

基于图文定位生成的育秧盘内密集稻种分割计数方法

欧阳盟¹, 邹荣^{1*}, 陈进¹, 李耀明², 陈宇航¹, 严昊¹

(1. 江苏大学机械工程学院, 江苏镇江 212013, 中国; 2. 江苏大学农业工程学院, 江苏镇江 212013, 中国)

摘要: [目的/意义] 随着工厂化育秧的快速发展, 育秧环节的智能化已成为提升水稻生产效率与品质的关键。育秧盘中穴孔内稻种数量的准确识别, 直接影响气振式育苗精密播种装置的作业效率与参数优化。然而, 在复杂环境下, 穴孔内稻种检测精度低, 难以实现单粒精确分割。[方法] 提出一种融合图文定位生成 (Caption Grounding and Generation, CGG) 模型与预训练模型的稻种实例分割方法, 通过图像与文本特征的联合对齐, 实现目标定位与语义理解的协同学习, 从而显著提升了稻种的检测与分割精度。[结果和讨论] 消融实验结果表明, 核心改进模块基于自举式语言-图像预训练模型生成的伪标签, 以及基于双向编码器表示的 Transformer 的词向量嵌入, 均能够提升模型性能。二者结合时表现出协同效应, 使分割准确率较基线提升超过 3 个百分点。所提 CGG 模型在交并比阈值为 0.5 时, 边界框检测平均精度达到 90.7%, 实例分割平均精度达到 91.4%, 显著优于 Mask R-CNN 和 Mask2Former 等主流模型。进一步在播种场景验证中, CGG 模型在单穴种子数量检测精度达 88%, 表现最优。其针对单粒稻种的误差指标, 包括均方根误差 16.8 粒、平均绝对误差 13.7 粒和平均绝对百分比误差 2.46%, 均显著低于对比模型。[结论] 该模型具备实时在线检测稻种数量的能力, 为后续需补种穴位的精准再播种提供了可量化操作依据, 具有良好的实际应用前景, 可为智慧农业和智能播种技术的发展提供有力支持。

关键词: 水稻播种; BERT; 实例分割; 预训练模型; 视觉语言模型; 图文定位生成; 大模型

中图分类号: S24

文献标识码: A

文章编号: SA202507030

引用格式: 欧阳盟, 邹荣, 陈进, 李耀明, 陈宇航, 严昊. 基于图文定位生成的育秧盘内密集稻种分割计数方法[J]. 智慧农业(中英文), 2026, 8(2): 188-199. DOI: 10.12133/j.smartag.SA202507030

OUYANG Meng, ZOU Rong, CHEN Jin, LI Yaoming, CHEN Yuhang, YAN Hao. CGG-Based Segmentation and Counting of Densely Distributed Rice Seeds in Seedling Trays[J]. Smart Agriculture, 2026, 8(2): 188-199. DOI: 10.12133/j.smartag.SA202507030 (in Chinese with English abstract)

0 引言

水稻是中国主要粮食作物之一, 2024 年种植总面积约达 2 900.7 万 hm^2 [1]。在传统水稻栽培模式中, 育秧移栽因其能够提高成苗率、便于田间管理与保证产量稳定, 而成为主要的栽培方式。其中, 秧苗培育环节对最终产量与品质具有关键作用。随着精准农业技术的发展, 播种精度的要求日益提高, 育秧盘播种质量检测逐渐成为调控气振式育苗精密播种装置播种量, 以及实现精准播种的基础环节 [2]。该检测技术为精准农业的推广和实施提供了重要支撑, 因此, 具有显著的研究与应用价值。

近年来, 图像处理技术在水稻秧盘播种检测领

域得到了广泛应用。现有研究主要分为两类: 一是基于传统机器视觉的方法; 二是基于深度学习的方法 [3]。传统机器视觉技术通常依赖阈值分割提取目标区域, 并结合形态学运算进行后处理。例如, DONG 等 [4] 采用色相-饱和度-亮度 (Hue-Saturation-Lightness, HSL) 色彩空间模型进行分割; BAI 等 [5] 提出基于投票机制的分割方法。然而, 这类方法易出现籽粒特征丢失, 难以处理密集重叠的籽粒, 导致漏检与误检率较高。相比之下, 深度学习方法在复杂场景下展现出更好的适应性。WANG 等 [6] 提出改进 YOLOv5s 算法实现秧盘稻种检测。YAN 等 [7] 利用 YOLOv5x 模型对番茄穴盘种子进行

收稿日期: 2025-07-21

基金项目: 国家自然科学基金项目 (31871528)

作者简介: 欧阳盟, 硕士研究生, 研究方向为物体实例分割。E-mail: amen610401404@163.com

*通信作者: 邹荣, 副教授, 研究方向为深度学习、人工智能与机器视觉等。E-mail: zr@ujs.edu.cn

copyright©2026 by the authors

(登录 www.smartag.net.cn 免费获取电子版全文)

识别，并建立了漏播检测与补种定位方法。

实例分割技术^[8]能够在识别物体类别的同时精准勾勒其边界，即使在密集分布的情况下，也能有效克服基于锚框的目标检测方法在重叠目标上的不足，从而提升检测精度。作为目标检测^[9]与语义分割^[10]的结合，实例分割方法实现了对象类别识别与像素级分割的统一。典型代表，如掩码区域卷积神经网络（Region-Based Convolutional Neural Networks for Instance Segmentation, Mask R-CNN）^[11]在快速区域卷积神经网络（Region-based Convolutional Neural Networks, Faster R-CNN）^[12]基础上引入掩膜分支，实现目标检测向实例分割的拓展。该方案虽保留两阶段检测框架的高精度特性，但仍存在检测速度较慢的问题。随着Transformer模型^[13]在计算机视觉领域的崛起，Mask2Former^[14]在MaskFormer^[15]架构中融入掩膜增强型Transformer模块，通过基于查询的机制直接生成分割掩膜，摆脱区域建议框依赖，在精度与速度层面实现双重突破。然而其参数化查询机制仍受训练数据语义范畴制约，对未见类别泛化能力有限。为满足更高精度与鲁棒性需求，视觉语言模型^[16]通过融合视觉-文本信息提升实例分割性能，为后续研究开辟新范式。

目前，已有学者将实例分割算法应用于种子分割任务。TODA等^[17]、YE等^[18]与GAO等^[19]利用实例分割技术对图像中种子进行分割并提取其表型性状，然而，这类研究多在背景单一、籽粒特征相对单纯的条件下开展。与此同时，不少研究聚焦于复杂环境中的作物实例分割^[20-22]，尽管更贴近实际

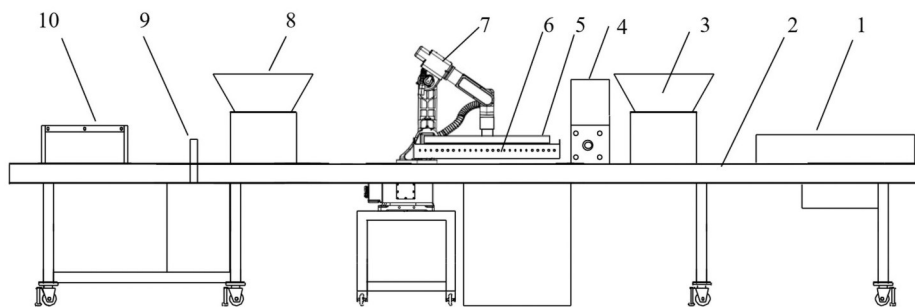
场景，但其目标对象普遍尺寸较大，分割难度相对较低。XING等^[23]探索了重叠条件下的种子实例分割问题，提出的方法在重叠区域的种子检测精度超过88%。WU等^[24]研究了密集分布水稻种子的实例分割，然而相比秧盘环境其复杂度仍显不足。在秧盘场景中，种子常呈密集重叠状态且受土壤局部遮挡，传统实例分割方法在此条件下难以取得理想效果。因此，融合语言文本特征以提升此类复杂场景下的分割性能与鲁棒性具有重要研究价值。

综上所述，本研究提出一种基于实例分割算法的水稻育秧盘内复杂背景下的播种效果检测方法。并通过与Mask R-CNN及Mask2Former这两种主流实例分割模型的性能对比，评估图文定位生成（Caption Grounding and Generation, CGG）模型的分割精度与适用性。本研究的成果可为水稻播种检测提供技术支撑，为智能化育秧盘监测奠定基础。

1 研究材料

1.1 稻种在气振式育苗精密播种装置中的处理流程

如图1所示，气振式育苗精密播种装置主要由供盘机、输送流水线、铺底土装置、吸种盘、供种盘、六轴机械臂、覆表土装置、洒水装置和叠盘机构等部分构成。其工作流程为：育秧盘经供盘机进入播种流水线，依次经过铺底土环节，随后由振动供种盘启动供种过程。六轴机械臂控制吸种盘完成下行吸种、水平移动至流水线上方，并将所吸附的稻种释放至育秧盘中。之后，育秧盘继续通过覆表土和洒水装置，最终输送至叠盘机进行收集。



注：1. 供盘机；2. 流水线；3. 铺底土装置；4. 流水线控制箱；5. 吸种盘；6. 供种盘；7. 机械臂；8. 覆表土装置；9. 洒水装置；10. 叠盘机。

图1 气振式育苗精密播种装置结构示意图

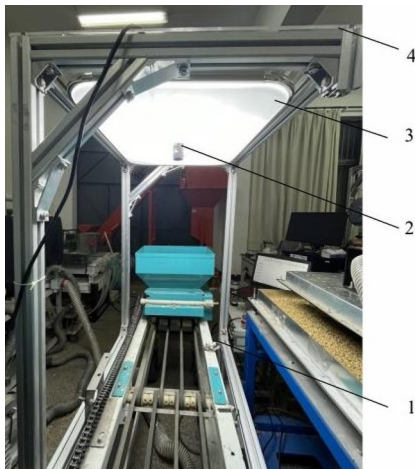
Fig. 1 Schematic structure of pneumatic vibration precision seeder

1.2 数据集采集与制作

为实现对气振式育苗精密播种装置播种效果的检测，在播种与覆表土工序之间增设育秧盘图像采

集装置，其结构如图2所示。该装置主要由相机、镜头、光源和漫反射式光电传感器组成。由于流水线运行过程中存在轻微抖动，为保障图像采集的稳

定性,在流水线两侧搭建铝型材支架以提供支撑。中孔面光源固定于流水线正上方50 cm处,相机及镜头安装于光源中心位置。当育秧盘运行至采集区域下方时,漫反射光电传感器检测到育秧盘并触发相机拍摄,从而获取播种后的育秧盘图像。



注:1.光电传感器;2.相机镜头;3.中孔面光源;4.支架。

图2 育秧盘图像采集装置图

Fig. 2 Image acquisition device for seedling tray

实验所用水稻种子为广两优476品种,育秧基质由泥炭土与蛭石按3:1配比制备,并控制湿度为65%~75%,以保证播种条件的一致性。从2023年4月初至2024年3月底,在江苏大学农业测量大楼按时间段对播种后的育秧盘进行图像采集。为保证数据集的多样性并提高训练后模型的鲁棒性,使用了两种设备:工业相机Basler acA1920-40uc(分辨率1920×1200,配套Computar 12 mm定焦镜头)固定于育秧盘上方进行垂直俯视拍摄,共采集2532张图像;同时,使用分辨率为3024×4032的iPhone 13手机手持拍摄,共采集906张图像,以增加数据多样性。两种设备共计采集原始图像3438张,用于构建高质量育秧盘稻种图像数据集。拍摄过程中采用环形LED光源(照度约15000 lux),并通过漫射罩实现均匀照明,以减少阴影和反射对图像质量的影响。

如图3a所示为所拍摄的样例,其中图3b~图3c为育秧盘内单个穴位的复杂情况。图3b中有轻微的土壤覆盖;图3c中存在种子被大量覆盖,以及密集重叠的情况;图3d中种子比较密集;图3e中种子密集重叠。这些复杂的情况,都体现实现精确实例分割的困难程度。

本研究的水稻种子实例数据集,采用开源工具Labelme^[25]对采集的原始图像进行人工标注,由5

名具备领域知识的研究人员使用多边形工具勾勒每粒稻种的外轮廓。为提升标注效率,引入预训练EfficientSAM模型^[26]辅助标注过程。标注结果包含每粒种子的左上角和右下角坐标及其轮廓多边形坐标,最终将标注数据转换为COCO数据集格式并保存为JSON文件,完成水稻种子实例数据集构建。数据按8:1:1比例随机划分为训练集(2750张)、验证集(344张)与测试集(344张)。

2 研究方法

2.1 CGG的网络结构

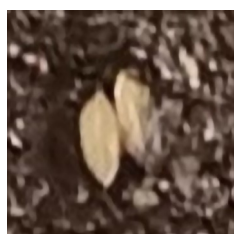
CGG^[27]的网络架构专注于开放词汇实例分割任务,通过联合对象定位和文本生成来处理图像中新类别的检测和分割。传统方法通过使用大规模的图像-文本对数据集训练模型来建立图像区域与文本词汇之间的一对一映射。CGG提出了一种更加高效的方法,通过对目标名词进行定向的匹配来提高学习效率。

为提升CGG网络在育秧盘图像中对密集遮挡稻种的语义理解能力,本研究引入大语言模型技术,并在网络结构中实现了多模态特征融合。首先,采用自举式语言-图像预训练(Bootstrapping Language-Image Pre-training, BLIP)模块对输入的原始育秧盘图像自动生成高层次语义描述,将视觉特征转化为可解释的文本表征。随后,利用双向编码器表示的Transformer(Bidirectional Encoder Representations from Transformers, BERT)模块对生成文本中的词汇进行上下文相关的向量编码,以获得语义一致性更强的词嵌入表示。此类语言嵌入与CGG主干网络所提取的视觉特征在特定融合层进行交互,实现了视觉与语言模态的深度融合与协同建模。该机制有效弥补了纯视觉特征在处理密集小目标及高遮挡区域时存在的语义表达能力局限,为稻种实例分割与计数任务提供了更鲁棒的判别基础。相较于传统仅依赖卷积视觉特征的方法,本研究通过将CGG与大语言模型有机集成,在复杂播种场景中显著增强了模型的语义感知能力与分割精度。

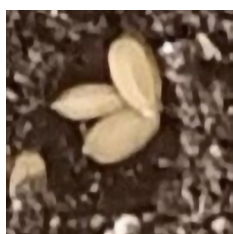
CGG网络架构图如图4所示,CGG主要分成图像模块和语言模块两个部分。在图像模块中,将原始的水稻种子的RGB图像转化成640×640×3的向量,输入到Mask2Former的基线当中。首先,图像会经过主干网络进行原始图像的特征提取得到图像特征,然后提取到的这些特征会被送入Trans-



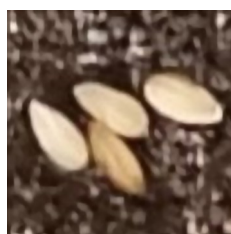
a. 播种后育秧盘原始图像



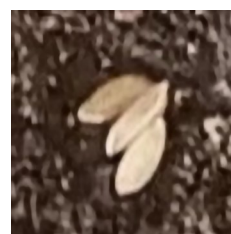
b. 轻微土壤覆盖



c. 密集重叠及土壤覆盖



d. 密集



e. 密集重叠

图3 播种后育秧盘原始图像示例

Fig. 3 Example of raw images of seedling trays after sowing

former解码器中处理。其得到的输出被输入到多层感知机（Multilayer Perceptron, MLP）中，该MLP会和像素级解码器的输出一起得到 N 个掩码预测。同时，将对象查询转换成 N 个多模态嵌入，然后计算这些嵌入和类别嵌入之间的相似度，从而得到 N 个类别的预测。在语言模块中，还分为两个部分：一部分是计算标题和图像的接地损失；另一部分是进行类别预测。得到的每张原始图像会经过BLIP模型生成5个标题文本信息，然后会从标题中提取到所需要的seed的名词。这些单词会经由单词编码器进行提取得到单词的特征，参与到接地损失的计算中。在类别预测中，使用预训练好的BERT模型，直接对seed这个单词进行提取得到预训练的词向量。通过计算类别的损失、掩码的损失，以及文本图像对齐的损失^[28]，来不断地训练网络模型。最终训练得到的模型，即可对水稻种子图像进行预测分割。

2.2 网络核心改进模块

在图4中，BLIP与BERT模块作为关键性增强组件被集成于CGG模型架构中，其在图中以青色

高亮标注。下文将针对这两个核心改进的模块展开系统性阐述，其结构如图5所示。

BLIP^[29]是一种视觉-语言预训练模型，能够深度融合图像与文本信息，具备强大的跨模态理解与生成能力。该模型基于大规模图像-文本对进行训练，通过对比学习和自引导伪标注策略，有效捕捉视觉与语言之间的语义关联，并提升泛化性能。其核心优势在于能够根据输入图像生成流畅且准确的文本描述，不仅识别显著视觉特征，还能融合上下文语义信息，增强对复杂场景的理解能力。针对水稻种子图像数据量大、重复性高、人工标注成本高的问题，本研究利用预训练BLIP模型对每幅图像自动生成5组描述文本，以提供高质量语义监督。

BERT^[30]是一种基于Transformer架构的语言表示模型，能够生成深层次上下文相关的词嵌入。本研究采用BERT基础版本，其结构包含12层Transformer编码器。在处理文本表征任务时，BERT依托其双向自注意力机制，同步整合词语的上下文信息，进而生成更具上下文依赖性的高质量语义表

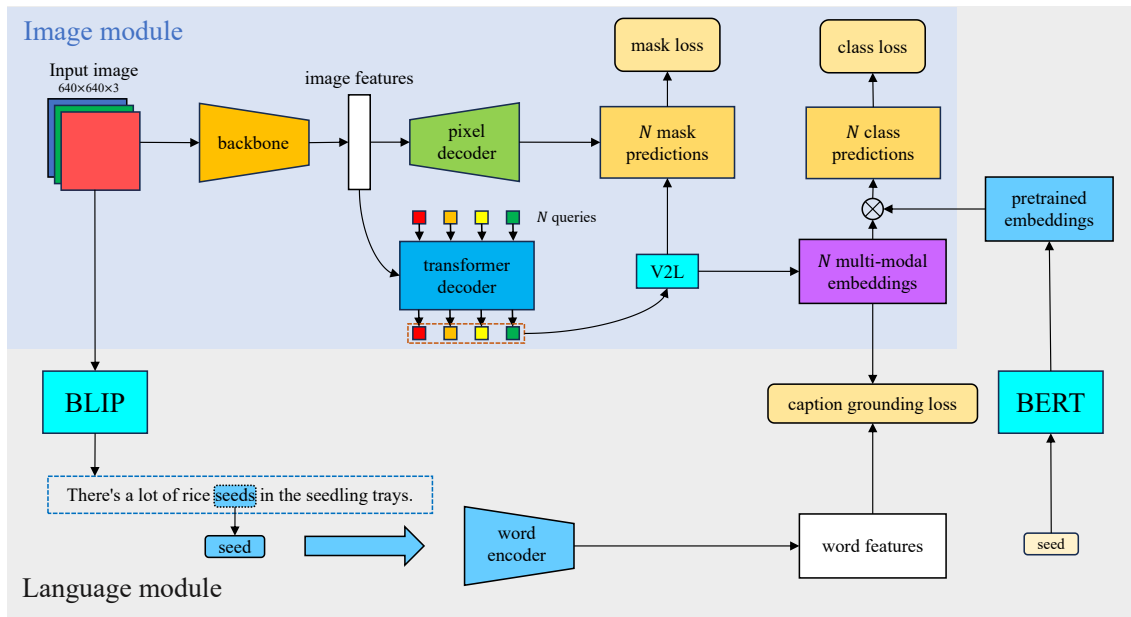
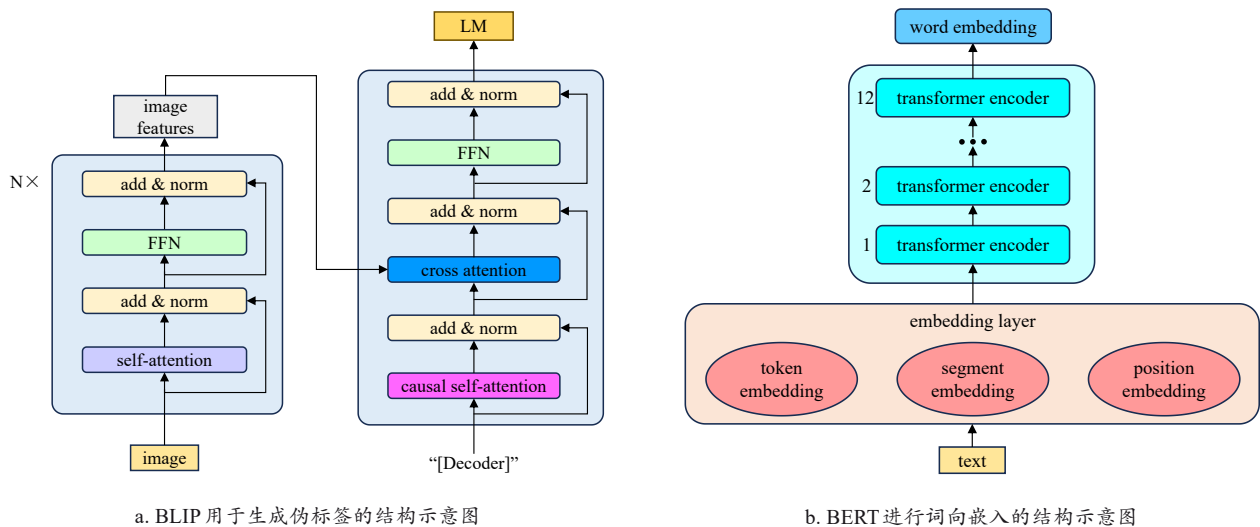


图4 CGG网络架构图

Fig. 4 CGG network architecture diagram



a. BLIP 用于生成伪标签的结构示意图

b. BERT 进行词向嵌入的结构示意图

图5 CGG模型中所采用的BLIP与BERT模块结构示意图

Fig. 5 Schematic diagram of the BLIP and BERT module structures employed in the CGG model

示。相比传统单向语言模型，BERT在多项自然语言处理任务中表现出更强的语义建模能力。本研究借助BERT对BLIP生成的文本描述进行编码，以获取具有上下文一致性的词向量表示。

2.3 评价指标

在对水稻种子图像进行检测分割的基础上，对水稻种子进行计数。所以前置的分割指标格外重要。以IoU阈值为0.5时的平均精度均值作为评价指标。其中， $mAP50^{bb}$ 用于评估边界框检测精度，

衡量预测框与真实框的重叠匹配程度； $mAP50^{seg}$ 则用于评估实例分割精度，反映预测掩码与真实掩码在像素级的贴合质量。二者共同表征了模型的定位能力与分割性能。如果标记多边形的IoU大于0.5，则将预测的边界框识别为真阳性（True Positive, TP），否则，将其归类为假阳性（False Positive, FP）。如果是模型漏检的目标，将其归类为假阴性（False Negative, FN）。精度是准确检测到的实例数与检测到的实例总数的比率，而召回率是准确检测到的实例与数据集中总体实例数的比率。 mAP 是衡

量模型在不同置信阈值下的总体性能。最终，这些指标可以定义成公式（1）~公式（4）。

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$AP = \frac{1}{11} \sum_{Recall \in [0, 0.1, \dots, 1]} Precision(Recall_i) \quad (3)$$

$$mAP = \frac{1}{c} \sum_{i=1}^c AP_i \quad (4)$$

式中： c 为类别数； $Precision$ 为精确率； $Recall$ 为召回率； AP 为平均精度。

此外，为了系统评估模型的整体性能，引入参数规模（Params）、计算量 GFLOPs 与推理速度（Frame rate）3项指标，分别用于反映模型规模、理论运算量和实际推理效率，如公式（5）~公式（7）所示。

$$Params = \sum_{l=1}^L Num(W_l) + Num(b_l) \quad (5)$$

式中： L 为网络层数； $Num(W_l)$ 为第 l 层权重矩阵的元素数量； $Num(b_l)$ 为第 l 层偏置向量的元素数量。

$$FLOPs = \sum_{l=1}^L 2 \times H_{out} \times W_{out} \times C_{in} \times C_{out} \times K^2 \quad (6)$$

式中： H_{out} 和 W_{out} 分别为输出特征图的高和宽，px； C_{in} 和 C_{out} 分别为输入和输出通道数； K 为卷积核尺寸；系数2表示一次乘加运算。

$$Frame\ rate = \frac{N}{T} \quad (7)$$

式中： N 为推理的图像数量，张； T 为推理所用时间，s。

为了比较CGG和其他实例分割算法得出种子数量的结果，使用平均绝对误差（Mean Absolute Error, MAE）、平均绝对百分误差（Mean Absolute Percentage Error, MAPE）和误差均方根（Root Mean Square Error, RMSE）等指标进行评估。RMSE常用于衡量预测值与真实值之间的偏差，它能够反映模型的预测误差大小；MAE用于衡量预测值与真实值之间的平均绝对偏差；MAPE常用于评价预测值与真实值之间的相对误差。计算如公式（8）~公式（10）所示。

$$RMSE = \sqrt{\frac{1}{n} \sum (N_{Actual_i} - N_{Predicted_i})^2} \quad (8)$$

$$MAE = \frac{1}{n} \sum |N_{Actual_i} - N_{Predicted_i}| \quad (9)$$

$$MAPE = \frac{1}{n} \sum \left| \frac{N_{Actual_i} - N_{Predicted_i}}{N_{Actual_i}} \right| \times 100\% \quad (10)$$

式中： N_{Actual_i} 为第 i 个样本真实的数量，粒； $N_{Predicted_i}$ 为第 i 个样本预测的数量，粒； n 为样本总数，个。

3 结果与讨论

3.1 试验与模型性能比较

本研究所采用的加速硬件为 NVIDIA GeForce RTX 3060 图形处理器，具备 12 GB 显存。模型构建所使用的编程语言为 Python（版本 3.8），深度学习框架为 PyTorch（版本 1.10），并结合 CUDA（版本 11.8）进行 GPU 加速。基础学习率设定为 0.000 1，优化器选用 AdamW，动量参数设置为 0.937。输入图像像素为 640×640 ，训练轮数为 200。

为评估所提出的 CGG 模型中改进模块的有效性，在 4 种不同配置下进行了消融实验，结果如表 1 所示。基线模型（第 1 组）为未引入伪标签与语义嵌入的 CGG 模型；第 2 组在训练过程中加入由 BLIP 生成的伪标签，实验结果表明各项准确率指标均有明显提升，说明 BLIP 伪标签有效增强了训练数据的语义丰富性，从而提升了模型的泛化能力。第 3 组在基线模型中引入基于 BERT 的词向量嵌入，进一步提升了稻种分割精度，表明语义特征嵌入增强了模型在复杂密集场景下的识别能力。第 4 组同时结合 BLIP 伪标签与 BERT 词向量嵌入，在各项评估指标中均表现出最优性能，验证了两者在模型中的互补性，共同促进了分割精度的提升。该配置在计算开销上略有增加，其主要原因在于：一方面，BLIP 伪标签的生成在训练阶段引入了额外的数据处理与语义对齐步骤；另一方面，BERT 模块作为多层 Transformer 架构，增加了参数量和前向推理计算量，从而导致整体浮点运算量上升，推理帧率（Frame Per Second, FPS）略有下降。尽管如此，该额外开销仍处于可接受范围内。此外，与基线模型相比，该组合策略在准确率方面提升超过 3 个百分点，展现出其在精量播种等实际应用中的显著优势。另外，下文中所提到的 CGG 均是加入了 BLIP 与 BERT 结构的模型。

为了验证所提出 CGG 模型的有效性，本研究选取了两类具有代表性的实例分割方法作为对比对象：一方面，Mask R-CNN 作为经典的两阶段实例分割模型，能够在目标检测与分割精度上提供较强

表1 CGG模型在不同配置下的消融实验
Table 1 Ablation study of the CGG model under different configurations

Group	BLIP	BERT	mAP50 ^{bb} /%	mAP50 ^{seg} /%	Params/M	计算量/GFLOPs	帧率/(帧/s)
1	×	×	87.5	88.3	33.9	225.1	15
2	√	×	89.8	90.6	34.5	229.2	15
3	×	√	90.3	91.1	34.0	202.1	18
4	√	√	90.7	91.4	35.7	207.5	18

注:√表示采用该模块;×表示不采用该模块。

的基准,但其推理速度较慢;另一方面,Mask2Former作为近年来基于Transformer的先进模型,同时也是CGG的主要框架,其通过引入查询机制实现了精度与速度的兼顾,代表了实例分割领域的发展趋势。通过与这两类主流方法的对比,可以更全面地评估CGG模型在复杂育秧盘场景下的性能优势。

表2 不同模型在稻种实例分割中的性能对比
Table 2 Performance comparison of different models in rice seed instance segmentation

Model	Input size	mAP50 ^{bb} /%	mAP50 ^{seg} /%	Params/M	计算量/GFLOPs	帧率/(帧/s)
Mask R-CNN	640	81.1	80.7	44.1	137.7	13
Mask2Former	640	84.4	86.1	44.4	226.0	16
CGG	640	90.7	91.4	35.7	207.5	18

在推理速度方面,对3种模型的帧率进行了量化对比。实验结果表明,Mask R-CNN的推理速度为13帧/s,Mask2Former为16帧/s,而所提出的CGG模型达到18帧/s。在保证分割精度提升的同时,CGG在推理效率上也表现出一定优势,验证了其在实际应用场景中具备更高的可用性。

精确率-召回率(Precision-Recall, PR)曲线反映了模型在不同置信度阈值下的检测能力,其中较高的Precision表示较少的FP,较高的Recall表示较少的FN。这两个指标通常相互制约:提高精确率往往会降低召回率,反之亦然。曲线越接近右上角,说明模型在保持较高精确率的同时也具备较高的召回率,表明其整体性能更优。图6展示了3种模型的PR曲线。可以明显看出,CGG模型的曲线更接近右上角,相较于另外两种模型表现更佳,进一步验证了CGG模型在实例分割任务中的优越性能。

3.2 育秧盘内稻种实例分割结果分析

图7展示了3种模型在同一张育秧盘图像上的分割结果,分别为Mask R-CNN、Mask2Former和CGG,对应图中标注的图7a、图7b和图7c。总体来看,3种模型均能够较为成功地分割出大部分稻

实验结果如表2所示,CGG模型在mAP50^{bb}指标上达到90.7%,分别较Mask R-CNN和Mask2Former高出9.6和6.3个百分点;在mAP50^{seg}指标上表现同样优异,达到91.4%,分别高出10.7与5.3个百分点。虽然CGG模型参数数量最少,但其计算开销相对较高。

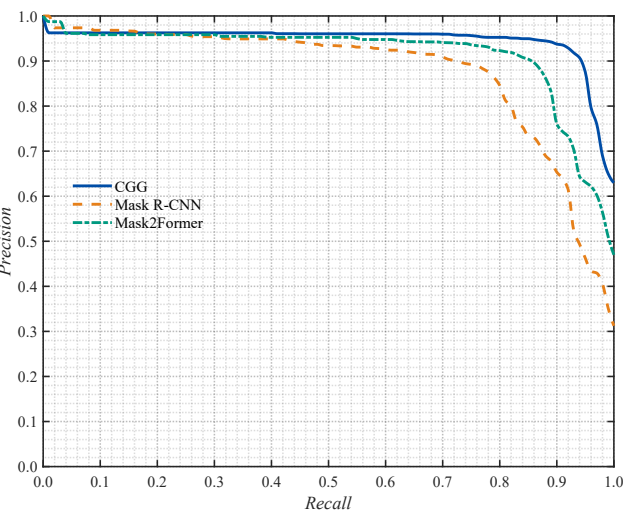


图6 不同模型在稻种实例分割任务中的PR曲线
Fig. 6 Precision-recall curves of different models for rice seed instance segmentation

种。然而,通过对比可以发现,各模型在结果中仍存在不同程度的漏分现象。其中,图7c中漏分的情况明显少于图7a和图7b。由于图像中包含大量密集重叠的小粒稻种,且部分区域存在土壤遮挡,在这种复杂情况下,人工目视评估在一定程度上具有主观性。为更直观地呈现差异,选取原图中1处具有代表性的复杂区域进行裁剪,该区域包含密集重

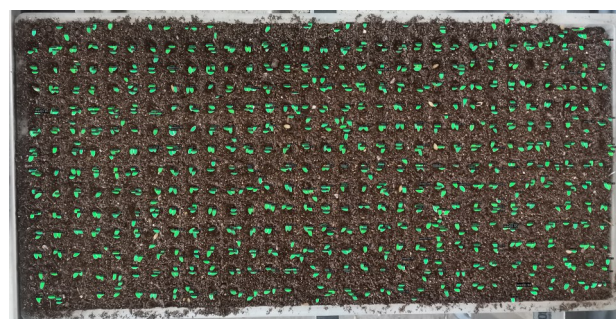
叠的稻种和土壤覆盖，便于对模型在复杂场景下的分割性能进行比较分析。



a. Mask R-CNN



b. Mask2Former

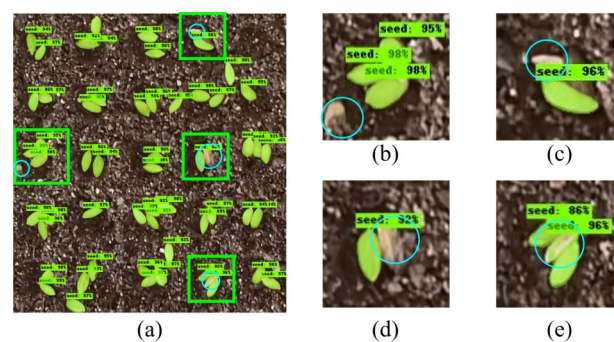


c. CGG

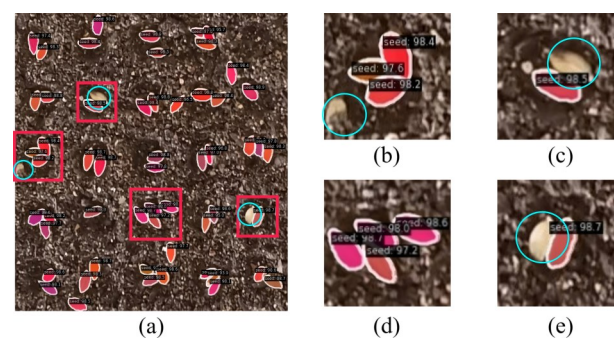
图7 不同模型对育秧盘内稻种的实例分割结果对比图
Fig. 7 Instance segmentation results in the entire seedling tray by different models

图8分别展示了Mask R-CNN、Mask2Former和CGG对育秧盘内同一个局部区域的分割结果，分别标注为a、b和c。每组中的子图(a)显示了在同一裁剪区域上的整体推理结果；每组子图(b)(c)(d)和(e)则对应于每组子图(a)中用边界框标记的4个局部区域，依序按从左至右、从上至下的顺序排列。

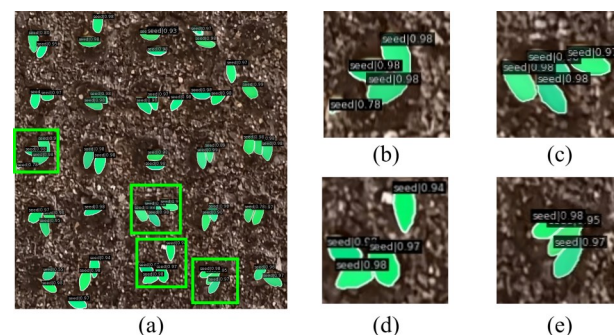
从图8各子图(a)可以看出，Mask R-CNN与Mask2Former存在较明显的漏检现象，其中漏检实例已在图中以青色圆圈标注，而CGG能完整识别该区域内的所有稻种，表明CGG模型在整体检测性



a. Mask R-CNN



b. Mask2Former



c. CGG

图8 不同模型在育秧盘局部区域的分割细节放大图
Fig. 8 Zoomed-in segmentation details of a local region in the seedling tray by different models

能上表现更优。进一步观察各组中的子图(b)~(e)，可发现Mask R-CNN在所有局部区域均存在漏检，尤其是子图8a(e)中的中间稻种，分割效果较差，几乎无明显掩膜覆盖，显示其分割不完整。Mask2Former在图8b(d)中实现了较为准确的分割，但在图8b(b)、图8b(c)和图8b(e)中仍有漏检现象，整体优于Mask R-CNN但仍存在不足。而CGG模型在4个子区域中均成功识别并分割出全部稻种，表现最为稳定。此外，尽管CGG在图8c(b)中对一粒被土壤部分遮盖的稻种分割置信度仅为0.78，略低于理想水平，但相较于Mask R-CNN与Mask2Former未能识别该种子的情况，CGG仍显示出更强的鲁棒性。这也反映出当

稻种被部分覆盖时，实例分割仍面临一定挑战。

综上所述，在稻种密集重叠且背景复杂的场景中，与其他两个模型相比，CGG模型在实例分割任务中展现出最优的性能，具有更高的检测完整性与分割精度。

3.3 育秧盘穴播种子量测试结果

为了进一步验证CGG算法的有效性，本研究在真实流水线上连续进行20盘实际播种试验，采用不同的算法模型及人工检测方法，对每盘及每个穴位中的稻种数量进行统计，最终汇总每盘检测出的稻种总数，并计算各类型穴位在整盘（共434个

穴位）中的比例，所得结果如表3所示。其中，误差指的是各算法模型与人工检测结果之间的直接数值差异，数值越接近0，表明算法检测结果越接近真实值。然而，该误差仅反映了表层的差距，未能充分考虑算法在某些特定情况下的识别偏差。例如，当算法将某一穴位检测为含1粒稻种，而实际该穴位可能包含2粒或3粒稻种时，仅以“1粒”计入误差将低估检测的实际偏差。因此，本研究进一步引入“真实误差”的概念，即将此类识别错误归入对应的真实类别（如2粒或3粒稻种），从而更准确地衡量检测算法与真实情况之间的偏差。

表3 不同检测方法下对单盘不同稻种穴位平均占比统计表

Table 3 Statistics on the average percentage of cavities of different rice species in a single disk

		单位：%							
单穴稻种数量		0	1	2	3	4	5	6	7
检测方法	Mask R-CNN	3.30	10.83	29.19	30.26	20.20	4.07	1.77	0.38
	Mask2Former	2.69	10.98	33.33	23.50	20.35	7.45	0.92	0.77
	CGG	2.38	10.22	33.10	22.50	22.73	6.99	1.38	0.69
	人工检测	2.07	9.52	32.41	23.12	22.12	8.76	1.08	0.92
真实误差	Mask R-CNN	0.00	1.23	3.69	4.84	7.83	5.99	1.15	0.54
	Mask2Former	0.00	0.61	3.69	3.53	3.46	3.15	0.46	0.15
	CGG	0.00	0.31	2.46	2.84	3.15	2.23	0.69	0.23
误差	Mask R-CNN	-1.23	-1.31	3.23	-7.14	1.92	4.69	-0.69	0.54
	Mask2Former	-0.61	-1.46	-0.92	-0.38	1.77	1.31	0.15	0.15
	CGG	-0.31	-0.69	-0.69	0.61	-0.61	1.77	-0.31	0.23

结果显示，真实误差普遍大于表层误差，反映出前者在评估算法性能方面具有更高的敏感性与准确性。此外，在“真实误差”中，稻种检测为空穴时对应栏目的误差值为0，这是由于算法对空穴位的识别均为正确。若某一穴位中实际存在1粒稻种但未被检测出，该错误将被统计至“1粒稻种”对应的真实类别中。

图9展示了3个模型在预测育秧盘中每个穴内稻种数量方面的混淆矩阵。结果表明，CGG模型在单个孔位上的预测准确率达到88%，明显高于Mask R-CNN（75%）和Mask2Former（85%）。在推理效率方面，CGG模型的帧率为18帧/s，优于Mask2Former的16帧/s和Mask R-CNN的13帧/s。这表明，CGG模型在保证较高预测准确率的同时，还具备更快的推理速度，从而在精度与效率之间实现了更优的平衡。综上，CGG模型在该任务中展现出显著的综合优势。

表3所反映的数据是以单个穴位为单位，对穴

内稻种数量的检测情况进行统计。为了进一步评估各模型在单粒稻种检测任务中的误差表现。表4列出了均方根误差、平均绝对误差，以及平均绝对百分比误差。表4列出了Mask R-CNN、Mask2Former与CGG这3种模型在各项误差指标上的表现，其中CGG模型的结果为RMSE=16.8颗，MAE=13.7颗，MAPE=2.46%，整体优于其他两种模型，取得了最优的检测效果。此外，这些误差指标均以稻种数量为基础进行统计，因此，实现对稻种精确的轮廓分割对于获得可靠的误差评估结果至关重要。

4 结 论

本研究提出了一种面向育秧盘稻种检测的实例分割模型。该模型融合了CGG网络架构与大型语言模型，实现了秧盘图像中单粒稻种的高精度分割。本模型在mAP50^{bb}和mAP50^{es}指标上分别达到90.7%与91.4%，显著优于Mask R-CNN与Mask2Former基准模型。同时，其RMSE、MAE和

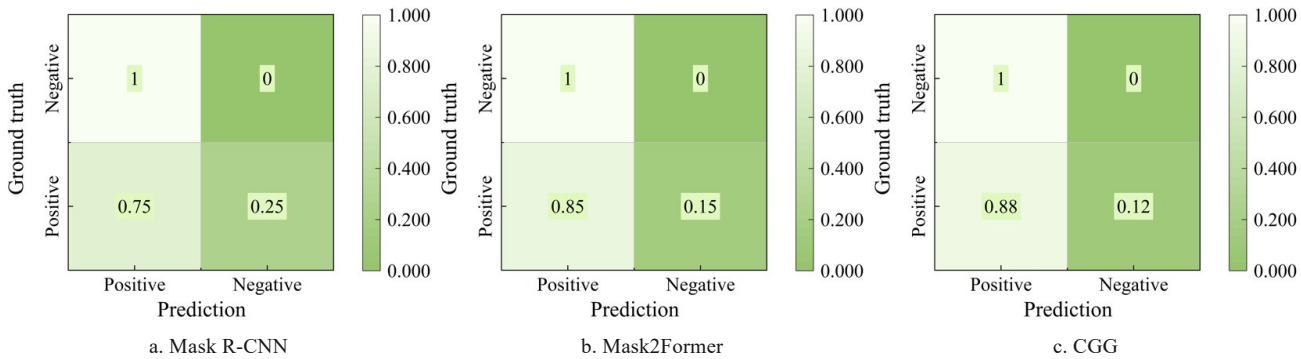


图9 不同模型在单穴稻种数量预测的混淆矩阵对比图

Fig. 9 Confusion matrix comparison of different models for per-cavity seed count prediction

表4 不同模型针对单粒稻种的误差结果

Table 4 Error analysis results for individual rice seeds by different models

Model	RMSE/颗	MAE/颗	MAPE/%
Mask R-CNN	49.6	44.4	7.27
Mask2Former	30.9	26.4	5.21
CGG	16.8	13.7	2.46

MAPE均低于对比模型，验证了模型在数量预测与误差控制上的优势。在效率表现上，所提模型在实现高精度分割的同时，仍保持了具有竞争力的推理速度，推理帧率达到18帧/s，展现出良好的实用性与可部署性。

本研究为精准农业实践及智能化补种系统的实施提供了可靠技术支撑。未来将进一步结合智能补种技术，解决育秧盘中漏播或播种量不足的问题，从而有效提升整体播种质量。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] 陈品, 徐春春, 纪龙, 等. 2024年我国水稻产业形势分析及2025年展望[J]. 中国稻米, 2025, 31(2): 1-5.
CHEN P, XU C C, JI L, et al. Analysis of China's rice industry in 2024 and the outlook for 2025[J]. China Rice, 2025, 31(2): 1-5.

[2] CHENG J H, LI Y M, CHEN J, et al. Design and realization of seeding quality monitoring system for air-suction vibrating disc type seed meter[J]. Processes, 2022, 10(9): 1745.

[3] 丁幼春, 王凯阳, 刘晓东, 等. 中小粒径种子播种检测技术研究进展[J]. 农业工程学报, 2021, 37(8): 30-41.
DING Y C, WANG K Y, LIU X D, et al. Research progress of seeding detection technology for medium and small-size seeds [J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(8): 30-41.

[4] DONG W H, MA X, LI H W, et al. Detection of performance of hybrid rice pot-tray sowing utilizing machine vision and machine learning approach[J]. Sensors, 2019, 19(23): 5332.

[5] BAI J Q, HAO F Q, CHENG G H, et al. Machine vision-based supplemental seeding device for plug seedling of sweet corn[J].

Computers and Electronics in Agriculture, 2021, 188: 106345.

[6] WANG S X, CHEN J, LI Y M, et al. The detection method of rice seedling tray hole seeding quantity based on improved YOLOv5s [C]// 2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI). Piscataway, New Jersey, USA: IEEE, 2023: 1165-1169.

[7] YAN Z Y, ZHAO Y M, LUO W S, et al. Machine vision-based tomato plug tray missed seeding detection and empty cell replanting [J]. Computers and Electronics in Agriculture, 2023, 208: 107800.

[8] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[M]// Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[9] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New Jersey, USA: IEEE, 2014: 580-587.

[10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2015: 3431-3440.

[11] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2017: 2980-2988.

[12] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-art natural language processing[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, PA, USA: ACL, 2020: 38-45.

[14] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2022: 1280-1289.

[15] CHENG B, SCHWING A, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation[C/OL]// 35th Conference on Neural Information Processing Systems. 2021.[2025-07-06]. <https://proceedings.neurips.cc/paper/2021/hash/950a4152c2b4aa3ad78bdd6b366cc179-Abstract.html>.

[16] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. Proceedings of Machine Learning Research. New York, USA: PMLR, 2021: 8748-8763..

[17] TODA Y, OKURA F, ITO J, et al. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping [J]. Communications Biology, 2020, 3(1): 173.

- [18] YE S, LIU W H, ZENG S, et al. SY-net: A rice seed instance segmentation method based on a six-layer feature fusion network and a parallel prediction head structure[J]. *Sensors*, 2023, 23(13): 6194.
- [19] GAO Q, LI H, MENG T Y, et al. A rapid construction method for high-throughput wheat grain instance segmentation dataset using high-resolution images[J]. *Agronomy*, 2024, 14(5): 1032.
- [20] TANG S X, XIA Z L, GU J N, et al. High-precision apple recognition and localization method based on RGB-D and improved SOLOv2 instance segmentation[J]. *Frontiers in Sustainable Food Systems*, 2024, 8: 1403872.
- [21] MA J, ZHAO Y K, FAN W P, et al. An improved YOLOv8 model for *Lotus* seedpod instance segmentation in the *Lotus* pond environment[J]. *Agronomy*, 2024, 14(6): 1325.
- [22] PENG Y, WANG A C, LIU J Z, et al. A comparative study of semantic segmentation models for identification of grape with different varieties[J]. *Agriculture*, 2021, 11(10): 997.
- [23] XING H, WAN Y K, ZHONG P, et al. Design and experimental analysis of real-time detection system for the seeding accuracy of rice pneumatic seed metering device based on the improved YOLOv5n[J]. *Computers and Electronics in Agriculture*, 2024, 227: 109614.
- [24] WU Z P, CHEN J, MA Z, et al. Development of a lightweight online detection system for impurity content and broken rate in rice for combine harvesters[J]. *Computers and Electronics in Agriculture*, 2024, 218: 108689.
- [25] MAO J, MA X, BI Y, et al. TJDR: A high-quality diabetic retinopathy pixel-level annotation dataset[EB/OL]. arXiv: 2312.15389, 2023.
- [26] XIONG Y Y, VARADARAJAN B, WU L M, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2024: 16111-16121.
- [27] WU J Z, LI X T, DING H H, et al. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2023: 21881-21891.
- [28] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]// 2016 Fourth International Conference on 3D Vision (3DV). Piscataway, New Jersey, USA: IEEE, 2016: 565-571.
- [29] LI J, LI D, XIONG C, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]// International Conference on Machine Learning. Proceedings of Machine Learning Research. New York, USA: PMLR, 2022: 12888-12900..
- [30] KOROTEEV M V. BERT: A review of applications in natural language processing and understanding[EB/OL]. arXiv:2103.11943, 2021.

CGG-Based Segmentation and Counting of Densely Distributed Rice Seeds in Seedling Trays

OUYANG Meng¹, ZOU Rong^{1*}, CHEN Jin¹, LI Yaoming², CHEN Yuhang¹, YAN Hao¹

(1. School of Mechanical Engineering, Jiangsu University, Zhenjiang 212013, China; 2. School of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: [Objective] The precise quantification of rice seeds within individual cavities of seedling trays constitutes a critical operational parameter for optimizing seeding efficiency and fine-tuning the performance of air-vibration precision seeders. Achieving high accuracy in this task directly impacts resource utilization, seedling uniformity, and ultimately crop yield. However, the operational environment presents significant challenges, including complex backgrounds, seed overlap, variations in lighting and seed orientation, and the inherent difficulty of distinguishing individual seeds within dense clusters. These factors often lead to suboptimal performance in existing automated detection systems, manifesting as low detection accuracy and an inability to achieve robust, precise instance segmentation of individual rice seeds. To address these persistent limitations and advance the state-of-the-art in precision seeding monitoring, an integrated framework for rice seed instance segmentation was proposed. The core innovation lies in the synergistic combination of a cross-modal grounding generation (CGG) network with a pretrained model, which is designed to leverage complementary information from visual and textual domains. **[Methods]** The proposed methodology fundamentally aimed to bridge the gap between visual perception and semantic understanding within the specific context of rice seed detection. The CGG-pretrained model framework achieved this through deep joint alignment of visual features extracted from seedling tray images and textual features derived from contextual knowledge. This cross-modal grounding enabled collaborative learning, where the visual processing stream (handling object localization and pixel-level segmentation) was continuously informed and refined by the semantic understanding stream (interpreting context and relationships). Specifically, the visual backbone network processes input imagery to generate feature maps, while the pretrained language model component, which utilized contextual embeddings, generated semantically rich textual representations. The CGG module acted as the fusion engine, establishing explicit correspondences between specific regions in the image (potential seeds or clusters) and relevant semantic concepts or descriptors provided by the pretrained model. This bidirectional interaction significantly enhanced the model's ability to disambiguate overlapping seeds, resolved occlusions, and accurately delineated individual seed boundaries under challenging conditions. Key technical innovations validated through rigorous ablation studies include: (1) The strategic use

of the bootstrapping language-image pre-training (BLIP) model for generating high-quality pseudo-labels from unlabeled or weakly labeled image data, facilitating more effective semi-supervised learning and reducing annotation burden, and (2) the application of bidirectional encoder representations from transformers (BERT)-based word embed to capture deep semantic relationships and contextual nuances within textual descriptors related to seeds and seeding environments. **[Results and Discussions]** The ablation experiments demonstrated a pronounced synergistic effect when the core improvements were combined, resulting in a segmentation accuracy improvement exceeding 3 percentage points compared to the baseline model that lacking the integration. Comprehensive experimental evaluation demonstrated the superior performance of the proposed CGG model against established benchmarks. Under the standard intersection over union (IoU) threshold of 0.5, the model achieved a mean average precision (mAP) of 90.7% for bounding box detection (denoted as mAP50^{bb} for detection) and an outstanding 91.4% mAP for instance segmentation (denoted as mAP50^{seg} for segmentation). These results represented a statistically significant improvement over leading contemporary models, including region-based convolutional neural network (Mask R-CNN) and Mask2Former, which highlighted the efficacy of the cross-modal grounding approach in accurately localizing and segmenting individual rice seeds. Further validation within realistic seeding trial scenarios, which involved direct comparison with meticulous manual annotations, confirmed the model's practical robustness. The CGG model attained the highest accuracy in two critical operational metrics: (1) Precision in segmenting individual seed instances (single-seed segmentation accuracy), and (2) accuracy in determining the exact seed count per cavity, and it achieved an average accuracy of 88% for per-cavity quantification. Moreover, the model exhibited superior performance in minimizing estimation errors for cavity seed counts, as evidenced by its significantly lower error metrics: a root mean square error (RMSE) of 16.8 seeds, a mean absolute error (MAE) of 13.7 seeds, and a mean absolute percentage error (MAPE) of 2.46%. These error values were markedly lower than those recorded by the comparison models, which underscored the CGG model's enhanced reliability in practical counting tasks. The discussion contextualized these results and attributed the performance gains to the model's ability to leverage semantic context to resolve ambiguities inherent in visual-only approaches, particularly in dense and overlapping seed scenarios common in precision seeding trays. **[Conclusions]** The developed CGG-pretrained model integration presents a significant advancement in automated monitoring for precision rice seeding. The model successfully addresses the core challenges of low detection accuracy and imprecise instance segmentation for seeds in complex environments. Its high accuracy in both individual seed segmentation and per-cavity seed count quantification, coupled with low error rates, demonstrates strong potential for practical deployment. Importantly, the model enables real-time detection of rice seeds during the image analysis stage, this functionality provides a quantifiable, data-driven basis for making immediate operational decisions, most notably enabling the targeted precision reseeding of empty or under-seeded cavities identified during the seeding process. By ensuring optimal seed placement and density from the outset, the technology contributes directly to improved resource efficiency (reducing seed waste), enhanced seedling uniformity, and potentially higher crop yields. Future work will focus on further optimizing inference speed for higher-throughput seeding lines and exploring generalization to other crop types and seeding mechanisms.

Keywords: rice seeding; BERT; instance segmentation; pre-trained model; visual-language model; caption grounding and generation; large model

Foundation item: The National Natural Science Foundation of China (31871528)

First author: OUYANG Meng, E-mail: amen610401404@163.com

***Corresponding author:** ZOU Rong, E-mail: zr@ujls.edu.cn