深度学习在目标视觉检测中的应用进展与展望

张 慧 [1, 2] 王坤峰 [1, 3] 王飞跃 [1, 4]

摘 要 目标视觉检测是计算机视觉领域的一个重要问题, 在视频监控、自主驾驶、人机交互等方面具有重要的研究意义和应用价值. 近年来, 深度学习在图像分类研究中取得了突破性进展, 也带动着目标视觉检测取得突飞猛进的发展. 本文综述了深度学习在目标视觉检测中的应用进展与展望. 首先对目标视觉检测的基本流程进行总结, 并介绍了目标视觉检测研究常用的公共数据集; 然后重点介绍了目前发展迅猛的深度学习方法在目标视觉检测中的最新应用进展; 最后讨论了深度学习方法应用于目标视觉检测时存在的困难和挑战, 并对今后的发展趋势进行展望.

关键词 目标视觉检测, 深度学习, 计算机视觉, 平行视觉

引用格式 张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, **43**(8): 1289−1305

DOI 10.16383/j.aas.2017.c160822

Advances and Perspectives on Applications of Deep Learning in Visual Object Detection

ZHANG Hui[1, 2] WANG Kun-Feng[1, 3] WANG Fei-Yue[1, 4]

Abstract Visual object detection is an important topic in computer vision, and has great theoretical and practical merits in applications such as visual surveillance, autonomous driving, and human-machine interaction. In recent years, significant breakthroughs of deep learning methods in image recognition research have arisen much attention of researchers and accordingly led to the rapid development of visual object detection. In this paper, we review the current advances and perspectives on the applications of deep learning in visual object detection. Firstly, we present the basic procedure for visual object detection and introduce some newly emerging and commonly used data sets. Then we detail the applications of deep learning techniques in visual object detection. Finally, we make in-depth discussions about the difficulties and challenges brought by deep learning as applied to visual object detection, and propose some perspectives on future trends.

Key words Visual object detection, deep learning, computer vision, parallel vision

Citation Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. *Acta Automatica Sinica*, 2017, **43**(8): 1289−1305

目标视觉检测是计算机视觉领域中一个非常重要的研究问题. 随着电子设备的应用在社会生产和人们生活中越来越普遍, 数字图像已经成为不可缺少的信息媒介, 每时每刻都在产生海量的图像数据. 与此同时, 对图像中的目标进行精确识别变得越来越重要[1]. 我们不仅关注对图像的简单分类, 而且希望能够准确获得图像中存在的感兴趣目标及其位置[2], 并将这些信息应用到视频监控、自主驾驶等一系列现实任务中, 因此目标视觉检测技术受到了广泛关注[3].

目标视觉检测具有巨大的实用价值和应用前景. 应用领域包括智能视频监控、机器人导航、数码相机中自动定位和聚焦人脸的技术、飞机航拍或卫星图像中道路的检测、车载摄像机图像中的障碍物检测等. 同时, 目标视觉检测也是众多高层视觉处理和分析任务的重要前提, 例如行为分析、事件检测、场景语义理解等都要求利用图像处理和模式识别技术, 检测出图像中存在的目标, 确定这些目标对象的语义类型, 并且标出目标对象在图像中的具体区域[4].

在自然环境条件下, 目标视觉检测经常遇到以下几个方面的挑战:

收稿日期 2016-12-15 录用日期 2017-03-16
Manuscript received December 15, 2016; accepted March 16, 2017
国家自然科学基金 (61533019, 61304200), 国家留学基金 (201504910397) 资助
Supported by National Natural Science Foundation of China (61533019, 61304200) and China Scholarship Council (201504910397)
本文责任编委 周涛
Recommended by Associate Editor ZHOU Tao
1. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190 2. 中国科学院大学 北京 100049 3. 青岛智能产业技术研究院 青岛 266000 4. 国防科学技术大学军事计算实验与平行系统技术研究中心 长沙 410073
1. State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. University of Chinese Academy of Sciences, Beijing 100049 3. Qingdao Academy of Intelligent Industries, Qingdao 266000 4. Research Center for Computational Experiments and Parallel Systems Technology, National University of Defense Technology, Changsha 410073

1) 类内和类间差异

对于很多物体, 它们自身就存在很大的差异性, 同类物体的不同实例在颜色、材料、形状等方面可能存在巨大的差异, 很难训练一个能够包含所有类内变化的特征描述模型. 另外, 不同类型物体之间又可能具有很大的相似性, 甚至非专业人员从外观上很难区分它们. 类内差异可能很大, 而类间差异可能很小, 给目标视觉检测提出了挑战.

2) 图像采集条件

在图像采集过程中, 由于环境、光照、天气、拍摄视角和距离的不同、物体自身的非刚体形变以及可能被其他物体部分遮挡, 导致物体在图像中的表观特征具有很大的多样性, 对视觉算法的鲁棒性提出了很高要求.

3) 语义理解的差异

对同一幅图像, 不同的人可能会有不同的理解, 这不仅与个人的观察视角和关注点有关, 也与个人的性格、心理状态和知识背景等有关, 这明显增加了从仿生或类脑角度来研究视觉算法的难度.

4) 计算复杂性和自适应性

目标视觉检测的计算复杂性主要来自于待检测目标类型的数量、特征描述子的维度和大规模标记数据集的获取. 由于在真实世界中存在大量的目标类型, 每种类型都包含大量的图像, 同时识别每种类型需要很多视觉特征, 这导致高维空间稀疏的特征描述[4]. 另外, 目标模型经常从大规模标记数据集中学习得到, 在许多情况下, 数据采集和标注很困难, 需要耗费大量的人力物力. 这些情况导致目标检测的计算复杂性很高, 需要设计高效的目标检测算法. 同时, 在动态变化的环境中, 为了提高目标检测精度, 还需要探索合适的机制来自动更新视觉模型, 提高模型对复杂环境的自适应能力.

为了克服上述挑战, 已经提出了许多目标视觉检测算法, 它们在目标区域建议、图像特征表示、候选区域分类等步骤采用了不同的处理策略. 近年来, 随着深度学习技术的发展, 很多基于深度学习的目标视觉检测方法陆续被提出, 在精度上显著优于传统方法, 成为最新的研究热点. 本文首先介绍目标视觉检测的基本流程, 然后重点介绍深度学习在目标视觉检测中的应用进展.

本文内容安排如下: 第 1 节介绍目标视觉检测的基本流程; 第 2 节对目标视觉检测研究常用的公共数据集进行概述; 第 3 节介绍深度学习技术在目标视觉检测中的最新应用进展; 第 4 节讨论深度学习技术应用于目标视觉检测时存在的困难和挑战, 并对今后的发展趋势进行展望; 第 5 节对本文进行总结.

1 目标视觉检测的基本流程

目标视觉检测的根本问题是估计特定类型目标出现在图像中的哪些位置. 如图 1 所示, 目标视觉检测技术在流程上大致分为三个步骤: 区域建议 (Region proposal)、特征表示 (Feature representation) 和区域分类 (Region classification). 首先对图像中可能的目标位置提出建议, 也就是提出一些可能含有目标的候选区域. 然后采用合适的特征模型得到特征表示. 最后借助分类器判断各个区域中是否含有特定类型的目标, 并且通过一些后处理操作, 例如非极大值抑制、边框位置回归等, 得到最终的目标边框. 该基本流程被许多工作所采用, 例如文献 [5] 提出的 HOG-SVM 检测方法、文献 [6] 提出的 Selective search 区域建议方法、目前在 PASCAL VOC、MS COCO、ImageNet 等数据集上取得领先精度的 Faster R-CNN[7] 检测方法以及 Faster R-CNN 采用的特征表示和区域分类方法 ResNet[8] 等.

本节接下来从区域建议、特征表示和区域分类

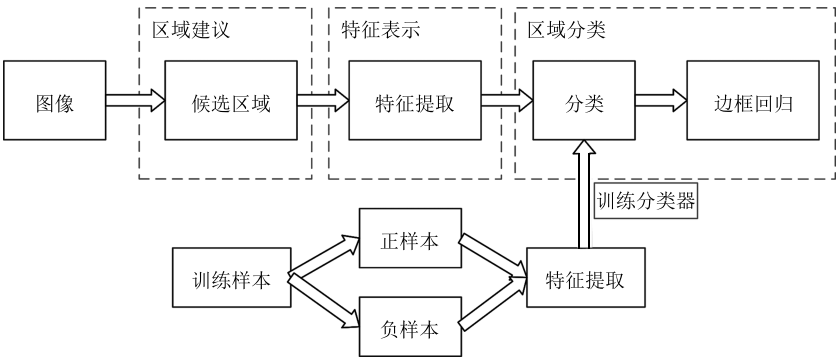

图 1 目标视觉检测的基本流程

Fig. 1 Basic procedure for object detection

三个方面来总结目标视觉检测的关键技术.

1.1 区域建议

目标检测要求获得目标的位置和尺度信息, 这需要借助区域建议来实现. 区域建议是指在输入图像中搜寻特定类型目标的可能区域的一种策略. 传统的区域建议策略包括三种[4]: 基于滑动窗的区域建议、基于投票机制的区域建议和基于图像分割的区域建议.

1.1.1 基于滑动窗的区域建议

基于滑动窗的方法是在输入图像所有可能的子窗口中执行目标检测算法来定位潜在的目标. 在文献 [5] 中, 检测窗口是一个给定大小的矩形框, 在整幅图像的所有位置和尺度上进行扫描, 并对区域分类结果做非极大值抑制. 基于滑动窗的区域建议方法采用穷举搜索, 原理简单, 易于实现, 但是计算复杂性高, 太过耗时. 于是一些研究者提出加快窗口搜索的方法. Lampert 等[9] 提出了一种高效的子窗口搜索策略 (简称为 ESS), 采用分支限界法来减少搜索范围. 但是它的性能在很大程度上取决于输入图像中的物体, 当没有物体出现时, 该算法退化到穷举搜索. An 等[10] 提出一种改进的 ESS 算法. Wei 等[11] 提出一种在直方图维度上具有常数复杂度的滑动窗口策略. Van de Sande 等[12] 引入图像分割信息, 将其作为目标假设区域, 从而只对这些假设区域进行目标检测.

1.1.2 基于投票机制的区域建议

基于投票机制的方法主要用于基于部件的模型, 通常投票机制的实现可归纳为两步[13−14]: 1) 找到输入图像与模型中各个局部区域最匹配的区域, 并最大化所有局部区域的匹配得分; 2) 利用拓扑评价方法取得最佳的结构匹配. 由于投票机制是一种贪心算法, 可能得不到最优的拓扑假设, 并且部件匹配通常采用穷举搜索来实现, 计算代价很高.

1.1.3 基于图像分割的区域建议

基于图像分割的区域建议建立在图像分割的基础上, 分割的图像区域就是目标的位置候选. 语义分割是一种最直接的图像分割方法, 需要对每个像素所属的目标类型进行标注[15]. 目前主要采用的方法是概率图模型, 例如采用 CRF[16] 或 MRF[17] 方法来鼓励相邻像素之间的标记一致性. 图像分割是一个耗时而又复杂的过程, 而且很难将单个目标完整地分割出来.

不同于以上策略, 文献 [6] 先将图片分割成若干小区域, 然后再聚合, 通过对聚合后的区域打分并排序, 获得较有可能是目标区域的窗口. 文献 [18−19] 中采用生成大量窗口并打分, 然后过滤掉低分的方法. 文献 [20] 对这些方法进行了讨论和比较. 这些方法存在的主要问题是, 采样数目较少时召回率不高、定位精度较低等. 对于一个目标检测系统来说, 少量的候选区域不仅可以减少运行时间, 而且使得检测准确率更高, 因此保证采样数目少的情况下召回率仍然很高是至关重要的. 为了解决这些问题, 一些研究者开始采用深度学习方法来产生候选区域. 在 MultiBox[21−22] 中, 通过采用深度神经网络回归模型定位出若干可能的包围边框. 在 Deepbox[23] 中, Kuo 等采用训练卷积神经网络模型来给通过 EdgeBoxes[19] 产生的候选区域进行排序. 在 DeepProposal[24] 中, Ghodrati 等评估了用卷积神经网络产生目标候选区域的质量, 发现最后一层卷积层可以以很高的召回率找到感兴趣的目标, 但是定位精度很低, 而第一层网络可以很好地定位目标, 但是召回率很低. 基于此发现, 他们设计了一种通过多层 CNN 特征由粗到细地串联来产生候选区域的方法. 文献 [7] 提出区域建议网络 (Region proposal network, RPN), 把产生候选区域和区域分类联合到一个深度神经网络, 通过端到端训练, 在提高精度的同时降低了计算时间. 最近, Gidaris 等[25] 使用概率预测方式来进一步提高目标检测的定位精度, 不同于边框位置回归的方法, 该方法首先将搜索区域划分成若干个水平区域和竖直区域, 然后给搜索区域的每列或每行分配概率, 利用这些概率信息来不断迭代获得更精确的检测框.

1.2 特征表示

特征表示是实现目标视觉检测必备的步骤, 选择合适的特征模型将图像区域映射为特征向量, 然后利用从训练样本学习到的分类器对该特征向量进行分类, 判断其所属类型. 特征的表达能力直接影响分类器精度, 决定了算法的最终性能. 特征模型主要分为手工设计的特征和自动学习的特征.

1.2.1 手工设计的特征

在深度学习热潮之前, 主要采用手工设计的特征. 手工特征数目繁多, 可以分为三大类: 基于兴趣点检测的方法、基于密集提取的方法和基于多种特征组合的方法.

1) 基于兴趣点检测的方法

兴趣点检测方法通过某种准则, 选择具有明确定义并且局部纹理特征比较明显的像素、边缘和角点等[3]. 其中 Sobel、Prewitt、Roberts、Canny 和 LoG (Laplacian of Gaussian) 等是典型的边缘检测算子[26−29]. 而 Harris、FAST (Features from accelerated segment test)、CSS (Curvature scale space) 和 DOG (Difference of Gaussian) 等是典型

的角点检测算子[30−32]. 兴趣点检测方法通常具有一定的几何不变性, 能够以较小的计算代价得到有意义的表达.

2) 基于密集提取的方法

密集提取方法主要提取局部特征. 区别于颜色直方图等全局特征, 局部特征有利于处理目标部分遮挡问题. 常用的局部特征有 SIFT (Scale-invariant feature transform)[33]、HOG (Histogram of oriented gradient)[5]、Haar-like[34] 和 LBP (Local binary pattern)[35−36] 等. 局部特征包含的信息丰富、独特性好, 并且具有较强的不变性和可区分性, 能够最大程度地对图像进行底层描述. 但是其计算一般比较复杂, 近些年图像的局部特征正在向快速和低存储方向发展.

3) 基于多种特征组合的方法

手工特征具有良好的可扩展性, 将兴趣点检测与密集提取相结合的多种特征组合方法, 能够弥补利用单一特征进行目标表示的不足. DPM (Deformable part-based model)[2] 提出了一种有效的多种特征组合模型, 被广泛应用于目标检测任务并取得了良好效果, 例如行人检测[37−38]、人脸检测[39−40] 和人体姿态估计[41] 等. 另外, 文献 [42] 提出了一种改进的 DPM 方法, 大大提升了检测速度.

依靠手工设计特征, 需要丰富的专业知识并且花费大量的时间. 特征的好坏在很大程度上还要依靠经验和运气, 往往整个算法的测试和调节工作都集中于此, 需要手工完成, 十分费力. 与之相比, 近年来受到广泛关注的深度学习理论中的一个重要观点就是手工设计的特征描述子作为视觉计算的第一步, 往往过早地丢失掉有用信息, 而直接从图像中学习到与任务相关的特征表示, 比手工设计特征更加有效[3].

1.2.2 自动学习的特征

近年来, 深度学习在图像分类和目标检测等领域取得了突破性进展, 成为目前最有效的自动特征学习方法. 深度学习模型具有强大的表征和建模能力, 通过监督或非监督的方式, 逐层自动地学习目标的特征表示, 将原始数据经过一系列非线性变换, 生成高层次的抽象表示, 避免了手工设计特征的繁琐低效. 深度学习在目标视觉检测中的研究现状是本文的核心内容, 将在第 3 节进行详细介绍.

1.3 区域分类

区域分类是指把候选区域的特征向量作为分类器输入, 预测候选区域所属的目标类型. 分类器在目标检测中的作用可以概括为: 先利用训练数据集进行模型学习, 然后利用学习到的模型对新的候选区域进行类型预测. 分类器一般是利用监督学习方法训练得到的, 常用的有支持向量机 (Support vector machine, SVM)、Adaboost、随机森林、神经网络等. 目前, 图像识别任务中广泛采用一对多 (One-vs-others) 的分类器训练方式[43], 就是把其中一类模式作为正样本, 其余模式作为负样本, 针对每一类模式分别训练一个分类器; 在测试阶段, 将图像特征分别输入到所有的分类器, 选择分类器响应最大的一类模式作为类型预测. Girshick 等[44] 就是采用这种方式, 提取候选区域的特征表示, 利用一对多 SVM 分类器实现对 PASCAL VOC 图像集 20 种目标的检测.

2 目标视觉检测的公共数据集

为了促进目标视觉检测的研究进展, 建设大规模的公共数据集成为必然要求. 目前, 目标视觉检测研究常用的公共数据集有 ImageNet、PASCAL VOC、SUN 和 MS COCO 等. 下面将从这些数据集包含的图像数目、类型数目、每类样本数等方面对它们进行介绍. 直观对比如图 2 所示.

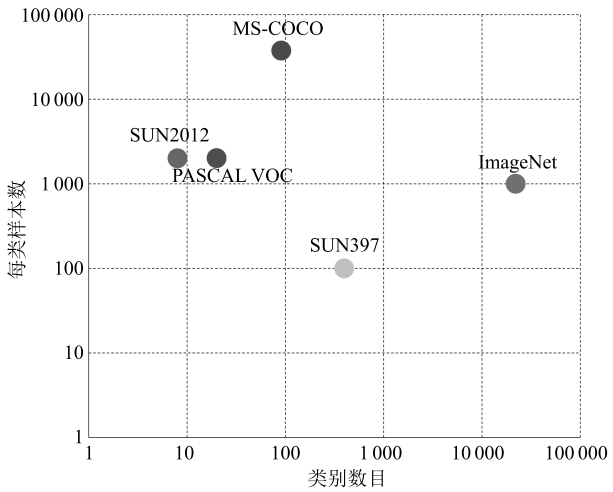

图 2 几种公共数据集的对比图

Fig. 2 Comparison of several common datasets

1) ImageNet 数据集[45]

该数据集是目前世界上最大的图像分类数据集, 包含 1 400 万幅图像、2.2 万个类型, 平均每个类型包含 1 000 幅图像. 此外, ImgeNet 还建立了一个包含 1 000 类物体, 有 120 万图像的数据集, 并将该数据集作为图像识别竞赛的数据平台.

2) PASCAL VOC 数据集[46]

2005 ∼ 2012 年, 该数据集每年都发布关于图像分类、目标检测和图像分割等任务的数据集, 并在相应数据集上举行算法竞赛, 极大地推动了计算机

视觉领域的研究进展. 该数据集最初只提供了 4 个类型的图像, 到 2007 年稳定在 20 个类; 测试图像的数量从最初的 1 578 幅, 到 2011 年稳定在 11 530 幅. 虽然该数据集类型数目比较少, 但是由于图像中物体变化极大, 每幅图像可能包含多个不同类型目标对象, 并且目标尺度变化很大, 因而检测难度非常大.

3) SUN 数据集[47]

该数据集是一个覆盖较大场景、位置、物体变化的数据集, 其中的场景名主要是从 WorldNet 中描述场景、位置、环境等任何具体的名词得来. SUN 数据集包含两个评测集: 一个是场景识别数据集, 称为 SUN 397, 共包含 397 类场景, 每类至少包含 100 幅图像, 总共有 108 754 幅图像; 另一个评测集为物体检测数据集, 称为 SUN 2012, 包含 16 873 幅图像.

4) MS COCO 数据集[48]

该数据集包含约 30 多万幅图像、200 多万个标注物体、91 个物体类型. 虽然比 ImageNet 和 SUN 包含的类型少, 但是每一类物体的图像多, 另外图像中包含精确的分割信息, 是目前每幅图像平均包含目标数最多的数据集. MS COCO 不但能够用于目标视觉检测研究, 还能够用来研究图像中目标之间的上下文关系.

3 深度学习在目标视觉检测中的应用进展

3.1 深度学习简介

深度学习模型具有强大的表征和建模能力, 通过监督或非监督的训练方式, 能够逐层、自动地学习目标的特征表示, 实现对物体层次化的抽象和描述. 1986 年, Rumelhart 等[49] 提出人工神经网络的反向传播 (Back propagation, BP) 算法. BP 算法指导机器如何从后一层获取误差而改变前一层的内部参数, 深度学习能够利用 BP 算法发现大数据中的复杂结构, 把原始数据通过一些简单的非线性函数变成高层次的抽象表达[50], 使计算机自动学习到模式特征, 从而避免了手工设计特征的繁琐低效问题. Hinton 等[51−52] 于 2006 年首次提出以深度神经网络为代表的深度学习技术, 引起学术界的关注. 之后, Bengio[53]、LeCun[54] 和 Lee[55] 等迅速开展了重要的跟进工作, 开启了深度学习研究的热潮. 深度学习技术首先在语音识别领域取得了突破性进展[56]. 在图像识别领域, Krizhevsky 等[57] 于 2012 年构建深度卷积神经网络, 在大规模图像分类问题上取得了巨大成功. 随后在目标检测任务中, 深度学习方法[7, 44, 58] 也超过了传统方法.

目前应用于图像识别和分析研究的深度学习模型主要包括堆叠自动编码器 (Stacked auto-encoders, SAE)[53]、深度信念网络 (Deep belief network, DBN)[51−52] 和卷积神经网络 (Convolutional neural networks, CNN)[59] 等.

SAE 模型的实质是多个自动编码器 (Auto-encoder, AE) 的堆叠. 一个自动编码器是由编码器和解码器两部分组成, 能够尽可能复现输入信号. 作为一种无监督学习的非线性特征提取方法, 其输出与输入具有相同的维度, 隐藏层则被用来进行原始数据的特征表示或编码. SAE 模型将前一层自动编码器的输出作为后一层自动编码器的输入, 逐层地对自动编码器进行预训练, 然后利用 BP 算法对整个网络进行微调. 目前基于 SAE 的扩展模型有很多, 例如, 堆叠去噪自动编码器 (Stacked denoising autoencoders, SDA)[60], 以及堆叠卷积自动编码器 (Stacked convolutional auto-encoders, SCAE)[61].

DBN 类似于 SAE, 它的基本单元是受限玻尔兹曼机 (Restricted Boltzmann machines, RBM), 整个网络的训练分为两个阶段: 预训练和全局微调. 首先以原始输入为可视层, 训练一个单层的 RBM, 该 RBM 训练完成后, 其隐层输出作为下一层 RBM 的输入, 继续训练下一层 RBM. 以此类推, 逐层训练, 直至将所有 RBM 训练完成, 通过这种贪婪式的无监督训练, 使整个 DBN 模型得到一个比较好的初始值, 然后加入数据标签对整个网络进行有监督的微调, 进一步改善网络性能.

CNN 是图像和视觉识别中的研究热点, 近年来取得了丰硕成果. 图 3 给出了由 LeCun 等[59] 提出的用于数字手写体识别的 CNN 网络结构, CNN 通常包含卷积层、池化层和全连接层. 卷积层通过使用多个滤波器与整个图像进行卷积, 可以得到图像的多个特征图表示; 池化层实际上是一个下采样层, 通过求局部区域的最大值或平均值来达到降采样的目的, 进一步减少特征空间; 全连接层用于进行高层推理, 实现最终分类. CNN 的权值共享和局部连接大大减少了参数的规模, 降低了模型的训练复杂度, 同时卷积操作保留了图像的空间信息, 具有平移不变性和一定的旋转、尺度不变性. 2012 年, Krizhevsky 等[57] 将 CNN 模型用于 ImageNet 大规模视觉识别挑战赛 (ImageNet large scale visual recognition challenge, ILSVRC) 的图像分类问题, 使错误率大幅降低, 在国际上引起了对 CNN 模型的高度重视, 也因此推动了目标视觉检测的研究进展.

3.2 AlexNet 及其改进模型

随着深度学习的发展, 人们将深度学习应用于

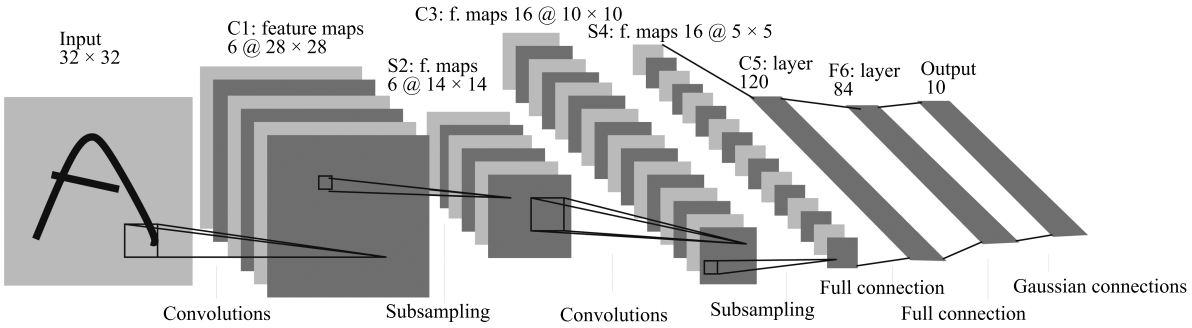

图 3 卷积神经网络的基本结构[59]

Fig. 3 Basic structure of convolutional neural network[59]

图像分类和目标检测任务中, 在许多公开竞赛中取得了明显优于传统方法的结果. Krizhevsky 等[57] 提出了一种新型卷积神经网络模型 AlexNet, 随后其他研究者相继提出 ZFNet[62]、VGG[63]、GoogLeNet[64] 和 ResNet[8] 等改进模型, 进一步提高了模型精度. 表 1 显示了几种经典 CNN 模型在图像分类任务中的性能对比. ILSVRC 的图像分类错误率每年都在被刷新, 如图 4 所示. 随着模型变得越来越深, 图像分类的 Top-5 错误率也越来越低, 目前已经降低到 3.08 % 附近[65]. 而在同样的 ImageNet 数据集上, 人眼的辨识错误率大约在 5.1 %. 尽管这些模型都是针对图像分类来做的, 但是都在解决一个最根本的问题, 即更强大的特征表示. 采用这些 CNN 模型得到更强大的特征表示, 然后应用到目标检测任务, 可以获得更高的检测精度.

表 1 经典 CNN 模型在 ILSVRC 图像分类任务上的性能对比

Table 1 Performance comparison of classical CNN model in image classification task of ILSVRC

CNN 模型	Top-5 错误率 (%)
AlexNet[57]	16.4
ZFNet[62]	14.8
VGG[63]	7.3
GoogLeNet[64]	6.7
ResNet[8]	3.57
Inception-v4, Inception-ResNet[65]	3.08

AlexNet[57] 在 ILSVRC 2012 图像分类任务上取得了 Top-5 错误率 16.4 %, 明显优于基于传统方法的第 2 名的结果 (Top-5 错误率 26.2 %). AlexNet 神经网络由 5 个卷积层、最大池化层、Dropout 层和 3 个全连接层组成, 网络能够对 1 000 个图像类型进行分类. 由于 AlexNet 的成功, 许多研究人员开始关注和改进 CNN 结构. Zeiler 等[62] 通过可视化 AlexNet 网络, 发现第 1 层滤波器是非常高频和低频信息的混合, 很少覆盖中间频率. 并且由于第 2 层卷积采用比较大的步长, 导致第 2 层出现混叠失真 (Aliasing artifacts). 为了解决这些问题, 他们将第 1 层滤波器的尺寸从 11 × 11 减小到 7 × 7, 将步长从 4 减小到 2, 形成 ZFNet 模型. ZFNet 在网络的第 1 层和第 2 层保留了更多信息, 降低了分类错误率.

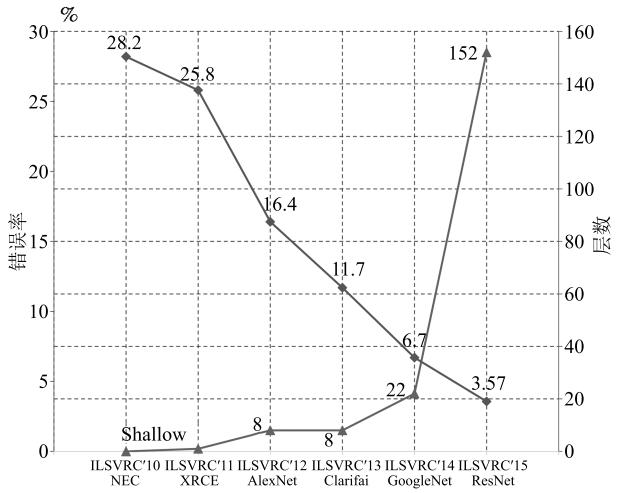

图 4 ILSVRC 图像分类任务历年冠军方法的 Top-5 错误率 (下降曲线) 和网络层数 (上升曲线)

Fig. 4 Top-5 error rate (descent curve) and network layers (rise curve) of the champion methods each year in image classification task of ILSVRC

Simonyan 等[63] 随后提出 VGG 网络, 探索在网络参数总数基本不变的情况下, CNN 随着层数的增加, 导致其性能的变化. 不同于 AlexNet, VGG 采用的滤波器尺寸是 3 × 3, 通过将多个 3 × 3 滤波器堆叠的方式来代替一个大尺寸的滤波器, 因为多个 3 × 3 尺寸的卷积层比一个大尺寸滤波器卷积层具有更高的非线性, 使模型更有判别能力, 而且多个 3 × 3 尺寸的卷积层比一个大尺寸的滤波器有更少

的参数. 通过加入 1×1 卷积层, 在不影响输入输出维数的情况下, 进一步增加网络的非线性表达能力.

Szegedy 等[64] 提出了一种新的深度 CNN 模型 GoogLeNet, 习惯上称为 Inception-v1. 只利用了比 AlexNet[57] 少 12 倍的参数, 但分类错误率更低. GoogLeNet 采用 Inception 结构, 上一层的输出经过 1×1、3×3、5×5 的卷积层和 3×3 的池化层, 然后拼接在一起作为 Inception 的输出. 并且在 3×3、5×5 卷积层之前采用 1×1 卷积层来降维, 既增加了网络的深度, 又减少了网络参数. Inception 结构既提高了网络对尺度的适应性, 又提高了网络计算资源的利用率. 但是深度网络在训练时, 由于模型参数在不断更新, 各层输入的概率分布在不断变化, 因此必须使用较小的学习率和较好的参数初值, 导致网络训练很慢, 同时也导致采用饱和的非线性激活函数 (例如 Sigmoid) 时训练困难. 为了解决这些问题, 又出现了 GoogLeNet 的续作 Inception-v2[66]. 它加入了批规范化 (Batch normalization) 处理, 将每一层的输出都进行规范化, 保持各层输入的分布稳定, 使得梯度受参数初值的影响减小. 批规范化加快了网络训练速度, 并且在一定程度上起到正则化的作用. Inception-v2 在 ILSVRC 2012 图像分类任务上的 Top-5 错误率降低到 4.8 %. 随着 Szegedy 等研究 GoogLeNet 的深入, 网络的复杂度也逐渐提高. Inception-v3[67] 变得更加复杂, 它通过将大的滤波器拆解成若干个小的滤波器的堆叠, 在不降低网络性能的基础上, 增加了网络的深度和非线性. Inception-v3 在 ILSVRC 2012 图像分类任务上的 Top-5 错误率降低到 3.5 %.

2015 年, He 等[8] 提出了深度高达上百层的残差网络 ResNet, 网络层数 (152 层) 比以往任何成功的神经网络的层数多 5 倍以上, 在 ImageNet 测试集上的图像分类错误率低至 3.57 %. ResNet 使用一种全新的残差学习策略来指导网络结构的设计, 重新定义了网络中信息流动的方式, 重构了网络学习的过程, 很好地解决了深度神经网络层数与错误率之间的矛盾 (即网络达到一定层数后, 更深的网络导致更高的训练和测试错误率). ResNet 具有很强的通用性, 不但在图像分类任务, 而且在 ImageNet 数据集的目标检测、目标定位任务以及 MS COCO 数据集的目标检测和分割任务上都取得了当时最好的竞赛成绩. 此后, Szegedy 等[65] 通过将 Inception 结构与 ResNet 结构相结合, 提出了 Inception-ResNet-v1 和 Inception-ResNet-v2 两种混合网络, 极大地加快了训练速度, 并且性能也有所提升. 除了这种混合结构, 他们还设计了一个更深更优化的 Inception-v4 网络, 单纯依靠 Inception 结构, 达到与 Inception-ResNet-v2 相近的性能. Szegedy 等[65] 将 3 个 Inception-ResNet-v2 网络和 1 个 Inception-v4 网络相集成, 在 ILSVRC 2012 图像分类任务上的 Top-5 错误率降低到 3.08 %.

3.3　深度学习在目标视觉检测中的应用

深度学习技术的发展, 极大推动了目标视觉检测研究. 目标检测与图像分类最主要的不同在于目标检测关注图像的局部结构信息, 而图像分类关注图像的全局表达. 与图像分类一样, 目标检测的输入也是整幅图像. 目标检测和图像分类在特征表示和分类器设计上有很大的相通性.

接下来, 我们从基于区域建议的方法和无区域建议的方法两方面来介绍深度学习在目标视觉检测中的研究现状.

3.3.1　基于区域建议 (Proposal-based) 的方法

Girshick 等[44] 提出的 R-CNN (Region-based convolutional neural networks) 方法, 是近年来基于深度学习的目标检测研究的重要参考方法. R-CNN 将目标区域建议 (Region proposal) 和 CNN 相结合, 在 PASCAL VOC 2012 上的检测平均精度 mAP (Mean average precision) 达到 53.3 %, 比传统方法有了明显改进. R-CNN 的基本流程如图 5 所示, 首先对每一幅输入图像, 采用选择性搜索 (Selective search)[6] 来提取候选区域; 然后用 CNN 网络从每个区域提取一个固定长度的特征向量, 这里采用 AlexNet[57] 结构, 图像经过 5 个卷积层和 2 个全连接层, 得到一个 4 096 维的特征向量; 接着把提取到的特征向量送入支持向量机进行分类. 由于一些区域存在高度交叠, Girshick 等采用非极大值抑制 (Non-maximum suppression) 来舍弃那些与更高得分区域的 IoU (Intersection-over-Union) 过大的区域. 为了得到更精确的结果, 还采用了边框回归方法来进一步改善检测结果. 在 R-CNN 模型的训练过程中, 由于目标检测标注数据集的规模不够, Girshick 等先将网络在大规模数据集 ImageNet 上进行预训练, 然后用 $N+1$ 类 (N 个目标类和 1 个背景类) 的输出层来替换 1 000 类的 Softmax 层, 再

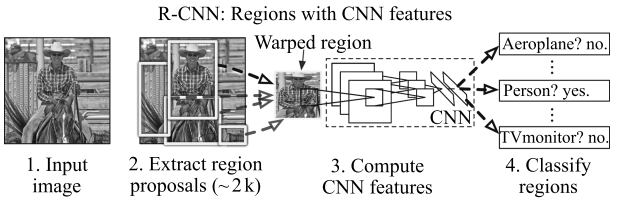

图 5　R-CNN 的计算流程[44]

Fig. 5　Calculation flow of R-CNN[44]

针对目标检测任务, 用 PASCAL VOC 数据集进行微调. 这种方法很好地解决了训练数据不足的问题, 进一步提升了检测精度. 得益于 CNN 的参数共享以及更低维度的特征, 整个检测算法更加高效. 但是, R-CNN 也存在一些不容忽视的问题: 1) 候选区域之间的交叠使得特征被重复提取, 造成了严重的速度瓶颈, 降低了计算效率; 2) 将候选区域直接缩放到固定大小, 破坏了物体的长宽比, 可能导致物体的局部细节损失; 3) 使用边框回归有助于提高物体的定位精度, 但是如果待检测物体存在遮挡, 该方法将难以奏效.

He 等[68] 针对 R-CNN 速度慢以及要求输入图像块尺寸固定的问题, 提出空间金字塔池化 (Spatial pyramid pooling, SPP) 模型. 在 R-CNN 中, 要将提取到的目标候选区域变换到固定尺寸, 再输入到卷积神经网络, He 等加入了一个空间金字塔池化层来避免了这个限制. SPP-net 网络不论输入图像的尺寸大小, 都能产生固定长度的特征表示. SPP-net 是对整幅图像提取特征, 在最后一层卷积层得到特征图后, 再针对每个候选区域在特征图上进行映射, 由此得到候选区域的特征. 因为候选区域的尺寸各不相同, 导致它们映射所得到的特征图大小也不同, 但 CNN 的全连接层需要固定维度的输入, 因此引入了空间金字塔池化层来把特征转换到相同的维度. 空间金字塔池化的思想来源于空间金字塔模型 (Spatial pyramid model, SPM)[43], 它采用多个尺度的池化来替代原来单一的池化. SPP 层用不同大小的池化窗口作用于卷积得到的特征图, 池化窗口的大小和步长根据特征图的尺寸进行动态计算. SPP-net 对于一幅图像的所有候选区域, 只需要进行一次卷积过程, 避免了重复计算, 显著提高了计算效率, 而且空间金字塔池化层使得检测网络可以处理任意尺寸的图像, 因此可以采用多尺度图像来训练网络, 从而使得网络对目标的尺度有很好的鲁棒性. 该方法在速度上比 R-CNN 提高 24 ∼ 102 倍, 并且在 PASCAL VOC 2007 和 Caltech 101 数据集上取得了当时最好的成绩. 但是它存在以下缺点: 1) SPP-net 的检测过程是分阶段的, 在提取特征后用 SVM 分类, 然后还要进一步进行边框回归, 这使得训练过程复杂化; 2) CNN 提取的特征存储需要的空间和时间开销大; 3) 在微调阶段, SPP-net 只能更新空间金字塔池化层后的全连接层, 而不能更新卷积层, 这限制了检测性能的提升.

后来, Girshick 等[58] 对 R-CNN 和 SPP-net 进行了改进, 提出能够实现特征提取、区域分类和边框回归的端到端联合训练的 Fast R-CNN 算法, 计算流程如图 6 所示. 与 R-CNN 类似, Fast R-CNN 首先在图像中提取感兴趣区域 (Regions of Interest, RoI); 然后采用与 SPP-net 相似的处理方式, 对每幅图像只进行一次卷积, 在最后一个卷积层输出的特征图上对每个 RoI 进行映射, 得到相应的 RoI 的特征图, 并送入 RoI 池化层 (相当于单层的 SPP 层, 通过该层把各尺寸的特征图统一到相同的大小); 最后经过全连接层得到两个输出向量, 一个进行 Softmax 分类, 另一个进行边框回归. 在微调阶段, Fast R-CNN 采用一种新的层级采样方法, 先采样图像, 再从采样出的图像中对 RoI 进行采样, 同一幅图像的 RoI 共享计算和内存, 使得训练更加高效. Fast R-CNN 采用 Softmax 分类与边框回归一起进行训练, 省去了特征存储, 提高了空间和时间利用率, 同时分类和回归任务也可以共享卷积特征, 相互促进. 与 R-CNN 相比, 在训练 VGG 网络时, Fast R-CNN 的训练阶段快 9 倍, 测试阶段快 213 倍; 与 SPP-net 相比, Fast R-CNN 的训练阶段快 3 倍, 测试阶段快 10 倍, 并且检测精度有一定提高. 然而, Fast R-CNN 仍然存在速度上的瓶颈, 就是区域建议步骤耗费了整个检测过程的大量时间.

为了解决区域建议步骤消耗大量计算资源, 导致目标检测不能实时的问题, Ren 等[7] 提出区域建议网络 (Region proposal network, RPN), 并且把 RPN 和 Fast R-CNN 融合到一个统一的网络 (称为 Faster R-CNN), 共享卷积特征. RPN 将一整幅图像作为输入, 输出一系列的矩形候选区域. 它是一个全卷积网络模型, 通过在与 Fast R-CNN 共享卷积层的最后一层输出的特征图上滑动一个小型网络, 这个网络与特征图上的小窗口全连接, 每个滑动窗口映射到一个低维的特征向量, 再输入给两个并列的全连接层, 即分类层 (cls layer) 和边框回归层 (reg layer), 由于网络是以滑动窗的形式来进行操作, 所以全连接层的参数在所有空间位置是共享的. 因此该结构由一个卷积层后连接两个并列的 1×1 卷积层实现, 如图 7 所示. 对于每个小窗口, 以中心点为基准点选取 k (作者采用 $k = 9$) 个不同尺度、不同长宽比的 Anchor. 对于每个 Anchor, 分类层输

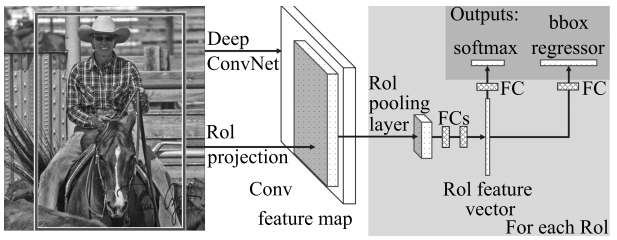

图 6 Fast R-CNN 的计算流程[58]

Fig. 6 Calculation flow of Fast R-CNN[58]

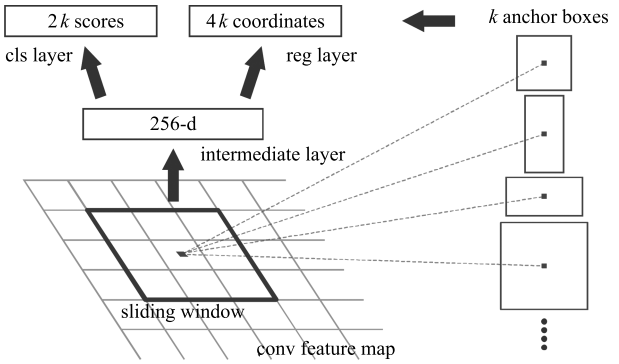

图 7　区域建议网络的基本结构[7]

Fig. 7　Basic structure of region proposal network[7]

出 2 个值, 分别表示其属于目标的概率与属于背景的概率; 边框回归层输出 4 个值, 表示其坐标位置. RPN 的提出, 以及与 Fast R-CNN 进行卷积特征的共享, 使得区域建议步骤的计算代价很小. 与以前的方法相比, 提取的候选区域数量大幅减少, 同时改进了候选区域的质量, 从而提高了整个目标检测网络的性能, 几乎可以做到实时检测. 在 PASCAL VOC 2007 和 2012、MS COCO 等数据集上, Faster R-CNN 取得了当时最高的检测精度. 但是由于深度特征丢失了物体的细节信息, 造成定位性能差, Faster R-CNN 对小尺寸物体的检测效果不好.

Bell 等[69] 提出的 ION (Inside-outside net) 也是基于区域建议的目标检测方法. 为了提高检测精度, ION 同时利用 RoI 的内部和外部信息. 其中内部信息是指多尺度的信息提取. 不同于以前的方法将最后一层卷积层输出作为特征图, Bell 等将不同卷积层的特征连接在一起, 作为一个多尺度特征用来预测, 这样做的目的是对于一些很小的物体, 不会丢失在低层的高分辨率信息. RoI 的外部信息是指上下文信息, 在视觉识别中上下文信息具有很重要的作用. 为了得到上下文特征, Bell 等采用沿着图像的横轴或纵轴独立地使用 RNN 的方法, 并把它们的输出组合在一起, 重复该过程得到的输出作为上下文特征. 最后把这两种特征组合在一起, 并调整到固定的大小输入到全连接层, 进行 Softmax 分类和边框回归. 该方法在检测小物体上的性能比以前的方法更好, 在 PASCAL VOC 2012 目标检测任务上将平均精度 mAP 从 73.9 % 提高到 76.4 %, 在 MS COCO 2015 目标检测任务上取得第 3 名的成绩.

Yang 等[70] 为了处理不同尺度的目标, 并且提高对候选区域的计算效率, 提出了两个策略, 统称为 SDP-CRC. 一个策略是采用与尺度相关的池化层 (Scale-dependent pooling, SDP), 由于不同尺寸的物体可能在不同的卷积层上得到不同的响应, 小尺寸物体会在浅层得到强响应, 而大尺寸物体可能在深层得到强响应. 基于这一思想, SDP 根据每个候选区域的尺寸, 从对应的卷积特征图上池化特征. 对于小尺度的候选区域, 从第三层卷积特征图上池化特征; 对于中等尺度的候选区域, 从第四层卷积特征图上池化特征; 对于大尺度的候选区域, 从第五层卷积特征图上池化特征. 另一个策略是采用级联拒绝分类器 (Cascaded rejection classifier, CRC), 快速排除一些明显不包含目标的候选区域, 只保留那些更可能包含目标的候选区域, 交由 Fast R-CNN 做最终分类. 与 Fast R-CNN 相比, 该方法能够更加准确地检测小尺寸目标, 在平均检测精度和检测速度上都有很大提升.

为了提高 Fast R-CNN 训练时的效率, Shrivastava 等[71] 提出了困难样本在线挖掘 (Online hard example mining, OHEM) 的思想, 该方法利用 Bootstrapping[72] 技术, 对随机梯度下降算法进行修改, 使得在训练过程中加入在线挖掘困难样本的策略. OHEM 机制的加入提高了 Fast R-CNN 方法在 PASCAL VOC 2007 和 2012 上的检测精度.

在 Faster R-CNN 基础上, Kong 等[73] 提出了 HyperNet, 计算流程如图 8 所示. 通过把不同卷积层得到的特征图像聚集起来得到超特征 (Hyper feature) 来获得质量更高的候选区域. 由于不同卷积层的输出尺寸不同, 较浅层的特征图像分辨率较高, 边框定位精度高, 但是召回率低; 较深层的特征图像分辨率低, 对小尺寸物体的边框定位精度低, 但是这些特征有利于提高召回率. 因此, 他们通过多层特征的融合, 解决了对小物体很难提取到精细特征的问题. 该方法在每幅图像中仅提取 100 个候选区域, 在 PASCAL VOC 2007 和 2012 数据集上获得了很好的检测效果.

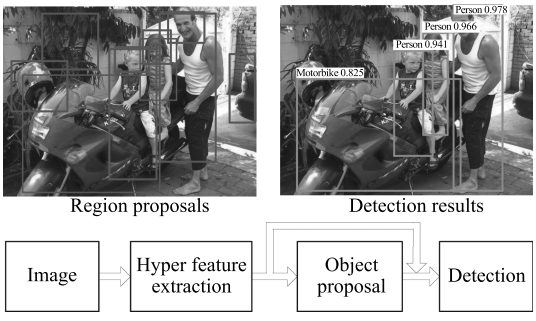

图 8　HyperNet 的计算流程[73]

Fig. 8　Calculation flow of HyperNet[73]

许多基于区域建议的目标检测方法存在一个共同问题, 就是有一部分子网络需要重复计算. 例如最早提出的 R-CNN, 每一个候选区域都要经历一次 CNN 网络提取特征, 这导致目标检测速度非常慢.

之后提出的 Fast R-CNN 和 Faster R-CNN 等方法, 在最后一个卷积层通过 RoI pooling 把每一个候选区域变成一个尺寸一致的特征图, 但是对于每一个特征图, 还要经过若干次全连接层才能得到结果. 于是, Dai 等[74] 提出了一种新的基于区域的全卷积网络检测方法 R-FCN. 为了给网络引入平移变化, 用专门的卷积层构建位置敏感的分数图 (Position-sensitive score maps), 编码感兴趣区域的相对空间位置信息. 该网络解决了 Faster R-CNN 由于重复计算全连接层而导致的耗时问题, 实现了让整个网络中所有的计算都可以共享.

最近, Kim 等[75] 提出 PVANET 网络, 在 TITAN X 上实现了基于轻量级模型的目标检测, 处理一幅图像仅需要 46 ms, 在 PASCAL VOC 2012 数据集上的检测平均精度达到 82.5 %. 为了减少网络参数, PVANET 采用了 Concatenated ReLU[76] 结构, 在不损失精度的情况下使通道数减少一半, 并在拼接操作之后加入了尺度变化和偏移. 网络中还加入了 Inception[64] 模型来更有效地捕捉各种尺度的物体, 以及 HyperNet[73] 中多尺度特征融合的思想, 来增加对细节的提取.

3.3.2 无区域建议 (Proposal-free) 的方法

基于区域建议的目标检测方法不能利用局部目标在整幅图像中的空间信息, 所以一些研究者开展了无区域建议的目标检测研究, 主要采用回归的思想. 早期提出的无区域建议的方法, 检测效果不太理想.

DPM 模型[2] 是一种性能较好的传统目标检测模型. 它对目标内在部件进行结构化建模, 可以更好地适应非刚体目标的较大形变, 大大提高了检测性能. 但是 DPM 模型的构建需要关于物体结构的先验知识 (例如部件个数), 并且模型训练也比较复杂. Szegedy 等[1] 将目标检测看做一个回归问题, 估计图像中的目标位置和目标类型概率. 作者通过采用基于深度神经网络 (Deep neural network, DNN) 的回归来输出目标包围窗口的二元掩膜 (Mask), 从掩膜中提取目标窗口. 该方法的运行框架如图 9 所示, 网络中采用的卷积神经网络是 AlexNet 结构, 但是用回归层代替最后一层. 基于 DNN 的回归不仅能学习到有利于分类的特征表示, 还能捕获到很强的目标几何信息, Szegedy 等还采用 DNN 定位器进一步提高了定位准确度. 由于用单一的掩膜很难区分出识别的前景是单个物体还是粘连的多个物体, 作者采用了多个掩膜, 为每种掩膜训练一个单独的 DNN, 这也使得网络训练复杂度很高, 很难扩展到多种目标类型.

Sermanet 等[77] 提出 Overfeat 模型, 把一个卷积神经网络同时用于分类、定位和检测这几个不同的任务. 卷积层作为特征提取层保持不变, 只需要针对不同的任务改变网络的最后几层为分类或回归层. Overfeat 的模型结构与 AlexNet 结构[57] 基本相同. 其中, 前面 5 个卷积层为不同任务的共享层, 其余的层则根据任务进行相应的调整, 并对网络做了一些改动. 为了避免图像的某些位置被忽略, Sermanet 等采用偏置池化层来替换最后一层池化层, 既实现了池化操作, 也减小了采样间隔. Overfeat 训练分类模型时只使用单个尺度 (221×221) 进行训练, 测试时使用多个尺度输入图像, 没有使用 AlexNet 中的对比归一化. 对于检测问题, 传统的方法是采用不同尺寸的滑动窗对整幅图像进行密集采样, 然后对每一个采样所得的图像块进行检测, 从而确定目标物体的位置. Overfeat 使用 CNN 来进行滑动窗操作, 避免了对各图像块的单独操作, 提高了算法效率; 而且将全连接层看作卷积层, 使得输入图像的尺寸不受限制. 但是 Overfeat 对于较小尺寸目标的识别依然存在困难.

近年来, Redmon 等[78] 提出了一种新的无区域建议的目标检测方法, 称为 YOLO (You only look once). 作为一种统一的、实时的检测框架, YOLO 的检测速度非常快, 可以达到 45 fps (Frame per second). YOLO 用一个单一的卷积网络直接基于整幅图像来预测包围边框的位置及所属类型, 首先将一幅图像分成 $S \times S$ 个网格, 每个网格要预测 B 个边框, 每个边框除了要回归自身的位置之外, 还要附带预测一个置信度. 置信度不仅反映了包含目标的

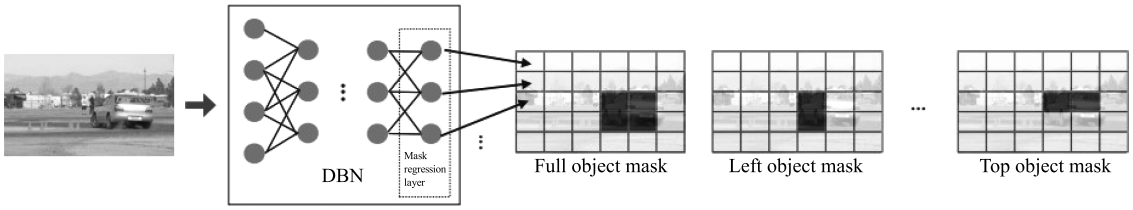

图 9 基于 DNN 回归的目标检测框架[1]

Fig. 9 Object detection framework based on DNN regression[1]

可信程度, 也反映了预测位置的准确度. 另外对每个网格还要预测 C 个类型的条件概率, 将这些预测结果编码为一个 $S \times S \times (B \times 5 + C)$ 维的张量 (Tensor). 整个网络的结构类似于 GoogLeNet, 包含 24 个卷积层和 2 个全连接层, 卷积层用来从图像中提取特征, 全连接层预测边框的位置坐标和类型概率. YOLO 模型通过采用空间限制, 减少了对同一目标的重复检测, 大大提高了效率, 能够达到实时的效果. 但是 YOLO 的整体性能不如 Fast R-CNN 和 Faster R-CNN, 并且对于相邻的目标和成群的小尺寸目标 (例如成群的鸟) 的检测效果不好, 对于新的或异常尺度的目标泛化能力较差.

与 YOLO 类似, Najibi 等[79] 提出的 G-CNN 模型也着重于检测速度的提升. 该方法将目标检测模型转化为迭代回归问题, 通过对整个图像进行不同尺度的网格划分得到初始检测框, 然后采用分段回归模型多次迭代, 不断提高边框准确度. G-CNN 使用了约 180 个初始边框, 经过 5 次迭代达到与 Fast R-CNN 相当的检测精度, 但是计算速度比 Fast R-CNN 快 5 倍.

针对 YOLO 存在的不足, Liu 等[80] 提出 SSD 模型, 在提高 mAP 的同时兼顾实时性的要求. SSD 使用卷积神经网络对图像进行卷积后, 在不同层次的特征图上生成一系列不同尺寸和长宽比的边框. 在测试阶段, 该网络对每一个边框中分别包含各个类型的物体的可能性进行预测, 并且调整边框来适应目标物体的形状. 在 PASCAL VOC、MS COCO 和 ILSVRC 数据集上的实验显示, SSD 在保证精度的同时, 其速度要比用候选区域的方法快很多. 与 YOLO 相比, 即使是在输入图像较小的情况下, SSD 也能取得更高的精度. 例如输入 300×300 尺寸的 PASCAL VOC 2007 测试图像, 在单台 Nvidia Titan X 上的处理速度达到 58 fps, 平均精度 mAP 达到 72.1 %; 如果输入图像尺寸为 500×500, 平均精度 mAP 达到 75.1 %.

与基于候选区域的方法相比, YOLO 定位准确率低且召回率不高. 因此, Redmon 等[81] 提出了改进的 YOLO 模型, 记作 YOLOv2, 主要目标是在保持分类准确率的同时提高召回率和定位准确度. 通过采用多尺度训练、批规范化和高分辨率分类器等多种策略, 提升了检测准确率的同时速度超过其他检测方法, 例如 Faster R-CNN 和 SSD. Redmon 等还提出了一种新的联合训练算法, 同时在检测数据集和分类数据集上训练物体检测器, 用检测数据集的数据学习物体的准确位置, 用分类数据集的数据增加分类的类别量, 提升健壮性, 采用这种方法训练出来的 YOLO9000 模型可以实时地检测超过 9 000 种物体分类.

3.3.3 总结

基于区域建议的目标检测方法, 特别是 R-CNN 系列方法 (包括 R-CNN、SPPnet、Fast R-CNN 和 Faster R-CNN 等), 取得了非常好的检测精度, 但是在速度方面还达不到实时检测的要求. 在不损失精度的情况下实现实时检测, 或者在提高检测精度的同时兼顾速度, 逐渐成为目标检测的研究趋势. R-FCN 比 Faster R-CNN 计算效率更高, 在检测精度和速度上平衡的很好. PVANET 是一种轻量级的网络结构, 通过调整和结合最新的技术达到最小化计算资源的目标. 无区域建议的方法 (例如 YOLO) 虽然能够达到实时的效果, 但是其检测精度与 Faster R-CNN 相比有很大的差距. SSD 对 YOLO 进行了改进, 同时兼顾检测精度和实时性的要求, 在满足实时性的条件下, 缩小了与 Faster R-CNN 检测精度的差距. YOLOv2 在检测精度和速度上都超过了 SSD. 一些目标视觉检测方法在公共数据集上的性能对比如图 10 所示.

4 思考与展望

近年来, 由于深度学习技术的迅猛发展和应用, 目标视觉检测研究取得了很大进展. 未来若干年, 基于深度学习的目标视觉检测研究仍然是该领域的主流研究方向. 不同于传统方法利用手工设计的特征, 可能忽视掉一些重要的特征信息, 深度学习方法可以通过端到端训练自动学习与任务相关的特征, 通过多层的非线性变换获得图像的高层次抽象表示. 尽管深度学习在目标视觉检测领域取得了一定成功, 但是还存在一些问题:

1) 深度学习理论还不完善

深度学习的优势之一是能够自动学习表达能力强的抽象特征, 不需要由专家手工进行特征设计和选择. 但是, 将深度学习模型应用于目标检测时还缺乏足够的理论支撑, 学习到的模型的可解释性较弱. 目前的研究通常是把深度学习模型当作一个黑盒子 (Black box) 来直接使用, 对于如何选择和构建模型、如何确定模型的深度以及深度学习的本质等基本问题还没有给出很好的解释. 理论的不完善导致研究时缺乏充分的原理性指导, 在设计新的模型时往往只能凭借经验和运气. Pepik 等[82] 利用 Pascal 3D+[83] 数据集对 R-CNN 方法进行分析, 结果表明卷积神经网络对于场景和目标的各种外观因素的变化不具有视觉不变性, 目前大多数深度学习方法在处理多目标遮挡和小尺寸目标等困难问题时效果还不是很好, 增加额外的训练数据并不能克服

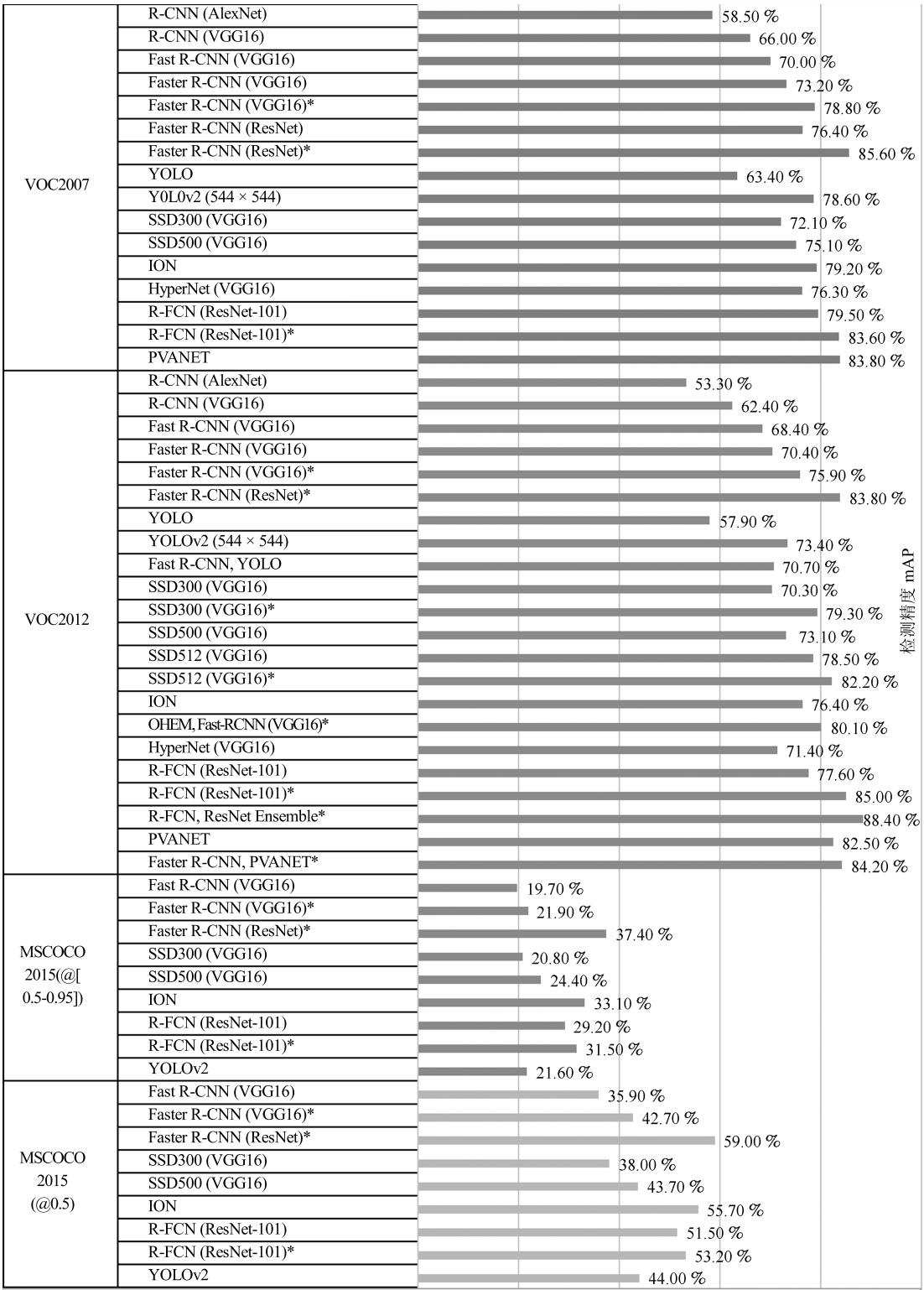

图 10 一些目标视觉检测方法在公共数据集上的性能比较

Fig. 10 Performance comparison of some object visual detection methods on public datasets

这些缺陷, 有必要对模型结构做出改变. 因此必须进一步完善深度学习理论, 为改进模型结构、加速模型训练和提高检测效果等提供指导.

2) 大规模多样性数据集还很缺乏

深度学习模型主要是数据驱动的, 依赖于大规模多样性的标记数据集. 对一个特定的任务, 增加训练数据的规模和多样性, 可以提高深度学习模型的泛化能力, 避免过拟合. 但是目前缺乏可用于目标检测的大规模多样性数据集, 即便是最大的公共数据集也只提供了很有限的标记类型, 比如 PASCAL VOC 有 20 个类型, MS COCO 有 80 个类型, ImageNet 有 1 000 个类型. 由人工采集和标注含有大量目标类型的大规模多样性数据集非常费时耗力, 并且由于光照、天气、复杂背景、目标外观、摄像机视角和物体遮挡等导致的复杂性和挑战性, 同一类型目标在不同图像中可能看起来非常不同, 使得人工标注变得困难甚至容易出错. 虽然可以采用众包方法 (例如 Amazon MTurk[84]) 进行数据标注, 但是同样要耗费大量的人力财力, 并且标注困难. 另外在一些特殊领域 (例如在医疗和军事等领域) 很难获得大规模实际图像. 标记数据集的不足, 可能导致训练出的目标检测模型的可靠性和鲁棒性达不到要求. 目前许多目标检测模型都采用先在 ImageNet 数据集上进行预训练, 再针对具体任务进行微调的方式. 如果针对具体的目标检测任务, 有大规模多样性的标记数据集可供使用, 那么目标检测效果可以得到进一步提高.

为了解决上述问题, 我们认为可以采用平行视觉[85−86] 的思路进行研究. 2016 年, 王坤峰等[85] 将复杂系统建模与调控的 ACP (Artificial societies, computational experiments, and parallel execution) 理论[87−89] 推广到视觉计算领域, 提出平行视觉的基本框架和关键技术. 其核心是利用人工场景来模拟和表示复杂挑战的实际场景, 通过计算实验进行各种视觉模型的设计与评估, 最后借助平行执行来在线优化视觉系统, 实现对复杂环境的智能感知与理解. 图 11 显示了平行视觉的基本框架. 为了解决复杂环境下的目标视觉检测问题, 我们可以按照平行视觉的 ACP 三步曲开展研究.

1) 人工场景 (Artificial scenes)

构建色彩逼真的人工场景, 模拟实际场景中可能出现的环境条件, 自动得到精确的目标位置、尺寸和类型等标注信息, 生成大规模多样性数据集. 另外, 实际场景通常不可重复, 而人工场景具有可重复性, 通过固定一些物理模型和参数, 改变另外一些, 可以定制图像生成要素, 以便从各种角度评价视觉算法. 人工场景可以不受现有实际场景的限制, 预见未来的实际场景, 为视觉算法设计与评估提供超前信息. 总之, 人工场景能够提供一种可靠的数据来源, 是对实际场景数据的有效补充.

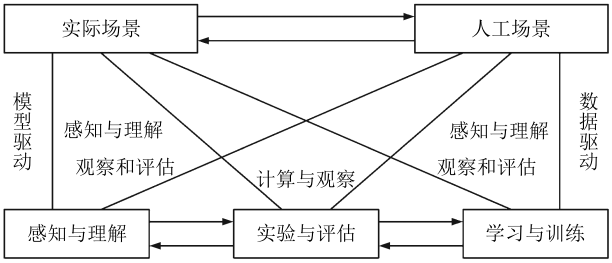

图 11　平行视觉的基本框架[85]

Fig. 11　Basic framework of parallel vision[85]

2) 计算实验 (Computational experiments)

结合人工场景数据集和实际场景数据集, 进行全面充分的计算实验, 把计算机变成视觉计算实验室, 设计和评价视觉算法, 提高其在复杂环境下的性能. 与基于实际场景的实验相比, 在人工场景中实验过程可控、可观、可重复, 并且可以真正地产生实验大数据, 用于知识提取和算法优化. 计算实验包含两种操作模式, 即学习与训练、实验与评估. 学习与训练是针对视觉算法设计而言, 实验与评估是针对视觉算法评价而言. 两种操作模式都需要结合人工场景数据集和实际场景数据集, 能够增加实验的深度和广度.

3) 平行执行 (Parallel execution)

将视觉算法在实际场景与人工场景中平行执行, 使模型训练和评估在线化、长期化, 通过实际与人工之间的虚实互动, 持续优化视觉系统. 由于应用环境的复杂性、挑战性和变化性, 不存在一劳永逸的解决方案, 只能接受这些困难, 在系统运行过程中不断调节和改善. 平行执行基于物理和网络空间的大数据, 以人工场景的在线构建和利用为主要手段, 通过在线自举 (Online bootstrapping) 或困难实例挖掘 (Hard example mining), 自动挖掘导致视觉算法失败或性能不佳的实例, 利用它们重新调节视觉算法和系统, 提高对动态变化环境的自适应能力.

目前, 已经有一些工作基于人工场景数据进行目标检测模型的训练. 例如, Peng 等[90] 利用 3D CAD 模型自动合成 2D 图像, 使用这种虚拟图像数据来扩大深度卷积神经网络的训练集非常有效, 尤其是在真实的训练数据很有限或不能很好地匹配目标领域的情况下, 避免了代价昂贵的大规模手工标注. Johnson-Roberson 等[91] 利用游戏引擎生成逼真的虚拟图像, 用于目标检测模型的训练. 实验表明, 在 KITTI 数据集上, 使用大规模的虚拟图像集

训练的模型比基于较小规模的真实世界数据集训练的检测器精度更高. 但是, 已有的工作主要集中在人工场景和计算实验, 忽视了平行执行. 我们认为, 将视觉算法在实际场景与人工场景中平行执行, 持续优化视觉系统, 提高其在复杂环境下的鲁棒性和适应性是非常重要的.

许多机器学习算法假设训练数据与测试数据具有相同的数据分布以及特征空间[92], 然而使用 ACP 时会遇到虚拟数据与真实数据的分布差异问题. 迁移学习[93] 能够很好解决分布差异问题. 通过迁移学习, 我们能够运用 ACP 中人工模拟出的虚拟数据来不断提高模型的精准度与鲁棒性.

另外, 在深度学习模型自身方面, 如何提高模型的可解释性, 改善模型结构, 设计新的优化方法, 降低模型训练和应用时的计算复杂性, 提高计算效率, 得到更加有用 (More effective) 和更加有效的 (More efficient) 深度学习模型, 这些问题都需要深入研究. 目前, 基于候选区域的目标检测方法精度最高, 而基于回归的 SSD 方法在实时性上表现最好, 如何将这两类方法相结合, 借鉴和吸收彼此的优点, 在检测精度和速度上取得新的突破还有待研究.

5 结论

目标视觉检测在计算机视觉领域具有重要的研究意义和应用价值, 深度学习是目前最热门的机器学习方法, 被广泛研究和应用. 本文综述了深度学习在目标视觉检测中的应用进展与展望. 首先说明了目标视觉检测的基本流程和常用的公共数据集, 然后重点介绍了深度学习方法在目标视觉检测中的最新应用进展, 最后对深度学习在目标视觉检测研究中的困难和挑战进行了分析, 对未来的发展趋势进行了思考与展望.

在今后的工作中, 还需要进一步完善深度学习理论, 提高目标视觉检测的精度和效率. 另外, 平行视觉作为一种新的智能视觉计算方法学, 通过人工场景提供大规模多样性的标记数据集, 通过计算实验全面设计和评价目标视觉检测方法, 通过平行执行在线优化视觉系统, 能够激发深度学习的潜力. 我们相信, 深度学习与平行视觉相结合, 必将大力推动目标视觉检测的研究和应用进展.

References

1 Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Proceedings of the 2013 Advances in Neural Information Processing Systems (NIPS). Harrahs and Harveys, Lake Tahoe, USA: MIT Press, 2013. 2553−2561

2 Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

3 Huang Kai-Qi, Ren Wei-Qiang, Tan Tie-Niu. A review on image object classification and detection. *Chinese Journal of Computers*, 2014, **37**(6): 1225−1240
(黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述. 计算机学报, 2014, **37**(6): 1225−1240)

4 Zhang X, Yang Y H, Han Z G, Wang H, Gao C. Object class detection: a survey. *ACM Computing Surveys (CSUR)*, 2013, **46**(1): Article No. 10

5 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA: IEEE, 2005, **1**: 886−893

6 Uijlings J R R, van de Sande K E A, Gevers T, Smeulders A W M. Selective search for object recognition. *International Journal of Computer Vision*, 2013, **104**(2): 154−171

7 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137−1149

8 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, Nevada, USA: IEEE, 2016. 770−778

9 Lampert C H, Blaschko M B, Hofmann T. Beyond sliding windows: object localization by efficient subwindow search. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska, USA: IEEE, 2008. 1−8

10 An S J, Peursum P, Liu W Q, Venkatesh S. Efficient algorithms for subwindow search in object detection and localization. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, Florida, USA: IEEE, 2009. 264−271

11 Wei Y C, Tao L T. Efficient histogram-based sliding window. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, 2010. 3003−3010

12 Van de Sande K E A, Uijlings J R R, Gevers T, Smeulders A W M. Segmentation as selective search for object recognition. In: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 1879−1886

13 Shotton J, Blake A, Cipolla R. Multiscale categorical object recognition using contour fragments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(7): 1270−1281

14 Leibe B, Leonardis A, Schiele B. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 2008, **77**(1−3): 259−289

15 Arbelaez P, Maire M, Fowlkes C, Malik J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(5): 898−916

16 Shotton J, Winn J, Rother C, Criminisi A. TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Proceedings of the 9th European Conference on Computer Vision (ECCV). Berlin, Heidelberg, Germany: Springer, 2006. 1−15

17 Verbeek J, Triggs B. Region classification with Markov field aspect models. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Minneapolis, Minnesota, USA: IEEE, 2007. 1−8

18 Cheng M M, Zhang Z M, Lin W Y, Torr P. BING: binarized normed gradients for objectness estimation at 300fps. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE, 2014. 3286−3293

19 Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 391−405

20 Hosang J, Benenson R, Schiele B. How good are detection proposals, really? arXiv: 1406.6962, 2014.

21 Szegedy C, Reed S, Erhan D, Anguelov D, Ioffe S. Scalable, high-quality object detection. arXiv: 1412.1441, 2014.

22 Erhan D, Szegedy C, Toshev A, Anguelov D. Scalable object detection using deep neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, Ohio, USA: IEEE, 2014. 2155−2162

23 Kuo W C, Hariharan B, Malik J. Deepbox: learning objectness with convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2479−2487

24 Ghodrati A, Diba A, Pedersoli M, Tuytelaars T, Van Gool L. Deepproposal: hunting objects by cascading deep convolutional layers. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2578−2586

25 Gidaris S, Komodakis N. Locnet: improving localization accuracy for object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 789−798

26 Lawrence G R. Machine Perception of Three-dimensional Solids [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 1963.

27 Canny J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986, **PAMI-8**(6): 679−698

28 Marr D, Hildreth E. Theory of edge detection. *Proceedings of the Royal Society B: Biological Sciences*, 1980, **207**(1167): 187−217

29 Pellegrino F A, Vanzella W, Torre V. Edge detection revisited. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2004, **34**(3): 1500−1518

30 Harris C, Stephens M. A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference. Manchester, UK: University of Sheffield Printing Unit, 1988. 147−151

31 Rosten E, Porter R, Drummond T. Faster and better: a machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(1): 105−119

32 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV). Kerkyra, Greece: IEEE, 1999, **2**: 1150−1157

33 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

34 Papageorgiou C P, Oren M, Poggio T. A general framework for object detection. In: Proceedings of the 6th International Conference on Computer Vision (ICCV). Bombay, India: IEEE, 1998. 555−562

35 Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision and Image Processing. Jerusalem, Israel, Palestine: IEEE, 1994, **1**: 582−585

36 Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 1996, **29**(1): 51−59

37 Yan J J, Lei Z, Yi D, Li S Z. Multi-pedestrian detection in crowded scenes: a global view. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, Rhode Island, USA: IEEE, 2012. 3124−3129

38 Yan J J, Zhang X C, Lei Z, Liao S C, Li S Z. Robust multiresolution pedestrian detection in traffic scenes. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, Oregon, USA: IEEE, 2013. 3033−3040

39 Yan J J, Zhang X C, Lei Z, Yi D, Li S Z. Structural models for face detection. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Shanghai, China: IEEE, 2013. 1−6

40 Zhu X X, Ramanan D. Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, Rhode Island, USA: IEEE, 2012. 2879−2886

41 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2011. 1385−1392

42 Yan J J, Lei Z, Wen L Y, Li S Z. The fastest deformable part model for object detection. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, Ohio, USA: IEEE, 2014. 2497−2504

43 Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY, USA: IEEE, 2006. 2169−2178

44 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, Ohio, USA: IEEE, 2014. 580−587

45 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C, Fei-Fei L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

46 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

47 Xiao J X, Hays J, Ehinger K A, Oliva A, Torralba A. Sun database: large-scale scene recognition from abbey to zoo. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, 2010. 3485−3492

48 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft COCO: common objects in context. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 740−755

49 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

50 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

51 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

52 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

53 Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2006. 153−160

54 LeCun Y, Chopra S, Hadsell R, Ranzato M, Huang F. A tutorial on energy-based learning. *Predicting Structured Data*. Cambridge, MA, USA: MIT Press, 2006.

55 Lee H, Ekanadham C, Ng A Y. Sparse deep belief net model for visual area V2. In: Proceedings of the 2007 Advances in Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada: MIT Press, 2007. 873−880

56 Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, **29**(6): 82−97

57 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: MIT Press, 2012. 1097−1105

58 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1440−1448

59 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

60 Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extracting and composing robust features with denoising Autoencoders. In: Proceedings of the 25th IEEE International Conference on Machine Learning (ICML). Helsinki, Finland: IEEE, 2008. 1096−1103

61 Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proceedings of the 21th International Conference on Artificial Neural Networks. Berlin, Heidelberg, Germany: Springer, 2011. 52−59

62 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 818−833

63 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

64 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, Massachusetts, USA: IEEE, 2015. 1−9

65 Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. arXiv: 1602.07261, 2016.

66 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv: 1502.03167, 2015.

67 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. arXiv: 1512.00567, 2015.

68 He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Proceedings of the 2014 European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 346−361

69 Bell S, Lawrence Zitnick C, Bala K, Girshick R. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2874−2883

70 Yang F, Choi W, Lin Y Q. Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2129−2137

71 Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 761−769

72 Sung K K. Learning and Example Selection for Object and Pattern Detection [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 1996.

73 Kong T, Yao A B, Chen Y R, Sun F C. HyperNet: towards accurate region proposal generation and joint object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 845−853

74 Dai J F, Li Y, He K M, Sun J. R-FCN: object detection via region-based fully convolutional networks. In: Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain: MIT Press, 2016. 379−387

75 Kim K H, Hong S, Roh B, Cheon Y, Park M. PVANET: deep but lightweight neural networks for real-time object detection. arXiv: 1608.08021, 2016.

76 Shang W L, Sohn K, Almeida D, Lee H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). New York, USA: IEEE, 2016. 2217−2225

77 Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv: 1312.6229, 2013.

78 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 779−788

79 Najibi M, Rastegari M, Davis L S. G-CNN: an iterative grid based object detector. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2369−2377

80 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S E, Fu C Y, Berg A C. SSD: single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer, 2016. 21−37

81 Redmon J, Farhadi A. YOLO9000: better, faster, stronger. arXiv: 1612.08242, 2016.

82 Pepik B, Benenson R, Ritschel T, Schiele B. What is holding back convnets for detection? In: Proceedings of the 2015 German Conference on Pattern Recognition. Cham, Germany: Springer, 2015. 517−528

83 Xiang Y, Mottaghi R, Savarese S. Beyond PASCAL: a benchmark for 3d object detection in the wild. In: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV). Steamboat Springs, Colorado, USA: IEEE, 2014. 75−82

84 Amazon Mechanical Turk [Online], available: https://www.mturk.com/, February 13, 2017

85 Wang Kun-Feng, Gou Chao, Wang Fei-Yue. Parallel vision: an ACP-based approach to intelligent vision computing. *Acta Automatica Sinica*, 2016, **42**(10): 1490−1500
(王坤峰, 苟超, 王飞跃. 平行视觉: 基于 ACP 的智能视觉计算方法. 自动化学报, 2016, **42**(10): 1490−1500)

86 Wang K F, Gou C, Zheng N N, Rehg J M, Wang F Y. Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives. *Artificial Intelligence Review* [Online], available: https://link.springer.com/article/10.1007/s10462-017-9569-z, July 18, 2017

87 Wang Fei-Yue. Parallel system methods for management and control of complex systems. *Control and Decision*, 2004, **19**(5): 485−489, 514
(王飞跃. 平行系统方法与复杂系统的管理和控制. 控制与决策, 2004, **19**(5): 485−489, 514)

88 Wang F Y. Parallel control and management for intelligent transportation systems: concepts, architectures, and applications. *IEEE Transactions on Intelligent Transportation Systems*, 2010, **11**(3): 630−638

89 Wang Fei-Yue. Parallel control: a method for data-driven and computational control. *Acta Automatica Sinica*, 2013, **39**(4): 293−302
(王飞跃. 平行控制: 数据驱动的计算控制方法. 自动化学报, 2013, **39**(4): 293−302)

90 Peng X C, Sun B C, Ali K, Saenko K. Learning deep object detectors from 3D models. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1278−1286

91 Johnson-Roberson M, Barto C, Mehta R, Sridhar S N, Rosaen K, Vasudevan R. Driving in the matrix: can virtual worlds replace human-generated annotations for real world tasks? arXiv: 1610.01983, 2016.

92 Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, **22**(10): 1345−1359

93 Taylor M E, Stone P. Transfer learning for reinforcement learning domains: a survey. *The Journal of Machine Learning Research*, 2009, **10**: 1633−1685

张 慧 中国科学院自动化研究所复杂系统管理与控制国家重点实验室博士研究生. 主要研究方向为智能交通系统, 目标视觉检测, 深度学习.
E-mail: zhanghui2015@ia.ac.cn
(**ZHANG Hui** Ph. D. candidate at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. Her research interest covers intelligent transportation systems, object vision detection, and deep learning.)

王坤峰 中国科学院自动化研究所复杂系统管理与控制国家重点实验室副研究员. 主要研究方向为智能交通系统, 智能视觉计算, 机器学习.
E-mail: kunfeng.wang@ia.ac.cn
(**WANG Kun-Feng** Associate professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interest covers intelligent transportation systems, intelligent vision computing, and machine learning.)

王飞跃 中国科学院自动化研究所复杂系统管理与控制国家重点实验室研究员. 国防科学技术大学军事计算实验与平行系统技术研究中心主任. 主要研究方向为智能系统和复杂系统的建模、分析与控制. 本文通信作者.
E-mail: feiyue.wang@ia.ac.cn
(**WANG Fei-Yue** Professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. Director of the Research Center for Computational Experiments and Parallel Systems Technology, National University of Defense Technology. His research interest covers modeling, analysis, and control of intelligent systems and complex systems. Corresponding author of this paper.)