基于改进 SSD 的轻量化小目标检测算法

吴天舒[1,2],张志佳[1*],刘云鹏[2],裴文慧[1],陈红叶[1]

(1. 沈阳工业大学 软件学院,辽宁 沈阳 110870;
2. 中国科学院沈阳自动化研究所,辽宁 沈阳 110016)

摘 要: 为提高 SSD 目标检测算法的小目标检测能力,提出在 SSD 算法中引入转置卷积结构,采用转置卷积将低分辨率高语义信息特征图与高分辨率低语义信息特征图相融合,增加低层特征提取能力,提高 SSD 算法的平均精准度。同时针对 SSD 算法存在模型过大,运行内存占用量过高,无法在嵌入式 ARM 设备上运行的问题,以 DenseNet 为基础,结合深度可分离卷积,逐点分组卷积与通道重排提出轻量化特征提取最小单元,将 SSD 算法特征提取部分替换为轻量化特征提取最小单元的组合后,可在嵌入式 ARM 设备上运行。在 PASCAL VOC 数据集和 KITTI 自动驾驶数据集上进行对比实验,结果表明改进后的网络结构在平均精准度上得到明显提升,模型参数数量得到有效降低。

关键词: 目标检测; 转置卷积; 深度可分离卷积; 嵌入式; PASCAL VOC 数据集; KITTI 数据集

中图分类号: TP391.4 **文献标志码:** A **DOI:** 10.3788/IRLA201847.0703005

A lightweight small object detection algorithm based on improved SSD

Wu Tianshu[1,2], Zhang Zhijia[1*], Liu Yunpeng[2], Pei Wenhui[1], Chen Hongye[1]

(1. School of Software, Shenyang University of Technology, Shenyang 110870, China;
2. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: In order to improve the small object detection ability of SSD object detection algorithm, the transposed convolution structure in SSD algorithm was proposed, the low resolution high semantic information feature map was integrated with high resolution low semantic information feature map using transposed convolution, which increased the ability of low level feature extraction and improved the average accuracy of SSD algorithm. At the same time for the problem that SSD algorithm model being large, running memory consumption high, without running on the embedded equipment ARM, a lightweight feature extraction minimum unit was proposed based on DenseNet, combining depthwise separable convolutions, pointwise group convolution and channel shuffle, running on the embedded equipment ARM cloud be realized. The comparative experiments on PASCAL VOC data set and KITTI autopilot data set show that the mean average is significantly improved by improved network structure,

收稿日期: 2018-02-10; **修订日期:** 2018-03-20

基金项目: 国家自然科学基金(61540069);装发部共用技术课题项目(Y6k4250401)

作者简介: 吴天舒(1993-),男,硕士生,主要从事目标检测与深度学习性能优化等方面的研究。Email:747112077@qq.com

通讯作者: 张志佳(1974-),男,教授,博士生导师,主要从事机器视觉检测技术、图像处理与模式识别方面的研究。Email:zhangzj@sut.edu.cn

and the number of model parameters is effectively reduced.

Key words: object detection; transposed convolution; depthwise separable convolution; embedded; PASCAL VOC data set; KITTI data set

0 引 言

目标检测是计算机视觉领域的重要研究方向之一，传统的目标检测方法是通过构建特征描述子提取特征后利用分类器对特征进行分类实现目标检测，如梯度方向直方图HOG (Histogram of Oriented Gradient) 和支持向量机 SVM(Support Vector Machine)[1]。随着深度学习在图像分类领域的优异表现，卷积神经网络在计算机视觉的各领域开始得到了广泛使用[2]，在目标检测领域中使用深度学习实现目标检测成为一个新的方向。

2015 年 Shaoqing Ren 等提出 Faster R-CNN 深度学习目标检测算法，在平均精准度 mAP(Mean Average Precision) 上高于特征描述子与分类器结合的传统方法，但 Faster R-CNN 存在检测速度慢的问题[3]。

2016 年，Joseph Redmon 等在 CVPR (IEEE Conference on Computer Vision and Pattern Recognition)会议上提出了目标检测算法 YOLO[4]。同年 Wei Liu 等在 ECCV 会议上 (European Conference on Computer Vision) 提出了目标检测算法 SSD[5]。YOLO 与 SSD 通过回归的方式完成目标检测，使利用深度学习的方式进行目标检测达到实时的检测速度。

Faster R-CNN 属于两步目标检测算法，通过分类与回归完成目标检测[6]。SSD 属于单步目标检测算法，通过回归直接实现目标检测，但 SSD 在达到高速的同时存在小目标检测能力差的问题。而在实际工程应用中待检测目标占图像比例较小的情况更为普遍。

SSD 虽然可以在 Titan X 等 GPU 服务器上达到实时的效果，但 SSD 模型参数过多，运行内存占用量过大，在显存容量较小的 GPU 设备或 ARM 等移动嵌入式设备上无法运行[7]。

文中以 SSD 为基础，借鉴 FPN 的网络结构对 SSD 进行改进[8]。在特征提取部分，利用深度可分离卷积，逐点分组卷积与通道重排提出轻量化特征提取最小单元[9-10]，达到减少网络模型参数数量、降低模型运行内存占用量和加速算法运行速度的目的。

文中提出的算法已完成程序实现，程序已开源在 github：https://github.com/canteen –man/MobileNet –SSD–Focal–loss。

1 改进SSD 检测模型

SSD 属于单步目标检测算法，在多尺度特征图上以回归的方式得到目标的类别和位置，但存在小目标检测能力弱的问题。文中分析导致 SSD 小目标检测能力弱的原因，并在 SSD 中引入转置卷积结构，提升 SSD 小目标检测能力。

1.1 SSD 小目标检测能力分析

若提升小目标检测检测能力可通过图像金子塔的方式形成多尺度目标检测，但卷积神经网络在特征提取部分计算量过大，如采用图像金字塔提升小目标检测能力将会增加大量计算量[11]。

SSD 以回归的方式得到目标的类别和位置。选取网络结构中六层不同尺度的特征图，对这六层特征图取不同尺寸的候选框。六层特征图通过回归得出目标的类别置信度和候选框与真实值之间的偏差[5]。尽管 SSD 采用多尺度特征图回归在特征图上实现类似图像金字塔的效果，有利于多尺度目标检测，但 SSD 依旧存在小目标检测能力弱的问题，提升 SSD 小目标检测能力将进一步提升 mAp。

SSD 在各检测层生成不同尺寸的预测框，选取 IOU(Intersection Over Union)大于 0.5 的预测框作为正样本，IOU 小于 0.5 的预测框作为负样本。因此大目标物体上覆盖 IOU 大于 0.5 的预测框多，正负样本均衡。而小目标上覆盖 IOU 大于 0.5 的预测框少，导致小目标物体正负样数量失衡，不利于小目标物体训练。Lin T Y 等在 2017 年 ICCV (IEEE International Conference on Computer Vision)会议上提出的 Focal Loss 可以有效解决正负样本不均

衡的问题。

文中进一步分析得出小目标检测能力弱的原因。当 SSD 输入图像分辨率为 300×300 时，38×38 的特征图由于其分辨率较高，主要用来检测图像中的小目标，但其特征表达能力仅取决于前 10 层卷积层，容易造成欠拟合的问题。而若选择在 38×38 特征图前加入大量卷积容易造成后续特征图过拟合的问题，出现增加高分辨率特征图特征提取能力时，低分辨率特征图容易过拟合的矛盾。

所以为提高小目标物体检测能力，特征图应一方面具备足够高的分辨率，以得到小目标的位置信息，另一方面具备更全局的信息及更强的特征提取能力，以学习小目标的类别信息。

1.2 引入转置卷积结构

文中结合 SSD 目标检测算法多尺度特征图检测的特点，通过转置卷积在特征图上实现多尺度特征融合。

文中采用转置卷积对特征图进行上采样，将低维特征映射成高维输入，与卷积操作的作用相反。转置卷积示意图如图 1 所示。

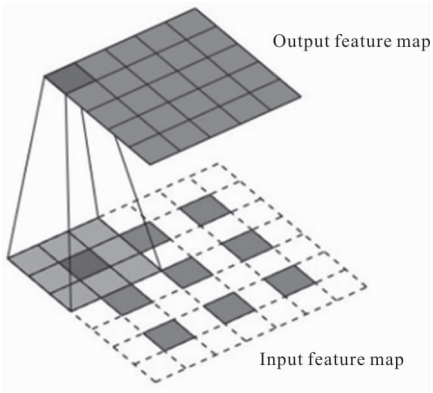

图 1 转置卷积示意图

Fig.1 Schematic diagram of deconvolution

转置卷积又称逆卷积，即为卷积神经网络中相对于卷积的逆过程。将特征图展为一维向量，卷积核展为稀疏矩阵，转置卷积运算即为矩阵乘法。转置卷积计算如公式(1)所示：

$$C^{\mathrm{T}}\cdot\vec{Y}=\vec{X} \tag{1}$$

式中：$\vec{X}$和$\vec{Y}$表示特征图展开得到的一维向量；C 表示卷积核展成的稀疏矩阵。

转置卷积的作用与上采样相同，但不同于固定参数值的上采样。转置卷积即为卷积神经网络中普通卷积前向传播的逆过程，将输入特征图与输出特征图位置进行互换，所以转置卷积的卷积核参数同普通卷积卷积核参数相同，可在训练过程中经过反向传播学习调整，使上采样参数更加合理。

通过转置卷积使高语义信息特征图与低语义信息特征图分辨率相同，将两种特征图拼接成多通道特征图。利用多通道卷积对经拼接而成的多通道特征图提取特征，实现特征融合。因为卷积核参数可通过反向传播学习调整，所以利用多通道卷积实现特征融合较直接对特征图相加实现特征融合的方式更有效。

改进后的网络结构低层高分辨率特征图具备更全局的信息与更强的拟合能力，同时高层特征拟合能力不变，不会产生过拟合的问题，转置卷积与特征融合过程如图 2 所示。

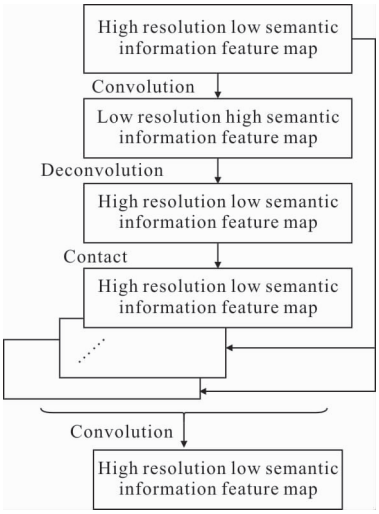

图 2 转置卷积与特征融合

Fig.2 Deconvolution and feature fusion

因为在 SSD 中，中型尺寸与大型尺寸目标检测能力较好，而小型尺寸目标上存在检测能力弱，所以文中只针对低层特征图进行改进，以避免引入过多转置卷积而增加计算量。当输入图像分辨率为 300×300 时，将 75×75 高分辨率低语义信息特征图添加进检测层。分别将分辨率为 19×19 与 38×38 的特征图做转置卷积，特征提取能力较原先的 7 层卷积和 10 层卷积层分别提升至 16 层卷积层和 17 层卷积层。

对低分辨率特征做转置卷积后，其特征图分辨率需与高分辨率特征图保持一致才能完成特征图合并。根据转置卷积特征图分辨率计算公式，结合SSD 网络结构实际特征图分辨率情况，文中对 19×19 分辨率的特征图采用卷积核为 2×2，步长为 2 的转置卷积参数，对 38×38 分辨率的特征图的转置卷积参数为 3×3 卷积核，步长为 2，扩充边缘为 1，转置卷积后特征图分辨率计算公式如公式(2)所示：

$$O=S\times(L-1)+H-2\times P \tag{2}$$

式中：O 为转置卷积输出特征图分辨率；S 为步长；L 表示输入特征图分辨率；H 表示卷积核尺寸；P 表示边缘补充尺寸。

但由于在网络结构中引入两层转置卷积，并利用两层卷积实现特征融合，所以增大了网络模型的参数数量与计算量。通过在特征提取部分使用文中提出的轻量化特征提取最小单元，会弥补引入转置卷积结构带来的计算量增加的问题。

2 特征提取轻量化设计

文中以 Densenet 网络结构为基础，结合深度可分离卷积与通道重排设计出轻量化特征提取最小单元，通过更改前端特征提取模型，降低网络结构中参数数量，节省运行内存占用量，加快模型运行速度。

2.1 深度可分离卷积与通道重排

文中使用深度可分离卷积与逐点分组卷积组合的方式代替 SSD 前端特征提取模型中的传统卷积。深度可分离卷积计算过程如公式(3)所示[10]：

$$G_{k,l,m}=\sum_{i,j}K_{i,j,m}\cdot F_{k+i-1,\ l+j-1,m} \tag{3}$$

式中：G 为输出特征图；K 为卷积核；F 为输入特征图；i,j 为特征图像素位置；k,l 为输出特征图分辨率；m 为通道数。

深度可分离卷积仅能提取对应特征图的特征，所以在深度可分离卷积后利用逐点分组卷积，对各个通道特征图进行特征融合。深度可分离卷积与逐点分组卷积组合结构较传统卷积所降低的参数量比例可以定义为公式(4)：

$$\frac{K\times K\times M+M\times\frac{N}{G}}{K\times K\times M\times N}=\frac{1}{N}+\frac{1}{GK^2} \tag{4}$$

式中：K 为深度可分离卷积的卷积核尺寸；N 为逐点分组卷积的卷积核数量；M 为输入特征图数量；G 为逐点分组卷积的分组数。由公式(4)可得出结论，输出特征图数量越多，逐点分组卷积的分组数目越多，较传统卷积参数数量压缩率越大。

但深度可分离卷积与逐点分组卷积均属于组卷积，容易造成通道间信息相互独立的情况，缺少通道间的特征融合。所以在进行深度可分离卷积计算前先进行通道重排，将不同组特征图进行交叉重排后再进行深度可分离卷积与逐点分组卷积，有利于特征提取能力的提升[9]，通道重排过程如图 3 所示。

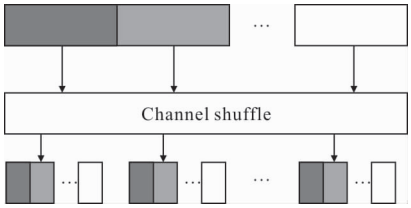

图 3 通道重排

Fig.3 Channel shuffle

2.2 轻量化特征提取最小单元

文中以 Densenet 网络结构为基础设计特征提取的最小单元[12]，在特征提取的最小单元中首先使用逐点分组卷积对输入特征图进行降维，以便减少深度可分离卷积的计算量。然后利用通道重排对多通道间特征图交叉重排，最后利用深度可分离卷积与逐点分组卷积的组合完成特征提取，与输入特征图进行特征拼接。特征拼接方式为将特征提取最小单元的输出特征图与输入特征图组合成为多通道特征图。再通过多通道卷积对多通道特征图提取特征，实现特征融合。

每一组最小单元之间都有跳跃连接，每一组最小单元的输入都是前面所有层输出的并集。而该最小单元所学习的特征图也会被直接传给其后面所有最小单元作为输入，实现特征的重复利用，文中设计的特征提取最小单元如图 4 所示。

文中提出的轻量化最小结构可以灵活应用在实际工程应用中，可以根据使用场景的需求达到 mAp 和 FPS 的平衡。大量使用轻量化最小结构与提高深度可分离卷积核的数量可以增加平均精准度，但同时会增加模型参数数量，降低检测速度。

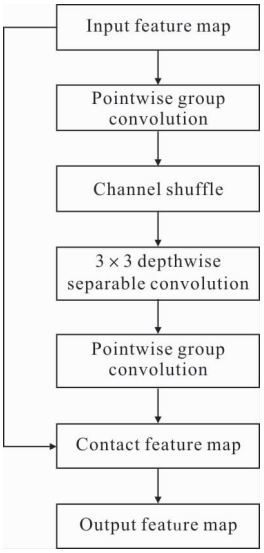

图 4 轻量化特征提取结构最小单元

Fig.4 Lightweight feature extraction minimum unit

在特征重复利用的前提下，最小单元中的深度可分离卷积可设计较窄，即只学习较少的特征图达到降低冗余性的目的。轻量化特征提取最小单元所学习的特征图数量应与网络深度成反比。

同时在各卷积层后加入 Batch Normalization 批量正则化，以防止梯度消失和梯度爆炸的出现。Batch Normalization 正则化首先对每一卷积层输出的特征向量根据每一维度的均值和方差进行归一化,然后对经过修正的输出特征向量做线性变换,使输出特征向量在满足高斯分布的同时保留原有的输出信息。满足高斯分布的特征向量保证了梯度反向传播始终不会出现过大或过小的情况，避免了梯度消失和梯度爆炸的出现。

3 实验结果与分析

在 PASCAL VOC 和 KITTI 两种数据集上进行对比实验。在 PASCAL VOC 数据集中各算法训练集为合并 PASCAL VOC 2007 训练集部分和 PASCAL VOC 2012 训练集部分，测试集为 PASCAL VOC 2007 测试集部分。KITTI 自动驾驶数据集中训练集和测试集为 2012 2D Object Detection left color images of object data，将 KITTI 数据集中 8 类目标合并为 3 类目标，分别为机动车,非机动车和行人。

实验软件环境操作系统为 Ubuntu 14.04，深度学习软件框架为 Caffe。实验硬件环境 CPU 为 Intel (R) Core(R) i7 6700 3.4 GHz×8,内存为 8 G。GPU 为 NVIDIA(R) GTX(R) 1080TI。嵌入式 ARM 设备为 1.2 GHz Cortex A53,1G RAM。

3.1 算法平均精准度结构分析

在前端特征提取模型均采用 VGG-16 的情况下对比不同结构。利用 PASCAL VOC 数据集进行训练和测试,实验结果如表 1 所示。

表 1 PASCAL VOC 数据集下各算法 mAp 对比

Tab.1 mAp comparison of each algorithm in PASCAL VOC data

Object detection algorithm	①	②	③	④	⑤	⑥
mAp	66.4%	73.2%	77.2%	77.4%	77.7%	79.6%
FPS	20	6	28	24	23	21

其中①为 YOLO[4],②为 Faster R-CNN[3],③为 SSD300 原始结构[5]。④为在 SSD300 的基础上将 19×19 特征图做转置卷积后与 38×38 特征图融合,特征融合方式为特征图相加。⑤与④结构相同,特征融合方式为特征图拼接后做卷积。⑥为 SSD300 的基础上将 19×19 特征图转置卷积后与 38×38 特征图融合，将融合后的特征图再转置卷积与 75×75 特征图融合,特征融合方式为特征图拼接后卷积。

在 KITTI 数据集上对经过文中改进的 SSD512 网络结构与原 SSD512 网络结构进行训练和测试，实验结果如表 2 所示。

表 2 KITTI 数据集下各算法性能对比

Tab.2 Properties comparison of each algorithm in KITTI data set

Object detection algorithm	mAp	FPS
①	68.1%	10
②	62.8%	16

①为采用文中提出改进方案的 SSD512 网络结构,②为原 SSD512 网络结构[5]。

经转置卷积改进的 SSD 网络结构在 PASCAL VOC 与 KITTI 数据集上检测小目标效果如图 5 所示,其中图 5(a)为 PASCAL VOC 数据集检测效果,图 5(b)为 KITTI 数据集上检测效果。

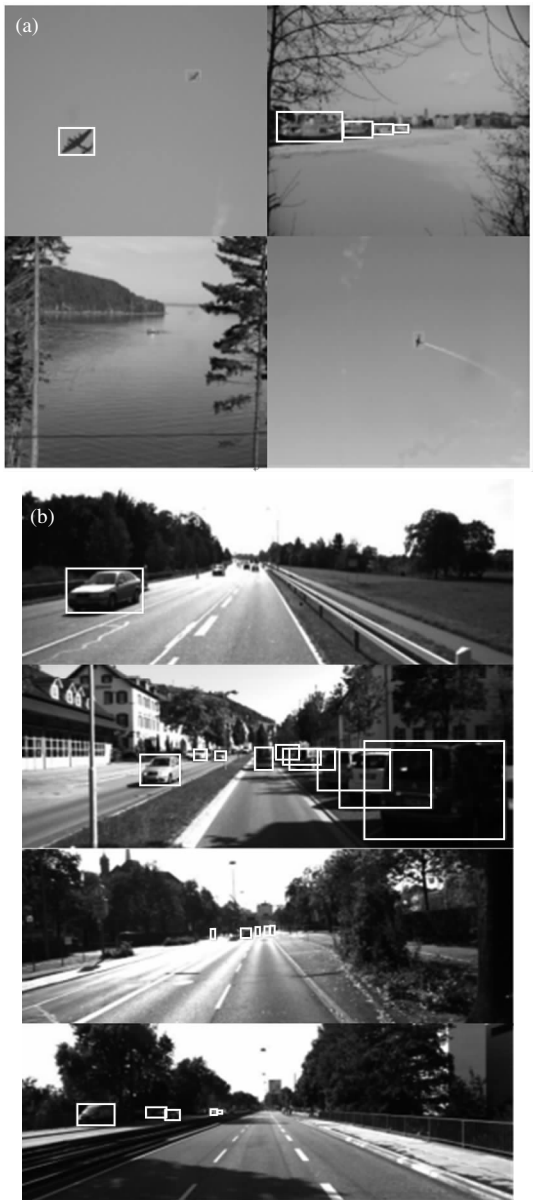

图 5 PASCAL VOC 数据集与 KITTI 数据集检测结果

Fig.5 Test result of PASCAL VOC data set and KITTI data set

通过在 PASCAL VOC数据集和 KITTI 数据集上与各算法做对比,经文中改进的 SSD 网络结构在平均精准度上有明显提升。

3.2 模型压缩与加速结果分析

在 ARM 硬件平台上采用 PASCAL VOC 数据集对比 Tiny YOLO 与使用 17 个轻量化结构并引入转置卷积的改进 SSD300。在单核单线程的情况下,对平均精准度,每帧检测用时与模型大小三项指标进行对比, 实验结果如表 3 所示, 其中①为改进 SSD300 网络结构,②为Tiny YOLO[4]。

表 3 PASCAL VOC 数据集下各算法对比

Tab.3 Properties comparison of each algorithm in PASCAL VOC data set

Object detection algorithm	①	②
mAp	67.2%	57.1%
Use time/s	11	39
Model size/MB	22	60

通过对比实验可知,当使用 17 个轻量化最小单元作为 SSD 特征提取模型时, 模型大小与在 ARM 上运行的速度较 Tiny YOLO 在模型有明显改善。同时由于轻量化特征提取最小单元参数数量少,可以大量使用以提高网络结构的特征提取能力, 所以使用轻量化特征提取最小单元的 SSD 在平均精准度上明显优于 Tiny YOLO。

由于在对比实验中采用单核单线程, 所以在计算性能有限的 ARM 平台上检测速度较慢。在实际使用中,由于特征提取轻量化最小单元模型参数少,所以在运行时占用内存也较少。因此可以根据实际 ARM 的多核数量,利用多核多线程实现帧与帧之间的并行处理,即每一核心处理一帧图像,同时检测多帧图像,以提升检测速度。但多帧并行计算会占用多倍的内存,而 Tiny YOLO 模型参数较多,在内存资源有限的 ARM 平台上无法开启多线程多帧并行计算,只能采用单帧计算,无法进一步提升检测速度。

4 结 论

文中利用转置卷积将高分辨率低语义信息特征图与低分辨率高语义信息特征图相融合,提高 SSD 算法的小目标检测能力。同时以 DenseNet 为基础,结合深度可分离卷积与通道重排提出轻量化特征提取最小单元。在 PASCAL VOC 和 KITTI 数据集上对比 Faster R-CNN,YOLO 和原 SSD 算法等,经转置卷积改进的 SSD 算法达到提升平均精准度的目的。将 Tiny YOLO 与经轻量化改进的 SSD 作对比, 改进后的 SSD 达到降低参数数量,减少内存占用和提升检测速度的目的,能够在 ARM 嵌入式设备上运行。

参考文献:

[1] Zhang Difei,Zhang Jinsuo,Yao Keming, et al. Infrared ship-

target recognition based on SVM classification [J]. ***Infrared and Laser Engineering***, 2016, 45(1): 0104004. (in Chinese)
张迪飞, 张金锁, 姚克明, 等. 基于 SVM 分类的红外舰船目标识别[J]. 红外与激光工程, 2016, 45(1): 0104004.

[2] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)
罗海波, 许凌云, 惠斌, 等. 基于深度学习的目标跟踪方法研究现状与展望[J]. 红外与激光工程, 2017, 46(5): 0502002.

[3] Ren S, He K, Girshick R B, 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.

[4] Redmon J, Divvala S K, Girshick R B, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multiBox detector [J]. ***European Conference on Computer Vision***, 2016: 21-37.

[6] Xiong Changzhen, Shan Yanmei, Guo Fenhong. Image retrieval method based on image principal part detection[J]. ***Optics and Precision Engineering***, 2017, 25(3): 792-798. (in Chinese)
熊昌镇, 单艳梅, 郭芬红. 结合主体检测的图像检索方法[J]. 光学 精密工程, 2017, 25(3): 792-798.

[7] Wu Xin, Zhang Jianqi, Yang Chen. Efficient infrared image background prediction with Jetson TK1 [J]. ***Infrared and Laser Engineering***, 2015, 44(9): 2615-2621. (in Chinese)
吴鑫, 张建奇, 杨琛. Jetson TK1 平台实现快速红外图像背景预测算法[J]. 红外与激光工程, 2015, 44(9): 2615-2621.

[8] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [J/OL]. 2016: arXiv: 1612.03144[cs.CV].

[9] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[J/OL]. 2017: arXiv: 1707.01083[cs.CV].

[10] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J/OL]. 2017, arXiv:1704.04861[cs.CV].

[11] Liu Zhi, Huang Jiangtao, Feng Xin. Action recognition model construction based on multi-scale deep convolution neural network[J]. ***Optics and Precision Engineering***, 2017, 25(3): 799-805. (in Chinese)
刘智, 黄江涛, 冯欣. 构建多尺度深度卷积神经网络行为识别模型[J]. 光学 精密工程, 2017, 25(3): 799-805.

[12] Huang G, Liu Z, Weinberger K Q, et al. Densely connected convolutional networks [J/OL]. 2016, arXiv: 1608.06993[cs.CV].