微电子学与计算机
Microelectronics & Computer
ISSN 1000-7180,CN 61-1123/TN

《微电子学与计算机》网络首发论文

题目：　改进 RT-DETR 的无人机小目标检测算法
作者：　苏佳，杨梦凡，张柏杨，常永浩，侯艳丽
网络首发日期：　2024-11-04
引用格式：　苏佳，杨梦凡，张柏杨，常永浩，侯艳丽．改进 RT-DETR 的无人机小目标检测算法[J/OL]．微电子学与计算机.
https://link.cnki.net/urlid/61.1123.tn.20241104.1032.004

网络首发时间：2024-11-04 16:23:37
网络首发地址：https://link.cnki.net/urlid/61.1123.tn.20241104.1032.004

改进 RT-DETR 的无人机小目标检测算法

苏佳，杨梦凡，张柏杨，常永浩，侯艳丽

（河北科技大学，信息科学与工程学院，河北石家庄 050018）

摘要:复杂无人机环境下的小目标检测存在目标分布密集和特征提取困难的问题，检测准确度仍有提升空间，为提高小目标检测效果，提出基于 RT-DETR 的无人机小目标检测改进算法 DRT-DETR。为提升模型计算效率和特征提取能力，引入快速多尺度注意力特征提取模块 Faster-EMA，显著降低模型参数量，增强特征提取效率。为提高多尺度特征的利用率，采用加权双向跨尺度特征融合模块 Bi-CCFM，优化特征传递与信息融合。为提升定位和识别的精确性，提出基于归一化高斯距离的回归损失函数 NWD，用 Wasserstein 距离来度量边界框之间的相似性，提升小目标检测的准确度。实验结果表明，DRT-DETR 在 VisDrone 数据集上的 mAP@0.5 达到了 48.4%，较改进前增长了 3.1%，参数量降低了 12.6%，实现了轻量化与精度提升的双重目标。

关键词：遥感；深度学习；小目标检测；RT-DETR

文献标志码:A　　中图分类号:TP391.41　文章编号：1000-7180 (2024) xx-xxxx-x

Enhancing RT-DETR for small object detection in UAV

SU Jia, YANG Mengfan, ZHANG Boyang, CHANG Yonghao, HOU Yanli

(College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract: In complex drone environments, small target detection faces challenges such as dense target distribution and difficult feature extraction, which leaves room for improvement in detection accuracy. To enhance the effectiveness of small target detection, we propose an improved algorithm for drone small target detection based on RT-DETR, named DRT-DETR. To improve the model's computational efficiency and feature extraction capability, we introduce a Fast Multi-Scale Attention Feature Extraction Module (Faster-EMA), which significantly reduces the model's parameter count while enhancing feature extraction efficiency. To improve the utilization of multi-scale features, we adopt a Weighted Bidirectional Cross-Scale Feature Fusion Module (Bi-CCFM) to optimize feature transmission and information fusion. To enhance the precision of localization and recognition, we propose a regression loss function based on Normalized Wasserstein Distance (NWD) that uses Wasserstein distance to measure the similarity between bounding boxes, consequently improving the accuracy of small target detection. Experimental results show that DRT-DETR achieves an mAP@0.5 of 48.4% on the VisDrone dataset, which represents a 3.1% increase compared to the previous version, while the parameter count has been reduced by 12.6%, achieving the dual goals of model lightweighting and accuracy improvement.

Keywords: remote sensing; deep learning; small target detection; RT-DETR

1 引言

随着计算机视觉与人工智能技术的迅猛发展，无人机已经成为现代诸多领域不可或缺的利器[1]，诸如军事侦察、精准农业、环保监测、交通管理以及紧急救援等，其轻便灵活的特性，使得无人机得以自如地从高空俯瞰，实时采集高分辨率图像，极大地丰富了数据分析手段。由于无规律变化的俯拍视角，导致了场景中包含大量小尺度的密集目标，且容易被错综复杂的地形所遮挡，加大了目标检测的难度。

目前的目标检测算法中，双阶段算法典型代表为 R-CNN[2]、Fast R-CNN[4]、Faster R-CNN[5]算法，通过先筛选出候选目标区域，再进行精细化的目标分类和定位来提升目标检测的精度。而单阶段检测算法，如 SSD[6]和 YOLO[7-13]系列，减少了计算步骤，实现了对目标的高效实时检测。尽管这两种算法各有特点，但都需要进行阈值筛选和非极大值抑制（NMS）[14]处理，在一定程度上降低了模型的鲁棒性和检测速度。随着 Transformer[15]架构在自然语言处理领域的显著成功，学者们开始尝试将其革新性理念应用于目标检测任务， Facebook 团队于 2020 年提出了基于 Transformer 的端到端目标检测算法 DETR，巧妙地消除了上述处理步骤，从而简化了处理流程。然而尽管 DETR 带来了高效且直接的检测体验，但其较大的参数量也成为了一个挑战，需要在保持模型性能的同时，寻求参数优化的策略。因此出现了大量 DETR 变体，旨在克服 DETR 参数量大和计算成本高的问题，Zhu 等人于 2021 年提出了 Deformable DETR[16]，改进注意力模块只关注参考点周围的关键采样点来提高性能；Roh 等人于 2022 年提出了 Sparse DETR[17]，通过 token 稀疏化方法减轻了编码器中的注意力复杂性，可以在相同的计算量下提高检测性能；Li 等人于 2022 年提出了 DN-DETR[18]，通过 DeNoising 去噪训练解决 DETR 二分图匹配不稳定的问题，加快模型收敛速度；Zhang 等人于 2022 年提出了 DINO[19]，通过改进编码器和解码器显著减少了模型大小；百度针对 DETR 高计算成本的问题，提出了基于 DETR 的实时检测方法 RT-DETR[20]，去除了阈值筛选和非极大值抑制，使得 RT-DETR 在更少的迭代次数下就能达到更高的训练精度。

但在复杂背景中的小目标检测上，特征提取困难导致漏检误检的问题会降低模型的检测准确率[21]。为了克服这些难题，学者们正在不断优化算法，提升复杂场景下小目标的识别和定位能力，实现更高效、准确的无人机应用。M. Muzammul 于 2024 年提出使用图像切片 SAHI 技术结合 RT-DETR[22]，在 VisDrone 数据集上实现了较好的检测效果。但是模型参数量较大，在实际应用方面效果不尽如人意，因此本研究以 RT-DETR 为基础提出 DRT-DETR，旨在通过改进模型的特征提取和特征融合模块减少模型参数量的同时通过改进损失函数提高小目标检测精度，提升模型的检测性能，主要工作如下：

（1）提出了快速多尺度注意力特征提取模块 Faster-EMA 来提升模型计算效率和特征提取能力。采用 FasterNet 对原本 BasicBlock 进行改进，通过部分卷积 PConv 实现模型的轻量化，减少了模型的参数数量，从而降低了计算复杂度，显著提升模型的推理速度，并通过融合多尺度注意力 EMA 模块，提高特征关注度，增强特征提取的效率，保留了小目标的特征细节，极大地提升了模型对小目标检测和识别的准确性。

（2）在颈部网络采用加权双向跨尺度特征融合模块 Bi-CCFM 来提高多尺度特征的利用率，通过在输入和输出节点间增设新的融合路径，动态调整特征权重聚焦关键特征，优化信息传递与特征融合，聚焦关键特征，减少冗余信息干扰。

（3）提出基于归一化高斯距离的回归损失函数 NWD 来为提升小目标定位和识别的精确性，NWD 通过衡量目标检测任务中边界框之间的相似性，增强对目标尺度变化的适应性，克服了传统 IoU 损失函数的局限性，在测量不重叠或相互包含的边界框之间时可以衡量相似度，显著提高小目标检测的性能。

2 RT-DETR 模型原理

RT-DETR 是一种创新的实时端到端目标检测框架，它依托 Transformer 架构，巧妙地应对了多尺度特征处理的挑战，其详细的网络设计如图 1 所示。

如图 1 所示，RT-DETR 模型主要由 ResNet-18 主干网络、AIFI 和 CCFM 组成的混合编码器，以及带有辅助预测头的解码器共同构成。ResNet-18 通过卷积层和 BasicBlock 模块实现初步的特征提取，将 S3、S4、S5 这三层的输出送入混合编码器来提供丰富的多层次信息。混合编码器内 AIFI 模块专注于处理高级图像特征，通过自注意力机制在 S5 特征图上进行内部尺度交互，提高模型在对象检测和识别方面的性能，CCFM 模块则利用自底向上和自顶向下的双路径融合策略，通过上采样和下采样有效整合 S3、S4、F5 三个特征图的多尺度特征。解码器部分包括 IoU-aware query 和 Decoder head 模块，其中 IoU 感知查询模块根据分类分数选择的排名靠前的 K 个预测框，将排名靠前的预测框输出，IoU 感知查询选择可以为对象查询提供更多具有准确分类(高分类分数)和精确位置(高 IoU 分数)的编码器特征，从而提高检测器的准确度。随后通过迭代优化过程，解码器逐步生成精确的边界框预测和对应的置信度评分，以完成高效且准确的检测任务。

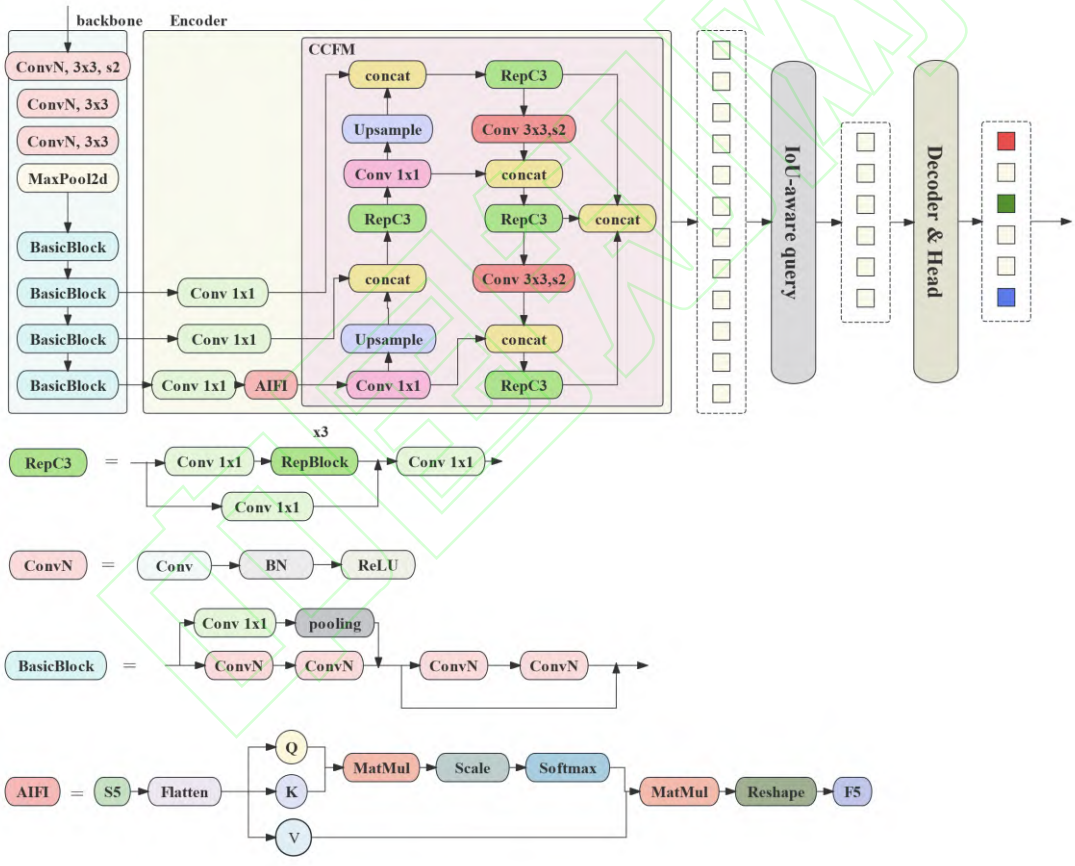


图 1 RT-DETR-R18 网络架构

Fig.1 RT-DETR-R18 Network Architecture

3 DRT-DETR 网络结构

本实验基于 RT-DETR 提出了改进网络结构 DRT-DETR，图 2 为 DRT-DETR 网络结构图，该网络旨在通过优化特征提取和融合过程，提高小目标检测的精度与效率。如图 2 所示，为了确保在提升小目标检测精度的同时，提高检测效率，在主干网络采用 Faster-EMA 模块替换原本的 BasicBlock 结构，在缩减模型参数的基础上提升检测精度，在颈部网络采用加权双向跨尺度特征融合结构 Bi-CCFM，优化多尺度特征融合，提高特征利用率，最后在解码器部分改进基于归一化

Wasserstein 距离（Normalized Wasserstein Distance, NWD）的回归损失函数，用于提高小目标检测的性能。

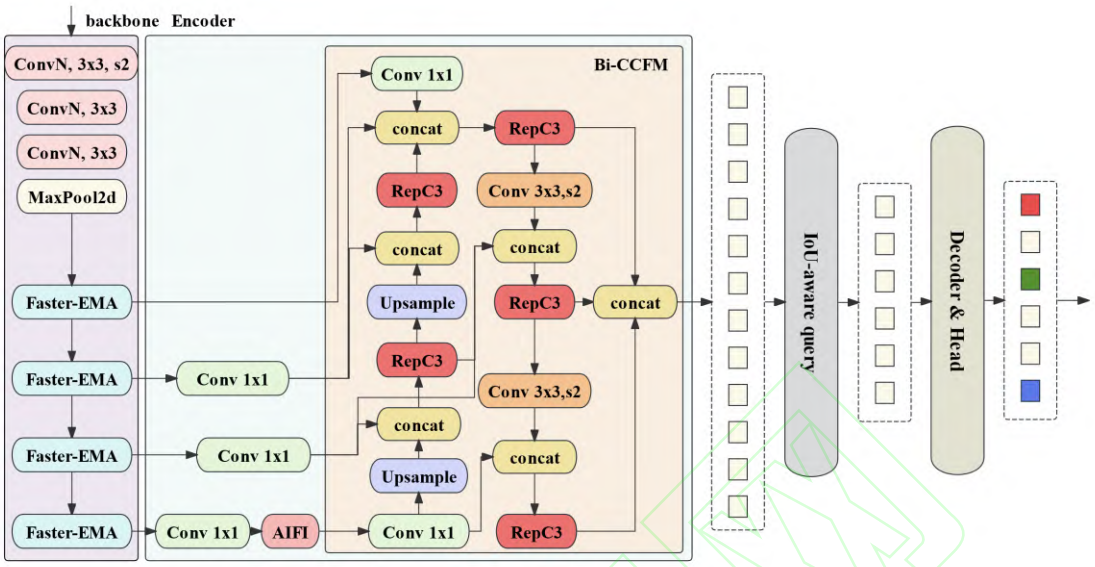


图 2 DRT-DETR 网络结构图

Fig.2 DRT-DETR Network Architecture Diagram

3.1 快速多尺度注意力特征提取模块（Faster-EMA）

RT-DETR 模型参数量较大，特征提取模块存在大量冗余计算，为了在减少模型参数量的基础上提高检测精度，提出了 Faster-EMA 模块，它结合了高效的多尺度注意力机制（Efficient Multi-Scale Attention, EMA）[23]和轻量级主干网络 FasterNet[24]中的 FasterBlock。FasterBlock 引入了 PConv 旨在降低计算冗余和内存访问，从而实现更高效且精简的特征提取过程。与传统的卷积运算不同，PConv 仅对输入特征的一部分通道执行卷积操作以捕获空间特征，而其他通道则保持不变。通常情况下 PConv 会选择第一个或最后一个连续的 c_p 个通道作为计算整个特征映射的代表，以此达到计算资源的优化利用。在保持模型性能的同时，减少了不必要的计算和内存访问，提升了模型的运行效率。图 3 为 Faster-EMA 模块以及 PConv 和普通卷积的对比。

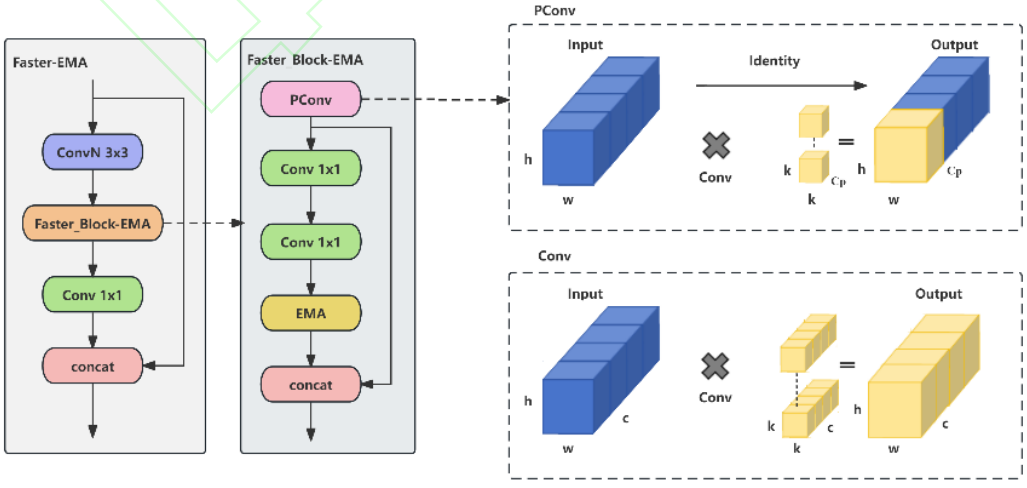


图 3 Faster-EMA 模块以及 PConv 和普通卷积的对比图

Fig.3 Comparison of Faster-EMA Module and PConv and Regular Convolution

因此 PConv 的 FLOPs 和内存访问量为：

$$h\times w\times k^2\times c_p^2 \quad (1)$$

$$h\times w\times 2c_p+k^2\times c_p^2\approx h\times w\times 2c_p \quad (2)$$

普通卷积的 FLOPs 和内存访问量为：

$$h\times w\times k^2\times c^2 \quad (3)$$

$$h\times w\times 2c+k^2\times c_p^2\approx h\times w\times 2c \quad (4)$$

上式中(c,h,w)为输入通道尺寸，k为卷积核大小，c为普通卷积的通道数，c_p为PConv卷积的通道数，且$\frac{c_p}{c}=\frac{1}{4}$。

因此可得 PConv 的 FLOPs 为普通卷积的$\frac{1}{16}$，内存访问量为普通卷积的$\frac{1}{4}$。

为了解决轻量化的特征提取模块造成精度降低的问题，引入了高效多尺度注意力模块 EMA。EMA 凭借其基于跨空间学习的高效多尺度注意力机制，特别有利于在小目标检测中提升检测性能实现高精度，与 CBAM（Convolutional Block Attention Module）[25]、SAM（Spartial Attention Module）[26]和 CAM（Channel Attention Module）[27]等常见的注意力机制相比，EMA 在保持相对较低的参数量的基础上，展现出了卓越的性能优势。具体来说 EMA 将部分通道重构为批处理维度，通过跨空间学习整合通道间的相关信息，生成更精确的注意力权重，不仅保证了模型的计算效率，还使得模型在关注关键特征的同时，避免了过度复杂的参数调整，从而在提升检测精度的同时，实现了轻量化的目标。

如图 4 所示为 EMA 的计算过程。从左至右三条分支分别代表两个 1×1 分支和一个 3×3 分支，两条 1×1 分支经过全局平均池化(Avg Pool)操作后进行拼接，全局平均池化操作的数学表达式为：

$$AvgPool=\frac{1}{H\times W}\sum_{j}^{H}\sum_{i}^{W}x_c(i,j) \quad (5)$$

其中H为输入特征图高度，W为输入特征图宽度，$x_c(i,j)$代表特征图在第c个通道上的像素值。接着通过 1×1 卷积，其输出被分解为两个向量，这两个向量分别用 Sigmoid 函数进行非线性拟合，然后采用重加权运算（Re-weight）将通道进行融合来促进两个 1×1 分支之间的跨通道信息交流。3×3 分支利用 3×3 卷积来捕捉多尺度特征信息，以扩大特征空间。然后将 1×1 分支重加权后的结果进行组归一化（GroupNorm），然后将输出和 3×3 分支的输出分别通过全局平均池化和 Softmax 函数，将经过归一化后的输出和 1×1 分支和 3×3 分支未经过上述处理前的输出分别进行点积运算（Matmul），这样可以得到保留准确空间信息的空间注意力图，就是通过跨空间学习(Cross-spatial learning)模块，将 1×1 分支和 3×3 分支的输出相结合，进一步增强全局空间信息编码，最后将整合后的结果通过 Sigmoid 函数拟合和 Re-weight 函数重加权来捕获全局上下文信息，EMA 的最终输出大小与输入相同，可以有效地堆叠到其他架构中。

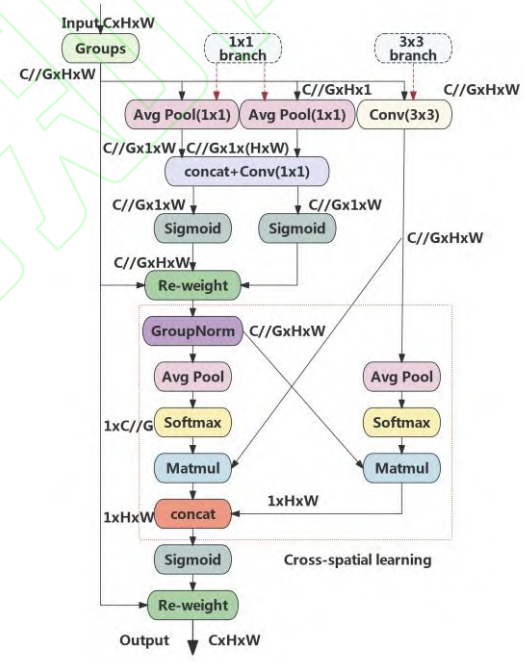


图 4 EMA 的计算过程

Fig.4 EMA Calculation Process

3.2 加权双向跨尺度特征融合模块（Bi-CCFM）

颈部网络用于融合不同层次的特征，结合了低层的细节信息和高层的语义信息，使得模型能够同时理解图像的局部特征和全局结构，通过信息传递，Neck 网络能够确保高层的语义信息能够传递到低层，同时低层的细节信息也能影响高层，增强了特征的丰富

度和表达力。传统的 FPN[28]首先从深层网络中提取出语义丰富的特征图，然后通过上采样（UpSample）再与浅层网络的特征图进行融合。PAN[29]在保持了 FPN 自顶向下结构的同时，引入了自底向上的路径增强，即在特征金字塔中加入了从低层到高层的直接连接。低层的信息可以传递到高层，增强了高层特征的细节感知能力。RT-DETR 中使用的 CCFM 结构实际是一个类似 PAN 的结构，所以在多尺度特征融合的策略上，基于 RT-DETR 的基线模型，创新性地采用了加权双向跨尺度特征融合结构 Bi-CCFM。

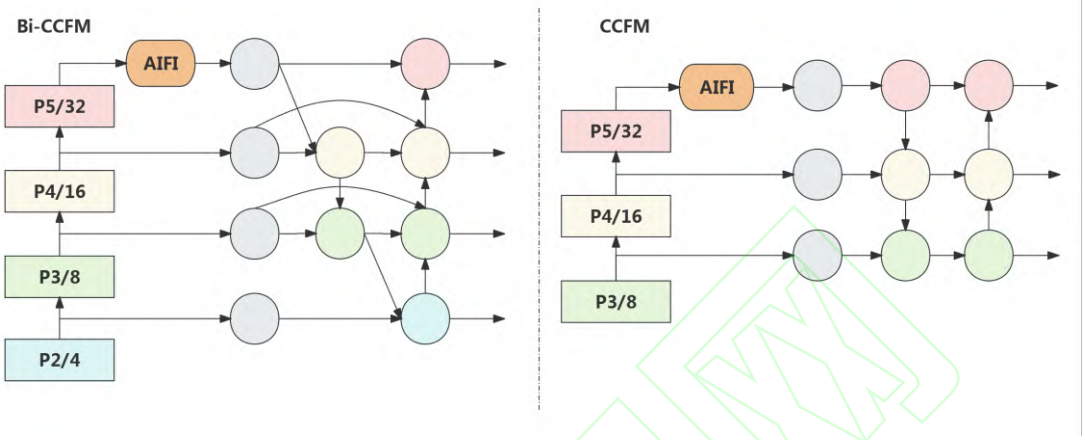


图 5 Bi-CCFM 和 CCFM 的区别

Fig.5 Difference Between Bi-CCFM and CCFM

图 5 为改进后的 Bi-CCFM 和 CCFM 的对比。相较于右侧的 CCFM 结构，Bi-CCFM 通过在输入和输出节点间增设新的融合路径，如图 5 所示，当原始输入节点和输出节点处于同一层时，在原始输入节点和输出节点之间添加一条额外的融合路径，因此模型能够巧妙地融合不同尺度的特征图，从而更深入地理解图像中的多尺度信息，对小目标检测的敏感度大为提升。在融合不同分辨率特征的过程中，Bi-CCFM 摒弃了简单相加的方式，引入了智能权重分配机制。每个输入特征都被赋予了独特的权重，使得网络能够自主学习并适应性地调整每个特征的重要性。具体计算过程为：

$$O=\sum_i \frac{w_i}{\varepsilon+\sum_j w_j}\cdot I_i \tag{6}$$

$$B_s^{td}=C\left(\frac{w_1 B_s^{in}+w_2 Resize\left(B_{(s+1)}^{in}\right)}{w_1+w_2+\varepsilon}\right) \tag{7}$$

$$B_s^{out}=C\left(\frac{w_1^{'}\cdot B_s^{in}+w_2^{'}\cdot B_s^{td}+w_3^{'}\cdot Resize\left(B_{(s-1)}^{out}\right)}{w_1^{'}+w_2^{'}+w_3^{'}+\varepsilon}\right) \tag{8}$$

在这一过程中，w_i 是学习到的权重，通过 ReLU 激活函数用于量化特征融合过程中不同特征的重要性，设定学习率常数 ε =0.0001，来确保权重更新的稳定性。B_s^{td} 即从上到下的表示每一层的中间特征，B_s^{out} 反映了自下而上的输出特征。*Resize* 操作通常涉及下采样或上采样，C 是深度可分离卷积，每个卷积后都添加了 BN 层和激活函数。

Bi-CCFM 结合了快速归一化融合和双向跨尺度连接，这使得网络能够动态调整特征的权重，更加聚焦于关键特征，从而更有效地整合了不同尺度的信息。这种策略显著提升了对各种目标的检测性能，使得模型在处理复杂场景时表现出色，Bi-CCFM 在小目标检测任务中展现出了卓越的优势。

3.3 归一化高斯距离回归损失函数（NWD）

由于 VisDrone 数据集中小目标较多，传统的 IoU（Intersection over Union）指标在评估小目标时往往不尽如人意，所以采用 Wasserstein 距离来更精确地衡量边界框的相似性，因此引入了更适合小目标检测的 NWD 回归损失函数[30]。 NWD 方法是将边界框视为二维高斯分布，通过计算两个分布之间的归一化 Wasserstein 距离来评估它们的相似度，这种度量方式能更好地捕捉小目标的细微差异，提升了检测精度和鲁棒性。

首先是二维高斯分布建模，就是将边界框建模为二维高斯分布，边界框 $R=(c_x,c_y,w,h)$，边界框的中心坐标为 (c_x,c_y)，边界框宽度为 w，高度为 h，常见的二维高斯分布的公式为：

$$f(x\mid\mu,\Sigma)=\frac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{2\pi|\Sigma|^{\frac{1}{2}}} \tag{9}$$

其中 $\mu=\begin{bmatrix}c_x\\c_y\end{bmatrix}$，$\Sigma=\begin{bmatrix}\frac{\omega^2}{4}&0\\0&\frac{h^2}{4}\end{bmatrix}$，$x$ 是高斯分布的坐标 (x,y)，T 是矩阵转置，μ 和 Σ 分别是高斯分布的均值向量和协方差矩阵。

然后是计算归一化高斯距离，$A=(cx_a,cy_a,w_a,h_a)$ 建模后高斯分布为 N_a，$B=(cx_b,cy_b,w_b,h_b)$ 建模后高斯分布为 N_b，对于这两个边界框计算高斯距离公式如下：

$$W_2^2(N_a+N_b)=\left\|\left[cx_a,cy_a,\frac{w_a}{2},\frac{h_a}{2}\right]^T,\left[cx_b,cy_b,\frac{w_b}{2},\frac{h_b}{2}\right]\right\|_2^2 \tag{10}$$

将距离公式归一化为相似度度量公式为：

$$NWD(N_a,N_b)=\exp\left(-\frac{\sqrt{W_2^2(N_a,N_b)}}{C}\right) \tag{11}$$

但是只使用 NWD 损失函数会导致模型收敛速度变慢，所以综合考虑将 IoU 损失函数和 NWD 损失函数进行加权组合，形成一个新的回归损失函数，公式如下：

$$Loss=IoU_{loss}\times0.5+NWD\times0.5 \tag{12}$$

使用新的 NWD 回归损失函数来进行小目标检测，增强了模型对小目标的识别能力，而且使得模型在复杂场景中展现出更强的适应性和准确性。

4 实验与数据分析

4.1 数据集介绍

本实验采用的数据集是来自天津大学机器学习与数据挖掘实验室通过无人机收集的 VisDrone 数据集。共含有 8629 张图片，采用其中的 6471 张图片作为训练集，548 张图片作为验证集，1610 张图片作为测试集，该数据集共包含十个类别分别为：行人、人、自行车、汽车、面包车、卡车、三轮车、遮阳篷-三轮车、公共汽车和摩托车。

4.2 实验环境

实验环境使用 Windows 10 操作系统、Python 3.8.19 和 PyTorch 深度学习框架，处理器型号为 Intel(R) Core(TM) i5-13400F，2.50 GHz，GPU 型号为 NVIDIA GeForce RTX 4060Ti，8GB 显存，CUDA 版本为 12.1。实验参数设置如下：训练轮次为 200 轮，批次大小为 4，输入图像尺寸为 640×640，采用 AdamW 优化器，初始学习率为 0.0001，最终学习率 0.1，其他均为默认设置。

4.3 评价指标

为了更好的评估模型检测性能，采用 mAP@0.5、mAP@0.5：0.95、参数量（Params）、准确率（Precision）和召回率（Recall）等关键指标，mAP@0.5 专注于 IoU 阈值为 0.5 时的各类别平均精度，mAP@0.5：0.95 则是计算在 IoU 阈值从 0.5 到 0.95 范围内模型的平均精度，用来评价小目标的检测性能。参数量用来衡量模型复杂度，计算量衡量模型运行时的计算复杂度。

TP 表示被正确识别为正例的实例数量，即那些实际上属于正类并且被分类器准确标记为正类的样本。FP 是模型可能出现的误判，指的是被错误地标记为正例的样本，FN 是模型的漏检，即那些实际上应被识别为正例的样本，却被分类器错误地标记为负例。准确率（P）表示模型在所有被判断为正例的样本中，真正是正例的比例，召回率（R）表示模型所有被预测正确的正例占总体正例样本的比例。AP 是指每个类别在 P-R 曲线上与横、纵坐标所围成的面积，是每个类别的平均准确率，具体计算公式如下：

$$P=\frac{TP}{TP+FP} \tag{13}$$

$$R=\frac{TP}{TP+FN} \tag{14}$$

$$AP=\int_0^1 P(R)dR \tag{15}$$

$$mAP=\frac{1}{n}\sum_{j=1}^{n}AP_j \tag{16}$$

上式中 n 为类别数，j 为第 j 个类别，mAP 为所有类别的平均精度值。

4.4 消融实验

在相同条件下为测试每个改进模块的有效性，通过消融实验来验证 Faster-EMA、Bi-CCFM、NWD 对算法性能的影响。消融实验具体设计和相应结果如表 1 所示。√表示改进添加的模块，-表示未添加。

表 1 消融实验

Table 1 ablation experiment

Faster-EMA	Bi-CCFM	NWD	Params/M	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%
-	-	-	19.8	59.8	43.7	45.3	27.5
√	-	-	16.9	60.2	45.1	46.5	28.3
-	√	-	20.3	62.2	44.4	46.2	28.4
-	-	√	19.8	60.7	43.9	46.3	28.2
√	√	-	17.3	61.6	45.1	47.1	28.8
√		√	16.9	61.4	45.4	46.8	28.3
	√	√	20.3	60.8	44.8	46.9	28.5
√	√	√	17.3	62.0	46.8	48.4	29.5

根据表格中的数据可以看出，集成 Faster-EMA 模块后，显著降低了模型的计算负担，通过引入高效的多尺度注意力机制，提升了特征提取能力，实现了轻量化与精度提升的双重目标。由实验结果可知，应用 Faster-EMA 模块之后，模型的参数量减少了 2.9M，同时检测精度提高 1.2%。

集成 Bi-CCFM 后，模型能更好融合不同尺度的特征图，理解图像中的多尺度信息。在保持计算复杂度相对较低的同时，显著提升了模型的检测精度。由实验结果可以看出尽管模型的参数量略有增加，但 mAP@0.5 提升 0.9%。

改进新的损失函数之后，能够更好地衡量目标检测任务中边界框之间的相似性，提高模型的检测精度。由实验数据可知，改进 NWD 之后，模型的 mAP@0.5 增加 1.0%。

同时改进 Faster-EMA 和 Bi-CCFM，模型可以更高效地提取和融合多尺度特征，强化信息交互和理解上下文信息，在处理复杂场景和小目标时，提高模型的泛化能力，在保证模型效率的同时提升了检测结果的可靠性。

同时集成 Faster-EMA 和 NWD 损失函数，强化了特征提取的效率，在处理小目标时能更加精确地衡量边界框之间的相似性，提高模型的定位准确性，在不增加计算负担的情况下，进一步提升了整体的检测性能。

改进 Bi-CCFM 结构和 NWD 损失函数后，增强了模型对不同尺度目标的识别能力，提升了模型在评估小目标边界框相似性时的准确性和鲁棒性，在保持模型效率的同时，增强了其在复杂场景下的检测性能。

通过消融实验对每个模块进行详细评估，同时集成三个改进点后 mAP@0.5、mAP@0.5:0.95 分别达到了 48.4%和 29.5%，较改进前的 RT-DETR 模型分别增长了 3.1%和 2.0%，参数量为 17.3M，较原版降低了 12.6%。由结果可知集成改进模块后显著提升了模型的检测精度，还减少了模型的参数量，从而证明了 DRT-DETR 算法在检测性能方面优于原版 RT-DETR 算法。

4.5 对比实验

为了更好的比较 DRT-DETR 与其他算法的性能，将与下列算法进行对比，主要包括 YOLOv5x、YOLOv8n、YOLOv8s、YOLOv10s、YOLOv10n、Deformable DETR、DINO、RT-DETR，实验结果如表 2 所示。从表 1 中可以看出，DRT-DETR 在性能方面表现最为出色，其 mAP@0.5 和 mAP@0.5:0.95 分别达到了 48.4%和 29.5%，均高于对比的目标检测算法，对于原版 RT-DETR 算法，由于采用了更高效的特征提取和特征融合模块，DRT-DETR 在 VisDrone2019 数据集上的 mAP@0.5 和 mAP@0.5:0.95 分别提高了 3.1%和 2.0%，参数量为 17.3M，较原版降低了 12.6%，由于采用了专用于小目标检测的损失函数 NWD，因此小目标检测精度 APs 达到了 18.9%，比基础模型提高了 2.0%。虽然 FPS 低于基础模型和部分 YOLO 算法，但是改进显著提升了模型的精度以及模型的参数量，依然满足航拍实时检测。相较于 Deformable DETR 和 DINO 算法，DRT-DETR

的参数量更低，准确率更高。相较于 YOLO 系列算法，DRT-DETR 的参数量和计算量有所上升，但检测精度显著提高，mAP@0.5 提高了 14.9%、11.1%、18.5%和 12.0%，可以证明 DRT-DETR 在小目标检测上达到了令人满意的成果。

表 2 各算法在 VisDrone2019 验证集上的检测结果对比

Table 2 Comparison of detection results of various algorithms on the VisDrone2019 val set

模型	Params/M	FLOPs/G	FPS	APs	mAP@0.5/%	mAP@0.5:0.95/%
YOLOv5x	86.2	203.8	34	15.8	42.5	25.2
YOLOv8n	3.0	**8.1**	119	9.3	33.5	17.8
YOLOv8s	11.2	28.7	85	12.4	37.3	21.6
YOLOv10n	**2.69**	8.2	111	8.6	29.9	17.1
YOLOv10s	8.04	24.5	143	12.2	36.4	21.4
Deformable DETR	40	196	29	15.6	42.2	27.1
DINO	47	279	24	17.1	46.2	29.4
RT-DETR[18]	19.8	57.0	101	16.9	45.3	27.5
DRT-DETR	17.3	58.8	73	**18.9**	**48.4**	**29.5**

4.6 可视化分析

为了深入剖析模型的检测效能，通过两个方面对模型进行详尽的评估，一是比较改进前后模型的 mAP 对比图，二是通过可视化检测结果来直观展示模型的表现。

首先如图 6 所示，mAP 曲线能够清晰地揭示优化前后模型在检测任务中的整体性能是否提升。从图 6 中可以明显观察到，DRT-DETR 模型在 mAP@0.5 指标上取得了显著的提升，较改进前提高了 3.1%，DRT-DETR 在目标检测各种阈值下的平均性能更好。检测性能的提升主要由于融合了 Faster-EMA、Bi-CCFM 以及 NWD 改进点，这些创新有效地增强了模型对小目标检测的敏锐度和准确性，从而在复杂场景下展现出更强的识别能力。

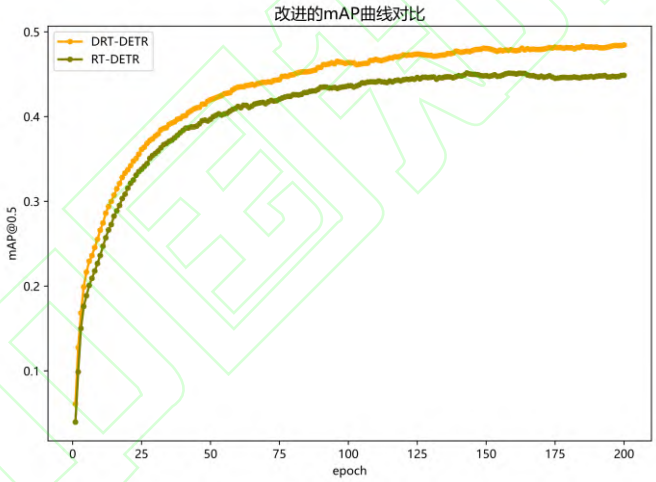


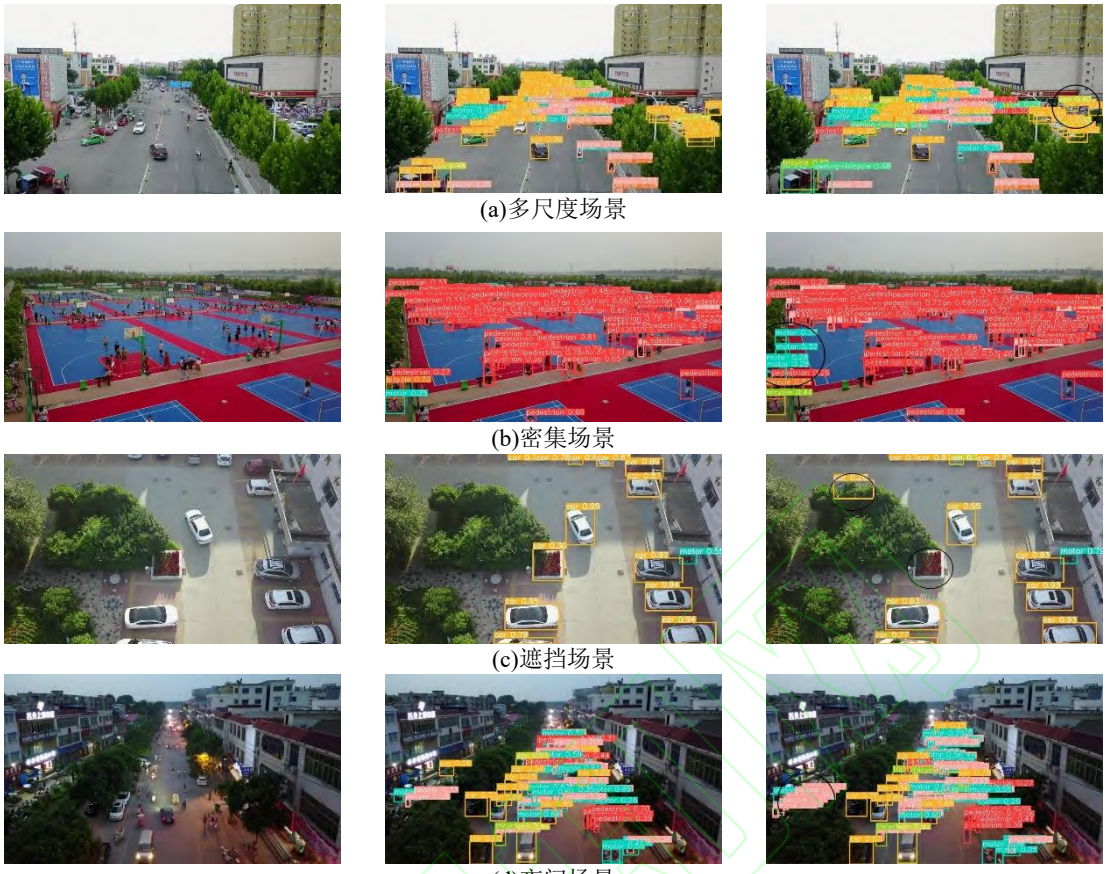
图 6 mAP 曲线对比

Fig.6 Comparison of mAP curves

通过可视化可以直观展示模型的检测效果，为了直观对比 DRT-DETR 模型与基准 RT-DETR 模型在小目标检测上的差异，在 VisDrone 数据集中挑选了四个具有挑战性的场景进行检测。

如图 7 所示，对于图 7（a）所示的多尺度目标场景，DRT-DETR 模型明显优于基准模型，特别是在检测图中较远的人群密集的区域时，展现出更高的精准度。在图 7（b）的密集目标场景中，DRT-DETR 模型能够有效地识别和定位众多目标，对密集对象识别更为准确，为处理复杂场景提供有力保障。在图 7（c）的遮挡场景中，面对被树木部分或完全遮挡的目标，DRT-DETR 也能准确地识别目标提供更加完整且准确的检测结果。在图 7（d）的夜间环境下，DRT-DETR 模型依然保持稳定性能，能够有效识别夜间密集人群。

原始图片　　RT-DETR 模型　　DRT-DETR 模型



(a)多尺度场景

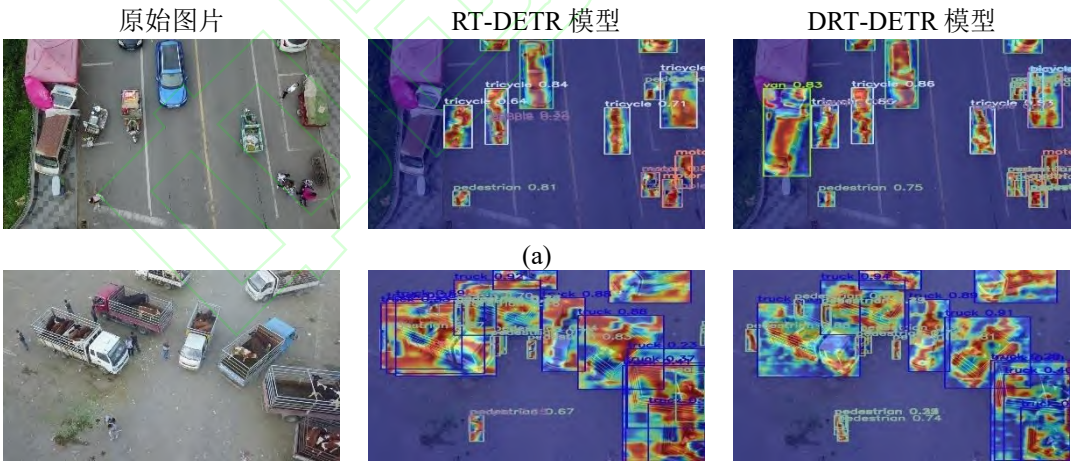
(b)密集场景

(c)遮挡场景

(d)夜间场景

图 7 原图、基准模型和 DRT-DETR 模型对比

Fig.7 Comparison of the original image, baseline model and DRT-DETR model



(a)

(b)

图 8 原图、基准模型和 DRT-DETR 模型热力图对比

Fig.8 Comparison of Heatmaps of Original Image, Baseline Model, and RT-DETR Model

为了能够更加直观的看到模型的性能，利用 GradCAM++技术进行热力图可视化操作。热力图的颜色代表了特征点的密集程度，其中红色为最密集，由图 8 可知可以清晰地看到相比基础模型，DRT-DETR 模型的特征点更密集的围绕在检测目标上，而非背景，因此 DRT-DETR 模型在复杂场景下对小目标的响应更强，能更好的识别漏检误检问题，也能更好地抑制背景干扰。

由上图 7 和图 8 结果及分析可得 DRT-

DETR 模型较基准模型准确性和鲁棒性都有所提升，在实际应用中优势更为明显。

5 结束语

提出一种无人机小目标检测算法 DRT-DETR，在减少模型参数量的基础上提高模型的准确性和鲁棒性。引入快速多尺度注意力特征提取模块，提高特征关注度，提升模型计算效率和特征提取能力，然后采用了加权双向跨尺度特征融合模块，融合不同尺度的特征图，提高多尺度特征的利用率，最后提出基于归一化高斯距离的回归损失函数，增强对目标尺度变化的适应性，在复杂环境下提高小目标检测精度。实验结果表明，较改进前的 RT-DETR 模型，DRT-DETR 在 VisDrone 数据集上的 mAP@0.5、mAP@0.5:0.95 分别增长了 3.1%和 2.0%，参数量较原版降低了 12.6%，验证了改进模块的有效性，通过实验结果分析可以得到，DRT-DETR 模型在小目标检测上的性能更为优异。在 VisDrone 数据集中多数场景的光照条件理想且目标分布相对集中，所以当遇到光照条件较差或目标分布不规律的情况时，模型可能造成漏检误检。因此后续拟通过数据增强技术，增强模型对不同环境的适应性，提高模型的泛化能力，以确保在各种复杂光照和目标分布条件下都能保持稳定且准确的检测性能。

参考文献：

[1] Li T ,Zikang L ,Xiaokai H , et al.A transformer-based UAV instance segmentation model TF-YOLOv7[J].Signal, Image and Video Processing,2024,18(4):3299-3308.

[2] Zhou L, Liu Z, Zhao H, et al. A multi-scale object detector based on coordinate and global information aggregation for UAV aerial images[J]. Remote Sensing, 2023, 15(14): 3468.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich featurehierarchies for accurate object detection and semanticsegmentation[C]//Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, Columbus, OH,USA, 2014: 580-587.

[4] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEEInternational Conference on Computer Vision (ICCV), Santiago,Chile, 2015:1440-1448.

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

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shotMultiBox detector[C]//Proceedings of the 14th EuropeanConference on Computer Vision (ECCV 2016), Amsterdam,The Netherlands, October 11-14, 2016: 21-37.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You onlylook once: unified, real-time object detection[C]//Proceedingsof the IEEE Conference on Computer Vision and PatternRecognition (CVPR), Las Vegas, NV, USA, 2016: 779-788.

[8] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition (CVPR), Honolulu, HI,USA, 2017: 6517-6525.

[9] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[10] WU T H, WANG T W, LIU Y Q. Real-time vehicle and distance detection based on improved yolo v5 network[C]//2021 3rd World Symposium on Artificial Intelligence (WSAI). IEEE, 2021: 24-28.

[11] Wu W, Guo L, Gao H, et al. YOLO-SLAM: A semantic SLAM system towards dynamic environment with geometric constraint[J]. Neural Computing and Applications, 2022: 1-16.

[12] Wang Y, Yan G, Meng Q, et al. DSE-YOLO: Detail semantics enhancement YOLO for multi-stage strawberry detection[J]. Computers and electronics in agriculture, 2022, 198: 107057.

[13] Gai R, Chen N, Yuan H. A detection algorithm for cherry fruits based on the improved YOLO-v4 model[J]. Neural Computing and Applications, 2023, 35(19): 13895-13906.

[14] Haoran X ,Shibin T ,Jia W , et al.Rock fracture identification algorithm based on the confidence score and non-maximum suppression[J].Bulletin of Engineering Geology and the Environment,2024,83(6).

[15] Bujagouni G K ,Pradhan S .Transformer based deep learning hybrid architecture for phase unwrapping[J].Physica Scripta,2024,99(7).

[16] Wang D, Li Z, Du X, et al. Farmland obstacle detection from the perspective of uavs based on non-local deformable detr[J]. Agriculture, 2022, 12(12): 1983.

[17] Shehzadi T, Hashmi K A, Stricker D, et al. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5840-5850.

[18] Li F, Zhang H, Liu S, et al. Dn-detr: Accelerate detr training by introducing query denoising[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13619-13627.

[19] Li F, Zhang H, Xu H, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3041-3050.

[20] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.

[21] Hao Z ,Chuanyan H ,Wanru S , et al.Adaptive Slicing-Aided Hyper Inference for Small Object Detection in High-Resolution Remote Sensing Images[J].Remote Sensing,2023,15(5):1249-1249.

[22] Muzammul M, Algarni A M, Ghadi Y Y, et al. Enhancing UAV aerial image analysis: Integrating advanced SAHI techniques with real-time detection models on the VisDrone dataset[J]. IEEE Access, 2024.

[23] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[24] Chen J, Kao S, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12021-12031.

[25] WOO S, PARK J, LEE J Y, et al. CBAM: convolutionalblock attention module[C]//Proceedings of the 15thEuropean Conference on Computer Vision. Munich:Springer, 2018: 3−19.

[26] Mohammad E ,Elham G .Self-attention (SA) temporal convolutional network (SATCN)-long short-term memory neural network (SATCN-LSTM): an advanced python code for predicting groundwater level.[J].Environmental science and pollution research international,2023,30(40):92903-92921.

[27] GuangboL ,Guolong S ,Jun J .YOLOv5-KCB: A New Method for Individual Pig Detection Using Optimized K-Means, CA Attention Mechani-sm and a Bi-Directional Feature Pyramid Network.[J].Sensors (Basel, Switzerland),2023,23(11).

[28] Lin T Y, Dollár P., Girshick, R, et al. Feature pyramid networks for object detection[C]//Procee-dings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[29] Liu S, Qi L, Qin H, et al. Path aggregation netw-ork for instance segmentation [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[30] Zhang J, Wei X, Zhang L, et al. YOLO v7-ECA-PConv-NWD detects defective insulators

on transmission lines[J]. Electronics, 2023, 12(18): 3969.

作者简介:
苏佳　博士，教授。
杨梦凡　硕士研究生。
侯艳丽（通信作者）　博士，副教授。E-mail：286285437@qq.com。