

基于 RT-WEDT 的麦穗检测与计数方法

李 婕¹, 杨子豪¹, 郑 权¹, 乔江伟², 涂静敏^{1*}

(1. 湖北工业大学电气与工程学院, 武汉 430068; 2. 中国农业科学院油料作物研究所, 武汉 430062)

摘 要: 小麦是重要的粮食作物之一, 麦穗计数对于预测麦穗产量至关重要。针对现有的检测计数方法在复杂农田环境下存在检测精度不足、模型参数量大等问题, 该研究提出一种轻量级麦穗检测模型 RT-WEDT (real-time wheat ear detection transformer)。首先, 选择基于 transformer 的轻量化网络 EfficientFormerV2 作为 RT-WEDT 的骨干网络, 以提升特征提取效率的同时学习麦穗图像的长距离特征; 其次, 设计三重特征融合模块 (triple feature fusion, TFF) 并引入尺度序列特征融合模块 (scale sequence feature fusion, SSFF) 以构建多尺度增强混合编码器 (multi-scale enhanced hybrid encoder, MSEHE), 达到浅层和深层特征充分融合, 提高模型在不同尺度上的检测精度; 最后, 采用 WIoUv3 损失函数作为边界框损失函数来优化模型对麦穗目标的定位准确度。在全球麦穗数据集上的试验结果表明, RT-WEDT 模型的交并比阈值 0.50 的平均精度 AP_{50} 为 90.2%, 高于传统的目标检测模型。在自建的无人机视角麦穗数据集 (drone perspective wheat spike dataset, DPWSD) 上的交并比阈值 0.50 的平均精度 AP_{50} 为 96.8%, 验证了模型有较好的普适性。此外模型的参数量为 12M, 检测速度为 79.7 帧/s, 可达到麦穗高通量实时检测的目的。该研究为实现高效、快速的小麦产量估计提供了技术支撑, 对推动智慧农业的发展具有重要意义。

关键词: 模型; 麦穗; 目标检测; transformer; 轻量化

doi: 10.11975/j.issn.1002-6819.202405200

中图分类号: S126

文献标志码: A

文章编号: 1002-6819(2024)-21-0146-11

李婕, 杨子豪, 郑权, 等. 基于 RT-WEDT 的麦穗检测与计数方法[J]. 农业工程学报, 2024, 40(21): 146-156. doi: 10.11975/j.issn.1002-6819.202405200 <http://www.tcsae.org>

LI Jie, YANG Zihao, ZHENG Quan, et al. Method for detecting and counting wheat ears using RT-WEDT[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(21): 146-156. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202405200 <http://www.tcsae.org>

0 引 言

小麦作为主要的粮食作物, 其产量对国家粮食安全至关重要^[1]。小麦每亩穗数和穗粒质量直接决定最终产量^[2-3], 而麦穗检测可以快速的预估麦田每亩的小麦穗数, 因此, 单位面积内麦穗的检测对长势评估和产量预测具有重要意义。

传统的麦穗检测计数多采用人工调查, 存在效率低下, 准确度不高、预测主观性强等问题^[4]。随着计算机技术的发展, 基于图像处理^[5-7]和机器学习^[8-9]的麦穗计数方法逐渐兴起。然而由于大田环境中麦穗形状、颜色和尺度等存在较大差异, 为高通量麦穗检测计数的实现带来了挑战。

深度学习模型凭借多层次的特征提取和抽象学习, 无需手工设计且模型的泛化能力较强, 近年来在农业领域的应用越来越广。例如国内外学者^[10-12]利用深度学习的方法得到麦穗密度图进而直接获得麦穗数量, 大幅度

提高了计数效率。但这类基于密度图直接计数的方法容易受到透视畸变及麦穗密集遮挡等影响^[13], 此外缺乏计数麦穗的位置信息因此无法反映麦穗的几何信息。为了解决此问题, 基于检测计数的方法逐渐成为研究的热点, WANG 等^[14]提出了一种基于改进 YOLOv5 的小麦幼苗检测计数方法; 杨蜀秦等^[15]改进了 YOLOX 网络, 提升了麦穗检测和计数的准确率; HE 等^[16]采用伪标记和数据扩充等方法来提升 YOLOv4 模型在麦穗检测计数的泛化能力; MENG 等^[17]引入小尺度检测层和卷积注意力模块, 提出了 YOLOv7-MA 模型, 实现了对复杂田间小麦穗的准确检测和计数; ZHANG 等^[18]提出了一种半监督的麦穗检测器, 解决了在麦穗检测任务中数据标注不足的问题; ZHANG 等^[19]引用旋转 YOLO 麦穗检测网络和简单空间注意力网络来解决检测小麦镰刀菌头枯病的问题; HARADA 等^[20]结合卷积神经网络和 transformer 网络提出一种混合麦穗检测模型来对麦穗进行精准的检测和计数。这类方法先对目标进行检测, 通过统计检测框的数量来达到计数的目的。但田间不同麦穗生长规律差异、且受到风和光线影响, 使得麦穗图像出现遮挡粘连、尺度不一、与背景难以区分等问题。全局信息可以帮助模型捕获麦穗图像的整体布局, 区分麦穗目标和周围的背景, 从而提升模型对麦穗目标定位的精准度。但由于卷积神经网络 (CNN) 的感受野通常较小, 不利于捕获长距离特征^[21], 这使得模型很难学习到全局信息^[22],

收稿日期: 2024-05-28 修订日期: 2024-10-09

基金项目: 国家自然科学基金项目 (42301515); 智能光电系统感知及应用四川省高校重点实验室 2023 年度开放基金 (ZNGD2308)

作者简介: 李婕, 博士, 副教授, 硕士生导师, 研究方向为计算机视觉。

Email: jielonline@hbut.edu.cn

*通信作者: 涂静敏, 博士, 讲师, 研究方向为计算机视觉、三维点云数据处理。Email: jingmin.tu@hbut.edu.cn

不利于麦穗目标的精确定位。

近年来,因 transformer^[23] 基于全局信息交互进行特征学习,能有效的对图像的全局依赖关系进行建模^[24],在水稻病害检测^[25]、木薯叶病检测^[26]上体现出比 CNN 更好的优势。但基于 transformer 的方法存在计算设备要求好,在资源有限的情况下部署难的问题^[27]。受此启发,本文提出一种 transformer 架构下,改进 RT-DETR (real-time detection transformer)^[28] 的麦穗检测网络 RT-WEDT (real-time wheat ear detection transformer)。此方法将轻量化的 EfficientFormerV2^[29] 结构作为骨干网络,设计三重特征融合模块 (triple feature fusion, TFF),结合尺度序列特征融合模块 (scale sequence feature fusion, SSFF) 构建了多尺度增强混合编码器 (multi-scale enhanced hybrid encoder, MSEHE),并采用 WIoUv3 边界框损失函数以达到在保持检测计数精度的同时降低运算量的目的。最后,在全球麦穗数据集和本文自建的无人机麦穗数据集 (drone perspective wheat spike dataset, DPWSD) 上验证 RT-WEDT 的检测性能,拟为自动、高通量的大田麦穗检测和计数提供理论依据。

1 数据与方法

1.1 数据集构建

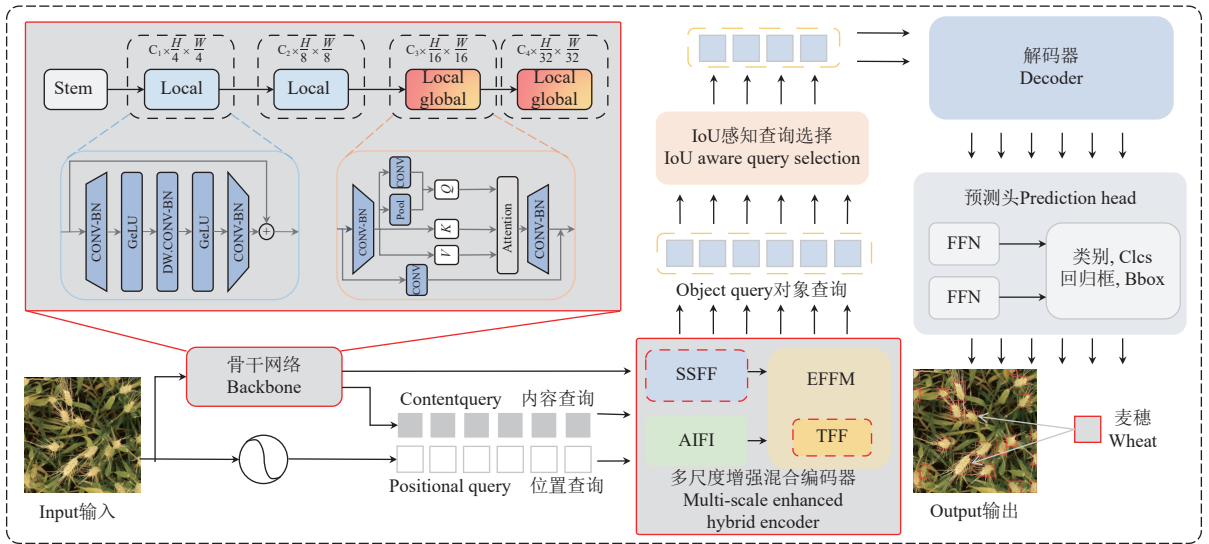
本文使用 2021 版的全球麦穗数据集 (global wheat head detection, GWHD)^[30] 作为模型的训练数据集。该数据集由拍摄于欧洲、亚洲、北美洲和大洋洲,且拍摄

时间、气候和麦穗品种各不相同的 6422 张麦穗图像 (jpg 格式) 构成,包含 275187 个边界框标签。全球麦穗数据集具有丰富的基因和环境多样性,非常适合用于神经网络的训练数据,有助于提高麦穗检测的准确性。根据机器学习数据划分原理,如果数据集边界框在 10^4 数量级,一般训练集和测试集的比例为 7:3 或是 8:2,本文考虑到较大的验证集能够提供更多样本用于模型配置的评估,从而更有利于在训练中调整最优参数^[31-33],所以将全球麦穗数据集按照 7:2:1 的比例随机划分为训练集、验证集和测试集。

1.2 RT-WEDT 网络结构

DetectionTransformer (DETR)^[34] 是首个将 Transformer 结构应用到目标检测领域的深度学习模型。DETR 将目标检测任务重构为一个序列预测问题,借鉴了 Transformer 的编解码器结构和基于二分图的匹配策略,摒弃了非极大值抑制 (non-maximum suppression, NMS) 的处理过程,避免了多余的手动操作步骤。相较于常见的使用非极大值抑制后处理的 YOLO 系列算法,其优化难度减小、鲁棒性得到增强,但 DETR 具有参数量大、模型收敛速度慢的问题,面对复杂任务时无法实现较好的实时检测效果。

RT-DETR 设计了一种高效的混合编码器,通过解耦尺度内交互和跨尺度融合来有效的处理多尺度特征,并提出 IoU 感知查询选择来改进对象查询的初始化以提高模型的运算效率。本文提出一种基于 RT-DETR 改进的轻量化网络模型 RT-WEDT,其结构如图 1 所示。



注: C_1 、 C_2 、 C_3 、 C_4 表示从骨干网络进行不同次数下采样输出的特征图; Stem 表示小内核卷积; H 表示输入图像的高; W 表示输入图像的宽; CONV 表示卷积; GeLU 表示激活函数; DW.CONV 表示深度卷积; BN 表示归一化; Pool 表示池化; Attention 表示注意力操作; Q 、 K 、 V 分别表示查询张量、键张量、值张量; SSFF 表示尺度序列特征融合模块; AIFI 表示尺度内特征交互模块; EFM 表示增强尺度融合模块; TFF 表示三重特征融合模块; FFN 表示前馈神经网络。

Note: C_1 , C_2 , C_3 , C_4 represents the feature maps output from the backbone network with different number of downsampling; Stem represents small kernel convolution; H represents the height of the input image; W represents the width of the input image; CONV represents convolution; GeLU represents activation function; DW.CONV represents deep convolution; BN represents normalization; Pool represents pooling; Attention represents Attention operation; Q , K , and V represents query tensor, key tensor, and value tensor; SSFF represents scale sequence feature fusion module; AIFI represents Intra-scale Feature Interaction module; EFM represents enhanced feature-fusion module; TFF represents triple feature fusion module; FFN represents feedforward neural network.

图 1 RT-WEDT 模型图

Fig.1 Diagram of the RT-WEDT(real-time wheat ear detection transformer) model

输入一张原始分辨率的麦穗图像, RT-WEDT 采用轻量级的 EfficientFormerV2 结构作为骨干网络来提取多

尺度特征图。然后为特征图添加位置编码并送入多尺度增强混合编码器 (MSEHE), 用来实现图像中每个像素

的长距离特征交互及图像局部特征交互,其在增加少量参数量的同时提升对不同尺度目标的检测能力;再将多尺度增强混合编码器处理后的特征图送入解码器进行解码,并通过两个前馈层分别预测检测框的位置和类别;为了提升模型对麦穗目标的定位准确性,将原有的 GIoU 损失函数更换为 WIoUv3 损失函数,提升模型对麦穗目标定位的准确度。

1.2.1 骨干网络

骨干网络是模型的重要组成部分,其选择对于模型的性能有着非常重要的影响。本文选择了 EfficientFormerV2 作为 RT-WEDT 的骨干网络。EfficientFormerV2 是由 LI 等^[29]提出的基于 Transformer 的轻量化骨干网络,被设计为 4 个分层阶段,分别获得输入分辨率的 $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ 尺寸的特征图, EfficientFormerV2 在保持较高精度的同时拥有较少的参数量,相应地降低了计算资源的需求和存储空间的占用,也有助于减少过拟合的风险。

1.2.2 多尺度增强混合编码器

在大田麦穗背景下,因受光照、土壤养分和湿度的影响,麦穗的尺寸变化较大,所以提升模型对不同尺度麦穗的检测能力至关重要。RT-DETR 提出了一种高效的混合编码器,有效降低了原始 DETR 编码器的参数量。但这种混合编码器由于缺乏最原始的浅层特征,导致 RT-DETR 在检测不同尺度目标时体现了局限性。浅层特征包含着麦穗图像小目标的表征细节和全局信息,可以帮助模型更好地捕捉和学习图像的整体框架,更有效地识别图像中小目标麦穗。对于大田麦穗这种尺度变化较大的目标,本文基于上述原因提出了一种多尺度增强混合

编码器 (multi-scale enhanced hybrid encoder, MSEHE), 增强模型多尺度的特征融合能力进而提升模型对不同尺度目标的检测能力。

多尺度增强混合编码器整体架构图如图 2 所示,由三个模块组成,即基于注意力的尺度内特征交互模块 (intra-scale feature interaction, AIFI)、尺度序列特征融合模块 (scale sequence feature fusion, SSFF) 和增强尺度融合模块 (enhanced feature-fusion module, EFFM)。首先利用骨干网络 4 个阶段的输出特征图 $\{C_1, C_2, C_3, C_4\}$ 作为编码器的输入,为了平衡计算量,尺度内特征交互模块 (AIFI) 仅在最小尺寸特征图 C_4 应用,将 C_4 特征图张量经过 Flatten 操作变成一维向量,得到用于计算注意力的 Q, K, V 。 F_4 为经过尺度内特征交互模块以计算自注意力后的输出特征图。骨干网络输出的 C_2, C_3, C_4 特征图经过尺度序列特征融合模块 SSFF, 得到输出特征图 F_1 。最后,将 F_1 、尺度内特征交互模块的输出 F_4 以及骨干网络输出的特征图一起经过增强尺度融合模块 EFFM, 得到整个多尺度增强混合编码器的输出结果 Output。具体过程可以表示为:

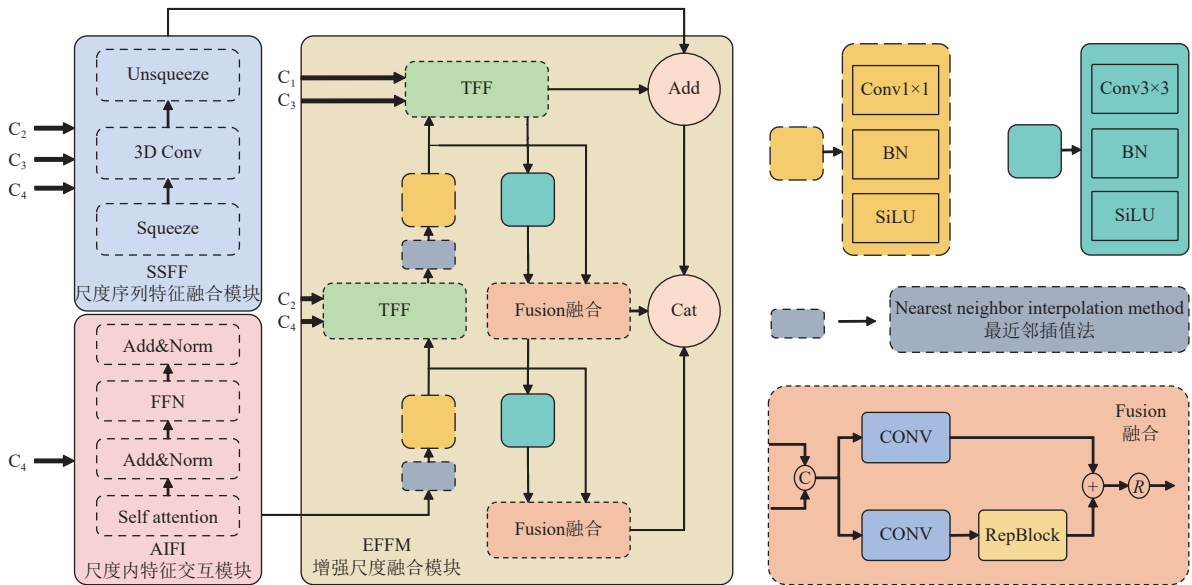
$$Q=K=V= \text{Flatten}(C_4) \quad (1)$$

$$F_4= \text{Reshape}(\text{AIFI}(Q,K,V)) \quad (2)$$

$$F_1= \text{SSFF}(C_2, C_3, C_4) \quad (3)$$

$$\text{Output} = \text{EFFM}(F_1, F_4, C_1, C_2, C_3, C_4) \quad (4)$$

其中, Reshape 表示将特征的形状恢复到与 C_4 相同,是 Flatten 的逆操作, AIFI 代表经过尺度内特征交互模块, SSFF 代表经过尺度序列特征融合模块, EFFM 代表经过增强尺度融合模块。

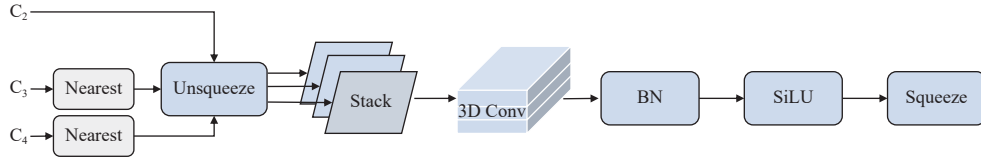


注: Unsqueeze 表示维度扩张操作; Squeeze 表示维度压缩操作; Add 表示叠加操作; Norm 表示归一化操作; Self Attention 表示自注意力操作; Cat 表示拼接操作; SiLU 表示激活函数; RepBlock 表示可重参数化块; R 表示重组操作; Conv1×1 表示卷积核为 1×1 的卷积; Conv3×3 表示卷积核为 3×3 的卷积。
Note: Unsqueeze represents dimension expansion operation; Squeeze represents dimension compression operation; Add represents overlay operation; Norm represents normalization operation; Self Attention represents self attention operation; Cat represents splicing operations; SiLU represents the activation function; RepBlock represents reparameterized block; R represents recombination operation; Conv1×1 represents the convolution kernel is a 1×1 convolution; Conv3×3 represents the convolution kernel is a 3×3 convolution.

图 2 多尺度增强混合编码器
Fig.2 multi-scale enhanced hybrid encoder (MSEHE)

1.2.3 尺度序列特征融合模块

尺度序列特征融合模块 (SSFF) [35] 是一个独立引入的分支用以融合不同尺度的特征图。该模块使用 3D 卷积提取 $\{C_2, C_3, C_4\}$ 特征图的尺度序列, 由于浅层特征 C_2 没有经过大幅度的下采样, 保留了原始麦穗图像中大量的细节, 因此, 该模块采用将高层特征图上采样到与 C_2 一致的分辨率, 以保留更多利好小目标检测的信息。图 3 展示了尺度序列特征融合模块的具体结构, 首先使



注: Nearest 表示最近邻插值法; Stack 表示堆叠操作。

Note: Nearest represents nearest neighbor interpolation method; Stack represents stacking operation.

图 3 尺度序列特征融合模块

Fig.3 Scale sequence feature fusion(SSFF)

1.2.4 三重特征融合模块

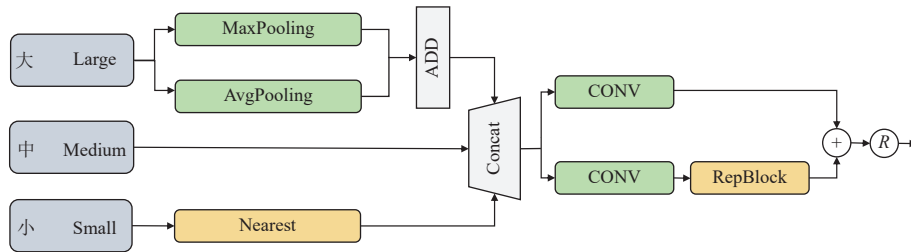
由于麦穗生长环境的差异以及其生长过程中的非一致性, 导致尺度差异较大。对于不同尺度的麦穗, 深层特征和浅层特征都非常重要。为了在保持计算量的同时, 把深层和浅层特征充分融合, 本文设计了三重特征融合模块。相对于原始的 fusion 模块中, 加入浅层的特征图, 给模型提供更全面的全局信息。

图 4 说明了三重特征融合模块 (TFF) 的实现过程: 大尺寸特征图 (Large) 采用最大池化和平均池化混合结构来进行下采样, 使分辨率尺寸与中等尺寸 (Medium)

的特征图保持一致, 这样做有利于保留高分辨率的特征。对于小尺寸 (Small) 的特征图, 使用最近邻插值方法进行上采样来与中等尺寸特征图保持一致, 这有助于保持低分辨率图像局部特征的丰富性, 防止小目标特征信息的丢失。最后将尺寸相同的大、中、小 3 个特征图在通道维度进行拼接。如式 (5) 所示:

$$F_{TFF} = \text{Concat}(F_l, F_m, F_s) \quad (5)$$

其中 F_{TFF} 表示 TFF 模块的输出, F_l, F_m, F_s 分别表示大尺寸、中等尺寸和小尺寸的特征张量。为了减少计算量, 本文仅对前两个融合使用三重特征融合。



注: MaxPooling 表示最大池化操作; AvgPooling 表示平均池化操作; Nearest 为最近邻插值法; ADD 为叠加操作; Concat 为维度拼接操作。

Note: MaxPooling represents the maximum pooling operation; AvgPooling represents the average pooling operation; Nearest represents the nearest neighbor interpolation method; ADD represents stacking operation; Concat represents dimension concatenation operation.

图 4 三重特征融合模块

Fig.4 Triple feature fusion (TFF)

增强尺度融合模块 (EFFM) 通过三重特征融合模块 (TFF) 和 Fusion 来进行尺度间的交互。其中三重特征融合模块 (TFF) 以三种尺度的特征图进行融合, Fusion 以两种尺度特征图来进行融合。从而达到保持模型计算量同时, 提升模型对不同尺度麦穗检测精度的目的。

1.2.5 损失函数

RT-DETR 使用的 GIoU 损失函数通过引入预测框和真实框的最小外接矩形来获取预测框、真实框在最小外接矩形区域中的比重, 从而解决了两个目标没有交集时

梯度为零的问题。但大田麦穗图像采集和人工标注会带来图像真实锚框标注不准确的问题, 而 GIoU 损失函数仅反映预测框和真实框的重合度, 并未考虑在低质量数据图像下真实锚框标注不准确的问题。WIoUv3 [36] 不仅考虑了重叠面积, 而且还引入了动态非单调聚焦机制, 结合了合理的梯度增益分配策略, 减少了极端样本中出现的大梯度或有害梯度。因此, 该损失函数更加关注普通质量的样本, 从而提高了网络模型的泛化能力和整体性能。基于此, 本文将原始的 GIoU 替换为 WIoUv3 边界框损失函数。其计算式为:

$$L_{WIoU_{v3}} = rR_{WIoU}L_{IoU}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \quad (6)$$

其中

$$R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \quad (7)$$

R_{WIoU} 用于放大普通适量锚框的 L_{IoU} , (x, y) 和 (x_{gt}, y_{gt}) 分别为预测框和真实框的中心点坐标, W_g 和 H_g 是最小包围框的尺寸, α 和 δ 为超参数。通过引入参数离群度 β 来描述锚框的质量, 它与锚框质量负相关, 离群度小意味着锚框质量高, 离群度的计算式为:

$$\beta = \frac{L_{IoU}^*}{L_{IoU}}, \beta \in [0, +\infty) \quad (8)$$

式中 L_{IoU}^* 为单调聚焦系数, L_{IoU}^* 随着 L_{IoU} 的减小而减小, $\overline{L_{IoU}}$ 为滑动平均值。引入 $\overline{L_{IoU}}$ 意味着可以根据训练进程动态调整最高梯度增益。对离群度较大的锚框分配较小的梯度增益, 将有效防止低质量图像数据产生较大的有害梯度。WIoUv3 采用合理的梯度增益分配策略, 动态优化损失中高质量和低质量锚框的权重, 使模型聚焦于平均质量样本, 增强模型对麦穗目标的定位准确度。

1.3 评价指标

本文采用标准的 COCO AP 评价指标来评价模型性能, 即: AP_{50} 、 AP_{75} 、 AP_{50-95} 、 AP_S 、 AP_M 、 AP_L 。AP(average precision) 是指 P-R 曲线围成的面积, 用来衡量模型对一个类检测的好坏。平均精度的公式表示为:

$$AP = \int_0^1 P(r)dr \quad (9)$$

IoU 是指预测框和真实框交集和并集的比值, 这个比值用于衡量预测框和真实框的重叠程度。 AP_{50} 、 AP_{75} 分别表示在 IoU 阈值为 0.50、0.75 下的平均精度值, AP_{50-95} 表示在 IoU 阈值 0.50~0.95 (步长 0.05) 下的平均精度均值, AP_{50} 、 AP_{75} 、 AP_{50-95} 是对模型精度的整体概括, AP_S 表示在小目标范围内的 AP_{50-95} 的值, AP_M 表示在中等目标范围内的 AP_{50-95} 的值, AP_L 表示在大目标

范围内的 AP_{50-95} 的值。

本文采用参数量 (parameters)、浮点数计算量 (floating point operations, FLOPs) 来衡量模型的复杂度, 采用帧率 (frames per second, FPS) 来衡量模型的实时性。

2 试验结果与分析

2.1 网络训练

本文试验采用 Ubuntu 操作系统、NVIDIA GTX 4 080 GPU, 并基于 PyTorch2.1.1、CUDA12.2、Python3.8 来实现。训练参数设置如下: 批次为 4, 训练轮次是 200, 学习率设置为 10^{-4} 且采用的是余弦退火对学习率进行调整, 采用 AdamW 优化器。

2.2 骨干网络消融试验

本文对骨干网络进行消融试验, 选择的有 RT-DETR 原始的骨干网络 HGNetV2、Resnet50 及轻量化骨干网络 MobilenetV3^[37]、Fasternet^[38]、EfficientViT^[39]、EfficientFormerV2。各个骨干网络对比的结果如表 1 所示, 在参数量仅为 10 M 的量级下, EfficientFormerV2 的平均精度值最高, 实现了计算量和精度的最好平衡。

表 1 加入不同骨干网络后模型训练对比结果

Table 1 The comparative results of model training after integrating different backbone networks

| 骨干网络 Backbone | 参数量 Parameters/M | 浮点数运算量 Floating point operations FLOPs/G | 平均精度值 Average precision /% |
|-------------------|---------------------|--|-------------------------------|
| HGNetV2 | 32.0 | 103.4 | 89.7 |
| Resnet50 | 42.0 | 129.5 | 90.4 |
| MobilenetV3 | 9.5 | 23.6 | 80.6 |
| Fasternet | 10.8 | 28.5 | 88.6 |
| EfficientViT | 10.7 | 27.2 | 88.1 |
| EfficientFormerV2 | 11.7 | 29.4 | 89.5 |

2.3 消融试验

本文采用全球麦穗数据集划分的 639 张麦穗图像作为测试集来进行消融试验, 逐一加入改进方案来验证提出的模块对于模型性能的影响。表 2 为添加每个模块的试验结果。

表 2 各改进模块消融试验结果

Table 2 Results of ablation experiments for various improvement modules

| 试验编号 Test No. | EfficientFormerV2 | MSEHE | WIoU | 参数量 Parameters/M | 浮点运算量 FLOPs/G | 帧率 Frames per second/ (帧·s ⁻¹) | $AP_{50-95}/\%$ | $AP_{50}/\%$ | $AP_{75}/\%$ | $AP_S/\%$ | $AP_M/\%$ | $AP_L/\%$ |
|------------------|-------------------|-------|------|---------------------|------------------|--|-----------------|--------------|--------------|-----------|-----------|-----------|
| 1 | × | × | × | 32.0 | 103.4 | 65.1 | 51.1 | 89.7 | 51.6 | 17.5 | 50.5 | 61.7 |
| 2 | √ | × | × | 11.7 | 29.4 | 83.2 | 50.8 | 89.5 | 51.2 | 14.8 | 50.3 | 61.3 |
| 3 | √ | √ | × | 12.0 | 33.1 | 79.5 | 51.8 | 89.6 | 53.1 | 17.0 | 51.1 | 62.8 |
| 4 | √ | √ | √ | 12.0 | 33.1 | 79.7 | 51.7 | 90.2 | 53.1 | 17.5 | 51.0 | 62.4 |

注: 试验 1 为 RT-DETR 模型; 试验 4 为 RT-WEDT 模型; “×” 表示未加入此模块; “√” 表示加入此模块; AP_{50-95} 表示交并比阈值 0.50~0.95 的平均精度均值; AP_{50} 、 AP_{75} 分别表示交并比阈值为 0.50、0.75 的平均精度值; AP_S 、 AP_M 、 AP_L 分别表示小、中、大目标的平均精度值。下同。

Note: Test 1 is the RT-DETR model; Test 4 is the RT-WEDT model; “×” represents that the module is not added; “√” represents that the module is added to this module; AP_{50-95} represents the average precision mean value for the intersection and merger ratio threshold of 0.50 to 0.95; AP_{50} and AP_{75} represents the average accuracy values for the intersection and merger ratio thresholds of 0.50 and 0.75, respectively; AP_S , AP_M and AP_L represents the average accuracy values for small, medium and large targets, respectively. The same below.

从表 2 可以看出, RT-DETR 把原始骨干网络 HGNetV2 更换成 EfficientFormerV2 后, AP_{50} 和 AP_{50-95} 分别下降了 0.2 和 0.3 个百分点, 其中骨干网络的替换对小目标的检测精度损耗较大, 降低了 2.7 个百分点, 但模型的整体参数量和浮点数运算量降低了 63.4% 和 71.6%, FPS 上升了 27.8%。在此基础上, 加入多尺度增强混合

编码器后参数量仅上升了 0.3M, 但 AP_{50} 和 AP_{50-95} 分别提高了 0.1 和 1.0 个百分点, 此外, 在小目标的检测精度指标 AP_S 上有 2.2 个百分点的提升, 在中等目标上的检测精度 AP_M 提升了 0.8 个百分点, 在大目标上的检测精度 AP_L 提升了 1.5 个百分点, 验证了本文提出的多尺度增强混合编码器可以提升模型对不同尺度麦穗目标的检

测能力。更换 $WIoU_{v3}$ 损失函数后，在参数量保持不变的情况下， AP_{50} 和小目标 AP_s 分别提高了 0.6 和 0.5 个百分点。相较于 RT-DETR，本文提出的模型 RT-WEDT 参数量和浮点数计算量分别降低了 62.5% 和 68%，FPS 达到 79.7 帧/s，检测速度提高了 22.4%，具备了较好的实时检测能力。由于本文模型考虑了轻量化后对小目标的精度损失，使得最终模型在大幅度降低模型参数的同时，检测精度指标 AP_{50} 达到了 90.2%， AP_{50-95} 提升了 0.6 个百分点。证明了本文所改进模块的有效性。

为了更加直观的验证改进模块的有效性，本文利用热力图来体现网络对麦穗目标特征的捕获能力，在 RT-DETR 的基准下逐渐加入本文采用的各个模块，并与原始的 RT-DETR 对比。图 5 的试验结果表明，原始的 RT-DETR 将注意力集中于麦穗、麦穗杆径和部分背景，而 RT-WEDT 将注意力更加偏向于麦穗，减少了背景和麦穗杆径的关注程度。

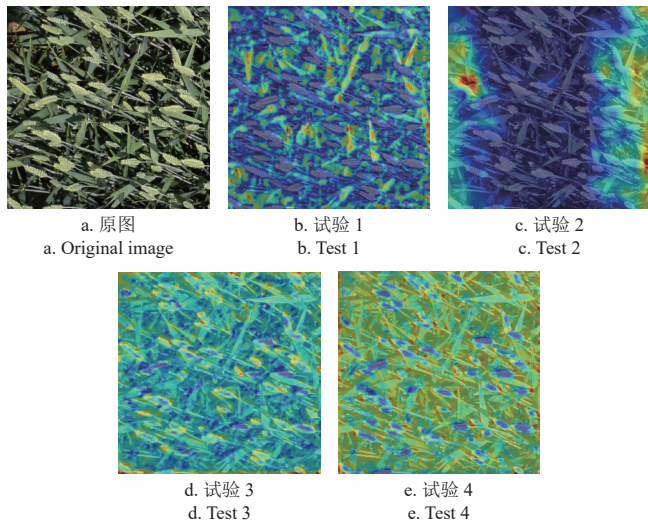


图 5 不同改进模块热力图可视化结果

Fig.5 Visualization results of heatmaps for different improvement module

2.4 与不同网络的对比

为比较本文模型与当前主流的目标检测模型在田间麦穗的检测效果，本文选择了 SSD^[40]、Faster-RCNN^[41]、YOLOv4^[42]、YOLOX^[43]、Centernet^[44]、Retinanet^[45] 等基于 CNN 的检测模型，以及基于 Transformer 的网络模型 DETR。

表 3 为本文所提出的 RT-WEDT 与其他主流目标检测模型在全球麦穗数据集上评价指标的对比结果，从表 3 可以看出，RT-WEDT 的参数量、FLOPs、 AP_{50-95} 和 AP_{50} 分别为 12M、33.1G、51.7% 和 90.2%。对比近年来比较有优势的轻量化 CNN 网络，RT-WEDT 的 AP_{50} 比 YOLOv5、YOLOv8 和 YOLOX 分别高 4.6、1.1 和 0.7 个百分点。与经典的 CNN 网络 Retinanet 和 YOLOv4 相比， AP_{50} 分别高出 5.1 个百分点和 2.3 个百分点。由于 CNN 这类模型缺乏浅层特征图，无法识别局部小目标麦穗，而 RT-WEDT 加入了更多的浅层特征，所以多数 CNN 模型在小目标上的检测精度 AP_s 都远低于 RT-WEDT。与 Transformer 网络 DETR 和 RT-DETR 相比，本文提出的模型参数量仅为它们的三分之一，不仅在参数上具有较大的优势， AP_{50} 也分别高出 8.3 和 0.5 个百分点。综上，表 3 的指标证明了本研究提出的模型的有效性，在大、中、小三种尺度上的精度均为最高。

图 6 为各模型在 GWHD 数据集上麦穗检测的细节图，红色矩形框为模型预测框，蓝色矩形框为模型漏检的麦穗，黄色矩形框为模型产生的冗余检测框。图 6 麦穗与背景相似且难以区分，导致了 CNN 模型均出现了漏检，而 YOLOv5、YOLOv8、YOLOX 对小目标检测能力具有局限性，所以这些网络对小目标麦穗漏检的数量较多。本文提出的 RT-WEDT 可以充分提取图像的全局特征，在图 6 的示例中综合检测效果最好，体现了本文算法的优越性。

表 3 本研究方案与主流模型对比

Table 3 Comparison of our research approach with mainstream models

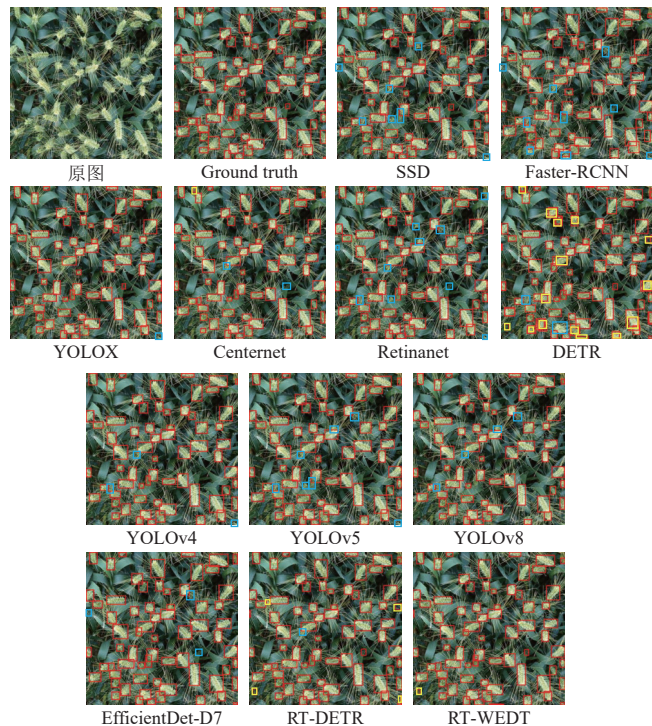
| 模型 Model | 骨干网络 Backbone | 参数量 Parameters/M | FLOPs/G | $AP_{50-95}/\%$ | $AP_{50}/\%$ | $AP_{75}/\%$ | $AP_s/\%$ | $AP_M/\%$ | $AP_L/\%$ |
|-----------------|-------------------|---------------------|---------|-----------------|--------------|--------------|-----------|-----------|-----------|
| SSD | Vgg | 23.6 | 60.8 | 31.5 | 70.1 | 22.8 | 0.8 | 28.1 | 50.1 |
| Faster-RCNN | Resnet50 | 28.3 | 948.1 | 22.0 | 62.3 | 9.50 | 0.40 | 17.4 | 40.2 |
| YOLOv4 | CSPDarknet | 64.3 | 60.5 | 43.9 | 87.9 | 37.7 | 13.2 | 43.1 | 53.6 |
| YOLOv5 | CSPDarknet | 7.1 | 16.5 | 43.5 | 85.6 | 38.8 | 9.9 | 42.7 | 53.9 |
| YOLOv8 | CSPDarknet | 11.1 | 28.6 | 49.3 | 89.1 | 48.9 | 15.7 | 48.6 | 59.5 |
| YOLOX | CSPDarknet | 8.9 | 26.8 | 50.0 | 89.5 | 49.8 | 17.5 | 49.2 | 59.5 |
| Centernet | Resnet50 | 32.7 | 109.7 | 46.2 | 87.8 | 43.3 | 12.0 | 45.2 | 56.9 |
| Retinanet | Resnet50 | 36.3 | 163.5 | 44.0 | 85.1 | 40.6 | 2.2 | 43.6 | 55.8 |
| DETR | Resnet50 | 36.7 | 73.6 | 39.3 | 81.9 | 32.1 | 6.8 | 37.9 | 51.4 |
| EfficientDet-D7 | EfficientNet | 51.5 | 629.9 | 37.9 | 80.7 | 29.3 | 9.0 | 38.6 | 44.0 |
| RT-DETR | HGNetV2 | 32.0 | 103.4 | 51.1 | 89.7 | 51.6 | 17.5 | 50.5 | 61.7 |
| RT-WEDT | EfficientFormerV2 | 12.0 | 33.1 | 51.7 | 90.2 | 53.1 | 17.5 | 51.0 | 62.4 |

2.5 不同场景下的精度分析

为了深入了解环境因素对模型检测精度的影响，本文从全球麦穗数据集中手动筛选出图 7 所示的 4 个场景下的麦穗图像进行检测精度分析，每个场景下各 20 幅麦穗图像，由于自然风、相机不稳等原因导致图片模糊是

让模型精度损失的最重要原因，对于整体而言，评价指标 AP_{50-95} 和 AP_{50} 分别降低了 7.8 个百分点和 1.9 个百分点；其次，强光、麦穗密集和重叠也是影响模型精度的原因之一， AP_{50-95} 和 AP_{50} 分别降低了 4.7 个百分点和 4.4 个百分点。而强光或弱光情况下麦穗图像，对模型精

度影响较小。图 7 展示了每个场景的原图、真实值图及预测图。从图 7 的可视化效果可以看出, 密集重叠导致麦穗遮挡难以被精确检测到、图像模糊和强光下会使得目标与背景难以区分, 网络对这几类情况下的麦穗检测精度会存在局限性, 与表 4 的客观评价指标一致。



注: 红色矩形框为模型预测框, 蓝色矩形框为模型漏检的麦穗, 黄色矩形框为模型产生的冗余检测框。
Note: The red rectangular box is the model prediction box, the blue rectangular box is the wheat missed by the model, and the yellow rectangular box is the redundant detection box generated by the model.

图 6 不同模型的麦穗检测结果
Fig.6 Wheat spike detection results of different models

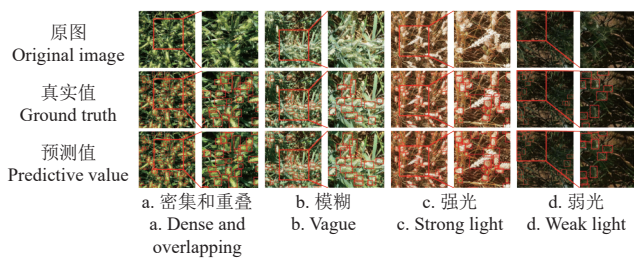


图 7 不同场景下麦穗检测情况可视化结果

Fig.7 Visualization results of wheat spike detection in different scenarios

2.6 计数性能分析

为了进一步体现本文 RT-WEDT 模型的计数性能, 本文选择与一些经典且性能优异的计数模型 MCNN^[46]、CSRnet^[47]、TasselNetV2^[3]、TasselNetV2plus 在计数性能对比。对图 8 所示的原图进行人工标注, 真实值为 56。采用平均绝对误差 (mean absolute error, MAE)、均方根误差 (root mean squared error, RMSE) 作为评估指标。其中, MAE 值越小, 代表着真实值与预测值误差越小; RMSE 值越小, 代表着预测值与真实值分散程度越小; R^2 代表趋势线的拟合程度, 其越接近于 1, 拟合程度越高。

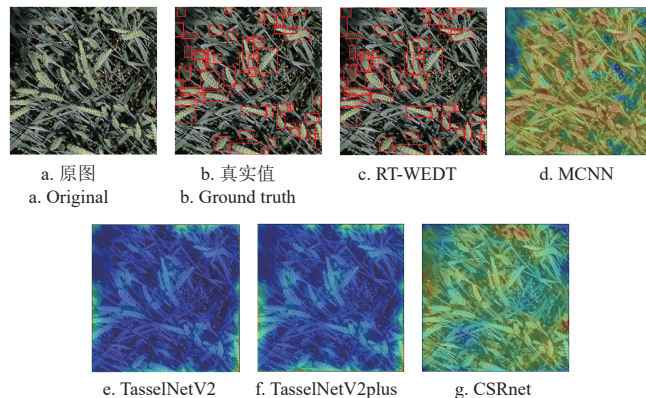


图 8 RT-WEDT 与不同计数模型的可视化结果

Fig.8 Visualization results of RT-WEDT and different counting model

所有的模型均在全球麦穗数据集上进行训练和测试, 不同方法对比结果如表 4 所示。从表 4 可以看出, 本文提出的 RT-WEDT 的 MAE 为 9.19, RMSE 为 11.62, R^2 为 0.94, 线性拟合程度较高, 计数性能略高于 MCNN、TasselNetV2 和 TasselNetV2plus。其中 CSRnet 的 MAE 为 6.84, RMSE 为 8.73, 在全球麦穗数据集测试集上表现最好。通过图 8 中对 4 种计数网络进行可视化展示可以看出, 相较于基于密度图回归的计数方法 (图 8d、图 8e、图 8f、图 8g), 检测计数的方法 (图 8c) 可以对麦穗目标进行精准的定位和识别, 从而了解到麦穗的大小、形状和姿态等表型信息, 有助于育种专家评估麦穗的品质和价值。

表 4 不同计数模型性能比较

Table 4 Comparison of the performance of different counting models

| 模型 Model | 预测值 Predicted value | MAE | RMSE | R^2 |
|-----------------|---------------------|-------|-------|-------|
| MCNN | 62.65 | 11.13 | 14.54 | 0.90 |
| TasselNetV2 | 58.91 | 9.70 | 13.41 | 0.92 |
| TasselNetV2plus | 59.95 | 9.67 | 13.40 | 0.92 |
| CSRnet | 62.65 | 6.84 | 8.73 | 0.96 |
| RT-WEDT | 60.00 | 9.19 | 11.62 | 0.94 |

2.7 模型鲁棒性验证

为了验证本研究的有效性, 本文将在全球麦穗数据集上训练的模型在自建的无人机视角麦穗数据集 DPWSD 上进行鲁棒性测试。数据集由 DJI Phantom 4 Pro V2.0 无人机于 2022 年 4 月 11 日 (灌浆期) 和 2022 年 5 月 3 日 (成熟期) 在中国农业科学院阳逻综合试验基地获取。无人机采集的原始图像分辨率为 5 472×3 648 像素, 飞行高度 8 m。考虑到 GPU 的运算内存, 本文将图像裁剪为 1 024×1 024 像素, 并从拍摄的图像中去除模糊和畸变严重的图像, 最后总共选出灌浆期小麦图像 28 幅, 成熟期小麦图像 23 幅。之后使用 labelImg 软件对小麦图像进行标注, 形成本文的 DPWSD 数据集。图 9 为本研究采集的无人机视角麦穗数据的试验田区域图。无人机视角麦穗数据集与全球麦穗数据集相比视角更广, 麦穗更为密集, 背景更加复杂, 有大量重叠^[48], 麦穗尺寸不一, 背景更加复杂, 很好模拟了自然田间俯拍麦穗图像的情况。

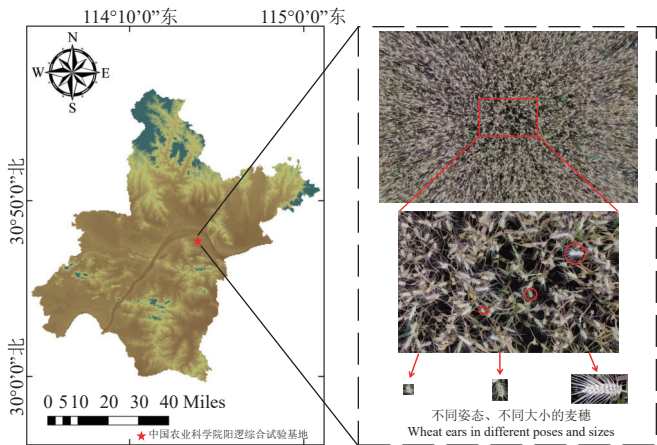


图 9 研究区域地理位置及无人机遥感麦穗图像
Fig.9 Geographic location of the study area and UAV remote sensing images of wheat ears

本文提出的模型 RT-WEDT 在 DPWSD 数据集上的测试结果如表 5 所示，在灌浆期和成熟期的 AP_{50} 可分别达到 97.4% 和 96.1%， AP_{50-95} 也可达到 60.2% 和 61.0%。由于本文制作的无人机视角麦穗数据集仅涉及两个生育期且麦穗品种单一，所以检测精度高于全球麦穗数据集，但是上述结果仍然进一步验证了该模型在不同视角下有较强的麦穗检测能力和泛化能力。

表 5 本研究方案在麦穗不同时期的指标

| 时期 Period | $AP_{50-95}/\%$ | $AP_{50}/\%$ | $AP_{75}/\%$ | $AP_S/\%$ | $AP_M/\%$ | $AP_L/\%$ |
|------------------------|-----------------|--------------|--------------|-----------|-----------|-----------|
| 总体 Population | 60.8 | 96.8 | 68.4 | 54.6 | 60.2 | 65.9 |
| 灌浆期 Grouting period | 60.2 | 97.4 | 66.8 | 20.1 | 60.6 | 61.8 |
| 成熟期 Mature period | 61.0 | 96.1 | 69.4 | 55.3 | 59.8 | 71.6 |

图 10 展示了本文方法在 DPWSD 数据集的检测结果。从图 10a 可以看出，虽然灌浆期麦穗形状大小不一，但是颜色特征较为明显，本研究提出的模型能够精准的检测到。从图 10b 可以看到，在成熟期时，由于此时麦穗和麦穗叶都呈现灰色，颜色特征相近，模型有少数误检的情况。

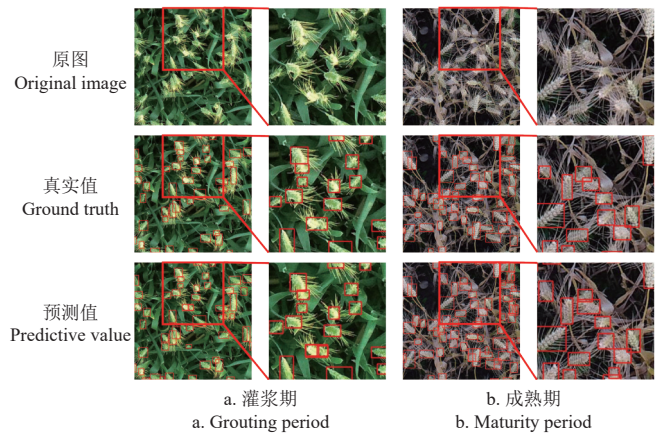


图 10 DPWSD 数据集检测结果可视化图
Fig.10 Visualization of the detection results of the DPWSD dataset

为了验证本文提出检测模型在自建数据集 DPWSD 上的计数效果，本文对 51 张 DPWSD 数据集中的麦穗图像进行计数试验，将计数结果与标签的真实值进行比对，结果如图 11 所示。从图 11 可以看出，无人机视角麦穗数据集的预测值和真实值拟合的 R^2 为 0.9499，线性拟合线可以有效反映预测值与真实值的关系，拟合程度较高。总体而言，本文提出的轻量化模型 RT-WEDT 鲁棒性较好，可以准确地对麦穗进行检测并计数，进而实现精准的小麦估产。

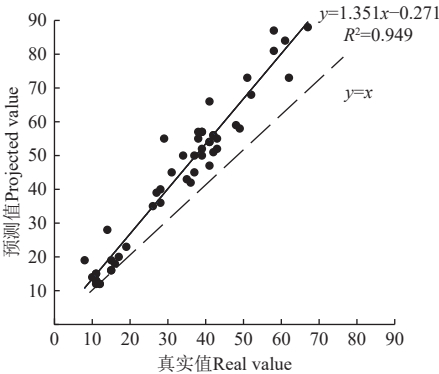


图 11 RT-WEDT 在无人机视角麦穗数据集上预测值和真实值的拟合结果
Fig.11 RT-WEDT fitting results of predicted and true values on the UAV view wheat dataset

3 结 论

麦穗的产量估计对于麦穗作物育种至关重要。本文提出了一种用于麦穗检测的轻量化模型 RT-WEDT。首先采用轻量化的骨干网络 EfficientFormerV2 作为 RT-WEDT 的主干网络，其次提出了一种新颖的多尺度增强混合编码器来更好的融合多个尺度特征图的信息，提升对不同尺度麦穗目标的检测能力，最后使用 Wise-IoU 边界框损失函数替换原有的 GIoU 边界框损失函数，实现了对麦穗的快速精准的实时检测的目的。RT-WEDT 的麦穗检测的平均精度为 90.2%、参数量为 12M、浮点数运算量为 33.1G、检测速度为 79.7 帧/s。模型相较于原始的 RT-DETR，模型参数量降低了 62.5%、浮点数计算量降低了 68%、交并比阈值 0.50~0.95 的平均精度均值提升了 0.6 个百分点、交并比阈值 0.50 的平均精度提升了 0.5 个百分点、检测速度提高了 22.4%。RT-WEDT 不仅参数量小，且对不同尺度目标麦穗有较好的检测效果，与 SSD、YOLOv4、YOLOv5s、YOLOv8、Centernet、DETR 等网络相比，RT-WEDT 的检测精度最高、漏检麦穗数最少、综合表现最优、更适合移动端部署。除此之外，本文还对 RT-WEDT 进行鲁棒性测试，使用 RT-WEDT 对无人机视角麦穗数据集进行验证，分别在麦穗的灌浆期和成熟期取得 97.4% 和 96.1% 的平均精度，验证了本模型的鲁棒性，可以为后续的麦穗产量估计提供更有效的技术支撑。

[参 考 文 献]

[1] ZHOU M, ZHENG H B, HE C, et al. Wheat phenology

- detection with the methodology of classification based on the time-series UAV images[J]. *Field Crops Research*, 2023, 292: 108798.
- [2] HASAN M M, CHOPIN J P, LAGA H, et al. Detection and analysis of wheat spikes using convolutional neural networks[J]. *Plant Methods*, 2019, 15: 27.
- [3] XIONG H P, CAO Z G, LU H, et al. TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks[J]. *Plant Methods*, 2019, 15(1): 150.
- [4] PAN Y Y, ZHU N Z, DING L, et al. Identification and counting of sugarcane seedlings in the field using improved Faster R-CNN[J]. *Remote Sensing*, 2022, 14(22): 5846.
- [5] 李毅念, 杜世伟, 姚敏, 等. 基于小麦群体图像的田间麦穗计数及产量预测方法[J]. *农业工程学报*, 2018, 34(21): 185-194. (in Chinese with English abstract).
- LI Yinian, DU Shiwei, YAO Min, et al. Method for wheatear counting and yield predicting based on image of wheatear population in field[J]. *Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE)*, 2018, 34(21): 185-194. (in Chinese with English abstract).
- [6] 刘涛, 孙成明, 王力坚, 等. 基于图像处理技术的大田麦穗计数[J]. *农业机械学报*, 2014, 45(2): 282-290. (in Chinese with English abstract).
- LIU Tao, SUN Chengming, WANG Lijian, et al. In-field Wheatear Counting Based on Image Processing Technology[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2014, 45(2): 282-290. (in Chinese with English abstract).
- [7] 杜颖, 蔡义承, 谭昌伟, 等. 基于超像素分割的田间小麦穗数统计方法[J]. *中国农业科学*, 2019, 52(1): 21-33. (in Chinese with English abstract).
- DU Ying, CAI Yicheng, TAN Changwei, et al. Field wheat ears counting based on superpixel segmentation method[J]. *Scientia Agricultura Sinica*, 2019, 52(1): 21-33. (in Chinese with English abstract).
- [8] 刘哲, 黄文准, 王利平. 基于改进 K-means 聚类算法的大田麦穗自动计数 [J]. *农业工程学报*, 2019, 35(3): 174-181.
- LIU Zhe, HUANG Wenzhun, WANG Liping Field wheat ear counting automatically based on improved K-means clustering algorithm [J]. *Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE)*, 2019, 35(3): 174-181. (in Chinese with English abstract).
- [9] 范梦扬, 马钦, 刘峻明, 等. 基于机器视觉的大田环境小麦麦穗计数方法[J]. *农业机械学报*, 2015, 46(S1): 234-239.
- FAN Mengyang, MA Qin, LIU Junming, et al. Counting method of wheatear in field based on machine vision technology[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2015, 46(S1): 234-239. (in Chinese with English abstract).
- [10] 鲍文霞, 张鑫, 胡根生, 等. 基于深度卷积神经网络的田间麦穗密度估计及计数[J]. *农业工程学报*, 2020, 36(21): 186-193. (in Chinese with English abstract).
- BAO Wenxia, ZHANG Xin, HU Gensheng, et al. Estimation and counting of wheat ears density in field based on deep convolutional neural network[J]. *Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE)*, 2020, 36(21): 186-193. (in Chinese with English abstract).
- [11] 孙俊, 杨锴锋, 罗元秋, 等. 基于无人机图像的多尺度感知麦穗计数方法[J]. *农业工程学报*, 2021, 37(23): 136-144. (in Chinese with English abstract).
- SUN Jun, YANG Kaifeng, LUO Yuanqiu, et al. Method for the multiscale perceptual counting of wheat ears based on UAV images[J]. *Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE)*, 2021, 37(23): 136-144. (in Chinese with English abstract).
- [12] WU W, ZHONG X C, LEI C K, et al. Sampling survey method of wheat ear number based on UAV images and density map regression algorithm[J]. *Remote Sensing*, 2023, 15(5): 1280.
- [13] 鲍文霞, 苏彪彪, 胡根生, 等. 基于 FE-P2Pnet 的无人机小麦图像麦穗计数方法[J]. *农业机械学报*, 2024, 55(4): 155-164. (in Chinese with English abstract).
- BAO Wenxia, SU Biaobiao, HU Gensheng, et al. Method for Counting Wheat Ears in UAV Images Based on FE-P2Pnet[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2024, 55(4): 155-164. (in Chinese with English abstract).
- [14] WANG S W, ZHAO J Q, CAI Y C, et al. A method for small-sized wheat seedlings detection: from annotation mode to model construction[J]. *Plant Methods*, 2024, 20(1): 15.
- [15] 杨蜀秦, 王帅, 王鹏飞, 等. 改进 YOLOX 检测单位面积麦穗[J]. *农业工程学报*, 2022, 38(15): 143-149. (in Chinese with English abstract).
- YANG Shuqin, WANG Shuai, WANG Pengfei, et al. Detecting wheat ears per unit area using an improved YOLOX[J]. *Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE)*, 2022, 38(15): 143-149. (in Chinese with English abstract).
- [16] HE M X, HAO P, XIN Y Z. A robust method for wheatear detection using UAV in natural scenes[J]. *IEEE Access*, 2020, 8: 189043-189053.
- [17] MENG X, LI C, LI J. YOLOv7-MA: Improved YOLOv7-based wheat head detection and counting[J]. *Remote Sensing*, 2023, 15(15): 3770.
- [18] ZHANG R, YAO M, QIU Z. Wheat teacher: a one-stage anchor-based semi-supervised wheat head detector utilizing pseudo-labeling and consistency regularization methods[J]. *Agriculture*, 2024, 14(2): 327.
- [19] ZHANG D Y, LUO H S, CHENG T. Enhancing wheat Fusarium head blight detection using rotation Yolo wheat detection network and simple spatial attention network[J]. *Computers and Electronics in Agriculture*, 2023, 211: 107968.
- [20] HARADA S, HAN X H. A Hybrid Wheat Head Detection model with Incorporated CNN and Transformer[C]// *Proceedings of the 18th International Conference on Machine*

- Vision and Applications (MVA). Hamamatsu, Japan: IEEE, 2023: 1-5.
- [21] ZHU J P, YANG G F, FENG X P, et al. Detecting wheat heads from UAV low-altitude remote sensing images using deep learning based on transformer[J]. *Remote Sensing*, 2022, 14(20): 5141.
- [22] ZHOU Q, HUANG Z I, ZHENG S J, et al. A wheat spike detection method based on Transformer[J]. *Frontiers in Plant Science*, 2022, 13: 1023924.
- [23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in neural information processing systems(NIPS). Long Beach, CA, 2017: 5998-6008.
- [24] KHAN S, NASEER M, HAYAT M, et al. Transformers in Vision: A Survey[J]. *ACM Computing Surveys*, 2022, 54(10s): 200.
- [25] SALAMAI A A, AJABNOOR N, KHALID W E, et al. Lesion-aware visual transformer network for paddy diseases detection in precision agriculture[J]. *European Journal of Agronomy*, 2023, 148: 126884.
- [26] THAI H T, LE K H, NGUYEN N L T. FormerLeaf: An efficient vision transformer for Cassava Leaf Disease detection[J]. *Computers and Electronics in Agriculture*, 2023, 204: 107518.
- [27] YE J X, YU Z H, WANG Y X, et al. WheatLFANet: In-field detection and counting of wheat heads with high-real-time global regression network[J]. *Plant Methods*, 2023, 19(1): 103.
- [28] LV W Y, XU S L, ZHAO Y, et al. DETRs beat YOLOs on real-time object detection [EB/OL]. arXiv preprint arXiv: 230408069, 2024. (2023-04-17). <https://arxiv.org/abs/2304.08069>.
- [29] LI Y Y, HU J, WEN Y, et al. Rethinking Vision Transformers for MobileNet Size and Speed[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 16843-16854.
- [30] DAVID E, SEROUART M, SMITH D, et al. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods[J]. *Plant Phenomics*, 2021: 9846158. doi: 10.34133/2021/9846158.
- [31] ZHAO J Q, ZHANG X H, YAN J W, et al. A wheat spike detection method in UAV images based on improved YOLOv5[J]. *Remote Sensing*, 2021, 13(16): 3095.
- [32] 鲍文霞, 谢文杰, 胡根生. 基于 TPH-YOLO 的无人机图像麦穗计数方法[J]. *农业工程学报*, 2023, 39(1): 155-161. (in Chinese with English abstract).
BAO Wenxia, XIE Wenjie, HU Gensheng, et al. Wheat ear counting method in UAV images based on TPH-YOLO[J]. *Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, 2023, 39(1): 155-161. (in Chinese with English abstract).
- [33] ZHAO J Q, CAI Y C, WANG S W. Small and oriented wheat spike detection at the filling and maturity stages based on WheatNet[J]. *Plant Phenomics*, 2023, 5: 0109.
- [34] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of the European Conference on Computer Vision(ECCV). Glasgow, UK: Springer, 2020: 213-229.
- [35] KANG M, TING C M, TING F F, et al. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell Instance segmentation [EB/OL]. arXiv preprint arXiv: 06458, 2023. (2023-12-11). <https://arxiv.org/abs/2312.06458>.
- [36] TONG Z J, CHEN Y H, XU Z W, et al. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism [EB/OL]. arXiv preprint arXiv: 230110051, 2023. (2023-04-08). <https://arxiv.org/abs/2301.10051>.
- [37] HOWARD A, SANDLER M, CHU G. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 1314-1324.
- [38] CHEN J R, KAO S H, HE H. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 12021-12031.
- [39] LIU X, PENG H, ZHENG N. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Vancouver, BC, Canada: IEEE, 2023: 14420-14430.
- [40] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision(ECCV). Amsterdam, the Netherlands: Springer, 2016: 21-37.
- [41] 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.
- [42] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection [EB/OL]. arXiv preprint arXiv: 200410934, 2020. (2020-04-23). <https://arxiv.org/abs/2004.10934>.
- [43] GE Z, LIU S T, WANG F, et al. YOLOX: Exceeding YOLO Series in 2021 [EB/OL]. arXiv preprint arXiv: 210708430, 2021. (2021-08-06). <https://arxiv.org/abs/2107.08430>.
- [44] DUAN K W, BAI S, XIE L X, et al. CenterNet: keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 6568-6577.
- [45] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2980-2988.
- [46] ZHANG Y, ZHOU D, CHEN S. Single-image crowd counting via multi-column convolutional neural network[C]//

- Proceedings of the Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 589-597.
- [47] LI Y, ZHANG X, CHEN D. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the Computer Vision and Pattern Recognition(CVPR). Las Vegas: IEEE, 2018: 1091-1100.
- [48] ZHAO J Q, YAN J W, XUE T J, et al. A deep learning method for oriented and small wheat spike detection (OSWSDet) in UAV images[J]. *Computers and Electronics In Agriculture*, 2022, 198: 107087.

Method for detecting and counting wheat ears using RT-WEDT

LI Jie¹, YANG Zihao¹, ZHENG Quan¹, QIAO Jiangwei², TU Jingmin^{1*}

(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China; 2. Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China)

Abstract: Wheat is one of the most widely cultivated staple food crops in the globe. Its yield prediction can share the profound implications for food security. Deep learning can be expected to detect and count wheat spikes, and then rapidly predict wheat yields. However, some challenges are still remained on in the low detection accuracy and a large number of model parameters under complex agricultural environments. This study aims to propose a lightweight wheat ear detection model, RT-WEDT (Real-Time Wheat Ear Detection Transformer), using RT-DETR. Firstly, EfficientFormerV2 was selected as the backbone network structure of RT-WEDT to fully capture both the long-range and local features of wheat ear images with the high computational efficiency. Secondly, a multiscale enhanced hybrid encoder (MSEHE) was introduced to take as the input feature maps at four scales output from the four downsampling stages of the backbone network. The MSEHE consisted of three sub-modules: the Attention-based intra-scale feature interaction (AIFI) module acted on the smallest feature maps to extract global features of the image; the Scale Sequence Feature Fusion (SSFF) module with multiscale fusion and 3D convolution was utilized to extract information about wheat ear targets at different scales. The outputs of these two modules were fed into the Enhanced Feature Fusion Module (EFFM) for feature fusion, in order to integrate the global and local information of the wheat ear image. Additionally, the localization accuracy was improved for wheat targets. WIoUv3 loss function was employed as the bounding box one to enhance the quality of the anchor frame. The detection dataset was obtained for the global wheat head. Experimental results demonstrate that the RT-WEDT model was had 12M parameters, a floating-point operation capacity of 33.1×10^9 G, an average accuracy of 90.2%, and a detection speed of 79.7 frames/s. Compared with RT-DETR, the RT-WEDT model was had 62.5% fewer parameters, 68% fewer floating-point operations, an AP_{50-95} increase of 0.6%, an AP_{50} increase of 0.5%, and a detection speed increase of 22.4%. The AP_{50-95} values were improved by 8.2%, 2.4%, and 1.7%, respectively, and the AP_{50} values were improved by 4.6%, 1.1%, and 0.7%, respectively, compared with YOLOv5, YOLOv8, and YOLOX with a similar parameter volume. Furthermore, samples were classified from the detection dataset of global wheat heads. The performance was then evaluated on wheat ear targets in various scenarios. The experimental results indicate that the dense and overlapping wheat ears were the most significant influencing factors on the performance of the model, followed by image blurriness. The intensity of light during photography shared the a minimal effect on the detection. Drone The drone perspective wheat spike dataset (DPWSD) was constructed for two periods, in order to verify the robustness of the improved RT-WEDT. And then, the RT-WEDT was directly tested on the drone perspective wheat dataset. Specifically, 60.2% AP_{50-95} and 97.4% AP_{50} were achieved during the filling stage; 61.0% AP_{50-95} and 96.1% AP_{50} were achieved during the maturity stage. The counting experiments were conducted on the test set from the global wheat dataset and the self-built drone perspective wheat ear dataset, respectively, in order to validate the counting effectiveness of RT-WEDT. The R^2 values of RT-WEDT on the global wheat head detection dataset and the DPWSD were 0.94, and 0.95, respectively, indicating an excellent fit between predicted and actual values. Therefore, the RT-WEDT was highly accurate for wheat ear detection and counting. The improved model was significantly reduced the complexity to maintain a high average accuracy, indicating the real-time detection of the wheat ear. This finding can provide the technical support for the efficient and rapid estimation of wheat yields in smart agriculture.

Keywords: model; wheat ear; object detection; transformer; lightweight