文章编号：2095-2163(2022)12-0070-05 中图分类号：TP391 文献标志码：A

基于目标检测模型的无人机影像识别技术

孙盼盼[1]，丁学文[1,3]，常黎玫[1]，蔡鑫楠[1]，董国军[2]
（1 天津职业技术师范大学 电子工程学院，天津 300222；2 天津市高速铁路无线通信企业重点实验室，天津 300350；
3 天津云智通科技有限公司，天津 300350）

摘 要：YOLOv5 具有较高的目标检测速度和检测精度，但在无人机影像小目标检测方面效果不太好。为解决在自然环境情况下小目标检测精度低及鲁棒性差等问题，本文以自然环境情况下无人机影像为研究对象，提出了一种改进的 YOLOv5 小目标检测模型。通过对特征图增加上采样处理，使特征图继续扩大，从而降低采样率和缩小感受野，提高模型对小目标的检测能力。改进的模型在天大无人机影像 VisDrone 数据集上进行了训练和测试。实验结果表明，改进 YOLOv5 的算法平均精度值为 46.4%，与原 YOLOv5 模型相比，平均精度值提升了 14.9%，改进 YOLOv5 在一定程度上改善了 YOLOv5 无人机影像识别率。

关键词：YOLOv5；无人机影像；上采样；改进算法；平均精度值

UAV images recognition technology based on target detection model

SUN Panpan[1], DING Xuewen[1,3], CHANG Limei[1], CAI Xinnan[1], DONG Guojun[2]
(1 School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China;
2 Tianjin High Speed Railway Wireless Communication Enterprise Key Laboratory, Tianjin 300350, China;
3 Tianjin Yunzhitong Technology Co., Ltd., Tianjin 300350, China)

【Abstract】 YOLOv5 has high target detection speed and detection accuracy, but it is not very effective in detecting small targets in UAV images. In order to solve the problems of low detection accuracy and poor robustness of small targets in natural environment, this paper takes UAV images in natural environment as the research object, and proposes an improved YOLOv5 small target detection model. By adding up-sampling processing to the feature map, the feature map continues to expand, thereby reducing the sampling rate and the receptive field, and improving the model's ability to detect small targets. The improved model is trained and tested on the VisDrone dataset of UAV images. The experimental results show that the average accuracy of the improved YOLOv5 algorithm is 46.4%. Compared with the original YOLOv5 model, the average accuracy is increased by 14.9%. The improved YOLOv5 can improve the YOLOv5 UAV images recognition rate to a certain extent.

【Key words】 YOLOv5; UAV images; upsampling; improved algorithm; average precision value

0 引 言

目标检测作为计算机视觉的研究热点之一，引起了各国学者的关注。近年来深度学习的目标检测算法得到了快速发展，识别精度和速度也在不断提升。基于深度学习的目标检测算法分为 2 类：一阶段和两阶段。其中，SSD[1-2] 和 YOLO（You Only Look Once）系列[3-6]是一阶段检测算法，R-CNN[7]、Fast R-CNN[8]和 Faster R-CNN[9]是两阶段检测算法。与两阶段识别相比，一阶段识别准确率略低，但识别速度要快上数百倍。在单阶段算法中，YOLOv5 比 SSD 快 2~3 倍，所以 YOLOv5 在开发人员中更受欢迎。目前，YOLOv5 已然广泛应用在对实时性要求较高的各种目标识别领域中。虽然 YOLOv5 具有良好的目标检测性能，但对无人机影像这类小目标的识别率却较低。与其他目标相比，容易发生漏检和误检，这在一定程度上限制了 YOLOv5 的使用。在实际应用场景中，会有相当多的对象都是小目标，小目标在图像中面积小、特征也不明显，采用多层卷积神经网络后，可能出现部分特征丢失的问题，从而导致识别率的下降。针对以上问题，本文提出改进的 YOLOv5 目标检测算法，该算法增加了有利于小目标的处理，从而提高了精确率、召回率和平均精确率。

1 YOLOv5 算法及改进

1.1 YOLOv5 网络结构

YOLOv5 按照网络深度和网络宽度的大小，可

基金项目：天津市科委科技特派员项目（20YDTPJC01110）。
作者简介：孙盼盼（1997-），女，硕士研究生，主要研究方向：智能视觉。
通讯作者：丁学文 Email：dingxw1@126.com
收稿日期：2022-03-25

以分为 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x。YOLOv5s 的网络结构最为小巧，同时图像推理速度最快达 0.007 s，故本文使用 YOLOv5s 模型。YOLOv5s 的网络结构如图 1 所示。由图 1 可知，YOLOv5s 的网络结构主要由输入端、基准网络、Neck 网络以及 Head 输出端四部分组成。对此将展开研究分述如下。

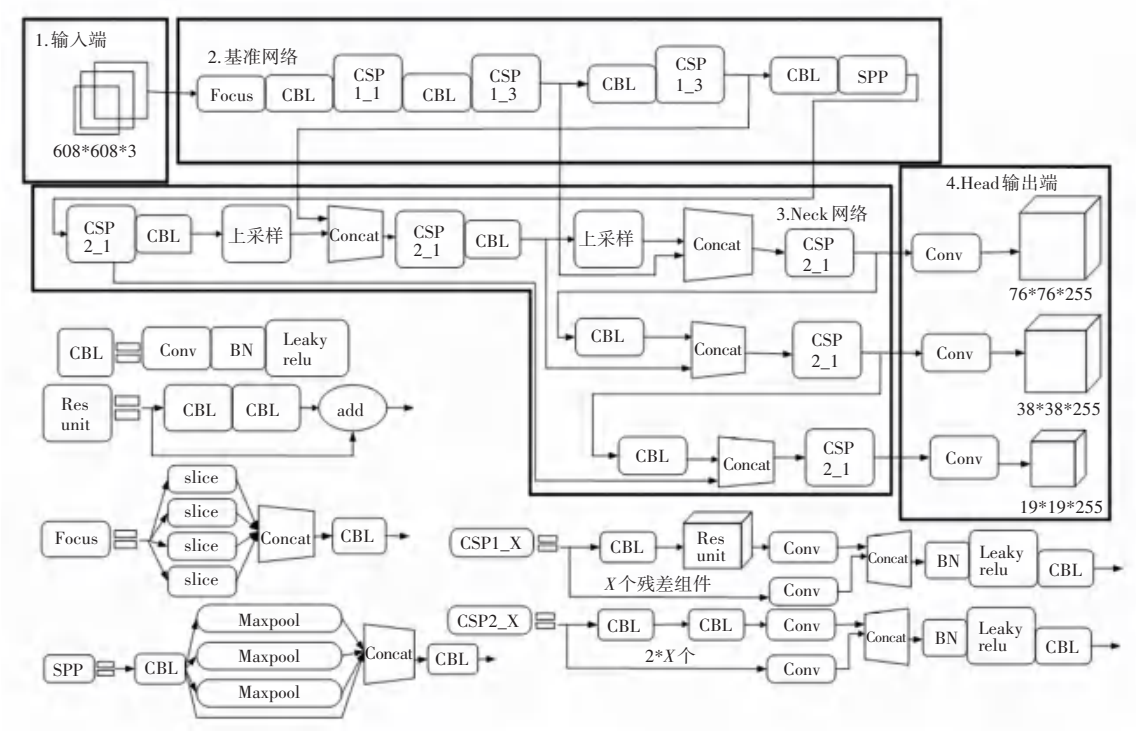


图 1　YOLOv5s 的网络结构

Fig. 1　Network structure of YOLOv5s

（1）输入端。表示输入的图片的部分。YOLOv5s 输入大小为 608 * 608 的图像，该阶段会把输入图像进行缩放，直至与网络的输入大小相等，再对图像进行归一化处理。在网络训练阶段，YOLOv5s 为了提高网络模型的训练速度和网络精度，在网络模型中增加了 Mosaic 数据增强操作；为了使数据集多样化以及减少 GPU 的占用，在网络模型中增加了自适应锚框的计算以及自适应图片缩放。

（2）基准网络。通常是提取一些通用的特征。YOLOv5s 中使用了 CSPDarknet53 和 Focus 网络结构作为基准网络，CSP 结构是用来进行下采样的，但和传统卷积的下采样不太相同，CSP 结构可以对 Focus 的计算量和普通卷积的下采样计算量进行比较。

（3）Neck 网络。通常位于基准网络和输入端之间的位置，利用 Neck 网络可以使提取的特征具有多样性及更好的稳定性。YOLOv5s 用到了 FPN+PAN 模块，FPN 层是自顶向下的特征卷积用于传达强语义特征，而特征金字塔是自底向上的特征卷积用于传达强定位特征，两两联合，从不同的主干层对不同的检测层进行参数聚合，进而达到很好的效果。

（4）Head 输出端。用来完成目标检测结果的输出。YOLOv5s 主要是 *GIoU_Loss* 代替 *IoU* 作为 bounding box 回归的损失，*IoU* 的缺点是不重合或者重合面积相等，而 YOLOv5s 的 *GIoU* 在计算时，不同位置的预测框都会对 *GIoU* 产生影响，从而弥补了 *IoU* 的不足，并进一步提升算法的检测精度。

1.2　YOLOv5 效果展示

不同版本的 YOLOv5 检测算法在 COCO2017 验证集与测试集上的检测效果如图 2 所示。

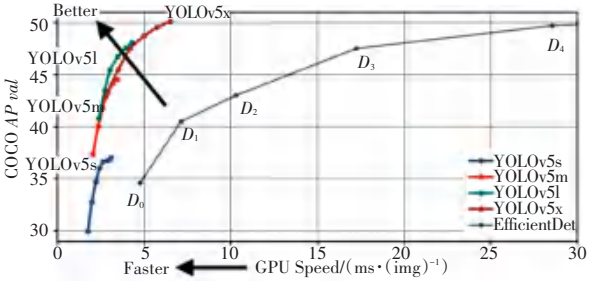


图 2　YOLOv5 效果展示

Fig. 2　YOLOv5 effect display

图 2 中,横轴表示 YOLOv5 算法在 GPU 上推理出每张图片需要的毫秒数,距原点越近、效果越好;纵轴表示 YOLOv5 算法在 COCO2017 的测试集上测试出的 *AP* 值,距离原点越远、效果越好。通过观察分析可以看出,相较于 EfficientDet,YOLOv5s 的 *AP* 值更高,而且推理速度更快;相较于 YOLOv5m、YOLOv5l、YOLOv5x,YOLOv5s 具有更高的速度,但 *AP* 值并不高,不过也在可接受范围内。

YOLOv5 的整体效果展示见表 1。由表 1 可知,YOLOv5s 的输入图片分辨率为 640 * 640,在 COCO 测试集与验证集上的 *AP* 指标为 36.8,AP50 指标为 55.6。该算法在 V100 GPU 上的推理速度仅仅需要 2.2 ms,帧率为 455 FPS,该网络的模型大小仅为 7.3 M。相对 YOLOv5m、YOLOv5l 及 YOLOv5x 模型来说,YOLOv5s 的速度更快、模型较小、且精度也较高。故而,在本文中拟选择 YOLOv5s 模型进行研究。

表 1 YOLOv5 效果展示

Tab. 1 YOLOv5 effect display

Model	*size*	*APval*	*APtest*	*AP*50	*Speedv*100 /ms	*FPSv*100	*params* /M	*GFLOPS*
YOLOv5s	640	36.8	36.8	55.6	2.2	455	7.3	17.0
YOLOv5m	640	44.5	44.5	63.1	2.9	345	21.4	51.3
YOLOv5l	640	48.1	48.1	66.4	3.8	264	47.0	115.4
YOLOv5x	640	50.1	50.1	68.7	6.0	167	87.7	218.8

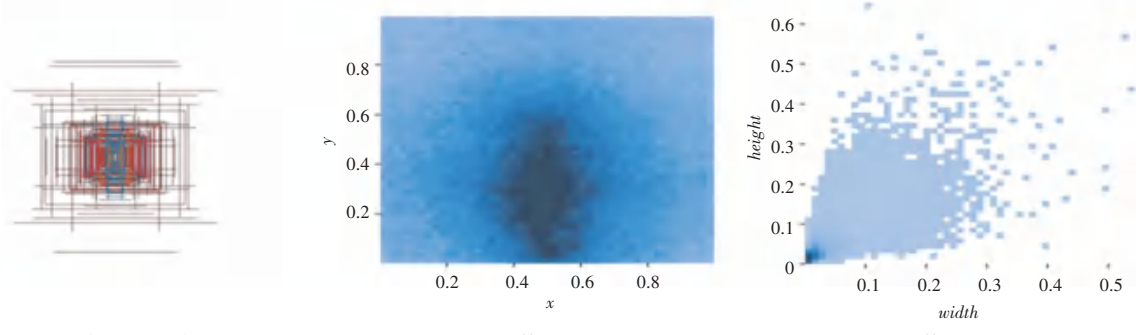
1.3 YOLOv5s 算法的改进

考虑到 YOLOv5s 对传统目标检测较好,但对小目标经常出现误检、漏检,从而造成精度较低的问题,本文提出了改进的 YOLOv5s。改进的 YOLOv5s 主要是在第 17 层后,对特征图增加上采样操作,使特征图继续扩大,如此一来就改善了小目标浅层语义信息不足的缺陷。

2 实验及结果分析

2.1 数据集的相关说明

本文实验采用无人机影像 VisDrone 数据集。VisDrone 数据集由中国天津大学机器学习和数据挖掘实验室的 AISKYEYE 团队收集并且标注的[10]。该数据集在采集时把摄像机架设在无人机上,在中国 14 个不同地区的城市和村庄以及不同的天气和光照下,采集稀疏程度不同的行人、小汽车、三轮车、自行车等不同的物体。VisDrone 目标检测数据集中包括 pedestrian、people、bicycle、car、van、truck、tricycle、awning-tricycle、bus、motor 共 10 类被标注的物体。其中,pedestrian 为直立姿势或者行走的人,除 pedestrian 以外的人定义为 people。通过对数据集进行分析,得到可视化结果,如图 3 所示。



(a) 标记框尺寸分布 (b) 物体中心点分布位置 (c) 物体大小分布

图 3 数据集可视化结果

Fig. 3 Visualization results of the dataset

由图 3(a)中可以看出,图像中的大多数都是较小的标记框。由图 3(b)中可以看出,物体中心点位置在 x 轴方向大多数分布在 0.4~0.6 之间,在 y 轴方向大多数分布在 0~0.6 之间。在图 3(c)中,横坐标表示物体的宽,纵坐标表示物体的高。综合图 3(b)和图 3(c)的分析可知,该数据集中小物体较多,并且存在一定程度的遮挡。

2.2 实验环境参数

本文实验采用的电脑硬件配置及 Pycharm 软件设置情况见表 2。

表 2　实验环境参数

Tab. 2　Experimental environmental parameters

实验环境	配置参数
操作系统	Ubuntu 16.04　LTS
GPU	GeForce RTX2070 SUPER
深度学习框架	Pytorch 1.7.1
GPU 加速库	CUDA 10.0.130　CUDANN 7.6.5
CPU	Intel® coreTM　i5-10400F

2.3　模型评价指标

预测值为正例，记为 P（Positive）；预测值为反例，记为 N（Negative）；预测值与真实值相同，记为 T（True）；预测值与真实值相反，记为 F（False）。改进的 YOLOv5s 采用平均精度（mean average precision，*mAP*）来验证所提模型相较于 YOLOv5s 模型的优越性。*mAP* 在计算时需用到 *Precision*、*Recall*、*AP*，对此可做阐释表述如下。

（1）精度。具体计算公式为：

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{1}$$

（2）召回率（*Recall*）。具体计算公式为：

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{2}$$

（3）*AP* 和 *mAP*。具体计算公式为：

$$mAP = \int_0^1 Precision \times Recall \, \mathrm{d}r \tag{3}$$

2.4　不同模型检测精度对比

本实验选用数据集中的 6 471 张图片作为训练集，548 张图片作为验证集训练 300 次，YOLOv5s 和改进 YOLOv5s 的 VisDrone 数据集的评估结果见表 3。

表 3　VisDrone 数据集结果评估

Tab. 3　Results evaluation of VisDrone data set

目标种类	YOLOv5s-*AP*	YOLOv5s-*MAP*	改进 YOLOv5s-*AP*	改进 YOLOv5s-*MAP*
prdestrian	0.353	0.315	0.545	0.464
people	0.276	0.315	0.419	0.464
bicycle	0.095	0.315	0.231	0.464
car	0.714	0.315	0.840	0.464
van	0.361	0.315	0.493	0.464
truck	0.295	0.315	0.437	0.464
tricycle	0.185	0.315	0.356	0.464
awning-tricycle	0.097	0.315	0.168	0.464
bus	0.418	0.315	0.624	0.464
motor	0.359	0.315	0.523	0.464

从仿真实验结果可以看出，改进 YOLOv5s 各个类别的 *AP* 值都有 10%～15%的提升，*mAP* 值提升了14.9%，由此可见改进的 YOLOv5s 确实对小目标有了很好的改善。

训练结束后，本文采用无人机重新捕获图片进行测试，运行的效果如图 4 所示。

（a）YOLOv5s　　（b）改进 YOLOv5s

图 4　测试结果

Fig. 4　Test results

为方便查看无人机影像的检测结果，从图像中选取图 4(a)的局部区域①、②、③，如图 5(a)～(c)所示，选取图 4(b)的局部区域①、②、③，如图 6(a)～(c)所示。

图 5(a)中把井盖误检为 bicycle，图 6(a)中此井盖没有被认为是标签中的物体；图 5(b)中漏检多辆被树木遮挡的 car，图 6(b)中被树木遮挡的 car 均被正确检出；图 5(c)中把 tricycle 误检为 car，漏检 pedestrian 和 people，car 的概率为 39%；图 6(c)改进 YOLOv5 测试结果中将该 car 的概率提升为 72%，pedestrian 和 people 均被正确检出，但此图却把 tricycle 误检为 motor。因此改进的 YOLOv5s 改善了漏检、误检以及检测效果不佳的问题，也仍有待进一步扩充数据集，并且进行更多训练来优化模型。总之，改进的 YOLOv5s 算法在小目标检测方面已经具有较好的检测性能。

(a) 测试结果 1 (b) 测试结果 2 (c) 测试结果 3

图 5 YOLOv5s 测试结果

Fig. 5 YOLOv5s test results

(a) 测试结果 1 (b) 测试结果 2 (c) 测试结果 3

图 6 改进 YOLOv5s 测试结果

Fig. 6 Improved YOLOv5s test results

3 结束语

针对自然环境使用 YOLOv5s 检测无人机影像时出现的漏检、误检以及检测效果欠佳等问题,本文提出了一种基于 YOLOv5s 模型改进的无人机影像检测模型。研究中,在 17 层后增加上采样模块,来弥补浅层特征语义信息的不足,从而提高了模型的特征提取能力,模型的检测精度也得以提升。改进后的 YOLOv5s 与原网络的无人机影像检测模型对比,获得了很好的检测结果,然而整体的平均精度稍微偏低。在今后的研究当中,将会在这一方向做更加深入的探讨,以利于有效提升最终效果。

参考文献

[1] LIU Wei, ANGUELOV D, ERHAN D , et al. SSD: Single shot multibox detector[C]//Europen Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[2] 孙彦,丁学文,雷雨婷,等. 基于 SSD_MobileNet_v1 网络的猫狗图像识别[J]. 天津职业技术师范大学报,2020,30(01):38-44.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:779-788.

[4] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, Hawaii:IEEE, 2017:6517-6525.

[5] REDMON J, FARHADI A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv: 1804.02767,2018.

[6] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimalspeed and accuracy of object TEtection[J]. arXiv preprint arXiv:2004.10934, 2020.

[7] 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:IEEE, 2014:580-587.

[8] GIRSHICK R. Fast R-CNN[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile:IEEE, 2015:1440-1448.

[9] 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.

[10] ZHU Pengfei, WEN Longyin, BIAN Xiao, et al. Vision meets drones: A challenge[J]. arXiv preprint arXiv:1804.07437,2018.