[文章编号] 1003－4684(2020)01-0021-04

基于改进 YOLOv3 的航拍目标实时检测方法

舒 军，吴 柯

（湖北工业大学电气与电子工程学院，湖北 武汉 430068）

[摘 要] 对于航拍图像中的小型目标，YOLOv3 算法模型对其识别精准度低，在目标被遮挡或目标较密集时存在漏检现象。针对上述问题提出了一种基于改进 YOLOv3 的航拍目标实时检测方法，该方法加入 104×104 特征分辨率的检测模块并删减了 13×13 特征分辨率的检测模块，同时增加了浅层网络的层数，用于提取更加细微的像素特征；在训练阶段针对 DOTA-v1.0 航拍数据集使用 K-means＋＋聚类得到 9 个先验框进行检测，用于提升整体网络的训练速度。实验结果表明：改进后的 YOLOv3 检测算法的检出率提升了 15.0%，mAP-50 提升了 10.5%。

[关键词] 航拍图像；YOLOv3；DOTA-v1.0；K-means＋＋

[中图分类号] TP751.1 [文献标识码] A

随着传感器技术、深度学习和图像处理技术的发展，航空摄影被广泛地应用于救援追踪、生态研究、军事领域等方面。航空摄影图像的分辨率大小通常超过 4096×2160，而待测目标在其中的像素可小至 16×10，在这样大范围的超高清分辨率[1]图像中检测识别出微小像素目标，特别是实时定位这些目标对象，仍然是一项艰巨的挑战。

目前基于深度学习的图像实时检测分类算法模型主要使用 SSD[2]、YOLO[3]、SqueezeDet[4]等算法框架；其中 YOLOv3 实时检测识别框架[5]由 Redmon，Joseph 和 Farhadi 于 2018 年提出，其识别速度快，检测精准度高，并能同时检测多尺度分辨率的目标，适用于同一镜头的多目标识别领域。但该方法直接用来检测航拍图像中的小分辨率目标时会出现遮挡漏检情况，对密集群体的检测效果亦不甚良好，需要对 YOLOv3 算法进行改进以适应航拍图像中的目标检测。

1 基本概念

1.1 YOLO 算法

YOLO(You Only Look Once)模型是第一个端到端(End-to-End)的深度学习检测算法，其借鉴了 GoogleNet[6]分类网络结构，提取特征的过程只需要一次前向运算，并同时进行训练和检测，因此以检测速度见长。YOLOv2[7]使用了 VGGNet[8]的网络结构，又称 YOLO9000，能满足 9000 多种不同类别目标的快速检测。YOLOv3 借鉴了 ResNet 残差网络[9]结构，并加入了 FPN[10](Feature Pyramid Networks)网络结构，在前两代的基础上提升了精准度，同时保证了检测速度。

YOLOv3 模型的骨干框架使用 53 层卷积的残差网络结构，故 YOLOv3 神经网络又被称为 Dark-Net-53，其由一系列的 3×3 和 1×1 的卷积层组成，其中包含 23 个残差模块和检测通道的全连接层。除训练框架外 YOLOv3 还划分三种不同栅格大小的特征检测通道，包括 52×52、26×26、13×13 栅格大小的特征图，分别对应小、中、大尺度特征的图像检测，因此网络在训练时保留了小型目标的像素特征信息，对小物体的检测效果良好。

1.2 识别对象分析

本文使用 DOTA-v1.0 数据集[11]，其为武汉大学遥感国重实验室与华中科技大学电信学院联合制作的大型遥感开源数据集，其内包含 2806 张遥感图像，分辨率 800×800 至 4000×4000 不等，共 15 种类别，包含了小型汽车、飞机、足球场等尺度跨越大的目标对象，因此该数据集适合作为航拍目标识别的实验研究对象。

在使用深度学习方法对传统目标进行训练检测时，其图像的分辨率大小通常只有数百，目标在图像中所占的比例大于 10%，且同一副图像中存在不同种类目标的数量较少；而航拍图像在进行训练检测时区别较大，需要考虑其以下几个特点。

1)超高清分辨率：原始图像像素分辨率太大，图像按原始尺寸输入神经网络造成内存不足，压缩输

[收稿日期] 2019－09－29

[第一作者] 舒 军(1973－)，男，湖北荆门人，工学博士，湖北工业大学副教授，研究方向为机器视觉，深度学习

入后又导致图像中小目标像素特征的丢失；

2)旋转不变性：实际高空拍摄的目标包含了各个方向的朝向，但训练的样本图像只包含了某部分的朝向；

3)样本数量少：该类图像样本数量相对少，且一幅图像中可能同时包含几十上百个待测目标对象，在训练过程中容易发生过拟合现象。

在对航拍数据集进行训练前，需要考虑以上因素来对深度学习算法进行改进。

2 YOLOv3 算法的改进

2.1 网络结构的改进

由于航拍图像数据集与传统的图像数据集差别较大，其包含的小型和中型目标多而大型目标少，一副样本图像包含的范围大，即使待测目标是一个足球场，相对整张图片所占的比例也不超过 20%，因此使用原 YOLOv3 网络结构将特征图大小划分为 52×52、26×26、13×13 显然已经不合适。

本研究划分特征图大小为 104×104、52×52、26×26，加入了划分更加细微的特征信息，同时删减了大型尺度的特征。改进后的网络整体结构如图 1，其中共包含 24 个残差模块，在浅层网络中增加了卷积层和残差连接结构的数量，这样可以提升更深层网络特征的细粒度，然后从此后部分更深层的网络中，提取三种不同尺度的特征信息分别输入检测通道，深层网络的检测通道向浅层进行上采样，以丰富浅层的特征信息，来提升小型目标的检测精准度。

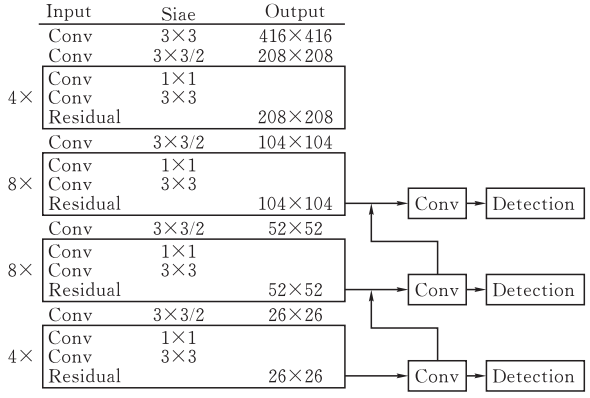


图 1 改进后 YOLOv3 网络的整体结构

2.2 基于数据集的聚类

本研究在训练 DOTA-v1.0 数据集前，使用 K-means++[12] 对 DOTA-v1.0 样本图片进行聚类，设置适合其数据集的 9 个大小不一的先验框，神经网络在训练阶段会优先使用这 9 个先验框去匹配图像中的目标，从而快速找到目标以提取像素特征信息。因此能够提升训练速度，缩短达到拟合的时间，聚类得到对应尺度的先验框大小见表 1。

表 1 三种尺度的先验框分辨率大小

特征图栅格	104×104	52×52	26×26
视觉尺度	小	中	大
先验框大小	(7×11)	(18×8)	(63×63)
	(15×18)	(32×32)	(87×173)
	(16×55)	(45×15)	(217×178)

3 实验与分析

3.1 数据集的预处理

由于改进 YOLOv3 使用了 104×104、52×52、26×26 栅格大小的特征输入检测通道，在训练输入的图像分辨率为 416×416 的情况下，能够检测的目标最低分辨率为 4×4(图像尺寸除以栅格大小，即 416/104)，但对于 DOTA-v1.0 数据集中 4000×4000 分辨率的原始图像，在以 416×416 分辨率输入时，因其像素特征缩小 10 倍，已经无法检测出原有的小目标，在训练前使用一种滑窗裁剪策略，在输入图像前，将一张原始图像滑窗裁剪为多张 608×608 分辨率的图像，然后再经过 416×416 压缩输入，这种方法使得原始图像中的像素特征信息得以尽可能地保留。

在滑窗裁剪时设置 10%的重叠率，防止将其中的目标对象一分为二时导致像素特征信息被破坏。本实验从 DOTA-v1.0 数据集中选取 1869 张带目标标签信息的图像，裁剪后得到图片 19219 张，其中 4/5 作为训练集参与网络训练，1/5 作为验证集来分析分类检测的效果。

3.2 模型的训练

本研究的实验硬软件环境：系统平台为 Ubuntu16.04，CPU 为 Intel Core i5-4590 ，GPU 为 GTX1080Ti，GPU 加速为 CUDA8.0 cuDNN6.0，编译环境为 gcc/g++5.0 OpenCV3.4.0，算法框架为改进 YOLOv3-416。

在进行航拍模型训练时，将训练集中的原始图像按 416×416 输入，每批次训练 64 张图像，分 16 组输入。每次输入时生成±1.5 倍饱和度和±1.5 倍曝光量的样本，以及±0.1 比例色调的样本，这些图像会首先经过骨干训练网络，然后在多尺度训练网络继续生成尺寸调整后的样本，整体训练流程见图 2。

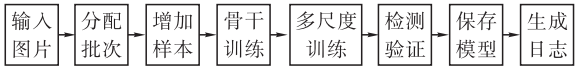


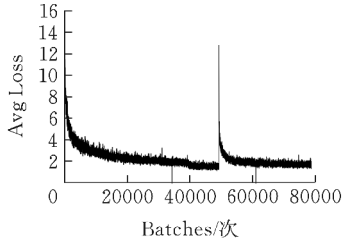
图 2 模型训练流程图

对于训练参数，设置最优化方法动量参数为 0.9，权重衰减正则项为 0.000 5，学习率 0.001，迭代到 45 000 次时学习率降低 10%，迭代到50 000次时

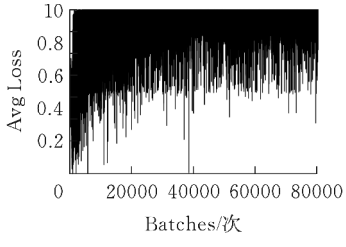
学习率再降低 10%，训练时每 1000 次迭代保存一次模型，超过 1000 次迭代数后每 10 000 次保存一次模型。实验首先迭代 50 000 次，后续在该基础上继续迭代 30 000 次，共迭代 80 000 次，总共训练图片 5 120 000 张。

3.3　模型的评估

训练完毕后，从导出的 log 日志文件中提取关键词信息，绘制出 loss 曲线以及 IoU 值统计图像（图 3）。



(a)模型平均 loss 值走势



(b)模型的 IoU 统计值

图 3　模型的 loss 曲线和 IoU 曲线

从图 3a 中可以看到，模型的 loss 值在前 10 000 次迭代中迅速降低到 3 左右，随后逐渐平滑，稳定在 2 左右；而在迭代 50 000 次时 loss 值有明显的起伏，这是由于实验在进行到 50 000 次时暂停了训练，后续又将这 50 000 次迭代模型作为初始权重模型继续迭代了 30 000 次，故 loss 出现突然的起伏后迅速回到正常范围，但后续的迭代过程 loss 值已经没有下降的趋势。整体来看，训练模型迭代到 80 000 次时模型已经趋于稳定，最终 loss 值在 2 左右，表明模型的检测性能良好。

从图 3b 中可以看到，随着迭代次数的增加，IoU 的值逐渐达到 0.6 以上，并有大量值接近 1.0，说明模型预测框与对象实际框的匹配度较高。

3.4　改进前后模型的对比

在使用改进 YOLOv3 网络与原网络进行对比时，都保留了相同的参数设置和预处理操作。在选取模型时，一般选取 IoU 及 mAP 值最高的权重模型作为对比，但训练到 80 000 次时，这些指标的变化差异已经非常小，为了方便直观地进行对比，都统一选取迭代 80 000 次的模型来进行对比，具体性能指标见表 2。

表 2　改进前后各项类别具体性能对比

类别	准确率/%		TP/FP	
	改进前	改进后	改进前	改进后
plane	68.14	80.71	3175/444	4073/415
ship	69.75	76.47	13770/2851	14999/2560
Storage-tank	48.42	52.07	2043/593	2047/281
baseball-diamond	40.38	59.99	160/98	253/115
tennis-court	90.20	90.39	1441/124	1416/74
basketball-court	49.04	57.45	159/208	170/101
ground-track-field	32.37	43.25	93/141	148/215
harbor	64.66	74.75	2960/1064	3481/1126
bridge	21.41	42.82	247/553	409/305
small-vehicle	48.47	55.13	5944/10747	5379/3312
large-vehicle	64.10	61.22	6438/3046	5001/1295
helicopter	19.52	29.47	21/21	59/134
roundabout	21.47	51.67	70/114	161/104
soccer-ball-field	34.01	45.47	112/110	154/130
swimming-pool	25.24	34.02	225/245	248/189

从表 2 可以看到，两种模型对飞机、船、网球场和港口的检测性能较好，原因是这几个类别的样本数量相对较多；而小型汽车的样本数量较多但检测效果一般，原因是小型汽车的分辨率太小，在输入原始图像检测时，一部分过小目标的像素特征变得不明显，而大型汽车的分辨率稍大，故检测效果比小型汽车更好。对于网球场，检测性能基本没有提升，这是由于网球场的分辨率本身就较大，即使将原图压缩后输入训练，提取的特征信息仍然比较完整。

整体来看，改进后的网络对 15 种类别的目标都有较大提升，这是由于改进后网络生成了更少的真实负样本(FP)参与验证，且样本数量越多，检测的准确率一般也会相应增加。

表 3　改进前后综合指标对比

检测模型	检出率	召回率	F1 分数	mAP-50	FPS
改进前	0.64	0.66	0.65	46.5	19.8
改进后	0.79	0.68	0.73	57.0	20.4

从表 3 可以看到，改进后模型与改进前进行对比时，检出率由 64%提升到 79%，提升了 15.0%；召回率由 66% 提升至 68%；F1 分数提升了 0.08；mAP-50 从 46.5%提升到 57.0%，提升了 10.5%；FPS 由 19.8 提升至 20.4。综合上述，实验证明了改进 YOLOv3 网络模型对比原网络拥有更好的实时检测性能。

4　结论

本文针对航拍目标识别提出了一种改进的 YOLOv3 算法，该方法加入 104×104 特征分辨率的检测模块并删减了 13×13 特征分辨率的检测模块，同时增加了浅层网络的层数，用于提取更加细微

的像素特征;并针对 DOTA-v1.0 数据集样本使用 K-means++聚类,得到 9 个先验框来参与模型训练。实验表明,改进后的 YOLOv3 检测算法的检出率提升了 15.0%,mAP-50 提升了 10.5%,对于航拍图像目标的实时检测具有一定的参考意义。

由于原始航拍图像分辨率太大,即使在训练时保留了分辨率特征信息,但将整张图片直接压缩输入后,难免会破坏原有分辨率特征,因此平均准确度不高;未来还应在保证实时检测速度的情况下,对检测时的图像输入模块进行优化。

[参 考 文 献]

[1] 房磊. 4K 超高清数字发展动态研究[J].广播与电视技术, 2014, 41(12):1-2.

[2] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[3] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[4] Wu B, Iandola F, Jin P H, et al. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 129-137.

[5] Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. IEEE Conference on Computer Vision and Vision and Pattern Recognition. IEEE, 2018: 89-95.[6] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[8] Simonyan, Karen, Zisserman, Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014: 542-564.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[11] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3974-3983.

[12] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.

Research on Real Time Detection Method of Aerial Targets Based on Improved YOLOv3

SHU Jun, WU Ke

(*School of Electrical and Electronic Engineering, Hubei Univ. of Tech., Wuhan* 430070, *China*)

Abstract: For small targets in aerial images, the YOLOv3 algorithm model had low recognition accuracy, and there was a missed detection when the target was occluded or dense. Aiming at these problems, this paper proposed a real time detection method based on improved YOLOv3 for aerial targets. This method added a detection module of 104×104 feature resolution and cut the detection module of 13×13 feature resolution, and added shallow layers, which was used to extract more subtle pixel features. In the training phase, K means++ clustering was used to obtain 9 prior boxes for the DOTA v1.0 aerial dataset, which was used to improve the training speed of the overall network. Experiments showed that the detection rate of the improved YOLOv3 detection algorithm increased by 15.0%, and the mAP 50 increased by 10.5%.

Keywords: aerial images; YOLOv3; DOTA v1.0; K means++

[责任编校: 张岩芳]