doi:10.3772/j.issn.1002 - 0470.2026.08.007

基于多尺度融合的无人机航拍小目标检测研究①

陈志旺②*** 王 婷*** 吕昌昊*** 张汉锋*** 彭 勇****

(*燕山大学 智能控制系统与智能装备教育部工程研究中心 秦皇岛 066004)

(**燕山大学 河北省工业计算机控制工程重点实验室 秦皇岛 066004)

(***燕山大学 燕山大学电力电子节能与传动控制河北省重点实验室 秦皇岛 066004)

(****燕山大学 电气工程学院 秦皇岛 066004)

摘 要 针对无人机航拍图像中小目标检测精度较低的问题,本文提出一种改进的小目标检测算法。首先,该算法在主干网络中引入可变形卷积,通过自适应调整卷积操作中的采样位置,提升对形态不规则小目标的特征提取能力;其次,在颈部网络增加一个小目标检测层,保留更多浅层细节特征,加强不同尺度特征层之间的信息融合;最后,使用基于最小点距离的损失函数(minimum point distance based IoU,MPDIoU)改进边界回归损失函数。实验部分,本文在 VisDrone 2019 数据集和 MS COCO 数据集上进行了全面验证。实验结果表明,改进后的算法在 VisDrone 2019 数据集上的 mAP@0.50 达到了 39.8%,较原算法提升了 2.3%;在 MS COCO 数据集的小目标检测平均精度均值(mean average precision,mAP)为 22.5%,提升了 2.8%。此外,本文还对实际场景图像和 VisDrone 2019 数据集的典型图像进行测试与可视化,进一步验证了所提算法的有效性。

关键词 无人机航拍图像; 小目标检测; 多尺度融合; 可变形卷积; 最小点距离的损失函数; 可视化分析

近年来,无人机航拍结合目标检测技术在农业监测[1]、灾情检测[2]以及城市规划[3]等领域得到广泛应用。无人机拍摄视角的独特性导致航拍图像中的目标尺度差异大,包含大量分布密集的小目标,这些目标易受形变、光线变化等因素影响[4]。

目标检测领域中,传统目标检测方法的效率较低[5],基于深度学习的目标检测算法成为当前研究的主流。基于深度学习的目标检测算法主要分为二阶段和一阶段算法 2 类。以区域卷积神经网络(region-convolutional neural network,R-CNN)[6]和快速的区域卷积神经网络(faster-region-convolutional neural net-work,Faster-R-CNN)[7]为代表的二阶段算法实时性较差;以只看一次(you only look once,YOLO)[8]算法为代表的单阶段算法应运而生,单阶段算法直接将图像像素映射到边界框坐标和类别概率,提高了检测速度。YOLO 系列算法不断更新迭代,YOLOv7[9]设计高效层聚合网络(efficient layer aggregation networks,ELAN),通过调整模型的深度和宽度拓展出不同规模的模型,在多种目标检测任务中实现精度和速度的平衡。YOLOv8[10]将分类和回归任务解耦,进一步提高目标检测的效果。YOLOv9[11]引入可编程梯度信息(programmable gradient information,PGI),辅助可逆分支生成可靠梯度。YOLOv10[12]通过无非极大值抑制(non-maximum suppression,NMS)训练、高效架构设计及自注意力融合等改进,实现了目标检测的精度与效率突

① 国家自然科学基金(61573305),河北省自然科学基金(F2022203038,F2019203511)和省级重点实验室绩效补助经费(22567612H)资助项目。

② 男,1978 年生,博士,副教授;研究方向:多旋翼飞行控制,计算机视觉;联系人,E-mail: czwaaron@ysu.edu.cn。

(收稿日期:2025 - 01 - 20)

破。YOLOv11[13]将 YOLOv8 中的 C2f 模块改进为 C3K2 模块,并设计了 C2PSA 模块,减低参数量的同时提高了检测精度。最新版本的 YOLOv12[14]将以注意力为核心的设计引入 YOLO 框架,实现了精度和效率的显著提升。

无人机航拍图像中通常包含高密度分布的中小尺度目标,使得检测任务面临着显著挑战。Yu 等人[15]提出了一种实时监测 Transformer(real-time detection Transformer,RT-DETR)算法,增加上下文引导下采样模块,用于无人机图像中的精确目标检测。伍智华等人[16]重构特征融合网络,采用跨层级联的方式实现深层信息和浅层信息的融合,提高小目标特征融合能力,但此算法仅针对人体姿态识别实验,并未在航拍图像中实验。Du 等人[17]使用稀疏卷积减少计算量,但算法的精度有所降低。Wang 等人[18]提出的 Gold-YOLO 收集分发机制以增强多尺度特征融合,避免模型在传输过程中丢失特征信息。综上所述,提升小目标的检测效果的改进方法可以从改善卷积方式和促进融合多尺度信息入手。

为提高无人机航拍图像中小目标检测精度,本文首先在模型主干网络部分中引入第 2 代可变形卷积,形成可以自适应感受野形状的 ELAN-d 结构,ELAN-d 为本文自行设计的引入可变形卷积(deformable convolution,DCN)的高效层聚合网络(efficient layer aggregation networks, ELAN),命名为 ELAN-d,其含义为可变形高效层聚合网络,以增强模型的特征提取能力;其次,增加一个小目标检测层,提升多尺度融合能力;最后,使用基于最小点距离的损失函数改进边界框回归损失函数,以更加准确地检测目标。

1 本文算法的网络结构介绍

本文的网络模型分为输入端(Input)、主干网络(Backbone)、颈部网络(Neck)和预测网络(Prediction)4 部分,图中使用加粗圆角矩形框重点标注网络中的改进部分。

如图 1 所示,输入端对输入的图片进行预处理操作后送入主干网络。主干网络使用可变形卷积模块替换 ELAN 模块中的部分标准卷积模块,对检测目标进行不同层级的特征提取,将改进后的结构命名为 ELAN-d。颈部网络使用空间金字塔池化-跨阶段局部网络(spatial pyramid pooling cross stage partial,SPPCSP)模块,结合特征金字塔网络(feature pyramid network,FPN)[19]和路径聚合网络(path aggregation network,PAN)[20]实现不同层级的特征信息融合,本文在颈部网络原有 80 × 80、40 × 40、20 × 20 这 3 个不同像素尺度的特征层基础上增加 160 × 160 的小目标特征检测层,进一步保留小目标的细节信息特征。在预测网络中,根据颈部网络生成的 4 个有效特征层进行分类回归预测,生成预测框和目标类别,本文引入基于最小点距离的损失函数(minimum point distance based IoU,MPDIoU)改进边界框回归损失函数,有效提高检测精度。

1.1 可变形卷积

卷积神经网络(convolutional neural network,CNN)在目标检测领域应用十分广泛,通过卷积模块提取图像的目标特征以进行分类和定位。标准卷积使用固定尺寸的 $N \times N$ 矩形卷积核对特征图进行特征提取。卷积核的形状和大小固定,感受野和采样位置也因此固定,导致卷积操作缺乏灵活性,提取特征能力不足。实际应用中的待检测图片中的目标形态复杂,绝大多数目标的形状不是简单的矩形,此外光照条件、拍摄角度以及无关背景干扰等因素也会造成检测目标的形态发生变化。

为改善标准卷积对检测物体形状适应能力不足的问题,Dai 等人[21]提出了可变形卷积(deformable convolution,DCN),可变形卷积通过在采样点增加可学习的偏移量,改进了传统标准卷积的卷积核。如图 2 所示,可变形卷积使得网格模型根据待检测物体自适应调整卷积核的尺度和形状,从而突破了固定大小矩形感受野的限制。这种改进不仅能更准确地贴近不规则形状物体的边界,还能更好地适应目标物体因形变或遮挡导致的外形变化,从而提升特征提取能力。

规则形状的网格定义感受野的大小,如图 3 所示,可以用网格定义一个的 3 × 3 卷积核的采样网格,也就是卷积核的覆盖面积:

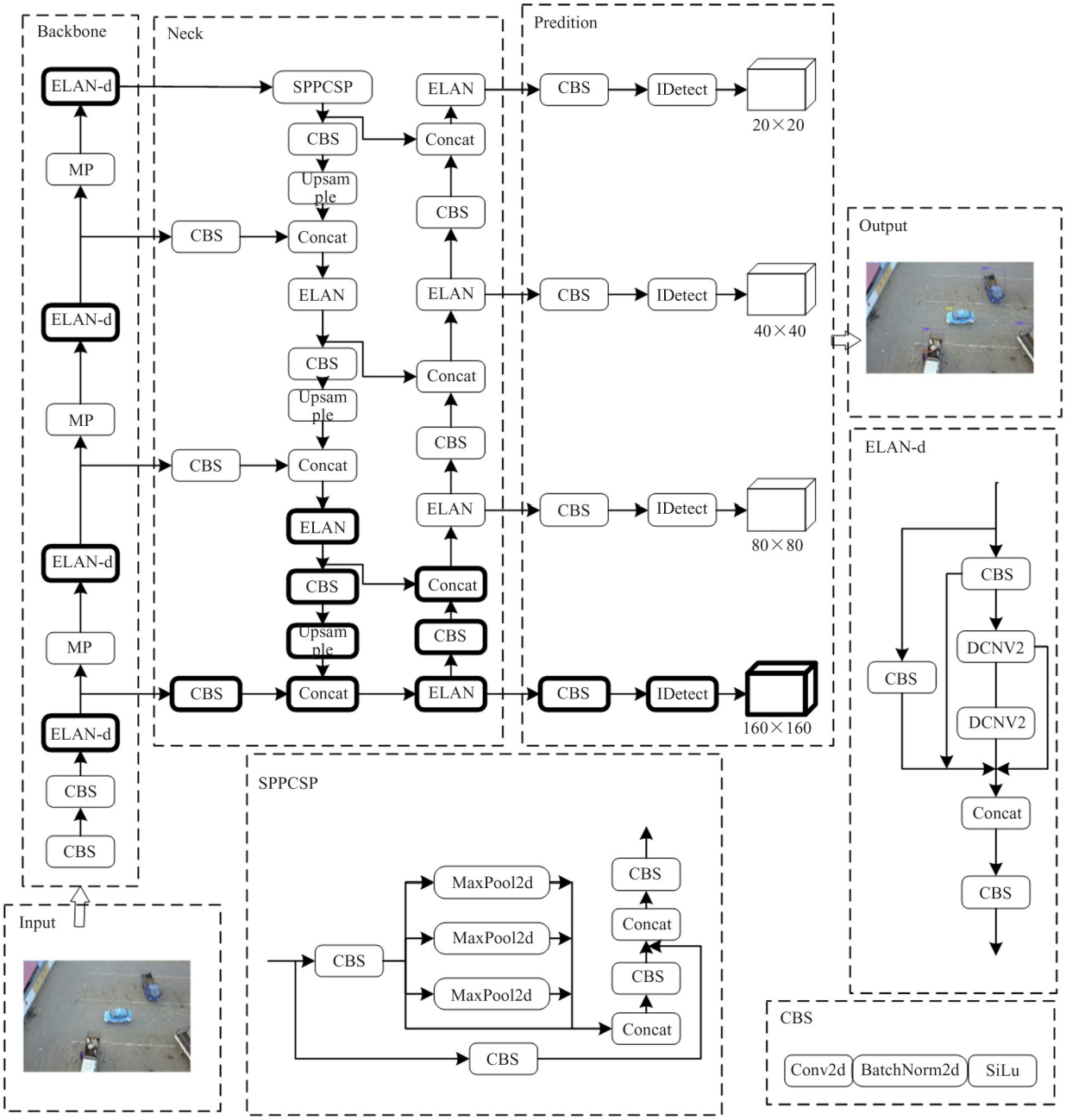


图 1　改进网络结构图

$$R=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \tag{1}$$

式中:R 中的每个元素为卷积核中所有位置相对于卷积核的中心位置的偏移量,其中(0,0)对应卷积核的中心位置。

标准卷积使用固有的矩形网格,按照一定步长在输入特征图上滑动采样,网格的每一个采样点与标准卷积的卷积核对应位置权重先相乘再累加求和,随后将结果保存到输出特征图对应位置。对于采样网格中心点 p_0,标准卷积的特征图输出为

$$y(p_0)=\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n) \tag{2}$$

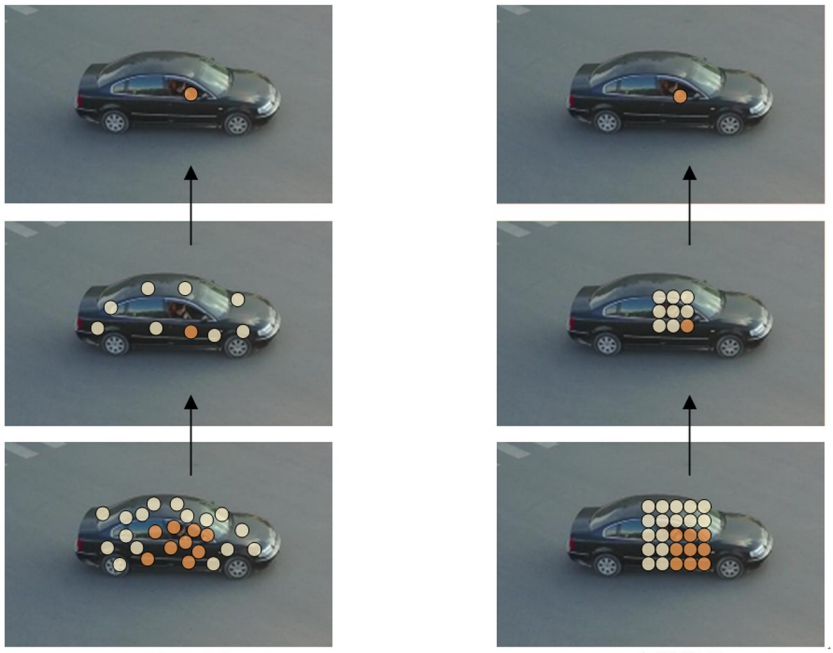
式中:x 为输入特征图,y 为输出特征图,p_0 为采样网格的中心点,p_n 为 R 内每个位置的枚举,$x(\cdot)$ 为输入特征图不同位置的像素值,$w(\cdot)$ 为卷积核与采样点对应位置的权重。

可变形卷积 DCN 引入了偏移量 $\{\Delta p_n \mid n=1,\cdots,N\}$,其中 $N=|R|$,可变形卷积的计算公式如

式(3)所示。

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad (3)$$

式中：Δp_n 为可以学习的偏移量，通常情况下是浮点数，偏移后的对应位置的像素值在输入特征图上可能不是一个存在的点，需要通过双线性插值法获得[22]。图 4 为标准卷积与可变形卷积采样示意图。

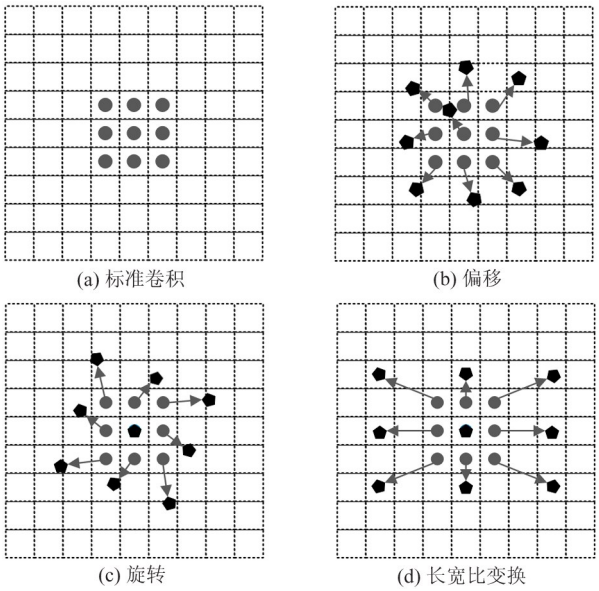


(a) 可变形卷积　　(b) 标准卷积

图 2　可变形卷积与标准卷积作用对比图

(−1,−1)	(0,−1)	(1,−1)
(−1,0)	(0,0)	(1,0)
(−1,1)	(0,1)	(1,1)

图 3　采样网格相对位置示意图



(a) 标准卷积　(b) 偏移　(c) 旋转　(d) 长宽比变换

图 4　标准卷积与可变形卷积采样示意图

如图 4 所示，可变形卷积通过引入可学习的偏移量，突破标准卷积的固定采样形状限制，图 4(a)为标准卷积操作，圆点为标准卷积采样点；图 4(b)为可变形卷积操作，箭头为偏移向量，五边形的点为偏移后的采样点；图 4(c)和图 4(d)为图 4(b)可变形卷积的 2 个特例，体现可变形卷积适应采样形状发生旋转和长宽比变换。

第1代可变形卷积在对目标进行检测时，存在无关背景信息干扰的问题，于是 Zhu 等人[23]提出了第 2 代可变形卷积(deformable convolutional networks v2，DCNv2)，在第 1 代可变形卷积的基础上增加惩罚项，能够更准确地集中于检测图像的感兴趣区域，

从而减少无关背景的干扰。DCNv2 不仅可以自适应地调整卷积核的偏移量,还增强了偏移的调制能力,使得模型能够对输入特征图不同采样点的自适应偏移的学习幅度进行调节,提升了网络模型的建模功能。DCNv2 的计算公式如式(4)所示。

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_n \tag{4}$$

式中:Δm_n 代表调制标量,Δm_n 的范围为[0,1]。

本算法使用可变形卷积 DCNv2 替换主干网络中的高效聚合层网络 ELAN 中的 3×3 标准卷积模块,改进后的结构称为 ELAN-d,以提高网络模型的特征提取能力,更好地关注不规则形状目标的特征,提高目标检测精度。

1.2 小目标检测层

浅层特征图的感受野较小,分辨率较高,包含丰富的边缘和纹理细节,能够提取到小目标的几何特征信息;相对而言,经过多层神经网络的特征提取之后,深层特征图感受野扩大,语义信息丰富,但分辨率较低,可能会丢失许多小目标的特征,导致几何信息表征能力不足[24]。

本文算法通过特征金字塔网络与路径聚合网络的协同设计实现跨尺度融合。首先,在主干网络中自底向上通过卷积与池化操作将浅层特征图逐步转化为深层特征图,降低空间分辨率的同时扩大感受野,提取到更多的语义特征,但在此过程中,深层特征图可能丢失小目标信息。于是根据特征金字塔网络原理,在颈部网络中,自顶向下通过上采样恢复深层特征的空间分辨率,将深层语义特征信息传达至高分辨率特征图,同时利用横向连接保留浅层的边缘和纹理信息。路径聚合网络作为特征金字塔网络的补充,进一步优化特征传播,再次自底向上通过卷积而非池化进行下采样,以保留更多细节特征和空间定位特征并传达至深层特征图,减轻深层特征的细节丢失问题,提升对小目标的定位精度。此阶段同样使用横向连接融合深层特征和浅层特征信息,使不同尺度的目标都能获得高质量的特征表示。图 5为算法多尺度融合示意图,图中将输出特征图尺寸相同的连续网络层的集合归为一个阶段,将主干网络中每阶段的最后一层输出的特征图记为 Ci,i 的取值为输出特征图大小为原始输入图片大小的 $1/2^i$ 倍。颈部网络中的 Ni 和 Pi 的定义方法与 Ci 相似,因此图中 C5 与 N5、N2 与 P2 并不代表同一层输出,而是不同阶段的最后一层的输出特征图。

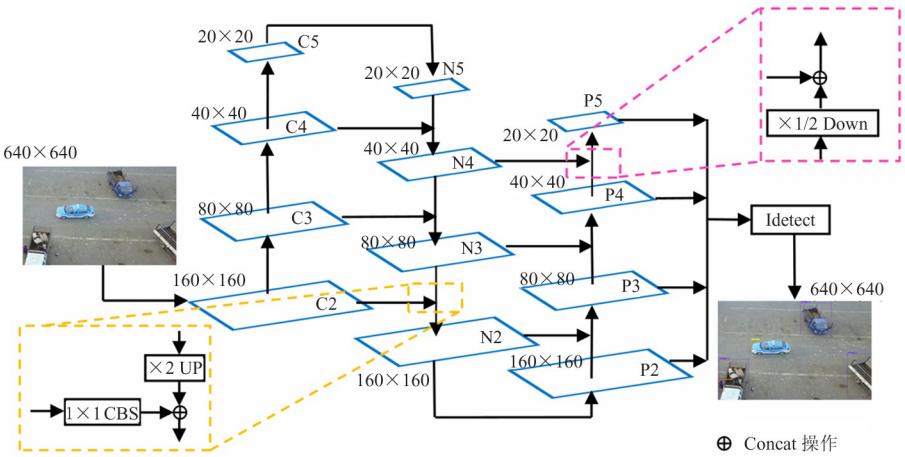


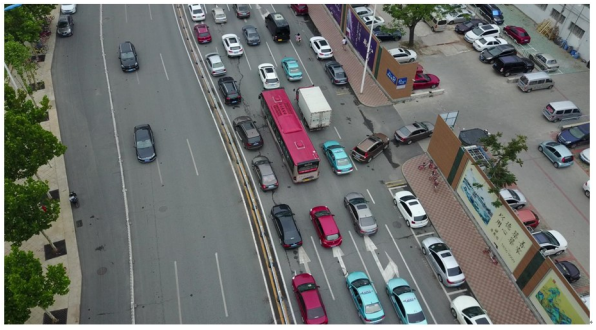
图 5 多尺度特征融合

本文算法的具体的多尺度融合过程如图 5 所示,输入的图片在主干网络中经过 4 倍、8 倍、16 倍、32 倍下采样后得到标准像素大小为 160×160、80×80、40×40 和 20×20 4 个不同尺度的输出特征图{C2,C3,C4,C5},其中 160×160 像素的浅层特征图 C2 分辨率高,可以保留较多小目标的细节特征信息;20×20 像素的深层特征图 C5 的语义特征信息更加丰富,但可能在多层特征提取后遗失小目标的特征信息。在颈部网络,深层特征图逐层上采样,将深层的语义信息传递到浅层特征图,得到 20×20、40×40、80×80 和 60×160 标准像素大小的特征图{N5,N4,N3,N2};随后自底向上将浅层特

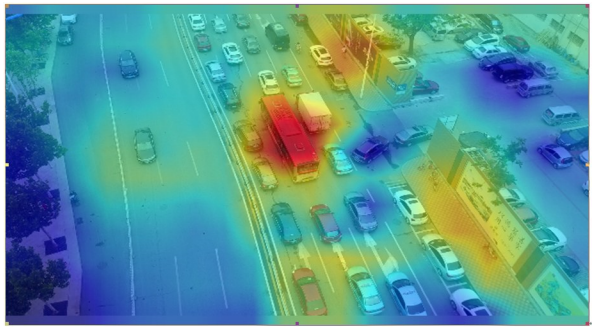
征图包含的丰富细节信息和定位信息向深层特征图传递,增强特征的定位表达能力。算法采用横向连接兼顾深层特征图和浅层特征图中的语义分类信息和细节几何信息,最终实现对特征的多尺度融合,得到{P2,P3,P4,P5}这 4 个不同尺度的目标检测层。其中,本算法新增加的 160×160 像素大小的检测层保留更多小目标的细节信息,提高网络对航拍图片中大量小目标的检测性能,进而提升整体目标检测效果。

可视化工具梯度加权类激活映射[25](gradient-weighted class activation mapping,Grad-CAM)通过生成热力图来可视化基于深度学习的卷积神经网络模型预测时关注的图像区域。

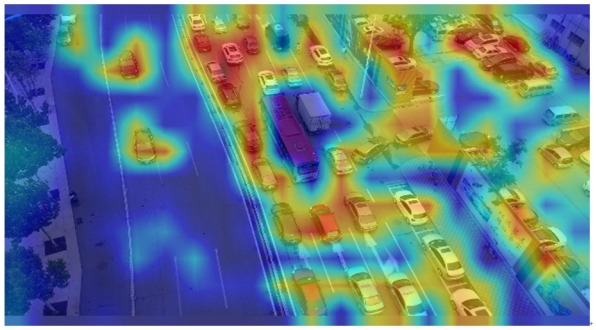
通过对比算法生成的热力图关注焦点分布差异,评估算法网络的合理性。图 6(a)是典型的无人机航拍图像,图中包含大量密集分布的小目标。图 6(b)展示了未增加小目标检测层的算法生成的热力图,结果显示模型主要关注图像中心的公交车,忽略了车道上的小目标。图 6(c)展示了增加小目标检测层后的算法生成的热力图,与图 6(b)相比,改进后的算法在检测中保持对小目标的关注,更好地识别了小目标,大多数小目标的热度值较高。



(a) 原图



(b) base



(c) base + 小目标检测层

图 6　热力图对比

1.3　MPD 损失函数

损失函数通过对比网络模型的检测结果与实际情况的差异来评估模型效果,是目标检测算法的重要组成部分。本文算法中的损失函数由边界框回归损失、目标置信度损失和分类损失 3 部分组成。其中边界框回归损失为 L_{CIoU},计算公式如式(5)所示,使用的完全交并比(complete intersection over union,CIoU)损失函数既考虑了预测框与真实框之间的交并比(图 7),又考虑到两者中心间的距离以及宽高纵横比,计算公式如下:

$$L_{\mathrm{CIoU}} = 1 - \mathrm{CIoU} \tag{5}$$

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2(b,b^{gt})}{c^2} - \alpha v \tag{6}$$

$$\mathrm{IoU} = \frac{B^{\mathrm{gt}} \cap B}{B^{\mathrm{gt}} \cup B} \tag{7}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{8}$$

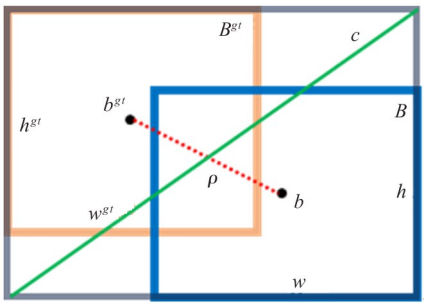


图 7　CIoU 边界框优化关系图

$$\alpha = \frac{v}{1 - \mathrm{IoU} + v} \tag{9}$$

式中:B 代表预测框,B^{gt} 代表真实框,b 代表预测框中心点,b^{gt} 代表真实框中心点,ρ 为预测框和真实框中心点的距离,c 代表预测框与真实框最小外接

矩形的对角线距离,α 是权重参数,v 表示宽高比一致性,w 和 h 代表预测框的宽和高,w^{gt} 和 h^{gt} 代表真实框的宽和高。

针对边界框回归损失,包括 CIoU 在内的大多数方法未考虑图像的尺寸,在预测框和真实框宽高纵横比一致,但实际宽度值和高度值不同的情况下表现不理想,所以本文使用基于最小点距离的损失函数[26]替代 CIoU 用于网络模型的边界框回归损失(图 8)。MPDIoU 的计算公式如下:

$$\mathrm{MPDIoU} = \mathrm{IoU} - \frac{d_1^2}{W^2 + H^2} - \frac{d_2^2}{W^2 + H^2} \tag{10}$$

$$d_1^2 = (x_1 - x_1^{gt})^2 + (y_1 - y_1^{gt})^2 \tag{11}$$

$$d_2^2 = (x_2 - x_2^{gt})^2 + (y_2 - y_2^{gt})^2 \tag{12}$$

$$L_{\mathrm{MPDIoU}} = 1 - \mathrm{MPDIoU} \tag{13}$$

式中:W 和 H 代表二维输入图片的宽和高,x_1^{gt} 代表真实框左上角水平方向坐标,x_1 代表预测框左上角水平方向坐标,y_1^{gt} 代表真实框左上角竖直方向坐标,y_1 代表预测框左上角竖直方向坐标,x_2^{gt} 代表真实框右下角水平方向坐标,x_2 代表预测框右下角水平方向坐标,y_2^{gt} 代表真实框右下角竖直方向坐标,y_2 代表预测框右下角竖直方向坐标。

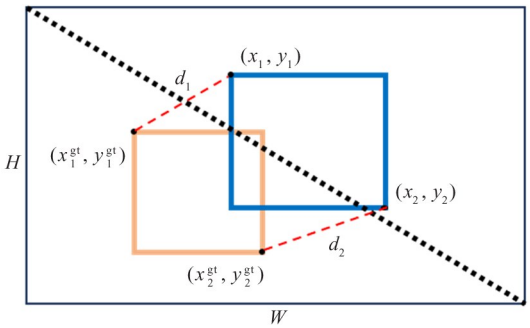


图 8　MPDIoU 边界框优化关系图

MPDIoU不仅考虑到现有损失函数考虑到的预测框与真实框的交并比、中心点距离、宽高纵横比等的相关因素,还考虑到当预测框和真实框宽高比一致,实际宽度和高度值不同的情况。

2　实验及结果分析

2.1　实验环境及参数配置

本文使用的硬件设备配置如下:CPU 的型号为 Intel Core i7-12700KF,GPU 的型号为 NVIDIA GeForce RTX 4080,显存为 16 GB。操作系统为 Ubuntu 20.04 操作系统,Python 版本为 3.10.13,Pytorch 版本为 2.0.1,CUDA 版本为 12.3。所有模型训练过程中均未使用预训练权重;在图像预处理阶段,采用图片自适应缩放,将训练图像调整为固定大小的 640×640 标准像素尺寸。

本文使用 VisDrone 2019-DET[27]和 MS COCO 数据集[28]。在 VisDrone 2019-DET 上的消融实验和对比实验共训练 300 个轮次,批次大小设置为 4。在 MS COCO 数据集上的实验共训练 450 个轮次,批次大小设置为 24。训练初始学习率为 0.001,优化器选用 Adam。

2.2　消融实验

本文为分析每个改进点对算法小目标检测性能的影响效果,以 YOLOv7-tiny 为基础(base)算法,在 VisDrone 2019-DET-train 数据集上进行 4 组实验,第 1 组为未改进的 base 算法,第 2 组增加用 P2 代表的 160×160 小目标检测层,第 3 组在第 2 组基础上使用 MPDIoU 替换边界框回归损失中的 CIoU,第 4 组在前 3 组基础上使用融合可变形卷积的 ELAN-d。本文通过对比验证集上的结果,分析各改进点在无人机航拍数据集上对小目标检测的性能影响。实验结果如表 1 所示。

表 1　不同改进点对总体效果影响

模型	参数/$\times 10^6$	FLOPs/G	mAP@0.50/%	mAP@0.50:0.95/%	FPS
base	6.0	13.1	37.5	19.6	476.2
base + P2	6.1	15.5	39.5	21.4	416.7
base + P2 + MPDIoU	6.1	15.5	39.7	21.3	416.7
base + P2 + MPDIoU + ELAN-d	6.4	13.2	39.8	21.7	384.6

注:FPS 表示每秒帧数(frames per second,FPS),用于衡量检测速度

由表 1 可知,相较于 base 算法,新增加一个小目标检测层后,模型的参数量变化不大,计算量有所增加,检测速度有所下降,FPS 较添加小目标检测层前降低了 12.5%,但是均值平均度得到明显提升,其中 mAP@0.50 提升了 2%,mAP@0.50∶0.95 提升了 1.8%。

在添加小目标检测层的基础上增加第 2 个改进点,使用 MPDIoU 损失函数替换原算法损失函数的边界框损失函数 CIoU 损失函数。由实验结果可知,改进前后,算法模型的参数量和计算量没有增加,FPS 改进前后也没有变化。MPDIoU 损失函数简化了几何约束条件,提高了低阈值下的粗定位能力,但削弱了精细化调整能力,导致 mAP@0.50 指标较改进提升了 0.2%,mAP@0.50∶0.95 却有 0.1% 的降低,但与未改进前的 base 算法相比,mAP@0.50∶0.95 仍然保持 1.7% 的增长。在前 2 个改进的基础上,将主干网络中的高效聚合层网络 ELAN 中的 3×3 传统标准卷积替换为可变性卷积 DCNv2,形成新的模块 ELAN-d,将得到的实验结果与上一组实验比较,增加少量参数,实现了计算量的降低,同时提升 mAP@0.50 和 mAP@0.50∶0.95 指标值,FPS 有所下降。

整体来看,本文改进的整体网络参数量较 base 实验略有增加,其中融合可变形卷积的 ELAN-d 是算法参数量增加的主要原因。增加小目标检测层会增加计算量浮点运算次数(floating point operations,FLOPs),而在使用 ELAN-d 模块后又减少,证明添加小目标层增加了算法的计算复杂度,ELAN-d 模块则降低计算复杂度,二者结合使得整体改进算法较 base 实验算法在计算量上基本没有变化,这表明改进算法提高检测精度的同时不会对计算资源产生太多负担。

由表 1 可知,0.50 置信度阈值下的均值平均精度 mAP@0.50 随着每一步改进逐步增加,其中增加小目标检测层对算法模型精度的提升效果明显优于其他改进点,验证了多尺度特征融合在检测中的关键作用。改进算法的 mAP@0.50 较 base 算法增加了 2.3%,明显提升了无人机航拍数据集中的小目标检测精确度。mAP@0.50∶0.95 与 mAP@0.50 的变化也整体呈现上升趋势,改进算法较改进前增加了 2.1%,这表明改进后的算法在综合多个较高置信度阈值下的检测精度指标上也有很好的表现。

2.3　VisDrone 2019-DET 数据集对比实验

2.3.1　主流算法对比实验

为了验证改进模型在无人机航拍数据集上的优越性,本文在 VisDrone 2019-DET-train 上对改进前的 YOLOv7-tiny 算法和本文改进算法进行训练,并在 VisDrone 2019-DET-val 得到验证结果。

本文选取单阶段多框检测器(single shot multi-box detector,SSD)、EffcientDet、RetinaNet、CenterNet、Faster R-CNN、YOLOv3-SPP、YOLOv5s、YOLOv5、YOLOv6 共 9 种算法与改进前后算法在无人机航拍数据集 VisDrone 2019-DET 的验证集上进行对比,将 mAP@0.50 和 mAP@0.50∶0.95 作为评价指标,以分析本文提出改进算法的检测效果。实验结果如表 2所示。

表 2　不同模型 VisDrone 2019-DET 数据集对比实验　%

算法	mAP@0.50	mAP@0.50∶0.95
SSD[29]	10.6	5.0
EffcientDet[30]	21.1	12.8
RetinaNet[31]	25.5	15.1
CenterNet[32]	29.1	14.0
Faster R-CNN[33]	35.6	19.6
YOLOv3-SPP[34]	18.8	10.6
YOLOv5s[35]	32.8	17.6
YOLOv5[31]	34.9	19.1
YOLOv6[36]	28.8	19.0
YOLOv7-tiny	37.5	19.6
本文	39.8	21.7

由表 2 数据可知,本文提出改进算法的 mAP@0.5 和 mAP@0.50∶0.95 指标表现最佳。本文改进算法在验证集中的结果 mAP@0.50 高达 39.8%,较 SSD 算法增加 29.2%,较 EffcientDet 算法增加 18.7%,较 RetinaNet 算法增加 14.3%,较 CenterNet 算法增加 10.7%,较 Faster R-CNN 算法增加 4.2%,较 YOLOv3-SPP 算法增加 21.0%,较 YOLOv5s 算法增加 7.0%,较 YOLOv5 算法增加 4.9%,较 YOLOv6 算法增加 11.0%,较未改进前的 YOLOv7-tiny 算法增加

2.3%,通过实验验证,精度得到了有效增强。

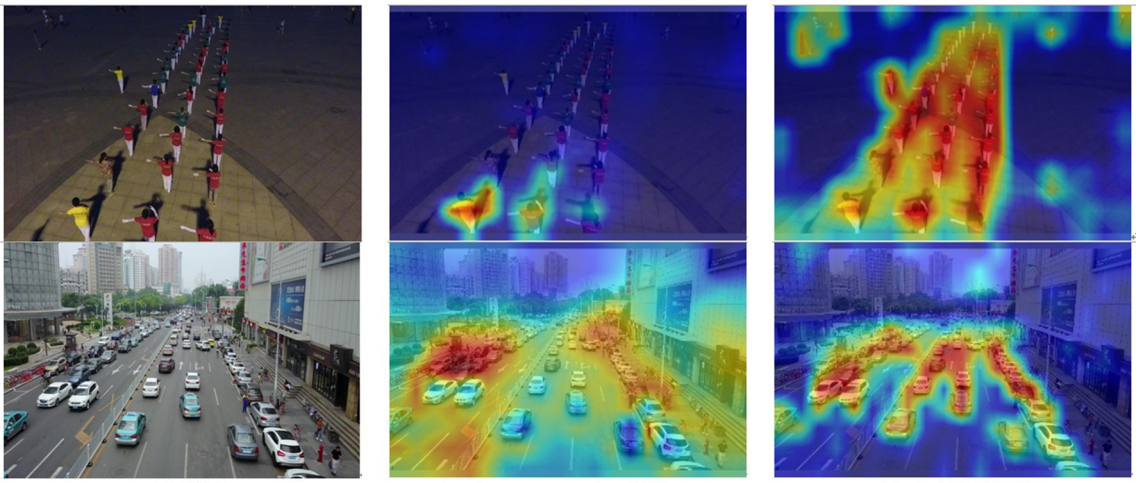
同时,本文改进算法的在 mAP@0.50∶0.95 指标上也优于其余 10 种算法,达到了 21.7%,其中相比于 SSD 算法增加了 16.7%,相比于 EffcientDet 算法增加 8.9%,相比于 RetinaNet 算法增加 6.6%,相比于 CenterNet 算法增加 7.7%,相比于 Faster R-CNN 算法增加 2.1%,相比于 YOLOv3-SPP 算法增加 11.1%,相比于 YOLOv5s 算法增加 4.1%,相比于 YOLOv5 算法增加 2.6%,相比于 YOLOv6 算法增加 2.7%,相比于未改进前的 YOLOv7-tiny 算法增加 2.1%。总而言之,VisDrone 2019-DET 数据集上的对比实验结果验证了本文算法对提升无人机航拍小目标检测精度性能的有效性。

2.3.2 具体图像结果分析

本文从 VisDrone 2019-DET 数据集中选取 2 张典型场景图像,每张图像分别对应小目标检测中的低光照环境、目标密集分布两类典型挑战场景。基于可视化工具 Grad-CAM 生成热力图,直观呈现 YOLOv7-tiny 算法与本文算法在典型的无人机航拍图像中的目标关注区域。

如图 9 所示,图 9(a)为原始检测图像,图 9(b)为 YOLOv7-tiny 算法生成的热力图,图 9(c)为本文改进算法生成的热力图。

图 9(a)中第 1 张图片展示了夜间广场舞场景,图像中主要部分为跳广场舞的人群。该场景有明显的亮度变化,远处和角落区域几乎处于黑暗状态。如 YOLOv7-tiny 算法对应的热力图所示,图中只有光照充足的 2 处热度较高,这 2 处仅左侧目标呈现高亮,右侧目标亮度相对较低,剩余目标呈现与无关背景一致的深蓝色,代表改进前算法仅关注光照条件好的 2 个目标,忽略图片中其余众多检测目标。而从本文改进算法对应的热力图中可以看出,改进后算法对光照条件不同的广场舞队伍中的每 1 个目标人物都给予了重点关注。同时,改进后算法不仅关注到广场舞队伍中光照条件差、尺寸小的远处目标,还关注到了图片角落中的路人目标。图片角落处的光照条件恶劣,路人目标小且不完整,检测十分困难,改进后算法的表现充分证明算法在恶劣光照环境下的优越性。



(a) 原图　(b) YOLOv7-tiny 算法　(c) 本文改进算法

图 9　算法改进热力图对比

图 9(a)中第 2 张图片检测场景为繁华路段道路,主道上行驶着较多车辆,两侧也几乎停放满了车辆,远处目标物体既密集又微小。面对如此密集的检测目标,图 9(b)中 YOLOv7-tiny 算法对应的热力图关注范围十分发散,高亮区域仅为图片左侧较少部分车辆,其余关注区域亮度较低,且几乎覆盖住整条道路,无法集中在检测目标上,而图片中最明显的近处车辆被算法错误识别为无关背景区域。图 9(c)中的最后一幅图为本文改进算法对应的热力图,相较于改进前算法,高亮区域集中在正在行驶和停放的车辆,而不是覆盖在整条道路上,边界清晰准确,充分证明了本文改进算法在含有密集小目标场景的检测优势。

综上所述,本文改进算法在小目标面对相似背

景干扰、光照不足、目标密集问题的检测难点时，对检测效果有明显的提升，进一步验证了本文改进算法的有效性。

2.4 MS COCO 数据集对比实验

MS COCO 数据集的目标类型丰富，检测场景复杂，可以充分反映算法模型的稳定性和有效性。本文将 YOLOv7-tiny 算法和本文改进算法在 MS COCO-train2017 训练集上训练，并在 MS COCO-val2017 验证集上对训练结果进行对比，同时选取具有代表性的 YOLOv4-tiny-3l 算法、YOLOv5-s 算法以及目标检测领域先进的 SAM-DETR 算法在 MS COCO-val2017 验证集的结果进行对比分析。

本文将参数量、mAP@0.50、mAP@0.50∶0.95 作为对比分析的评价指标，同时增加平均精度均值(mean average precision，mAP)以分析 MS COCO 数据集中小目标的检测精度。实验结果如表 3 所示。

表 3 不同模型 MS COCO 数据集对比实验 %

模型	参数/$\times 10^6$	mAP@0.50	mAP@0.50∶0.95	mAP
YOLOv4-tiny-3l[9]	8.7	47.3	30.8	10.9
YOLOv5-s[36]	7.2	56.8	37.5	21.1
SAM-DETR[37]	58.0	63.2	41.8	22.1
Yolov7-tiny	6.0	52.8	34.3	19.7
本文	6.4	54.7	36.4	22.5

由表 3 可知，受限于参数量，YOLOv7-tiny 算法和本文改进算法在 mAP@0.50、mAP@0.50∶0.95 的表现稍逊一筹，本文改进算法的均值平均精度 mAP@0.50 除了高于改进前的算法 1.9%，高于参数量为 8.7×10^6 的 YOLOv4-tiny-3l 算法 7.4% 外，低于参数量为 7.2×10^6 的 YOLOv5-s 算法和参数量 59.0×10^6 的 SAM-DETR 算法，同样本文改进算法的 mAP@0.50∶0.95 也高于改进前的算法2.1%，高于参数量为 8.7×10^6 的 YOLOv4-tiny-3l 算法 5.6%，低于 YOLOv5-s 和 SAM-DETR 算法。

本文主要针对小目标做出改进，而 MS COCO 数据集中包含多种尺度的检测目标，因此增加了 mAP 指标。由表 3 数据可知，本文改进算法拥有最高的 mAP 值，达到 22.5%。值得注意的是，改进算法在 mAP@0.50 和 mAP@0.50∶0.95 2 个指标上低于 YOLOv5-s 和 SAM-DETR 算法，但在小目标的检测精度 mAP 却高于这 2 种优秀算法。YOLOv5-s 算法参数量为 7.2×10^6，高出本文改进算法 12.5%，mAP 低于本文算法 1.4%；SAM-DETR 算法的参数量高达本文算法的 9 倍，mAP 指标却低于本文算法 0.4%。

表 3 数据充分证明了本文改进算法在检测场景和目标种类更加丰富的情况下对小目标检测性能的提升效果，与改进前算法结果对比，证明本文改进算法在权威通用数据集上的有效性与参数量远超本文改进算法的 SAM-DETR 算法对比，验证了本文提出方法对小目标检测精度的提升效果。

3 实际场景检测

3.1 检测设备

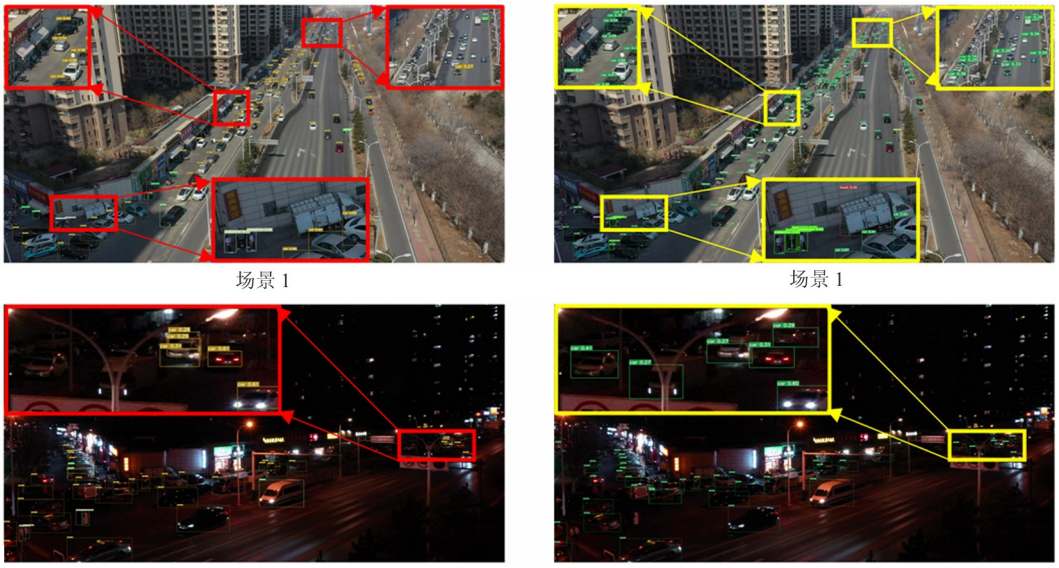
为了验证本文改进算法在实际检测场景中的有效性，本文使用改进前后的算法模型对无人机 DJI Mavic2 Pro 拍摄的图片在操作系统为 Ubuntu 20.04，GPU 为 NVIDIA GeForce RTX 4080 的设备上进行检测。本文使用无人机录制 4K 分辨率的录像，导出图片大小为 3 840×2 160 像素。

3.2 实际检测效果可视化

为检测本文改进算法在实际应用场景中对提升小目标检测效果的能力，本文使用 YOLOv7-tiny 算法和改进算法对选取图片中的目标进行检测并对比检测结果。由于实际拍摄场景单一，数据量相对较少，且实际拍摄场景与 VisDrone 2019 数据集中的道路场景相似，所以使用在 2.5 节中训练得到的最佳权重作为检测权重。

改进前后 2 种算法对 5 种不同场景的实地拍摄

图片检测结果可视化如图 10 所示,其中图 10(a)为 YOLOv7-tiny 算法检测结果,图 10(b)为改进算法检测结果。由于图片为无人机航拍所得,图中存在大量小目标对象,不易观察检测结果,为方便对比 2 种算法的检测效果,本文对同一场景图片对应位置作局部放大处理。



(a) YOLOv7-tiny 算法　　(b) 本文改进算法

图 10　改进前后算法检测可视化

图 10 中的场景 2 为夜晚繁华地段的路口,此处光照条件十分恶劣,缺乏稳定明亮光源且存在较多亮度较低的干扰光源,路边密集停放着许多车辆,正值晚高峰,道路车辆严重拥挤,检测条件十分困难。由图中局部放大处可以观察到改进后算法在此条件下仍能提高检测的成功率,证明了本文算法的有效性。综合对实际检测场景检测结果的可视化分析,本文改进算法减少了小目标的错检、漏检情况,对较小目标的检测效果有明显的提升,当小目标面对遮挡、密集、光照条件不适宜等难以检测的情况时,一定程度上改善了检测的效果。根据以上对现实场景拍摄照片的检测效果分析,有力地证明了本文提出的改进方法的有效性。

4　结　论

本文针对无人机航拍图像中存在的小目标物体检测效果不佳的问题,提出了一种基于多尺度融合与可变形卷积的小目标检测算法模型。本文首先通过在特征提取部分引入可变形卷积 DCNv2 替换 3×3 标准卷积,突破传统卷积核的固定几何约束,通过动态调整感受野形状,提升车辆、行人等小目标的特征提取精度,有效解决了无人机视角下目标形变问题。其次,通过新增小目标检测层,保留小尺度目标更多的浅层特征信息,加强不同尺度的信息融合,实现浅层细节与深层语义的协同优化。最后,通过改进边界框损失函数,使用基于边界框优化策略的 MPDIoU,进一步提高了小目标的检测效果。

图 10 中场景 1 中改进后算法不仅提高了道路远处和门市旁边车辆的检测成功率,还提高了小目标检测的正确率。由于场景 1 左下角局部放大处的正在聊天的人处在阴影处且距离较近,改进前算法虽然成功检测到人,但错误地检测为 2 个人,遗漏检测目标,改进后算法则能正确地识别为 4 个人。

通过实验验证,改进后算法与改进前算法及其他经典的算法相比,在小目标物体检测性能上拥有更好的检测精度。本文改进算法能够有效应对无人机航拍图像中的小目标检测问题,在作出的改进中,增加针对更小目标的检测层是提升改进算法网络检测精度的关键性因素。本文研究的创新性在于通过改进网络结构和损失函数,解决了无人机航拍小目标检测中的关键问题,并在现实场景图像中进行实验验证,结果证明改进算法具有重要的理论价值和

实际应用意义。

本文改进算法提升了检测精度,但同时增加了算法模型的参数量,加重了计算负担,一定程度上影响检测的速度。无人机通常搭载显存和计算资源有限的嵌入式设备,如果检测算法的计算需求过大,将直接影响无人机的实时处理能力。随着移动设备、嵌入式系统等边缘设备的快速发展,未来研究将聚焦于如何在资源受限的环境下实现目标检测技术的实时部署,实现检测精度和速度协同优化。

参考文献

[1] Ye J, Yu Z. Fusing global and local information network for tassel detection in UAV imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024,17:4100 - 4108.

[2] Zhang R, Luo B, Su X, et al. GA-Net: accurate and efficient object detection on UAV images based on grid activations[J]. Drones, 2024,8(3):74.

[3] Liu Y C, Shi G, Li Y X, et al. M-YOLO based detection and recognition of highway surface oil filling with unmanned aerial vehicle[C]//The 7th International Conference on Intelligent Computing and Signal Processing. Xi'an, China: IEEE, 2022:1884 - 1887.

[4] 江波,屈若锟,李彦冬,等. 基于深度学习的无人机航拍目标检测研究综述[J]. 航空学报,2021,42(4):137 - 151.

[5] 李雅雯,孙浩然,胡跃明,等. 基于注意力机制与多尺度特征融合的电极缺陷 YOLO 检测算法[J]. 控制与决策,2023,38(9):2578 - 2586.

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014:580 - 587.

[7] Ren S,He K,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, 2016,39(6):1137 - 1149.

[8] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE,2016:779 - 788.

[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]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023:7464 - 7475.

[10] Reis D, Kupec J, Hong J, et al. Real-time flying object detection with YOLOv8[EB/OL]. (2024 - 05 - 22)[2025 - 01 - 09] https://arxiv.org/pdf/2305.09972.

[11] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: learning what you want to learn using programmable gradient information[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2025:1 - 21.

[12] Wang A, Chen H, Liu L, et al. YOLOv10: real-time end-to-end object detection[EB/OL]. (2024 - 10 - 30)[2025 - 01 - 09] https://arxiv.org/pdf/2405.14458.

[13] Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements[EB/OL]. (2024 - 10 - 23)[2025 - 01 - 09] https://arxiv.org/pdf/2410.17725.

[14] Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors[EB/OL]. (2025 - 02 - 18)[2025 - 03 - 24]https://arxiv.org/pdf/2502.12524.

[15] Yu C, Shin Y. Object detection in UAV images based on RT-DETR with CG downsampling and CCFMP[C]//2024 IEEE VTS Asia Pacific Wireless Communications Symposium. Singapore: IEEE, 2024:1-4.

[16] 伍智华,程江华,刘通,等. 激光透窗低质量图像人体姿态识别技术研究[J]. 中国激光,2025,52(6):02.

[17] Du B, Huang Y, Chen J, et al. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023:13435 - 13444.

[18] Wang C C, He W, Nie Y, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism[EB/OL]. (2023 - 10 - 23)[2025 - 01 - 09] https://arxiv.org/pdf/2309.11331.

[19] 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. Honolulu, USA: IEEE, 2017:936 - 944.

[20] Liu S,Qi L,Qin H,et al. Path aggregation network for instance segmentation[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA:IEEE, 2018:8759 - 8768.

[21] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017:764 - 773.

[22] Li S, Cao S, Hui L, et al. A computational-efficient deformable convolution network accelerator via hardware and algorithm co-optimization[C]//2022 IEEE Workshop on Signal Processing Systems, Rennes: IEEE,2022:1 - 6.

[23] Zhu X, Hu H, Lin S, et al. Deformable ConvNets V2: more deformable, better results[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:9308 - 9308.

[24] 桑嘉更,张志佳,肖传民,等. 基于 YOLOv8s 模型改进的道路交通目标检测方法研究[J]. 红外与激光工程,2024,53(11):303 - 316.

[25] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE,2017:618 - 626.

[26] Siliang M,Yong X. MPDIoU:a loss for efficient and accurate bounding box regression)[EB/OL]. (2023 - 07 -

14)[2025-01-09] https://arxiv.org/pdf/2307.07662.
[27] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea: IEEE, 2019: 213-226.
[28] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//European Conference on Computer Vision. Zurich, Switzerland: ECCV, 2014: 740-755.
[29] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Amsterdam: Netherlands, 2016:21-37.
[30] Sun C, Chen Y, Xiao C, et al. YOLOv5s-DSD: an improved aerial image detection algorithm based on YOLOv5s[J]. Sensors, 2023,23(15):6905.
[31] Lu S, Lu H, Dong J, et al. Object detection for UAV aerial scenarios based on vectorized IOU[J]. Sensors, 2023,23(6):3061.
[32] Duan K, Bai S, Xie L, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:6568-6577.
[33] Ren S,He K,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, 2016,39(6):1137-1149.
[34] Redmon J,Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2020-04-08)[2025-01-09]. https://arxiv.org/pdf/1804.02767.
[35] Ding K,Li X,Guo W,et al. Improved object detection algorithm for drone-captured dataset based on yolov5[C]// 2022 2nd International Conference on Consumer Electronics and Computer Engineering. Guangzhou, China:IEEE, 2022:895-899.
[36] Li C, Li L, Jiang H, et al. YOLOv6: a single-stage object detection framework for industrial applications[EB/OL]. (2022-09-07)[2025-01-09] https://arxiv.org/pdf/2209.02976.
[37] Hou X, Liu M, Zhang S, et al. Salience DETR: enhancing detection transformer with hierarchical salience filtering refinement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2024:17574-17583.

Research on small target detection for UAV aerial images based on multi-scale fusion

Chen Zhiwang[*,**], Wang Ting[*,**], Lyu Changhao[***], Zhang Hanfeng[*,**], Peng Yong[****]
([*]Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao 066004)
([**]Key Laboratory of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004)
([***]Key Laboratory of Power Electronics for Energy Conservation and Drive Control of Hebei Province, Yanshan University, Qinhuangdao 066004)
([****]School of Electrical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

To address the problem of low detection accuracy of small target in unmanned aerial vehicle(UAV) aerial images, an improved small target detection algorithm is proposed in this paper. First, the algorithm introduces deformable convolution into the backbone network, and enhances the feature extraction ability of irregular small targets by adaptively adjusting the sampling position in the convolution operation. Secondly, a small target detection layer is added to the neck network to retain more shallow-layer detail features and strengthen the information fusion between different scale feature layers. Finally, the bounding box regression loss function is optimized by adopting MPDIoU (minimum point distance intersection over union). Extensive experiments are conducted on the VisDrone 2019 and MS COCO datasets. The experimental results show that the mAP@0.50 of the improved algorithm on the VisDrone2019 dataset reaches 39.8%, which is 2.3% higher than that of the original algorithm, and the small target detection mAP (mean average precision) on the MS COCO dataset is 22.5%, which is 2.8% higher than that of the original algorithm. Furthermore, visualizations of both real-world scenarios and VisDrone 2019 samples are provided, with qualitative analyses validating the algorithm's effectiveness.

Key words: unmanned aerial vehicle aerial images, small object detection, multi-scale integration, deformable convolution, minimum point distance intersection over union, visual analysis