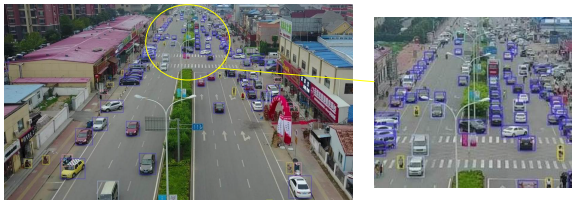
Opto-Electronic Engineering
光 电 工 程

Article
2022 年, 第 49 卷, 第 3 期

DOI: 10.12086/oee.2022.210372

基于改进 YOLOv5s 的无人机图像实时目标检测



陈　旭，彭冬亮，谷　雨*

杭州电子科技大学自动化学院，浙江 杭州 310018

摘要：针对无人机图像背景复杂、分辨率高、目标尺度差异大等特点，提出了一种实时目标检测算法 YOLOv5sm+。首先，分析了网络宽度和深度对无人机图像检测性能的影响，通过引入可增大感受野的残差空洞卷积模块来提高空间特征的利用率，基于 YOLOv5s 设计了一种改进的浅层网络 YOLOv5sm，以提高无人机图像的检测精度。然后，设计了一种特征融合模块 SCAM，通过局部特征自监督的方式提高细节信息利用率，通过多尺度特征有效融合提高了中大目标的分类精度。最后，设计了目标位置回归与分类解耦的检测头结构，进一步提高了分类精度。采用 VisDrone 无人机航拍数据集实验结果表明，提出的 YOLOv5sm+模型对验证集测试时交并比为 0.5 时的平均精度均值(mAP50)达到了 60.6%，相比于 YOLOv5s 模型 mAP50 提高了 4.8%，超过 YOLOv5m 模型的精度，同时推理速度也有提升。通过在 DIOR 遥感数据集上的迁移实验也验证了改进模型的有效性。提出的改进模型具有虚警率低、重叠目标识别率高的特点，适合于无人机图像的目标检测任务。

关键词：无人机图像；实时目标检测；YOLOv5sm+

中图分类号：TP391.41　　**文献标志码：**A

陈旭，彭冬亮，谷雨. 基于改进 YOLOv5s 的无人机图像实时目标检测 [J]. 光电工程，2022，**49**(3): 210372
Chen X, Peng D L, Gu Y. Real-time object detection for UAV images based on improved YOLOv5s[J]. *Opto-Electron Eng*, 2022, **49**(3): 210372

Real-time object detection for UAV images based on improved YOLOv5s

Chen Xu, Peng Dongliang, Gu Yu*

School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China

Abstract: As unmanned aerial vehicle (UAV) image has the characteristics of complex background, high resolution, and large scale differences between targets, a real-time detection algorithm named as YOLOv5sm+ is proposed in this paper. First, the influence of network width and depth on UAV image detection performance was analyzed, and an improved shallow network based on YOLOv5s, which is named as YOLOv5sm, was proposed to improve the detection accuracy of major targets in UAV image through improving the utilization of spatial features extracted by residual dilated convolution module that could increase the receptive field. Then, a feature fusion module SCAM was designed, which could improve the utilization of detailed information by local feature self-supervision and could improve classification accuracy of medium and large targets through effective feature fusion. Finally, a detection head structure consisting with decoupled regression and classification head was proposed to further improve the

收稿日期：2021−11−22；收到修改稿日期：2022−01−04
基金项目：浙江省自然科学基金资助项目 (LY21F030010)
通信作者：谷雨，guyu@hdu.edu.cn。

classification accuracy. The experimental results on VisDrone dataset show that when intersection over union equals 0.5 mean average precision (mAP50) of the proposed YOLOv5sm+ model reaches 60.6%. Compared with YOLOv5s model, mAP50 of YOLOv5sm+ has increased 4.1%. In addition, YOLOv5sm+ has higher detection speed. The migration experiment on the DIOR remote sensing dataset also verified the effectiveness of the proposed model. The improved model has the characteristics of low false alarm rate and high recognition rate under overlapping conditions, and is suitable for the object detection task of UAV images.

Keywords: UAV image; real-time object detection; YOLOv5sm+

1 引 言

“无人机+行业应用”逐渐成为社会刚需，实现无人机图像的目标准确实时检测与跟踪是在安防巡警、农业防害、电力检修、物联网运输等领域广泛应用需要解决的核心问题之一。与通用目标检测不同，无人机视角图像小目标多且密集，不同类型目标间尺度差异大、背景复杂等特点[1]严重影响了目标检测的精度，而无人机图像的高分辨率特点对目标检测模型优化设计提出了挑战。

随着深度学习理论与技术的飞速发展，基于深度学习的通用目标检测性能取得了远超传统方法的性能[2-4]，基于深度学习的通用目标检测算法可分为 R-CNN[2] 系列双阶段算法和 YOLO[3]、SSD[4] 系列单阶段算法。单阶段检测器有着端到端的性能优势，但是在小目标定位识别上精度偏低。双阶段目标检测器以先定位后识别的方式，在精度方面优于单阶段，但实时性较差。

深度目标检测模型算法的训练需要大量的数据，目前主要的无人机图像目标检测数据集包括 VisDrone[5]、UAVDT[6] 等。VisDrone 无人机数据集由多架无人机倾斜俯视拍摄而成，涵盖了中国 14 个城市景观，包含 10000 张图像以及 260 万标注信息，对于检测、跟踪任务而言仍然是一个难度较高的数据集。VisDrone 数据集中图片分辨率高达 2000×1500，包含 10 种目标类别，其中 people 类和 pedestrians 类极易混淆，图像的尺度、方向多样性、强度不均匀、退化严重等特点对算法设计提出了极大挑战。UAVDT 无人机数据集是一个大规模的目标检测跟踪数据集，由 100 个航拍视频、4 万张图片及其 84.15 万标注信息组成，图片大小为 1080×540，包含不同天气状况、飞行高度、摄像机视图和遮挡等 14 种不同场景下的三类车辆图像，UAVDT 数据集中图像的背景复杂度高于 VisDrone 数据集。

直接将通用目标检测算法应用于无人机图像目标检测时，由于无人机图像的上述特性，检测性能通常会有较大降低，因此研究学者进行了有针对性的改进，主要从优化双阶段检测网络、进行数据增强、优化无锚方法、优化轻量化网络等几个方面展开。

为充分利用双阶段网络在小目标检测上的优势，文献 [7] 针对提高 IOU 训练阈值存在的问题，提出了一种级联指导 IOU 重采样的网络结构 Cascade R-CNN，显著提升了小目标检测精度，但推理速度有所降低。基于无人机图像目标聚集的特点，文献 [8] 基于 R-CNN 改进算法，提出了一种多阶段集群检测网络 ClusDet。该网络使用区域聚类、切片检测、尺度适应的方法，提高了双阶段目标检测网络在高分辨率无人机图像上的运行速度与小目标检测率。Singh 等人[9]使用小目标缩放、均衡正负样本的训练策略，提高了双阶段 R-CNN 的精度水平；文献 [10] 使用多方法联合增强的训练策略，解决了训练网络的过程中存在的尺度变化、目标稀疏、类别不均衡等问题，在不牺牲推理速度的情况下大幅提高精度。近年新兴的无锚网络十分适用于无人机图像小目标检测：Duan 等人提出的 CenterNet[11] 提出的视定位为检测中心点及其偏移的任务，使用预测焦点的方式进行回归，并从中心回归的偏移参数得到实际的位置信息，该方法有效增强了小目标的检出率，但高分辨率特征图也降低了算法实时性。从实时性角度出发，谷歌采用深度可分离卷积替换传统卷积，提出了 MobileNet[12] 骨干网络，大幅降低计算量，广泛应用在边缘设备上。随后也出现了大量轻量化网络：Pelee[13]、EfficientDet[14]、GhostNet[15] 等。此外基于 L1 正则化的模型剪枝、特征组之间的模型蒸馏加速方法也备受关注，提速效果明显，可兼容各种边缘设备，但是精度会出现较大损失。

结合无人机图像特点和单阶段 YOLO 系列算法的实时性和准确性，本文充分利用 YOLOv5s 的优势

解决了其深度宽度不均衡、分类精度不足等问题，有效提高了无人机场景下小模型实时检测的精度，主要创新点包括以下几点：

1) 为解决无人机图像目标尺度差异大、小目标检测率低的问题，分析了深度模型中模型深度和宽度对于无人机图像检测的性能增益，提出了可显著提高感受野的混合残差空洞卷积模块，并结合无人机图像特点对 YOLOv5s 模型进行改进，设计了 YOLOv5sm 模型；

2) 为进一步优化改进模型的实时性与识别率，设计了一种基于目标局部部件特征信息的注意力机制，提出了一种跨阶段注意力特征融合模块 SCAM；

3) 考虑到目标检测任务中位置回归与分类任务之间的矛盾，通过对 YOLO 检测头进行改进，单独对分类分支进行特征后处理，实现位置回归与分类任务的隔离解耦；

4) 最后采用 VisDrone 和 DIOR[16] 数据集验证了提出算法的有效性与适用性。

2 基于 YOLOv5s 改进算法

2.1 YOLOv5s 基准算法

基于全卷积的单阶段 YOLO 目标检测算法，以其简洁、快速、易部署的优点，被广泛应用于工业领域的目标检测、跟踪、分割，其中 YOLOv5s 十分适用于无人机场景实时目标检测。与其他检测算法相比，YOLOv5s 有着如下特点。

2.1.1 骨干网络

采用 CSPDarkNet53 表征学习，在检测性能上优于 ResNet 基准算法。其深度、宽度均衡化的特性兼容了不同设备、数据集。残差网络避免了深度网络学习中的梯度消失的问题，以及 CSPNet[17](cross stage partial network) 在不丢失模型精度的条件下，加速推理 44%。在最后一个尺度的表征学习前添加 SPP[18]，极大提高感受野，提高大目标检出率和平移鲁棒性。

2.1.2 特征融合

沿用特征金字塔网络 (feature pyramid network, FPN)[19] 辅以 PANet[20] 多尺度特征融合策略，如图 2 所示，在不同特征层输出检测结果，提高了各尺度目标的检出率与定位、识别精度。结合 CSPNet 融合特征，优化特征融合速度。

2.1.3 其他特性

采取一系列数据增强方法，并沿用目标检测算法中的先验框思想，在目标数据集上主动学习得到预设先验框，使得目标定位更加精确，训练更加快速稳定。

2.2 YOLOv5sm+算法

尽管 YOLOv5s 性能优异，在无人机场景上有巨大优势，但是精度上相较 YOLOv4[21]、EfficientDet 等一流模型差距较大，故本文从骨干网络、特征融合、检测头三个方面改进 YOLOv5s，提出了一个均衡化的实时检测算法 YOLOv5sm+，力图保持运行速度的同时提高检测精度。

2.2.1 YOLOv5sm 骨干网络

深度卷积通过不断叠加卷积模块来提高检测精度，但是无人机图像小目标众多，分辨率高，一味增加深度将严重降低算法实时性，给深度网络带来难以承受的推理、后处理计算成本，而且难以在硬件中实际部署。而 YOLOv5s 模型低级特征映射少、感受野小，导致各大目标的召回率、精度偏低，故需针对无人机图像对网络进行调整。

深度网络可以映射更深层次的语义信息，这对分类有益，却不利于回归，回归框的优劣极大影响样本判定，即召回率，进而影响整体精度。宽度网络不但可以保存更多的历史信息，降低神经网络的灾难性遗

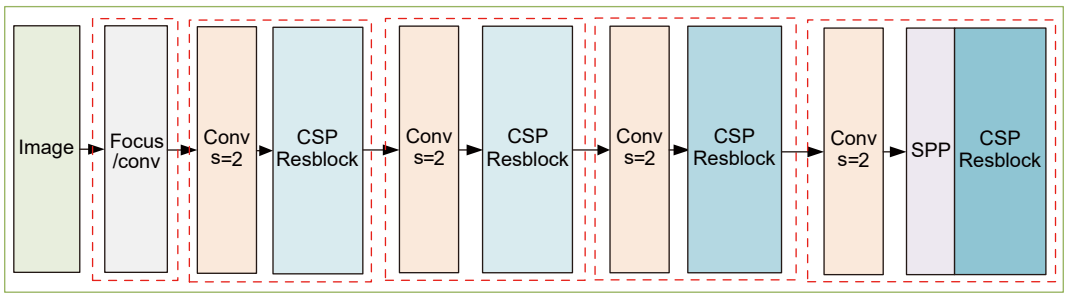


图 1 YOLOv5 骨干网络架构图

Fig. 1 YOLOv5 backbone network architecture diagram

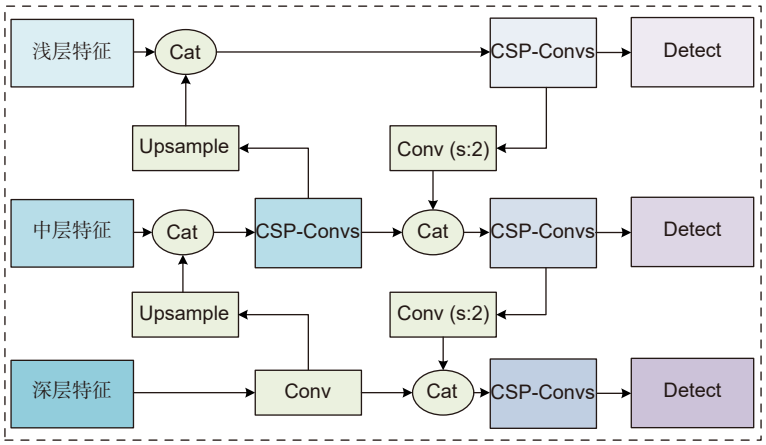


图 2　特征融合模块结构图

Fig. 2　Structure diagram of feature fusion module

忘，而且可以映射更细微的特征信息，即相似特征之间微弱的差异、偏移。这些对于无人机图像定位、识别来说尤为重要。

基于以上分析，本文对 YOLOv5s 增宽 50% 以增加模型容量，去除 Focus 模块降低对小目标定位的影响，由于低层特征提取模块内部仅有 32 维特征，故替换为残差块以增加低层内部特征的容量、信息。为了解决低层特征感受野较小的问题，本文提出了混合残差空洞卷积模块 (Res-DConv)，如图 3 (a)，通过有效提高感受野来增强背景信息对回归、分类的指导，并避免降低局部细节信息损失，提高回归的精度。如图 3 (b)，该模块 (空洞率为 3) 等价于四层普通卷积的感受野，即可以一半的计算量实现相同深度。最终提出 YOLOv5sm 轻量化骨干网络，具体架构如表 1 所示。

其次考虑到锚的尺寸需受到特征感受野、下采样次数的约束。首先根据实际数据集中目标的长宽分布，采用 K-Means 聚类确定预设先验框的大致范围，再根据表 2 中的框预设值范围对先验框进行归类。基于 YOLO 系列模型，VisDrone 数据集的预设锚点要参考实际目标出现频次、长宽先验信息与模型预测输出的最值进行判断取舍。采用契合数据集的超参数设置方法，将三个预设框增加为四个，可增加硬件设备兼容性，提高训练速度，也可细分样本的尺度变化，增大锚与真实样本框的拟合度，提升训练样本召回率，利于检测框的回归，提升小目标的检测精度。

2.2.2 SCAM 特征融合模块

为了保持小目标分类性能的同时平衡计算成本和物体尺度的方差、强化大目标的识别精度，受基于部件的细粒度目标分类[22]、注意力机制[23] 启发，本文提出了 SCAM 特征融合模块，其主要思想是基于低分辨率特征图的空间注意力对高分辨率特征图进行加权筛选，用以增强目标的部件特征，提高特征利用率，增强检测器的分类性能。本文称之为跨阶段注意力模块 (stage crossed attention module，SCAM)。

本 SCAM 模块可取代下采样模块：首先低分辨率特征经过最大池化和均值池化，连接后经过混合空洞卷积后得到注意力掩码图像 Mask；然后对高分辨率特征按照尺度转通道进行处理 (下转换) 结合 Mask 掩码对高分辨率特征进行加权，后经过通道注意力[24] 调整通道得到待融合特征；最后将高阶特征与处理后

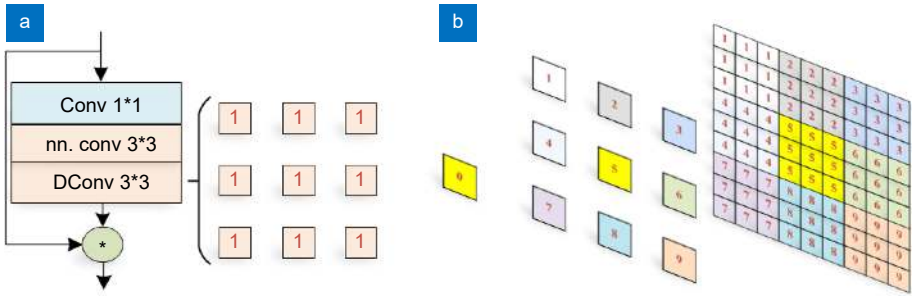


图 3　(a) Res-DConv 模块；(b) 感受野映射

Fig. 3　(a) Res-DConv module; (b) Receptive field mapping

表 1 感受野分析表

Table 1 Receptive field analysis table

YOLOv5s	感受野	通道	YOLOv5sm	感受野	通道
Focus	6	32	Conv 3*3 (stride:2)	3	24
			Conv3*3 (dilation:2)	15	48
下采样	10	64	Conv3*3 (stride:2)	19	96
			Res-Block	27	96
C3_x1	18	64	Res-Dconv	51	96
下采样	26	128	Conv 3*3 (stride:2)	59	192
C3_x3	74	128	C3_x3	107	192
下采样	90	256	Conv3*3 (stride:2)	123	384
C3_x3	186	256	C3_x3	219	384
下采样	218	512	Conv3*3 (stride:2)	251	768
Spp	218~634	512	Spp	251~667	768
C3_x1	282~698	512	C3_x1	315~731	768

表 2 呼应感受野、下采样的锚点预设置

Table 2 Pre-setting anchors in response to the receptive field and down-sampling

下采样因子	3	4	5
最大感受野/pixel	111	255	731
先验框范围	8*8~37*37	32*32~85*85	96*96~365*365

的低阶特征按维度级联融合得到融合特征。具体模块结构如图 4 中 SCAM 模块所示。

2.2.3 SDCM 检测头解耦模块

目标检测中分类和回归的矛盾本质上是卷积的平移、尺度的不变性和恒等性之间的矛盾。分类任务希望目标状态经平移、旋转、光照和尺度改变后，类别信息不变，即平移和尺度的不变性，而对于回归任务，需要目标的状态变化皆反映在特征上，进而回归出准确位置，即平移和尺度的恒等性。基于 Retina-Net[24]、Double-Head R-CNN[25] 等文献对检测头解耦的做法，本文提出 SDCM(split de-couped module) 模块，可分阶段地执行不同的任务，防止特征共用，第一阶段完成回归任务，第二阶段借助跨阶段卷积模块协助完成分类任务，从而缓解了这种互斥矛盾，提高了细类别的分类精度。

2.2.4 YOLOv5sm+模型架构

混合残差空洞卷积的高感受野和高维特征，降低了实际所需的卷积层数，使得低层特征具有较大感受野的同时包含较多的细节信息，辅以 SCAM 特征融合模块和 SDCM 检测头，用以提高检测速度与定位识别精度，改进模型并称之为 YOLOv5sm+，和 YOLOv5 相似，该模型有着四种不同的容量大小，以

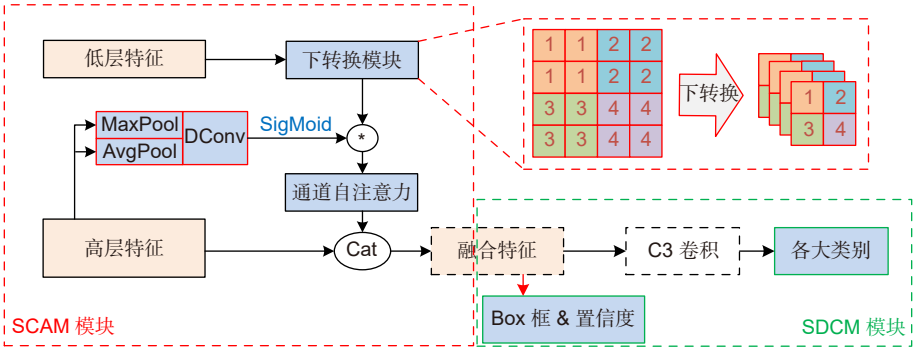


图 4 改进模块结构

Fig. 4 Improved module structure

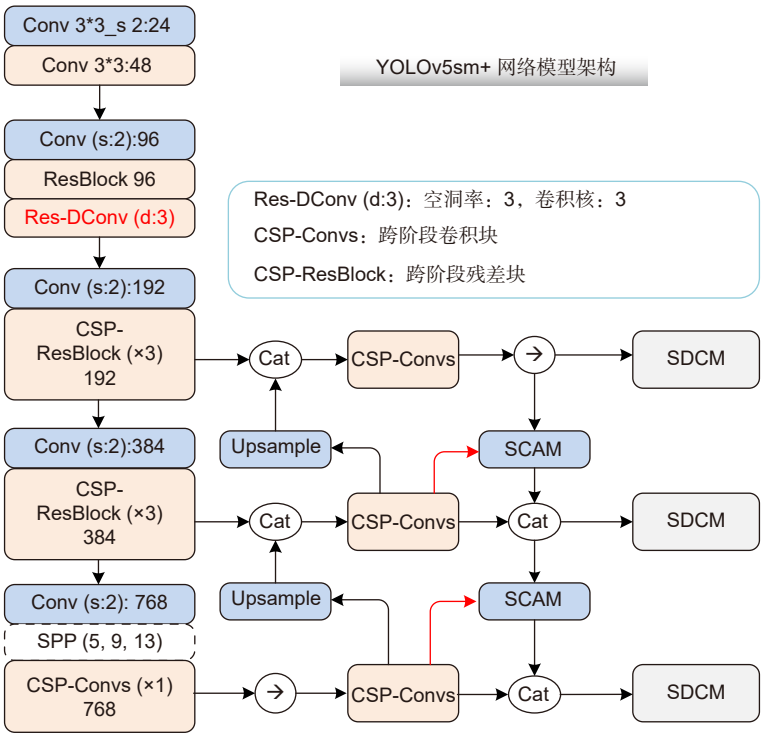


图 5 YOLOv5sm+ 模型架构

Fig. 5 YOLOv5sm+ model architecture

匹配不同设备、数据集。尤其在无人机场景中，性能优异的轻量化的模型结构十分重要。

3 实验设计

3.1 实验准备

3.1.1 数据处理

本文选取 VisDrone2019-DET 数据集进行实验，训练、验证数据集均以步长为 600，切分为 800×800 的图像，其中训练数据集有样本 25447 张图片及其标注，验证集样本 1115 张图像及其标注信息，测试集为 547 张验证集原图。

经过数值统计分析，由表 3、图 6 可知，VisDrone 数据集类别分布不均衡、小目标众多、大目标稀少，部分类间方差较小，类别混淆严重，是一个极具挑战性的数据集。

3.1.2 实验设置

实验中采用的服务器配置如下：Intel(R) i7-6850K 的 CPU，64 G 内存，NVIDIA GeForce GTX 3090 图形处理器，Ubuntu 18.04 操作系统。

所有模型训练使用双卡分布式混合精度训练，并使用单卡单批次方式进行测试。实验代码基于 ultralytics 的 YOLOv5 工程第四个版本和 yolov3-archive 工程融合改进，同时支持 yaml 模型文件和 cfg 模型文件，所有算法皆为官方模型在本工程的迁移实现。训练轮次 (epoch) 初始为 200，批大小为 16；采用 SGD 梯度下降优化器，初始学习率 0.01，动量为 0.949，采用 one-cycle 学习率衰减，其它为默认设置。

3.2 评价指标

为了准确评估深度模型在无人机空域图像上的检测性能，本文采用检测算法评估公认度最高的平均精度均值 (mean average precision, mAP)，即数据集中各类精度的平均值。每个类别根据准确率和召回率可绘制一条曲线，该曲线与坐标轴的面积则为 AP 值。其中准确率 (precision，P)、召回率 (recall，R) 定义如下。其中 TP 为真正例，FP 为假正例，FN 为假反例。

表 3 不同类型目标数量统计

Table 3 Statistics of different types of objects

目标种类	Small (0×0~32×32)	Mid (32×32~96×96)	Large (96×96~)
数量	44.44	18.63	1.704

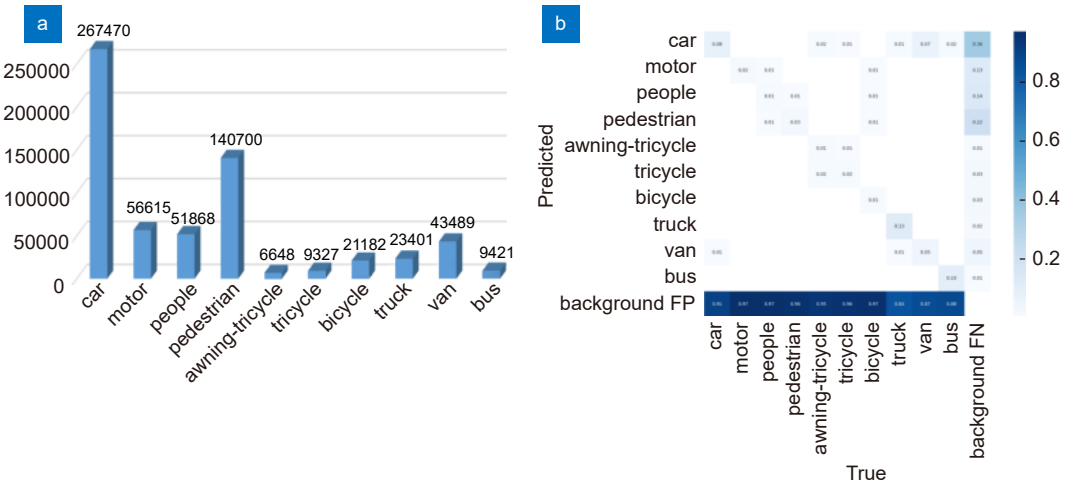


图 6 (a) VisDrone 数据集类别实例总计; (b) YOLOv5m 算法下的类混淆矩阵

Fig. 6 (a) Total number of category instances on the VisDrone dataset; (b) Classes confusion matrix of YOLOv5m algorithm

$$R=\frac{TP}{TP+FP},$$
$$P=\frac{TP}{TP+FN}. \tag{1}$$

实验中采用 COCO 评价标准[26]，使用 pycocotools 工具对检测结果进行评估分析。当检测框与真值的交并比 (intersection over union，IOU) 大于 0.5 认为该目标被准确预测，分别在 IOU 取值为 0.5、0.75、0.5:0.95 条件下的计算总类别的平均精度 (mAP50，mAP75，mAP)，并且在 IOU 为 0.5 的条件下分别统计大、中、小三种尺度目标的平均精度 (AP-large, AP-mid, AP-small)。模型实时性评估采用单张图片的平均推理时间。

4 数据分析

4.1 消融实验

4.1.1 目标检测模型中深度和宽度对检测精度影响实验分析

首先本文探索了深度和宽度对 VisDrone 数据集算法精度的增幅，为了保证对比模型的计算量、参数量一致，模型设置如下：1) 深度为 1.33、宽度为 0.5 的深度模型；2) 深度为 0.33、宽度为 0.75 的宽度模型；3) 深度为 0.33，宽度为 0.5 的基准模型 YOLOv5s。实验使用处理后的训练集以及相同的默认参数进行训练，在 800×800 的裁剪后的验证数据集上进行单尺度测试，评价模型的表征能力。由表 4 实验结果可知：1) YOLOv5s 的模型容量不足以容纳 VisDrone 数据集的知识总量；2) 在 VisDrone 无人机数据集上，相比深度网络模型，宽度网络模型对精度提升增益更大。

为了验证混合残差空洞卷积模块的有效性，设置实验如下，模型分别为 YOLOv5s 和更改残差块为 Res-Dconv 的 YOLOv5s+Res-Dconv 模型，实验条件同上。由表 5 实验结果可知，在 VisDrone 数据集上，相较原始 YOLOv5s 模型，改进模型平均精度提升 1.4%，验证了本模块可增大感受野，缩减网络深度，进而提高性能。

鉴于以上实验，本文提出了 YOLOv5sm 骨干网络，并与官方 s、m 模型进行对比实验，在 1536×1536 的图片分辨率下测试各项指标，如表 6 的 1、2 行所示，在 s 模型基准下，改进骨干的 mAP50 提高了 4.1 个百分点，优于 s 模型的 0.548，验证了改进模型在无人机图像上具有可行性。

4.1.2 特征融合 SCAM 模块对比实验

为验证 SCAM 模块的有效性，本文以 YOLOv5s 为基准模型，并以 SCAM 模块替换下采样特征融合模块进行对比实验，实验参数默认，使用单尺度训练，在 1536×1536 分辨率下进行测试，由表 6 第 1、3 行可知，相较于 YOLOv5s 基准模型，SCAM 模块提升 mAP50 近 0.7%，且目标越大，提升越明显，同时参数量和推理时间也低于基准模型。

4.1.3 SDCM 检测头解耦模块对比实验

为验证 SDCM 模块的可行性，本文仍然以 YOLOv5s 为基准模型，在此基础之上加入 SDCM 模块进行对比实验。SDCM 模块可直接替换 YOLOv5s 的检测头结构，实验条件同上，通过比较表 6 中的

第 1、4 行，得出 SDCM 模块在轻量级 s 模型基础上将性能指标 mAP 提升 1.4%，也验证了对于回归、分类解耦的可行性。

4.2 VisDrone 数据集检测结果

为了验证本文方法的有效性，本文以 YOLOv5s 模型作为基准算法，然后以 Scaled-YOLOv4[27]、YOLOv3[28] 探索精度水平，MobileNetv3[29] 探索速度基准，MobileViT[30] 试验 Transformer 算法的性能表现，YOLOX[31] 试验无锚检测算法在 VisDrone 数据集上的性能表现。实验结果如表 7 所示。

对比发现，在 1536×1536 分辨率下，基准算法 YOLOv5s 的 mAP50 精度为 54.8%，实时性最好；Scaled-YOLOv4 精度最高、YOLOv3 次之，同时模型复杂度最高，推理时间达不到无人机平台算法实时性

表 4　深度、宽度模型性能对比实验结果

Table 4　Performance comparison experiment results of depth and width models

深度	宽度	mAP50	mAP	BFLOPs
0.33	0.5	0.502	0.288	16.5
0.33	0.75	0.540	0.319	36.3
1.33	0.5	0.525	0.311	35.4

表 5　Res-Dconv 模块验证实验结果

Table 5　Verification experiment results on Res-Dconv module

Baseline	Res-Dconv	mAP50	mAP	BFLOPs
√		0.502	0.288	16.5
√	√	0.516	0.299	19.8

表 6　本文算法模块在 VisDrone 数据集上的消融实验结果

Table 6　The ablation experiment results of our algorithm modules on the VisDrone dataset

Baseline	SM	SCAM	SDCM	mAP	mAP50	BFLOPs	Infer	AP-small	AP-medium	AP-large
YOLOv5s				0.319	0.548	16.5	4.8	0.220	0.437	0.495
	√			0.358	0.589	30.1	8.3	0.280	0.476	0.495
√		√		0.324	0.555	14.7	3.8	0.225	0.446	0.511
√			√	0.333	0.555	19.5	4.9	0.250	0.448	0.482
	√		√	0.356	0.593	38.0	9.0	0.278	0.475	**0.512**
	√	√	√	**0.360**	**0.596**	30.8	**7.7**	**0.281**	**0.479**	0.505

注：加粗字体为该列最优值。

表 7　不同算法在 VisDrone 数据集上的检测性能

Table 7　Detection performance of different algorithms on VisDrone dataset

算法	mAP50	mAP	mAP75	AP-small	AP-mid	AP-large	BFLOPs	Infer/ms
YOLOv3	0.609	0.389	0.417	0.297	0.496	0.545	154.9	27.8
Scaled-YOLOv4	0.620	0.400	0.428	0.305	0.514	0.626	119.4	27.1
ClusDet[1]	0.562	0.324	0.316	-	-	-	-	-
HRDNet[1]	0.620	0.355 1	0.351	-	-	-	-	-
YOLOv5s	0.548	0.319	0.317	0.220	0.437	0.495	16.5	4.8
YOLOv5m	0.595	0.365	0.372	0.285	0.482	0.525	50.4	9.8
YOLOX-s	0.535	0.314	0.317	0.225	0.415	0.485	41.65	5.1
MobileNetv3	0.554	0.329	0.329	0.245	0.443	0.495	23.8	8.0
MobileViT	0.555	0.333	0.337	0.249	0.442	0.418	-	13.7
YOLOv5sm+	0.596	0.360	0.369	0.281	0.479	0.505	30.8	7.7
YOLOv5sm+*	0.606	0.367	0.378	0.295	0.478	0.439	-	-

注：+为添加改进模块的模型，*为多尺度测试结果，包含引用文献实验结果。

的要求；对于轻量级网络 MobileNetv3 来说，精度优于 YOLOv5s 模型，但是速度上次于 YOLOv5s；基于注意力的 Transformer 轻量级网络性能上并不占优势；无锚检测器 YOLOX 的精度最低，即在轻量骨干网络下，无锚检测器回归精度低，性能较差。

YOLOv5sm+模型较基准模型 mAP50 高 5.6%，优于 m 模型且推理速度提升 21.4%，也验证了改进算法在无人机数据集上的有效性。由图 7 可知，本文 YOLOv5sm+模型在远景小目标的检出率优于 s 和 m 模型，由图 8 可知，在重叠度高的目标集群中，本文算法可以更准确的检测出实际的目标，虚警低于对比模型。

4.3 DIOR 数据集迁移实验

为了充分验证本文方法有效性和鲁棒性，本文在 DIOR 遥感数据集上进行了对比验证。该数据集是西北工业大学于 2019 年发布了一个大规模的空域遥感数据集，数据集在不同成像条件、天气和季节下采集而成，覆盖 20 个目标类别，类间相似、类内多样，

图 7 不同算法在 VisDrone 无人机场景下的检测实例。

(a) YOLOv5m 模型; (b) YOLOv5sm+模型; (c) YOLOv5s 模型

Fig. 7 The detection examples of different algorithms in the VisDrone UAV scene.

(a) YOLOv5m model; (b) YOLOv5sm+ model; (c) YOLOv5s model

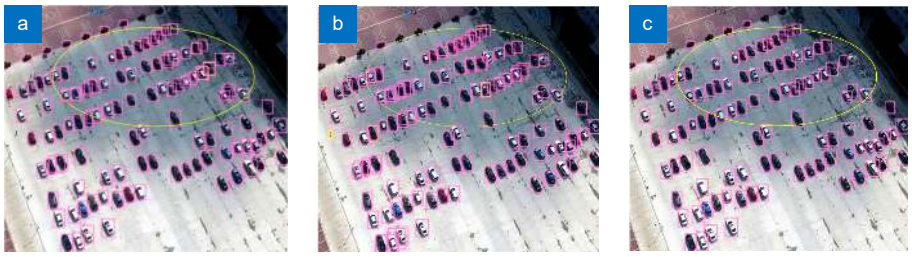


图 8 三种算法在密集车辆场景的检测结果对比图。

(a) YOLOv5m; (b) YOLOv5s; (c) YOLOv5sm+

Fig. 8 Comparison of the detection effects of three algorithms in dense vehicle scenes.

(a) YOLOv5m; (b) YOLOv5s; (c) YOLOv5sm+

尺度差异性大，背景复杂，适合迁移验证本文算法有效性。训练参数按照默认实施，使用多尺度训练模型 200 个轮次。DIOR 数据集使用官方的数据划分，分为 5876 张训练集，876 张验证集，14885 张测试集，采用默认超参数设置。

实验结果如表 8 所示，在 20 类的 DIOR 遥感数据集上，相对于 YOLOv5s 模型，改进模型检测精度提升近 4.2%，达到了 66.7%，优于 Faster R-CNN 两阶段算法。由如图 9 中的部分检测实例可知，本模型在虚警率、密集目标分辨率上表现优于 YOLOv5s 模型。实验表明，本文算法对于小目标众多、尺度差异大、目标重叠度的数据集可以实现较好的鲁棒性。

5 结 论

本文以无人机监视场景为背景，分析了 VisDrone 无人机视角数据集的目标分布规律。首先探索了在 UAV 数据集上深度和宽度对 YOLOv5 模型的精度增幅，实验结果表明，在无人机数据集上，虽然深度模型的深层语义提高了模型精度，但是由于内部特征匮乏，深度模型性能差于宽度模型，主要影响精度水平的是内部特征映射量。基于混合残差空洞卷积模块，提出了一种均衡化的实时目标检测模型 YOLOv5sm，精度高于 YOLOv5s 模型 4.1 个百分点。SCAM 特征融合模块提高了特征空间利用率和特征融合速度，进一步提升了检测精度。在 VisDrone 数据集上的结果表明，目标尺度越大，精度提升越明显。最后基于解耦的思想改进检测头结构，进一步提升了精度水平。通过与 Scaled-YOLOv4、MobileNetv3 轻量化网络、MobileViT 注意力网络、YOLOX 无锚检测器对比可知，改进模型可显著提高模型精度，验证集 mAP50 高达 60.6%，优于 m 模型且速度提升 21.4%，基本满足无人机边缘设备上的性能、精度要求。在 DIOR 数

表 8 不同算法在 DIOR 数据集上的检测性能

Table 8 Detection performance of different algorithms on DIOR dataset

模型	BackBone	mAP50
Faster R-CNN [33]	VGG16	0.541
PANet [20]	ResNet50	0.638
Retina-Net[24]	ResNet50	0.685
文献[32]	ResNet50	**0.732**
CAT-Net[34]	ResNet50	0.763
YOLOv5sm+(ours)	-	0.667

注：加粗字体为该列最优值，包含其他文献对比结果。

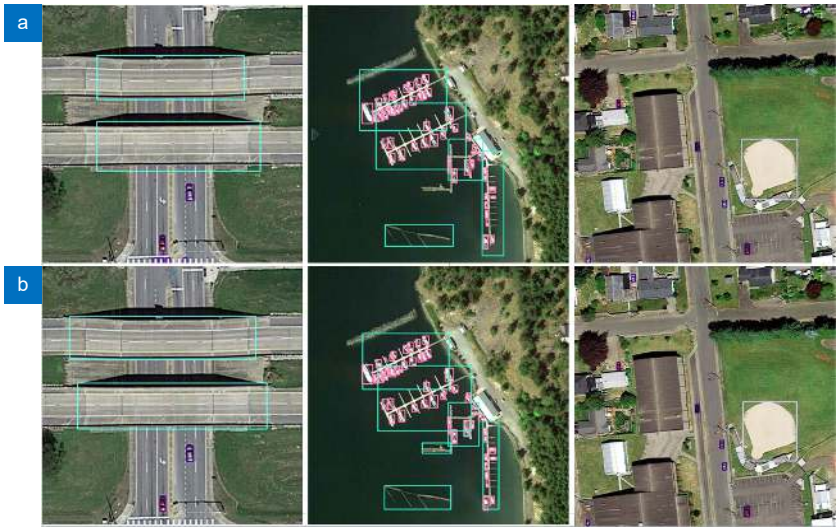


图 9 改进算法在 DIOR 数据集的检测对比。

(a) YOLOv5s; (b) YOLOv5sm+

Fig. 9 Detection comparison of improved algorithm in DIOR dataset.

(a) YOLOv5s; (b) YOLOv5sm+

据集上的迁移实验表明，改进模型相较于 YOLOv5s 基准模型，mAP50 提升 4.2%，验证了算法的有效性和鲁棒性。虽然改进模型的精度、速度在 VisDrone 数据集上较为可观，但后期工作仍需关注召回对目标的精度影响以及在无人机并行设备上的实际测试部署工作。

参考文献

[1] Wu X, Li W, Hong D F, et al. Deep learning for UAV-based object detection and tracking: a survey[EB/OL]. arXiv: 2110.12638. https://arxiv.org/abs/2110.12638.

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788. doi: 10.1109/CVPR.2016.91.

[4] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 21–37. doi: 10.1007/978-3-319-46448-0_2.

[5] 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 2019 IEEE/CVF International Conference on Computer Vision Workshop*, 2019: 213–226. doi: 10.1109/ICCVW.2019.00030.

[6] Du D W, Qi Y K, Yu H Y, et al. The unmanned aerial vehicle benchmark: object detection and tracking[C]//*Proceedings of the15th European Conference on Computer Vision*, 2018: 375–391. doi: 10.1007/978-3-030-01249-6_23.

[7] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 6154–6162.

[8] Yang F, Fan H, Chu P, et al. Clustered object detection in aerial images[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 8310 –8319. doi: 10.1109/ICCV.2019.00840.

[9] Singh B, Najibi M, Davis L S. SNIPER: efficient multi-scale training[C]//*Proceedings of Annual Conference on Neural Information Processing Systems 2018*, 2018: 9333–9343.

[10] Wei Z W, Duan C Z, Song X H, et al. AMRNet: chips augmentation in aerial images object detection[EB/OL]. arXiv: 2009.07168. https://arxiv.org/abs/2009.07168.

[11] Duan K W, Bai S, Xie L X, et al. CenterNet: Keypoint triplets for object detection[C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 6568–6577. doi: 10.1109/ICCV.2019.00667.

[12] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. arXiv: 1704.04861. https://arxiv.org/abs/1704.04861.

[13] Wang R J, Li X, Ao S, et al. Pelee: a real-time object detection system on mobile devices[C]//*Proceedings of the 6th International Conference on Learning Representations*, 2018.

[14] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10778–10787.

[15] Han K, Wang Y H, Tian Q, et al. Ghostnet: more features from cheap operations[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 1580–1589. doi: 10.1109/CVPR42600.2020.00165.

[16] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: a survey and a new benchmark[J]. *ISPRS J Photogr Remote Sens*, 2020, **159**: 296−307.

[17] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 1571 –1580. doi: 10.1109/CVPRW50498.2020.00203.

[18] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2015, **37**(9): 1904−1916.

[19] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944.

[20] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8759–8768. doi: 10.1109/CVPR.2018.00913.

[21] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. arXiv: 2004.10934. https://arxiv.org/abs/2004.10934.

[22] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//*Proceedings of the 13th European Conference on Computer Vision*, 2014: 834–849. doi: 10.1007/978-3-319-10590-1_54.

[23] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 3–19. doi: 10.1007/978-3-030-01234-2_1.

[24] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 2999 –3007. doi: 10.1109/ICCV.2017.324.

[25] Wu Y, Chen Y P, Yuan L, et al. Rethinking classification and localization for object detection[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10183 –10192. doi: 10.1109/CVPR42600.2020.01020.

[26] Chen X L, Fang H, Lin T Y, et al. Microsoft coco captions: data collection and evaluation server[EB/OL]. arXiv: 1504.00325. https://arxiv.org/abs/1504.00325.

[27] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-YOLOv4: scaling cross stage partial network[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 13024 –13033. doi: 10.1109/CVPR46437.2021.01283.

[28] Farhadi A, Redmon J. Yolov3: an incremental improvement[C]//*Proceedings of Computer Vision and Pattern Recognition*, 2018: 1804–2767.

[29] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3 [C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 1314 –1324. doi: 10.1109/ICCV.2019.00140.

[30] Mehta S, Rastegari M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[EB/OL]. arXiv: 2110.02178. https://arxiv.org/abs/2110.02178.

[31] Ge Z, Liu S T, Wang F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. arXiv: 2107.08430. https://arxiv.org/abs/2107.08430.

[32] Yao Y Q, Cheng G, Xie X X, et al. Optical remote sensing image object detection based on multi-resolution feature fusion[J]. *J Remote Sens*, 2021, **25**(5): 1124−1137.
姚艳清, 程塨, 谢星星, 等. 多分辨率特征融合的光学遥感图像目标检测[J]. 遥感学报, 2021, **25**(5): 1124−1137.

[33] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137−1149.

[34] Liu Y, Li H F, Hu C, et al. CATNet: context AggregaTion network for instance segmentation in remote sensing images[EB/OL]. arXiv: 2111.11057. https://arxiv.org/abs/2111.11057.

作者简介

陈旭 (1997-)，男，硕士研究生，主要从事图像目标识别、检测与模型优化。

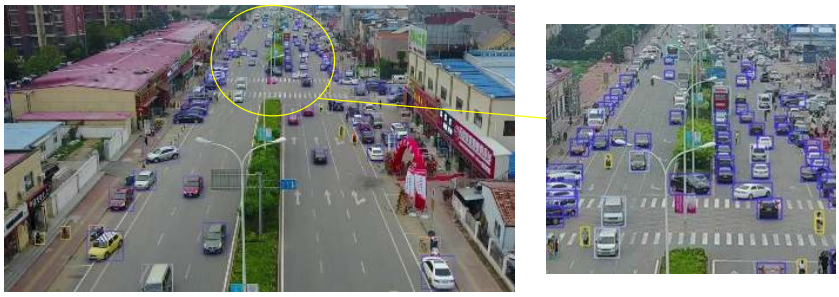
E-mail: cxu@hdu.edu.cn

【通信作者】谷雨 (1982-)，男，博士，副教授，主要从事多源信息融合、遥感图像目标检测与识别方面的研究。

E-mail: guyu@hdu.edu.cn。

Real-time object detection for UAV images based on improved YOLOv5s

Chen Xu, Peng Dongliang, Gu Yu*



YOLOv5sm+ model detection result on VisDrone UAV dataset

Overview: The real-time object detection under unmanned aerial vehicle (UAV) scenario has a wide range of military and civilian applications, including traffic monitoring, power line detection, etc. As UAV image has the characteristics of complex background, high resolution, and large scale differences between targets, how to meet the requirements of both detection accuracy and real-time performance is one of key problems to be solved. Thus, a balanced real-time detection algorithm based on YOLOv5s, which is named as YOLOv5sm+ is proposed in this paper. First, the influence of network width and depth on UAV image detection performance was analyzed. The experimental results on VisDrone datasets show that, due to the less internal feature mapping, the detection performance improves with model depth rather than model width. Moreover, with the depth of the model grows, semantic information can improve detection accuracy under generic object detection scenarios. An improved shallow network based on YOLOv5s, which is named as YOLOv5sm, was proposed to improve the detection accuracy of major targets in UAV image through improving the utilization of spatial features extracted by residual dilated convolution module that could increase the receptive field. Then, a cross-stage attention feature fusion module (SCAM) was designed, which could improve the utilization of detailed information by local feature self-supervision and could improve classification accuracy of medium and large targets through effective feature fusion. Finally, a detection head structure consisting with decoupled regression and classification head was proposed to further improve the classification accuracy. The first stage completes the regression task, and the second stage uses the cross-stage convolution module to assist in the classification task. The contradiction between regression and classification was alleviated, and the accuracy of the fine-grained classification was improved. Under the synergistic of the balanced light-weight feature extraction network (YOLOv5sm), the cross-stage attention feature fusion module (SCAM) and the improved detection head, the algorithm named YOLOv5sm+ was proposed. The experimental results on VisDrone dataset show that when intersection over union equals 0.5 mean average precision (mAP50) of the proposed YOLOv5sm+ model reaches 60.6%. Compared with YOLOv5s model, mAP50 of YOLOv5sm+ has increased 4.1%. In addition, YOLOv5sm+ has higher detection speed. The migration experiment on the DIOR remote sensing dataset also verified the effectiveness of the proposed model. The improved model has the characteristics of low false alarm rate and high recognition rate under overlapping conditions, and is suitable for the object detection task of UAV images.

Chen X, Peng D L, Gu Y. Real-time object detection for UAV images based on improved YOLOv5s[J]. *Opto-Electron Eng*, 2022, **49**(3): 210372; DOI: 10.12086/oee.2022.210372

Foundation item: Natural Science Foundation of Zhejiang Province, China (LY21F030010)

School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China

* E-mail: guyu@hdu.edu.cn