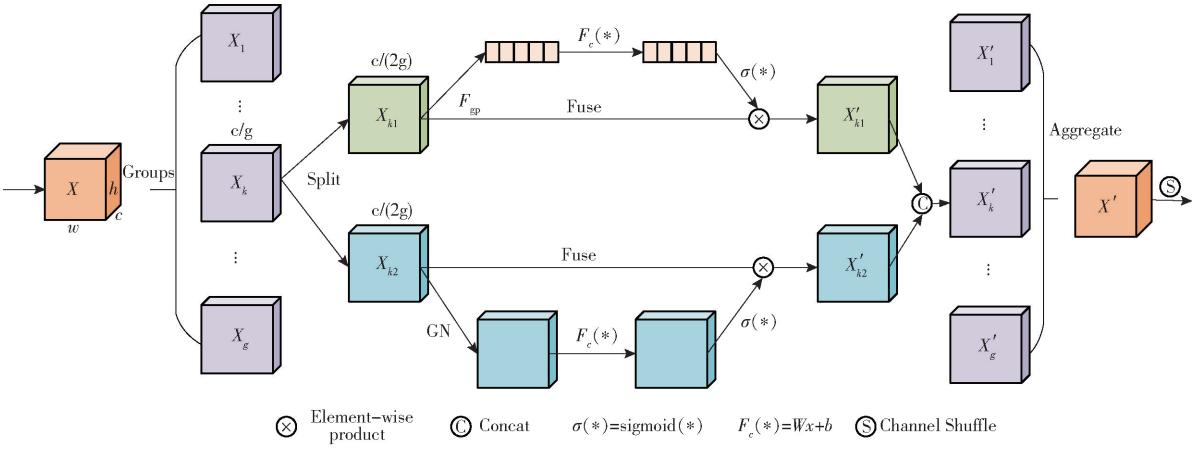

图 5　置换注意力模块

Fig. 5　Shuffle attention module

注意力机制能够使神经网络准确关注输入的所有相关特征，主要分为空间注意力和通道注意力两种机制，由于仅仅对两种机制简单融合，而没有进一步优化两种机制的互补信息，如 GAM[25]、CA[26]、TAM[27]，虽然能够一定程度上提高模型的预测精度，但大大增加了计算开销以及丢失了跨通道信息。SA 可以有效捕捉输入特征图中任意位置之间的依赖关系，并且自适应计算特征间的相似性，不需要固定的感受野，增强了特征之间的全局交互能力。

SA 设计了 Shuffle Unit，来表示特征空间和通道维度上的特征依赖关系，对输入的特征首先沿通道维度划分为 g 组，其中每个子特征 X_k在训练过程中逐渐捕获特定的语义信息，然后生成对应的重要度系数。具体来说，X_k沿通道维度划分为两个分支 X_{k1}和 X_{k2}，X_{k1}利用通道间的关系生成通道注意力图，X_{k2}利用特征的空间关系生成空间注意力图；为了得到一个轻量级的模块，仅使用全局平均池化(GAP)来嵌入全局信息，以提取通道维度上的全局信息，即

$$F_{gp}(x_{k1}(i,j)) = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} x_{k1}(i,j) \quad (10)$$

利用 sigmoid 获得了一个高分辨率的特征，以提供精确和自适应指导，通道注意力部分的输出为

$$X_{k1}' = \delta(w_1 F_{gp}(X_{k1}(i,j)) + b_2) X_{k1} \quad (11)$$

与通道注意力不同，空间注意力关注的是特征位置，与通道注意力互补，首先使用 Group Norm (GN)来获得空间统计信息，然后使用激活函数(sigmoid)来增强 X_{k2}的特征表示，最终输出为

$$X_{k2}' = \delta(w_1 \mathrm{GN}(X_{k2}) + b_2) X_{k2} \quad (12)$$

最后将这两个分支的输出特征进行通道级的拼接融合，通过 Channel Shuffle 算子实现了特征跨组交流信息，最终输出的大小与输入相同；相较于不引入置换注意力 SA，在几乎不增加计算开销的情况下，增强了网络的特征表达能力，使网络能够更加专注于区分番茄遮挡以及枝叶重叠，增强网络对其边界信息的分析感知能力，通过捕获空间和通道维度上的特征依赖关系，显著提高了网络深度预测的性能。

1.2.4　深度细化模块

在深度图像预测过程中，传统方法采用若干个 5×5 的卷积，但由于小核卷积的感受野有限，这种卷积只能聚集每个像素的局部特征，难以充分利用全局上下文信息，而导致估计出的深度图像边界模糊；而真实的成熟番茄采摘时需要精确的深度信息，

因此本文融合了深度细化模块 DRM[28],其结构如图 6 所示;使用两个条纹卷积在水平和垂直方向聚集边界信息的像素,卷积核大小分别为 3×11 和 11×3,由于沿着正交方向的全局上下信息对表达相对深度有着显著作用,因此更好地识别遮挡番茄和重叠枝叶之间的深度变化;然后,利用 3×3 的卷积对两次条纹卷积提取的特征进行融合,再采用 5×5 的卷积来细化最终的深度图像,最终,为了更准确预测深度图像,通过跳跃连接将卷积前的融合特征传递到最后一个卷积。

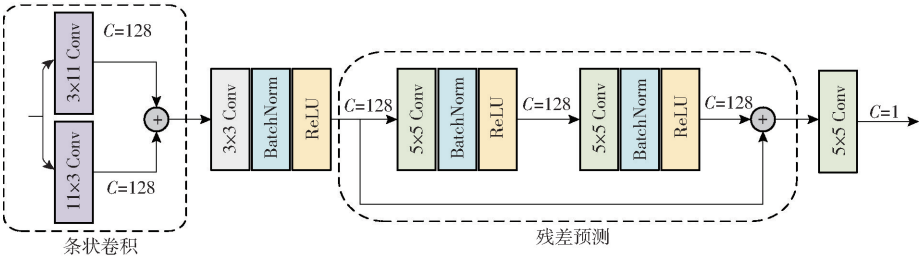

图 6 深度细化模块

Fig. 6 Depth refinement module

相比于传统的卷积直接估计深度,本研究在进行最终深度预测前先明确特征边界附近的深度变化,并挖掘细节特征;较好地解决了预测深度图像中番茄植株边界模糊问题。

1.2.5 损失函数

网络训练过程使用损失函数 L 进行优化,由数据损失 L_d 和梯度损失 L_g 组成,即

$$L=\alpha L_d(y,y^*)+\beta L_g(y,y^*) \tag{13}$$

式中 y、y^*——预测深度和真实深度

α、β——L_d、L_g 的平衡因子,设置为 10 和 0.1

由于深度相机自身的限制,深度数据在近距离时较为密集,而在远距离时则变得稀疏;为了降低背景等较深区域对预测结果的影响,利用文献[16]中提出的损失函数的平方根作为数据损失,用于计算预测深度与真实深度在对数空间的差异,即

$$L_d(y,y^*)=\sqrt{\frac{1}{n}\sum_{i\in V}d_i^2-\frac{\lambda}{n^2}\left(\sum_{i\in\nu}^{N_v}d_i\right)^2} \tag{14}$$

其中 $$d_i=\lg y_i-\lg y_i^*$$

式中 V——深度图中的有效像素集合

N_v——有效像素总数

λ——平衡因子,取 0.85

为了增强局部细节,特别是在深度边界处,定义深度图的梯度损失来减轻边界模糊现象,即

$$L_g(y,y^*)=\frac{1}{N}\sum_i^N|y_{h,i}-m(y^*)_{h,i}|+|y_{v,i}-m(y^*)_{v,i}| \tag{15}$$

式中 $m(\cdot)$——文献[29]中使用的插值函数

$y_{h,i}$、$m(y^*)_{h,i}$——水平方向上预测深度图和插值后的真实深度图第 i 个像素梯度

$y_{v,i}$、$m(y^*)_{v,i}$——垂直方向上预测深度图和插值真实深度图第 i 个像素梯度

N——深度图像素总数

2 结果与分析

2.1 评估指标

为评估本网络深度预测的性能,使用 7 个指标来评估预测深度图像的完整性和准确性,这些指标分别为:阈值准确率 $\delta_i(i=1,2,3)$、平均相对误差(Absoluter relative error, Abs Rel)、对数均方根误差(Root mean squared logarithmic error, $\mathrm{RMSE_{log}}$)、均方根误差(Root mean squared error, RMSE)、对数平均误差(Relative logarithmic error)。其中,阈值准确度越高表明网络精度越高,误差越小表明网络精度越高。

2.2 硬件平台以及训练策略

在模型的训练过程中,需要对不同改进进行消融评估,为保证训练、评估的公平性以及训练的效率,本试验将模型的所有训练任务放到同一硬件平台上。在深度学习工作站上完成了训练任务,其内存为 128 GB,中央处理器主频为 4.8 GHz,中央处理器核心数为 18,图形处理器(Graphic processing unit, GPU)为 NVIDIA GeForce RTX 3090,图形处理器显存为 24 GB,图形处理器核心数为 10 752。使用 AdamW 作为训练优化器,功率和动量分别为 0.9 和 0.999,编码器权重的衰减率设为 0.000 5,解码器部分则无权重衰减,初始学习率为 0.000 1,然后使用幂为 0.5 的多项式衰减到 0.000 01,共训练 35 个周期,每一次迭代的批量大小设置为 8。模型总参数量为 7.35×10^7,其中编码器和解码器分别占 5.79×10^7 和 1.55×10^7。

2.3　不同主干网络对比试验分析

为了评估不同特征提取主干网络对预测深度图像的影响，本研究在保持其他试验条件恒定的前提下，仅替换了网络的特征提取主干网络。具体来说，将本研究采用的 ResNext101 与目前使用较为广泛的主流主干网络，如 InceptionV3[30]、DenseNet161[31]以及 ResNet101[32]相比，在相同的超参数设置下进行训练直至模型收敛。随后，在同一测试集上进行了性能评估，各评估指标的结果如表 1 所示。根据表 1 中的数据，可以观察到在控制试验条件一致的情况下，采用 ResNext101 作为特征提取主干网络相较于 InceptionV3、DenseNet161 和 ResNet101，均方根误差(RMSE)分别降低了 6.15%、2.40% 和 3.17%。ResNext101 相比于 ResNet101 引入了分组卷积，能够更好地捕捉丰富的特征，在参数量相对不变的情况下提升了模型的性能，并且简化了 InceptionV3 中的分支设计方法，因此实现了比 InceptionV3 在结构上更加简单同时有着更好的性能，与 DenseNet161 相比有着更小的模型参数量以及更小的计算量，证明了 ResNext101 更加适用于本网络。

表 1　不同主干网络的评估结果

Tab. 1　Evaluation results of different backbones

主干网络	阈值准确率			平均相对误差	均方根误差/m	对数均方根误差	对数平均误差
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$				
InceptionV3	0.877	0.943	0.966	0.166	0.130	0.270	0.050
DenseNet161	0.885	0.945	0.967	0.154	0.125	0.261	0.046
ResNet101	0.883	0.945	0.967	0.154	0.126	0.263	0.046
ResNext101	0.891	0.952	0.975	0.139	0.122	0.253	0.043

可视化深度图像如图 7 所示，无论是番茄果实比较集中、植株枝叶较为密集、还是有障碍物的场景中，RDCN 均可以获得相当清晰准确的深度图像，这为基于单目传感器的温室场景深度估计视觉系统的设计提供了帮助。

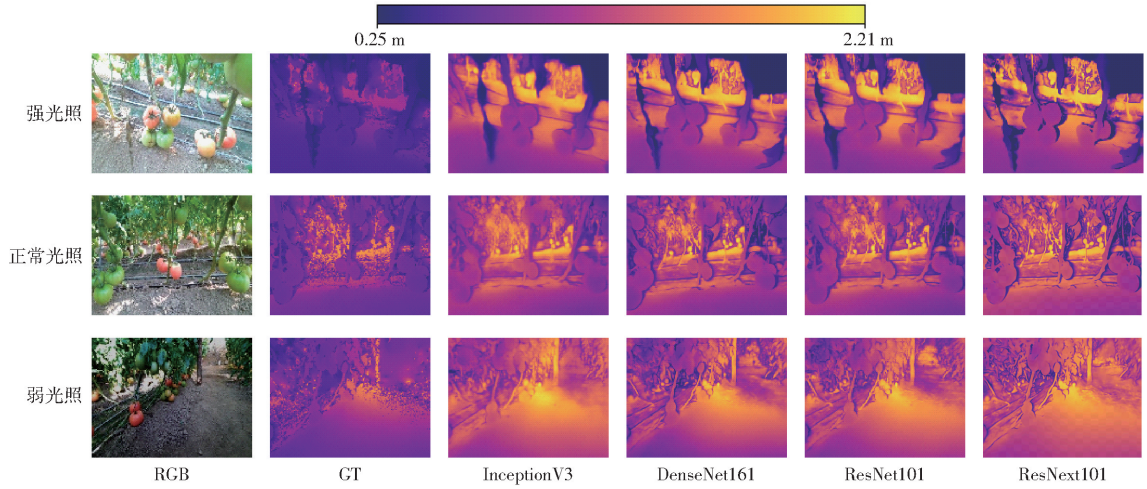

图 7　不同主干网络的深度预测结果

Fig. 7　Depth prediction results of different backbones

2.4　不同引入模块的对比试验分析

为了验证本研究引入的置换注意力模块 SA 和深度细化模块 DRM 的有效性，在控制其他条件不变的情况下，对不同引入模块后的网络进行了消融训练。训练完成后，各网络在平均相对误差、均方根误差、对数均方根误差、对数平均误差以及在不同阈值下的准确率结果如表 2 所示。

表 2　不同模块的评估结果

Tab. 2　Evaluation results of different modules

改进网络	阈值准确率			平均相对误差	均方根误差/m	对数均方根误差	对数平均误差
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$				
Baseline	0.863	0.941	0.965	0.175	0.136	0.276	0.055
Baseline + SA	0.885	0.945	0.967	0.158	0.125	0.262	0.048
Baseline + DRM	0.882	0.945	0.966	0.171	0.129	0.271	0.051
RDCN	0.891	0.952	0.975	0.139	0.122	0.253	0.043

从表 2 中可以看出引入置换注意力模块后均方根误差下降了 8.09%，引入深度细化模块后均方根误差下降了 5.15%，同时引入两种模块后，均方根误差下降了 10.3%，深度估计网络取得了最准确的精度，试验结果充分证明了 SA 和 DRM 对于提升深度估计网络性能的有效性。

为了研究不同改进对网络预测性能的影响，图 8 展示了不同引入模块的预测深度图像与真实深度图像对比。从可视化结果可见，基础模型重建的深度图像呈现出较为粗糙的特征，具有显著的深度失真和细节模糊现象。当引入 SA 注意力模块后，深度图像的清晰度显著提升，深度失真和模糊现象有所减少；通过嵌入深度细化模块 DRM 后，能够更为准确地恢复图像深度信息，其分辨率也得到增强。而本研究提出的单目深度估计网络 RDCN 所生成的深度图像轮廓清晰可见，整体效果部分超越了真实深度图像。

此外，为了评估不同光照条件对深度信息预测

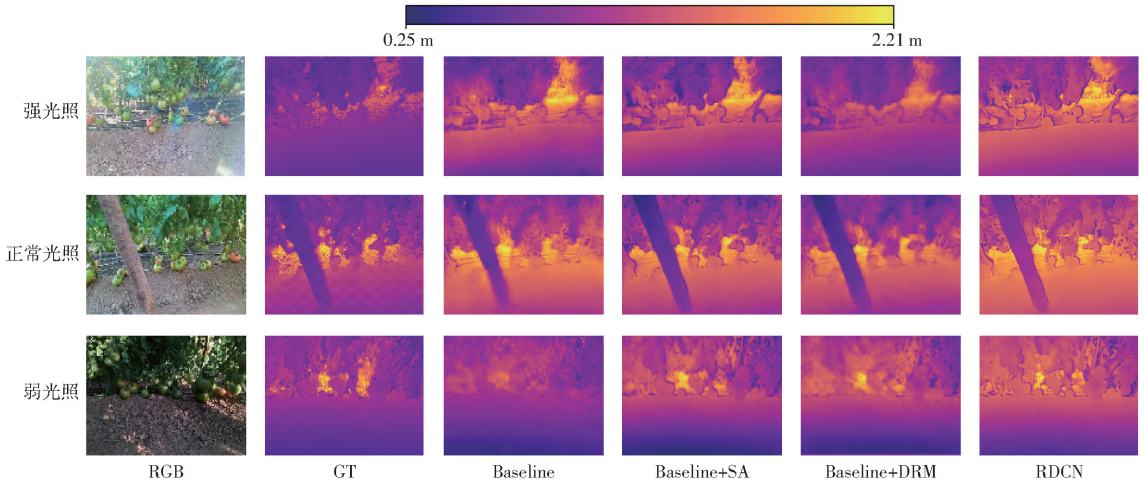

图 8 不同引入模块的深度预测结果

Fig. 8 Different prediction results of introducing modules

的影响，选取了强光、正常光照以及弱光条件下的深度图像进行分析。从图 8 可以看出，强光条件下，尽管 RGB 图像的亮度和对比度较高，SA 能够使模型将注意力集中在关键区域，从而在亮度过度的区域减少误差，保持较好的深度预测效果；正常光照条件下，SA 帮助模型充分利用图像的全局和局部信息，实现最佳的深度预测效果；弱光条件下，模型利用 SA 对阴影区域和低亮度区域进行合理处理，确保预测结果在噪声较多的环境中仍然保持较好的稳定性和一致性。试验表明，SA 注意力机制通过自适应调整模型对不同图像区域和通道的关注度，有效减轻了光照变化对深度估计的影响，增强了模型在复杂场景中的特征提取能力和光照适应性。

2.5 不同模型对比试验分析

为了进一步证明本研究提出的 RDCN 在复杂日光温室的有效性和鲁棒性，将其与目前主流的模型 Newcrfs[33]、Binsformer[34]、Bts[35] 进行试验对比，深度图像预测的评估结果如表 3 所示。

表 3 不同模型的评估结果

Tab. 3 Evaluation results of different models

模型	阈值准确率			平均相对误差	均方根误差/m	对数均方根误差	对数平均误差
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$				
Newcrfs	0.875	0.943	0.965	0.173	0.133	0.273	0.053
Binsformer	0.882	0.946	0.966	0.162	0.128	0.266	0.048
Bts	0.874	0.942	0.966	0.161	0.130	0.267	0.049
RDCN	0.891	0.952	0.975	0.139	0.122	0.253	0.043

从表 3 可以看出，在相同试验条件下，本研究提出的 RDCN 均方根误差相比于 Newcrfs、Binsformer、Bts 分别降低了 8.27%、4.69%、6.15%，验证了本研究所提算法的优越性。

可视化的深度预测图像如图 9 所示，展示了不同视角下的番茄植株图像，与其它 3 个模型相比，RDCN 有着更高的预测精度，更清晰的深度边缘，Newcrfs 生成的深度图像在物体边界处有着微小的连结，很难分辨，Binsformer、Bts 生成的深度图在局部有着微弱的模糊很难辨别。本研究提出的 RDCN 网络预测的深度图像连续性更好，局部和物体边界更清晰，番茄整体形状也更完整清晰，效果更好，证

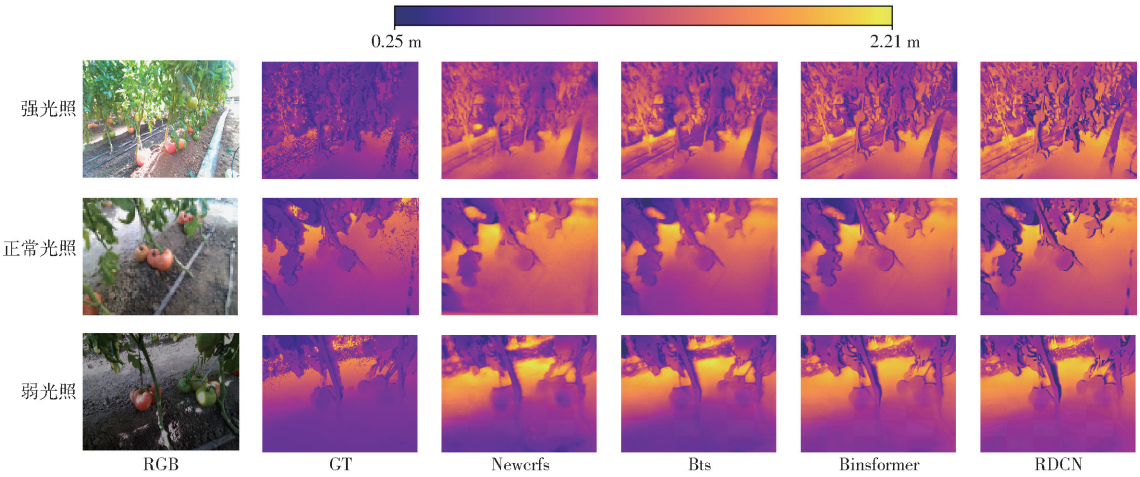

图 9　不同模型的深度预测结果

Fig. 9　Depth prediction results of different models

明了本研究提出的单目深度网络 RDCN 有着更高的精度，并且还有着更好的鲁棒性。

3　结论

（1）通过引入置换注意力模块以及深度细化模块，实现了对温室场景深度信息精确的预测；与改进前基准模型相比，从可视化深度图明显可以看出在复杂几何形状以及具有显著深度变化的区域时，表现出了较好的视觉效果，评估指标均方根误差降低了 10.3%、平均相对误差降低了 20.5%、对数均方根误差降低了 8.3%、对数平均误差降低了 21.8%，在 1.25、1.25^2、1.25^3 阈值下的准确率分别提高了 3.2%、1.2% 和 1.0%；充分验证了本研究算法的有效性。

（2）通过嵌入置换注意力模块，强化了全局特征捕获能力，增强了网络对深度边界信息的感知，提高了特征融合的深度，相比于不引入该模块，均方根误差降低了 8.09%，证明了置换注意力模块对提升深度估计性能的有效性。

（3）通过引入深度细化模块，优化了特征图的边界细节信息，解决了边界模糊问题，相比于不引入该模块，均方根误差降低了 5.15%，证明了深度细化模块对提高深度预测精度的有效性。

（4）验证了不同光照强度下深度估计的结果，试验表明本算法面对光照强度变化依然能获得准确的深度信息，具有良好的鲁棒性。

（5）针对日光温室环境下的番茄图像深度估计任务，本研究提出的单目深度估计网络 RDCN 可以实现这一复杂任务，同时可为基于单目传感器的温室场景农机导航以及深度图像在多模态任务中的应用提供技术支持。

参　考　文　献

[1] 宗泽，郭彩玲，张雪，等. 基于深度相机的玉米株型参数提取方法研究[J]. 农业机械学报，2015，46(增刊)：50－56.
ZONG Ze, GUO Cailing, ZHANG Xue, et al. Maize plant type parameters extraction based on depth camera[J]. Transactionsof the Chinese Society for Agricultural Machinery, 2015, 46(Supp.): 50－56. (in Chinese)

[2] COLL-RIBES G, TORRES-RODRIGUEZ I J, GRAU A, et al. Accurate detection and depth estimation of table grapes and peduncles for robot harvesting, combining monocular depth estimation and CNN methods[J]. Computers and Electronics in Agriculture, 2023, 215: 108362.

[3] 胡广锐，孔微雨，齐闯，等. 果园环境下移动采摘机器人导航路径优化[J]. 农业工程学报，2021，37(9)：175－184.
HU Guangrui, KONG Weiyu, QI Chuang, et al. Optimization of the navigation path for a mobile harvesting robot in orchard environment[J]. Transactions of the CSAE, 2021, 37(9): 175－184. (in Chinese)

[4] 仇瑞承，苗艳龙，季宇寒，等. 基于 RGB－D 相机的单株玉米株高测量方法[J]. 农业机械学报，2017，48(增刊)：211－219.
QIU Ruicheng, MIAO Yanlong, JI Yuhan, et al. Measurement of individual maize height based on RGB－D camera[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.): 211－219. (in Chinese)

[5] 赵永杰，蒲六如，宋磊，等. 基于 Shuffle－ZoeDepth 单目深度估计的苗期[J]. 农业机械学报，2024，55(5)：235－243，253.
ZHAO Yongjie, PU Liuru, SONG Lei, et al. Measurement method of seedling stage maize height based on Shuffle－ZoeDepth monocular depth estimation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(5): 235－243, 253. (in Chinese)

[6] 徐胜勇，易同舟，秦子轶，等. 基于单视角 RGBD 图像的柑橘果实三维重建与表型检测方法[J]. 农业机械学报，2025，56(3)：80－90.
XU Shengyong, YI Tongzhou, QIN Ziyi, et al. Three-dimensional reconstruction and phenotype detection of citrus fruits based

on single-view RGBD images[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(3): 80 - 90. (in Chinese)

[7] 王纪章,顾容榕,孙力,等. 基于 Kinect 相机的穴盘苗生长过程无损监测方法[J]. 农业机械学报,2021,52(2):227 - 235.
WANG Jizhang, GU Rongrong, SUN Li, et al. Non-destructive monitoring of plug seedling growth process based on Kinect camera[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(2): 227 - 235. (in Chinese)

[8] 麦春艳,郑立华,孙红,等. 基于 RGB - D 相机的果树三维重构与果实识别定位[J]. 农业机械学报,2015,46(1):35 - 40.
MAI Chunyan, ZHENG Lihua, SUN Hong, et al. Research on 3D Reconstruction of fruit tree and fruit recognition and location method based on RGB - D camera[J]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(1): 35 - 40. (in Chinese)

[9] LI Y, WEI X, FAN H. Attention mechanism used in monocular depth estimation: an overview[J]. Applied Sciences, 2023, 13(17): 9940.

[10] ZHAO C, SUN Q, ZHANG C, et al. Monocular depth estimation based on deep learning: an overview[J]. Science China Technological Sciences, 2020, 63(9): 1612 - 1627.

[11] ZHANG N, NEX F, VOSSELMAN G, et al. Lite-mono: lightweight cnn and transformer architecture for self-supervised monocular depth estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18537 - 18546.

[12] 周云成,邓寒冰,许童羽,等. 基于稠密自编码器的无监督番茄植株图像深度估计模型[J]. 农业工程学报, 2020, 36(11): 182 - 192.
ZHOU Yuncheng, DENG Hanbing, XU Tongyu, et al. Unsupervised deep estimation modeling for tomatoplant image based on dense convolutional auto-encoder[J]. Transactions of the CSAE, 2020, 36(11): 182 - 192. (in Chinese)

[13] GODARD C, MAC O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 270 - 279.

[14] ZHAO C, POGGI M, TOSI F, et al. GasMono: Geometry-aided Self-supervised Monocular depth estimation for indoor scenes [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 16209 - 16220.

[15] 刘佳,林潇,陈大鹏,等. 轻量化自监督单目深度估计[J]. 计算机系统应用, 2023,32(8):116 - 125.
LIU Jia, LIN Xiao, CHEN Dapeng, et al. Lightweight self-supervised monocular depth estimation[J]. Computer Systems Applications, 2023,32(8):116 - 125. (in Chinese)

[16] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]// Proceedings of the 27th International Conference on Neural Information Processing System, 2014: 2366 - 2374.

[17] XU D, RICCI E, OUYANG W, et al. Multi-scale continuous crfs as sequential deep networks for monocular depthestimation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5354 - 5362.

[18] 龙燕, 高研, 张广犇. 基于改进 HRNet 的单幅图像苹果果树深度估计方法[J]. 农业工程学报, 2022, 38(23): 122 - 129.
LONG Yan, GAO Yan, ZHANG Guangben. Depth estimation of apple tree in single image using improved HRNet[J]. Transactions of the CSAE, 2022, 38(23):122 - 129. (in Chinese)

[19] SONG M, LIM S, KIM W. Monocular depth estimation using laplacian pyramid-based depth residuals[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(11): 4381 - 4393.

[20] FU H, GONG M, WANG C, et al. Deep ordinal regression network for monocular depth estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2002 - 2011.

[21] 凌莉,胡佑璞,邹承明. 基于多层次 CNN 特征融合的深度估计算法[J]. 计算机应用与软件, 2023, 40(10): 291 - 296.
LING Li, HU Youpu, ZOU Chengming. Depth estimation algorithm based on multi-level CNN feature fusion[J]. Computer Applications and Software, 2023, 40(10): 291 - 296. (in Chinese)

[22] LI Z, CHEN Z, LIU X, et al. Depthformer: exploiting long-range correlation and local information for accurate monocular depth estimation[J]. Machine Intelligence Research, 2023, 20(6): 837 - 854.

[23] XIE S, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1492 - 1500.

[24] ZHANG Q L, YANG Y B. SA-net: shuffle attention for deep convolutional neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2235 - 2239.

[25] LIU Y, SHAO Z, HOFFMANN N. Global attention mechanism: retain information to enhance channel-spatial interactions[J]. arXiv Preprint, arXiv: 2112.05561, 2021.

[26] 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 Recognitio, 2021: 13713 - 13722.

[27] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate toattend: convolutional triplet attention module[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 3139 - 3148.

[28] XUE F, CAO J, ZHOU Y, et al. Boundary-induced and scene-aggregated network for monocular depth prediction[J]. Pattern Recognition, 2021, 115: 107901.

(下转第 574 页)

[17] 段璞, 刘春江. RSSI 在 RFID 读写器中的应用[J]. 计算机工程, 2010, 36(22): 289 - 290.
DUAN Pu, LIU Chunjiang. Application of received signal strength indicator in RFID reader[J]. Computer Engineering, 2010, 36(22): 289 - 290. (in Chinese)
[18] 李昂, 章勇, 张瑞祥, 等. 基于 RSSI 的弦相交加权质心定位算法[J]. 机电工程技术, 2024, 53(1): 171 - 175.
LI Ang, ZHANG Yong, ZHANG Ruixiang, et al. String intersection weighted centroid location algorithm based on RSSI[J]. Electrical Engineering TEC, 2024, 53(1): 171 - 175. (in Chinese)
[19] 张宏伟, 周志强. 基于双边与高斯滤波混合分解的图像融合方法[J]. 系统工程与电子技术, 2016, 38(1): 8 - 13.
ZAHNG Hongwei, ZHOU Zhiqiang. Image fusion based on a hybrid decomposition via bilateral and Gaussian filters[J]. Systems Engineering and Electronics, 2016, 38(1): 8 - 13. (in Chinese)
[20] 胡学平, 王柳柳, 杨瑞, 等. NSD 随机阵列加权和最大值的收敛性[J]. 吉林大学学报, 2021, 59(5): 1101 - 1106.
HU Xueping, WANG Liuliu, YANG Rui, et al. Convergence properties for maximum weighted sums of NSD random arrays[J]. Journal of Jilin University, 2021, 59(5): 1101 - 1106. (in Chinese)
[21] KHAN A, BANGASH J I. Cuckoo search-based SVM (CS - SVM) model for real-time indoor position estimation in IoT networks[J]. Security and Communication Networks, 2021, 2021: 1 - 7.
[22] 张建华. 一种简便易用的基尼系数计算方法[J]. 山西农业大学学报, 2007, 6(3): 275 - 278.
ZHANG Jianhua. An convenient method to calculate Gini coefficient[J]. Journal of Shanxi Agricultural University, 2007, 6(3): 275 - 278. (in Chinese)
[23] 夏晶晶. 基于 KNN 神经网络的无人机作业操控系统研究[J]. 农机化研究, 2024, 46(7): 183 - 187.
XIA Jingjing. Research on UAV operation control system based on KNN neural network[J]. Journal of Agricultural Mechanization Research, 2024, 46(7): 183 - 187. (in Chinese)
[24] 张创基, 林伟烜, 陈运胜. 基于 KNN 算法的计算机网络入侵数据识别方法[J]. 长江信息通信,2023, 36(8): 120 - 122.
[25] 王建琴. 基于改进决策树的数据挖掘与分析算法设计[J]. 电子设计工程, 2024, 32(4): 84 - 88.
WANG Jianqin. Design of data mining and analysis algorithm based on improved decision tree[J]. Electronic Design Engineering, 2024, 32(4): 84 - 88. (in Chinese)
[26] 徐颖晋, 庞振宇. 基于改进支持向量机的致密砂岩储层参数预测研究[J]. 现代电子技术, 2024, 47(5): 132 - 138.
XU Yingjin, PANG Zhenyu. Research on tight sandstone reservoir parameter prediction based on improved support vector machine[J]. Modern Electronics Technique, 2024, 47(5): 132 - 138. (in Chinese)
[27] 刘潭, 李子默, 冯帅, 等. 基于 LMPSO - SVM 的高光谱水稻稻瘟病害分级检测[J]. 农业机械学报, 2023,54(11): 208 - 216, 235.
LIU Tan, LI Zimo, FENG Shuai, et al. Classification detection of hyperspectral rice blast disease based on LMPSO - SVM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(11): 208 - 216, 235. (in Chinese)
[28] LI Lihua, ZHAO Yang, OLIVEIRA J, et al. A UHF RFID system for studying individual feeding and nesting behaviors of group-housed laying hens[J]. Transactions of the ASABE, 2017, 60(4): 1337 - 1347.
[29] CARMICHAEL N L, WALKER W, HUGHES B O, et al. Laying hens in large flocks in a perchery system: influence of stocking density on location, use of resources and behaviour[J]. British Poultry Science, 1999, 40(2): 165 - 176.
[30] ALI A B A, CAMPBELL D L M, KARCHER D M, et al. Influence of genetic strain and access to litter on spatial distribution of 4 strains of laying hens in an aviary system[J]. Poultry Science, 2016, 95(11): 2489 - 2502.

(上接第 508 页)

[29] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]// The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. IEEE, 2003: 1398 - 1402.
[30] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition, 2016: 770 - 778.
[31] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818 - 2826.
[32] 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.
[33] YUAN W, GU X, DAI Z, et al. Neural window fully-connected crfs for monocular depth estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 3916 - 3925.
[34] LI Z, WANG X, LIU X, et al. Binsformer: Revisiting adaptive bins for monocular depth estimation[J]. arXiv Preprint, arXiv:2204.00987, 2022.
[35] LEE J H, HAN M K, KO D W, et al. From big to small: multi-scale local planar guidance for monocular depth estimation[J]. arXiv Preprint,arXiv:1907.10326, 2019.