控制与决策
Control and Decision

基于多尺度融合和高分辨特征增强的无人机航拍目标检测

陈志旺, 肖迪创, 吕昌昊, 李思哲, 彭勇

引用本文:
陈志旺, 肖迪创, 吕昌昊, 等. 基于多尺度融合和高分辨特征增强的无人机航拍目标检测[J]. 控制与决策, 2025, 40(7): 2290-2299.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.1392

基于多尺度融合和高分辨特征增强的无人机航拍目标检测

陈志旺[1,2†]，肖迪创[1,2]，吕昌昊[3]，李思哲[1,2]，彭 勇[4]

(1. 燕山大学 智能控制系统与智能装备教育部工程研究中心，河北 秦皇岛 066004；
2. 燕山大学 河北省工业计算机控制工程重点实验室，河北 秦皇岛 066004；
3. 燕山大学 河北省电力电子节能与传动控制重点实验室，河北 秦皇岛 066004；
4. 燕山大学 电气工程学院，河北 秦皇岛 066004)

摘 要: 无人机飞行高度的动态变化使得航拍图像中往往包含大量小目标, 同时目标尺度变化显著, 这些问题给目标检测任务带来了挑战. 针对上述问题, 提出一种基于多尺度融合和高分辨特征增强的无人机航拍目标检测方法. 首先, 在骨干网络中引入多尺度结构重参数化特征提取模块, 利用普通卷积块和结构重参数化的大核卷积块对多个分支进行不同尺度的卷积运算, 有效提取不同感受野下的特征信息; 然后, 在颈部网络中引入基于特征金字塔网络的多维特征自适应融合模块, 以优化其自下而上的特征聚合过程, 实现对浅层特征中的精细细节和深层特征中的上下文信息的自适应选择, 从而更有效地应对目标尺度显著变化; 最后, 在颈部网络中引入多尺度特征融合小目标增强模块, 以捕捉无人机航拍图像中小目标物体在不同尺度上的变化. 通过在 VisDrone2019 和 TinyPerson 两个公开数据集上进行大量的实验, 表明了所提出方法的有效性和优越性.

关键词: 无人机航拍图像；目标检测；结构重参数化；多维特征自适应融合；高分辨特征增强

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1392

引用格式: 陈志旺, 肖迪创, 吕昌昊, 等. 基于多尺度融合和高分辨特征增强的无人机航拍目标检测 [J]. 控制与决策, 2025, 40(7): 2290-2299.

UAV aerial target detection based on multi-scale fusion and high-resolution feature enhancement

CHEN Zhi-wang[1,2†]，*XIAO Di-chuang*[1,2]，*LV Chang-hao*[3]，*LI Si-zhe*[1,2]，*PENG Yong*[4]

(1. Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment，Yanshan University，Qinhuangdao 066004，China；2. Hebei Key Laboratory of Industrial Computer Control Engineering，Yanshan University，Qinhuangdao 066004，China；3. Key Laboratory of Power Electronics for Energy Conservation and Drive Control of Hebei Province，Yanshan University，Qinhuangdao 066004，China；4. School of Electrical Engineering，Yanshan University，Qinhuangdao 066004，China)

Abstract: The dynamic change of UAV flight height makes aerial images often contain a large number of small targets, and the target scale changes significantly. These problems bring challenges to the target detection task. In view of the above problems, this paper proposes a UAV aerial target detection method based on multi-scale fusion and high-resolution feature enhancement. Firstly, a multi-scale structure re-parameterized feature extraction module is introduced into the backbone network. The convolution operation of different scales is performed on multiple branches by using ordinary convolution blocks and structure re-parameterized large-core convolution blocks, and the feature information under different receptive fields is effectively extracted. Then, a multi-dimensional feature adaptive fusion module based on the feature pyramid network is introduced into the neck network to optimize its bottom-up feature aggregation process, so as to realize the adaptive selection of fine details in shallow features and context information in deep features, so as to deal with the significant change of target scale more effectively. Finally, a multi-scale feature fusion small target enhancement module is introduced in the neck network to capture the changes of small target objects in UAV aerial images at different scales. Extensive experiments on two public datasets, VisDrone2019 and TinyPerson, demonstrate the effectiveness and superiority of the proposed method.

Keywords: UAV aerial images；target detection；structural reparameterization；multidimensional feature adaptive fusion；high resolution feature enhancement

收稿日期: 2024-12-02；**录用日期:** 2025-02-24.

基金项目: 国家自然科学基金项目 (61573305)；河北省自然科学基金项目 (F2022203038, F2019203511)；河北省级重点实验室绩效补助经费项目 (22567612H).

†通信作者. E-mail: czwaaron@ysu.edu.cn.

引　言

随着目标检测和无人机技术的快速发展, 无人机视角的目标检测已成为计算机视觉领域的前沿课题, 在多个领域中得到了广泛应用. 在森林火灾监控中[1], 无人机提供的实时数据传输、广域覆盖和精准定位, 有助于快速响应火情, 降低扑灭火灾的时间成本, 并减少救援人员风险; 在农作物检测中[2], 无人机结合目标检测算法可识别病虫害和营养缺乏, 能够帮助农民及时采取措施防止病害扩散, 保障作物健康; 在输电线路检测中[3], 无人机能够在高风险区域执行关键部件的检测, 提高了巡检安全性; 此外, 在交通管理中[4], 无人机通过灵活机动性和先进目标检测算法, 对行人和车辆进行实时检测和监控, 缓解了交通拥堵. 无人机结合目标检测技术的广泛应用显著提高了监测效率和安全性, 推动了各行业的智能化和现代化进程.

在深度学习目标检测领域, 现有的算法主要可划分为 3 大类: 1) 两阶段算法, 如 R-CNN 系列[5]; 2) 单阶段算法, 如 YOLO 系列[6] 和 SSD 系列[7]; 3) 端到端算法, 如 DETR 系列[8]. 这些检测算法的发展显著推动了目标检测技术的进步. 然而, 研究表明, 这些方法直接应用于无人机图像目标检测任务时效果并不理想. 无人机作为独特的遥感平台, 其图像通常具有小目标比例高、目标分布密集以及目标尺度差异较大的问题. 针对这些问题, 赵亮等[9] 提出了一种全局与局部图像特征自适应融合的一阶段小目标检测算法 SODet, 以提升小目标检测精度; Zhu 等[10] 提出了一种基于 Transformer 预测头改进的 YOLOv5 无人机捕获场景目标检测模型 TPH-YOLOv5, 以解决无人机图像中目标尺度变化较大的问题; Peng 等[11] 提出了一种基于高效卷积和多尺度特征融合的小型目标检测器 PS-YOLO, 以解决小目标的信息丢失问题; 周葳楠等[12] 提出了一种增强弱特征表达的一阶段轻量级小目标检测算法 SA-YOLO, 以应对复杂背景下小目标特征被背景噪声淹没的问题; Zhang[13] 提出了一种基于 YOLOv8 模型的多尺度无人机图像目标检测算法 Drone-YOLO, 以解决大场景尺寸和小检测对象的问题; Xu 等[14] 提出了一种基于 Vision Transformer 主干的车辆目标检测模型 YOLO-HV, 以解决传统 CNN 网络无法整合上下文信息而导致的多尺度目标识别性能差的问题.

为了更好地解决无人机图像中小目标比例高、目标分布密集以及目标尺度差异较大的问题, 实现对目标物体的精确检测, 本文提出一种基于多尺度融合和高分辨特征增强的无人机航拍目标检测算法. 首先, 设计多尺度结构重参数化特征提取模块 (MSRFEM), 以增强骨干网络对小目标信息的特征提取能力; 然后, 设计多维特征自适应融合模块 (MDFAFM), 用于自适应地选择浅层特征中的精细细节和深层特征中的上下文语义信息, 以提高对于不同目标尺度的适应性; 最后, 设计多尺度特征融合小目标增强模块 (MFFSTEM), 以捕捉小目标在不同空间尺度上的变化, 从而提升小目标的检测效果.

1　本文方法

所提出基于多尺度融合和高分辨特征增强的无人机航拍目标检测算法整体结构如图 1 所示.

该模型首先引入了一个分辨率为160 × 160的小目标检测层P_2, 以增强对小尺寸目标的检测性能; 然后, 在骨干网络的最后一次下采样后引入了多尺度结构重参数化特征提取模块 (MSRFEM), 其能够在不同感受野下提取目标的特征信息; 接着, 在颈部网络中引入了基于特征金字塔网络的多维特征自适应融合模块 (MDFAFM), 以有效应对无人机图像中目标尺度差异较大的问题; 最后, 在颈部网络中引入了基于高分辨特征图P_2的多尺度特征融合小目标增强模块 (MFFSTEM), 以进一步提升网络模型对小目标的检测能力.

1.1　小目标检测层

当输入的无人机图像中存在大量小于8 × 8像素的小尺度目标时, 就会出现检测精度低和漏检的问题. 为了解决此问题, 本文在保持输入图像尺寸不变的前提下, 引入了一个分辨率为160 × 160的小目标检测层P_2. 该检测层能够捕捉低至4 × 4像素的小目标, 从而显著提升对小目标的检测能力.

1.2　多尺度结构重参数化特征提取模块

在传统的骨干网络中, 随着网络深度的增加, 小目标的特征信息往往容易丢失, 进而可能会影响深层特征对小目标与背景的区分能力. 为解决这一问题, 本文设计了多尺度结构重参数化特征提取模块 (MSRFEM), 并将其集成至如图 1 所示的骨干网络中的最后一次下采样之后. 如图 2 所示, MSRFEM 利用普通卷积块和结构重参数化的大核卷积块[15], 对多个分支进行不同尺度的卷积运算, 从而能够在不同大小感受野下提取目标的特征信息, 减少特征图在下采样过程中特征信息的丢失, 更好地理解目标的上下文信息, 减少误检并提高检测准确性.

结构重参数化的大核卷积块 (DRB) 中的超参数包括大核卷积的大小K, 并行小核卷积大小k和膨胀

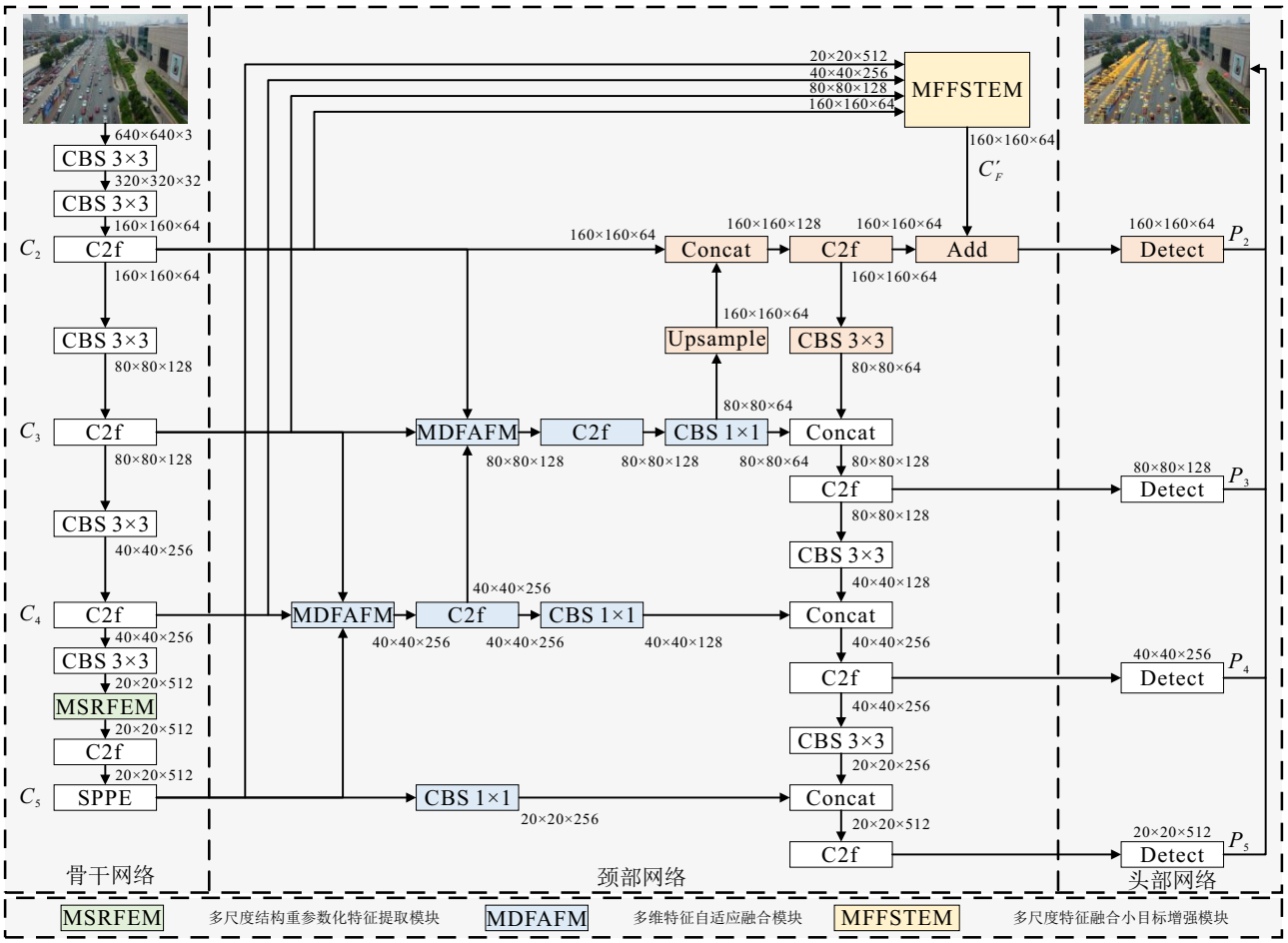


图1 模型网络结构

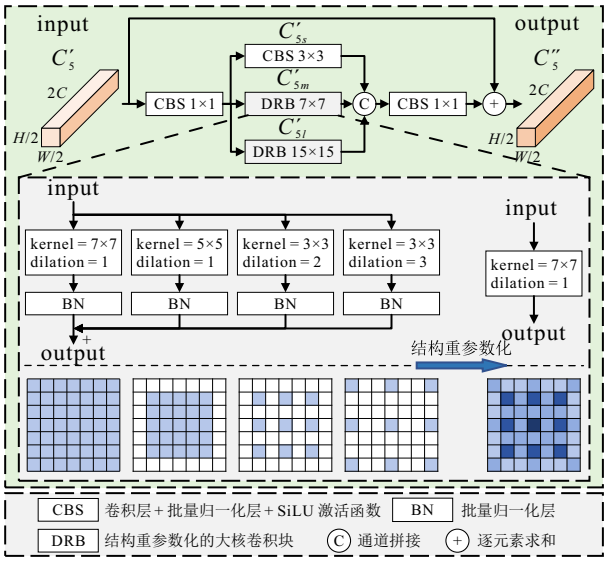


图2 多尺度结构重参数化特征提取模块

率r. 通过在像素间插入零填充, DRB 能够有效扩展小核卷积的感受野, 其等效卷积核大小可表示为$(k-1)r+1$. 图 2 重点展示了$K=7$时的 DRB 结构, 其包含 3 个并行小核卷积分支, 卷积核大小为$k=(5,3,3)$, 膨胀率为$r=(1,2,3)$, 对应的等效感受野大小为$(5,5,7)$. 当$K=15$时, 对应的并行分支为$k=(5,7,3,3,3)$, $r=(1,2,3,5,7)$, 等效感受野大小为$(5,13,7,11,15)$. 在推理阶段, 采用结构重参数化方法将并行分支上的卷积层和对应的批量归一化层融合, 并将膨胀率$r>1$的每个扩展层转化为单一卷积核后相加. 这种组合能够增强模型对局部和全局特征的提取能力. 在后文第 2.3.2 节中, 详细对比了 MSRFEM 中不同 DRB 分支组合的检测效果.

MSRFEM 包括 4 个分支, 每个分支对应不同的感受野大小. 第 1 个分支采用残差结构[16], 以形成等效映射, 保留小目标的关键特征. 在执行剩余 3 个分支前, 为了降低参数量并控制模型的复杂度, 首先利用1×1卷积将输入特征图C'_5的通道数由$2C$降低为C. 然后, 第 2 个细粒度分支通过3×3卷积提取小感受野特征, 生成输出特征图C'_{5s}, 有

$$C'_{5s}=\mathrm{CBS}_{3\times3}(\mathrm{CBS}_{1\times1}(C'_5)). \tag{1}$$

第 3 个近距离上下文分支使用$K=7$的 DRB 模块提取中感受野特征, 生成输出特征图C'_{5m}, 如下所示:

$$C'_{5m}=\mathrm{DRB}_{7\times7}(\mathrm{CBS}_{1\times1}(C'_5)). \tag{2}$$

第 4 个远距离上下文分支使用$K=15$的 DRB 模块提取大感受野特征, 生成输出特征图C'_{5l}, 即

$$C'_{5l}=\mathrm{DRB}_{15\times15}(\mathrm{CBS}_{1\times1}(C'_5)). \tag{3}$$

其中: $\mathrm{CBS}_{1\times1}(\cdot)$和$\mathrm{CBS}_{3\times3}(\cdot)$分别为$1\times1$和$3\times3$

卷积操作批量归一化和 SiLU 激活函数; $\mathrm{DRB}_{7\times7}(\cdot)$ 和 $\mathrm{DRB}_{15\times15}(\cdot)$ 分别为 7×7 和 15×15 的结构重参数化的大核卷积块.

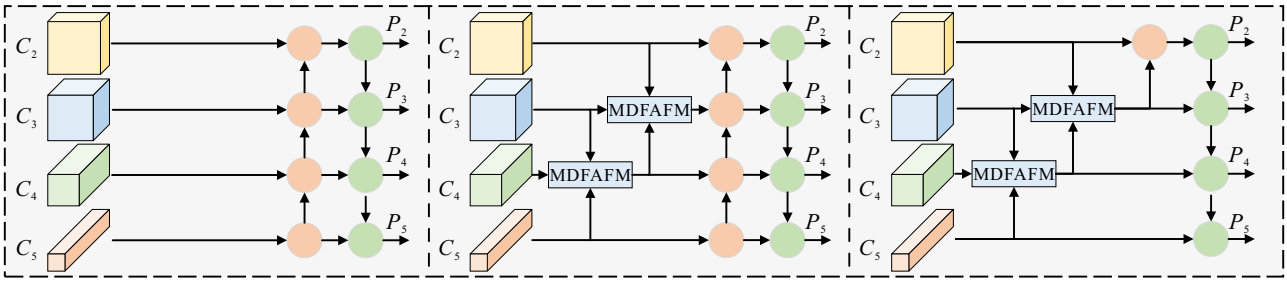
将输出特征图 C'_{5s}、C'_{5m} 和 C'_{5l} 在通道维度上进行 Concat 拼接后通过 1×1 卷积将其通道数调整至 $2C$, 并与输入特征图 C'_5 进行逐元素相加, 得到最终的特征图 C''_5 为

$$C''_5 = \mathrm{CBS}_{1\times1}(\mathrm{Concat}(C'_{5l}, C'_{5m}, C'_{5s})) \oplus C'_5, \quad (4)$$

其中 $\oplus$ 表示逐元素相加.

1.3　多维特征自适应融合模块

在处理无人机图像中目标尺度差异较大的问题时, 传统特征金字塔网络 (FPN) 由于其固定的信息传递模式, 可能会导致关键信息丢失, 影响目标检测效果. 在传统的 FPN 架构中, 层间信息交互仅局限于相邻层, 无法充分融合浅层特征中的空间位置信息和深层特征中的语义信息, 限制了其在检测目标尺度差异较大时的性能表现. 针对上述问题, 本文设计基于 FPN 的多维特征自适应融合模块 (MDFAFM). 在该模块集成到特征金字塔网络自下而上的特征聚合过程中, 根据目标物体的大小和特征自适应地选择合适的特征进行融合, 从而更有效地应对目标尺度变化显著的挑战. 在后文第 2.3.3 节中, 详细对比分析了图 3 中 3 种改进 FPN 网络结构的检测性能.



(a) FPN+PAN 网络结构　　(b) MDFAFM+FPN+PAN 网络结构　　(c) MDFAFM+PAN 网络结构

图3　特征金字塔网络 3 种改进方法的结构对比示意图

图 4 为多维特征自适应融合模块. 如图 4 所示, 以输出特征图 (C_3, C_4, C_5) 为例, 在特征自适应融合前, 需要调整各特征图的通道数和尺寸以与中间层特征图匹配. C_3 层输出的大尺寸特征图通过 1×1 卷积使其通道数由 $C/2$ 变为 C, 然后采用最大池化和平均池化的并行结构对其进行下采样, 生成输出特征图 $\bar{C}_3$. 这种并行结构有助于在下采样过程中保持无人机图像中目标物体的高分辨特征, 其过程如下所示:

$$\bar{C}_3 = \mathrm{Max}(\mathrm{CBS}_{1\times1}(C_3)) \oplus \mathrm{Avg}(\mathrm{CBS}_{1\times1}(C_3)). \quad (5)$$

其中: $\mathrm{Max}(\cdot)$ 为全局最大池化, $\mathrm{Avg}(\cdot)$ 为全局平局池化, $\oplus$ 表示逐元素相加.

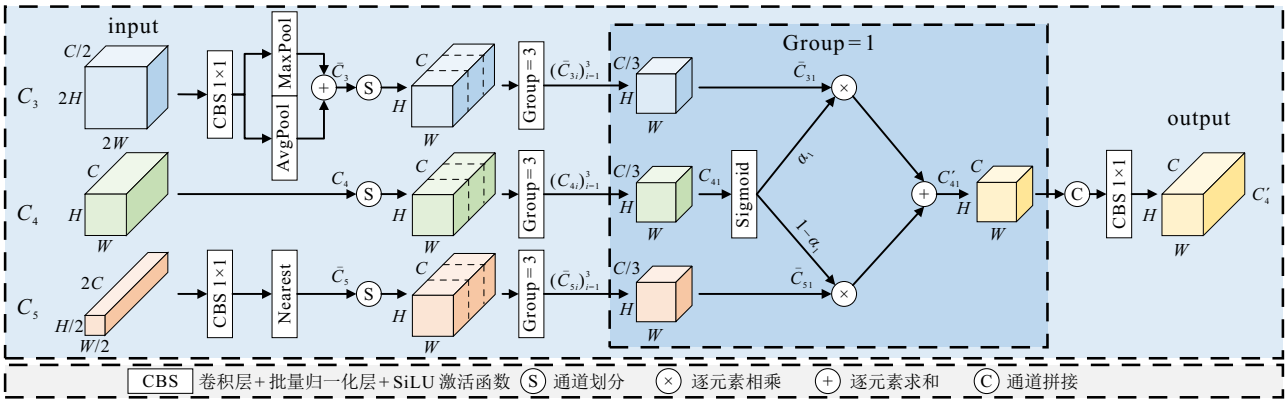


图4　多维特征自适应融合模块

C_5 层输出的小尺寸特征图也通过 1×1 卷积使其通道数由 $2C$ 变为 C, 然后采用最近邻插值法进行上采样得到输出特征图 $\bar{C}_5$. 此过程有助于保持低分辨率图像的局部特征丰富性, 减少小目标特征信息的丢失, 其过程如下所示:

$$\bar{C}_5 = \mathrm{NN}(\mathrm{CBS}_{1\times1}(C_5)), \quad (6)$$

其中 $\mathrm{NN}(\cdot)$ 为最近邻插值法.

将输出特征图 $\bar{C}_3$、$\bar{C}_5$ 以及中间层特征图 C_4 在通道维度上划分为 3 个相等的子特征图, 分别记为 $(\bar{C}_{3i})_{i=1}^3$, $(C_{4i})_{i=1}^3$ 和 $(\bar{C}_{5i})_{i=1}^3$. 通过特征自适应选择融合机制, 对特征图 C_4 的 3 个相等的通道子特征图 $(C_{4i})_{i=1}^3$ 分别应用 Sigmoid 函数计算出选择权重

$(a_i)_{i=1}^3$, 从而使得模型能够动态地调整浅层特征与深层特征的贡献比例. 通过通道划分策略使得模型能够针对每个子集进行自适应地特征选择和融合, 提升其灵活性和效率, 其过程如下所示:

$$\alpha_i = \text{Sigmoid}(C_{4i}), \tag{7}$$

$$C'_{4i} = \alpha_i \otimes \bar{C}_{3i} + (1-\alpha_i) \otimes \bar{C}_{5i}. \tag{8}$$

其中: Sigmoid$(\cdot)$为激活函数, $\otimes$表示逐元素相乘. 最后, 将自适应融合后的特征图C'_{41}、C'_{42}和C'_{43}在通道维度上进行 Concat 拼接, 并通过1×1卷积来增强通道间的特征交互, 最终得到输出特征图C'_4, 即

$$C'_4 = \text{CBS}_{1\times1}(\text{Concat}(C'_{41}, C'_{42}, C'_{43})), \tag{9}$$

这里Concat$(\cdot)$为通道拼接操作.

1.4 多尺度特征融合小目标增强模块

尽管本文在算法中增加了小目标检测层P_2来提高对无人机图像中小目标的检测能力, 但是仍然存在一定局限性. 首先, 小目标特征在经骨干网络的多层卷积操作后会被逐渐稀释, 难以在检测层中有效恢复; 其次, 小目标检测P_2有可能没有充分利用所有有用的特征层, 导致一些浅层特征图中包含的细节信息在特征融合过程中未被充分利用, 从而影响检测效果. 为了解决上述问题, 本文设计了基于高分辨特征图P_2的多尺度特征融合小目标增强模块(MFFSTEM), 如图 1 所示. 将经 MFFSTEM 输出的特征图C'_F与小目标检测层P_2前的输出特征图进行逐元素相加, 相加后的多尺度特征图能够有效地捕捉无人机图像中小目标物体在不同空间尺度上的变化, 提升对小目标物体的检测能力.

图 5 为多尺度特征融合小目标增强模块. 如图 5 所示, 本文首先将骨干网络中提取到的输出特征图C_3、C_4和C_5利用逐步降维法 (SDR) 进行归一化处理. 具体而言, 通过多个1×1卷积逐步降低通道数, 并结合 SiLU 非线性激活函数, 使得信息整合更加平滑, 同时逐步提取有效特征, 避免一次性大幅度降维导致的信息丢失. 然后, 采用最近邻插值法对归一化后的特征图进行上采样, 使得输出特征图$\hat{C}_3$、$\hat{C}_4$和$\hat{C}_5$的形状与特征图C_2的形状大小保持一致, 其过程如下所示:

$$\hat{C}_3 = \text{NN}(\text{CBS}_{1\times1}(C_3)), \tag{10}$$

$$\hat{C}_4 = \text{NN}(\text{CBS}_{1\times1}(\text{CBS}_{1\times1}(C_4))), \tag{11}$$

$$\hat{C}_5 = \text{NN}(\text{CBS}_{1\times1}(\text{CBS}_{1\times1}(\text{CBS}_{1\times1}(C_5)))). \tag{12}$$

将输出特征图C_2、$\hat{C}_3$、$\hat{C}_4$和$\hat{C}_5$在通道维度上进行 Concat 拼接, 形成多层次特征图, 并通过通道洗牌操作 (CS)[17] 重新排列通道顺序, 以增强不同层次特征的融合与交互. 随后, 使用1×1卷积将特征图的通道数由C降低至$C/4$, 以整合通道间的特征并提升计算效率, 得到输出特征图C_F为

$$C_F = \text{CBS}_{1\times1}(\text{CS}(\text{Concat}(C_2, \hat{C}_3, \hat{C}_4, \hat{C}_5))), \tag{13}$$

其中CS$(\cdot)$为通道洗牌操作.

为了进一步提取细致的特征并减少信息丢失, 使用3×3卷积来处理输出特征图C_F, 在保持高分辨率的同时提升特征表达能力. 这为 CPCA 通道先验卷积注意模块[18] 提供了更高质量的输入, 以便其能够更精确地计算通道和空间注意力. CPCA 模块结合通道与空间注意力, 实现了两者维度上的动态权重分配, 有效增强了对特征的提取能力.

3×3卷积模块和 CPCA 模块的组合能够使得模型更加有效地捕捉和处理无人机图像中的重要特征, 提高模型的检测能力. 其过程如下所示:

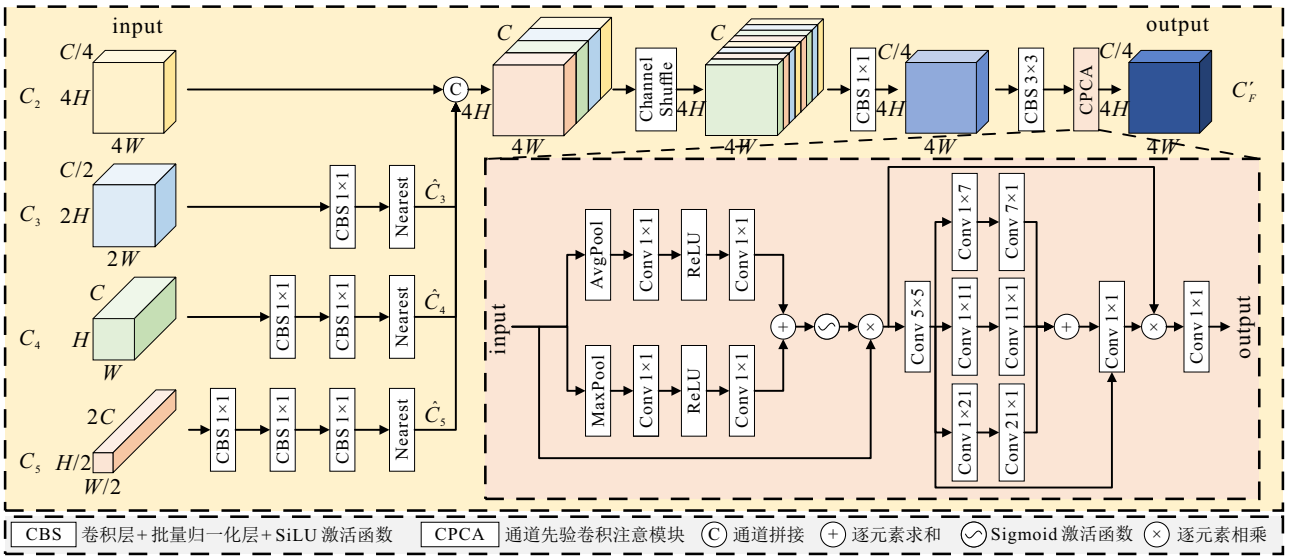


图5 多尺度特征融合小目标增强模块

$$C'_F = \mathrm{CPCA}(\mathrm{CBS}_{3\times3}(C_F)), \tag{14}$$

其中CPCA(·)为通道先验卷积注意模块. 最后, 将输出特征图C'_F与P_2检测层前的特征图进行逐元素相加, 从而增强对小目标的检测能力.

2　实验结果与分析

2.1　实验设置与评价指标

本文实验使用一张显存为 12 GB 的 NVIDIA RTX 3080Ti 显卡, 操作系统为 Ubuntu20.04, 显卡驱动版本 525.125.06, CUDA 版本 11.3. 实验中使用 Python 版本 3.8.18, Pytorch 版本 1.11. 实验将输入图像尺寸调整为640 × 640像素, 设置学习率初始值为 0.01, batchsize 为 4, 训练 epoch 为 300, 采用 SGD 随机梯度下降法优化模型训练.

为了全面评估目标检测模型的性能, 实验采用了 8 种评价指标: 精确率 Precision, 召回率 Recall, 调和平均数F_1-score, 平均精确率均值 mAP50% 和 mAP50-95%, 模型参数量 Params, 计算量 FLOPs, 以及每秒处理的帧数 FPS.

2.2　实验数据集

VisDrone2019 数据集[19] 是由 AISKYEYE 团队采集, 专为无人机视觉目标检测设计的. 该数据集包含 10209 张静态图像, 划分为训练集 6471 张, 验证集 548 张和测试集 3190 张, 共约 260 万个标注边界框, 涵盖 10 个目标类别. 由于官方评估服务器已关闭, 本文在验证集上来评估算法性能.

TinyPerson 数据集[20] 则是针对低分辨率小目标检测而设计的, 主要用于检测低分辨率的海上和陆地两类人类目标, 总计包含 1610 张图像和 72651 个标注对象. 在实验设置中, 将 1610 张图像按照 8 : 2 的比例划分为训练集 1288 张图像和验证集 322 张图像.

2.3　消融实验

2.3.1　小目标检测层P_2的有效性

表 1 和表 2 分别为在 VisDrone2019 数据集和 TinyPerson 数据集上消融实验的对比结果. 如表 1 和表 2 所示: 将P_2层加入 Baseline 模型后, 其在 VisDrone2019 数据集上的 mAP50% 和 mAP50-95% 分别提高了 4.9% 和 3.7%; 在 TinyPerson 数据集上分别提高了 3.5% 和 1.18%.

2.3.2　MSRFEM 的有效性

表 3 为 MSRFEM 中不同 DRB 分支组合的性能对比. 由表 3 中的对比实验结果可以看出, MSRFEM 中不同 DRB 分支组合的性能并不一定随着参数量或卷积核大小的增加而提高. 当 MSRFEM (K = 7, 15) 时, 模型达到了最佳的检测性能.

为了进一步分析 MSRFEM 在特征提取中的效果, 通过可视化对比展示了引入该模块前后热力图的差异, 如图 6 所示. 图 6(a) 为原始图像, 图 6(b) 为 Baseline +P_2对应的热力图, 图 6(c) 为 Baseline + P_2 集成 MSRFEM 后的热力图. 由图 6 对比结果表明,

表1　VisDrone2019 数据集上消融实验的对比结果

Baseline	P_2	MSRFEM	MDFAFM	MFFSTEM	Precision	Recall	F_1-score	mAP50%	mAP50-95%	Params/M	GLOPS/G	FPS
√					50.6	38.2	43.5	39.6	23.6	11.13	28.5	160.50
√	√				54.5	42.4	47.7	44.5	27.3	10.63	36.7	144.50
√	√	√			55.1	42.9	48.2	44.8	27.6	12.54	38.2	134.31
√	√		√		55.7	43.1	48.6	45.5	28.0	10.68	37.3	126.88
√	√			√	54.6	43.9	48.7	45.6	28.1	10.92	40.7	119.81
√	√	√	√		55.2	43.6	48.7	46.3	28.4	12.58	38.8	115.63
√	√	√		√	55.9	43.6	49.0	45.8	28.0	12.84	42.3	111.61
√	√		√	√	55.8	43.7	49.0	46.2	28.5	10.82	41.3	113.37
√	√	√	√	√	**56.3**	**44.6**	**49.8**	**46.8**	**28.8**	12.73	42.8	101.53

表2　TinyPerson 数据集上消融实验的对比结果

Baseline	P_2	MSRFEM	MDFAFM	MFFSTEM	Precision	Recall	F_1-score	mAP50%	mAP50-95%	Params/M	GLOPS/G	FPS
√					39.9	26.7	32.0	24.5	7.90	11.13	28.4	175.48
√	√				40.9	31.5	35.6	28.0	9.08	10.63	36.6	152.36
√	√	√			42.6	31.8	36.4	28.7	9.32	12.54	38.2	150.15
√	√		√		42.1	31.6	36.1	28.3	9.22	10.67	37.3	124.02
√	√			√	43.3	31.4	36.4	28.7	9.43	10.92	40.7	102.47
√	√	√	√		44.9	31.6	37.1	29.2	9.42	12.59	38.8	134.16
√	√	√		√	44.7	32.1	37.4	29.0	9.33	12.84	42.2	102.02
√	√		√	√	43.3	32.1	36.9	29.1	9.51	10.82	41.2	97.64
√	√	√	√	√	**45.0**	**32.5**	**37.7**	**30.2**	**9.73**	12.74	42.8	97.82

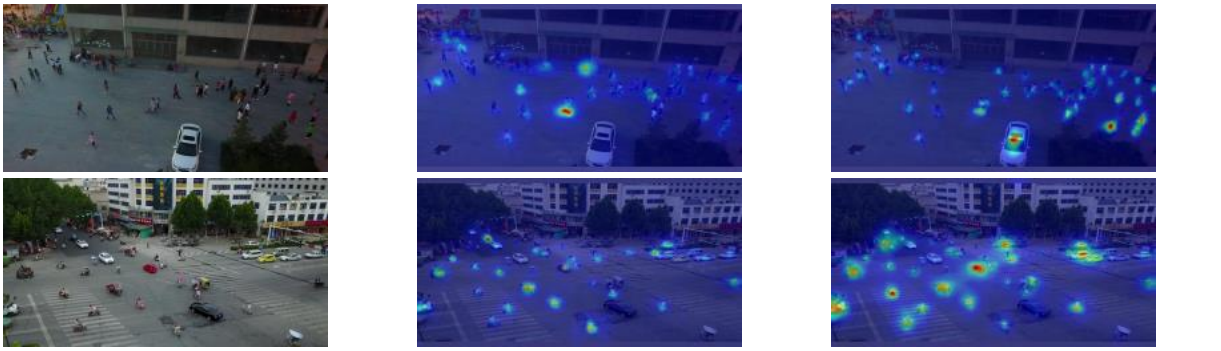
表3 MSRFEM 中不同 DRB 分支组合的性能对比

methods	Precision	Recall	F_1-score	mAP50%	mAP50-95%	Params/M
$K=7,13$	56.1	44.3	49.5	46.5	**28.8**	12.72
$K=7,15$	56.3	**44.6**	**49.8**	**46.8**	**28.8**	12.73
$K=7,17$	**57.7**	43.6	49.7	46.7	**28.8**	12.75
$K=9,13$	57.5	43.8	49.7	46.5	28.7	12.73
$K=9,15$	55.3	44.4	49.3	46.1	28.3	12.74
$K=9,17$	56.2	43.8	49.2	46.3	28.6	12.76
$K=11,13$	55.3	**44.6**	49.4	46.4	28.5	12.74
$K=11,15$	57.4	44.0	**49.8**	46.6	**28.8**	12.75
$K=11,17$	56.3	43.5	49.1	46.1	28.3	12.77

集成 MSRFEM 后的网络模型能够更准确地区分小目标物体与背景信息.

2.3.3 MDFAFM 的有效性

为了验证 MDFAFM 在特征金字塔网络中的有效性, 本节在 VisDrone2019 数据集上对第 1.3 节中图 3 所示的图 3(a) FPN + PAN, 图 3(b) MDFAFM + FPN + PAN 与图 3(c) MDFAFM + PAN 进行了性能对比. 实验结果如表 4 所示, 集成 MSRFEM 与 MFFSTEM 后的网络结构 (c) 在 F_1-score, mAP50% 和 mAP50-95% 指标上均优于集成后的网络结构 (a) 和 (b).



(a) 原始图像 (b) Baseline+P_2对应热力图 (c) Baseline+P_2集成MSRFEM后的热力图

图6 MSRFEM 集成前后的热力图效果对比

表4 MDFAFM 集成到特征金字塔网络的性能对比

methods	F_1-score	mAP50%	mAP50-95%	Params/M	GLOPS/G
(a)	49.0	45.8	28.0	12.84	42.3
(b)	49.7	46.3	28.4	13.06	44.1
ours(c)	**49.8**	**46.8**	**28.8**	12.73	42.8

2.3.4 MFFSTEM 的有效性

为了优化 MFFSTEM 在无人机图像中捕捉小目标物体在不同空间尺度上的能力, 本节对特征图 C_3、C_4、C_5进行了两种降维方式的对比实验: 一种是通过1×1卷积将通道维数降至$C/4$, 另一种是通过 SDR 方法将通道维数降至$C/4$. 实验结果如表 5 所示, 在 VisDrone2019 数据集上使用 SDR 方法后, F_1-score 和 mAP50% 分别提高了 0.3% 和 0.4%.

表5 MFFSTEM 中不同降维方式的性能对比

methods	F_1-score	mAP50%	mAP50-95%	Params/M	GLOPS/G
CBS1 × 1	49.5	46.4	28.8	12.70	42.7
ours	**49.8**	**46.8**	**28.8**	12.73	42.8

2.3.5 各模块协同工作的有效性

为了验证所提出 3 个模块协同工作的有效性和优越性, 在 Baseline + P_2模型的基础上进行了单模块、双模块和三模块的对比实验, 实验结果如表 1 和表 2 所示. 随着模块的逐步引入, 模型在两个数据集上的F_1-score、mAP50% 和 mAP50-95% 均实现了提升. 最终, 集成 3 个模块的模型 ours 与基线模型相比, 在 VisDrone2019 数据集上的F_1-score、mAP50% 和 mAP50-95% 分别提高了 6.3%、7.2% 和 5.2%; 在 TinyPerson 数据集上分别提高了 5.7%、5.7% 和 1.83%. 上述结果验证了所提出模块在无人机图像目标检测中的有效性, 以及三者在提升小目标检测性能和应对目标尺度变化方面的良好协同性.

2.4 对比实验

2.4.1 VisDrone2019 数据集上的对比结果

表 6 为所提出算法与其他主流目标检测算法在 VisDrone2019 验证集上的性能对比结果. 在计算成本相对较低的情况下, 所提出算法在 Recall、F_1-score、mAP50% 和 mAP50-95% 指标上分别达到了 44.6%、49.8%、46.8% 和 28.8% 的成绩, 均优于其他主流算法. 但是, 由于所提出 3 个模块均涉及到多尺度结构, 这种全面的特征提取可能会增加模型的误检风险, 从而限制 Precision 的提升. 由表 6 实验结果可见, 所提出算法的 Precision 比 TPH-YOLOv5 模型低了 1.7%, 但是, 在 Recall 和F_1-score 上分别高出了 1.9% 和 0.6%.

本文在 12 GB 显存的 NVIDIA RTX 3080 Ti 显卡上测试, 检测速度达 101.53 FPS. 同时, 在算力较低但是同为 12 GB 显存的 NVIDIA RTX 1080 Ti 显

表6　不同目标检测方法在 VisDrone2019 数据集上的对比结果

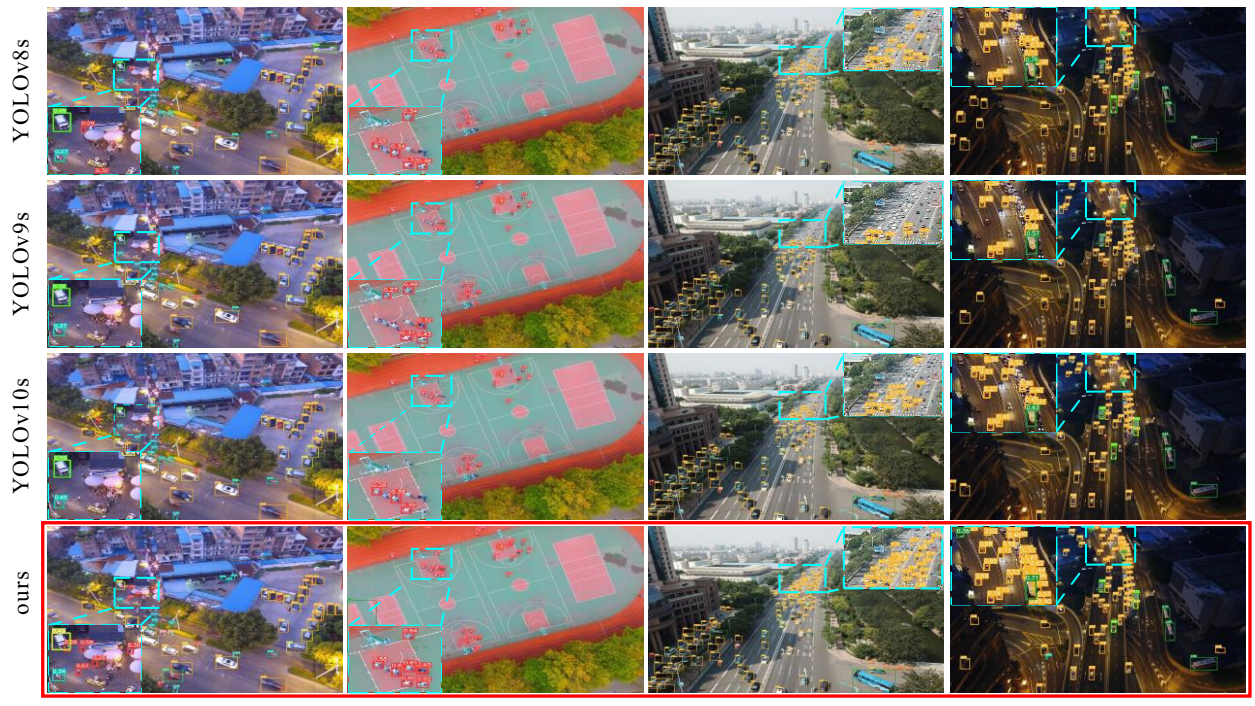
methods	Precision	Recall	F_1-score	pedestrian	people	mAP50%	mAP50-95%	Params/M	GLOPS/G	FPS
Faster R-CNN[5]	–	–	–	–	–	36.3	20.6	41.0	207.0	–
Centernet[21]	–	–	–	–	–	29.0	14.0	31.8	206.0	–
EfficientDet[22]	–	–	–	–	–	21.2	12.9	6.6	6.1	–
YOLOv5-M	49.0	37.0	42.2	45.0	36.3	37.0	20.9	20.89	48.0	109.89
TPH-YOLOv5[10]	**58.0**	42.7	49.2	54.2	42.8	45.5	27.0	60.43	145.7	33.44
EdgeYOLO[23]	–	–	–	–	–	44.8	26.4	40.5	–	–
YOLOv8-S	50.6	38.2	43.5	43.5	33.1	39.6	23.6	11.13	28.5	160.50
YOLOv8-M	54.6	41.3	47.0	47.8	36.2	43.2	26.4	25.84	78.7	99.02
Drone-YOLO-S[13]	–	–	–	–	–	44.3	27.0	10.9	–	–
YOLOv9-S[24]	54.3	39.6	45.8	43.6	36.9	41.8	25.7	9.79	39.8	67.11
YOLOv9-M[24]	55.5	42.9	48.4	47.6	36.8	45.2	27.9	32.57	130.8	63.29
YOLOv10-S[25]	50.5	38.3	43.6	42.1	33.0	39.1	23.5	8.04	24.5	181.82
YOLOv10-M[25]	54.2	40.2	46.3	45.1	35.0	42.2	25.8	16.46	63.5	149.25
PS-YOLO-M[11]	50.2	38.9	43.8	–	–	37.6	22.3	42.2	88.0	–
YOLO-HV[14]	48.0	38.8	42.9	–	–	38.1	19.9	38.5	111.9	–
ours	56.3	**44.6**	**49.8**	**54.3**	**44.1**	**46.8**	**28.8**	12.73	42.8	101.53

卡上测试, 其检测速度为 61.27 FPS. 两种平台下的检测速度均超过了 30 FPS[26] 的实时处理标准, 验证了模型的 FPS 与硬件性能有关.

为了直观地展示所提出算法的性能优势, 将其与现有算法 YOLOv8s、YOLOv9s、YOLOv10s 进行了对比分析, 如图 7 所示. 图 7 中各目标类别采用不同颜色标注, 关键性能差异放大展示, 并通过 4 种典型场景直观地呈现不同算法的检测表现: 图 7(a) 为复杂环境检测: 在复杂背景下, 所提出算法准确检测出遮阳棚下的人员; 图 7(b) 为密集人群检测: 在小规模密集人群中, 准确检测出操场上的目标人员; 图 7(c) 为尺度变化目标检测: 对于远处小尺寸车辆的检测, 所提出算法显著优于其他算法; 图 7(d) 为低光照环境检测: 在低光照条件下, 能够有效识别更多车辆. 上述结果表明, 所提出算法在处理各种复杂检测场景中具有明显优势.

2.4.2　TinyPerson 数据集上的对比结果

表 7 比较了所提出算法与其他主流目标检测算法在 TinyPerson 验证集上的性能. 实验结果表明, 所提出算法在该数据集上取得了领先的检测效果. 同



(a) 复杂环境检测　(b) 密集人群检测　(c) 尺度变化目标检测　(d) 低光照环境检测

图7　VisDrone2019 数据集中 4 种典型无人机捕获场景在不同方法下的视觉检测结果

表7 不同目标检测方法在 TinyPerson 数据集上的对比结果

methods	Precision	Recall	F_1-score	pedestrian	people	mAP50%	mAP50-95%	Params/M	GLOPS/G	FPS
YOLOv5-S	42.5	28.0	33.8	29.1	21.0	25.0	7.43	7.02	15.8	163.93
YOLOv5-M	**46.1**	29.2	35.8	30.9	23.1	27.0	8.18	20.86	47.9	120.48
TPH-YOLOv5[10]	45.2	30.2	36.2	30.5	23.9	27.2	7.94	60.36	145.3	44.64
YOLOv8-S	39.9	26.7	32.0	27.9	21.0	24.5	7.90	11.13	28.4	175.48
YOLOv8-M	41.1	26.6	32.3	28.6	22.0	25.3	8.33	25.84	78.7	109.88
YOLOv9-S[24]	35.9	25.5	29.8	24.7	18.9	21.8	7.23	9.78	39.8	62.72
YOLOv9-M[24]	39.6	24.9	30.6	24.9	19.9	22.4	7.34	32.74	131.7	58.82
YOLOv10-S[25]	40.0	26.1	31.6	25.0	19.3	22.2	6.95	8.04	24.4	149.25
YOLOv10-M[25]	42.5	25.7	32.0	25.3	19.6	22.5	7.05	16.45	63.4	125.00
ours	45.0	**32.5**	**37.7**	**32.7**	**27.8**	**30.2**	**9.73**	12.74	42.8	97.82

时, “sea person” 和 “earth person” 两类目标的 AP50% 分别达到了 32.7% 和 27.8%, 也均优于其他主流算法.

此外, 在 12 GB 显存的 NVIDIA RTX 3080 Ti 和 NVIDIA RTX 1080 Ti 显卡上进行了测试. 测试结果显示, 其检测速度分别达到了 97.82 FPS 和 59.61 FPS, 也均高于 30 FPS 的实时处理性能标准.

3 结 论

针对无人机图像中小目标比例高、分布密集以及目标尺度差异较大的问题, 本文提出了一种基于多尺度融合和高分辨特征增强的无人机航拍目标检测算法. 首先, 引入了小目标检测层P_2, 以提升对小目标的检测效果; 然后, 构建了多尺度结构重参数化特征提取模块, 以增强骨干网络对小目标信息的提取能力; 接着, 利用多维特征自适应融合模块来提高多尺度特征间的交互和整合能力; 最后, 通过多尺度特征融合小目标增强模块, 进一步提高对小目标的检测能力. 通过在两个公开数据集上进行大量的实验验证了所提出方法的有效性和优越性. 在未来工作中, 计划对网络模型进行剪枝和知识蒸馏等方面的研究, 以减少模型的参数和计算成本, 从而使其更适合部署于资源受限的应用场景中.

参考文献 (References)

[1] Marques T, Carreira S, Miragaia R, et al. Applying deep learning to real-time UAV-based forest monitoring: Leveraging multi-sensor imagery for improved results[J]. Expert Systems with Applications, 2024, 245: 123107.

[2] Feng H L, Li Q, Wang W, et al. Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework[J]. Information Fusion, 2024, 112: 102555.

[3] Zhang K, Zhou R H, Wang J C, et al. Transmission line component defect detection based on UAV patrol images: A self-supervised HC-ViT method[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024, 54(11): 6510-6521.

[4] Tahir N U A, Long Z, Zhang Z P, et al. PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8[J]. Drones, 2024, 8(3): 84.

[5] 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.

[6] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J/OL]. 2018, arXiv: 1804.02767.

[7] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot MultiBox detector[C]. Computer Vision — ECCV 2016: Proceedings of the 14th European Conference. Amsterdam, 2016: 21-37.

[8] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[J/OL]. 2021, arXiv: 2010.04159.

[9] 赵亮, 刘世鹏. 全局与局部图像特征自适应融合的小目标检测算法[J]. 控制与决策, 2023, 38(4): 935-943.
(Zhao L, Liu S P. Small object detection algorithm based on adaptive fusion of global and local image features[J]. Control and Decision, 2023, 38(4): 935-943.)

[10] Zhu X K, Lyu S C, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Montreal, 2021: 2778-2788.

[11] Peng S F, Fan X, Tian S W, et al. PS-YOLO: A small object detector based on efficient convolution and multi-scale feature fusion[J]. Multimedia Systems, 2024, 30(5): 241.

[12] 周葳楠, 吴治海, 张正道, 等. 基于弱特征增强的轻量化小目标检测方法[J]. 控制与决策, 2024, 39(2): 381-390.
(Zhou W N, Wu Z H, Zhang Z D, et al. Lightweight small target detection method based on weak feature enhancement[J]. Control and Decision, 2024, 39(2): 381-390.)

[13] Zhang Z X. Drone-YOLO: An efficient neural network

method for target detection in drone images[J]. Drones, 2023, 7(8): 526.

[14] Xu S Z, Zhang M J, Chen J Y, et al. YOLO-HyperVision: A vision transformer backbone-based enhancement of YOLOv5 for detection of dynamic traffic information[J]. Egyptian Informatics Journal, 2024, 27: 100523.

[15] Ding X H, Zhang Y Y, Ge Y X, et al. UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 5513-5524.

[16] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[17] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6848-6856.

[18] Huang H J, Chen Z G, Zou Y, et al. Channel prior convolutional attention for medical image segmentation[J]. Computers in Biology and Medicine, 2024, 178: 108784.

[19] Du D W, Zhu P F, Wen L Y, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, 2019: 213-226.

[20] Yu X H, Gong Y Q, Jiang N, et al. Scale match for tiny person detection[C]. IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, 2020: 1257-1265.

[21] Duan K W, Bai S, Xie L X, et al. CenterNet: Keypoint triplets for object detection[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 6569-6578.

[22] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10781-10790.

[23] Liang S Y, Wu H, Zhen L, et al. Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 25345-25360.

[24] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[J/OL]. 2024, arXiv: 2402.13616.

[25] Wang A, Chen H, Liu L H, et al. YOLOv10: Real-time end-to-end object detection[J/OL]. 2024, arXiv: 2405.14458.

[26] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv4: Scaling cross stage partial network[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13029-13038.

作者简介

陈志旺 (1978−), 男, 副教授, 博士, 主要研究方向为运动物体目标检测与跟踪、多旋翼飞行器导航及控制, E-mail: czwaaron@ysu.edu.cn;

肖迪创 (2000−), 男, 硕士, 主要研究方向为计算机视觉中的目标检测, E-mail: 2797736860@qq.com;

吕昌昊 (1996−), 男, 硕士, 主要研究方向为智能电网的优化和控制, E-mail: 316998054@qq.com;

李思哲 (1997−), 男, 硕士, 主要研究方向为计算机视觉中的目标跟踪, E-mail: ysu15lsz@163.com;

彭勇 (1963−), 男, 教授, 博士, 博士生导师, 主要研究方向为生物机器人控制和计算机视觉中的目标检测, E-mail: PY81@sina.com.