基于深度卷积神经网络的目标检测技术的研究进展

王慧玲[1,2] 綦小龙[1,2] 武港山[2]
(伊犁师范学院电子与信息工程学院 新疆 伊宁 835000)[1]
(南京大学计算机科学与技术系 南京 210023)[2]

摘 要 目标检测是计算机视觉领域中的一个研究热点。近年来,深度学习中的卷积神经网络在目标检测任务上表现突出。文中综述了深度学习在目标检测技术中的研究进展。首先,介绍了目标检测的两种方法和常用数据集,并分析了基于深度学习的方法在目标检测任务上所具有的优势。其次,根据深度学习的目标检测方法的发展过程,介绍了该方法所使用的经典卷积神经网络模型,并分析了各网络模型的特点。然后,从获取特征的能力、检测的速度及所使用的关键技术等方面进行了分析和总结。最后,根据基于深度学习的目标检测方法中存在的困难和挑战,对未来的发展趋势做了思考和展望。

关键词 深度学习,目标检测,卷积神经网络
中图法分类号 TP183 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.002

Research Progress of Object Detection Technology Based on Convolutional Neural Network in Deep Learning

WANG Hui-ling[1,2] QI Xiao-long[1,2] WU Gang-shan[2]
(Department of Electronics and Information Engineering, Yili Normal University, Yining, Xinjiang 835000, China)[1]
(Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China)[2]

Abstract Object detection is a hot topic in the field of computer vision. In recent years, convolutional neural network in deep learning has performed prominently in object detection tasks. This paper surveyed the research progress of deep learning in object detection. Firstly, two methods and commonly datasets of object detection were introduced and the advantages of deep learning based on object detection tasks were analyzed. Secondly, according to the development process of the object detection method based on deep learning, the classical convolutional neural network model used in this method was introduced, and the characteristics of each network model were analyzed. Then the aspects of the ability to acquire features, the speed of detection, and the used key technologies were analyzed and summarized. Finally, according to the difficulties and challenges existing in the object detection method based on deep learning and the future development trend, the thinking and outlook were made.

Keywords Deep learning, Object detection, Convolution neural network

1 引言

目标检测即计算机模拟人眼在图像中检索获取感兴趣的目标对象。目标检测要完成目标类别的判断和目标所处位置的标定,这对于人来说是一项基础的视觉处理任务,但对于计算机来说却非常困难。一张图像输入到计算机后被转化成一组0～255的二进制数值,计算机要从这组数据中抽象出目标类别的高级语义信息,并确定目标所在的位置,而目标又会因视角、光照、对象间的遮挡及自遮挡、噪声等的影响呈现出不同程度的形变,这都增加了目标检测的难度。目标检测虽然存在诸多困难,但却是让计算机"睁眼看世界"处理高级视觉任务的第一步,例如智能视频监控[1]、基于内容的图像检索[2]、机器人导航[3]、增强现实[4]和场景理解[5]等。因此,目标检测在计算机视觉领域和实际应用中都具有重要意义。

近年来,目标检测取得了较大的进步。目前,已有的目标检测方法分为两类:人工设计的特征+特征分类的传统目标检测方法,基于深度学习的目标检测方法。传统的目标检测方法的重点在于设计特征提取器和分类器,但由于人工设计的特征提取方法提取的特征存在表达能力不足及泛化能力弱等缺点,从而限制了传统目标检测方法的发展。近年来,各种深度学习的方法得到了广泛的应用和探讨[6-10]。基于深度学习的目标检测方法主要是利用深度学习中的卷积神经网络来从大量数据中学习特征,学习到的特征更能刻画数据内在的丰富信息,提高了特征的表达能力,而且卷积神经网络将特征提取、特征选择和特征分类融合在同一模型中,通过端到端的训练进行全局优化,增强了特征的辨别力[11-12],使其在分类

到稿日期:2017-12-12 返修日期:2018-05-17 本文受国家自然科学基金(61663045)资助。
王慧玲(1981－),女,博士生,讲师,主要研究方向为计算机视觉、图像分析与处理,E-mail:dg1633019@smail.nju.edu.cn;綦小龙(1981－),男,博士生,讲师,主要研究方向为机器学习、模式识别;武港山(1967－),男,教授,博士生导师,主要研究方向为媒体内容分析、多媒体信息检索,E-mail:gswu@nju.edu.cn(通信作者)。

和定位的准确性上显著优于传统方法，成为处理目标检测任务的研究热点。

本文重点对基于深度学习的目标检测方法的研究工作进行了分析和综述，通过对这些研究工作的分析和比较，总结了目标检测发展的现状，并提出了一些在目标检测中的前瞻性研究方向。

2 目标检测研究的框架

目标检测任务可分为两个部分：目标分类和目标定位。目标分类负责判断输入图像中的目标的类别；目标定位负责确定输入图像中的目标所在的位置。通过输出物体的包围盒或物体中心或物体的闭合边界等确定对象的位置，最常用的是方形包围盒。如图 1 所示，我们用不同颜色的方形包围盒确定出目标所在的位置（一种颜色代表一种类别），并给出所属类别的概率和名称。

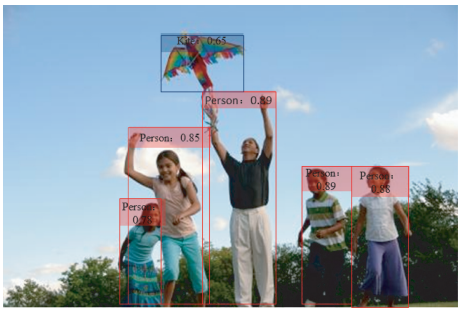

图 1 目标检测结果

Fig. 1 Results of object detection

2.1 传统的目标检测方法

传统的目标检测方法的流程如图 2 所示。首先，对输入的图像进行图像去噪、图像增强、色彩空间转换等预处理；其次，利用不同尺寸的滑动窗口在图像上滑动，将滑动窗口中的图像作为候选区域；然后，使用 Sift[13]，HOG[14]，Harr[15] 等特征提取方法提取候选区域的相关视觉特征；最后，利用分类器进行类别的判别，例如利用 AdaBoost[16]，SVM[17] 和 DPM[18] 等分类器对目标进行分类。

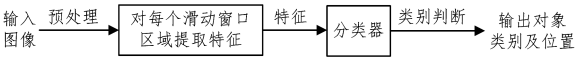

图 2 传统的目标检测框架

Fig. 2 Traditional object detection framework

传统的目标检测方法存在两方面的缺陷：1）人工设计特征的好坏是整个系统的性能瓶颈。这些人工设计的特征与检测目标对应，例如，人脸检测用 Harr 特征，使用 AdaBoost 分类器；行人检测用 HOG 特征，使用 SVM 分类器；一般性物体检测用 HOG 特征，使用 DPM 分类器。面对多样性的目标，人工设计的特征表达能力不足且鲁棒性较差，目前尚没有形成一个统一的有效算法。2）基于滑动窗口的区域选择方法产生的窗口没有针对性且窗口大量冗余，导致计算复杂，时间复杂度较高。

2.2 基于学习的目标检测方法

深度学习方法具有强大的表征和建模能力，通过监督、半监督或无监督的训练方式，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。基于深度学习的目标检测流程如图 3 所示，对输入的图像进行去均值、标准化等预处理，然后把图像输入深度学习模型，卷积神经网络从大量的输入数据中学习目标特征和目标位置的定位，最后通过 softmax 等方法判定类别。

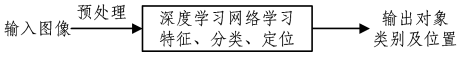

图 3 基于深度卷积神经网络的目标检测框架

Fig. 3 Object detection framework based on deep learning of convolutional neural network

基于深度学习的目标检测方法的优点是网络能从大量的数据中学习特征，而学习到的特征具有鲁棒性和较强的泛化能力，这对目标检测十分重要。

2.2.1 卷积神经网络

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卷积神经网络（Convolutional Neural Networks，CNN）[19]、深度信念网络（Deep Belief Network，DBN）[20] 和堆叠自动编码器（Stacked Auto-Encoders，SAE）[21] 等。而卷积神经网络在检测精度和速度上优于后两种方法。

卷积神经网络是由生物学家休博尔和维瑟尔在早期关于猫视觉皮层的研究中提出来的，是多层感知机（MLP）的变种。纽约大学的 LeCun 于 1998 年提出了卷积神经网络的模型 LeNet[19] 并将其应用于手写字符的检测中。Hinton 教授于 2012 年带领学生 Krizhevsky 将卷积神经网络模型 AlexNet[22] 应用于图像的分类任务中，并在 ImageNet[23] 大规模视觉识别挑战赛（ILSVRC）上获得了第一名，使 Top-5 错误率降低到了 15.3%；本次挑战赛的第二名采用的是传统的 DPM 模型，它的 Top-5 错误率高达 26.2%。ILSVRC 使得深度学习再次引起了大家的关注。目前，微软的 ResNet[24] 和谷歌的 Inception V4[25] 将 Top-5 错误率降到了 4% 以内，超越了人在这个特定任务上的表现。Girshick 于 2014 年将区域预测＋卷积神经网络的方法（R-CNN）[26] 应用到了目标检测任务中，使得在 Pascal VOC 2007 测试集上的 mAP 提升至 58.5%。

卷积神经网络（CNN）在目标检测任务中存在如下优势。

1）多卷积核。添加多个卷积核，可以学习多种特征来更好地描述图像。

2）局部感知。一般来说，人对外界的认知是从局部到全局的，而图像的空间联系也是邻近的像素区域联系较为紧密，距离较远的像素的相关性则较弱。因此，每个神经元先对局部区域进行感知然后在更高层综合局部信息，从而得到全局信息。通过局部感知还可以减少训练的参数数目。

3）参数共享。在局部连接中，每个神经元对应的滤波器的参数是相同的，这在很大程度上减少了训练参数的数目。

4）池化。使用了多核卷积后，提取的特征维数是巨大的。为了减少维数，采用最大值或平均值等方法进行降维。池化后的特征在缩放或方向变化时保持不变。

5）稀疏性限制。隐藏层的神经元数据较大，对此加入稀疏性限制，让大部分神经元处于被抑制状态，只有少数神经元被激活，从而有助于我们用少量的神经元提取出图像更加本质的特征。这大幅减少了网络的参数数量，加快了训练速度。

以上优点使得 CNN 处理图像时，能自行抽取图像特征，

包括颜色、纹理、形状及图像的拓扑结构；CNN 在处理二维图像问题，特别是在处理识别位移、缩放及其他形式的扭曲不变性的应用时，具有良好的鲁棒性和运算效率。

2.2.2　大规模的数据库

基于深度学习的目标检测方法通过卷积神经网络提取目标特征。如果卷积神经网络太浅，则其识别能力往往不如普通的 SVM 和 boosting 等浅层模型；如果卷积神经网络较深，则需要大量数据进行训练，否则在学习中就会出现过拟合等情况。1990 年到 2010 年间，互联网数据以指数级增长，为构建大规模数据集提供了数据保证。目前，在目标检测任务中常用到以下 3 个公开数据集。

1）PASCAL VOC 数据集[27]。2005 年建立的 PASCAL VOC 数据集是图像分类、识别和目标检测等任务的一个基准测试集，它提供了检测算法和学习性能的标准图像注释数据集以及标准的评估系统。该数据集包含 VOC2007（430 M）和 VOC2012（1.9G）两个子集。2005 年到 2012 年间，每年都会基于这个数据集举行关于图像识别的 PASCAL VOC 挑战赛。

2）ImageNet 数据集[23]。2009 年公布的 ImageNet 数据集是目前最大的图像分类数据集，包含了 1400 万幅图像和 2.2 万个类，平均每个类包含 1000 幅图像。其成为了图像分类、定位、目标检测等研究工作的标准数据集。2010 年到 2017 年间，每年都会举行关于图像分类、目标检测、分割等任务的 ImageNet 挑战赛（ILSVRC）。在这些挑战赛中涌现出了许多经典的深度学习的网络模型，极大地推动了计算机视觉的发展。

3）MS COCO 数据集[28]。2014 年发布的 MS COCO 数据集以场景理解为目标，收集的图像大多是从复杂的日常场景中截取的，图像中的目标通过精确的分割进行位置的标定。该数据集包括 80 类物体、32.8 万幅图像和 250 多万个标注物体，虽然比 ImageNet 包含的类型少，但是每一类物体的图像较多，是目前每幅图像平均包含目标数最多的数据集。MS COCO 不仅用于目标检测研究，还用于图像中目标之间的上下文关系和目标的精确定位。

大量的人工标注数据使得有监督训练成为可能。计算机硬件加速器（例如 GPU，FPGA 和 Intel Xeon Phi 等）也为深度卷积神经网络在大数据上的训练提供了计算保障，减少了模型的训练时间。为了能够方便、快捷地构建深度网络，研究者们设计了许多构建深度网络的架构，尤其是 TensorFlow[1)]，Caffe[2)]，Torch[3)]，MXNet[4)] 等在定义网络模型的几何空间、预构建层的可用性等功能上具有优势。

基于深度学习的目标检测方法在高性能的计算机硬件支撑下，采用方便、快捷的网络构架搭建的卷积神经网络，在大规模的带有标注的数据集上训练后获取了抽象的特征，使之在检测精度和速度上远远高于传统的目标检测方法。如图 4 所示，HOG＋DMP 方法是传统方法中精度最高的算法，而基于深度学习的目标检测算法在精度上比 HOG＋DMP 方法有了大幅的提升。例如，以 AlexNet 和 VGG 为主干网的 R-CNN 方法的精度分别提升至 58％和 66％；以 ResNet-101 为主干网络的 Faster R-CNN 方法的精度达到了 83.8％，是 HOG＋DMP 方法的 2.4 倍。

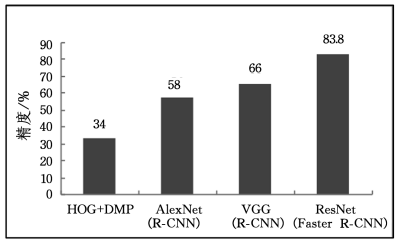

图 4　两种目标检测框架在 PASCAL VOC 2007 上的精度比较

Fig. 4　Accuracy comparison of two object detection frameworks on PASCAL VOC 2007

3　基于深度卷积神经网络的目标检测技术

在大型数据集和高性能的计算机硬件支撑下，基于深度学习的目标检测方法有了一系列突破性的进展，从而在很大程度上提升了图像处理的能力。纵观目标检测技术的研究进展，我们主要从以下两方面来介绍目标检测技术的研究现状。

3.1　CNN 的网络模型

研究者们不断对 CNN 网络模型进行改进以提高目标检测的精度，降低目标检测的时间复杂度和计算复杂度。在 ImageNet 挑战赛（ILSVRC）上的优秀成果[31-37]都证实了网络模型是极其重要的。如图 5 所示，随着网络层数的增多，提取特征的能力变强，使得模型获得了很强的拟合能力和泛化能力。计算机视觉中很多其他的任务[37-39]在很大程度上也得益于网络模型。

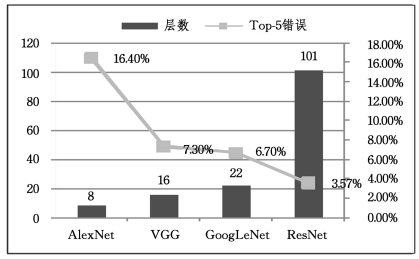

图 5　在 ImageNet 数据集上卷积神经网络结构对目标检测精度的影响

Fig. 5　Effect of structure of CNN on accuracy of object detection on ImageNet dataset

LeNet-5[19]是 LeCun 于 1998 年设计的用于手写数字识别的卷积神经网络，是早期卷积神经网络中最具代表性的网络结构之一。该模型是为了识别手写字体和计算机打印字符而设计的，曾被广泛应用于美国银行支票中手写体的识别。LeNet-5 共有 7 层（不包括输入层），通过对 LeNet-5 的网络结构进行分析，可以直观地了解一个卷积神经网络的构建方法，为分析和构建更复杂、更多层的卷积神经网络做准备。

1) https://www.tensorflow.org

2) http://caffe.berkeleyvision.org

3) http://torch.ch

4) http://mxnet.io

Krizhevsky 等所提的 AlexNet[22] 网络模型在 ImageNet 数据集上进行了大规模的对象识别，赢得了 ILSVRC 2012 的冠军，Top-5 错误率降低到 15.3%，该错误率比使用传统算法的第二名低了大约 10%，使得卷积神经网络重新激起了学术界的研究热情。AlexNet 的优势在于：网络增大（5 个卷积层＋3 个全连接层＋1 个 softmax 层），同时加入 dropout[39] 来解决过拟合问题，并且利用多 GPU 进行加速计算。

卷积神经网络的网络结构于 2014 年有了很大的改进，深度学习技术也在这一年快速成长。通过增加卷积神经网络的层数，深度网络能够找到图像中更多的高维特征，提取的特征也更抽象。但是，网络的加深使得需要存储和运算的参数也增多，研究者们在有限的硬件条件下尝试使用小的卷积核代替大的卷积核，这不仅提升了精度，而且大大减少了训练参数。Simonyan 和 Zisserman 实现的 VGGNet 结构[31] 在 ILSVRC 2014 中获得了第二名，它将 Top-5 的错误率降低到了 7.32%。VGGNet 结构的优点是选择带有小滤波器（3 * 3 和 1 * 1）的卷积层组合，而不是选用一个带有大卷积核的卷积层，这样不仅可以获得更有力的特征，而且减少了训练时间。VGG 虽然比 AlexNet 有更多的参数和更深的层次，但是只需要很少的迭代次数就可以收敛。Szegedy 和 Liu 实现的 GoogLeNet[32] 在 ILSVRC 2014 中获得了第一名，将 Top-5 的错误率降低到了 6.67%。GoogLeNet 是一个 22 层的深度网络，它不仅增加了网络的深度，而且增加了网络的宽度，最大化了网络的信息流。受 Network In Network[33] 的启发，GoogleNet 的 Bottleneck layer 减少了特征的数量，这种设计架构节省了大量的计算成本，在不大量增加计算量的前提下提高了性能。目前，GoogleNet 包括的模型有 Inception-V1，Inception-V2[34]，Inception-V3[35] 以及带有 ResNet 特性的 Inception-V4[25] 和 Inception-ResNet-V2[36]。

网络深度的加深也会带来一个致命的问题，即梯度消失（爆炸）[37-38]，它从根本上妨碍了网络的收敛。这个问题已被广泛讨论，文献[39-41]试图通过调参等办法加以解决，但又暴露了一个所谓的退化（degradation）问题，即随着网络深度的增加，准确率（accuracy）增长的速率会很快达到饱和但又会很快下降。文献[42-43]指出此类退化问题并不是由于过拟合造成的，而是网络层数的增加导致产生了越来越大的训练误差和测试误差。在 ILSVRC 2015 大赛上，何凯明提出的深度残差学习网络——ResNet[24] 通过对残差的学习缓解了深度网络的退化问题，使得超深度残差网络更便于优化，也更容易提高精度。该网络将 Top-5 的错误率降低到了 3.57%，获得了本次大赛的冠军。ResNet 在 ImageNet 数据集上验证了深达 152 层的残差网络，虽然其深度是 VGG 网络的 8 倍，但复杂度却相对较低。

表 1 经典 CNN 网络深度、模型大小及 Top-5 错误率的比较

Table 1 Comparison of depth, model size and Top-5 error rate of classical CNN networks

网络名	层数	Top-5 错误率/%	Caffe 模型大小/M
AlexNet	8	16.40	约 200
VGG	19	7.30	约 550
GoogLeNet	22	6.70	约 50
ResNet	101	3.57	约 170

从表 1 中的 Caffe 模型大小可以看出，网络的加深并不代表网络模型会变大，但却能有效提高目标检测的精度。ResNet-101 网络模型现在已成为目标检测的基础网络。

以上所述的卷积神经网络主要从两个方面来优化检测效果：1）加深网络（如 ResNet 解决了深层网络的梯度消失问题）；2）加宽网络（如 GoogleNet 的 Inception）。黄高等提出的 DenseNet[44] 则是从特征入手，网络中每一层的输入来自前面所有层的输出，而该层所学习的特征图也会被直接传给其后面的所有层作为输入，实现了特征的重复利用，同时把网络的每一层设计得特别“窄”，即只学习非常少的特征图，使得参数量和计算量显著减少，从而达到降低冗余性的目的。这种层与层之间的连接实现了信息流的整合，避免了信息在层间传递时丢失及梯度消失的问题（还抑制了某些噪声的产生）。

3.2 网络的优化

尽管大型网络具有潜在的泛化和学习能力，但是在训练数据量少或训练时间短的情况下，会出现过度拟合、收敛速度慢等问题。因此，在训练网络时需要一些策略和技巧来弥补这些缺点。

1）数据增强（Data Augmentation）。在不改变图像类别的情况下增加数据量。该方法能提高模型的泛化能力。自然图像的数据增强有很多方式，如常用的水平翻转，一定程度的位移、裁剪和颜色抖动等。

2）正则化（Regularizations）。加入正则化后不仅可以减少过拟合，而且能有效避免陷入局部最小点，更有效地重现实验结果。常用的方法是加入 L1、L2 和最大范数。除此以外，也使用 dropout[45] 改变神经网络的结构，以减少对神经元的依赖性，使神经网络在学习过程中学习到更加鲁棒的特征。

3）预处理（Pre-Processing）。输入图像需要对图像进行去均值、标准化、PCA 和白化等处理，以剔除一些无用和冗余的信息，加快检测速度。

4）初始化（Initializations）。CNN 的初始化主要是针对卷积层和输出层的卷积核的权重和偏置，以达到加快学习速度的目的。以前常用的是对卷积核和权重进行随机初始化，而对偏置进行全 0 初始化。目前常用的是 Batch Normalization[34] 方法，它强制在训练前将激活值设置为高斯分布，使得在网络训练过程中每一层的输入数据保持相同分布。

5）激活函数（Activation Functions）。是否激活神经元的函数称为激活函数，用来加快收敛速度，使得网络在训练时更稳定。常用的激活函数有 Sigmoid[46] 函数和 Tanh 函数，但它们都存在梯度饱和的问题。ReLU[47] 能使 SGD 的收敛速度加快，减少了梯度饱和，而且计算简单。Maxout[48] 的拟合能力非常强，它可以拟合任意的凸函数。

卷积神经网络的不断优化，使得网络训练过程更加稳定、快速，网络从图像中提取的特征越来越鲁棒。

4 基于深度卷积神经网络的目标检测算法

AlexNet 于 2012 年在 ILSVC2012 图像分类中取得成功后，也给目标检测任务带来了新的算法，因为目标检测任务也要对输入图像的目标进行分类。Girshick 于 2014 年使用区域预测＋CNN 代替传统目标检测方法使用的滑动窗口＋手

工设计特征，设计出了 R-CNN，其在 Pascal VOC 2007 测试集上达到了 58.5%的 mPA。之后，目标检测工作大都转移到使用深度学习的方法，并涌现出了许多研究成果，使得检测的精度和速度得到了大幅提升。

这些研究成果大致可以分为两类：1)将检测问题转化成对图片局部区域进行分类的问题，即基于候选框的目标检测；2)将检测问题看作是对整张图像的回归问题，即基于回归问题的对象检测。

4.1　基于候选框的目标检测算法

Girshick 等提出的 R-CNN 方法是基于候选框的目标检测算法的典型代表。针对传统方法中滑动窗口作为特征提取的候选区域而导致计算量大的缺点，R-CNN 算法使用 Selective Search[49] 方法对每张图像产生待分类的候选区域，CNN 在这些候选区域上提取特征。为了使定位更加准确，R-CNN 还训练了一个线性回归模型来对候选区域的坐标进行修正，该过程被称为检测框的回归。该模型在 PASCAL VOC 的目标检测数据集上取得了比传统算法高约 20%的正确率。鉴于 PASCAL VOC 数据集比 ImageNet 数据集小，R-CNN 使用 ImageNet 数据集对其中的卷积神经网络进行预训练，再在 PASCAL VOC 数据集上对模型进行微调，最终取得了较好的训练效果。这种微调方法也成为了训练中常用的预处理手段。

R-CNN 模型也存在以下缺点：1)输入的图像需要裁剪或拉伸到 227 * 227 大小，造成了图像的失真和信息的损失；2)Selective Search 方法产生大约 2000 个候选区域，仍存在大量冗余，而且一张图像需要进行多次卷积操作，计算量非常大；3)训练需要分为候选框的提取、CNN 特征提取、SVM 分类、边框回归多个阶段，步骤繁琐、训练耗时；4)卷积得到的特征数据还需要单独存储，占用大量的磁盘空间。

He 等提出的 SPP-Net[50] 算法将空间金字塔的思想加入到 R-CNN 中，实现了图像的多尺度输入。SPP-Net 在最后一个卷积层和全连接层之间加入了一个空间金字塔池化层(SPPlayer)(空间金字塔池化层的池化程度随输入特征的大小而调整)，最终得到一个固定长度的特征表示。这样做，避免了对图像的裁剪和拉伸而造成的失真和数据损失。相对 R-CNN 需要对大量的重叠候选区域多次提取特征的缺陷，SPP-Net 只需对原图进行一次卷积，得到整张图的特征图，然后将候选框在原图的位置映射到特征图上，弥补了对重叠候选区域多次提取特征的缺陷，节省了大量的计算时间。相比 R-CNN，R-CNN 大约有 100 倍的提速，这为后来的基于 CNN 的目标检测方法在速度提升方面提供了依据。

该算法存在如下缺点：1)依然存在提取候选区域、提取图像特征、根据特征进行分类和边框回归多个阶段；2)只能对全连接层进行微调；3)需要额外的空间来存储提取出来的特征，以供分类器使用。

Girshick 提出的 Fast R-CNN[51] 融合了 R-CNN 和 SPP-NET 的优点，同时使用了多任务损失函数(multi-task loss)将边框回归加入到了 CNN 网络中，与 region 分类合并成为了一个多任务模型，使得这两个任务能够共享卷积特征，弥补了 R-CNN 中分多个阶段训练和存储开销大的缺点。其在最后一个卷积层后加了一个 ROI 池化层，它把不同大小的候选区域所对应的特征划分为大小相同的块，比 SPP-Net 中的空间金字塔池化层更简洁。

该算法的缺点在于：提取候选框仍是目标检测过程中耗时最多的步骤，无法满足实时应用。

针对 Fast R-CNN 检测速度的问题，Ren 等将提取候选区域的任务也交给 CNN，从而提出了 Faster R-CNN[52]。Faster R-CNN 设计了区域预测网(Region Proposal Network，RPN)，并用其代替了费时的 Selective Search 方法。RPN 使用 9 个不同长宽比的 anchor boxes 在最后一个卷积层输出的特征图上滑动，且后接边框回归，使得这 9 种 anchor boxes 得到的候选框能得到一个与目标较为接近的候选区域；然后将 RPN 的输出再送入 Fast R-CNN 做更精细的分类和检测框的位置修正。RPN 在提取预测框时不仅没有时间成本，而且还提高了预测的精度。使用 VGG Net 作为主干网络后，在 PASCAL VOC 2007 上达到了 73%的 mAP。

该算法的缺点在于：以 ROI 池化层为界，前面的子网络共享计算，用于特征提取；后面的子网络用于目标检测，该子网络必须对每个候选区域单独处理，训练的超参数数量大，导致检测速度达不到实时处理的要求。

Faster R-CNN 算法的产生至今已有两年，但它仍是目标检测领域的主流算法之一。之后也出现了很多改进算法，例如将特征提取网络更换为 ResNet，PVAVet[53] 和 HpyerNet[54] 等基础网络，虽然该改进使得精度都有所提升，但存在内存占用量大、训练困难等问题。在 Faster R-CNN 的改进算法中，R-FCN[55] 和 FPN[56] 算法在速度和精度上的提升较为显著。

Dai 等设计了基于区域的全卷积网络 R-FCN，提出了一种新的卷积层——位置敏感的打分图(Position Sensitive Score Map)，它对 Conv5 层的特征图执行全卷积操作，然后对该特征图进行池化得到对位置敏感的打分图。打分图保证了位置信息的不变性。该方法把前面所有的卷积层都放在了共享的子网络中，只用最后一层卷积来预测，大大减少了计算量，比 Faster R-CNN 快 2.5～20 倍。

之前的目标检测算法只使用具有丰富语义信息的高层特征做预测，但高层特征图的分辨率低，存在目标定位不精确和小目标对象丢失的问题。低层特征图的分辨率高，可以获得更多细粒度的信息，目标定位准确。最近，Lin 提出了多尺度的目标检测算法 FPN(Feature Pyramid Networks)，它在获取特征时分为两个过程：随着卷积层的加深，对图像进行上采样以得到不同分辨率的特征图；在获取了最高层特征后，对特征图进行下采样。下采样是把高层的特征图与下一层特征图做元素相加的横向操作，把高层的语义信息传递给低层特征图。最终获得的融合特征图既有低层的分辨率又有高层的强语义信息，使得每一层都可以独立地对对应分辨率大小的物体做检测。此方法只是在原网络基础上增加了额外的跨层连接，在实际应用中几乎不增加额外的时间和计算量。该算法不仅可以每层独立做预测，而且对小目标的检测有显著的提升。

基于候选框的目标检测算法是将目标检测的问题转化到分类上，采用了预测候选框＋CNN 网络提取特征＋分类和定

位的思路。其主要从以下两方面来提升检测的效率:1)使用卷积神经网络获取分类特征,随着网络深度的增加,特征的表征能力增强;2)不断探索获取候选框的方法,试图找到一种能减少处理区域预测框的复杂度、数量更少和召回率更高的区域预测框方法。基于候选框的目标检测算法不但能加快目标检测的速度,还能提高目标检测的性能(假阳例少)。如表 2 所列,SPP-Net 的金字塔池化层和 Faster R-CNN 的 RPN 对减少计算的复杂度和提高准确性做出了贡献。但是,这类算法有一个很大的缺陷,即无法达到实时性的检测。

表 2 基于候选框的目标检测方法的性能比较

Table 2 Performance comparison of object detection methods based on proposals

方法	mAP(VOC 2007)/%	采用的 CNN 网络
R-CNN	58.8	AexNet
SPP-Net	60.9	ZF-5
Fast R-CNN	70.0	VGG16
Faster R-CNN	76.3	VGG16
R-FCN	78.8	Resnet-101
FPN	79.5(COCO AP@0.5)	ResNet-101

4.2 基于回归位置的目标检测算法

制约基于候选框的目标检测算法的速度提升的关键是将目标检测问题转化成了对图像局部区域的分类问题,不能充分利用局部对象在整个图片中的上下文信息。对此,研究者们又提出了基于回归位置的目标检测算法。如图 6 所示,基于回归位置的目标检测算法只使用一个卷积网络便完成了在整个图像上对检测框的回归和类别概率的预测,网络结构简单,实现了端到端的优化,且使检测速度有了很大的提升,能达到对视频的实时处理。

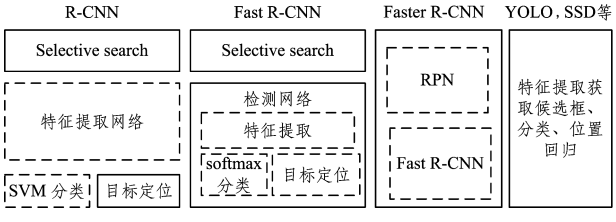

图 6 基于候选框的目标检测算法与基于回归位置的目标检测算法的结构比较

Fig. 6 Comparison of object detection algorithms based on proposal and based on regression location

Redmon 等提出的 YOLO[57] 采用了单网络结构,训练和检测均是在一个单独的网络中进行。YOLO 对候选框的处理很简单,将输入图片分为 7 * 7 的网格,每个网格有 2 个预测框,总共只有 98 个候选框。这种空间约束避免了在一些区域重复提取候选框,预测的目标窗口使用的是全图信息,使得负样本比例大幅降低,在候选框的处理上节省了大量的时间。YOLO 将物体检测作为一个回归问题进行求解,输入图像经过一次处理,便能得到图像中所有对象的位置和其所属类别及相应的置信度。最快的速度可以达到 154bps。

YOLO 存在如下缺点:1)空间约束的同时也限制了模型在邻近物体上的预测,导致密集型目标的检测会出现漏检;2)只使用了最高层的特征进行预测,使得小物体不能被检测到;3)YOLO 在目标定位时的准确度比 FastR-CNN 降低了 10%,导致它的精度不高。

YOLO 在检测速度方面提升较大,使得目标检测任务可以达到实时检测。紧随其后,Redmon 等和 Liu 等针对 YOLO 在召回和定位方面存在的问题分别提出了 YOLOv2[58] 和 SSD[59]。考虑到效率问题,YOLOv2 设计了一个新的分类网络(Darknet-19)作为 YOLOv2 的基础网络,尝试在将网络简单化的同时提高准确率。为了保证特征图的分辨率,移除了一个池化层,同时移除全连接层以应对各种分辨率的输入,通过使用不同分辨率的图像进行训练,实现了多尺寸的检测。YOLOv2 引入了 anchor box,将 YOLO 中回归检测框的位置简化为学习 anchor box 的相对偏移量;添加了 passthrough 层以获得细粒度特征;YOLO9000 采用 WordTree 综合了 ImageNet 数据集和 COCO 数据集进行训练,使之可以实时地检测 9000 多类对象。通过一系列的改进,相比 YOLO,YOLOv2 更快、更准、更强壮、检测的类别更多,检测速度达到了 45fps,同时在 VOC2007 上的 mAP 还能达到 78.6%。

SSD 旨在高精度地快速检测对象。与 YOLO 不同,该算法不使用整个特征图的全局信息来回归所有位置的检测框,而是使用了大量的小卷积核(1×1,3×3)获得的局部特征来回归各个位置的检测框,提高了定位的准确性。SSD 抽取了 5 个不同层的特征图来检测不同大小的物体,在不同的特征图上对 anchor box 的纵横比做了调整,以适应特征图的分辨率。其达到了 72.1%的 mAP 和 58fps 的速度。

SSD 的缺陷在于:虽然引入了低层的特征图,但没有引入高层的语义信息,对小物体的检测效率仍不够理想。

DSSD[60] 沿用了 SSD 利用多特征融合进行目标检测的思路,首先把基准网络从 VGG 换成了 ResNet-101,增强了特征提取能力。为了让低层抽取的特征图具有高层较强的语义信息,DSSD 添加了反卷积层(deconvolution layer),反卷积的特征具有高层的全局语义信息,将反卷积获得的特征与相同大小的卷积层特征图相融合,融合后的特征图不仅保存了原有的分辨率,而且具有高层的语义信息;添加了预测模块来提高每一个子网的准确率。最终,DSSD 提升了目标检测精度,尤其是在小物体的检测上表现突出。

从表 3 可以看出,基于回归位置的目标检测方法使用单网络完成了检测框位置的回归和目标的分类,与基于候选框的目标检测方法相比,在检测速度上提升了 2~10 倍,且通过引入 anchor box 使得精度也不断被提升。

表 3 基于回归问题的目标检测方法的性能比较

Table 3 Performance comparison of object detection methods based on regression

方法	mAP(VOC 2007+2012)%	主干网络	FPS
YOLO	63.4	GoogLeNet	45
SSD300	74.3	VGG16	46
YOLOv2 352	73.7	Darknet-19	81
DSSD321	76.3	Resnet-101	9.5

这些目标检测算法的设计相互之间有一定的联系,同时也有各自的特点。首先,基于深度学习的方法使目标检测任务的精度从 2013 年的 58.8%提升到了 2017 年的 79%(在 ImageNet 数据集上),这是一个突破性的提升;其次,产生检

测窗口的方式发生了很大的变化，从传统方法中逐一遍历整张图像的滑动窗口，转变成为寻找最有可能出现目标的窗口，最后只对少量的窗口进行类别的判定，从而使得检测速度有了飞跃；最后，检测框位置的回归不仅有利于目标的定位，还有助于得到更为准确的检测框。

5　思考与展望

随着计算机视觉领域中各项技术的不断成熟及应用范围的不断扩大，目标检测任务也变得越来越重要，并日益得到了学术界和工业界的广泛关注。虽然研究人员针对该问题取得了大量的研究成果，但我们认为该领域还存在着许多值得进一步关注的研究问题。

1）目标检测的鲁棒性主要受到类间差异和类内差异的影响，大的类内差异和小的类间差异通常会降低目标检测方法的鲁棒性。类内差异是指同类不同个体间的变化，例如，不同的椅子个体在大小、形状、颜色、纹理、姿态等方面存在着很大的差异。即使是同一把椅子，在光照、背景、姿态、视点的变化和遮挡的影响下，其视觉也会非常不同。Wang 等[61]将对抗学习的思路应用于图像识别问题中，通过对抗网络生成遮挡和变形图片样本来训练检测网络，并取得了不错的效果。构建具备泛化能力的特征模型仍然是一个开放性的研究问题。

2）对现有的经典网络模型和算法进行优化。使用了深度学习后，目标检测的检测精度和速度都得到了提高，但仍需进一步的优化。文献[56，62]都对网络结构进行了优化；文献[63]提出的 Facol loss 解决了样本不均衡和难分样本的处理问题；文献[64]对非极大值抑制环节进行了优化，从而可以检测出两个相邻的相同物体。

3）目前，对稀疏的目标检测方法的研究较多，而稠密的多个目标检测会因目标间的相互遮挡、拍摄的角度、图像的分辨率等受到很大的影响，如漏检或把多个目标当作一个对象。文献[65]利用图像特征归约得到目标密度，以此确定区域中的目标数量，再结合稀疏目标的检测方法。这种方法对检测精度的提升依赖于目标密度的准确性。文献[66]将传统的方法与深度学习相结合，发现这两类不同的方法间存在一定的互补性，对提升稠密的多目标检测具有一定的效果。因此，如何有效融合这两类不同的方法以进一步提高稠密的多目标检测精度，具有一定的研究意义。

4）将已有的目标检测方法相结合，进行多任务的学习，如将它们与语义分割技术相结合[67-68]，利用分割来去除检测区域内的背景，从而提高对象检测的性能。MaskR-CNN[69]在基础特征网络之后又加入了全连接的分割子网，在有效检测图像中的目标的同时，为每个实例生成一个高质量的分割，完成了目标检测和对象分割两个任务。文献[70]在基础特征网络中添加了对姿态估计的损失，在目标检测的同时完成了粗略的姿态估计。如何把用于目标检测的特征共享给其他任务，设计出新颖、高效的多任务集成学习是一个重要的研究方向。

5）基于视频的目标检测可以看作是单帧图像的目标检测。基于视频的目标检测可以利用时间上下文信息消除帧率较高时的信息冗余，同样也可以利用时间上下文信息来补充单帧图像上信息的不足，从而实现更准、更快的检测。基于视频的目标检测在自动驾驶、智能视频监控等领域都十分重要，因此，不管是从实用性，还是从学术研究的角度来说，它都具有重要的研究价值。

6）基于无监督和动态数据的持续学习。目前讨论的许多内容都是有监督的学习，使得基于深度学习的目标检测方法对数据集有很强的依赖性。今后，随着目标检测方法在监督的学习达到一定程度后，深度学习在无监督的学习方面很可能是未来的发展趋势。此外，目标检测、对象的识别等都是在收集好的静态数据上训练、在线预测完成的。这些静态数据如果没有及时更新，就会导致定期学习无法适应持续动态变化的环境，如自动驾驶在遇到突发情况时要做出及时的反应，这就需要检测任务能够持续地学习和适应异步的变化。

结束语　目标检测任务是计算机视觉领域中的基础研究问题之一，具有重要的研究意义和应用价值。把深度学习的方法应用于目标检测任务后，目标检测的准确率和检测速度都有了大幅的提升。本文首先介绍了目标检测的两个框架，并分析了基于深度学习的方法在目标检测任务上具有的优势，然后重点介绍了基于深度学习的目标检测方法的最新研究进展，最后分析了基于深度学习的目标检测方法中的困难和挑战，并对未来的发展趋势做了进一步的思考和展望。

参 考 文 献

[1] AGGARWAL J K，RYOO M S. Human activity analysis：A review[J]. ACM Computing Surveys (CSUR)，2011，43(3)：16.

[2] DATTA R，JOSHI D，LI J，et al. Image Retrieval：Ideas，Influences，and Trends of The New Age[J]. ACM Computing Surveys (CSUR)，2008，40(2)：5.

[3] KRÜGER V，KRAGIC D，UDE A，et al. The Meaning of Action：a Review on Action Recognition and Mapping[J]. Advanced Robotics，2007，21(13)：1473-1501.

[4] PALMESE M，TRUCCO A. From 3-D Sonar Images to Augmented Reality Models for Objects Buried on The Seafloor[J]. IEEE Transactions on Instrumentation and Measurement，2008，57(4)：820-828.

[5] LI L J，SOCHER R，LI F F. Towards Total Scene Understanding：Classification，Annotation and Segmentation in an Automatic Framework[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2010：49-56.

[6] BENGIO Y. Learning Deep Architectures for AI[J]. Foundations and Trends@ in Machine Learning，2009，2(1)：1-127.

[7] DENG L. A Tutorial Survey of Architectures，Algorithms，and Applications for Deep Learning[J]. APSIPA Transactions on Signal and Information Processing，2014，3(1)：1-29.

[8] SCHMIDHUBER J. Deep Learning in Neural Networks：An Overview[J]. Neural networks，2015，61(1)：85-117.

[9] BENGIO Y. Deep Iearning of Representations：Looking forward [C]// Proceedings of International Conference on Statistical Language and Speech Processing. Heidelberg：SpringerPress，2013：1-37.

[10] BENGIO Y，COURVILLE A，VINCENT P. Representation

learning:A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798-1828.

[11] LECUN Y. Learning Invariant Feature Hierarchies[C]// Proceedings of European Conference on Computer Vision. Heidelberg:Springer,2012:496-505.

[12] MOHAMED A,DAHL G,HINTON G. Deep Belief Networks for Phone Recognition[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge:MIT Press,2009:39-48.

[13] LOWE D G. Object Recognition FromLocal Scale-Invariant Features[C]// Proceedings of IEEE International Conference on Computer Vision. New York:IEEE Press,1999:1150-1157.

[14] DALAL N,TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press,2005:886-893.

[15] HARRIS C,STEPHENS M. A Combined Corner and Edge Detector[C]// Proceedings of AlveyVision Conference. Manchester:Springer,1988:147-151.

[16] COLLINS M,SCHAPIRE R E,SINGER Y. Logistic Regression,AdaBoost and BregmanDistances[J]. Machine Learning, Springer,2002,48(1-3):253-285.

[17] JOACHIMS T. Making large-scale SVM learning practical: Technical Report,SFB 475[R]. Komplexitätsreduktion in Multivariaten Datenstrukturen,Universität Dortmund,1998.

[18] FELZENSZWALB P F,GIRSHICK R B,MCALLESTER D, et al. Object Detection with Discriminatively Trained Part-Based Models[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2010,32(9):1627.

[19] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[20] HINTON G E,SALAKHUTDINOV R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786):504-507.

[21] BENGIO Y,LAMBLIN P,POPOVICI D,et al. Greedy layer-wise training of deep networks[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge:MIT Press,2006:153-160.

[22] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.

[23] DENG J,DONG W,SOCHER R,et al. Imagenet:A Large-Scale Hierarchical Image Database[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press,2009:248-255.

[24] HE K,ZHANG X,REN S,et al. Deep Residual Learning for Image Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press, 2016:770-778.

[25] SZEGEDY C,IOFFE S,VANHOUCKE V,et al. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning[C]// Proceedings of AAAI Conference on Artificial Intelligence. Menlo Park,CA :AAAI Press,2017:4-12.

[26] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2014: 580-587.

[27] EVERINGHAM M,VAN GOOL L,WILLIAMS C K I,et al. The Pascal Visual Object Classes (voc) Challenge[J]. International Journal of Computer Vision,2010,88(2):303-338.

[28] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft COCO: common objects in context[C]// Proceedings of European Conference on Computer Vision. New York: Springer, 2014: 740-755.

[29] SERMANET P,EIGEN D,ZHANG X,et al. Overfeat:Integrated Recognition, Localization and Detection Using Convolutional Networks[C]// International Conference on Learning Representations. New York:IEEE Press,2014:368-384.

[30] ZEILER M D,FERGUS R. Visualizing and Understanding Convolutional Networks[C]// Proceedings of European Conference on Computer Vision. New York:Springer,2014:818-833.

[31] SIMONYAN K,ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]// International Conference on Learning Representations. New York: IEEE Press,2015:1264-1278.

[32] SZEGEDY C,LIU W,JIA Y,et al. Going Deeper With Convolutions[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2015:1-9.

[33] LIN M,CHEN Q,YAN S. Network in network[C]// International Conference on Learning Representations. New York: IEEE Press,2014:1567-1577.

[34] IOFFE S,SZEGEDY C. Batch normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]// Proceedings of International Conference on Machine Learning. Heidelberg:Springer Press,2015:448-456.

[35] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the Inception Architecture for Computer Vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2016:2818-2826.

[36] XIE S,GIRSHICK R,DOLLÁR P,et al. Aggregated residual transformations for deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2017:5987-5995.

[37] BENGIO Y,SIMARD P,FRASCONI P. Learning Long-Term Dependencies with Gradient Descent is Difficult[J]. IEEE transactions on neural networks,1994,5(2):157-166.

[38] GLOROT X,BENGIO Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. New York:IEEE Press,2010:249-256.

[39] LECUN Y,BOTTOU L,ORR G B,et al. Efficient backprop[M]// Neural Networks: Tricks of the Trade. Berlin: Springer Berlin Heidelberg,1998:9-50.

[40] SAXE A M，MCCLELLAND J L，GANGULI S，et al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks[C]// ICLR. 2014：1-22.

[41] BA J，FREY B. Adaptive Dropout for Training Deep Neural Networks[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge：MIT Press，2013：3084-3092.

[42] HE K，SUN J. Convolutional Neural Networks at Constrained Time Cost[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2015：5353-5360.

[43] SRIVASTAVA R K，GREFF K，SCHMIDHUBER J. Highway Networks [C]// International Conference on Learning Representations. New York：IEEE Press，2015：567-573.

[44] HUANG G，LIU Z，WEINBERGER K Q，et al. Densely Connected Convolutional Networks [J/OL]. https：//arXiv. org/abs/1608. 06993.

[45] BALDI P，SADOWSKI P J. Understanding Dropout[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2013：2814-2822.

[46] YIN X，GOUDRIAAN J，LANTINGA E A，et al. A flexible Sigmoid Function of Determinate Growth[J]. Annals of Botany，2003，91(3)：753-753.

[47] XU B，WANG N，CHEN T，et al. Empirical Evaluation of Rectified Activations in Convolutional Network[J/OL]. https：//arxiv. org/abs/1505. 00853 ，2015-3-5/2015-11-27.

[48] GOODFELLOW I J，WARDEFARLEY D，MIRZA M，et al. Maxout Network[C]// ICML 2013. 2013：1319-1327.

[49] UIJLINGS J R R，SANDE K E A V D，GEVERS T，et al. Selective Search for Object Recognition[J]. International Journal of Computer Vision，2013，104(2)：154-171.

[50] HE K，ZHANG X，REN S，et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[C]// Proceedings of European Conference on Computer Vision. Heidelberg：Springer Press，2016：21-37.

[51] GIRSHICK R. Fast R-CNN [C]// Proceedings of IEEE International Conference on Computer Vision. New York：IEEE Press，2015：1440-1448.

[52] REN S，HE K，GIRSHICK R. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[C]// Proceedings of International Conference on Neural Information Processing Systems. MIT Press，2015：91-99.

[53] KIM K H，HONG S，ROH B，et al. Pvanet：Deep but lightweight neural networks for real-time object detection[J/OL]. https：//arxiv. org/abs/1608. 08021，2016-8-29/2016-9-30.

[54] KONG T，YAO A，CHEN Y，et al. Hypernet：Towards accurate region proposal generation and joint object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. New York：IEEE Press，2016：845-853.

[55] DAI J，LI Y，HE K，et al. R-fcn：Object Detection Via Region-Based Fully Convolutional Networks[C]// Proceedings of the International Conference on Neural Information Processing Systems. Cambridge：MIT Press，2016：379-387.

[56] LIN T Y，DOLLÁR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：936-944.

[57] REDMON J，DIVVALA S，GIRSHICK R，et al. You Only Look Once：Unified，Real-Time Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2016：779-788.

[58] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：101-110.

[59] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single shot multibox detector[C]// European conference on computer vision. Cham ，Springer，2016：21-37.

[60] FU C Y，LIU W，RANGA A，et al. DSSD：Deconvolutional Single Shot Detector[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：2301-2312.

[61] WANG X，SHRIVASTAVA A，GUPTA A. A-Fast-RCNN：Hard Positive Generation via Adversary for Object Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：2606-2615.

[62] HE X，ZHANG C，ZHANG L，et al. A-Optimal Projection for Image Representation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2016，38(5)：1009-1015.

[63] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal loss for dense object detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. NewYork：IEEE Press，2017：2999-3007.

[64] BODLA N，SINGH B，CHELLAPPA R，et al. Improving Object Detection With One Line of Code[J/OL]. https：//arxiv. org/abs/1704. 04503.

[65] RODRIGUEZ M，LAPTEV I，SIVIC J，et al. Density-Aware Person Detection and Tracking in Crowds[C]// Proceedings of International Conference on Computer Vision. New York：IEEE Press，2011：2423-2430.

[66] TANG S，ANDRILUKA M，SCHIELE B. Detection and Tracking of Occluded People[J]. International Journal of Computer Vision，2014，110(1)：58-69.

[67] REN J，CHEN X，LIU J，et al. Accurate Single Stage Detector Using Recurrent Rolling Convolution[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：5420-5428.

[68] SHRIVASTAVA A，SUKTHANKAR R，MALIK J，et al. Beyond Skip Connections：Top-Down Modulation for Object Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：5421-5431.

[69] HE K，GKIOXARI G，DOLLÁR P，et al. Mask R-CNN[C]// Proceedings of International Conference on Computer Vision. New York：IEEE Press，2017：2980-2988.

[70] POIRSON P，AMMIRATO P，FU C Y，et al. Fast Single Shot Detection and Pose Estimation. Fast Single Shot Detection and Pose Estimation[C]// Proceedings of 3D Vision (3DV). New York：IEEE，Press，2016：676-684.