基于 FD-SSD 的遥感图像多目标检测方法

朱敏超 冯 涛* 张 钰
(杭州电子科技大学电子与信息学院 浙江 杭州 310018)

摘 要 针对遥感图像中目标物体过小,不易检测的难点,提出对 SSD 的改进网络 FD-SSD(Feature Fusion and Dilated Convolution Single Shot Multibox Detector)。FD-SSD 去掉了 SSD 网络数据预处理层的随机剪裁步骤,并结合 FSSD 将具有高分辨率的低层特征图和具有高语义信息的高层特征图进行融合。使用空洞卷积增大第三层特征图的感受野,利用具有高分辨率的低层特征图对小目标进行预测。同时不再使用 1×1 的顶层特征图产生目标框。模型训练阶段将原始遥感图像进行"二次切割"处理,增加训练样本量。在预测阶段先将原始图像进行切割预测,再将目标框映射回原图,并对原图所有的目标框进行二次非极大值抑制(NMS),保留最优目标框。FD-SSD 在 DOTA 数据集上有良好的表现,比原始 SSD 的 mAP 提升 31%。

关键词 遥感图像 目标检测 特征融合 空洞卷积 深度学习

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.042

REMOTE SENSING IMAGE MULTI-TARGET DETECTION METHOD BASED ON FD-SSD

Zhu Minchao Feng Tao* Zhang Yu
(*School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China*)

Abstract Aiming at the difficulty that the object in remote sensing image was too small to be detected easily, FD-SSD (feature fusion and dilated convolution single shot multibox detector), an improved network of SSD, was proposed. FD-SSD removed the random tailoring steps in SSD network data preprocessing layer. It was combined with FSSD to integrate the feature map with high resolution in lower layer into the feature map with high semantic information in higher layer. Dilated convolution was adopted to enlarge the receptive field of the feature map in the third layer. The feature map with high resolution in lower layer was used to predict the small targets. The top-level feature map with dimension of 1×1 was no longer used to generate the target box. In the model training phase, FD-SSD performed the secondary cutting for the original remote sensing image to increase the training samples. In the prediction stage, it cut and predicted the original image, and mapped the target frame back to the original image. All the target frames of the original image were subjected to quadratic NMS to preserve the optimal target frame. FD-SSD has an excellent performance on the DOTA dataset. Compared with mAP of the previous SSD, it increases by 31%.

Keywords Remote sensing image Target detection Feature fusion Dilated convolution Deep learning

0 引 言

随着深度学习应用于目标检测,目标检测算法在 Pascal VOC 数据集上的准确率已达到了一定的水准。其中表现优异且最具有代表性的算法有 Faster-RCNN[5]系列、SSD[1]系列和 YOLO[6]系列。Faster-RCNN 接受的图片输入大小为[1 000,600],SSD 接受的图片输入大小为[300,300]或[512,512],YOLO 接受的图片数据大小为[416,416]或[544,544]。在速度上,SSD 和 YOLO 明显优于 Faster-RCNN。其中 SSD 保留了 Faster-RCNN 中类似 Anchor 的使用,且它通过多尺度预测,这使得它在检测速度和检测精度上都有良好的表现。

收稿日期:2018-08-06。国家自然科学基金项目(61372156)。**朱敏超**,硕士生,主研领域:深度学习目标检测。**冯涛**,副研究员。**张钰**,副教授。

相较于 Pascal VOC 数据集和 ImageNet 数据集中的图像,标定好的遥感图像数据相对较少,且图中的目标具有多方向性,这就需要网络模型具有良好的旋转不变性,而且相对于整张图而言,目标所占像素比例较小。若将整张遥感图像直接输入网络进行训练和预测,当图像经过数据预处理层,图像尺度将缩小为网络所需大小,图像中的目标经过尺度变换后将失去很多原有的信息,这将导致网络模型训练时难以收敛,且其泛化能力也非常弱。针对这个问题文献[7]提出了一种解决方案,YOLT 在对于遥感图像的检测时选用了 YOLO 模型,并将 YOLO 网络裁剪为 22 层的网络,在应对遥感图像中目标过小的问题,YOLT 加入了一个穿透层,使网络发现更多细粒度的特征。同时 YOLT 为了应对遥感图像目标尺度的巨大差别,训练了两种尺度的模型进行检测。针对不同的目标采用不同的训练数据,汽车的训练数据来源于 COWC[8],建筑物的训练集来源于 SpaceNet 数据集,同时又在 DigitalGlobe 图像上标定了船舶、飞机和机场的数据,用这些数据训练了两个不同尺度模型。在测试阶段,YOLT 将一张大尺度高分辨率的遥感图像切割成为具有 15% 重叠率的小图片,然后对小图片进行目标检测,再将目标框还原到原图相应的位置,并对 15% 重叠部分的图像进行 NMS 处理。

不同于 YOLT,本文采用 SSD 网络模型,SSD 在运算速度和检测精度上都有良好的表现。但在 DOTA[4](A Large-scale Dataset for Object Detection in Aerial Images)中,SSD 却表现得略逊色于其他网络模型。本文对原始遥感图像进行二次切割处理,既最大可能的保留了全部图像信息,同时也减小了切割图像带来的不利影响。在网络改进中,本文提出的 FD-SSD 去掉了数据增强的随机剪裁操作,保留图像的全部信息。网络结构上借鉴 FPN[9] 网络和 FSSD[2] 网络的方法,将高层特征图和低层特征图进行融合,增加低层特征图的空间语义信息。同时引入了空洞卷积[3],让第三层特征也参与预测;去掉网络中明显不适用于遥感图像目标检测的网络层。FD-SSD 网络在 DOTA 上的表现不逊色于同类型网络,且相较于原始 SSD 网络模型,mAP 提高了 31%。

1　样本数据切割

1.1　数据集介绍

DOTA 数据集有 2 806 张遥感卫星图,其中 1/6 作为验证集,1/3 作为测试集。DOTA 标定了 15 种类别,包含 188 282 个实例目标,包括飞机、船只、储蓄罐、棒球内场、网球场、篮球场、田径场、海港、桥梁、大型车辆、小型车辆、直升飞机、足球场、环形公路和游泳池。DOTA 的图像数据的尺寸大小为 800 ~ 4 000,而且图中的目标具有多方向性,尺寸大小也非常丰富。图像中的目标以中、小目标居多,且比例接近为 1∶1。DOTA 以目标框水平的高度作为目标大小的度量,因此小目标的大小为 10 ~ 50(像素),中等大小的目标为 50 ~ 300,而且对与同一类型的小目标而言,目标密度很高,例如停车场的汽车、机场的飞机、码头的船舶等。DOTA 数据的标定信息存放于文本文件中,不同于 Pascal VOC 的 xml 文件。DOTA 数据集的标注信息只有训练集和验证集,而测试集的标注信息并未给出。

1.2　二次切割

本文先将 DOTA 的文本格式标注信息转换为 xml 格式,将图像的宽、高、通道数以及标注框信息(ground truth)写入 xml 中,同时移除 DOTA 数据中的异常标注框(DOTA 的某些标注框存在越界的错误)。由于遥感图像的尺度过大,直接将原图送入卷积网络进行训练会导致网络训练无法收敛,即使收敛了,最终模型的检测效果和泛化能力也不理想。针对这种情况,本文采取 YOLT 和 DOTA 中提到的图像切割法。不同于 YOLT 和 DOTA 的切割,本文采用“二次切割”,切割示意图如图 1 所示。

图 1　二次切割

“二次切割”的做法即对原图从两个方向进行两次切割,第一次先从原图左上角开始(即①方向)往右下角开始切割,切割大小为 1 024,切割的重叠比率为 50%。若原图的长度小于 1 024,则在水平方向上不做切割,同理判断宽度是否小于 1 024,再确定垂直方向上是否做切割;若切割到原图的最右边和最底边时,余下的尺寸小于 1 024,则放弃这部分图像,如图 1 中点划线右边部分和点划线下面部分。第二次则从右下角

开始(即②方向)往左上角开始切割,切割方法与第一次相同。若切割后的图像中不存在任何目标,这张图像将不作为训练数据,直接移除。通过"二次切割"后的训练数据从原来的 1 411 张增加为 25 356 张。同时在切割过程中对标注框信息进行映射,使标注框能在切割后的图像中正确标定相应的目标,映射公式如下所示:

$$f(x_{\min}) = \begin{cases} 0 & x_{\min} \leqslant w_s \\ x_{\min} - w_s & x_{\min} > w_s \end{cases} \tag{1}$$

$$f(y_{\min}) = \begin{cases} 0 & y_{\min} \leqslant h_s \\ y_{\min} - h_s & y_{\min} > h_s \end{cases} \tag{2}$$

$$f(x_{\max}) = \begin{cases} C & x_{\max} \geqslant C \\ x_{\max} - w_s & x_{\max} < w_{ed} \end{cases} \tag{3}$$

$$f(y_{\max}) = \begin{cases} C & y_{\max} \geqslant C \\ y_{\max} - h_{ed} & y_{\max} < h_{ed} \end{cases} \tag{4}$$

其中 $x_{\min}$、$y_{\min}$、$x_{\max}$、$y_{\max}$为标注框的信息,w_s 表示切割图像在原图上横坐标的起始位置,w_{ed}表示切割图像在原图上横坐标的结束位置,h_s 表示切割图像在原图上纵坐标的起始位置,h_{ed}表示切割图像在原图上纵坐标的结束位置,C 表示切割图像的尺寸大小(本文 C = 1 024)。若目标处在切割线上,如图 2 所示。

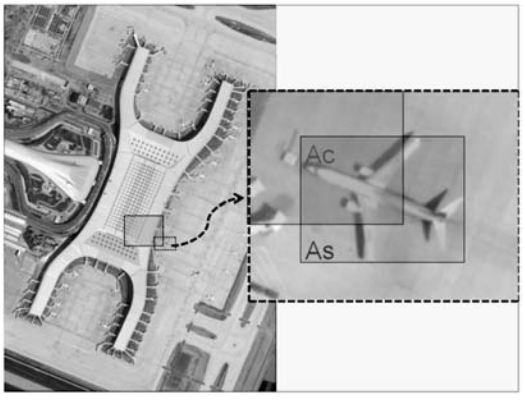

图 2 被切割的目标

对于图 2 出现的情况,先计算切割后的目标标注框面积与原目标标注框面积的比率 P 如式所示:

$$P = \frac{A_c}{A_s} \tag{5}$$

式中:A_c 表示切割后的标注框面积,A_s 表示切割前的标注框面积。标注框的面积 A 的计算如下所示:

$$A = (x_{\max} - x_{\min})(y_{\max} - y_{\min}) \tag{6}$$

式中:$x_{\min}$、$y_{\min}$、$x_{\max}$、$y_{\max}$为标注框的信息。

若 $P \geqslant 0.7$,则完全保留此目标的标注信息;若 $0.3 \leqslant P < 0.7$,则将标注框保留,但将 xml 文件中的 difficult 项置 1,若 $P < 0.3$ 则将此目标的标注信息移除。通过这种办法既可以保证最大可能的保留切割后的目标信息,也可以排除目标被切割后由于留下的信息量过少,混入过多的背景信息而导致误检增加。

采用"二次切割"的方式对原图进行划分,可以做到以下几点好处:(1) 最大可能的保留原图的所有信息,若一张图的原始尺寸不能被切割的尺寸整除,这将导致原图上很大一部分的信息流失。尤其是在遥感图像中,检测目标相对较小,若直接丢弃这部分信息,将使得模型对小目标的检测能力下降。(2) 最大可能的保留处于切割线上的目标信息框,同时在切割后可以增加被切割目标的样本多样性。切割结果及其标注框信息映射结果如图 3 所示。

图 3 二次切割结果

2 FD-SSD 网络模型

2.1 基础结构

FD-SSD 网络的基础结构是 SSD 和 FSSD。SSD 的网络架构主要分为两个部分:(1) 使用位于网络前端的基础网络提取图像的特征,如 VGG[10]、ResNet[11]等。其中 SSD 采用 VGG 网络并将 VGG 网络的第六层和第七层全连接层改为了卷积层。(2) 用于多尺度[12]检测的级联网络,其中使用 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Conv11_2 这六个尺度的特征层进行预测输出。

FSSD 网络是在 SSD 模型上的改进,主要思路源自 FPN 网络。FSSD 认为 SSD 把不同层级的特征层当作同一级别的特征层来进行预测,使得 SSD 在预测的时候,低层特征图缺少全局语义信息。但低层特征层又具有较高分辨率,保留较多细节信息,对检测小目标有非常大的帮助,因此将具有低语义高分辨率的低层特征和具有高语义低分辨顶层特征融合来解决这个问题。

2.2 FD-SSD 结构

本文针对遥感图像的特点,提出了 FD-SSD。由于遥感图像中目标尺寸较小,若直接使用 300 × 300 的图像作为输入,在网络数据预处理层进行图片的压缩时,

小目标将失去很大一部分信息,因此本文选用 512 × 512 作为网络的输入尺寸。同时 SSD 的数据增强主要是做一个随机的剪裁的过程,这一过程会使得原本信息量就少的小目标随机性地再失去一部分信息,这将导致最后训练的模型对小目标的检测能力下降。因此本文直接将原图作为输入,不再做随机剪裁,本文认为,数据的"二次切割"可以弥补样本丰富性不足的缺点。

FSSD 对 SSD 的改进使得 SSD 对小目标的检测能力有了一定的提高,但 FSSD 在构建特征融合时并未加入 Conv3,FSSD 尝试加入 Conv3 特征层后反而使得效果下降,SSD 也未使用 Conv3。主要原因在于 Pascl VOC 的目标大部分为大目标或为中等目标,Conv3 特征层的感受野过小,同时不具备较好的语义信息,因此使用了 Conv3 后模型的检测效果并未明显提升。但对于 DOTA 遥感数据集而言,图中大部分为小目标,Conv4 已经过了 3 个下采样的过程,特征图的分辨率已经不足以去检测遥感图像的小目标。因此 Conv3 的使用是非常有必要的。

本文提出的 FD-SSD 使用 Conv3,让其直接预测输出目标框。虽然 Conv3 分辨率较高,但是其感受野较小,语义信息不够,若直接使用大卷积核或多个小卷积核进行卷积来增加特征图的语义信息,这将增加网络模型的计算量。本文在保证计算量和速度的前提下,又希望提高特征层的感受野大小,增加语义信息,因此本文直接使用空洞卷积对 Conv3 进行卷积操作,FD-SSD 网络结构如图 4 所示。

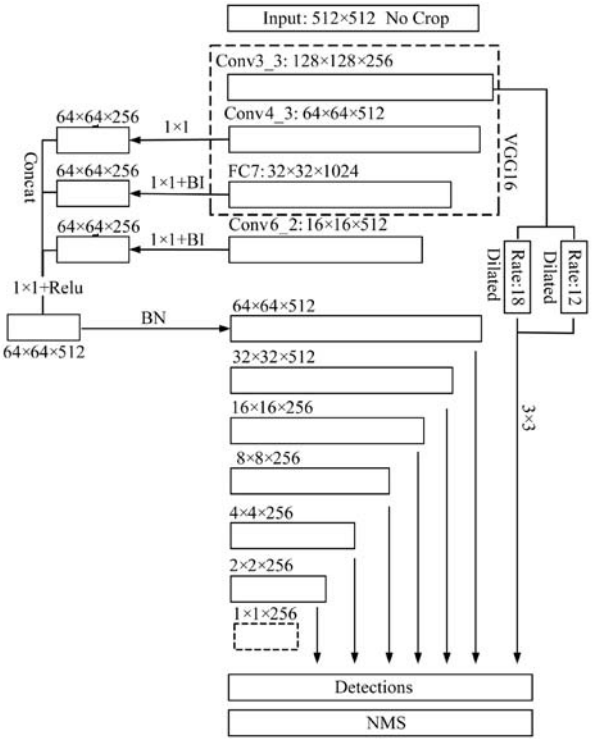

图 4　FD-SSD

FD-SSD 的基础网络为 VGG16,使用的模型基础结构是 SSD 结构,特征融合方法使用的是 FSSD 所用的方法,但与 FSSD 稍有不同的是,本文在对不同层的特征进行叠加后直接采用 1 ×1 的卷积进行降维卷积,同时引入 Relu 非线性激活函数,提升网络的表达能力。由于网络顶层的 1 ×1 特征图上的默认框过大,对遥感图像中的目标检测并没有实际的意义,且为了兼顾速度,本文直接去掉顶层的特征图。对于 Conv3 的处理采用两个 rate 分别为 12 和 18 的空洞卷积进行卷积,然后将卷积的到的具有不同感受野大小的特征图进行融合,融合的方式采用像素点相加法,同时使用一个 3 ×3 的卷积来消除不同特征图融合带来的混叠效应。融合之后的特征图既保留了低层特征图的高分辨率,又具有较好的语义信息。

2.3　FD-SSD 默认框参数设计

FD-SSD 每张特征图上的默认框的尺度计算如下所示:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1) \quad k \in [1,m] \tag{7}$$

式中:$s_{\min}$ 表示默认框的尺度占原图尺度的最小值,$s_{\max}$ 表示默认框的尺度占原图尺度的最大值,m 表示特征图的数量,本文设定 $s_{\min}=0.1$、$s_{\max}=0.9$、$m=7$。

默认框的纵横比计算公式如式(8)和式(9)所示:

$$w_k^a = s_k\sqrt{a_r} \tag{8}$$

$$h_k^a = \frac{s_k}{\sqrt{a_r}} \tag{9}$$

式中:$a_r \in \left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$,$s_k$ 是由前面的步骤计算的得到的,针对纵横比为 1 的情况增加一个默认框的尺度 $s'_k = \sqrt{s_k s_k + 1}$,默认框的坐标中心为 $\left(\frac{i+0.5}{|f_k|},\frac{j+0.5}{|f_k|}\right)$,$|f_k|$ 为对应特征图的尺寸大小,$i,j \in [0,|f_k|)$,这边可以使得中心坐标归一化到[0,1]以方便特征图上的坐标和原图进行映射。特征图上的默认框和原图的坐标映射公式如式(10) - 式(13)所示:

$$x_{\min} = \left(\frac{i+0.5}{|f_k|} - \frac{w_k}{2}\right)w_s \tag{10}$$

$$y_{\min} = \left(\frac{j+0.5}{|f_k|} - \frac{h_k}{2}\right)h_s \tag{11}$$

$$x_{\max} = \left(\frac{i+0.5}{|f_k|} + \frac{w_k}{2}\right)w_s \tag{12}$$

$$y_{\max} = \left(\frac{j+0.5}{|f_k|} + \frac{h_k}{2}\right) h_s \tag{13}$$

式中：w_k 和 h_k 是计算得到的特征图上默认框的宽和高；w_s 和 h_s 是原图的宽和高，即为映射到原图的预测框的坐标信息。

FD-SSD 虽然去掉了顶层的特征图，但在计算 FD-SSD 默认框尺度的时候，仍将顶层的特征图计入特征图总数量中。Conv3 不计入其中，Conv3 的默认框尺度进行手动设定，设定的默认框的尺度范围为原图的 2% ~4%。在计算起始特征图（本文指 Conv4）的默认框的尺度范围时，设定其为原图的 4% 到 $s_{\min}$。因此本文在网络设计时，默认框的大小如表 1 所示。

表 1 特征层默认框尺度对照表

特征层尺度	默认框尺度范围
128 ×128	[10.24, 20.48]
64 ×64	[20.48, 51.20]
32 ×32	[51.20, 133.12]
16 ×16	[133.12, 215.04]
8 ×8	[215.04, 296.96]
4 ×4	[296.96, 378.88]
2 ×2	[378.88, 460.80]

2.4 FD-SSD 模型训练测试

本文在对 FD-SSD 进行训练时采用 Adam[13] 优化算法，初始学习率为 0.01，训练的 epoch 数为 10，同时使用在 ImageNet 上训练的 VGG-16 模型作为网络的预调模型进行迁移学习。测试阶段数据的处理方式与训练阶段稍有不同，测试阶段数据的切割只保留 20% 的重合率，同时切割方法也有所改变。第一次切割方式与训练数据预处理的切割方式相同，只是切割的重合率改为 20%，切割大小改为 1 000。第二次的切割方向改为从图 1 中的③方向进行切割，且只切割一块，大小为 1 000。这种切割方式的好处有两个：1）保证切割后的图像能包含全部的原图信息。2）能尽量减少切割带来的重复计算量。对于切割后的图像上的预测框，采用式（1）－式（4）将其映射回原图，同时对全图的预测框采用非极大值抑制（NMS）去除重复框，为了减少目标的误检和重检，本文采用的 IOU 阈值为 0.2，检测的置信度为 0.75。整体方案流程如图 5 所示。

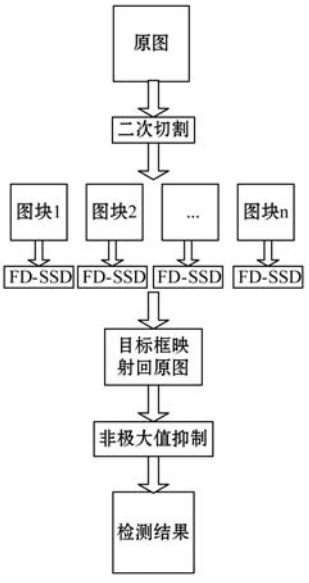

图 5 方案流程

3 实验设计与分析

本文实验环境：显卡为 GTX1070_8G，系统内存为 8 GB，CPU 型号为 ES 版，主频为 3.4 GHz，深度学习框架为 TensorFlow。

3.1 mAP 效果对比

作为对比实验的网络均采用 DOTA 所述方法进行训练测试。YOLOv2 的基础网络为 GoogLeNet[14]、RFCN[15] 和 Faster-RCNN 的基础网络均为 ResNet-101，SSD 的基础网络为 InceptionV2[16]。训练时均采用预调模型进行迁移学习，使用 SGD 优化算法，初始学习率为 0.01，训练的 epoch 为 10。训练和测试对原图均只采用单次切割（即图 1 中的①方向），切割大小为 1 024，重叠比率为 50%，同时在测试阶段最后使用全图 NMS 去除重叠部分的重复检测框。实验 mAP 结果对比如表 2 所示，效果对比如图 6 所示。

表 2 mAP 结果对照表

	YOLOv2	R-FCN	FR	SSD	FD-SSD
Plane	76.90	79.33	80.23	44.74	71.98
BD	33.87	44.26	77.55	11.21	32.50
Bridge	22.73	36.58	32.83	6.22	20.18
GTF	34.88	53.53	68.13	6.91	46.93
SV	38.73	39.38	53.66	2.00	41.56
LV	32.02	34.15	52.49	10.24	35.82
Ship	52.37	47.29	50.04	11.34	46.61
TC	61.65	45.66	90.41	15.59	77.28
BC	48.54	47.74	75.05	12.56	41.48

续表 2

	YOLOv2	R-FCN	FR	SSD	FD-SSD
ST	33.91	65.84	59.59	17.94	51.92
SBF	29.27	37.92	57.00	14.73	29.17
RA	36.83	44.23	49.81	4.55	38.98
Harbor	36.44	47.23	61.69	4.55	38.25
SP	38.26	50.64	56.46	0.53	46.71
HC	11.61	34.90	41.85	1.01	0.10
mAP	39.20	47.24	60.46	10.94	41.95

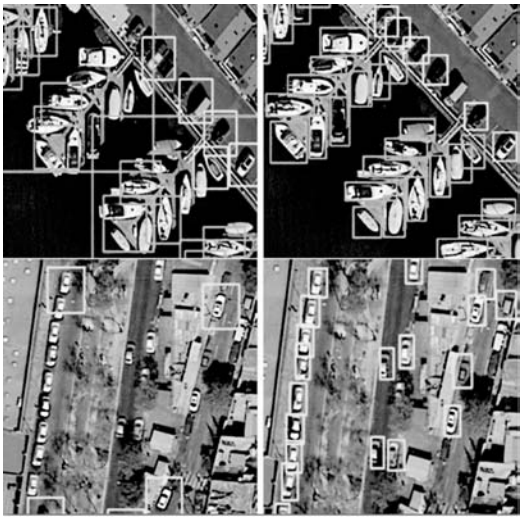

图 6　效果对比图

表 2 中的缩写表示为:BD-Baseball diamond, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, TC-Tennis court, BC-Basketball court, SC-Storage tank, SBF-Soccer-ball field, RA-Roundabout, SP-Swimming pool, HC-Helicopter, FR-Faster-RCNN。从表中可以看出,相较于 SSD,FD-SSD 的检测效果有明显的提升,尤其是图像中的小目标的检测,如 small-vehicle 提升了 39.56%,swimming-pool 提升了 46.18%,整体 mAP 提升了近 31%,同时 FD-SSD 超过 YOLOv2 的 mAP 2.75%,接近于 R-FCN 和 Faster-RCNN。

图 6 左边的是原始 SSD 的检测效果部分演示,右边的是 FD-SSD 的检测效果部分演示,可以明显看出,原始 SSD 对于小目标的检测效果并不理想,即使检测到了,其目标框也不够精准。FD-SSD 不仅提高了检出率,同时目标框也更加精准了。

3.2　其他遥感数据集对比实验

为了测试本文所提方法的可用性,本文在 NWPU VHR-10 dataset[17-19] 数据集和 RSOD-Dataset[20-21] 数据集上进行了 FD-SSD 和 SSD 的对比实验。两个网络使用相同的超参数进行学习,即使用 Adam 优化算法,初始学习率为 0.01,epoch 数为 10。不同的是 SSD 不采用本文所述的"二次切割"发对数据进行处理,FD-SSD 使用本文所述方法进行训练和测试。实验结果如表 3 所示。

表 3　实验结果对照表

数据集	网络模型	mAP	FPS
NWPU	SSD	44.77%	18
	FD-SSD	63.81%	10.7
RSOD	SSD	31.30%	18
	FD-SSD	50.69%	12

从表 3 中可以看出,FD-SSD 在两个数据集上的表现均比 SSD 有明显的提升,在 NWPU 上 FD-SSD 的 mAP 提升了 19.04%,在 RSOD 上 FD-SSD 的 mAP 提升了 19.39%。在速度方面,FD-SSD 比 SSD 略慢,主要原因在于网络的复杂性和对原图的二次切割造成的时间消耗增加。

3.3　速度测试

为了测试 FD-SSD 在实时性上的可用性,与本文相比网络均为原始网络,YOLOv2 图片的输出尺度为 416,采用 Darknet-19 基础网络,SSD-512 图片的输入尺寸为 512,采用 VGG-16 基础网络,Faster-RCNN 图片的输入尺寸为(1 000,600),采用 VGG-16 基础网络。由于遥感图像的尺度相差较大,因此本文针对不同尺度范围内的图像进行测试,即将图像的分辨率设为 0 ~ 1 000,1 000 ~ 2 000,2 000 ~ 3 000,3 000 ~ 4 000。测试结果如图 7 所示。

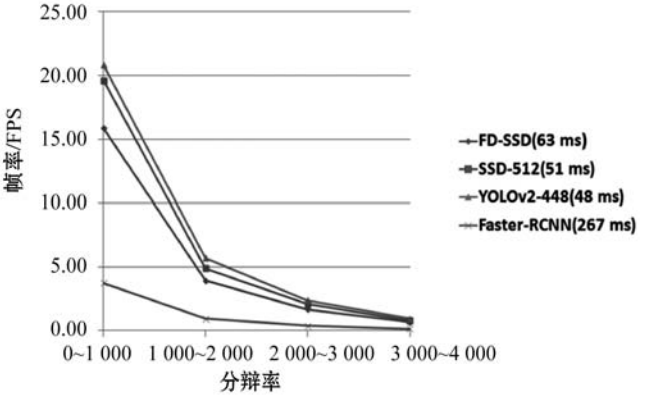

图 7　速度对比

从图 7 中可以看出,由于对图像切割的原因,当图像尺度大于切割尺度时,速度下降比较明显。FD-SSD 单次前向推理的时间为 63 ms,与原始 SSD 相比推理时间增加 12 ms,造成速度下降的主要原因是由于 FD-SSD 增加了网络的复杂度。与原始的 YOLOv2 相比推理时间增加了 15 ms。与原始的 Faster-RCNN 相比,FD-SSD 在速度上仍有明显优势。因此可以看出,

FD-SSD 在提高精度的同时,并没有过多的损失速度。

4 结 语

本文提出了针对遥感图像检测的改进网络 FD-SSD,且给出了遥感图像检测的整体方案。本文通过对遥感图像进行切割处理,使图像在进行尺度缩放后仍能保留大部分的小目标信息,同时通过去掉网络的随机剪裁处理,消除了训练过程中随机剪裁对小目标带来的不利影响。针对遥感图像中小目标检测,FD-SSD 引入了空洞卷积,通过空洞卷积的多尺度融合,增加特征图的语义信息,使得低层高分辨率的特征图能够用于目标检测。为了保证速度,FD-SSD 直接去掉了顶层的特征图。在进行预测时,对原图进行二次非极大值抑制去除重叠目标框。本文所提的改进方法,在尽可能保证 SSD 作为实时检测网络速度的前提下,使得 FD-SSD 对遥感图像的目标检测效果有了明显的提升。

参 考 文 献

[1] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Springer International Publishing, 2016:21-37.

[2] Li Z X, Zhou F Q. FSSD: Feature fusion single shot multibox detector[EB]. arXiv:1712.00960v3, 2018.

[3] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4):834-848.

[4] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[EB]. arXiv: 1711.10398v2, 2017.

[5] Ren S, He K, 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,2015, 39(6):1137-1149.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[EB]. arXiv: 1506.02640v5,2015.

[7] Van Etten A. You only look twice: rapid multi-scale object detection in satellite imagery[EB]. arXiv:1805.09512v1, 2018.

[8] Mundhenk T N, Konjevod G, Sakla W A, et al. A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning[C]//European Conference on Computer Vision. Springer, Cham, 2016:785-800.

[9] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[EB]. arXiv: 1612.03144v2, 2016.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv:1409.1556v6, 2014.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[12] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//European Conference on Computer Vision—ECCV 2016. 2016:354-370.

[13] Kingma D, Ba J. Adam: a method for stochastic optimization[C]//The 3rd International Conference for Learning Representations, San Diego, 2015.

[14] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[15] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks[EB]. arXiv:1605.06409v2,2016.

[16] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on International Conference on Machine Learning. JMLR.org, 2015.

[17] Cheng G, Han J, Zhou P, et al. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2014, 98(1):119-132.

[18] Cheng G, Han J. A survey on objectdetection in optical remote sensing images[J]. ISPRS Journal of Photogrammetry andRemote Sensing, 2016,117: 11-28.

[19] Cheng G, Zhou P, Han J. Learningrotation-invariant convolutional neural networks for object detection in VHRoptical remote sensing images[J]. IEEE Transactions on Geoscience and RemoteSensing,2016, 54(12): 7405-7415.

[20] Long Y, Gong Y, Xiao Z, et al. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks[J]. IEEE Transactions on Geoscience & Remote Sensing, 2017, 55(5):2486-2498.

[21] Xiao Z, Liu Q, Tang G, et al. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images[J]. International Journal of Remote Sensing, 2015, 36(2):27.