基于 Cascade R-CNN 的并行特征金字塔网络无人机航拍图像目标检测算法

刘英杰，杨风暴*，胡鹏
中北大学信息与通信工程学院，山西 太原 030051

摘要 在目标检测领域，小目标的检测识别一直都是研究的难点，导致模型提取到的特征并不具有良好的表达能力，因此对小目标的检测结果不佳。为此，提出一种基于特征金字塔网络(FPN)的改进算法。在原有基础上增加并行分支，再融合两种不同上采样方法的特征信息以加强小目标特征的表达能力。同时，增加多阈值检测器(Cascade R-CNN)强化小目标定位能力。基于无人机航拍数据集进行实验，实验结果表明，在 MS COCO 数据集下，所提算法的平均精确率相比原始 FPN 算法提高了 9.7 个百分点，具有良好的检测性能。

关键词 机器视觉；目标检测；深度学习；特征金字塔；级联网络
中图分类号 TP751 **文献标志码** A **doi**：10.3788/LOP57.201505

Parallel FPN Algorithm Based on Cascade R-CNN for Object Detection from UAV Aerial Images

Liu Yingjie, Yang Fengbao*, Hu Peng
School of Information and Communication Engineering, North University of China, Taiyuan, Shanxi 030051, *China*

Abstract The detection and recognition of small targets are always difficult for researchers in the field of target detection, resulting in the feature extracted from the model not having good expression ability, so the detection result of small targets is poor. This paper presents a modified algorithm based on feature pyramid network(FPN). Specifically, the parallel branch is devised on the original basis to fuse the feature information of two different up-sampling methods to enhance the representation ability of small objects. Meanwhile, a multiple threshold detector named Cascade R-CNN is added to prompt the localization ability of small objects. Experiments are conducted on UAV aerial image datasets. The experimental results reveal that the average precision of the proposed algorithm under MS COCO dataset increases by 9.7 percentage compared to that of the initial FPN algorithm; hence, the proposed algorithm has a good detection performance.

Key words machine vision; object detection; deep learning; feature pyramid; cascade network
OCIS codes 150.0155; 100.4996; 100.2000

1 引 言

无人机(UAV)最早出现于 20 世纪 20 年代，主要用于军事领域，具有体积小、使用方便、战场生存能力强等诸多优点[1]。近年来，随着无人机技术逐渐成熟，制造成本和使用门槛不断降低，民用无人机市场得以迅速发展[1]。同时，得益于人工智能和计算机视觉技术的飞速进步，智能无人机在商业、农业及工业等领域开始发挥着越来越重要的作用。

伴随着计算机视觉领域的多方面突破，目标检测技术取得了突飞猛进的进步。传统检测方法通常需先对滑动窗口提取手工特征，再利用分类器对窗口进行分类，然而这些手工设计的特征需要大量的先验知识，且泛化能力不足，其结果是通用性严重受限，如 SIFT 方法[2]、HOG 方法[3]和 DPM 方法[4]等。有别于传统检测方法，基于深度学习的方法[5-9]

收稿日期：2019-12-10；**修回日期**：2020-01-13；**录用日期**：2020-02-25
基金项目：山西省研究生教育创新项目(2019BY108)
* **E-mail**：yfengb@163.com

则是从大量数据中自动学习特征，特征的泛化能力得到显著提升。此外，这类方法能充分发挥大数据的优势，进而大幅提高了目标的检测精度。然而，在无人机航拍图像中，基于深度学习的目标检测方法（如 Faster R-CNN、SSD 等）尚且面临严峻的挑战。以广泛使用的 Faster R-CNN[10-11] 为例，其只利用最后一层（即高层特征）进行目标预测，没有充分考虑其他层的特征信息，导致其对小目标的检测能力明显不足。为了解决该问题，从不同层抽取不同尺度特征进行预测的 SSD 算法[12-13] 被相继提出，但其从较高的层开始预测，依然没有兼顾低层特征，所以 SSD 算法对小目标的检测能力仍不够理想。

对小目标而言，随着层数的增加，特征出现弥散，结果是高层特征所提供的目标信息往往过于单一（仅有大目标特征），由此可见，低层特征信息在小目标检测过程中尤为重要。为了解决小目标检测的难点，Lin 等[14] 在 2017 年提出一种新型的检测模型，即特征金字塔网络（FPN），该模型在 Faster R-CNN 的基础上加入上采样结构，并融合高、低层特征图之间的特征信息，大幅提升了小目标的检测精度。然而，无人机航拍图像中目标尺寸偏小、分布密集且背景复杂，原始 FPN 提取到的目标特征并不具有良好的区分度。因此，如何加强待检测目标特征信息的表征能力是提高检测精度的关键所在。

为此，本文针对 FPN 的不足提出新的改进算法。在 FPN 算法的基础上使用不同的上采样方式，增加一个并行金字塔结构，之后对两个金字塔层进行特征融合，丰富特征信息，加强目标的特征表达能力；在原有基础上引入定位精度优异的 Cascade R-CNN 算法[15]，以进一步提升模型的精确定位能力；设置合适的预设框尺寸。所提改进算法有效提高了无人机航拍图像目标检测的准确性和鲁棒性。

2 无人机航拍图像特点分析

图 1 为常规图像数据集 MS COCO[16] 与无人机航拍数据集 VisDrone[17] 中不同尺寸目标数量的占比，横坐标代表目标的像素尺寸，纵坐标代表目标数量的占比。从图 1 不难发现：在常规数据集 MS COCO 中，各种尺寸的目标分布 A 较为均匀，尽管小目标数量最大，但其占比仅为 41%；而在无人机航拍数据集 VisDrone 中，小目标占比达到了 62%。由此可见，无人机航拍图像中的目标尺寸整体偏小。

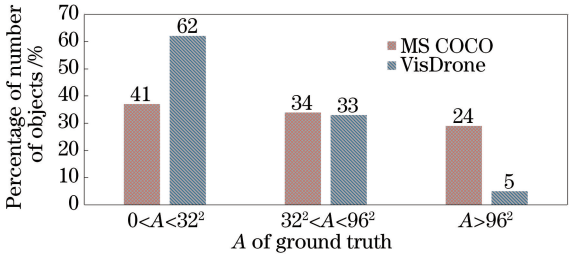

图 1 目标面积与目标数量的关系

Fig. 1 Object area versus number of objects

图 2 给出了 VisDrone 及 MS COCO 数据集的目标数量分布情况，可以看出，VisDrone 数据集每张图片中的目标数量平均达到 53.0 个，远大于 MS COCO 数据集每张图片的平均目标数量，但目标数量的增加会加剧目标密集程度，加大检测难度。

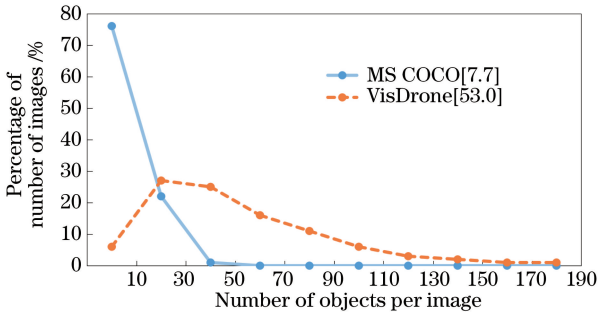

图 2 目标数量与所占图片数量的关系

Fig. 2 Number of objects per image versus percentage of number of images

图 3 展示了一组实际航拍结果，从图 3(a)可以发现，无人机航拍图像中目标分布非常密集，在图 3(b)中，存在树木等遮挡物。

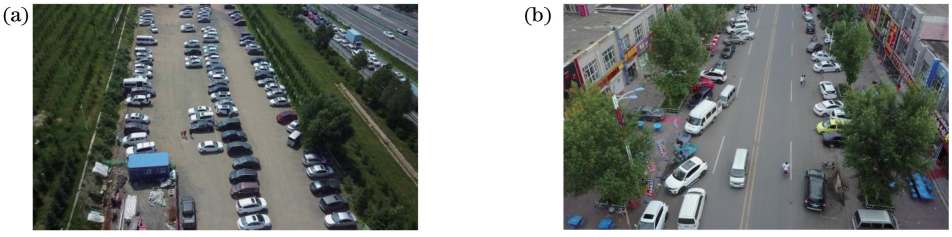

图 3 实际航拍图像实例

Fig. 3 Example of actual aerial images

通过上述分析可知，无人机航拍图像中的待检测目标（如行人、车辆等）通常存在以下特点：目标数量无规律且较为密集；图像幅面大，但目标所占像素较小，存在大量的小目标；航拍角度易发生改变；背景复杂，目标周围容易出现遮挡物。这 4 个方面极大影响了无人机航拍图像的检测精度，因此航拍图像的检测难度要大于普通常规图像，挑战性更大。基于此，本文针对无人机航拍图像中目标分布密集且尺寸过小的问题，在 FPN 的基础上提出一种新的网络框架。该框架包含两个阶段：在一阶段采用并行结构加强目标特征的表达能力；在二阶段使用级联结构 Cascade R-CNN 加强对目标的定位能力。

3 网络结构与算法原理

3.1 FPN 目标检测算法

FPN 作为一个多尺度检测器，可以方便地与各种网络结构结合，如图 4 所示，该网络结构主要由三部分构成，即自底向上（bottom-up）结构、自顶向下（top-down）结构、横向连接（lateral-connection）结构。自底向上结构主要利用特征提取网络进行特征提取，ResNet[18] 可根据特征图大小的变化将流程分为 5 个阶段，选取第 2 到第 5 个阶段中残差块的最后一层特征图记为 $\{C_2、C_3、C_4、C_5\}$，分别对应下采样倍数 $\{4、8、16、32\}$。自顶向下分支使用最近邻上采样法对高层特征图进行 2 倍上采样。横向连接结构利用 1×1 卷积统一特征图通道数为 256。图 4 中利用相加操作融合高层和低层的特征，接着使用 3×3 卷积来减弱混叠效应，得到对应金字塔层 $\{P_2、P_3、P_4、P_5\}$，对 P_5 进行最大值池化操作得到 P_6。P_6 仅用于生成固定尺寸的预设框和区域生成网络（RPN）阶段的预测，不参与第二阶段的分类和回归。各金字塔层 $\{P_2、P_3、P_4、P_5、P_6\}$ 经 3×3 卷积操作后分别生成对应尺度和长宽比的 anchor，共 15 种，之后由 RPN 对生成的 anchor 进行前景、背景框的分类和回归。

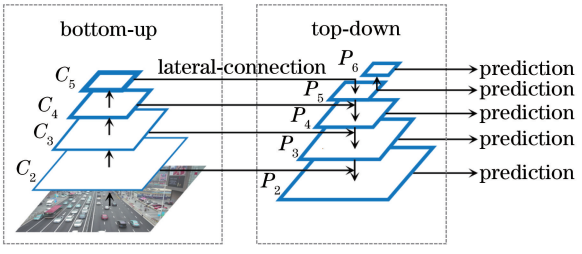

图 4 FPN 框架

Fig. 4 FPN frame

3.2 网络框架与算法原理

3.2.1 并行 FPN

FPN 算法通过融合高低层特征的方式来进行多尺度检测，虽然有良好的检测性能，但经 32 倍下采样后，小目标的语义信息已经出现较大的弥散，而单一的上采样方式难以有效恢复小目标的语义信息。为此，新增加一种上采样方式，利用其带来的不同特征信息的特点，增加特征的多样性，之后融合两种上采样方式的不同特征信息，以达到丰富和增强特征信息的目的。

在原有金字塔层的基础上加入一组新的并行金字塔层，原始金字塔结构使用最近邻上采样法，并行金字塔结构增加双线性插值上采样法，之后利用求和的方式融合所得到的不同特征，以进一步提升对待检测目标特征的表征能力。如图 5 所示，并行金字塔层分别为 $\{P_{c2}、P_{c3}、P_{c4}、P_{c5}\}$，依然沿用自顶向下结构和横向连接结构，之后与原金字塔层 $\{P_2、P_3、P_4、P_5\}$ 进行特征融合，然后取其均值得到平衡后的语义特征，公式为

$$P_i=\frac{1}{2}\mathrm{Conv}_{3\times3}(P_i+P_{ci}), \tag{1}$$

式中：i 为金字塔的层数；$\mathrm{Conv}_{3\times3}(\cdot)$ 为使用 3×3 卷积来减弱混叠效应的函数。最终得到融合后的金字塔层 $\{P_2、P_3、P_4、P_5\}$。

对 P_5 进行最大值池化得到 P_6，接着进行 anchor 的生成并由 RPN 进行分类和回归，之后对感兴趣区域（RoI）进行筛选并选择对应金字塔层中的候选框进行第二阶段精确的分类和回归。关于金字塔层选择的公式为

$$k=\left[k_0+\log_2\left(\sqrt{wh}/224\right)\right], \tag{2}$$

式中：k_0 为基准金字塔层，一般设置为 4；w 和 h 分别为 RoI 的宽和高；k 为所选金字塔层；224 为预训练模型数据集 ImageNet 的图片尺寸，即 224×224。所选金字塔层数的上限和下限设置为 5 和 2，即所选金字塔层为 $\{P_2、P_3、P_4、P_5\}$。在（2）式下，较大尺度的目标选择更高层的金字塔，较小尺度的目标选择低层金字塔，这充分利用了金字塔层各尺度目标的特征信息。

3.2.2 级联网络

原始 FPN 在第二阶段使用的是单阈值检测器，即只有一个正样本 IoU 阈值为 0.5 的检测网络，此网络的弊端在于设置较低的 IoU 阈值会带来大量的冗余框，不利于精确定位和回归，而直接加大 IoU

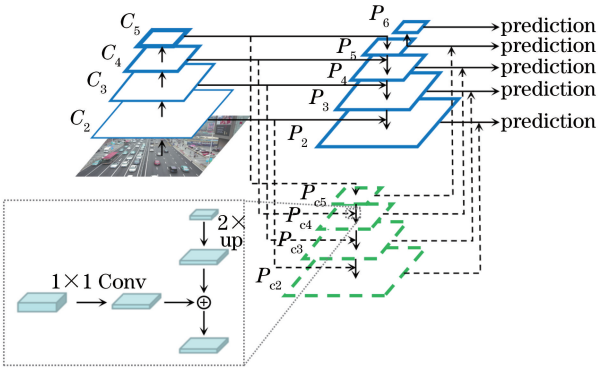

图 5 并行 FPN 框架

Fig. 5 Parallel FPN frame

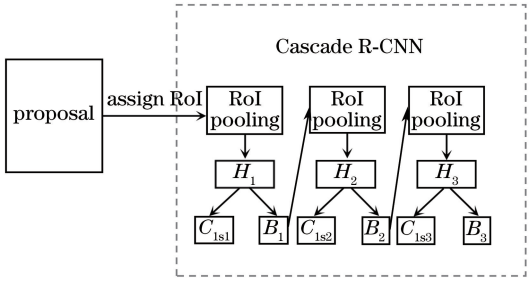

图 6 Cascade R-CNN 结构

Fig. 6 Structure of Cascade R-CNN

阈值会使正负样本比例不均衡，在训练时会出现样本过拟合情况。

针对上述问题，引入级联网络结构 Cascade R-CNN，通过级联多个 IoU 阈值递增的检测器可大幅提高检测器的分类和回归效果，如图 6 所示，$\{H_1、H_2、H_3\}$代表检测器的头部，$\{B_1、B_2、B_3\}$代表回归后的检测框，$\{C_{ls1}、C_{ls2}、C_{ls3}\}$代表检测框的具体分类类别。具体检测过程为：在金字塔层中选择相应的候选框后，通过级联 3 个阈值递增的检测器进行不断的分类和回归，并将上一个检测器回归后的边界框作为下一个检测器的输入，这样可以保证每个检测器的正样本质量稳步提升，从而不断提高检测精度。

图 7 给出了训练过程中级联网络的可视化正样本变化过程，正样本 IoU 阈值分别设置为 0.5，0.6，0.7。在图 7(a)中，正样本 IoU 阈值设为 0.5，不难看出，此阶段虽然有大量的正样本，但同时存在大量的冗余框，这些冗余框不利于对目标进行高质量分类和回归，因此需进一步提升正样本的质量并保证其数量。在图 7(b)中，经过上一阶段的分类和回归过程后，将 IoU 阈值提升到 0.6，这样在剔除冗余框的同时尽可能地保留了高质量正样本的数量。图 7(c)中的 IoU 阈值为 0.7，表明进一步加强了此过程。

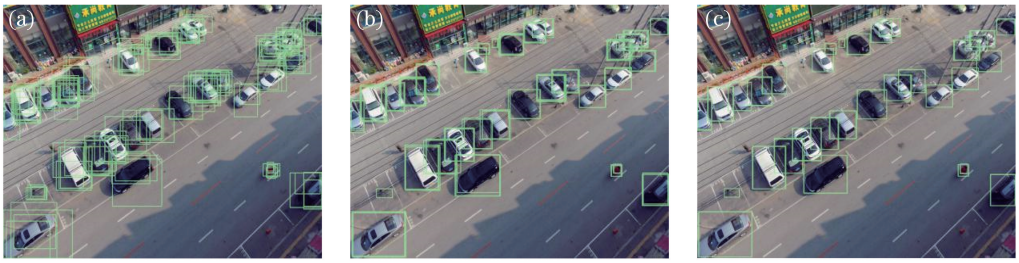

图 7 各阶段下正样本变化情况。(a) IoU 阈值为 0.5；(b) IoU 阈值为 0.6；(c) IoU 阈值为 0.7

Fig. 7 Change of positive proposals in every stage. (a) IoU threshold is 0.5; (b) IoU threshold is 0.6; (c) IoU threshold is 0.7

4 算法流程与实验

4.1 算法流程

模型的详细框架结构如图 8 所示。该模型主要由两部分组成，即并行 FPN 结构和级联网络结构 Cascade R-CNN。检测流程为：由并行 FPN 对输入图片进行特征提取，并在特征图上生成对应 anchor；由 RPN 对各金字塔层特征图中的 anchor 进行第一阶段的预测；第二阶段中，利用 Cascade R-CNN 对一阶段筛选得到的候选框进行高质量的分类和回归，最终得到检测结果。

4.2 实验与分析

4.2.1 数据集及评价指标

使用无人机航拍图像开源公开数据集 VisDrone，该数据集是 2018 年由天津大学机器学习与数据挖掘实验室 AISKYEYE 团队收集的关于航拍无人机图像的数据集，取材地点遍布国内 14 个不同的城市，共包括 4 个任务：基于图像的目标检测；基于视频的目标检测；单目标跟踪；多目标跟踪。基于图像的目标检测共包含 10209 张静态图像(6471 张图片用于训练，548 张图片用于验证，3190 张图片用于测试)和10类常见目标(行人、货车、小汽车、公

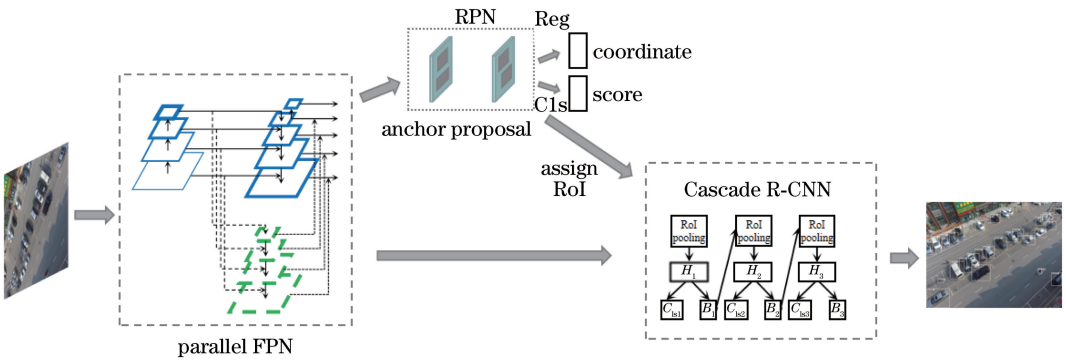

图 8 模型整体框架图

Fig. 8 Overall framework of proposed model

交车、自行车、摩托车等)。

在 MS COCO 数据集中,评价指标主要包括 AP、$AP_{0.5}$、$AP_{0.75}$、AP_S、AP_M、AP_L 等,其中 AP 指标是指 IoU 阈值从 0.5 到 0.95,以 0.05 为间隔,分别检测该阈值下的平均精确率,最后取 10 个值的平均结果;而 $AP_{0.5}$、$AP_{0.75}$ 分别表示单一 IoU 阈值为 0.5 和 IoU 阈值为 0.75 下的平均精确率;AP_S、AP_M、AP_L 分别表示目标的绝对像素面积在小(面积小于 32^2)、中(面积大于 32^2,小于 96^2)、大(面积大于 96^2)下的平均精确率。VisDrone 数据集的指标建立在 MS COCO 数据集的基础上,包括 AP、$AP_{0.5}$、$AP_{0.75}$,代表含义与 MS COCO 数据集指标相同。不同的是,MS COCO 数据集指标的最大检测目标数为 100,而 VisDrone 数据集指标的单张图片最大检测目标数为 500,同时去掉了 AP_S、AP_M、AP_L 指标。

4.2.2 实验环境

实验所需硬件资源主要由云服务平台 OpenBayes 提供。本实验使用一块 NVIDIA Tesla T4 GPU,开发框架 TensorFlow1.12,Python3.6。每个 minibatch 训练一张图片,每张图片有 256 个 anchors,512 个 RoIs,输入图片的短边长为 800 pixel。使用的优化器为 MomentumOptimizer,正则化系数为 0.0001,动量系数为 0.9。学习率设置情况为:首个 2.7×10^4 步为 0.001,接下来的 2.7×10^4 步为 0.0001,最后 1.8×10^4 步为 0.00001。采用随机左右翻转的数据增强方式。采用的预训练模型为 ResNet-101-v1d,该模型是计算机视觉库 GluonCV 中非常优秀的预训练模型,具有优秀的特征提取能力。

4.2.3 实验结果与分析

1)深度学习常用目标检测算法比较。对常用单阶段、二阶段经典目标检测算法及所提算法进行对比实验,采用 MS COCO 数据集的评价指标,结果如表 1 所示。可以看出,所提算法的 AP 值达到 17.5,比使用单一上采样方法(FPN w nearest,FPN w bilinear)的 AP 值分别高 0.5 和 0.3 个百分点。使用单一上采样方式的检测结果均不如融合不同上采样特征金字塔的检测结果,这说明所提算法加强了待检测目标的特征信息。此外,使用 ResNet-101-v1d 预训练模型整体提高了检测精度,使所提算法 AP 值达到 20.6。

表 1 经典算法比较

Table 1 Comparison of classical algorithms

%

Algorithm	Backbone	AP	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L
Retina-Net[19]	ResNet-101	7.1	13.2	7.0	2.9	12.4	17.6
Faster R-CNN		4.6	10.1	3.7	2.7	7.1	7.8
R-FCN[20]		7.9	16.6	6.5	4.5	12.4	18.1
FPN w nearest		17.0	37.3	13.6	11.4	25.6	27.7
FPN w bilinear		17.2	37.4	14.0	11.5	25.9	28.0
Proposed algorithm		17.5	37.5	14.5	11.5	26.2	28.4
FPN w nearest	ResNet-101-v1d	20.1	40.9	17.7	14.0	29.1	32.5
FPN w bilinear		20.3	41.0	18.1	14.0	29.3	32.9
Proposed algorithm		20.6	41.1	18.7	14.1	29.6	33.3

2)级联结构选取。在引入 Cascade R-CNN 级联结构后,采用 MS COCO 数据集的评价指标对所提算法进行验证,如表 2 所示。可以看出:未使用级联结构时,AP 值为 20.6;加入两组级联结构后,AP 值为 25.6;而在加入三组级联结构后,AP 值达到 26.1,虽然 $AP_{0.5}$ 略微降低(从 46.6%下降至 46.4%),但整体精度明显提高;然而,在增加第四组级联结构时,检测结果出现倒退,AP 值下降 0.4 个百分点(从 26.1%下降至 25.7%),$AP_{0.5}$ 下降 0.2 个百分点,$AP_{0.75}$ 下降 0.4 个百分点。这说明在加入第四组级联结构后,IoU 阈值设置过高,使得正样本数量大幅减少,造成样本数量不平衡,从而出现过拟合情况。

表 2 级联结构数量对参数的影响

Table 2 Impact of number of cascading stages on parameters %

Network		AP	$AP_{0.5}$	$AP_{0.75}$
Without Cascade R-CNN		20.6	41.1	18.7
With Cascade R-CNN	Stage 1-2	25.6	46.6	25.4
	Stage 1-3	26.1	46.4	25.7
	Stage 1-4	25.7	46.2	25.3

3)预设框选取。二阶段目标检测算法通常需要提前设定预设框的大小,合适的预设框尺寸往往可以带来更好的检测结果。重点对比三种预设框的选取方案,方案 A 选取{512,256,128,64,32},方案 B 选择{256,128,64,32,16},方案 C 选择{128,64,32,16,8}。上述方案长宽比均设置为{1∶2,2∶1,1∶1}。如表 3 所示,方案 B 结果最佳,AP 达到 26.1。

表 3 所提算法的 anchor 尺寸比较

Table 3 Comparison of anchor size of proposed algorithm

Anchor scheme	AP /%	AP_S/%	AP_M/%	AP_L/%
A	25.0	17.4	35.7	36.8
B	26.1	18.3	37.0	38.3
C	25.8	18.0	36.4	36.4

4)多尺度训练。在所提模型的基础上加入多尺度训练,设短边长为{600,800,1000},如表 4 所示。在 VisDrone 数据集的指标下,未加入多尺度训练之前,AP 值为 27.34,加入之后的 AP 值提升了 0.64 个百分点,达到 27.98。而在 MS COCO 数据集中,AP 值从 26.1 提升到 26.7,两种数据集中的指标均验证了所提算法的有效性。

表 4 多尺度训练检测结果

Table 4 Detection result of multi-scale training

Dataset	Multi-scale training	AP /%	$AP_{0.5}$/%	$AP_{0.75}$/%
MS COCO	×/√	26.1/26.7	46.4/48.1	25.7/26.1
VisDrone	×/√	27.34/27.98	52.31/52.54	24.84/25.62

从整个实验可以看出,在 MS COCO 数据集的指标下,FPN 的 AP 值从原始的 17.0%提升至 26.7%,提升了 9.7 个百分点。图 9 给出了部分实例的检测结果,不难发现,所提算法在检测小目标和密集目标时效果是比较不错的。

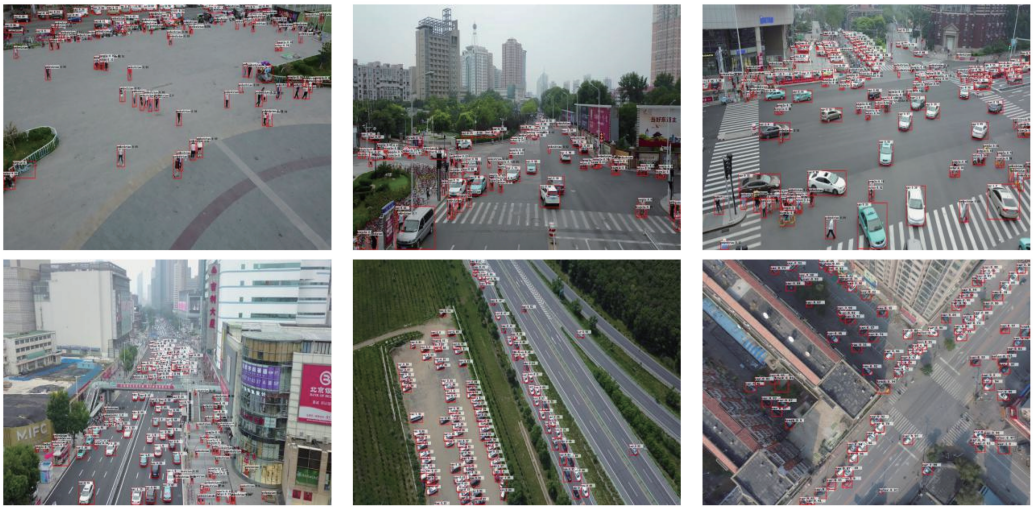

图 9 航拍图像可视化检测结果

Fig. 9 Visual detection results of aerial images

5 结　论

针对无人机航拍图像领域中存在的小目标和目标密集问题，对 FPN 算法进行改进，提出一种新的并行特征金字塔结构。该结构核心思想是对目标特征信息的不足之处进行补充和加强，有效提高了检测的精度。此外，针对 FPN 算法在第二阶段定位能力不足的问题，替换原模型中单 IoU 阈值检测结构，采用级联网络结构加强对小目标的定位能力。所提模型结构提升了目标检测在航拍图像领域中的准确性和鲁棒性。然而从实验指标可以看出，精度依然存在很大的上升空间，这是由航拍图像本身难度大所决定的，如何进一步解决复杂背景下小目标特征表达能力不佳的问题是解决航拍图像领域诸多问题的关键所在，同时也是下一步的研究方向。

参 考 文 献

[1] Fahlstrom P G, Gleason T J. Introduction to UAV systems[M]. Wu H P, Shi Z S, Ding Y F, et al, Transl. 2nd ed. Beijing: Electronic Industry Press, 2003.
Fahlstrom P G, Gleason T J. 无人机系统导论[M]. 吴汉平, 施自胜, 丁亚非, 等, 译. 二版. 北京:电子工业出版社, 2003.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 886-893.

[4] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model [C] // 2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE, 2008: 1-8.

[5] 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, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[6] Girshick R. Fast R-CNN [EB/OL]. (2015-09-27) [2019-12-09]. https://arxiv.org/abs/1504.08083.

[7] 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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6517-6525.

[9] Redmon J, Farhadi A. YOLOv3: an incremental improvement [EB/OL]. (2018-04-08) [2019-12-09]. https://arxiv.org/abs/1804.02767.

[10] 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.

[11] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved Faster R-CNN [J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测 [J]. 光学学报, 2018, 38(6): 0615004.

[12] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector [M] // Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[13] Hua X, Wang X Q, Wang D, et al. Multi-objective detection of traffic scenes based on improved SSD [J]. Acta Optica Sinica, 2018, 38(12): 1215003.
华夏, 王新晴, 王东, 等. 基于改进 SSD 的交通大场景多目标检测 [J]. 光学学报, 2018, 38(12): 1215003.

[14] 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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[15] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 6154-6162.

[16] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context [M] // Fleet D,

Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[17] Zhu P F, Wen L Y, Bian X, et al. Vision meets drones: a challenge[EB/OL]. (2018-04-23)[2019-12-09]. https://arxiv.org/abs/1804.07437.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[19] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[20] Dai J F, Li Y, He K M, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems, December 5-10, Barcelona, Spain. New York: Curran Associates, 2016: 379-387.