李新晨, 李晓飞. 基于 YOLOv5s 的无人机拍摄场景下的目标检测[J]. 智能计算机与应用,2025,15(4):151-157. DOI:10.20169/j.issn.2095-2163.250421

基于 YOLOv5s 的无人机拍摄场景下的目标检测

李新晨[1], 李晓飞[2]

(1 南京邮电大学 宽带无线通信技术教育部工程研究中心, 南京 210003;
2 南京邮电大学 通信与信息工程学院, 南京 210003)

摘 要: 针对无人机在高空视角捕获的画面中目标尺寸变化大、背景繁杂、目标所占像素少且其数量居多导致的目标检测精度过低的问题,提出一种改进 YOLOv5s 的目标检测算法。首先,提出一种可以提取多种尺度特征且减少参数量的 C3R 模块,在减少参数量的同时提高检测精度;其次,引入 ASFF 模块,从最佳融合中增强特征表达并自适应地学习,有利于高难度目标的检测;最后,改进多尺度融合与检测部分,增强小目标特征表达。在 VisDrone2019 数据集上进行实验,检测精度(mAP_{50}/%)比原先的模型提升了 6.5%,证明了改进模型的优越性。

关键词: 目标检测; YOLOv5; Res2Net; 自适应空间特征融合; 多尺度融合与检测
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)04-0151-07

Object detection in Unmanned Aerial Vehicle scenes based on YOLOv5s

LI Xinchen[1], LI Xiaofei[2]

(1 Engineering Research Center of Broadband Wireless Communication Technology, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; 2 School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: To address the issue of low accuracy in object detection caused by significant variations in object size, complex backgrounds, and a large number of targets with limited pixels in aerial scenes captured by Unmanned Aerial Vehicles (UAVs), an improved object detection algorithm based on YOLOv5s is proposed. Firstly, a C3R module is introduced to extract multi-scale features and reduce the number of parameters, thereby improving detection accuracy while reducing parameters. Secondly, an ASFF module is employed to adaptively learn the optimal fusion for enhanced detection of challenging objects. Lastly, the multi-scale fusion and detection components are improved to enhance the feature representation of small objects. Experimental results on the VisDrone2019 dataset demonstrate a 6.5% improvement in mean Average Precision (mAP_{50}/%), highlighting the superiority of the improved model.

Key words: object detection; YOLOv5; Res2Net; adaptive spatial feature fusion; multi-scale fusion and detection

0 引 言

近年来,随着无人机技术的进步,无人机配合深度学习在诸多领域都有发展,包括但不限于安全监控[1]、航空拍摄[2]、农业管理[3]。在上述场景中,无人机使用其自带的摄影功能进行拍摄,及时了解周围环境变化。Velusamy 等学者[4]指出了无人机在农业作物和害虫管理方面的挑战。Lygouras 等学者[5]提出了一种利用无人机实时人体探测执行搜救任务的方法。Almagbile[6]提出了一种基于加速分割测试(FAST)的算法,用于检测使用不同相机方向和位置拍摄的无人机图像中的人群特征。

目标检测作为无人机应用中的重要环节,在无人机捕获的图像中加入目标检测技术尤为关键[7]。近年来,目标检测技术从传统的图像处理技术逐渐过渡到深度卷积神经网络(CNN),主要有 2 种:一种是以 Faster-RCNN[8]为代表的两阶段算法,还有一种是以 YOLO[9]系列为代表的单阶段检测算法。在无人机拍摄的图像中,这些方法难以有效发挥作用。具体表现如下。

作者简介: 李新晨(1999—),男,硕士研究生,主要研究方向:目标检测。
通信作者: 李晓飞(1964—),男,教授,主要研究方向:信息网络与多媒体技术。Email:lixf@njupt.edu.cn。

收稿日期: 2023-09-11

(1)无人机在高空视角下拍摄的物体尺度变化极大,且随着高度变化而变化。

(2)从较高位置拍摄的图像大多包含高密度物体,物体的尺寸非常小,导致其特征不明显。

(3)高速飞行致使图像模糊,且背景复杂。

为了解决这些问题,Jiang 等学者[10]通过扩展目标检测量表和引入注意力机制,平均准确率比原来的 YOLOv4 网络提高了 5.09%。Sharoze 等学者[11]提出了一种改进的 YOLOv4 模型用于基于视觉的小目标检测,通过连接上采样层,并将上采样层的特征与原始特征连接,来增强小目标的特征,以获得更细粒度的小对象特征。奉志强等学者[12]在 YOLOv5 主干网络加入 Transformer 结构并将空间注意力和通道注意力结合最后进行多尺度特征融合提高了检测精度。Tan 等学者[13]在 EfficientDet 网络中提出了 BiFPN 的结构,该结构由加权双向特征金字塔网络构成,增加了跨尺度连接来增强特征的表示能力,给每个输入添加相应的权重来更好地进行小目标检测任务。

综上所述,为了进一步提高无人机目标检测的性能和效果,本文选择 YOLOv5s 作为基线模型,提出一种基于改进 YOLOv5s 的目标检测算法。具体来说,主要从 3 个方面进行了优化:提出一种多尺度特征表示 C3R 模块、引入 ASFF 模块、改进多尺度融合与检测。通过在 VisDrone2019[14]数据集上进行实验验证,证明了改进后的算法可以提高目标检测精度。

1 相关理论研究

YOLO 系列算法在整个深度学习目标检测领域发展史上占据着至关重要的地位,至今已经迭代推出多个版本,各版本性能也有所不同。其中,YOLOv5 凭借着轻量化的网络结构、不错的检测精度和速度,得到了广泛的应用。

YOLOv5s 网络结构如图 1 所示。YOLOv5 在输入端采用 Mosaic 数据增强方式,有效丰富了图片的背景[15],有助于小目标的检测,其中训练图像的输入尺寸为 640×640。骨干网络部分包括 Conv 模块、CSPDarkNet53 和 SPPF 模块。YOLOv5 的 Neck 采用特征金字塔(Feature Pyramid Network,FPN[16])+路径聚合网络(Perceptual Adversarial Network,PAN[17])结合的方式进行特征融合,该结构通过跨层特征融合,从而获得更加丰富的语义信息,提高目标检测的准确率,然后将提取到的特征传入到检测层,输出尺寸分别为 80×80、40×40、20×20 来实现大中小目标的分类与定位,最后使用非极大值抑制[18]等后处理操作输出置信度得分最高的物体类别。

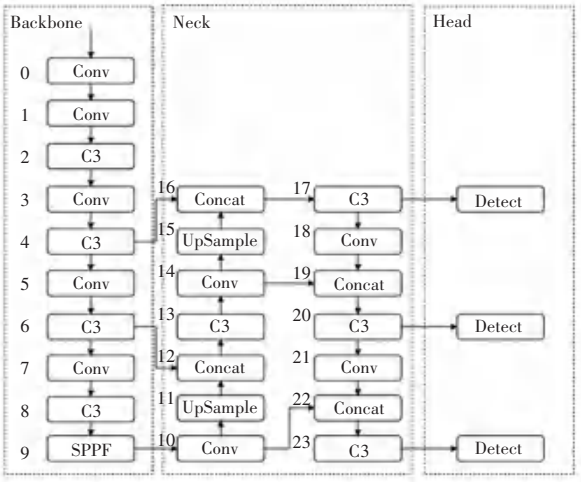


图 1 YOLOv5s 网络结构图

Fig. 1 YOLOv5 structure diagram

2 改进的 YOLOv5 目标检测算法

本文针对无人机拍摄这种复杂场景,目标在图像中占比小、目标尺寸变化大的问题,选择 YOLOv5s 作为基线模型,对其进行改进,网络结构如图 2 所示。对此研究改进,将展开分述如下:

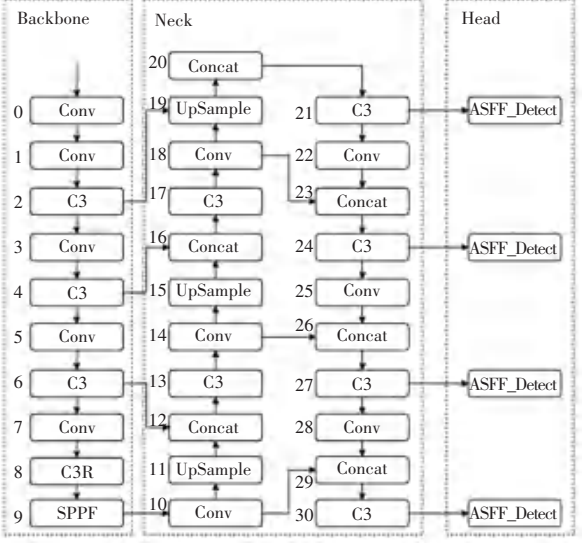


图 2 改进后的 YOLOv5s 网络结构图

Fig. 2 Improved YOLOv5s structure diagram

(1) 提出一种可以提取多尺度特征并减少参数量的 C3R 模块,可以很好地解决无人机拍摄场景下的不同尺度的问题。

(2)引用一种金字塔特征融合策略(Adaptively Spatial Feature Fusion, ASFF)[19],能在空域过滤冲突信息以抑制不一致特征,用来解决一阶检测器中

特征金字塔内部的不一致性。

(3)在 Neck 层增加 P2 层的融合，用于更好地融合小目标的特征信息。

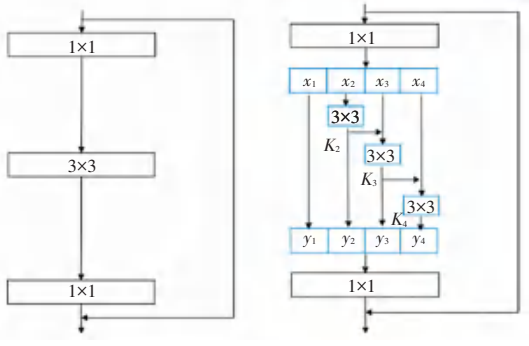
2.1　C3R 模块

Res2Net[20] 提出了一种在更细层面上进行多尺度处理的方法。Bottleneck 和 Res2Net 模块比较如图 3 所示。由图 3 可知，对比 Bottleneck 模块，Res2Net 模块使用跨层连接的方式将滤波器组进行融合，实现了增加多尺度表示的目的。

具体来说，将输入特征分为 s 组，分别记作 x_i，$i \in 1,2,\cdots,s$；每组特征图的通道数均为输入特征图通道数的 $1/s$。除 x_1 外，每组特征图都会经过 1 个 3×3 卷积，将该卷积操作记作 $\{x_j, j \leqslant i\}$。除了 x_1 和 x_2 外，第 i 组的特征图先与前一组 $K_{i-1}(\ )$ 的输出相加，将相加后的结果进行 $K_i(\ )$ 操作。上述操作可用如下公式表示：

$$y_i = \begin{cases} x_i, & i = 1 \\ K_i(x_i), & i = 2 \\ K_i(x_i + y_{i-1}), & 2 < i \leqslant s \end{cases} \tag{1}$$

接下来将这 s 组的输出再进行拼接，然后进行 1×1 的卷积、旨在完全融合信息，由于跨层带来的组合效应，可以得到诸多特征尺度。



(a) Bottleneck 模块　　(b) Res2Net 模块

图 3　Bottleneck 和 Res2Net 模块比较图

Fig. 3　Comparison diagram of Bottleneck and Res2Net modules

综上，基于 Res2Net 模块与 YOLOv5 中的 C3 模块，提出一种 C3R 多尺度特征提取模块，即将 Res2Net 替换 C3 中的 Bottleneck，以适应 YOLOv5 网络模型，通过增加模型的感受野并极大程度地减少参数量，成功提取了高效、有用的多尺度特征，实现了高精度的目标检测效果。

2.2　ASFF 模块

在传统的目标检测算法中，通常会使用多尺度特征来提高检测的准确率，但是这些特征的维度不同，难以直接融合。ASFF 通过学习动态权重，将多个不同尺度的特征图融合成一个具有适应性的特征图，从而更好地应对多尺度目标检测的挑战。

ASFF 的网络结构包括多个分支和融合模块。每个分支对应着特征金字塔中的一个层，分别提取不同尺度的特征图。在融合模块中，ASFF 使用一个自适应权重分配器来对不同尺度的特征图进行动态权重分配，使得每个尺度的特征图都能够得到充分利用。这种动态权重分配的方法可以自适应地调整权重，以适应不同的场景和目标，从而提高了检测的准确率。具体来说，ASFF 首先将每个分支的特征图进行上采样和下采样，调整尺寸并保持相同。然后，ASFF 使用一个全局平均池化层将每个分支的特征图降维成一个向量。接着，ASFF 使用一个自适应权重分配器来学习每个特征图的权重，该权重可以根据特征图的质量和场景的需求进行自适应调整。最后，ASFF 将所有特征图按照其权重进行加权融合，得到一个最终的特征图。这个特征图被用于后续的目标检测过程中。ASFF 网络结构如图 4 所示。以 ASFF-1 为例，ASFF-1 接受来自 3 个不同大小的特征层再乘以各自的权重参数 α、β、χ，然后进行相加，如下所示：

$$y_{ij}^1 = \alpha_{ij}^1 \times x_{ij}^{1\to1} + \beta_{ij}^1 \times x_{ij}^{2\to1} + \chi_{ij}^1 \times x_{ij}^{3\to1} \tag{2}$$

其中，y_{ij}^1 表示通过 ASFF-1 得到的新特征图；α_{ij}^1、β_{ij}^1、χ_{ij}^1 分别表示来自不同层的权重；$x_{ij}^{1\to1}$、$x_{ij}^{2\to1}$、$x_{ij}^{3\to1}$ 分别表示来自不同层的输出。由于采用相加的方式，需要保证 ASFF 层接收不同层的输出特征维度相同，通道数目也相同。采用卷积核大小为 1×1 的卷积层来调整成相同的通道数，将高层特征经过上采样，底层特征通过下采样，并调整成相同维度计算。其中，权重参数 α、β、χ 的和为 1，通过归一化函数将值锁定[0,1]。

总地来说，ASFF 通过对多尺度特征图进行动态权重分配，实现了特征的自适应融合，从而提高了检测的准确性。利用该模块，模型可以自适应地学习特征融合，更容易检测高难度的物体。

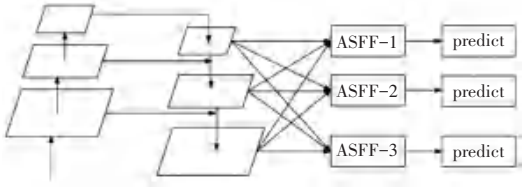


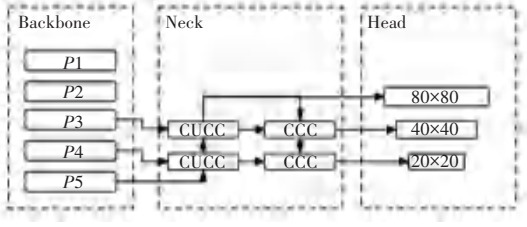
图 4　ASFF 网络结构图

Fig. 4　ASFF structure diagram

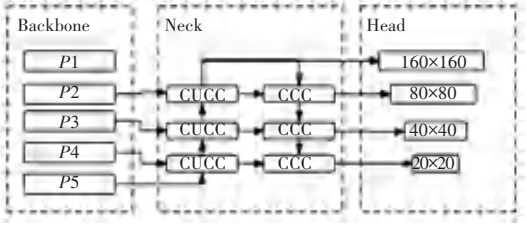
2.3　多尺度融合与检测

YOLOv5 特征融合与检测模型如图 5(a)所示。

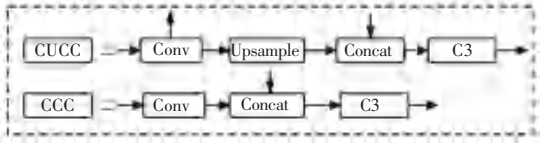
YOLOv5s 模型在骨干网络进行 5 次下采样卷积得到对应的 $Pi(i=1,2,3,4,5)$ 层特征图。在 Neck 层对 $P3$、$P4$、$P5$ 进行多尺度特征融合。最后,将 $P3$、$P4$、$P5$ 送入 Head 检测与识别。在小目标检测任务中,经常有很小的目标需要检测。在 VisDrone2019 数据集中,像素占比少的目标较多。经过 5 次下采样后,这种目标的特征信息基本已经丢失,即使是 $P3$ 检测层也难以捕捉得到,所以有必要在 $P2$ 层就对目标信息进行采集融合。本文针对无人机拍摄的复杂环境下小目标的检测与识别,对原有的 YOLOv5s 特征融合进行了改进,改进后的 YOLOv5 特征融合与检测模型如图 5(b)所示。通过上采样和 Concat 的运算融合了 $P2$ 层的特征,将融合后的信息特征送入检测,用于检测较小的目标。在检测头上应用更高分辨率的特征图后,小目标可以占据更多的像素,因此更容易被检测到,提高了模型对小目标的检测性能。



(a) YOLOv5 特征融合与检测模型



(b) 改进后的 YOLOv5 特征融合与检测模型



(c) 公共部分

图 5 YOLOv5 特征融合与检测模型比较图

Fig. 5 Comparison diagram of YOLOv5 feature fusion and detection model

3 实验与结果分析

3.1 实验环境及参数设置

本实验使用的操作系统为 Ubuntu 18.04 LTS,GPU 为 NVIDIA GeForce RTX 1080Ti,显存 11 G,CUDA 为 10.1,深度学习框架为 Pytorch1.10。本实验选择 YOLOv5s 作为预训练模型,训练参数如下:输入尺寸为 640×640,*epoch* 设置 300,*batch_size* 为 16,其余超参数和改进前的设置相同。

3.2 数据集及评价指标

本实验所用的数据集是 VisDrone2019。该数据集包含了各种无人机捕获的图片,地点涵盖了中国 14 个不同城市,环境包括城市和乡村,捕获的目标种类丰富,主要有 10 个类别:行人、人、轿车、货车、公共汽车、卡车、摩托车、自行车、遮阳篷三轮车和三轮车。本实验使用 VisDrone2019-DET-train、共 6 471 张图片作为训练集,VisDrone2019-DET-val,共 548 张图片作为验证集,其各类标签数量见表 1。

表 1 VisDrone2019 各类标签数量

Table 1 Amounts of various labels in VisDrone2019 dataset

类别	数量
pedestrian	79 055
people	26 962
bicycle	10 389
car	144 620
van	24 899
trunk	12 875
tricycle	4 812
awning-tricycle	3 245
bus	5 917
motor	29 618

本实验采用平均精度(Average Precision,*AP*)和平均精度均值(mean Average Precision,*mAP*)作为评价指标,包括 $mAP_{50}/\%$、$mAP_{50:95}/\%$,其中 $mAP_{50}/\%$ 表示所有目标类别的 *IoU* 阈值在 0.5 时的平均检测精度,可以反映算法对不同类别的检测精度;$mAP_{50:95}/\%$ 代表以步长为 0.05,计算 *IoU* 阈值从 0.50 ~ 0.95 的所有 10 个 *IoU* 阈值下的检测精度的平均值。计算公式见如下:

$$P=\frac{TP}{TP+FP} \tag{3}$$

$$R=\frac{TP}{TP+FN} \tag{4}$$

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

$$mAP=\frac{\sum_{i=1}^{k} AP_i}{k} \tag{6}$$

其中,*TP* 表示被模型预测为正类的正样本;*FN* 表示被模型预测为负类的负样本;*FP* 表示被模型预测为正类的负样本;*TN* 表示被模型预测为负类的正样本;k 表示类别总数量,这里 k 为 10。

3.3　结果分析

为了验证所提出 C3R 模块的有效性，分别将其加入骨干网络中的不同部分进行实验，实验结果见表 2。

表 2　C3R 对比实验

Table 2　Comparative trial of C3R

位置/层	mAP_{50}/%	$mAP_{50:95}$/%	参数量	计算量
Original	34.2	18.8	7 037 095	15.8
2	34.5	19.0	7 034 805	15.7
4	33.5	18.5	7 018 539	15.6
6	34.2	18.8	6 925 051	15.5
8	34.6	19.1	6 887 231	15.7
2、4	33.6	18.6	7 016 249	15.5
2、6	33.9	18.7	6 922 761	15.4
2、8	34.6	18.8	6 884 941	15.6
4、6	33.8	18.6	6 906 495	15.2
4、8	34.2	18.7	6 868 675	15.5
6、8	34.5	18.8	6 775 187	15.4
2、4、6	33.6	18.4	6 904 205	15.1
2、4、8	34.2	18.6	6 866 385	15.4
2、6、8	34.5	18.9	6 772 897	15.2
4、6、8	34.1	18.8	6 756 631	15.1
2、4、6、8	33.9	18.5	6 754 341	15.0

由表 2 可知，C3R 处于不同的骨干网络层，其结果也不同，但共同点都是减少了参数量以及计算量。将其加入骨干网络第 8 层，综合效果最佳，在减少了参数量和计算量的同时，还提升了平均精度，所以本文选择将其加入骨干网络第 8 层。

最终本文通过改进后的 YOLOv5s 模型，在 VisDrone2019 数据集上进行消融实验，来验证改进模型的优越性。首先使用提出的具有多尺度表示特征能力的 C3R 模块，其次引入 ASFF 模块，替换原先的 Detect，最后改进多尺度融合与检测，即增加 P2 层的融合与检测，得到最终的算法模型，与原始模型进行对比，实验结果见表 3。

表 3　消融实验结果

Table 3　Results of ablation experiment

方法	mAP_{50}/%	$mAP_{50:95}$/%
YOLOv5s	34.2	18.8
YOLOv5s+C3R	34.6	19.1
YOLOv5s+C3R+ASFF	35.0	19.4
YOLOv5s+C3R+ASFF+4detector	40.7	23.2

由表 3 实验结果可以看出，相比原始 YOLOv5s 模型，在使用了 C3R 模块后，mAP_{50}/% 提升 0.4%，$mAP_{50:95}$/% 提升了 0.3%，可见提出的模块的有效性；在此基础上引入 ASFF 模块后，mAP_{50}/% 提升了 0.4%，$mAP_{50:95}$/% 提升了 0.3%，证明了引入 ASFF 模块的有效性；最后改进多尺度融合与检测，mAP_{50}/% 提升 5.7%，$mAP_{50:95}$/% 提升 3.8%，证明了改进多尺度融合与检测对提升精度的有效性。最终改进后的模型比原模型 mAP_{50}/% 提升了 6.5%，$mAP_{50:95}$/% 提升了 4.4%。

基于 VisDrone2019 数据集，得到的每个类别的 *AP* 见表 4。由表 4 可以看到 10 个类别的平均精度不同程度地得到了提升，尤其是 pedestrian、car、bus 类提升较为显著，也有如 bicycle 和 awning-tricycle 这 2 类的提升不明显，因为这 2 类的特征较小，信息更难捕捉，特征信息提取难度过大，但就总体而言 mAP_{50}/% 和 $mAP_{50:95}$/% 得到了提升，说明模型性能得到了改善。

表 4　模型结果比较

Table 4　Comparison of results

类别	原模型 *AP*	改进模型 *AP*
pedestrian	39.1	47.5
people	31.1	36.4
bicycle	12.2	16.4
car	71.5	79.3
van	36.0	43.6
trunk	32.7	38.2
tricycle	21.5	29.3
awning-tricycle	12.5	14.8
bus	45.7	55.2
motor	39.3	46.2
All	34.2	40.7

将本文改进的模型与 VisDrone2019 榜单上提交的一些算法及其他经典目标检测最新算法进行对比，结果见表 5。

表 5　不同算法在 VisDrone2019 数据集的检测结果比较

Table 5　Comparison of detection results of different algorithms in VisDrone2019 dataset

方法	mAP_{50}/%
YOLOv5s	34.20
RetinaNet	28.70
PP-YOLOE+	34.63
DRONE-YOLO	36.14
CBNetv2	36.98
Drone-DINO	37.14
Faster-RCNN	33.20
TPH-YOLOv5	41.50
Deformable DETR	43.10
本文模型	40.70

由表 5 可知，本文改进的模型不仅比原先的

YOLOv5 模型精度更高,也比排行榜上的一些算法检测精度更有优势,在 $mAP_{50}/\%$ 这项指标上胜过别的算法,证明了本文改进模型的优越性。

本文对 VisDrone2019 数据集上的测试结果进行了可视化展示,为了更直观地感受改进模型的优越性,如图 6 所示。

由图 6 可以看出,本文算法的检测效果要更好,其检测框与目标贴合更为紧密,同时能够检测出更多的目标,显著降低了目标的漏检率,证明了本文算法的优越性。

(a) YOLOv5s 检测结果

(b) 改进算法检测结果

图 6 可视化检测结果图

Fig. 6 The results of the visualization

4 结束语

本文针对无人机拍摄的复杂场景下的目标检测问题,提出了一种基于改进 YOLOv5 的目标检测算法。该算法旨在解决目标尺度变化大、目标图像占比小的情况下检测精度低的问题,以 YOLOv5s 为基础,使用提出的 C3R 模块对目标信息进行更好的多尺度特征表示;引入 ASFF 模块,自适应地对其他特征级别进行空间滤波,保留信息以进行融合,达到了自适应融合目标特征的目的;改进 YOLOv5 特征融合与检测模型,使算法更好地获取小目标的信息,有效提高小目标检测的精度。在 VisDrone2019 数据集上,检测精度达到了 40.7%,比原 YOLOv5 提升了 6.5%,减少了目标漏误检。与其他目标检测算法相比,本文改进的算法更具优越性,但是整体检测精度还是不高,且计算量较大,仍亟待后续改进。未来将

从更多角度去尽可能地提取更多的目标特征信息来达到提升检测精度、同时减少计算量的目的。

参考文献

[1] BHASKARANAND M, GIBSON J D. Low-complexity video encoding for UAV reconnaissance and surveillance[C]// 2011 Military Communications Conference. Piscataway, NJ: IEEE, 2011:1633-1638.

[2] HUANG Chong, YANG Zhenyu, KONG Yan, et al. Through-the-lens drone filming[C]// Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2018: 4692-4699.

[3] LINNA P, HALLA A, NARRA N. Ground-penetrating radar-mounted drones in agriculture[C]// Proceedings of the New Developments and Environmental Applications of Drones. Cham: Springer, 2022:139-156.

[4] VELUSAMY P, RAJENDRAN S, MAHENDRAN R K, et al. Unmanned Aerial Vehicles (UAV) in precision agriculture: Applications and challenges[J]. Energies, 2022, 15: 217.

[5] LYGOURAS E, SANTAVAS N, TAITZOGLOU A, et al. Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations[J]. Sensors, 2019, 19:3542.

[6] ALMAGBILE A. Estimation of crowd density from UAVs images based on corner detection procedures and clustering analysis[J]. Geo-Spatial Information Science, 2019, 22: 23-34.

[7] WU Xiongwei, SAHOO D, HOI S C H. Recent advances in deep learning for object detection[J]. Neurocomputing, 2020, 396:39-64.

[8] REN Shaoqing, HE Kaiming, 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.

[9] 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. Piscataway, NJ: IEEE, 2016: 779-788.

[10] JIANG Zicong, ZHAO Liquan, LI Shuaiyang, et al. Real-time object detection method based on improved YOLOv4-tiny[J]. arXiv preprint arXiv, 2011.04244, 2020.

[11] SHAROZE A, SIDDIQUE A, ATEŞ H F, et al. Improved YOLOv4 for aerial object detection[C]//2021 29th Signal Processing and Communications Applications Conference (SIU). Piscataway, NJ: IEEE, 2021:1-4.

[12] 奉志强，谢志军，包正伟，等. 基于改进 YOLOv5 的无人机实时密集小目标检测算法[J]. 航空学报，2023，44(7)：251-265.

[13] TAN Mingxing, PANG Ruoming, LE Q V. EfficientDet: Scalable and efficient object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10781-10790.

[14] BAI Haoyue, WEN Song, CHAN S H G. Crowd counting on images with scale variation and isolated clusters[C] // 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway, NJ: IEEE, 2019:18-27.

[15] HERMAN J R, BERGEN J R, PELEG S, et al. Method and apparatus for mosaic image construction [P]. USA: CA2261128A1, 2000-06-13.

[16] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117-2125.

[17] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8759-8768.

[18] NEUBECK A, GOOL V L. Efficient non-maximum suppression [C] // Proceedings of the 18th International Conference on Pattern Recognition (ICPR06). Piscataway, NJ: IEEE, 2006, 3: 850-855.

[19] LIU Songtao, HUANG Di, WANG Yunhong. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv, 1911.09516, 2019.

[20] GAO Shanghua, CHENG Mingming, ZHAO Kai, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2):652-662.