计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):153−158 [doi: 10.15888/j.cnki.csa.006941]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于卷积神经网络的复杂场景目标检测算法①

王晓宁, 宫法明, 时念云, 吕轩轩

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)
通讯作者: 宫法明, E-mail: gfming@163.com

摘　要: 海上石油平台监控环境复杂, 采油工作平台摄像头监控角度不同, 海上环境复杂多变, 雨雾等天气下, 摄像头图片模糊不清. 针对上述增加了目标检测的难度的问题, 提出了一种基于卷积神经网络的复杂场景目标检测算法 (简称 ODCS) 来检测图像中的特定对象. 该方法结合不同分辨率的特征图预测来自然处理各种尺寸的对象, 消除了特征重新采样阶段, 并将所有计算封装在单个网络中, 这样易于训练且可以直接集成到需要检测组件的系统中. 实验结果表明, 相对于传统的方法, 该方法检测在准确率和召回率上明显提高, 且检测效率能够满足实时应用的要求.

关键词: 计算机视觉; 复杂场景; 目标检测; 深度学习; 卷积神经网络

引用格式: 王晓宁,宫法明,时念云,吕轩轩.基于卷积神经网络的复杂场景目标检测算法.计算机系统应用,2019,28(6):153–158. http://www.c-s-a.org.cn/1003-3254/6941.html

Object Detection Algorithm of Complex Scenario Based on Convolution Neural Network

WANG Xiao-Ning, GONG Fa-Ming, SHI Nian-Yun, LYU Xuan-Xuan

(College of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: The monitoring environment of offshore oil platforms is complex, the monitoring angle of the oil production working platform is different, the marine environment is complex and changeable, and the camera pictures are blurred in the weather such as fog and rain. To solve the above problem of increasing the difficulty of object detection, the object detection algorithm based on Convolutional Neural Network (CNN) in complicated scenario (ODCS) is proposed to detect specific objects in the image. This method integrates feature map prediction with different resolutions to naturally process objects of various sizes, eliminates the feature re-sampling phase, and encapsulates all calculations in a single network. This is easy to train and can be integrated directly into the system that needs to detect components. The experimental results show that compared with the traditional methods, the detection accuracy of this method and the recall rate are significantly improved, and the detection efficiency can meet the requirements of real-time applications.

Key words: computer vision; complicated scenes; object detection; deep learning; Convolutional Neural Network (CNN)

1　概述

海上石油平台环境复杂, 实现海上石油平台的目标检测和定位, 其难点主要体现在如下几个方面: 现有运行的采油工作平台类型非常多, 摄像头的安装位置各不相同, 加之海上环境复杂多变, 使得视频背景及海上石油平台工作人员在视频中出现的位置复杂多变; 海上情况复杂, 受天气影响严重, 雾天、雨天等天气下, 摄像头图片模糊不清, 增加了目标检测的难度; 海上石油平台中采油设备密集, 摄像头可选择的安装位置有限, 这一限制条件决定了视频画面中的待检测目标遮

① 收稿时间: 2018-12-07; 修改时间: 2018-12-25; 采用时间: 2019-01-15

挡程度各异、截断情形较多; 摄像头安装位置不定, 导致所得到的图片的拍摄角度各异, 部分摄像头甚至从头部向下俯拍, 此类目标难以良好检测.

随着计算机视觉技术的不断发展与进步, 深度学习作为计算机视觉领域中越来越火热和成熟的部分, 人们希望深度学习能够代替人类完成更多的工作, 甚至能在某些领域有更大的突破. 目标检测是计算机视觉中的基本挑战难题之一. 近几年, 目标检测等领域取得了重大进展, 这主要得益于深度学习. R-CNN[1], SPP-NET[2], Fast R-CNN[3]和 Faster R-CNN[4]等结合了候选区域和卷积神经网络 (CNN) 实现目标检测. YOLO[5]和 SSD[6]等将目标检测转换为回归问题的目标检测框架.

目前基于深度学习的目标检测研究很多, 应用在目标检测中效果十分突出, 现有的方法检测精确, 但计算量相对较大, 检测效率达不到实时应用的要求. 这些方法的检测速度通常以秒/帧为单位, Faster R-CNN 只能以每秒 7 帧的速度工作. 目前为止, 速度的显著提高会大大降低检测精度. 复杂场景目标检测算法不仅在检测精度上取得了一定提高, 同时检测效率方面也能够满足实时应用的要求.

复杂场景目标检测方法 (简称 ODCS) 可以有效地实现海上石油平台的目标检测和定位, 从输入的监控视频中, 确定检测目标在图像中的位置和大小, 将基于深度学习的目标检测应用于视频监控实际应用中. ODCS 消除了特征重采样, 从而避免了破坏原来的像元值, 最终使计算量大大减少, 提高了算法的检测效率.

本文的组织结构如下: 在第 2 节中, 介绍现阶段目标检测的相关研究; 在第 3 节中, 介绍本文提出的方法思想以及本文模型的训练过程; 第 4 节中, 针对石油平台的监控视频数据集进行实验验证; 第 5 节中, 对本文内容进行总结.

2 相关工作

在卷积神经网络出现之前, 目标检测方法最先进的领域——可变形部分模型 (DPM)[7]和选择性搜索[8]——具有类似的性能. 然而, 在 R-CNN 带来戏剧性的改善后, 结合选择性搜索候选区域和基于卷积网络分类, 区域提案的目标检测方法变得流行.

R-CNN 在各种方式上都得到了改进, 提高了后分类的质量和速度. 但是因为它需要对数千幅图像进行分类, 既昂贵又耗时. SPP-NET 引入了空间金字塔池层, 对区域大小和规模更有鲁棒性, 并且允许分类层重用在几种图像分辨率生成的特征映射上计算出来的特征, 大大加快了 R-CNN. Fast R-CNN 延伸 SPP-NET 使它可以很好的通过最小化为信心和包围盒回归损失调整所有层的端到端的, 这是第一次在 MultiBox[9,10]学习对象.

利用卷积神经网络可以提高提案生成的质量, 比如 MultiBox 基于低层图像特征的选择性搜索区域方案被直接由一个单独的卷积神经网络生成的正函数所取代. 这进一步提高了检测的准确性, 但需要训练两个神经网络之间的依赖关系. Faster R-CNN 取代了从区域建议网络 (RPN) 中学习的选择性搜索方案, 并提出了一种将 RPN 与 Fast R-CNN 相结合的方法, 在这两种网络之间交替调整共享卷积层和预测层. 这样, 区域提案被用于集合中层特征, 最后的分类步骤更便宜. SSD 非常类似于 R-CNN 中的区域提案网络 (RPN), 因为 SSD 也使用固定的 (默认) 框来预测, 类似 RPN 中的锚框. 但是, SSD 可以同时为每个框中的每个对象类别生成一个分数, 而不是使用这些到池特征并对另一个分类器进行评估. 因此, SSD 避免了 RPN 与 Fast R-CNN 合并的复杂性, 并且更容易训练, 更快、更直接地集成到其他任务中.

OverFeat[11]完全跳过建议步骤, 直接预测多个类别的边界框和置信度, 在了解基本对象类的置信度后直接从最高的特征图预测一个包围盒. YOLO 使用整个最顶层特征图来预测多个类别的置信度和边界框. SSD 没有建议步骤, 直接使用默认框, 比之前的方法更灵活, 因为可以在每个特征位置上使用不同的纵横比的默认框, 在不同的尺度上使用多个特征映射. 如果在最上面的功能图中每个位置只使用一个默认框, SSD 就会拥有类似于 OverFeat 的架构. 如果使用整个上面的特征映射和添加一个完全连接层, 而不是 SSD 的卷积因子的预测, 并没有明确地考虑多方面的比例, 可以近似地再现 YOLO.

为了处理不同的对象尺度, 一些方法[12]建议首先处理不同尺寸的图像, 然后将结果合并. 然而, 通过利用单个网络中几个不同层的特征映射进行预测, 我们可以模拟相同的效果, 同时还可以跨所有对象尺度共享参数. 已有工作[13,14]已经表明, 使用低层的特征图可

以提高语义分割的质量, 因为低层的特征图能够捕捉输入对象的更多细节. 同样, 文献[15]表明, 从特征映射汇集的全局上下文可以帮助平滑分割结果. 受这些方法的启发, 我们使用较低和较高的特征图进行检测.

3 复杂场景目标检测算法

3.1 算法概述

复杂场景目标检测算法 (简称 ODCS) 的完整流程如图 1 所示, 步骤总结如下.

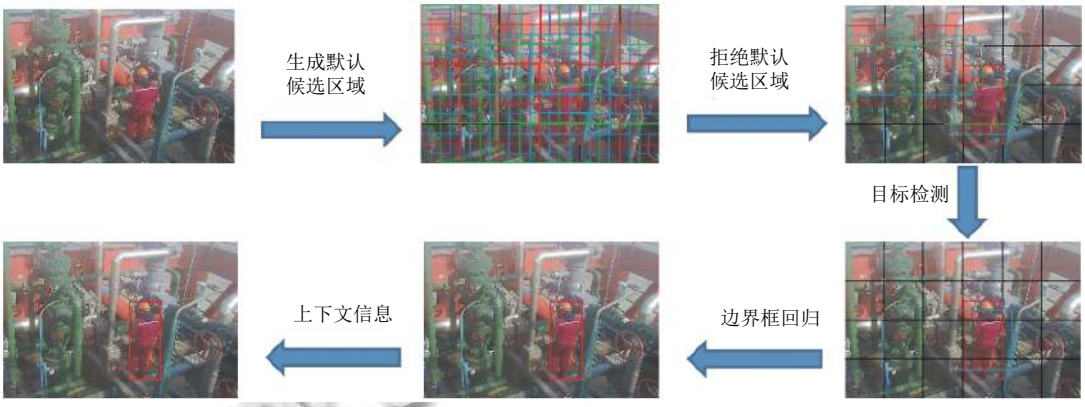

图 1 算法流程图

Step 1: 对原始图像进行图像预处理, 并在多级特征图上生成默认候选区域.

Step 2: 拒绝最有可能是背景的默认候选区域.

Step 3: 多级特征图输入到目标检测网络中, 获得检测分数, 每个检测分数包含一个特定对象类的检测置信度.

Step 4: 使用边界框回归进一步提炼边界框, 减少定位误差.

Step 5: 利用深度模型的全图像分类评分作为上下文信息来细化每个对象框的检测.

ODCS 结合不同分辨率的多个特征图的预测来自然处理各种尺寸的对象, 每个特征图上的每一个小格子为一个特征图单元. 本算法中, 首先定义默认候选区域为在特征图的每个小格上的一系列固定大小的包围盒. 假设每个特征图单元有 k 个默认候选区域, 那么对于每个默认候选区域都需要预测 c 个类别得分和 4 个偏移坐标, 那么如果一个特征图的大小为 $m \times n$. 经过候选区域拒绝后, 保留下 $q(q<m \times n \times k)$ 个候选区域, 那么这个特征图一共有 $(c+4) \times q$ 个输出. 然后, 利用全图像分类评分作为上下文信息细化对每个对象框的检测, 以提高检测精确度. 最后, 使用边界框回归精确定位, 以减少定位误差.

ODCS 实现了海上石油平台的目标实时检测, 从输入的监控视频中, 确定检测目标在图像中的位置和类别, 将基于深度学习的目标检测应用于视频监控实际应用中. ODCS 消除了特征重采样, 从而避免了破坏原来的像元值, 最终大大减少了计算量, 提高了目标检测效率. ODCS 将计算封装在单个网络中, 这样易于训练且可以直接集成到需要检测组件的系统中.

3.2 匹配策略

在训练期间, 需要建立真实标签框和候选区域之间的对应关系. 每个真实标签框都是从候选区域中选择, 这些候选区域根据位置和长宽比而变化. 首先, 我们将每个真实标签框与具有 Jaccard 相似系数的候选区域匹配. 将任何具有 Jaccard 相似系数高于阈值的候选区域与真实标签相匹配, 这简化了学习问题, 多个重叠候选区域时, 网络预测获得高置信度, 而不是要求它仅挑选具有最大重叠的一个, 所以一个真实标签可能对应多个候选区域.

在预测阶段, 直接预测每个候选区域的偏移以及对每个类别相应的得分, 最后通过 NMS 得到最终的结果. 对于每个候选区域, 同时预测它的偏移坐标和所有类的置信度. 因此, 对于每个特征图单元而言, 一共有 4 种候选区域. 可以看出这种候选区域在不同的特征层有不同的大小, 在同一个特征层又有不同的纵横比, 因此基本上可以覆盖输入图像中的各种形状和大小的对象.

第 i 个候选区域与第 p 个第 j 个真实标签框匹配的指示器:

$$x_{ij}^{p}=\{1,0\} \tag{1}$$

在上面的匹配策略中, 有$\sum_i x_{ij}^{p}\geq 1$.

整体目标损失函数是本地化损失:

$$L=\frac{1}{N}(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)) \tag{2}$$

其中, N 是匹配的默认框的数量. 如果 $N=0$, 则将损失设置为 0. 定位损失是预测框 (l) 和真实标签框 (g) 参数之间的平滑损失. 置信度损失是多级置信度 (c) 的最大值损失.

预训练和微调阶段的深层结构仅在用于预测标签的最后完全连接层中不同. 除了最后完全连接的分层以外, 在训练阶段学习的参数直接作为微调阶段的初始值.

3.3 默认候选区域

已知网络中不同层次的特征图具有不同大小的接受区域 . 在我们的框架内, 候选区域不需要对应于每个层的实际接受字段. 我们设计默认候选区域的平铺, 以便特定的特征图学习响应物体的特定尺度. 假设我们想使用 m 个特征映射进行预测, 每个特征映射的候选区域的尺度计算如下:

$$s_k=s_{\min}+\frac{s_{\max}-s_{\min}}{m-1}(k-1),k\in[1,m] \tag{3}$$

其中, $s_{\min}$ 为 0.2, $s_{\max}$ 为 0.9, 中间的所有层均匀分布.

通过将许多特征图的所有位置的预测误差都包含在不同的误差范围之内, 我们有了各种各样的预测, 涵盖了输入对象的各种大小和形状.

3.4 边界框回归

对于窗口一般使用四维向量 (x, y, w, h) 来表示, 分别表示窗口的中心点坐标和宽高. 对于图 2, 红色的框 P 代表原始的窗口, 绿色的框 O 代表目标的真实窗口, 我们的目标是寻找一种关系使得输入原始的窗口 P 经过映射得到一个跟真实窗口 G 更接近的回归窗口 Q.

3.5 上下文建模

对于图像分类任务学习的深度模型考虑到场景信息, 而深度模型的对象检测集中在局部边界框. 将 1000 级图像分类分数作为上下文特征, 与 200 级对象检测分数连接形成 1200 维特征向量, 在此基础上学习线性支持向量机对 200 个分类检测分值进行优化.

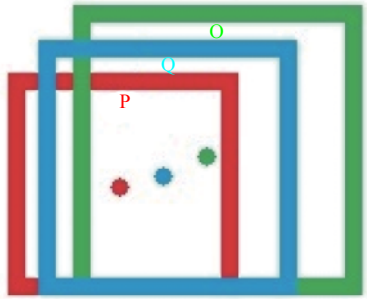

图 2 边界框回归

4 实验结果与分析

使用图像分类任务, 使用海上石油平台 281 个摄像头的监控视频数据的图像进行注释, 预训练深度模型. 本次选取的摄像头安装的位置、角度各不形同, 视频涵盖全天 24 小时所有时间段, 数据总量达到 28891 组. 微调对象检测任务的深度模型, 使用来自对象检测训练数据的对象级别注释.

本实验的原始数据来自海洋采油厂的流媒体服务器, 各个海洋平台上的监控设备保持固定不动, 以海洋工作平台作为监控场景, 并通过微波的方式将实时的监控视频传输并存储到流媒体服务器中. 在原始视频库数据集上, 使用关键帧图像提取法选取带有目标的图像数据集, 即在 1 秒的间隔内将首帧、中间帧和尾帧视为关键帧图像, 然后通过人工标注图像形成目标检测所使用的标签数据库. 该数据库存储了目标的标签类型和位置信息, 包含了 2 万张目标图像, 由 204 路摄像头采集各个场景的图像组成.

本文使用同一组训练数据分别对 Faster R-CNN、SSD 和本文提出的复杂场景目标检测方法进行训练, 并使用同一组测试数据在相同配置的计算机上对两种方法进行检测. 实验所用的训练数据和测试数据均取自石油平台监控视频. 数据涵盖了石油平台 281 个摄像头, 图 3 为复杂场景目标检测算法所得到的实验结果, 分别展示了不同场景下人员、车辆和船只的检测结果. 其中, 红色框代表检测结果为人员, 绿色框代表检测结果为车辆, 蓝色框代表检测结果为船只.

设计实验将 ODCS 与 Faster R-CNN、SSD 的训练迭代次数与错误率进行对比实验. 采用随机梯度下降的方法输入样本进行训练, 学习率为 0.01. 图 4 为实验结果对比. 由图 4 可以看出, 与其他两种方法相比, ODCS 的识别错误率更低, 当训练迭代次数达到 30 次时, ODCS 基本达到收敛状态, 其识别错误率已低于 1%.

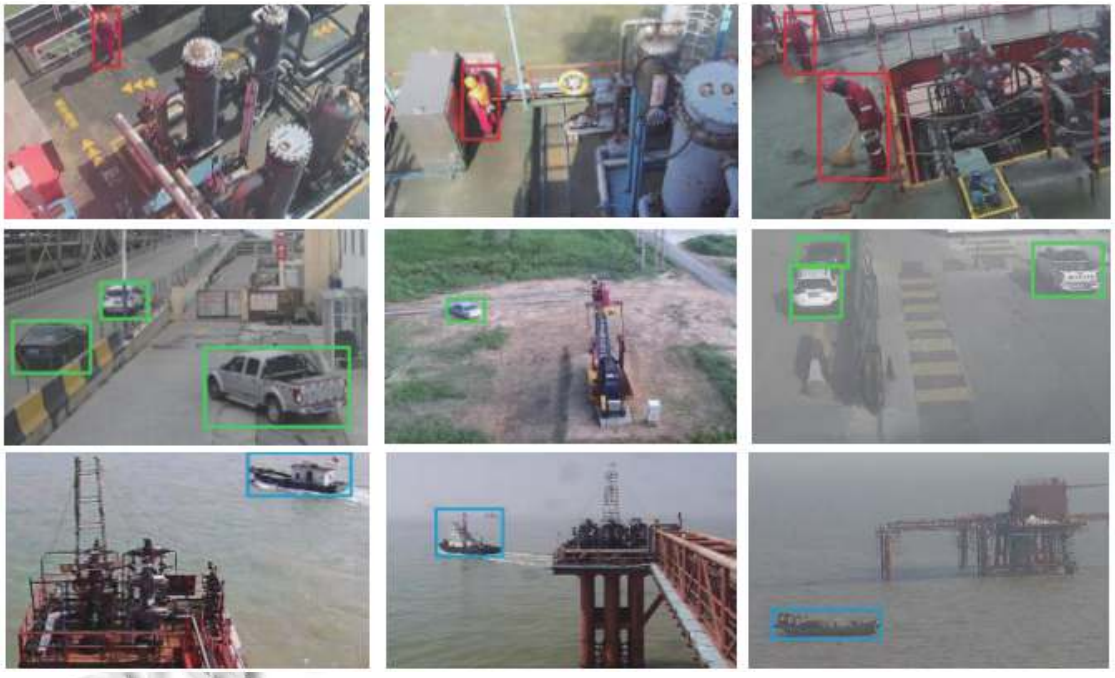

图 3 ODCS 实验结果图

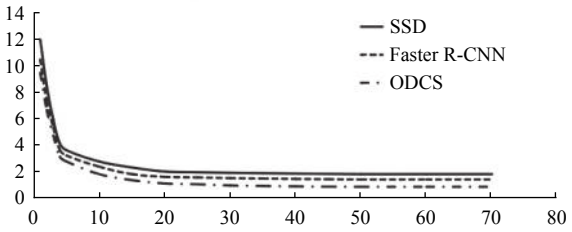

图 4 训练迭代次数与错误率

为了客观地验证复杂场景目标检测算法的检测效果如何, 针对检测结果现在做如下的定义: TP 定义为人 (车或船) 的图片被正确识别成人 (车或船); TN 定义为未包含目标的图片没有被识别出来, 系统正确地认为无目标; FP 无目标的图片被错误地识别为人 (车或船); FN 人 (车或船) 的图片没有被识别出来, 系统错误地认为无目标. 采用 3 个参数指标对本文提出的算法进行比较分析, 分别为精确率 (Precision)、召回率 (Recall) 和误检率 (PBC). 依据上述 3 个参数指标得到的对比实验结果, 表 1 为实验结果分析表, 图 5 所示为三种方法的准确率、召回率和误检率对比结果图.

表 1 目标检测实验结果对比

检测方法	检测目标	TP	TN	FP	FN	精确率 (%)	召回率 (%)	误检率 (%)
Faster R-CNN	人员	757	767	233	243	76.46	75.70	23.80
	车辆	799	877	123	201	86.66	79.90	16.20
	船只	782	876	124	218	86.31	78.20	17.10
SSD	人员	763	774	226	237	77.15	76.30	23.15
	车辆	802	865	135	198	85.59	80.20	16.65
	船只	781	873	127	219	86.01	78.10	17.30
ODCS	人员	798	796	204	202	79.64	79.80	20.30
	车辆	810	885	115	190	87.57	81.00	15.25
	船只	813	894	106	187	88.47	81.30	14.65

由图 5 实验结果可以直观的看出, 复杂场景目标检测算法在整体性能上要优于其他 2 种算法: 精确率平均提高了 2% 左右, 召回率提高了大约 3%, 误检率也降低了 2% 左右.

由上述实验结果得知, 三种方法对车辆船只的检测效果优于对人员的检测, 分析原因如下: 首先, 车辆、船只等本身属于刚性物体, 不存在形变, 人员属于非刚性目标, 存在形变等现象; 其次, 人员检测中的检测

环境存在很大的误导性, 周围环境复杂, 人员遮挡严重.

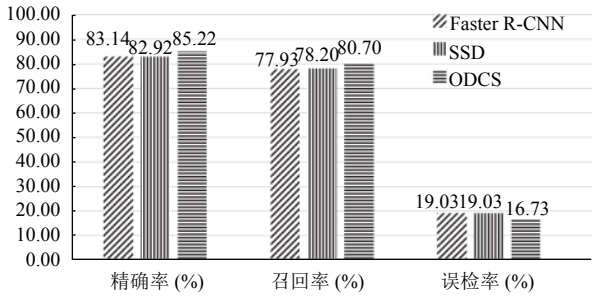

图 5 实验结果对比表

在目标检测中, 光线对于目标检测的结果影响很大, 光照强度的变化可能会导致结果的误判, 如光的反射或者光照射角度所产生的阴影的不同; 同样, 天气原因如雨雾天等会使视频中目标模糊, 使得提取的特征有细微的变化从而影响检测结果. 本文提出的复杂场景目标检测算法具有很好的鲁棒性, 可以在不同的复杂场景下进行目标检测, 检测结果良好.

5 结束语

如何从复杂场景中实现目标的识别与检测则成为更加重要和困难的问题. 针对该问题, 我们提出了一种基于卷积神经网络来检测图像中的对象的方法. 我们的方法将边界框的输出空间离散化为根据不同长宽比和每个特征映射位置缩放的一组默认框. 在预测时, 网络会在每个默认框中为每个对象类别的出现生成分数, 并对框进行调整以更好地匹配对象形状. 另外, 网络结合不同分辨率的多个特征图的预测来自然处理各种尺寸的对象. 相对于需要对象提议的方法, 我们的方法非常简单, 因为它完全消除了提案生成和随后的像素或特征重新采样阶段, 并将所有计算封装在单个网络中. 这使得我们的方法易于训练和直接集成到需要检测组件的系统中. 与其他单级方法相比, 即使输入图像尺寸较小, 我们的方法也具有更高的精度.

参考文献

1 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.

2 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 346–361.

3 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448.

4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada. 2015. 91–99.

5 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.

6 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 21–37.

7 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. 2008. 1–8.

8 Uijlings JRR, van de Sande KEA, Gevers T, *et al*. Selective search for object recognition. International Journal of ComputerVision,2013,104(2):154–171.[doi:10.1007/s11263-013-0620-5]

9 Erhan D, Szegedy C, Toshev A, *et al*. Scalable object detection using deep neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 2155–2162.

10 Szegedy C, Reed S, Erhan D, *et al*. Scalable, high-quality object detection. arXiv: 1412.1441, 2014.

11 Sermanet P, Eigen D, Zhang X, *et al*. OverFeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229, 2013.

12 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

13 Hariharan B, Arbeláez P, Girshick R, *et al*. Hypercolumns for object segmentation and fine-grained localization. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 447–456.

14 Liu W, Rabinovich A, Berg AC. ParseNet: Looking wider to see better. arXiv:1506.04579, 2016.

15 Zhou BL, Khosla A, Lapedriza A, *et al*. Object detectors emerge in deep scene CNNs. arXiv:1412.6856, 2014.