无人机视角下施工场景目标检测性能评估

宋琢[1]，卢德辉[1]，黄志超[1]，田时雨[1]，颜嵘龙[2]，邓逸川[2,3]

(1. 广州一建建设集团有限公司，广东 广州 510060；
2. 华南理工大学土木与交通学院，广东 广州 510641；
3. 亚热带城市与建筑科学全国重点实验室，广东 广州 510641)

摘 要：施工现场的组织管理是工程管理的关键环节，但传统的人力监管方法限制多、效率低。近年国家多部委发布有关政策，呼吁促进人工智能与实体经济深度融合，以人工智能推动经济高质高效发展。计算机视觉(CV)技术的准确性、高效性和自动化等优点使 CV 技术在施工监理领域的应用逐渐广泛，特别是无人机能高效获取复杂多变的施工场景视觉数据的特性显示出其在基于 CV 技术的施工监管任务中的应用潜力。但当前基于无人机的施工场景目标检测研究有限，且稀缺的无人机视角下的施工场景图像数据集限制着有关研究的深入发展。因此，采用大疆 Mavic 3T 无人机用于获取施工现场图像，以建立开源的施工场景俯拍图像数据集 UB-CSD。选用多种先进目标检测算法在 UB-CSD 数据集上进行对比实验，从模型流程设计、计算原理和任务场景特性等维度分析各算法性能差异原因。各算法的 mAP 检测结果为 YOLOv8 和 YOLOv10 (96.1%)，YOLOv9 (96.0%)，YOLO11 (95.7%)，DETR (95.3%)，Faster-RCNN (76.3%)和 RetinaNet (72.1%)。分析结果表明，YOLO 系列算法是基于无人机的施工场景目标检测任务算法的最优选。通过构建全新的开源专用数据集和开展对比实验得出的以上数据及结论，将为建筑业安全生产管理与日后相关检测研究提供有效数据与实验案例。

关 键 词：施工场景；无人机；目标检测；YOLO；Faster-RCNN；DETR；RetinaNet

中图分类号：TU 71 **DOI**：10.11996/JG.j.2095-302X.2026010068

文献标识码：A **文 章 编 号**：2095-302X(2026)01-0068-10

Performance evaluation of construction site object detection under drone-captured perspective

SONG Zhuo[1], LU Dehui[1], HUANG Zhichao[1], TIAN Shiyu[1], YAN Ronglong[2], DENG Yichuan[2,3]

(1. Guangzhou No. 1 Construction Group Co. Ltd., Guangzhou Guangdong 510060, China;
2. School of Civil Engineering and Transportation, South China University of Technology, Guangzhou Guangdong 510641, China;
3. State Key Laboratory of Subtropical Building and Urban Science, Guangzhou Guangdong 510641, China)

Abstract: The organizational management of construction sites is a critical aspect in engineering management; however, traditional human supervision method is constrained by many environment limitations and low efficiency. In recent years, multiple government departments have issued relevant policies advocating deep integration of artificial intelligence with the real economy to promote high-quality and efficient economic development. The accuracy, efficiency, and automation advantages of Computer Vision (CV) technology have gradually led to its widespread application in the field of construction supervision. Meanwhile, the drones, which can efficiently obtain complex and varied visual data of construction scene, demonstrate their application potential in CV-based construction supervision

收稿日期：2025-03-19；**定稿日期**：2025-07-23；**通信作者**：邓逸川，E-mail：ctycdeng@scut.edu.cn

Received：19 March，2025；**Finalized**：23 July，2025；**Corresponding author**：DENG Yichuan，E-mail：ctycdeng@scut.edu.cn

基金项目：国家自然科学基金(52308314)；广东省自然科学基金-青年提升项目(2023A1515030169)；广东省住房和城乡建设厅科技创新计划项目(20250305J0004)；广州市建筑集团有限公司科技计划项目([2023]-KJ008)

Foundation items：National Natural Science Foundation of China (52308314); Youth Enhance Project of Natural Science Foundation of Guangdong Province (2023A1515030169); Technology Innovation Program of Guangdong Provincial Department of Housing and Urban-Rural Development (20250305J0004); Technology Program Project of Guangzhou Municipal Construction Group CO. LTD([2023]-KJ008)

tasks. However, the current researches on drone-based construction scene detection are limited, and the lack of overhead-perspective construction-scene image datasets restricts further development in the field. Therefore, the DJI Mavic 3T drone was utilized to obtain construction-site images to establish an open-source overhead image dataset for construction scene UB-CSD. Several advanced object-detection algorithms were selected for comparative experiments on the UB-CSD dataset, and the reasons for performance differences were analyzed from multiple dimensions such as model workflow design, computation principle, and task characteristics. The mAPs of every algorithm's detection result were YOLOv8 and YOLOv10 (96.1%), YOLOv9 (96.0%), YOLO11 (95.7%), DETR (95.3%), Faster-RCNN (76.3%) and RetinaNet (72.1%). The analysis results indicated that the YOLO series algorithm constituted the most optical algorithm for drone-based object detection tasks in construction scenes. By establishing a new open-source special dataset and conducting comparative experiments, the conclusion drawn provided effective data and experimental cases to support future safety production management and object-detection algorithm research in the construction industry.

Keywords: construction scene; drones; object detection; YOLO; Faster-RCNN; DETR; RetinaNet

《十四五建筑业发展规划》指出，“十三五”期间建筑业年均增长值为 5.1%，占国内生产总值比重保持在 6.9%以上，为经济增长与人民生活水平提升做出重大贡献。但现阶段建筑业仍存在发展方式粗放、建筑总体品质不高等多方面问题。为实现建筑业由高速发展向高质量发展的转型升级，做到对建筑项目施工现场的精细化智能化监管是关键步骤之一。

2022 年 1 月三部门联合印发的《十四五民用航空发展规划》中提出探索无人机产业新生态，呼吁无人机行业创新发展、拓展无人机服务应用领域；7 月六部委联合印发的《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》中鼓励在制造、农业和物流等重点行业深入挖掘人工智能技术应用场景，围绕关键场景促进智能经济高端、高效发展。建筑业传统的人工巡检在耗费大量人力和时间成本的同时，会因为施工现场环境复杂性、高风险施工任务限制等因素[1]出现监管盲区。计算机视觉(Computer Vision, CV)技术因可实现目标识别、姿态评估、运动追踪等关键任务，已被逐渐在施工监管领域应用并得到发展[2-3]。此外，由于快速发展的无人机(Unmanned Aerial Vehicle, UAV)具有机动性强、操作灵活、监视范围大和能获取多视角厘米级分辨率图像和视频[4]等应用优势，可有力支撑 CV 技术的数据获取。

施工监管领域已出现一些无人机与 CV 技术的应用案例，如精细化虚拟施工场景还原[5]和基于无人机数据的不安全行为检测与安全状态分析[6]。然而关于无人机技术在施工监管领域应用的调研结果显示：在知网中分别以“施工管理”“无人机应用”为关键词可得到 14.12 万与 4.4 万篇文献，但“无人机”分别与“施工”“施工管理”“施工监测”关联的搜索结果仅有 2 304 篇、170 篇和 30 篇文献；Web of Science 平台的搜索结果显示，“UAV Application”主题相关文献有 31 834 篇，而关联“Construction Management”主题后的文献仅有 326 篇，占比约 1%。调研结果表明，建筑业虽然已有结合无人机与 CV 技术开展智能监管的成功案例，但相关领域仍显研究欠缺有待填补发展。

为此，本研究将使用在多个建筑项目采集的施工场景图构建无人机视角下的施工场景目标开源数据集 UB-CSD，并通过多种先进的检测算法在该数据集中进行预测结果对比，分析并明确无人机视角下施工场景中工人与施工机械检测任务的最佳算法，为后续开展基于 CV 技术的施工行为安全智能管理提供支持。具体技术路线如图 1 所示。

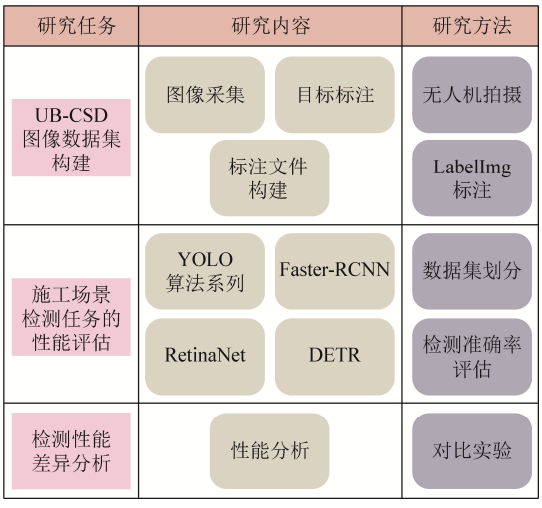

图 1　研究技术路线图

Fig. 1　Research technique route

1 研究综述

1.1 目标检测算法研究

目标检测算法的发展分为：①传统目标检测算法；②基于深度学习的目标检测算法 2 个阶段。传统目标检测算法受限于当时不具备有效表达图像信息的技术，通常基于人工设计的样本特征利用不同尺寸的滑动窗口确定感兴趣区域，再采用尺度不变特征变换、方向梯度直方图等算法提取特征，最后通过基于机器学习的分类器进行区域分类回归得到结果[5]。

卷积神经网络(Convolution Neural Networks, CNNs)诞生于上世纪 90 年代，相比传统目标检测方法具有自主学习特征提取参数、降低网络参数量和复杂度的权值共享机制、同时可进行多类学习任务等多方面优势[1]。由此基于深度学习的目标检测算法开始快速发展，目前常用方法包括两阶段算法和单阶段算法 2 类。两阶段算法包括 R-CNN 系列[7-9] (R-CNN[7], Fast-RCNN[8], Faster-RCNN[9])、SPPNet[10]和 FPN(Feature Pyramid Network)[11]等；单阶段算法主要有 YOLO (You Only Look Once)系列[12-13]、SSD (Single-Shot Multibox Detector)[14]、RetinaNet[15]、DETR (Detector Transformer)[16]等。2 类算法的架构区别在于：两阶段检测算法先根据图像生成建议区域，再通过对建议区域的分类与回归得到预测结果；单阶段检测算法基于提取的全图特征信息，直接通过分类与回归输出目标边界框与类别。因为计算步骤更少，后者检测速度一般更快，而前者则在保障检测精度的同时将耗费更长的训练时间和更多的存储空间。

1.2 基于无人机图像的检测任务挑战

无人机空中移动和大范围监视的特性会导致无人机俯拍图像时通常同时包含不同朝向、不同尺度规模及不同类别的复数待检测目标，甚至部分待检测目标会受周边复杂环境的影响而只有部分可见或被遮挡。相比于固定监控系统如塔吊视频监控系统等拍摄的高空视角图像，无人机虽然能够更灵活地获取任意视角下的感兴趣区域影像，但是小体积、低质量的无人机平台更易受到外界环境因素干扰而严重影响图像质量。WU 等[17]将上述难题概括为 5 类，包括数据质量、目标尺度多样性、目标朝向多样性、实时检测效率及目标检测。TANG 等[18]则指出，在无人机图像检测任务中复杂环境对检测结果也会产生显著影响。

该部分将简要阐述这 6 类应用难题：

1) 数据质量。无人机图像质量易受到光照、天气、相机类别、快门拍摄设置和飞行稳定性等因素[19]影响，而引发图像模糊、畸变、过曝和色彩失真等影响检测效果的质量问题。

2) 目标尺度多样性。无人机较大的监视范围与不同的飞行高度将使检测算法面对大量不同尺度表征的目标，对不同尺度下的特征信息提取、辨识与表示能力有更高要求。

3) 目标朝向多样性。俯拍视角下无人机与目标姿态的变化可能导致图像中各目标有不同旋转角朝向的边界框，要求检测算法能够学习边界框的旋转特征或在大边界框中提取目标特征。

4) 实时检测效率。无人机的载荷限制及稳定性要求与实时检测任务的高效精确的检测要求明显存在冲突，通常需牺牲部分性能或提高成本预算。

5) 复杂背景[19]。当目标周边密集分布较多物体或图像有较多背景噪声时，目标将因受遮挡或模糊化而难以被检测。

6) 小目标检测。该问题本质上是目标尺度多样性问题的外延，但由于可利用特征有限、易受环境干扰、类别不平衡和辨率限制等诸多挑战，小目标检测是本领域的检测热点与难度，小目标检测效果也是衡量目标检测算法性能的重要评价指标之一。

针对上述应用难题，学者们从多方面开展了研究以提升算法检测效果，如 DING 等[20]将变形卷积计算引入 Faster-RCNN 模型的主干网络 ResNet50 中 Cov3_x 至 Conv5_x 层以学习震后现场中倒塌建筑的不规则几何特征，提高灾情评估与救援效率；CHEN 和 JAHANSHAHI[21]提出基于主动旋转卷积核(Active Rotation Filter, ARF)[22]的 ARF-Crack 神经网络以捕捉裂缝的旋转不变特征，在混凝土裂缝数据集上取得多种检测算法中的最优综合表现；为应对患病松树与周边环境难以区分和不同株形体差异大的难题，HU 等[23]为 YOLOv5 算法添加通道注意力模块使其专注于病患特征，并以膨胀卷积提取不同尺度层级的上下文信息，在 F1 值上比 YOLOv5 高出 9.71%；BASHIR 和 WANG[24]结合循环生成对抗网络与残差聚合模块，提出一种超分辨率技术框架，并结合 YOLOv3 算法实现对小目标的高效检测。

上述改进算法的检测模型通常为适应特定任务性质而进行指向性地设计，但建筑施工场景与线

路巡检、农林植保和环境检测等无人机常用场景[25]在任务和场景特征上存在较大差异，所以现有算法对施工场景目标检测任务的性能有待评估。由于目前相对缺乏使用无人机与CV技术的智能施工监管研究与实践案例，本研究将选取多种先进检测算法，通过在无人机视角下施工场景图像数据集上的预测结构明确无人机视角下施工场景的工人与施工机械检测任务的最佳算法，从而为后续基于 CV 技术的施工行为安全管理研究提供技术支撑，填补建筑业相关领域的研究与实践空缺。

1.3　无人机图像数据集的发展现状

目标检测数据集通过提供足量的训练和验证图像及预先完成的相应人工标注，帮助检测算法学习待检测目标的特征和可能的负样本信息，以获取特定任务领域的最优检测参数。为实现基于无人机的计算机视觉技术在施工监测领域中的应用，构建相应场景的检测数据集是必要的第一步。

本小节从数据模态、图像及目标信息、应用任务场景和是否开源获取等 7 个维度，总结了 10 个现有的应用较广的代表性无人机图像数据集，见表 1。可以发现，当前已建立的基于无人机图像的检测数据集的应用目的主要是一般生活场景下的人、车检测与追踪；此外有少量服务于应急搜救任务的数据集，如 SeaDroneSee[26]和文献[27]。但可应用于建筑施工场景监测任务的无人机图像数据集领域尚处于空白状态，即使其他数据集涉及施工场景的图像，但因总量不足难以组建可有效训练的数据集。

表 1　无人机俯拍图像数据集调研表

Table 1　Survey on drone-captured image datasets

数据集	模态	图像数/k	图像尺寸	目标数/k	类别数	任务场景	开源与否
CARPK[28]	可见光	1.45	1 280×720	89.78	1	车辆计数	是
UAVDT[29]	可见光	80.00	1 080×540	840.00	3	车辆检测追踪	是
VisDrone[30]	可见光	10.21	2 000×1 500	540.00	10	多类目标检测	是
DAC-SDC[31]	可见光	150.00	640×360	—	95	多类目标检测	是
AU-Air[32]	多模态	32.82	1 920×1 080	132.00	8	交通监测	是
UVSD[33]	可见光	5.87	960×540~5 280×2 970	58.60	1	车辆检测	是
MOHR[34]	可见光	10.63	5 472×3 078/7 360×4 192/8 688×5 792	90.01	5	多类目标检测	否
DroneVehicle[35]	可见光 红外线	56.88	840×712	819.00	5	车辆检测	是
SeaDroneSee[26]	多光谱	54.00	3 840×2 160~5 456×3 632	400.00	6	海上人员检测	是
ManipalUAV[36]	可见光	13.46	1 280×720	153.11	1	行人检测	是

将正视角拍摄的施工场景检测数据集迁移至基于无人机实现的施工场景检测任务的方法可行性较低，因为无人机俯拍图像和一般正视图中人、物的特征之间存在显著差异[37]。以人为例，如图 2 所示，正视图中的人像包括头部、双臂、躯干和腿部等重要部位，而俯视图像的特征信息相比于正视图有较大差异且数量少。

图 2　平视与俯拍视角下的人体姿态对比

((a) 平视视角下人体姿态；(b) 俯拍视角下人体姿态)

Fig. 2　Human posture comparison between frontal and downward perspectives ((a) Human posture from frontal perspective; (b) Human posture under downward perspective)

2　检测方法

2.1　UB-CSD 数据集构建

由爬虫技术等方法从电子网站渠道获取的图片将面临著作权、肖像权等法律问题和图片质量不高、缺乏待检测目标物等技术问题，以及现有数据集中施工场景图片的匮乏，本研究将仅使用无人机实地采集的施工场景俯拍图像，用于构建全新开源的基于无人机俯拍图像的施工场景检测数据集 UB-CSD。

本研究使用大疆 Mavic 3T 无人机，分别对 4 个位于广州市天河区、南沙区、黄浦区的不同建筑项目的施工现场作业及周边街道交通场景进行拍摄。经过图像去重及过滤后，总计 12 862 张图像被用于建立 UB-CSD 数据集，其图像格式均为 JPG，分辨率被统一调整为 960×756，部分图像如图 3 所示。

考虑到实际施工过程中夜间作业和监测需求，UB-CSD 数据集共收集了 7 435 张白天及 5 427 张夜间施工活动照片。

为提高检测算法的鲁棒性与泛化能力，数据采集过程中操作者通过旋转调整无人机姿态、水平平移无人机和改变无人机飞行高度的方式获得部分目标旋转、平移和缩放后的图像。并于不同时段以相同视角在不同光照条件下拍摄同一目标，从而对 UB-CSD 数据集间接地应用数据增强技术，提高数据集图像多样性。部分“增强图像”如图 4 所示。

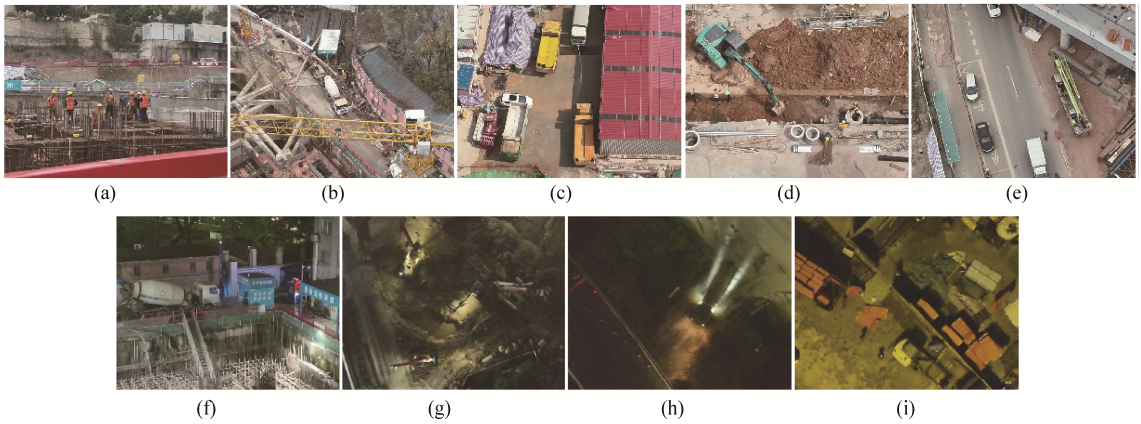

图 3 UB-CSD 数据集部分图像展示((a)~(e) 白天拍摄图像；(f)~(i) 部分夜间拍摄图像)

Fig. 3 Presentation of some images in UB-CSD dataset ((a)~(e) Daytime images; (f)~(i) Nighttime images)

图 4 部分增强图像组((a) 旋转增强；(b) 平移增强；(c) 缩放增强；(d) 亮度增强)

Fig. 4 Part of enhanced figure groups ((a) Rotation enhancement; (b) Translation enhancement; (c) Scaling enhancement; (d) Brightness enhancement)

UB-CSD 数据集覆盖人员、轿车、混凝土搅拌车、卡车、混凝土泵车、旋挖钻机、挖掘机、起重车及挖沟机共 9 种施工场地常见目标物体，是项目现场开展施工活动的主要行为主体，通过对上述目标的精确检测为后续研究施工现场的行为安全管理、实现施工场景的异常行为识别提供研究基础。图像标注作业使用基于 Python 的开源图像注释软件 LabelImg 创建目标标注框，并分别导出 PASCAL VOC 与 TXT 格式的图像标注文件。

目标中心点分布情况如图 5 所示，图中像素点颜色越深表示该位置的目标分布密度越高。可见图像中绝大部分区域均存在目标分布，且中心区域的目标分布最密集。因此基于 UB-CSD 数据集训练的算法不会出现过分关注图像局部区域的过拟合问题。

各类目标数占比如图 6 所示，数据集内总样本数为 40 296，其中人员作为施工活动主体是施工安全管理的关键要素，目标数占比最高；轿车、混凝土搅拌车、卡车和挖掘机 4 类设备在施工全过程中

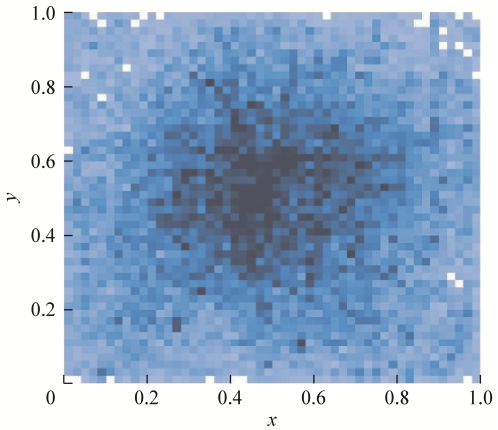

图 5 目标中心点分布图

Fig. 5 Target center point distribution

参与度较高，因此被采集的目标数占比较高；剩余 4 种施工机械的应用通常限于施工阶段的部分任务，故相应的采集样本数偏低。符合样本框的相对长宽值均小于 0.1 的定义的小目标数量为 15 402，占总目标数的 38.22%；各类别中人员类小目标有 13 981 个，占小目标总数的 90.77%，占本类别目标数的 71.29%，2 项指标均为各类别最高，所以人

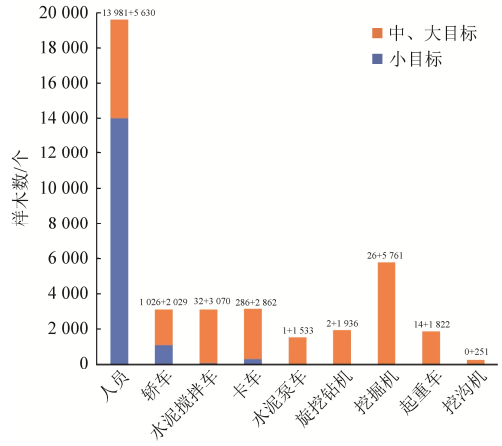

图 6　UB-CSD 数据集目标类别及尺度统计图

Fig. 6　Statistical chart about objects category and scale in US-CSD dataset

员类目标的检测难度最大；轿车类小目标在小目标总体及本类别占比分别为 6.88%和 34.32%，而卡车类为 1.86%和 9.09%，检测难度稍小于人员检测；其他 6 类的小目标数量均小于 50，可忽略其对算法训练结果和本类别目标检测效果的影响。

2.2　检测算法架构简介

为探究适用于面向施工场景的基于无人机的目标检测算法，本研究根据检测精度、检测速率、网络复杂度和训练成本等多方面指标，分别选择 YOLO 系列 v8~11 版本、Faster-RCNN、RetinaNet、DETR 共 7 种算法在 UB-CSD 数据集训练后进行测试。

2.2.1　YOLO 算法系列

YOLO 算法于 2015 年推出初始版本，目前该系列最新版本为 2024 年 9 月发布的 YOLO11。YOLO 算法的基本实现思想在于将目标检测任务简化为回归任务，仅通过一个神经网络即输出预测边界框和类别。因此 YOLO 算法系列在具有较高的检测速度的同时保证了在检测精度、召回率等指标上的较好表现。CAO 等[38]经过统计发现，YOLO 算法系列是当前实时检测算法领域使用频率最高的算法族。

v7 版本后的 4 种 YOLO 算法均放弃了锚框预测法而采取无锚的中心点预测，以减小正负样本不平衡对算法优化的负面影响。此外，这 4 种算法分别在不同的基准模型上通过融入多种优化策略和功能模块，如 v8 版本的空间金字塔快速池化模块、v9 版本提出的可编程梯度信息体系、v10 版本设计的部分自注意力机制模块等[12]，从而在 COCO 通用数据集上达到优于前代版本的检测效果。所以本文选取 YOLO 系列的 v8~v11 版本算法开展研究。

2.2.2　R-CNN 类算法

R-CNN 类算法主要包括 R-CNN，Fast-RCNN 和 Faster-RCNN 算法，属于两阶段检测算法。这类算法需先获得一定数量的建议区域，再对上述区域进行边界回归与目标分类。Faster-RCNN 通过提出区域建议网络(Region Proposal Network, RPN)取代前代算法的选择性搜索策略以提升检测效率。两阶段检测流程使得当时 R-CNN 类算法的检测精度优于单阶段算法，不过其检测效率低于单阶段检测算法。

2.2.3　RetinaNet 算法

RetinaNet 算法于 2017 年被提出，该算法的贡献在于提出了 Focal Loss 损失函数，其计算式[15]为

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log p_t \tag{1}$$

式中：p_t 表示目标类别的预测概率值，即置信度；α_t 表示正负样本平衡参数；γ 表示易分类样本的调制因子参数，且 $\gamma \geqslant 1$。Focal Loss 函数通过自适应调整不同判别难度样本的预测损失值对整体损失的影响，从而提高算法对难分样本的关注度以提升检测能力。因此 RetinaNet 算法在无明显架构改进的前提下取得与当时两阶段算法相当的准确度，但由此该算法敏感于样本标注的精确度及背景噪声。

2.2.4　DETR 算法

DETR 是一种基于 Transformer 模型的全新的端到端目标检测算法。其结构包括 CNN 主干网络、Transformer 编码器和解码器 3 部分[16]：CNN 主干网络负责提取输入图像的相应特征图；编码器对特征图作全局解析，生成相应的特征向量；解码器通过预学习的目标特征向量与编码器输出匹配，获得最终的预测结果。DETR 与一般方法的不同在于，其放弃非最大抑制算法等后处理操作和锚框预测而全面采用自注意力机制。这在简化 DETR 的检测流程和降低其对第三方库依赖的同时，可充分发挥图像全局信息对局部目标识别的作用。然而自注意力机制计算与结构复杂度更高、训练时间更长。此外，这一计算方式导致 DETR 对特征信息更少，且小目标检测性能不如大目标[39]。

3　结果分析

3.1　实验设置

本研究所使用设备的硬件为 Intel(R) Xeon(R)

CPU E5-2680 v4 @2.40 GHz 处理器与 NVIDIA GeForce RTX 2080Ti 显示适配器，所有算法均在 PyTorch 框架下搭建并运行。受限于可用的硬件环境，各算法采用的训练参数见表 2，表中“是否使用预训练权重”项为是否使用算法构建团队提供的预训练权重文件。

表 2 算法训练参数表

Table 2 Algorithm train parameter table

训练模型	是否使用预训练权重	训练世代	批处理规模	初始学习率	动量
YOLO 系列	否	100	16	0.010 0	0.9
Faster-RCNN	否	10 000	256	0.001 0	0.9
RetinaNet	否	冻结阶段：50 解冻阶段：50	冻结阶段：16 解冻阶段：8	0.000 1	0.9
DETR	是	100	4	0.000 1	0.9

完成数据集标注后，鉴于 CV 技术领域的已有研究案例[9,40-41]，分别采用 8∶1∶1，7∶2∶1 及 1∶1∶2 的比例随机划分应用上述数据集的训练集、验证集和测试集。根据可用硬件环境、预测损失与训练拟合度确认 8∶1∶1 为各算法最佳性能的对应划分比例，以该比例随机划分应用于数据集为训练集(10 289 张)、验证集(1 286 张)与测试集(1 287 张)。

目标检测算法性能的常见评估指标包括：精度(Precision, P)、召回率(Recall, R)、F1 值、平均精度(Average Precision, AP)、综合平均精度(mean Average Precision, mAP)等。P 用于衡量所有被算法检测出的目标中真实存在的目标数占比；R 旨在评估算法检测出所有目标的能力。检测结果的 P 值越高，则算法误警率越低；R 值越高，则算法的漏检率越低。理论上本研究希望算法训练结果中 2 个指标值较高，但实际应用中两者数值间通常表现出负相关，由此提出了 F1 值以综合考虑 2 种指标，正确评估算法的最终检测性能。P，R 和 F1 的计算公式为

$$P = TP/(TP+FP) \tag{2}$$

$$R = TP/(TP+FN) \tag{3}$$

$$F_1 = 2\times P\times R/(P+R) \tag{4}$$

式中：*TP* (真阳性)表示被正确预测的正样本；*FP* (假阳性)表示被错误预测的负样本；*FN*(假阴性)表示被错误预测的正样本。

AP 通过考虑不同 P-R 值组合，及计算 P-R 曲线积分来综合评估某类目标的检测效果；mAP 即所有目标类别的 AP 平均值，即

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

$$mAP = \left(\sum_{i=1}^{N} AP_i\right)\Big/N \tag{6}$$

使用 mAP 同样能较全面地反映算法性能。本研究考虑各指标效益，选择 mAP@50 作为主要评价指标，该指标表示用于判断预测结果的交并比取 50% 时的 mAP 值。

3.2 对比实验结果

按上述参数配置完成训练后，7 种检测算法在测试集上推理结果的指标值见表 3。图 7 为各算法相应指标值的三维柱状统计图。

mAP 统计结果显示：YOLO 系列 v8~11 版本算法展现的性能效果最佳，达到 95.7%~96.1%，其中 YOLO11 为系列最低值 95.7%，YOLOv9 结果为 96.0%，YOLOv8 和 v10 取得系列最高值 96.1%；DETR 推理结果的 mAP 值略小于 YOLO 系列，为 95.3%；Faster-RCNN 和 RetinaNet 算法检测结果的 mAP 则显著落后于前 5 种算法，分别为 76.3%和 72.1%。因此根据 mAP 值，目前无人机俯拍的施工

表 3 算法检测性能指标对比

Table 3 Comparison on detection performance indicators between algorithms

算法	人员	轿车	水泥搅拌车	卡车	水泥泵车	旋挖钻机	挖掘机	起重车	挖沟机	mAP
YOLOv8	**0.888**	0.921	**0.981**	**0.973**	0.987	**0.994**	**0.993**	**0.985**	0.931	**0.961**
YOLOv9	0.887	0.913	0.980	0.968	**0.988**	**0.994**	0.992	0.984	0.934	0.960
YOLOv10	**0.888**	0.921	**0.981**	**0.973**	0.987	**0.994**	**0.993**	**0.985**	0.931	**0.961**
YOLO11	0.885	0.911	0.980	0.970	0.984	**0.994**	**0.993**	0.982	0.911	0.957
Faster-RCNN	0.484	0.743	0.888	0.754	0.784	0.871	0.812	0.799	0.733	0.763
RetinaNet	0.360	0.654	0.885	0.666	0.838	0.757	0.877	0.705	0.744	0.721
DETR	0.802	**0.951**	0.973	0.940	0.979	0.990	0.988	0.979	**0.979**	0.953

注：加粗数据表示最优值。

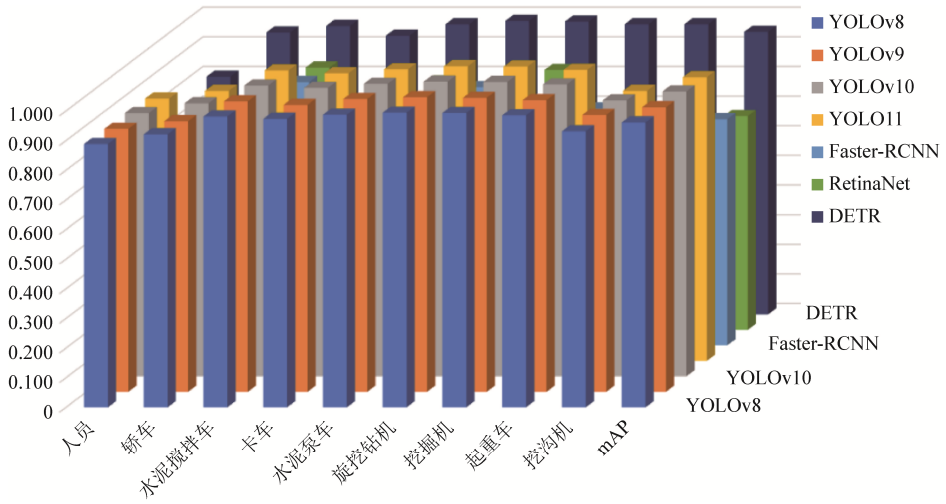

图 7　mAP 与 AP 指标统计图

Fig. 7　Statistical chart about mAP and AP parameter

场景的检测任务应采用 YOLOv8~v10 算法；YOLO11 和 DETR 算法表现与前三者相差无几，仍是可考虑的备选项；Faster-RCNN 和 RetinaNet 并未表现出优秀的性能，不易宜选择。

从各类目标检测结果来看：在小目标占比最高、检测难度最大的人员类检测结果中，YOLO 系列算法表现最好，AP 值均超过 88.0%；DETR 的结果仅为 80.2%，与 YOLO 算法存在一定差距；Faster-RCNN 和 RetinaNet 对人员检测结果的 AP 值未达到 50%，无法满足使用要求。在另外 8 类目标上，DETR 算法在轿车与挖沟机类的检测中取得高于 YOLO 系列算法的结果，对水泥搅拌车、水泥泵车、旋挖钻机、挖掘机和起重车 5 类目标检测得到的 AP 值比 YOLO 系列均值的差距不到 1%，仅在卡车类的检测结果与 YOLO 系列均值存在略大差距；而 Faster-RCNN 与 RetinaNet 算法得出的 8 类目标 AP 值均比 YOLO 系列与 DETR 算法结果的最小值低 9%以上，并不是第一选择。

由于人员在施工现场安全管理中的重要性，所以自 YOLOv8 版本起的 YOLO 系列算法是基于无人机俯拍施工场景检测任务的优选，其中 YOLOv8 与 YOLOv10 具有最强的性能表现。DETR 算法同样是施工场景检测任务的可选项，但当涉及人员检测方面时，需要对 DETR 进行适当改进以保证足够的检测效能。Faster-RCNN 与 RetinaNet 算法检测结果较差，表明 2 种算法并不适用于无人机俯拍图像的施工场景检测任务，在实际研究中不建议采用。

3.3　结果分析

本文以 YOLO 系列为基准，对上述实验结果中的指标值差异进行探讨，以分析导致性能差距的架构影响与未来可能的算法改进措施。

1) DETR 对人员目标的检测效果不佳。如第 2 节所述，UB-CSD 数据集中 19 611 个人员类目标里共有 71.29%属于小目标，因此算法对人员类的检测结果可以视作该模型对建筑施工场景中小目标的检测性能的等效评价。由此，DETR 对人员目标检测效果不佳的结果应看作 DETR 对施工场景下的小目标检测不佳的表现。

文献[39]为提升 DETR 的训练收敛速率开展消融实验，发现删去解码器、仅采用编码器架构能够显著提升改进的 DETR 算法对小目标的检测性能，同时算法对大目标的检测性能出现一定下降。原因应在于大目标存在更多的潜在特征匹配点，即大目标相比于小目标具有的可识别特征信息更丰富，从而在通过多头交叉注意力机制层的计算后能够更准确地被识别出。

自注意力机制模型对于小目标的检测性能同样可能产生负面影响。一方面，自注意力机制的计算原理使得该步骤的计算复杂度为 $O(N^2)$，其中 N 为特征图像素数。更复杂的计算过程导致 DETR 在处理多尺度目标图时取得的表现将不如其他一阶段检测算法，特别是关于小目标的检测结果。另一方面，初始化的自注意力机制为全图像素点分配的权重值基本相同，需要 DETR 通过大量训练过程来达到充分的损失收敛，造成占用像素点数少、特征信息少的小目标检测任务难度更大。

根据上述分析，可发现 Transformer-based 的检测算法应当充分考虑检测任务的特性，基于场景可能的待检测对象合理取舍模型中的解码器架构，并

且考虑基于查询对象的重要性、潜在风险和类别比重等因素自适应地为检测目标提供不同权重。

2) 关于 Faster-RCNN 与 RetinaNet 算法的整体性能不佳。虽然 Ultralytics 团队提供的 YOLO 算法训练文件默认为数据集图像应用一定程度的随机数据增强技术[13]，但不采用以上增强措施的 YOLO 算法推理性能与默认设置下的 YOLO 算法基本相当，并且 UB-CSD 数据集通过人工调整拍摄设置的方式被等效增强因此 YOLO 算法的默认数据增强操作并不是这 2 类算法表现差距较大的主要原因。

主要原因在于 Faster-RCNN 与 RetinaNet 2 种算法提出的网络结构与实现方法，已经为当前主流算法学习融入至自身模型，甚至得到进一步改进。2 种算法分别以 RPN 模型和 FPN 模型输出目标的预测边界框。RPN 模型仍采用基于锚框的预测与回归方法，更易受目标尺度差异、正负样本不平衡等因素的影响而不能取得较好的表现。FPN 模型架构能有效利用并融合浅层特征图的空间信息与深层特征图的语义上下文信息，但易感于图像噪声与标注精确度的 Focal Loss 函数和单阶段的流程设计使得 RetinaNet 算法对混凝土泵车、旋挖钻机、挖掘机、起重车和挖沟机等通常包含较多无关背景信息的中大型施工机械边界框的检测表现落后于 Faster-RCNN，更不如于 YOLO 算法。而如今的 YOLO 算法主要使用基于 FPN 模型改进的 PANet 模型和无锚的中心点预测法，且各版本 YOLO 算法进一步结合其他算法提出的功能模块，如注意力机制、变形卷积和扩张卷积等，以在前代版本的基础上进一步提升性能。

综上所述，Faster-RCNN 与 RetinaNet 算法不适合作为基于无人机的施工场景检测任务的主干模型，但其中的特定模块如 Faster-RCNN 构建的 RPN 网络与 RetinaNet 提出的 Focal Loss 函数，对于未来的算法设计与改进方法仍有学习与引入的价值。

4 研究总结

本研究通过采集 4 个项目施工现场及周边活动场景图像，建立了开源的基于无人机俯拍图像的施工场景目标检测图像数据集 UB-CSD，并选择 YOLOv8~11，Faster-RCNNRetinaNet 和 DETR 共 7 个有代表性的先进算法在 UB-CSD 数据集训练及测试。经过实验结果对比，认为 YOLO 系列中 YOLOv8 与 YOLOv10 是最适合无人机俯拍的施工场景的检测算法；Faster-RCNN 与 RetinaNet 算法则不能满足任务需求，不予考虑；剩余 3 种算法 YOLOv9，YOLO11 与 DETR 的性能表现略逊于 YOLOv8 与 YOLOv10，仍具备实际应用与深入研究的潜能。之后本文从架构设计与计算方式等技术因素出发，深入分析 DETR，Faster-RCNN 与 RetinaNet 算法表现逊色于 YOLO 算法系列的原因，并提出未来的可能提升措施。

但本研究还存在一定不足，主要在于对比实验选用的算法虽具有一定代表性，但并未覆盖目标检测领域当前所有的优秀算法，因此本文对比实验得出的结论存在局限性。未来将基于 YOLO 系列算法和本领域的任务特性，进一步开发基于无人机的施工场景实时监测系统，并投入实际项目验证和应用。

参考文献 (References)

[1] 朱密. 基于图像语义的建筑施工风险场景识别[D]. 大连. 大连理工大学, 2020.
ZHU M. Recognition of high-risk scenarios in building construction based on image semantics[D]. Dalian: Dalian University of Technology, 2020 (in Chinese).

[2] 崔自强, 杨淑娟, 于德湖. 人工智能在建筑施工领域应用研究进展[J]. 山东建筑大学学报, 2023, 38(4): 117-125, 134.
CUI Z Q, YANG S J, YU D H. Research progress on the application of artificial intelligence in the field of building construction[J]. Journal of Shandong Jianzhu University, 2023, 38(4): 117-125, 134 (in Chinese).

[3] PANERU S, JEELANI I. Computer vision applications in construction: current state, opportunities & challenges[J]. Automation in Construction, 2021, 132: 103940.

[4] 吴一全, 童康. 基于深度学习的无人机航拍图像小目标检测研究进展[J]. 航空学报, 2025, 46(3): 30848.
WU Y Q, TONG K. Research advances on deep learning-based small object detection in UAV aerial images[J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(3): 30848 (in Chinese).

[5] 尹东. 基于无人机和计算机视觉的智慧工地管理方法研究[D]. 长沙: 湖南大学, 2022.
YIN D. Study of intelligent construction site management based on UAV and computer vision[D]. Changsha: Hunan University, 2022 (in Chinese).

[6] 石智强. 基于无人机遥感数据的施工现场不安全行为检测和安全状态分析研究[D]. 宜昌: 三峡大学, 2023.
SHI Z Q. Research on unsafe behavior detection and safety state analysis of construction site based on UAV remote sensing data[D]. Yichang: China Three Gorges University, 2023 (in Chinese).

[7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[8] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision . New York: IEEE Press, 2015: 1440-1448.

[9] 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.

[10] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[11] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 936-944.

[12] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements[EB/OL]. [2025-03-05]. https://arxiv.org/pdf/2410.17725.

[13] JOCHER G. Ultralytics YOLO[EB/OL]. [2025-03-05]. https://github.com/ultralytics/ultralytics.

[14] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//The 14th European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[15] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[16] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//The 16th European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[17] WU X, LI W, HONG D F, et al. Deep learning for unmanned aerial vehicle-based object detection and tracking: a survey[J]. IEEE Geoscience and Remote Sensing Magazine, 2022, 10(1): 91-124.

[18] TANG G Y, NI J J, ZHAO Y H, et al. A survey of object detection for UAVs based on deep learning[J]. Remote Sensing, 2024, 16(1): 149.

[19] XIANG T Z, XIA G S, ZHANG L P. Mini-unmanned aerial vehicle-based remote sensing: techniques, applications, and prospects[J]. IEEE Geoscience and Remote Sensing Magazine, 2019, 7(3): 29-63.

[20] DING J J, ZHANG J H, ZHAN Z Q, et al. A precision efficient method for collapsed building detection in post-earthquake UAV images based on the improved NMS algorithm and faster R-CNN[J]. Remote Sensing, 2022, 14(3): 663.

[21] CHEN F C, JAHANSHAHI M R. ARF-Crack: rotation invariant deep fully convolutional network for pixel-level crack detection[J]. Machine Vision and Applications, 2020, 31(6): 47.

[22] ZHOU Y Z, YE Q X, QIU Q, et al. Oriented response networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 4961-4970.

[23] HU G S, YAO P, WAN M Z, et al. Detection and classification of diseased pine trees with different levels of severity from UAV remote sensing images[J]. Ecological Informatics, 2022, 72: 101844.

[24] BASHIR S M A, WANG Y. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network[J]. Remote Sensing, 2021, 13(9): 1854.

[25] 蒋文全, 高豪云, 郑佳秋, 等. 无人机在民用行业应用研究综述[J]. 机电工程技术, 2025, 54(9): 119-124, 183.
JIANG W Q, GAO H Y, ZHENG J Q, et al. Review of researches on the application of UAV in the civilian industry[J]. Mechanical & Electrical Engineering Technology, 2025, 54(9): 119-124, 183 (in Chinese).

[26] VARGA L A, KIEFER B, MESSMER M, et al. SeaDronesSee: a maritime benchmark for detecting humans in open water[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE Press, 2022: 3686-3696.

[27] DENG J N, SHI Z G, ZHUO C. Energy-efficient real-time UAV object detection on embedded platforms[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 39(10): 3123-3127.

[28] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network[C]//2017 IEEE International Conference on Computer Vision. New York: IEEE Press, 2017: 4165-4173.

[29] DU D W, QI Y K, YU H Y, et al. The unmanned aerial vehicle benchmark: object detection and tracking[C]//The 15th European Conference on Computer Vision. Cham: Springer, 2018: 375-391.

[30] ZHU P F, WEN L Y, BIAN X, et al. Vision meets drones: a challenge[EB/OL]. [2025-03-05]. https://arxiv.org/abs/1804.07437.

[31] XU X W, ZHANG X Y, YU B, et al. DAC-SDC low power object detection challenge for UAV applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 392-403.

[32] BOZCAN I, KAYACAN E. AU-AIR: a multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance[C]//2020 IEEE International Conference on Robotics and Automation. New York: IEEE Press, 2020: 8504-8510.

[33] ZHANG W, LIU C S, CHANG F L, et al. Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images[J]. Remote Sensing, 2020, 12(11): 1760.

[34] ZHANG H J, SUN M S, LI Q, et al. An empirical study of multi-scale object detection in high resolution UAV images[J]. Neurocomputing, 2021, 421, 173-182.

[35] SUN Y M, CAO B, ZHU P F, et al. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713.

[36] AKSHATHA K R, KARUNAKAR A K, SHENOY B S, et al. Manipal-UAV person detection dataset: a step towards benchmarking dataset and algorithms for small object detection[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 195: 77-89.

[37] AHMED I, AHMAD M, ADNAN A, et al. Person detector for different overhead views using machine learning[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(10): 2657-2668.

[38] CAO Z, KOOISTRA L, WANG W S, et al. Real-time object detection based on UAV remote sensing: a systematic literature review[J]. Drones, 2023, 7(10): 620.

[39] SUN Z Q, CAO S C, YANG Y M, et al. Rethinking transformer-based set prediction for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 3591-3600.

[40] SZELISKI R. Computer vision: algorithms and applications[M]. 2nd ed. New York: Springer, 2022: 30-35.

[41] DEAN J, CORRADO G S, MONGA R, et al. Large scale distributed deep networks[C]//The 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2012: 1223-1231.