软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(3):738−758 [doi: 10.13328/j.cnki.jos.005695]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

改进的 SSD 航拍目标检测方法*

裴 伟[1], 许晏铭[2], 朱永英[3], 王鹏乾[2], 鲁明羽[2], 李 飞[2]

[1](大连海事大学 环境科学与工程学院,辽宁 大连 116026)

[2](大连海事大学 信息科学技术学院,辽宁 大连 116026)

[3](大连海洋大学 海洋与土木工程学院,辽宁 大连 116023)

通讯作者: 鲁明羽, E-mail: lumingyu@dlmu.edu.cn

摘 要: 近年来,无人机技术的快速发展使得无人机地面目标检测技术成为计算机视觉领域的重要研究方向,无人机在军事侦察、交通管制等场景中具有普遍的应用价值.针对无人机场景下目标分辨率低、尺度变化大、相机快速运动、目标遮挡和光照变化等问题,提出一种基于残差网络的航拍目标检测算法.在 SSD(single shot multibox detector)目标检测算法的基础上,用表征能力更强的残差网络进行基准网络的替换,用残差学习降低网络训练难度,提高目标检测精度;引入跳跃连接机制降低提取特征的冗余度,解决层数增加出现的性能退化问题.同时,针对 SSD 目标检测算法存在的目标重复检测和小样本漏检问题,提出一种基于特征融合的航拍目标检测算法.算法引入不同分类层的特征融合机制,把网络结构中低层视觉特征与高层语义特征有机地结合在一起.实验结果表明,算法在检测准确性和实时性方面均具有较好的表现.

关键词: 深度学习;无人机;深度残差网络;特征融合

中图法分类号: TP18

中文引用格式: 裴伟,许晏铭,朱永英,王鹏乾,鲁明羽,李飞.改进的 SSD 航拍目标检测方法.软件学报,2019,30(3):738−758. http://www.jos.org.cn/1000-9825/5695.htm

英文引用格式: Pei W, Xu YM, Zhu YY, Wang PQ, Lu MY, Li F. The target detection method of aerial photography images with improved SSD. Ruan Jian Xue Bao/Journal of Software, 2019,30(3):738−758 (in Chinese). http://www.jos.org.cn/1000-9825/5695.htm

The Target Detection Method of Aerial Photography Images with Improved SSD

PEI Wei[1], XU Yan-Ming[2], ZHU Yong-Ying[3], WANG Peng-Qian[2], LU Ming-Yu[2], LI Fei[2]

[1](College of Environmental Science and Engineering, Dalian Maritime University, Dalian 116026, China)

[2](College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

[3](College of Ocean and Civil Engineering, Dalian Ocean University, Dalian 116023, China)

Abstract: In recent years, the rapid development of UAV (Unmanned Aerial Vehicle) technology makes UAV ground target detection technology become an important research direction in the field of computer vision. UAV has a wide range of applications in military investigation, traffic control, and other scenarios. Nevertheless, the UAV images have many problems such as low target resolution, scale changes, environmental changes, multi-target interference, and complex background environment. Aiming at the above difficulties, derived from the original SSD target detection algorithm, this study uses a residual network with better characterization ability to replace

* 基金项目: 国家自然科学基金(1001158, 61272369, 61370070); 辽宁省自然科学基金(2014025003); 辽宁省教育厅科学研究一般项目(L2012270); 大连市科技创新基金(2018J12GX043); 辽宁省重点研发计划

Foundation item: National Natural Science Foundation of China (61001158, 61272369, 61370070); Liaoning Provincial Natural Science Foundation of China (2014025003); Scientific Research Fund of Liaoning Provincial Education Department (L2012270); Science and Technology Innovation Foundation of Dalian (2018J12GX043); Key Research and Development Plan Program of Liaoning Province

本文由智能数据管理与分析技术专刊特约编辑樊文飞教授、王国仁教授、王朝坤副教授推荐.

收稿时间: 2018-07-20; 修改时间: 2018-09-20; 采用时间: 2018-11-01

the basic network and a residual learning to reduce the network training difficulty and improve the target detection accuracy. By introducing a hopping connection mechanism, the redundancy of the extracted features is reduced, and the problem of performance degradation after the increase of the number of layers is solved. The effectiveness of the algorithm is verified through experimental comparison. Aiming at the problem of target repeated detection and small sample missing detection of the original SSD target detection algorithm, this study proposes an aerial target detection algorithm based on feature information fusion. By integrating information with different feature layers, this algorithm effectively makes up for the difference between low-level visual features and high-level semantic features in neural networks. Results show that the algorithm has sound performance in both detection accuracy and real-time performance.
Key words: deep learning; unmanned aerial vehicle; deep residual network; feature fusion

近年来,目标检测和识别技术一直是业界研究的热点.目标检测技术主要有两种研究方向:1) 基于传统方法的目标检测.主要步骤为目标特征提取、训练分类器、输出结果,对标注的训练样本进行特征提取并将其送到分类器中进行训练;2) 基于深度学习的目标检测.

目前,基于深度学习的目标检测主要分为基于候选区域的目标检测算法和基于回归的目标检测算法:

基于候选区域的目标检测算法的计算过程是:首先,根据区域选择算法从输入图像中提取出 N(N 远大于真实提取出的目标个数)个感兴趣区域(region of interest,简称 ROI);然后,利用多层卷积神经网络(convolutional neural network,简称 CNN)对上述的感兴趣区域进行特征提取,对提取到的特征进行分类;最后,利用 Bounding-box 回归器对输出窗口进行更正,得到最终结果.2014 年,Girshick 提出了 Region CNN[1](R-CNN)目标检测算法;2015 年,Girshick 提出了 Fast R-CNN[2]和 Faster R-CNN[3];2017 年,何凯明提出了基于 Faster-RCNN 框架的 Mask R-CNN[4]目标检测算法等.

上述的基于候选区域的目标检测算法虽然精度很高,但实时性差,而不使用 RPN(region proposal network)网络的目标检测方法在速度方面更具优势,即基于回归的目标检测算法:对于给定的输入图像,直接在图像的多个位置上回归出这个位置的目标边框以及目标类别.Redmon 在 2016 年提出了 YOLO[5]目标检测算法.随后,在此基础上,作者提出了改进版的 YOLOv2[6].2016 年,Liu 提出的 SSD[7]算法结合了 YOLO 速度快和 Faster R-CNN 候选区域的优点,SSD 在不同特征图上进行分割,然后采用类似 RPN 的方式进行回归,在 VOC2007 数据集上最高可达到 74.3%的准确率,处理速度为 46 帧/s.该算法不仅保证了检测速度,也提高了检测的准确率. 2017 年提出的 DSOD[8]基于 SSD 算法引入 DenseNet[9]思想,mAP(mean average precision)为 77.7%,与 SSD300 相当,但检测速度为 17.4 帧/s,较 SSD300 的 46 帧尚有较大差距[10].2017 年,Jeong 提出 Rainbow SSD[11]算法对传统的 SSD 算法进行了改进,一方面利用分类网络增加了不同特征层之间的关联度,有效减少了重复区域;另一方面增加了特征金字塔中的特征图的个数,使其适用于小目标检测,其 mAP 达到了 77.1%,同时,速度也提高到 48.3 帧/s,效果比较明显.但该算法在融合不同特征层的特征信息时覆盖了整个网络结构,这样势必会引入冗余信息,增加了计算的复杂度.2018 年提出的 YOLOv3[12]算法通过多级预测方式改善了小目标检测精度差的问题,同时,采用简化的 residual block 取代了原来 1×1 和 3×3 的 block,为无人机目标检测的应用场景的落地提供了更多的可能.

无人机技术的快速发展,使得无人机地面目标检测技术已成为计算机视觉领域的重要研究方向,无人机在军事侦察、交通管制等场景中具有普遍的应用价值.国外很早便开展了针对无人机检测跟踪系统的研究.1997 年,美国就启动了 VSAM(video surveillance and monitoring)项目,在高处架设摄像机对地面目标进行全方位的检测和跟踪.2006 年,美国 DAPRA 部门设计了一套无人机监控系统 COCOA,该系统能对无人机下视的车辆、行人等目标进行实时的检测、识别和跟踪,捕获目标的连续运动序列,使用帧间对齐技术对运动序列进行背景补偿,然后对其进行背景建模并最终实现运动目标的跟踪[13].Ibrahim 于 2010 年提出了 MODAT(moving objects detection and tracking)系统,使用 SIFT 特征进行航拍目标的检测和跟踪[14].

虽然国内该方面的相关研究起步较晚,但发展速度很快[15].2008 年,张恒设计了一套无人机平台运动目标检测和跟踪系统,能对无人机拍摄图像进行特征提取,并对机载相机运动进行自适应消除[16].谭熊等人于 2011 年提出了基于区域的航拍目标跟踪算法,该算法计算速度快、精度高,能满足实时运算的要求[17].2013 年,董晶等人设计了一套地面运动目标实时监测及跟踪系统,提取特征点进行运动目标的检测,并将检测和跟踪相结合来

进行移动目标的定位,适用于误检和目标跟踪失效的情况[18].汤轶等人设计的无人机视频中,运动目标检测与跟踪系统使用 RANSAC 算法对背景运动进行补偿,粒子群优化算法进行目标中心位置的定位,该思路确保了算法的准确性和实时性[19].

无人机产业发展日益蓬勃,其应用领域仍在不断拓展.但无人机在执行军事侦察、消防、救灾、搜救等实时任务时,目标检测的精度和实时性决定了无人机飞行任务是以机毁人亡,还是以生命财产的延续而结束,成败就在一瞬间.受到负载、续航、航行环境、计算力等限制,无人机目标检测在这方面的研究进展缓慢,已成为制约无人机发展的瓶颈问题之一.当前,无人机目标检测算法面临以下难点和问题[20].

(1) 无人机快速移动的不稳定性造成航拍图像具有图像模糊、噪声多、运动目标可提取的特征信息少、易出现重复检测、目标误检等问题;

(2) 无人机从制高点进行图像采集时,图像中的检测目标一般较小,易出现小目标漏检情况;

(3) 随着无人机的不断移动,外界环境(比如光照、云、雾、雨等)的变化将会导致图像中目标特征的剧烈变化,增加了后续特征提取的难度[21];

(4) 无人机目标检测算法需要快速准确地检测出运动目标,因此算法应满足实时计算的要求.

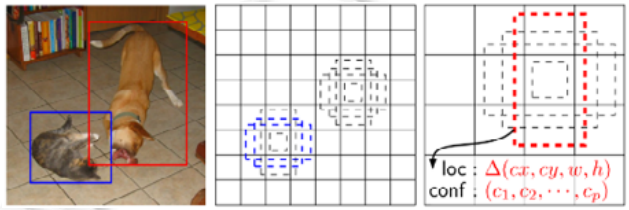
针对无人机场景下目标分辨率低、目标遮挡和光照变化等导致的可提取特征不多的问题,本文在 SSD 目标检测算法的基础上对原始基准网络 VGG-16[22]进行替换,提出了基于深度残差网络(deep residual network,简称 Resnet[23])的航拍目标检测方法(R-SSD),增强网络的特征提取能力;同时,针对 SSD 算法目标重复检测和小样本漏检问题,本文为特征提取层选取高层的语义信息和低层的视觉信息进行特征融合,提出一种基于特征融合的航拍目标检测方法(CI-SSD).

1 SSD 快速检测原理

SSD 的速度优势在于:该算法是在前馈 CNN 网络的基础上实现的,把网络的计算量封装在一个端到端的单通道中.针对单枚输入图像,SSD 会产生多个固定大小的 Bounding Box 和框中对象类别的得分,然后进行非极大值抑制(non-maximum suppression,简称 NMS)操作,得到最后的预测结果,显著提高了检测速度.网络前半部分为基础网络,主要用来进行图像分类;网络后半部分为多尺度卷积层,卷积层尺寸逐层减少,主要用于多尺度下目标特征的提取和检测.

SSD 网络中的任何一个特征层都使用一组卷积过滤器与输入进行卷积操作,产生一系列固定的预测集合,该集合包括预测目标框的 4 个偏移量和 21 个种类的置信度得分.

每个特征图都与一组不同尺度的默认边界框相绑定,在每个单元格中,预测结果为相对于默认边界框的位置偏移和类别得分.如在某个已知位置的 k 个边界框中,每个边界框都需计算相对于当前位置的 4 个坐标偏移量和 c 个类的分数,因此每个位置都有 $(c+4)\times k$ 个过滤器,对于 $m\times n$ 的输入图像,该操作总计会产生 $k\times m\times n\times(c+4)$ 个结果.如图 1 所示:图 1(a)为含有真实坐标框的输入图像,图 1(b)和图 1(c)分别是尺度为 8×8 和 4×4 的特征图.



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig.1 SSD framework for detection

图 1 SSD 的检测结构

在进行卷积操作时,每个位置都需要进行默认框(如图 1(b)和图 1(c)中 4 个不同宽高比的边界框)的计算,预测所有类别的得分和坐标偏移值.

SSD 算法中,目标损失函数的思想类似于 MultiBox[24],但 SSD 将其扩展为可处理多个类别的目标函数.

$X_{ij}^{p}=1$代表针对类别 p,第 i 个默认框和第 j 个真实框的结果保持一致; $X_{ij}^{p}=0$ 表示不一致. $\sum_i X_{ij}^{p}\geqslant 1$则表示对于类别 p 的第 j 个真实标签框,可能有多个默认框与之匹配.总的目标损失函数为

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \tag{1}$$

其中,N 为与真实标签框相匹配的默认框的个数,L_{loc} 和 L_{conf} 分别为位置和置信度的损失量,α为两者的权重,x 为输入图像,c 为目标类别,l 为预测框,g 为真实标签框.

2 残差网络模型

2.1 残差网络

深度残差网络(ResNet)是在 2015 年 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)大赛上由微软亚洲研究院(MSRA)何凯明团队提出的一种卷积神经网络,该网络赢得了当年图像分类、检测、定位和分割的第 1 名.如图 2 所示,随着网络层数的增加,网络结构的特征提取能力越来越强,识别错误率也越来越低. ResNet 在 ImageNet 数据集上可达到 3.57%的识别错误率,远低于 VGG 网络和人眼实测的错误率,这为后续的残差网络替换提供了研究基础.残差网络最深可达 152 层,与传统网络相比,深度网络带来了更好的泛化能力,同时还具有更低的复杂性.

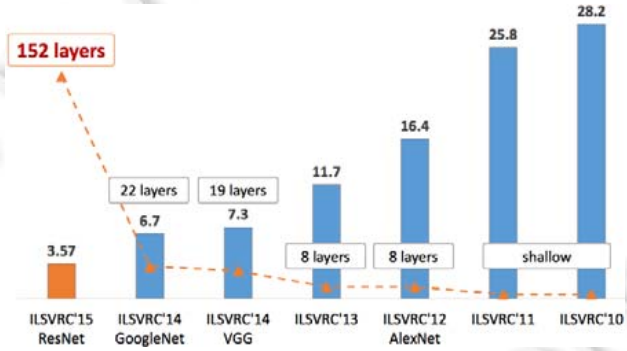


Fig.2 Top-5 error rate (%) of ILSVRC over the years

图 2 ILSVRC 历年的 Top-5 错误率(%)

2.2 残差网络的学习策略

残差网络引入了一种残差学习框架来应对传统网络的退化问题.如图 3 所示,该学习策略对多层的残差映射进行拟合,$H(x)$称为几个网络层堆叠的期望映射,x 为当前堆叠块的入口,relu 激活函数的使用缩短了学习周期.假设 n 个非线性层可近似地表达为某个复杂函数(残差函数),则把堆叠的网络层拟合成另一个映射 $F(x)=H(x)-x$,那么最终的基础映射便为 $H(x)=F(x)+x$.与通过叠加网络层来拟合期望的原始映射相比,尽管这两种方式都能近似得到残差函数,但残差映射更容易调优.通过构建残差学习,残差网络可将多个非线性连接的系数逼近零来近似成更优的期望映射.

图 3 中,公式 $H(x)=F(x)+x$ 可由带有跳跃连接(跳过一层以上的层间连接)的前馈神经网络来实现,跳跃连接执行恒等映射并将计算结果添加至其指向的输出层,这种计算方式没有导入其他系数,计算量没有明显的增加.残差学习框架的引入,可大幅降低提取特征的重复度,减少网络模型的计算量.这种跨层共享参数和重复利用中间特征的方式,可解决层数增加之后出现的性能退化问题.

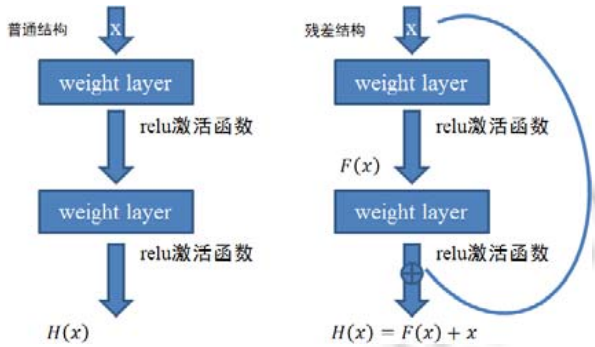
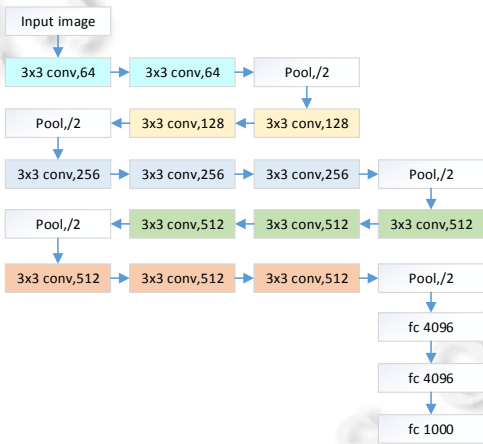


Fig.3 Comparison of common structure and residual structure

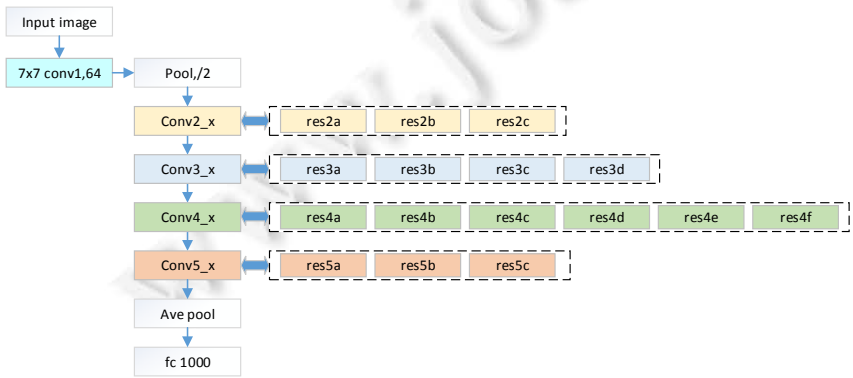
图 3 普通结构和残差结构对比

2.3 残差网络的结构设计

SSD 使用的基准网络 VGG16 的结构如图 4(a)所示,用 3×3 的卷积核来增大网络的感受野范围,用多个包含过滤器的卷积层来减少参数的引入和提高网络的拟合能力.VGG16 共 16 层,网络的前半部分为卷积层的叠加,后半部分为全连接层,最后为进行归一化的 Softmax 层.



(a) VGG-16



(b) 残差网络-50

Fig.4 Comparison of network architectures

图 4 网络结构对比图

图 4(b)所示为 Resnet-50 的网络结构,图中虚线框为不同层块的残差结构,ResNet 中的每个卷积块都包含不同数目的残差单元,每个残差单元进行 3 次卷积操作.残差网络使用身份快捷连接(identity shortcut connection)进行卷积层的跨连,它解决了网络层数加深但检测精度不升反降的问题.与传统 VGG 相比,残差网络具有更少的滤波器和更低的计算代价,这也是将基准网络替换为残差网络的原因.

3 基于残差网络的航拍目标检测算法

3.1 前置网络替换

用于图像分类的标准神经网络称为前置网络(base network),前置网络的理论基础是生成模型,生成模型可自适应地从输入图像中学习重要特征,这在很大程度上解决了某些模型(如传统的全连接网络)特征提取能力不足的问题.但生成模型提取到的特征信息冗余太多,有用信息提取困难.因此,通过前置网络对输入数据进行特征提取,为后续网络层提供输入信息,可加快后续训练速度,提高网络的表达能力.

ResNet 通过引入残差学习来提高模型的检测性能,合并 n 个堆叠块,进而构造一个残差学习模块.构造块定义为

$$y=F(x,W_i)+x \tag{2}$$

其中,x 和 y 分别为当前计算层的输入和输出,函数 $F(x,W_i)$代表当前网络想要学习的残差结构.如图 3 所示,第 1 层公式 $F=W_2\sigma(W_1x)$中,σ为 Relu 激活函数;第 2 层则通过快捷连接来执行 $F+x$ 操作.公式(2)中的输入向量 x 和函数 F 的维度应保持一致,否则,我们需要对输入向量 x 执行线性投影来实现维度匹配:

$$y=F(x,W_i)+W_sx \tag{3}$$

对无人机下视目标图像采集速度和机载硬件计算能力进行综合考虑,本文选用 Resnet-50 残差结构进行网络替换.选取的特征提取层为 conv2_x(分别使用大小为 1×1×64,3×3×64,1×1×256 的卷积核),conv3_x(分别使用大小为 1×1×128,3×3×128,1×1×512 的卷积核),conv4_x(分别使用大小为 1×1×256,3×3×256,1×1×1024 的卷积核),conv5_x(分别使用大小为 1×1×512,3×3×512,1×1×2048 的卷积核),conv7_x,conv8_x,conv9_x.Resnet 中的身份快捷连接没有增加额外计算量,因此,我们可以公正地对原始网络和残差网络进行实验对比.图 5 为原始的 SSD 和经过网络替换的 R-SSD 的网络结构对比图.

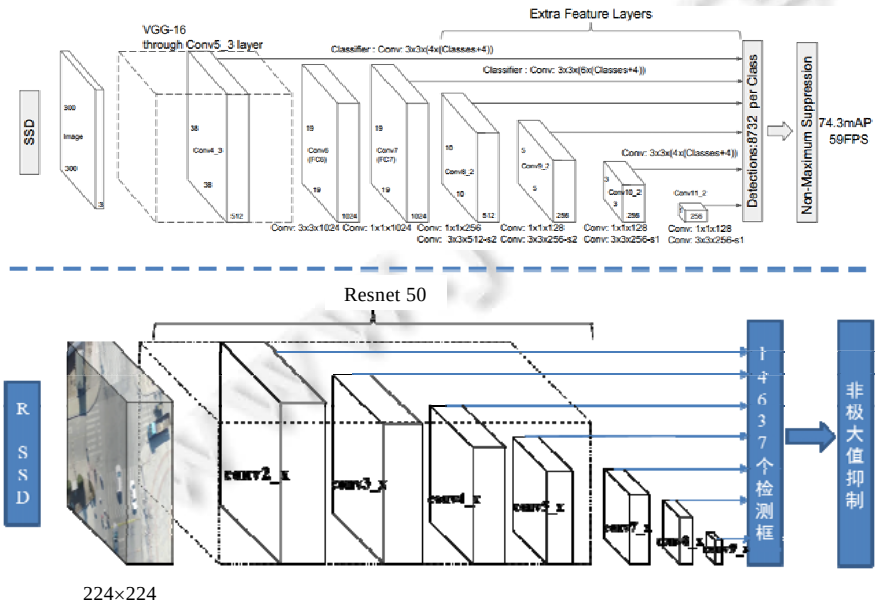


Fig.5 Comparison of SSD network and R-SSD network

图 5 SSD 网络和 R-SSD 网络对比

3.2 训练参数设置

(1) 选择默认框参数

为了能对不同尺度的目标进行正确检测,某些算法将输入图像转为不同的尺度,然后对转换后的图像进行处理,并将检测结果进行融合[25,26].使用若干个不同输出尺寸的特征图进行预测,同样可以得到上述的输出结果,而且在端对端的单一网络中可以进行参数的共享传递,转换效率更高.

在一个卷积神经网络中,位于不同层的特征图有着不同大小的感受域(特征图上输出的某个节点,其对应的输入图像中的某块区域).在此处采用的策略是默认框不用一对一的与特征图的感受域相映射,不同位置的默认框对应不同的区域和目标尺寸.假设用来预测的特征图有 m 个,则每个特征图中默认框的尺寸为

$$S_i = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(i-1), i \in [1, m] \tag{4}$$

其中,$S_{\min}$为网络结构中最底层的默认框尺度,值为 0.2;$S_{\max}$为最高层的默认框尺度,值为 0.95,不同层以一定规则间隔排序.默认框的宽高比取值 $a_r \in \{1,2,3,1/2,1/3\}$,则每一个默认框的宽、高分别为

$$W_i^a = S_i\sqrt{a_r} \tag{5}$$

$$h_i^a = S_i / \sqrt{a_r} \tag{6}$$

默认框的中心点为 $\left(\frac{i+0.5}{|f_k|}, \frac{j+0.5}{|f_k|}\right)$,其中,$i,j \in [0,|f_k|]$,$|f_k|$是第 k 个特征图的尺寸.对具有不同尺寸和宽高比的所有默认框进行特征提取和结果预测,将得到的预测结果进行整合,该集合覆盖待检测目标的各种尺寸和形状,可解决不同尺度的目标检测问题.

(2) 确定匹配策略

在生成 R-SSD 检测模型时,需要为每个真实标签框都选择默认框与其进行匹配.原始的 MultiBox 匹配思想是从所有的候选默认框中为每一个真实标签框找到一个最高的 Jaccard(用于比较样本之间的相似性和差异性)重叠值,该方法确保了每个真实标签框都有一个与其匹配的默认框.本文的匹配策略是在 MultiBox 思想的基础上将 Jaccard 重叠系数调整为 0.5,这一调整弱化了学习过程,允许网络模型自适应地计算多个默认框的重叠情况,而不只限于 Jaccard 重叠率最高的那个默认框.

(3) 选择损失函数

在进行模型训练时,始终存在一个目标函数,算法持续对该函数进行优化,直至损失值最低,这个目标函数称为损失函数.损失函数用来衡量网络模型的输出值 $\hat{y}_i$ 和真实值 y_i 的差异程度,损失函数的目的是使损失值最小化,其公式为

$$L = \sum_{i=1}^{N} l(y_i, \hat{y}_i) \tag{7}$$

本文基于深度学习框架 Caffe[27]建立了 R-SSD 训练模型,通过对比实验选择了 Softmax 作为损失函数,其公式为

$$L_i = -\log\left(\frac{e^{S_{y_i}}}{\sum_j e^{S_j}}\right) = -S_{y_i} + \log\left(\frac{e^{S_{y_i}}}{\sum_j e^{S_j}}\right) \tag{8}$$

其中,S_j 为类别 j 的得分,y_i 为目标的真实标签.不同类别上的目标分值离散程度越高,损失值越低,模型性能越好.针对航拍数据特点,对公式(8)进行如下变形可进一步提高精度:

$$\frac{e^{S_{y_i}}}{\sum_j e^{S_j}} = \frac{Ce^{S_{y_i}}}{C\sum_j e^{S_j}} = \frac{e^{S_{y_i}+\log C}}{\sum_j e^{S_j+\log C}} \tag{9}$$

通过计算损失函数 L_i,得到一个用于分类的 Softmax 模型.

3.3 总体流程图

在训练阶段,首先对输入的目标图像进行预处理操作,包括数据增强和图像去雾等,然后对输入图像中的目

标信息进行标注,得到真实目标的位置信息和类别信息,再进行模型的训练,生成最终的 R-SSD 目标检测模型.

在检测阶段,每枚测试图像都生成 N 个可能包含目标的框图,利用训练阶段生成的 R-SSD 模型对其进行真实坐标偏移和所属类别得分的计算,每枚图像都会得到 N 个分类结果,再利用非极大值抑制算法输出最终结果.整体流程如图 6 所示.

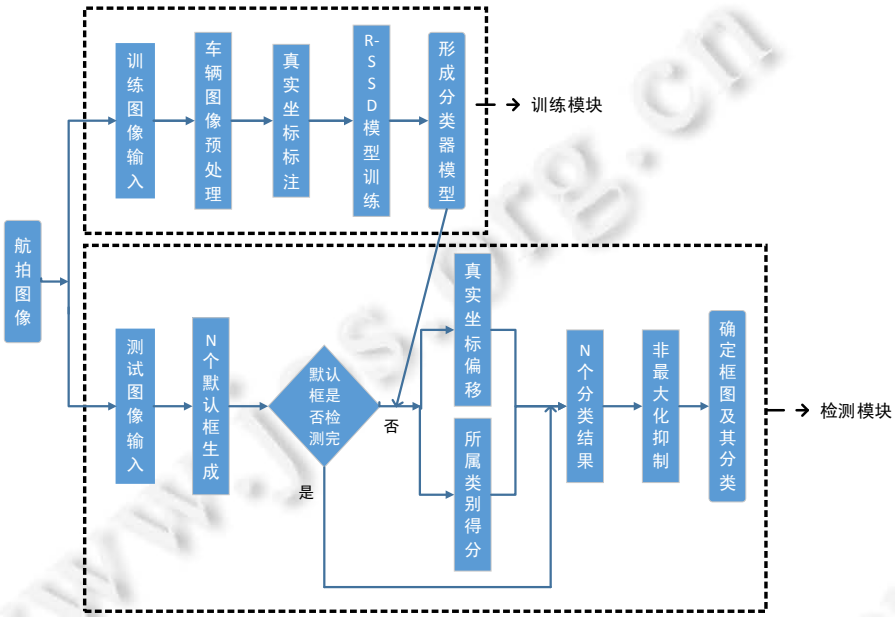


Fig.6 Flow diagram of object detection based on aerial photography

图 6 航拍目标检测流程图

4 基于特征融合的航拍目标检测算法

上一节提出了一种基于残差网络的航拍目标检测算法 R-SSD,使用 Resnet-50 网络替换原始前置网络 VGG-16,在原始 SSD 算法的基础上提高了精度,但实时性差,还存在误检、小目标漏检、重复检测的问题.传统的 SSD 目标检测算法在速度和精度方面表现出色,充分利用了多卷积层的优势来对目标进行检测,对目标的尺度变化具有较好的鲁棒性.但 SSD 网络结构的缺点是对小目标漏检,如图 7 所示,每个卷积层都当作后续分类网络的输入,即每个层对应一个目标的尺度,忽略了层与层之间的关联关系.如图 7 中的 conv4_3 特征层,从该层开始,随着网络层数和深度的增加,卷积层的尺度逐步减小,表征能力越来越强,语义信息也越来越丰富,但底层的 conv4_3 没有利用高层的语义信息,导致检测小目标效果较差.因此,本文利用特征融合技术对该网络进行改进.

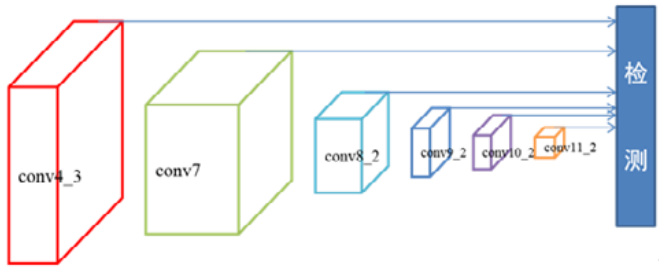


Fig.7 SSD extra feature layer

图 7 SSD 提取特征层

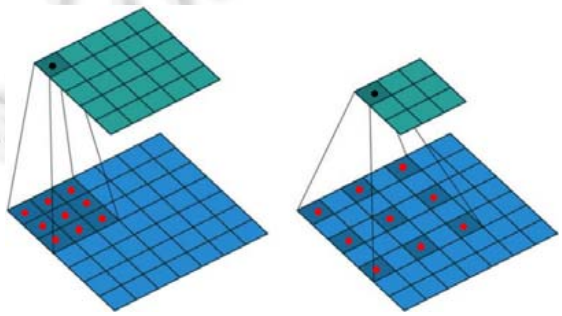
4.1 空洞卷积操作

当把图像输入到用于分割的FCN网络[28]时,FCN网络会先进行卷积操作再进行池化(通过缩小图像尺寸来增大感受野范围)操作,然后将池化后尺寸变小的图像进行上采样增大到原始图像尺寸进行结果预测.但在图像尺寸缩小再增大的过程中,池化层会造成图像部分信息的缺失.如果没有池化层,高层网络中尺寸较小的卷积层其感受野范围也相对较小,缺少图像的整体特征,模型学不到全局信息;如果加上池化层,图像原有的信息特征会遭受损失,降低模型的精度.所以我们采用空洞卷积方法(dilated convolution)[29]来解决这一问题,空洞卷积在不损失信息的前提下加大了卷积层的感受野范围.SSD 结构的缺点在于缺少图像的全局信息,利用空洞卷积进行特征下采样可以改善小目标检测精度不高的问题.

图 8(a)是卷积核大小为 3×3、扩张(dilation)为 1 的空洞卷积操作,该操作等同于卷积操作,3×3 的点状区域为当前卷积的感受野范围.图 8(b)是卷积核大小为 3×3、扩张为 2 的空洞卷积操作,即一个 7×7 的区域但只有 9 个点和 3×3 大小的卷积核发生了卷积操作,其余点的权值为 0.虽然该操作的卷积核大小只有 3×3,但与图 8(a)相比,感受野范围扩大到了 7×7.执行空洞卷积之后感受野的大小为

$$F_{dilation}=[2^{(dilation/2)+2}-1]\times[2^{(dilation/2)+2}-1] \tag{10}$$

其中,扩张值为当前卷积核中每个计算点的半径.如图 8(b)中扩张值为 2,则 $F_{dilation}=7\times7$.



(a) 扩张值为 1 (b) 扩张值为 2

Fig.8 Dilated convolution operation

图 8 空洞卷积操作

本文将图 7 中不同层之间的相互关系考虑在内,较低层的特征图通过空洞卷积操作连接到较高层的特征图上,并对其进行尺度归一操作,保持通道数目不变,改进之后的结构如图 9 所示.

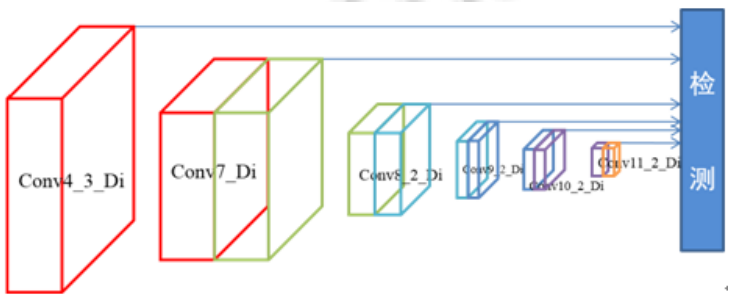


Fig.9 Extra feature layer after dilated convolution

图 9 空洞卷积操作之后的提取特征层

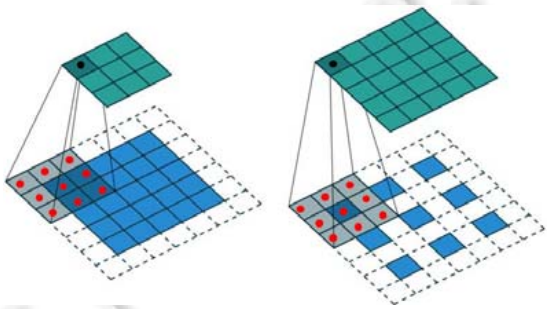
4.2 反卷积操作

为了让训练模型学到更多的上文信息,对分类特征层执行反卷积(deconvolution)[30]操作,卷积操作可以用

来对高维向量进行低维特征的计算.图 10(a)是输入尺寸为 5×5、滤波器大小为 3×3、步长为 2、扩充为 1 的卷积计算过程,输出尺寸为 3×3.反卷积操作刚好相反,它可以将低维的局部特征映射成高维向量,因 SSD 网络结构中高层(低维特征)的特征图含有丰富的语义特征,我们可对其进行反卷积操作映射到低层网络中,用来增强卷积层的表征能力.图 10(b)为卷积操作所对应的反卷积过程,其输入尺寸为 3×3.在给定的特征单元之间进行 0 值的插入上采样,然后采用步长间隔为 1 的 3×3 的滤波器进行反卷积计算,反卷积的输入输出关系为

$$F_{decon}=[s(i-1)+k-2p]\times[s(i-1)+k-2p] \tag{11}$$

其中,s 为移动步长,i 为输入的特征大小,k 为滤波器大小,p 为扩充值.例如图 10(b)中 i 为 3,s 为 1,k 为 3,p 为 0,则 F_{decon}=5×5.



(a) 卷积操作 (b) 反卷积操作

Fig.10 Convolution and deconvolution operation

图 10 卷积与反卷积操作

上采样的方式将语义信息更强的高层特征融入到低层特征图中,增强了网络的辨识度.反卷积操作不仅增加了特征图大小,也使低层特征可以学到更为丰富的语义信息.在原始 SSD 网络的基础上,将较高层的特征图通过反卷积操作连接到较低层的特征图上,并对其进行尺度归一操作,保持通道数目不变,改进之后的结构如图 11 所示.

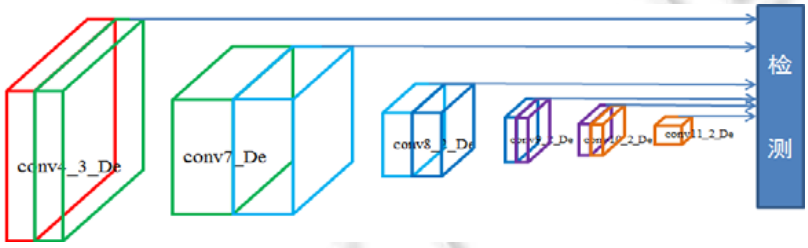


Fig.11 Extra feature layer after deconvolution

图 11 反卷积操作之后的提取特征层

4.3 网络结构

原始 SSD 算法没有计算不同尺度特征层之间的映射关系,在对同一目标进行检测时,SSD 会生成多个不同尺度的预测框,对小目标检测效果差.本文提出的 CI-SSD 网络结构在 SSD 目标检测算法的基础上进行了改进,保留了该算法的前置网络 VGG-16,将 conv4_ci,conv7_ci,conv8_ci,conv9_ci,conv10_ci,conv11_ci 作为预测的特征层.CI-SSD 充分利用了不同特征层之间的相互关系,用空洞卷积操作将低层的特征图和高层的特征图融合,可显著提高分类网络的感受野范围,有利于模型学习到更多的全局信息;反卷积操作将高层的特征图和低层的特征图融合,有助于低层特征层进行小目标的检测,增强了模型的语义表征能力.这种连接方式使 CI-SSD 网络可在同一特征层上将目标的不同尺度考虑在内,增强模型的泛化能力.

如图 12 所示,conv4_ci 特征层由 512 个 38×38 的特征图组成,其中:前 256 个特征图是由 conv4_3 经过卷积运算生成的,所用的卷积核大小为 3×3,步长为 1,扩充为 1,特征图尺度未发生变化;后 256 个特征图是由 conv7 经过反卷积上采样操作生成的,所用的卷积核大小为 2×2,步长为 2,扩充为 0,特征图尺度扩大一倍.

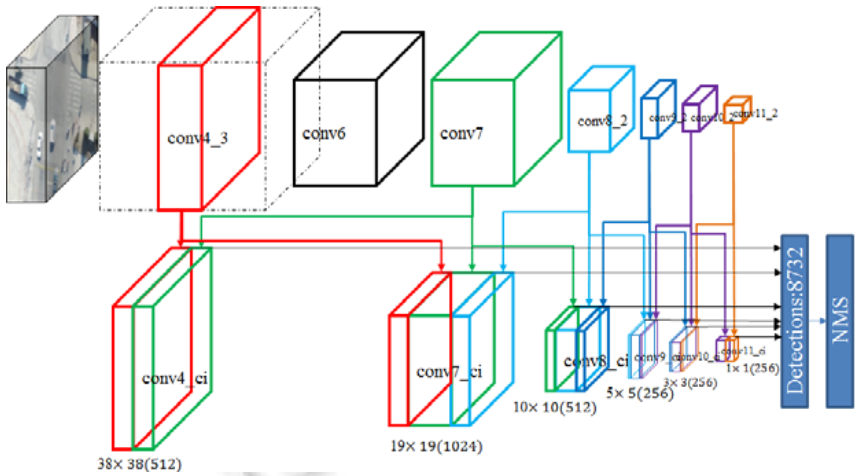


Fig.12 CI-SSD object detection network

图 12 CI-SSD 目标检测网络

conv7_ci 特征层由 1 024 个 19×19 的特征图组成,其中:前 256 个特征图是由 conv4_3 经过空洞卷积下采样运算生成的,所用的扩张值为 2,卷积核大小为 3×3,步长为 2,扩充为 2,特征图尺度减少一倍;中间的 512 个特征图是由 conv7 经过卷积运算生成的,所用的卷积核大小为 3×3,步长为 1,扩充为 1,特征图尺度未发生变化;后 256 个特征图是由 conv8_2 经过反卷积上采样操作生成的,所用的卷积核大小为 3×3,步长为 2,扩充为 1,特征图尺度扩大一倍,其余特征层类似.conv7_ci 的多层融合如图 13 所示.

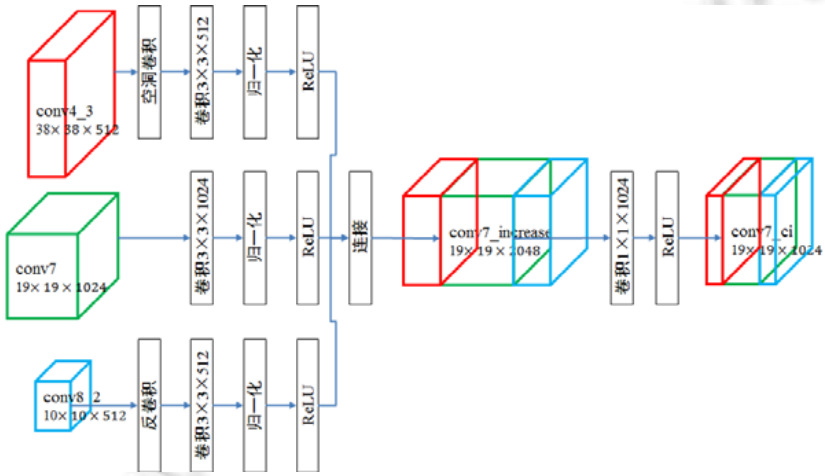


Fig.13 Multi-layer fusion of conv7_ci layer

图 13 conv7_ci 层的多层融合

为使 conv4_3 层和 conv8_2 层的特征图尺寸与 conv7 层相同,我们对 conv4_3 层进行下采样空洞卷积操作,对 conv8_2 层进行上采样反卷积操作,然后使用 3×3 的卷积层学习融合特征.因 VGG-16 基础网络的低特征层与高层数据维度分布差距较大,直接融合效果不好,所以加入 BN 层(batch normalization layer)进行归一化处理,3 个特征图在融合之前进行激活操作,最后使用 1×1 的卷积核进行降维操作,生成最终的特征融合层.

5 实验及结果分析

第 5.1 节介绍实验的运行环境及算法评估指标;第 5.2 节为基于残差网络的航拍目标检测算法 R-SSD 的对比实验;第 5.3 节为基于特征融合的航拍目标检测算法 CI-SSD 的对比实验.

5.1 实验环境及算法评估指标

实验运行环境见表 1.

Table 1 Runtime environment of this experiment

表 1 实验运行环境

类别	环境条件
电脑类型	台式电脑
CPU	Intel(R) Core(TM) i7-6700
显卡	GeForce GTX 1080
内存	8G
操作系统	64 位 Windows 7 旗舰版
深度学习框架	Caffe
CUDA 版本	CUDA 8.0
cuDNN 版本	cuDNN 5.1
运行环境	Visual Studio 2013
脚本语言	Python 2.7.12

本文所用的算法评估指标如下所述:mAP 由精度、召回率和平均值这 3 部分组成.

- 精度 P(precision)也称正确率,广泛应用在信息检索领域.正确率指返回结果中相关类别占总返回结果的比例,定义为正确率=返回结果中相关类别的数目/总返回结果的数目;
- 与正确率一同使用的是召回率 Recall,召回率指返回结果中相关类别占总的相关类别的比例,定义为召回率=返回结果中相关类别的数目/总的相关类别的数目;
- 由正确率和召回率可求出每一类别的 AP 曲线,再对所有类的 AP 取平均值,即可求得 mAP:

$$mAP=\frac{\sum_{k=1}^{N}P(k)\Delta r(k)}{m} \tag{12}$$

其中,N 代表测试集中图像数,$P(k)$表示识别 k 枚图像的精度值,$\Delta r(k)$表示识别图像枚数从 k–1 变化到 k 时 Recall 值的变化量,m 为所有图像的类别数.

本文所用的实验数据来自 UAV123(包括行人和车辆共 13.7G)[31],VEDAI(1 270 枚图像)[32]等公共数据集和大疆 M100 无人机在大连海事大学心海湖附近拍摄的图像数据(1 600 枚),将数据分为汽车、卡车、船、飞机、行人等 5 个类别,训练样本如图 14 所示,其中,训练样本为 19 256 枚,测试样本为 3 000 枚.

(a) 公共数据集 (b) 无人机图像

Fig.14 Training image sample

图 14 训练图像示例

训练参数设置见表 2.

Table 2 Training parameter setting

表 2 训练参数设置

参数名称	取值
base_lr	0.0001
max_iter	60000
lr_policy	“step”
gamma	0.1
momentum	0.9
weight_decay	0.0005
image_size	300×300
type	“SGD”

5.2 基于残差网络的航拍目标检测实验

(1) 前置网络替换实验

前置网络主要用来进行特征提取,并将产生的目标特征传递到后续的卷积层中进行模型训练.针对航拍图像目标尺度小、分辨率低等问题,我们将原始 SSD 算法的前置网络 VGG-16 替换为 Resnet50,对输入图像进行归一化处理,并增加特征提取层数来提高特征提取能力.前置网络 VGG-16 和 Resnet50 的具体结构见表 3,其中,每个单元块的值为选择的特征提取层和对应的输出尺寸.

Table 3 Pre-network parameter comparison table

表 3 前置网络参数对比表

比较类别	VGG-16	Resnet50
输入图像尺寸	300×300	224×224
第 1 级特征层		conv2_x,56×56
第 2 级特征层	Conv4_3,38×38	conv3_x,28×28
第 3 级特征层	Conv7,19×19	conv4_x,14×14
第 4 级特征层	Conv8_2,10×10	conv5_x,7×7
第 5 级特征层	Conv9_2,5×5	conv7_x,4×4
第 6 级特征层	Conv10_2,3×3	conv8_x,2×2
第 7 级特征层	Conv11_2,1×1	conv9_x,1×1
检测框个数	8 732	14 637

如图 15 所示,使用 Resnet50 网络的模型与原始 SSD 算法相比具有更高的 mAP 值,尤其在卡车、飞机这两类数据集上精度提升比较明显.

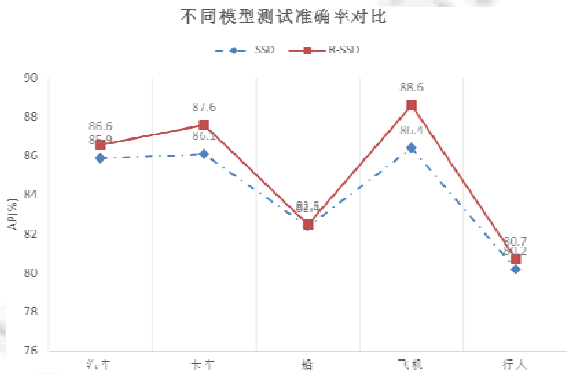


Fig.15 Comparison of detection accuracy of different models

图 15 不同模型的检测准确率对比

表 4 统计了使用不同基础网络(VGG-16 和 Resnet50)的两种算法在无人机数据集上的目标检测结果,测试集为包含 5 大类别的 3 000 枚图像.使用 Resnet50 作为前置网络的 R-SSD 模型取得了 85.2%的 mAP.两个模型

在检测较大物体时(如飞机、卡车等)都有较高的准确率,在识别飞机这一类别上,R-SSD 模型达到了最高的 mAP 值,为 88.6%,高于 SSD 模型的 86.4%,提高了 2.2%.这是因为 Resnet50 网络层更深,特征提取能力更强,检测效果也更好.对于行人这一类,两个模型的表现都不是很好,因无人机下拍摄的行人目标较小,发生形变,不利于特征的提取和表达.

Table 4 R-SSD object detection results

表 4 R-SSD 目标检测结果

模型	基础网络	mAP(%)	汽车(%)	卡车(%)	船(%)	飞机(%)	行人(%)
SSD	VGG-16	84.2	85.9	86.1	82.4	86.4	80.2
R-SSD	Resnet50	85.2	86.6	87.6	82.5	88.6	80.7

(2) 默认框参数实验

默认框的参数设置直接影响着模型处理不同尺度目标的检测性能,默认框横宽比 r 的分布也会影响目标的检测准确率.当 r 分布较为集中时,网络计算负担加重但检测精度却没有明显的提升;当 r 分布较为分散时,模型的表征学习能力不足.为了测试不同尺寸和横宽比的默认框对模型的影响,本文设计了如下实验.

如图 16 所示,横坐标为使用的默认框横宽比的集合,例如,r=[1/2,1,2]时表示针对当前输入图像采用的默认框的横宽比分别为 1/2,1,2(如图 17 所示,实线框为目标的真实坐标框,虚线框为选取的默认框的范围).因 R-SSD 模型使用了 7 个卷积层作为后续目标分类网络的输入,所以图 16 中的 4 条折线分别代表了当前选取的卷积层使用的不同默认框的个数,其中,折线[3×7]代表 7 个卷积层中默认框的个数都为 3,即默认框横宽比分布为[1/2,1,2].当默认框个数 n 和默认框横宽比分布 r 取值为([3,5×6],[1/3,1/2,1,2,3])和([5×7],[1/3,1/2,1,2,3])时 mAP 值较高,准确率分别为 85.2%和 85.4%.但采用[5×7]分布的模型与[3,5×6]相比需额外计算 6 272 个检测框,这增加了计算复杂度,但性能却只提高了 0.2%,得不偿失.因此,本文选择的默认框参数见表 5.

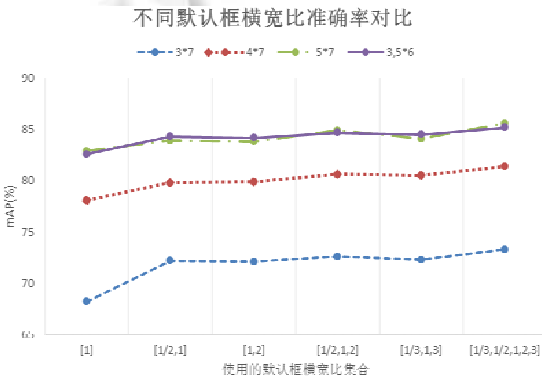


Fig.16 Comparison of detection accuracy of different default boxes aspect ratio

图 16 不同默认框横宽比的检测准确率对比

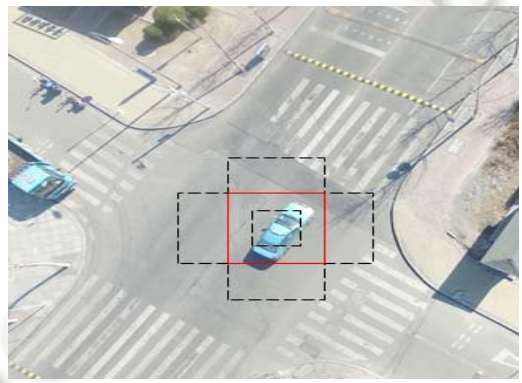


Fig.17 Sample of default boxes aspect ratio

图 17 默认框横宽比样例

Table 5 Model detection results

表 5 默认框的数量及横宽比

卷积层	默认框个数	默认框横宽比	默认框尺寸
conv2_x	3	[1/2,1,2]	56×56
conv3_x	5	[1/3,1/2,1,2,3]	28×28
conv4_x	5	[1/3,1/2,1,2,3]	14×14
conv5_x	5	[1/3,1/2,1,2,3]	7×7
conv7_x	5	[1/3,1/2,1,2,3]	4×4
conv8_x	5	[1/3,1/2,1,2,3]	2×2
conv9_x	5	[1/3,1/2,1,2,3]	1×1

(3) 综合性能对比

为了综合评估模型的检测能力,本文将 R-SSD 模型与 SIFT+SVM 和 Faster R-CNN 等目前较为流行的检测算法进行对比,得到的结果见表 6.

Table 6 Accuracy comparison of different methods

表 6 不同方法的准确率对比

方法	基础网络	原始数据准确率(%)	数据增强准确率(%)	FPS(GTX 1080)
SIFT+SVM	传统方法	61.9	62.4	40
Faster R-CNN[3]	VGGNet	73.9	76.1	6
SSD[7]	VGG-16	79.3	84.2	**46**
R-SSD	Resnet50	**82.5**	**85.2**	11

从表 6 可以看出:本文改进的方法无论在原始数据集(未进行数据增强)还是在增强的数据集上都取得了较好的检测效果,分别取得了 82.5%和 85.2%的准确率,准确率比传统方法高出近 30 个百分点.但在速度方面,R-SSD 没有 SSD 和传统方法速度快,这是由于 R-SSD 为了提高特征提取能力,增加了特征提取层数,牺牲了算法速度.图 18 为 R-SSD 算法的部分实验截图,图中的矩形框为模型预测的目标全局位置,矩形框左上方为预测的类别和分值.

Fig.18 Detection result of R-SSD algorithm

图 18 R-SSD 算法检测结果图

图 19 分别为 R-SSD 算法的目标误检图、小样本漏检图和目标重复检测图.

(a) 目标误检

(b) 小目标检测(黑框为小目标,难检测,手动标注)

(c) 重复检测

Fig.19 Error detection result of R-SSD algorithm

图 19 R-SSD 算法的误检结果图

5.3 基于特征融合的航拍目标检测实验

(1) 目标类别检测实验

CI-SSD 算法的类别检测结果如图 20 所示,与传统的 SSD 算法相比,CI-SSD 目标检测算法在各个类别的检测精度上均有了大幅度的提升,其中,行人类别的准确率提升最为明显,提高了 6%.这是因为 CI-SSD 网络结构融合了高层特征向量的语义信息和低层特征向量的位置和边缘信息,使得模型在保持原有检测精度的前提下对行人等小目标检测具有更强的适应性.

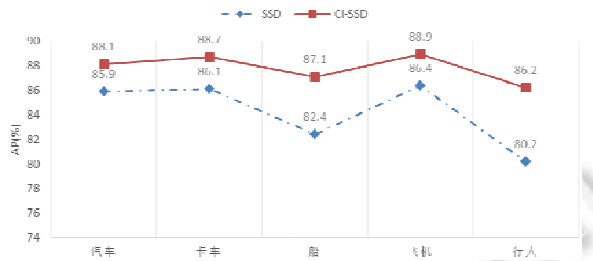


Fig.20 Comparison of detection accuracy of different models

图 20 不同模型的检测准确率对比

表 7 统计了 SSD 和 CI-SSD 两种算法在无人机数据集上的目标检测结果.本文的 CI-SSD 目标检测算法的准确率达到了 87.8%,较 SSD 算法提高了 3.6%,较上一节的 R-SSD 算法提高了 2.6%.

Table 7 SSD and CI-SSD object detection results

表 7 SSD 和 CI-SSD 目标检测结果

模型	基础网络	mAP(%)	汽车(%)	卡车(%)	船(%)	飞机(%)	行人(%)
SSD	VGG-16	84.2	85.9	86.1	82.4	86.4	80.2
CI-SSD	VGG-16	**87.8**	88.1	88.7	87.1	88.9	86.2

(2) 特征融合实验

为验证本文特征融合的有效性,设计了以下几组对比实验.

- 第 1 组为 SSD-Diconv:在 CI-SSD 网络结构的基础上去掉反卷积上采样,保留特征图空洞卷积下采样操作,网络结构如图 9 所示;
- 第 2 组为 SSD-Deconv:在 CI-SSD 网络结构的基础上去掉空洞卷积下采样,保留特征图反卷积上采样操作,网络结构如图 11 所示;
- 第 3 组为 SSD-Pooling:在 CI-SSD 网络结构的基础上去掉反卷积上采样和空洞卷积下采样操作,用池化层进行特征图下采样操作.

改进后的 CI-SSD 算法与 SSD-Diconv,SSD-Deconv,SSD-Pooling 结果对比如图 21 所示.

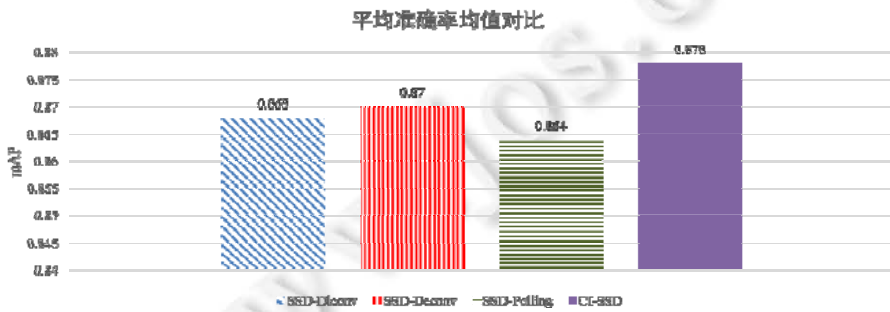


Fig.21 Comparison of experimental results

图 21 实验结果对比

由图 21 可以看出,SSD-Diconv,SSD-Deconv 和 SSD-Pooling 在检测精度方面表现均优于 SSD,其中,SSD-Deconv 表现最为优秀,达到 87%的准确率,这说明高层的语义和低层的边缘纹理等信息均能提高模型的检测精度.其中,SSD-Deconv 精度高于 SSD-Pooling,说明本文采用的空洞卷积操作与池化操作相比,在进行特征融合时保存了更多的图像信息.实验结果表明,本文提出的基于特征融合的航拍目标检测算法 CI-SSD 准确率最高,为 87.8%.

(3) 综合性能对比

为了评估 CI-SSD 算法的综合性能,本节将 CI-SSD 与 SSD,R-SSD 进行对比实验,结果见表 8.本文改进的方法无论在原始数据集上还是在增强数据集上都取得了最高的检测精度,分别为 84.1%和 87.8%.在速度方面,因需要进行不同特征层的信息融合,速度略有下降,但高于 R-SSD 算法的处理速度,满足实时性的要求.

Table 8 Comprehensive comparison of different methods

表 8 不同方法的综合对比

方法	基础网络	原始数据准确率(%)	数据增强准确率(%)	FPS(GTX 1080)
SSD	VGG-16	79.3	84.2	**46**
R-SSD	Resnet50	82.5	85.2	11
CI-SSD	VGG-16	**84.1**	**87.8**	39

图 22 为部分实验截图,测试数据集与上节相同,图中的矩形框为模型预测的目标位置,矩形框左上方为预测的类别和分值,不同的边框颜色代表不同的目标分类.

(a) 目标误检纠正

(b) 重复检测纠正

(c) 重复检测+小目标检测纠正

Fig.22 Comparison of detection result between R-SSD and CI-SSD

图 22 R-SSD 和 CI-SSD 的检测结果对比

(d) 小目标检测纠正

Fig.22 Comparison of detection result between R-SSD and CI-SSD (Continued)

图 22 R-SSD 和 CI-SSD 的检测结果对比(续)

图 22(a)中,R-SSD 误将地标建筑检测成车,CI-SSD 纠正了这一误标.在图 22(b)中,针对卡车这一目标,R-SSD 将其检测为卡车和车,出现了重复检测,而 CI-SSD 没有出现该错误.图 22(d)中,R-SSD 漏检了上方小车,CI-SSD 成功将其检测成车.与传统 SSD 目标检测算法相比,CI-SSD 算法检测精度更高,尤其在小目标物体的检测上更有优势.实验结果表明,本文改进的算法有效提高了无人机图像中目标检测的准确率.

6 总 结

针对无人机目标检测分辨率低、遮挡、小目标漏检、重复检测、误检等精度低下问题,本文在 SSD 算法的基础上,用表征能力更强的残差网络进行基准网络的替换,用残差学习降低网络训练难度,提高目标检测精度;引入跳跃连接机制降低提取特征的冗余度,解决层数增加出现的性能退化问题;引入不同分类层的特征融合机制,把网络结构中低层视觉特征与高层语义特征有机地结合在一起.算法的准确率达到了 87.8%,较 SSD 算法提高了 3.6%.实验结果表明,图像预处理、特征融合能够提高目标检测的精度,满足实时性要求;增加网络层次和深度虽能提高目标检测的精度,但是计算量的增加严重影响了目标检测实时性.接下来,将裁减基础网络,优化特征融合的程度,以期进一步提高检测精度和实时性,促进无人机核心技术的快速发展.

References:

[1] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: O'Conner L, ed. Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus, Ohio: IEEE Computer Society, 2014. 580−587.

[2] Girshick R. Fast R-CNN. In: O'Conner L, ed. Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE Computer Society, 2015. 1440−1448.

[3] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,39(6):1137−1149.

[4] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: O'Conner L, ed. Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE Computer Society, 2018. 2980−2988.

[5] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: O'Conner L, ed. Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2016. 779−788.

[6] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: O'Conner L, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017. 6517−6525.

[7] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot MultiBox detector. In: Leibe B, ed. Proc. of the 2016 European Conf. on Computer Vision. Amsterdam: Springer Int'l Publishing, 2016. 21−37.

[8] Shen ZQ, Liu Z, Li JG, Jiang YG, Chen YR, Xue XY. DSOD: Learning deeply supervised object detectors from scratch. In: O'Conner L, ed. Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE Computer Society, 2017. 1937−1945.

[9] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: O'Conner L, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE Computer Society, 2017. 2261−2269.

[10] Fang LP, He HJ, Zhou GM. Research overview of object detection methods. Computer Engineering and Applications, 2018,54(13): 11−18,33 (in Chinese with English abstract).

[11] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection. In: Proc. of the British Machine Vision Conf. (BMVC). 2017. https://arxiv.org/abs/1705.09587

[12] Redmon J, Farhadi A. YOLOv3: An incremental improvement. https://arxiv.org/abs/1804.02767

[13] Ali S, Shah M. COCOA: Tracking in aerial imagery. In: Proc. of the Society of Photo-optical Instrumentation Engineers (SPIE). Florida, 2006. 6209. http://spie.org/Publications/Proceedings/Paper/10.1117/12.667266

[14] Ibrahim AWN, Pang WC, Seet GLG, Lau WSM, Czajewski W. Moving objects detection and tracking framework for UAV-based surveillance. In: Werner B, ed. Proc. of the 2010 4th Pacific-rim Symp. on Image and Video Technology. Singapore: IEEE Computer Society, 2010. 456−461.

[15] Tong XM. The research of moving object detection and tracking methods based on aerial video [Ph.D. Thesis]. Xi'an: Northwestern Polytechnical University, 2015 (in Chinese with English abstract).

[16] Zhang H. Researches on UAV based moving targets detection and tracking and vision aided UAV landing system [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2008 (in Chinese with English abstract).

[17] Tan X, Yu XC, Liu JZ, Huang WJ. Object fast tracking based on unmanned aerial vehicle video. Bulletin of Surveying and Mapping, 2011,(9):32−34,41 (in Chinese with English abstract).

[18] Dong J, Fu D, Yang X. Real-time moving object detection and tracking by using UAV videos. Journal of Applied Optics, 2013, 34(2):255−259 (in Chinese with English abstract).

[19] Tang Y, Zhou PC, Xiao X, Chang C, Liu YL, Pan F. Researches of moving targets detection and tracking algorithm based on UAV. Robot Technique and Application, 2017(3):35−37 (in Chinese with English abstract).

[20] Rani LD, Prasad CGVN, Rao CK. Aerial image analysis using dynamic bayesian network. Int'l Journal of Research, 2014,1(8): 909−915.

[21] Teutsch M, Kruger W. Detection, segmentation, and tracking of moving objects in UAV videos. In: O'Conner L, ed. Proc. of the IEEE 9th Int'l Conf. on Advanced Video and Signal-based Surveillance. Beijing: IEEE Computer Society, 2012. 313−318.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

[23] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: O'Conner L, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE Computer Society, 2016. 770−778.

[24] Szegedy C, Reed S, Erhan D, Erhan D, Anguelov D, Loffe S. Scalable high quality object detection. https://arxiv.org/abs/1412. 1441

[25] Sermanet P, Eigen D, Zhang X, Michael M, Fergus R, LeCun Y. OverFeat: Integrated recognition, localization and detection using convolutional networks. https://arxiv.org/abs/1312.6229

[26] He KM, Zhang XY, Ren SQ, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(9):1904−1916.

[27] Jia YQ, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. 2014. 675−678. https://dl.acm.org/citation.cfm?id= 2654889

[28] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: O'Conner L, ed. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE Computer Society, 2015. 3431−3440.

[29] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proc. of the Int'l Conf. on Learning Representation (ICLR). 2016. https://arxiv.org/abs/1511.07122

[30] Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional networks. In: O'Conner L, ed. Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE Computer Society, 2010. 2528−2535.

[31] https://ivul.kaust.edu.sa/Pages/Dataset-UAV123.aspx

[32] https://downloads.greyc.fr/vedai/

附中文参考文献:

[10] 方路平,何杭江,周国民.目标检测算法研究综述.计算机工程与应用,2018,54(13):11−18,33.

[15] 仝小敏.航拍视频运动目标检测与跟踪方法研究[博士学位论文].西安:西北工业大学,2015.

[16] 张恒.无人机平台运动目标检测与跟踪及其视觉辅助着陆系统研究[博士学位论文].长沙:国防科技大学,2008.

[17] 谭熊,余旭初,刘景正,黄伟杰.基于无人机视频的运动目标快速跟踪.测绘通报,2011(9):32−34,41.

[18] 董晶,傅丹,杨夏.无人机视频运动目标实时检测及跟踪.应用光学,2013,34(2):255−259.

[19] 汤轶,周鹏程,肖璇,常成,刘益麟,潘峰.基于无人机平台的运动目标检测与跟踪算法研究.机器人技术与应用,2017(3):35−37.

裴伟(1977－),男,山东枣庄人,博士,副教授,CCF 专业会员,主要研究领域为人工智能,图像处理,模式识别.

王鹏乾(1997－),男,学士,主要研究领域为通信技术,信号传输.

许晏铭(1992－),男,硕士,主要研究领域为图像处理.

鲁明羽(1963－),男,博士,教授,博士生导师,主要研究领域为人工智能,模式识别.

朱永英(1978－),女,博士,副教授,主要研究领域为模式识别,海洋环境.

李飞(1995－),女,硕士,CCF 学生会员,主要研究领域为图像处理.