https://doi.org/10.1007/s11263-019-01247-4



**通用目标检测的深度学习：综述**

# 摘要

目标检测是计算机视觉中最基本和最具挑战性的问题之一，它试图从自然图像中的大量预定义类别中定位目标实例。深度学习技术已经成为一种直接从数据中学习特征表示的强大策略，并在通用目标检测领域取得了显著突破。鉴于这一快速发展的时期，本文的目标是对深度学习技术在这一领域的最新成就进行全面的调查。本次调查包括300多项研究贡献，涵盖了一般对象检测的许多方面：检测框架、对象特征表示、对象建议生成、上下文建模、培训策略和评估指标。我们通过确定未来研究的前景来完成调查。

**关键词**目标检测·深度学习·卷积神经网络·目标识别

# 1导言

作为计算机视觉中一个长期存在的、基本的和具有挑战性的问题，目标检测（如图1所示）几十年来一直是一个活跃的研究领域（Fis）-

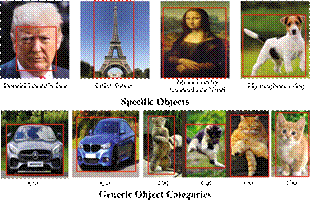
chler和Elschlager，1973年）。目标检测的目标是确定图像中是否存在来自给定类别（如人类、汽车、自行车、狗或猫）的任何对象实例，如果存在，则返回每个对象实例的空间位置和范围（例如，通过边界框Everingham et al.2010；Russakovsky et al.2015）。作为Cornerston图像理解和计算机视觉，目标检测构成了解决复杂或高级视觉任务（如分割、场景理解、目标跟踪、图像字幕、事件检测和活动识别）的基础。目标检测支持广泛的应用，包括机器人视觉、消费电子、安全、自动驾驶、人机交互、基于内容的图像检索、智能视频监控和增强现实。

最近，深度学习技术（Hinton和Salakhutdinov，2006；LeCun等人，2015）已经成为从数据中自动学习特征表示的强大方法。具体而言，这些技术在对象检测方面提供了重大改进，如图3所示。

如图2所示，对象检测可分为两种类型之一（Grauman和Leibe 2011；Zhang等人2013）：检测特定实例与检测广泛类别。第一种类型旨在检测特定对象的实例（如唐纳德·特朗普的脸、埃菲尔铁塔或邻居的狗），本质上是一个匹配问题。



**图1**2016年至2018年ICCV和CVPR会议论文中最常见的关键词。每个单词的大小与该关键字的频率成正比。我们可以看到，目标检测近年来受到了广泛的关注



**图2**对象检测包括定位特定对象的实例（顶部），以及概括检测一般对象类别（底部）。本综述主要关注后一类目标检测问题的最新进展

第二种类型的目标是检测某些预定义对象类别（例如人类、汽车、自行车和狗）的实例（通常是以前看不到的）。历史上，目标检测领域的大部分工作都集中在单个类别（通常是人脸和行人）或几个特定类别的检测上。相比之下，在过去几年中，研究界已开始朝着更具挑战性的目标迈进，即建立通用目标检测系统，在该系统中，目标检测能力的广度可与人类相媲美。

Krizhevsky等人（2012a）提出了一种称为AlexNet的深度卷积神经网络（DCNN），该网络在大规模视觉识别挑战赛（ILSVRC）中实现了破纪录的图像分类精度（Russakovsky等人，2015）。从那时起，计算机视觉的大部分领域的研究重点都是具体的深度学习方法，实际上包括通用目标检测领域（Girshick等人2014；He等人2014；Girshick 2015；Sermanet al.2014；Ren等人2017）。虽然已经取得了巨大的进步，如图3所示，但我们不知道这方面的全面调查

|  |
| --- |
|  |
|  |  |

**图3**近期目标检测性能概述：我们可以观察到，自2012年进行深入学习以来，性能（以平均精度衡量）有了显著提高。2007-2012年VOC2007竞赛获奖作品的检测结果，以及2013-2017年ILSVRCB顶级目标检测竞赛结果（两个面板中的结果仅使用提供的培训数据）

在过去的5年中，鉴于进展速度异常迅速，本文试图跟踪最新进展并总结其成就，以便更清楚地了解当前通用目标检测的全景图。

## 1.1与以往审查的比较

如表1所示，已经发布了许多著名的物体检测调查。这些调查包括许多关于特定物体检测问题的优秀调查，如行人检测（Enzweiler和Gavrila，2009年；Geronimo等人，2010年；Dollar等人，2012年）、人脸检测（Yang等人，2002年；Zafeiriouetal，2015年）、车辆检测（Suntetal，2006年）和文本检测（Ye和Doermann 2015）。除了Zhang等人（2013年）就对象类检测主题进行的调查外，最近直接关注通用对象检测问题的调查相对较少。然而，Grauman和Leibe（2011年）、Andreopoulos和Tsotsos（2013年）对该研究进行了审查而Zhang等人（2013年）大多在2012年之前，因此在深度学习和相关方法最近取得显著成功并占据主导地位之前。

深度学习允许计算模型学习极其复杂、微妙和抽象的表示，推动了广泛问题的重大进展，如视觉识别、目标检测、语音识别、自然语言处理、医学图像分析、药物发现和基因组学works，DCNNs（LeCun et al.1998，2015；Krizhevsky et al.2012a）在图像、视频、语音和音频处理方面取得了突破。可以肯定的是，已经发表了许多关于深度学习的调查，包括Bengioteal.（2013）、LeCunetal.（2015）、Litjensetal.（2017）、Gu et al.（2018），最近在ICCV和CVPR的教程中。

相比之下，尽管已经提出了许多基于深度学习的目标检测方法，但我们并不知道

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 不 | 调查标题 | 工具书类 | 年 | 地点 | 内容 |
| 1 | 单目行人检测：  调查与实验 | 恩兹韦勒和加夫里拉（2009年） | 2009 | 帕米 | 三种行人检测器的评价 |
| 2 | 高级驾驶员辅助系统行人检测综述 | Geronimo等人（2010年） | 2010 | 帕米 | 高级驾驶员辅助系统行人检测综述 |
| 3 | 行人检测：最新技术的评价 | Dollar等人（2012年） | 2012 | 帕米 | 单目图像中探测器的全面详细评估 |
| 4 | 图像中的人脸检测：综述 | 杨等人（2002年） | 2002 | 帕米 | 基于单幅图像的人脸检测研究综述 |
| 5 | 野外人脸检测研究现状与展望 | Zafeiriou等人（2015年） | 2015 | CVIU | 2000年以来野外人脸检测研究概况 |
| 6 | 道路车辆检测综述 | Sun等人（2006年） | 2006 | 帕米 | 基于视觉的道路车辆检测系统综述 |
| 7 | 图像中的文本检测与识别：综述 | 叶和多尔曼（2015） | 2015 | 帕米 | 彩色图像中的文本检测与识别综述 |
| 8 | 面向类别级目标识别 | 庞塞等人（2007年） | 2007 | 书 | 关于对象分类、检测和分割的代表性论文 |
| 9 | 对象分类的演变与图像抽象的挑战 | Dickinson等人（2009年） | 2009 | 书 | 四十年来物体分类的发展轨迹 |
| 10 | 基于上下文的对象分类：一个重要综述 | Galleguillos和Belongie（2010年） | 2010 | CVIU | 面向对象分类的上下文信息综述 |
| 11 | 物体识别50年：  前进方向 | Andreopoulos和Tsotsos（2013年） | 2013 | CVIU | 50年来物体识别系统的发展回顾 |
| 12 | 视觉目标识别 | Grauman和Leibe（2011年） | 2011 | 辅导的 | 实例和类别对象识别技术 |
| 13 | 对象类检测：综述 | 张等（2013） | 2013 | ACM CS | 2011年前通用目标检测方法综述 |
| 14 | 基于统计学习的目标检测特征表示方法综述 | 李等人（2015b） | 2015 | 公共关系 | 基于统计学习的目标检测中的特征表示方法，包括手工制作和基于深度学习的特征 |
| 15 | 显著目标检测综述 | Borji等人（2014年） | 2014 | arXiv | 显著目标检测综述 |
| 16 | 表征学习：回顾与新视角 | Bengio等人（2013年） | 2013 | 帕米 | 无监督特征学习和深度学习、概率模型、自动编码器、流形学习和深度网络 |
| 17 | 深度学习 | LeCun等人（2015年） | 2015 | 本性 | 深度学习与应用导论 |
| 18 | 医学图像分析深度学习研究综述 | Litjens等人（2017年） | 2017 | 米娅 | 医学图像分析中图像分类、目标检测、分割和配准的深度学习研究综述 |
| 19 | 卷积神经网络的研究进展 | 顾等（2018） | 2017 | 公共关系 | CNN及其在计算机视觉、语音和自然语言处理中的应用的最新进展综述 |
| 20 | 教程：用于有效对象检测的工具 | − | 2015 | ICCV15 | 目标检测短期课程，仅涵盖最近的里程碑 |

**表1**2000年以来相关物体探测调查摘要

**表1**继续的

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 不 | 调查标题 | 工具书类 | 年 | 地点 | 内容 |
| 21 | 教程：对象和场景的深入学习 | − | 2017 | CVPR17 | 对最近关于物体和场景视觉识别的深度学习工作的高水平总结 |
| 22 | 教程：实例级识别 | − | 2017 | ICCV17 | 简要介绍实例级识别的最新进展，包括对象检测、实例分割和人体姿势预测 |
| 23 | 教程：视觉识别和超越 | − | 2018 | CVPR18 | 关于图像分类、对象检测、实例分割和语义分割背后的方法和原理的教程 |
| 24 | 基于深度学习的通用目标检测 | 我们的 | 2019 | 维西 | 面向一般目标检测的深度学习综述 |

任何全面的近期调查。彻底审查和总结现有工作对于目标检测的进一步进展至关重要，特别是对于希望进入该领域的研究人员。由于我们的重点是通用目标检测，因此针对特定目标检测的DCNN的广泛工作，如人脸检测（Lietal.2015a；Zhangetal.2016a；Huetal.2017）、行人专用区检测（Zhangetal.2016b；Hosangetal.2015）、车辆检测（Zhou等人2016b）和交通标志检测（Zhu等人2016b）将不予考虑。

## 1.2范围

关于基于深度学习的通用目标检测的论文数量惊人。事实上，有如此之多的论文，以至于编写任何关于最新技术的综合评论都超出了任何合理篇幅的论文的范围。因此，有必要制定选择标准，这样我们就把重点限制在顶级工作上了urnal和会议论文。由于这些限制，我们真诚地向那些其作品未包含在本文中的作者致歉。有关相关主题的工作调查，读者请参阅表1中的文章。本调查主要关注过去5年的主要进展，我们将注意力限制在静态图片上，留下im视频目标检测作为一个重要的课题，将在未来单独考虑。

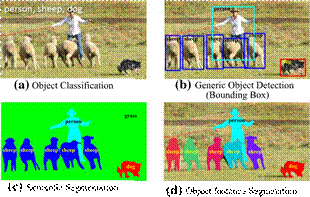
本文的主要目标是提供基于深度学习的通用对象检测技术的全面综述，并主要基于流行的数据集、评估指标、上下文建模和检测建议方法，呈现一定程度的分类、高层次的视角和组织自我化有助于读者理解各种策略之间的相似性和差异。建议的分类法为研究人员提供了一个框架，以了解当前的研究，并确定未来研究的开放挑战。

本文的其余部分组织如下。第2节总结了相关背景和过去20年取得的进展。第3节简要介绍了深度学习。第4节总结了流行的数据集和评估标准。我们描述了第5节中的里程碑目标检测框架。第s、 第6至第9节，讨论了设计目标探测器所涉及的基本子问题和相关问题。最后，在第10节中，我们对目标探测、最新性能和未来研究方向进行了全面的讨论，以结束本文。

# 2通用目标检测

## 2.1问题

*通用目标检测*，也称为通用对象类别检测、对象类别检测或对象类别检测（Zhang et al.2013），定义如下。给定一幅图像，确定其是否为预定义类别中对象的实例（通常为许多类别，例如ILSVRC对象检测挑战中的200个类别）如果存在，则返回每个实例的空间位置和范围。更强调的是检测范围广泛的自然类别，而不是特定的对象类别检测，其中只有更窄的预定义兴趣类别（例如，人脸、行人或汽车）可能存在。尽管数千个物体占据了我们生活的视觉世界，但目前研究界主要对高度结构化的物体（如汽车、人脸、自行车和飞机）和艺术的定位感兴趣-

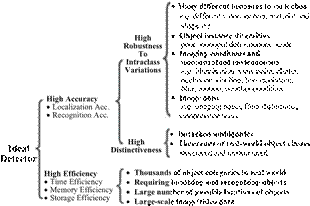


**图4**与通用对象检测相关的识别问题：a图像级对象分类、b边界框级通用对象检测、cpixel wisesemanticsegmentation、DinStanceLevel语义分割

非结构化对象（例如人、牛和马）而不是非结构化场景（例如天空、草地和云）。

对象的空间位置和范围可以使用边界框（紧密包围对象的轴对齐矩形）（Everingham等人2010；Russakovsky等人2015）、精确的像素分割遮罩（Zhang等人2013）或闭合边界（Lin等人2014；Russell等人2008）粗略定义，如图4所示。据我们所知，对于通用目标检测算法的评估，当前文献中使用最广泛的是边界框（Everingham et al.2010；Russakovsky et al.2015），因此这也是我们在本次调查中采用的方法。然而，随着研究团体不断深入了解（从图像级对象分类到单对象定位，到通用对象检测，再到像素级对象分割），预计未来的挑战将在像素级（Lin等人，2014年）。

有许多问题与一般目标检测密切相关。目标分类或目标分类的目标（图4a）是评估图像中给定对象类集合中对象的存在性；即，为给定图像指定一个或多个对象类标签，无需定位即可确定对象的存在性。在图像中定位实例的附加要求使检测比分类更具挑战性。对象识别定位问题是指识别/定位图像中存在的所有对象的更一般问题，包括对象检测和分类问题（Everingham et al.2010；Russakovsky et al.2015；Opelt）[[6]](" \l "_ftn6" \o ")



**图5**通用目标检测中的挑战分类

一般对象检测与语义图像分割密切相关（图4c），其目的是将图像中的每个像素分配给语义类标签。对象实例分割（图4d）目的是区分同一对象类的不同实例，而不是不区分的语义分割。

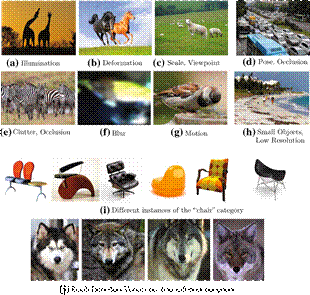
## 2.2主要挑战

通用目标检测的理想是开发一种通用算法，以实现高质量/精度和高效率这两个相互竞争的目标（图5）.如图6所示，高质量检测必须准确定位和识别图像或视频帧中的对象，以便能够区分现实世界中的各种对象类别（即，高区分性），并且来自同一类别的对象实例（受类内外观变化的影响）可以进行本地化和识别（即，高鲁棒性）。高效率要求整个检测任务以可接受的内存和存储需求实时运行。

### 2.2.1与准确性相关的挑战

挑战（1）广泛的类内变化和（2）大量的对象类别导致的检测准确性系统。

类内变化可分为两种类型：内在因素和成像条件。就内在因素而言，每个对象类别可以有许多不同的对象实例，可能在颜色、纹理、材质、形状和大小的一个或多个方面有所不同，例如“椅子”类别如图6i所示。即使在更狭义的类别中，例如人或马，对象实例也可以以不同的姿势出现，受到非刚性变形或添加衣服的影响。



**图6**在成像条件（a–h）发生变化的情况下，类的外观发生了变化。单对象类的含义（i）发生了惊人的变化。相比之下，j中的四幅图像看起来非常相似，但实际上来自四个不同的对象类。大多数图像来自ImageNet（Russakovsky et al.2015）和MS COCO（林等人，2014年）

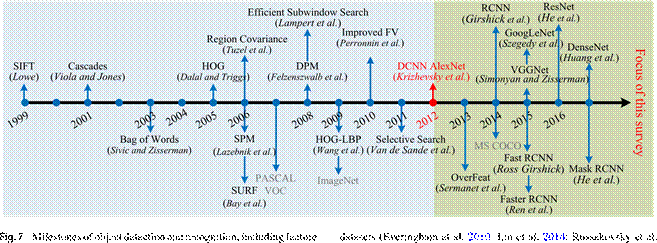
成像条件的变化是由不受约束的环境对对象外观的显著影响引起的，例如照明（黎明、白天、黄昏、室内）、物理位置、天气条件、相机、背景、照明、遮挡和视距。所有这些条件都会在对象外观上产生显著变化，如照明、姿势、比例、遮挡、杂波、阴影、模糊和运动，示例如图6a–h所示。di可能会增加进一步的挑战数字化伪影、噪声破坏、低分辨率和过滤失真。

除了组内变化外，在104–105顺序上的大量对象类别要求检测器具有强大的辨别能力，以区分细微不同的组内变化，如图6j所示。实际上，目前的探测器主要关注结构化对象类别，例如PASCAL VOC（Everingham等人2010）、ILSVRC（Russakovsky等人2015）和MS COCO（Lin等人2014）中的20、200和91个对象类别。显然，现有基准数据集中考虑的对象类别的数量比人类所能识别的要少。

### 2.2.2与效率和可扩展性相关的挑战

社交网络和移动/可穿戴设备的普及导致对分析视觉数据的需求不断增加。然而，移动/可穿戴设备的计算能力和存储空间有限，因此有效的目标检测至关重要。

效率挑战源于定位和识别的需要，计算复杂性随着对象类别的数量（可能很大）和单个图像中的位置和比例的数量（可能很大）而增长，如图6c、d中的示例。

  
另一个挑战是可伸缩性：一个架构师应该能够处理以前看不见的对象，未知的位置-

陈述（Csurka et al.2004；Dalal and Triggs 2005；He et al.2016；Krizhevsky et al.2012a；Lazebnik et al.2006；Lowe 1999，2004；Perronnin et al.2010；Simonyan and Zisserman 2015；Sivic and Zisserman2003；Szegedyetal.2015；ViolaandJones2001；Wangetal.2009），检测框架（Felzenszwalb等人2010b；Girshick等人2014；Sermanet等人2014；Uijlings等人2013；Viola和Jones 2001）和2015）。截至2012年的时间段主要是手工制作的特征，2012年，随着克里茨耶夫斯基（Krizhevskyetal）（2012a）为图像分类开发DCNNs，过渡发生，2012年后的方法主要是相关的深度网络。大多数列出的方法都被高度引用，并获得了重要的ICCV或CVPR奖。见A节。2.3了解详情

数据量大，数据传输率高。随着图像数量和类别数量的持续增长，手动注释图像可能变得不可能，这迫使依赖于弱监督策略。

## 2.3过去20年的进展

物体识别的早期研究是基于模板匹配技术和简单的基于零件的模型（Fischler和Elschlager 1973），重点关注空间布局大致僵硬的特定物体，如人脸。1990年之前，目标识别的主导范式是基于几何表示（Mundy2006；Ponceetal.2007），随后重点从几何和先前模型转向使用统计分类器[如神经网络（Rowley et al.1998）、SVM（Osuna et al.1997）和Adaboost（Viola和Jones 2001；Xiao et al.2003）]基于外观特征（Murase和Nayar 1995a；Schmid和Mohr 1997）。这一成功的目标探测器系列为该领域的大多数后续研究奠定了基础。

图7显示了目标检测或接收的里程碑，其中突出显示了两个主要ERA（SIFT与DCNN）。外观特征从全局表示法（Murase和Nayar 1995b；Swain和Ballard 1991；Turk和Pentland 1991）转移到局部表示法，这些局部表示法设计为对平移、缩放、旋转、照明、视点和遮挡的变化保持不变。手工制作的局部变异特征从尺度不变特征变换（SIFT）特征（Lowe 1999）开始获得了广泛的应用，各种视觉识别任务的进展主要基于局部描述符的使用（Mikolajczyk和Schmid 2005），如Haarlike特征（Viola和Jones 2001）、SIFT（Lowe 2004），形状背景（Belongie et al.2002），梯度直方图（HOG）（Dalal和Triggs 2005），局部二元模式（LBP）（Ojalaetal.2002），andregioncovariances（Tuzeletal.2006）。这些局部特征通常通过简单的拼接或特征池编码器聚合，如Sivic和Zisserman（2003）和Csurka等人（2004）提出的视觉单词包方法、弓模型的空间金字塔匹配（SPM）（Lazebnik等人2006）和Fisher向量（Perronnin等人2010）。

多年来，手工制作的局部描述符和鉴别分类器的多级手动调整管道主导了计算机视觉的各个领域，包括目标检测，直到2012年DCNNs（Krizhevsky et al.2012a）在图像分类方面取得了破纪录的结果，这是一个重大转折点。

CNN用于检测和定位（Rowley et al.1998）的使用可以追溯到20世纪90年代，有少量的隐藏层用于目标检测（Vaillant et al.1994；Rowley et al.1998；Sermanet al.2013），在人脸检测等受限领域取得了成功。然而，最近，更深层次的CNN在更一般的目标类别的检测方面取得了突破性的进步，这一转变发生在DCNNs在图像分类中的成功应用（Krizhevsky et al.2012a）被转移到目标检测时，产生了里程碑式的基于区域的CNN（RCNN）Girshick等人（2014）的检测器。

深度探测器的成功在很大程度上依赖于庞大的训练数据和具有数百万甚至数十亿参数的大型网络。具有极高计算能力和大规模检测数据集的GPU的可用性[如ImageNet（Deng等人，2009年；Russakovsky等人，2015年）和MS COCO（Lin等人，2014年）]对其成功起着关键作用。大型数据集使研究人员能够针对具有大类内变异和类间相似性的图像中更现实和复杂的问题（Lin等人2014；Russakovsky等人2015）。然而，精确的注释是劳动密集的获得，因此检测器必须考虑的方法，可以减轻注释困难或可以学习较小的训练数据集。

研究界已经开始朝着一个具有挑战性的目标迈进，即建立通用目标检测系统，其检测多种目标类别的能力与人类相匹配。这是一个重大挑战：根据认知科学家的说法，人类总共可以识别大约3000个世纪级别的类别和30000个视觉类别，而可与领域专业知识区分的类别数量可能达到105个左右（Biederman 1987a）。尽管过去几年取得了显著的进步，但设计一个精确、稳健、高效的检测和识别系统，在104–105个类别上接近人类水平的性能，无疑是一个尚未解决的问题。

# 3.深度学习简介

深度学习已经彻底改变了机器学习任务的范围，从图像分类和视频处理到语音识别和自然语言理解。鉴于这一惊人的快速发展，最近有许多关于深度学习的调查论文（Bengio等人2013；Goodfelle等人2016；Gu等人2018；LeCun等人2015；Litjens等人2017；Pouyanfar等人2018；Wu等人2019；Young等人2018；Zhang等人2018d；Zhou等人2018a；Zhu等人2017）。这些调查从不同角度回顾了深度学习技术（Bengioetal.2013；Goodfello et al.2016；Gu et al.2018；LeCun et al.2015；Pouyanfar et al.2018；Wu et al.2019；Zhou et al.2018a），或将其应用于医学图像分析（Litjens et al.2017）、自然语言处理（Young et al.2018），语音识别系统（Zhang等人，2018d）和遥感（Zhu等人，2017）。

|  |
| --- |
| **（a） （b）** |

**图8a**典型CNN重复应用的三种操作的说明：与多个线性滤波器的卷积；非线性（如ReLU）；和本地池（例如最大池）。使用N个不同的过滤器（此处显示为大小3×3×）将前一层的M个特征贴图卷积，步长为1。然后，将生成的N个特征图通过非线性函数（例如ReLU）并合并（例如，取最大值超过2×2个区域），以降低分辨率给出N个特征图。b VGGNet体系结构示意图（Simonyan和Zisserman 2015），一个典型的CNN，具有11个权重层。*M*

一个具有3个颜色通道的图像作为输入。该网络有8个卷积层、3个全连接层、5个最大池层和一个softmax分类层。最后三个完全连接的层以向量形式将来自顶部卷积层的特征作为输入。最后一层是C路softmax函数，C是类的数量。通过随机梯度下降优化目标函数（如均方误差或交叉熵损失），可以从标记的训练数据中学习整个网络（在线彩色图）

卷积神经网络（CNN）是最具代表性的深度学习模型，可用于揭示自然信号的基本特性：平移不变性、局部连通性和成分层次结构（LeCun et al.2015）。图8所示的典型CNN具有层次结构，由多个层组成，用于学习具有多个抽象级别的数据表示（LeCun et al.2015）。我们从卷积开始

***x****L*−∗ (1)1 *L*

在输入特征映射xl之间−在上一层的要素地图上−1，使用DConvolutionalKernel（或过滤器或权重）wl进行卷积。这种卷积出现在一系列层上，受到非线性运算σ的影响，因此1

***x****LJ*, ,                                         (2)

现在Nl之间有一个卷积−输入要素地图−和相应的内核wli，加上偏差项blj。元素非线性函数σ（·）通常是每个元素的校正线性单元（ReLU），1 ***x****L我*1 ,*J*

σ（）=max{，0}。（3） 最后，池对应于特征映射的下采样/上采样。这三种操作（卷积、非线性、池）在图8a中示出；具有大量层的cnn，即“深”网络，被称为深cnn（DCNNs），其典型的DCNN架构如图8b所示。*xx*

大多数图层都与自然贴图的数量一致，其中每个像素都像一个神经元。卷积层中的每个神经元通过一组权值wi（基本上是一组2D滤波器）连接到前一层的特征映射。如图8b所示，早期CNN层通常由卷积层和池层组成，后期层通常完全连接。从早期到后期，输入图像被反复卷积，随着每一层，感受野或支持区域增加。一般来说，最初的CNN层提取低级特征（如边缘），随后的层提取更加复杂的一般特征（Zeiler和Fergus 2014；Bengio等人2013；LeCun等人2015；Oquab等人2014）。,*J*

DCNN有许多突出的优点：一种分层结构，用于学习具有多个抽象级别的数据表示，能够学习非常复杂的函数，以及直接和自动地从数据中学习特征表示，只需最少的领域知识。使DCNNs特别成功的是大规模标记数据集和具有极高计算能力的GPU的可用性。

尽管取得了巨大的成功，但知识仍然存在。特别是，人们迫切需要有标签的培训数据，需要昂贵的计算资源，并且仍然需要相当多的技能和经验来选择合适的学习参数和网络架构。训练有素的网络解释能力差，对退化缺乏鲁棒性，许多DCNN表现出严重的攻击脆弱性（Goodfello等人，2015年），所有这些都限制了DCNN在现实应用中的使用。

# 4数据集和性能评估

## 4.1数据集

在目标识别研究的整个历史中，数据集发挥了关键作用，不仅是衡量和比较竞争算法性能的共同基础，而且还将该领域推向了日益复杂和具有挑战性的问题。特别是最近，深度学习技术极大地解决了许多视觉认知问题，并提供了大量的注释数据，这些数据对这些问题的成功起着关键作用。通过访问互联网上的大量图像，可以建立全面的数据集，以捕获大量丰富多样的对象，从而实现前所未有的对象识别性能。

对于一般目标检测，有四个著名的数据集：PASCAL VOC（Everingham et al.2010、2015）、ImageNet（Deng et al.2009）、MS COCO（Lin et al.2014）和Open Images（Kuznetsova et al.2018）。表2中总结了这些数据集的属性，图9显示了选定的样本图像。创建大规模Notated数据集有三个步骤：确定目标对象类别的类型，收集一组不同的候选图像以表示互联网上选定的类别，以及注释收集的图像，通常通过设计众包策略。认识到空间限制，我们建议感兴趣的读者参考原始论文（Everingham et al.2010、2015；Lin et al.2014；Russakovsky et al.2015；Kuznetsova et al.2018），了解这些数据集在结构和性质方面的详细描述。

这四个数据集构成了各自检测挑战的主干。每个挑战包括一个公开的图像数据集、地面真相注释和标准化评估软件，以及一个年度竞赛和相应的研讨会。表3给出了检测挑战的培训、验证和测试数据集中的图像和对象实例数量的统计信息。VOC、COCO、ILSVRC和开放式图像检测数据集中最常见的对象类如表4所示。[[7]](" \l "_ftn7" \o ")

PASCAL VOC Everingham等人（2010年、2015年）多年来一直致力于分类和目标检测基准数据集的创建和维护，开创了在FannualCompetition之前对识别算法进行标准化评估的先例。从2005年的四个分类开始，该数据集已增加到日常生活中常见的20个类别。自2009年以来，图像的数量每年都在增长，但保留了所有以前的图像，以便每年对测试结果进行比较。由于ImageNet、MS COCO和Open Images等大型数据集的可用性，PASCAL VOC已逐渐过时。

ILSVRC，ImageNet大规模视觉识别挑战（Russakovsky et al.2015），源自ImageNet（Deng et al.2009），将PASCAL VOC的检测算法标准化培训和评估目标在对象类别和图像数量上提升了一个数量级以上。ImageNet1000是ImageNet图像的一个子集，包含1000个不同的对象类别和120万张图像，现已修复，为ILSVRC图像分类挑战提供了标准化基准。

COCO女士回应了ImageNet的批评，即其数据集中的对象往往较大且居中，这使得ImageNet数据集在现实场景中具有非典型性。为了推动更丰富的图像理解，研究人员创建了MS COCO数据库（Lin et al.2014），该数据库包含复杂的日常场景，这些场景中的普通对象在其自然环境中更接近真实生活，其中对象使用完全分割的实例进行标记，以提供更准确的检测器评估。COCO对象检测挑战赛（Lin et al.2014）具有两个对象检测任务：使用边界框输出或对象实例分割输出。COCO提出了三项新挑战：

1.    它包含范围广泛的物体，包括高比例的小物体（Singh和Davis，2018）；

2.    物体不那么具有标志性，并且处于杂乱或严重遮挡中；

3.    评估指标（见表5）鼓励更准确的对象定位。

就像当时的ImageNet一样，MS COCO已经成为今天目标检测的标准。

OICOD（开放图像挑战目标检测）源自OpenImageSv4（2014年为NowV5）（Kuznetsova等人，2018年），目前是最大的公开对象

**表2**用于对象识别的流行数据库

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集名称 | 总图像 | 按类别对图像进行分类按图像大小对对象进行分类 | | | | 开始年度亮点 | |
| 帕斯卡  挥发性有机化合物（2012年）  （埃弗林-  哈姆等人。  2015) | 11,540 | 20 | 303–4087 | 2.4 | 470 × 380 | 2005 | 仅涵盖日常生活中常见的20个类别；大量的训练图像；接近真实世界的应用程序；显著较大的组内变异；场景语境中的对象；一幅图像中的多个对象；包含许多困难的样本 |
| ImageNet（Russakovsky等人，2015年） | 1400万+ | 21,841 | − | 1.5 | 500 × 400 | 2009 | 对象类别多；每个图像有更多实例和更多类别的对象；比PASCAL VOC更具挑战性；ILSVRC挑战的骨干力量；图像是以对象为中心的 |
| 可可女士  （林等人，2014年） | 328,000+ | 91 | − | 7.3 | 640 × 480 | 2014 | 更接近真实世界的场景；每个图像包含更多的对象实例和更丰富的对象注释信息；包含ImageNet数据集中不可用的对象分段表示法数据 |
| 地点  （Zhou等人，2017a） | 1000万+ | 434 | − | − | 256 × 256 | 2014 | 用于场景识别的最大标记数据集；四个子集  地点365标准，地点365  挑战，地点205和  排名88作为基准 |
| 打开  图像（Kuznetsova等人，2018年） | 900万+ | 6000+ | − | 8.3 | 多种多样的 | 2017 | 使用图像级别标签、对象边界框和视觉关系进行注释；Open Images V5支持大规模对象检测、对象实例分割和可视关系检测 |

来自PASCAL VOC、ImageNet、MS COCO和Open图像的示例图像如图9所示



**图9**来自PASCAL VOC、ILSVRC、MS COCO和Open images的带有对象注释的一些示例图像。有关这些数据集的摘要，请参见表2

检测数据集。OICOD不同于以前的大型类、图像、边界框注释和实例级对象检测数据集，如ILSVRC和MS COCO、分割掩码注释，但它不仅在显著增加的过程数量方面考虑注释。在ILSVRC和MS COCO中，所有

**表3**常用目标检测数据集的统计信息

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 质询对象类图像数  训练 | | | 瓦尔 | 试验 | 注释对象的数量    训练 | 瓦尔 | 汇总（列车+Val） | |  |
| 图像 | 盒 | 方框/图像 |
| VOC07 | *PASCAL VOC目标检测挑战*  20 | 2501 | 2510 | 4952 | 6301(7844) | 6307(7818) | 5011 | 12,608 | 2.5 |
| VOC08 | 20 | 2111 | 2221 | 4133 | 5082(6337) | 5281(6347) | 4332 | 10,364 | 2.4 |
| VOC09 | 20 | 3473 | 3581 | 6650 | 8505(9760) | 8713(9779) | 7054 | 17,218 | 2.3 |
| 挥发性有机化合物10 | 20 | 4998 | 5105 | 9637 | 11,577(13,339) | 11,797(13,352) | 10,103 | 23,374 | 2.4 |
| 挥发性有机化合物11 | 20 | 5717 | 5823 | 10,994 | 13,609(15,774) | 13,841(15,787) | 11,540 | 27,450 | 2.4 |
| 挥发性有机化合物12 | 20 | 5717 | 5823 | 10,991 | 13,609(15,774) | 13,841(15,787) | 11,540 | 27,450 | 2.4 |
| *ILSVRC对象*  ILSVRC13 | *检测试剂盒*  200 | *通用电气*  395 909 | 20,121 | 40,152 | 345,854 | 55,502 | 416 030 | 401,356 | 1.0 |
| ILSVRC14 | 200 | 456 567 | 20,121 | 40,152 | 478,807 | 55,502 | 476 668 | 534,309 | 1.1 |
| ILSVRC15 | 200 | 456 567 | 20,121 | 51,294 | 478,807 | 55,502 | 476 668 | 534,309 | 1.1 |
| ILSVRC16 | 200 | 456 567 | 20,121 | 60,000 | 478,807 | 55,502 | 476 668 | 534,309 | 1.1 |
| ILSVRC17 | 200 | 456 567 | 20,121 | 65,500 | 478,807 | 55,502 | 476 668 | 534,309 | 1.1 |
| *MS-COCO目标检测chall*  COCO15 80女士 | | *恩格*  82,783 | 40,504 | 81,434 | 604,907 | 291,875 | 123,287 | 896,782 | 7.3 |
| Coco1680女士 | | 82 783 | 40,504 | 81,434 | 604,907 | 291,875 | 123 287 | 896,782 | 7.3 |
| COCO17 80女士 | | 118 287 | 5000 | 40,670 | 860,001 | 36,781 | 123 287 | 896,782 | 7.3 |
| COCO18 80女士 | | 118,287 | 5000 | 40,670 | 860,001 | 36,781 | 123 287 | 896,782 | 7.3 |
| *开放图像挑战目标检测*（OICOD）（基于开放图像V4 Kuznetsova等人。  OICOD18 500 1643042 100000 9999 11498734 696410 | | | | | | | 1,743,042 | 12,195,144 | 7.0 |

VOC挑战的对象统计信息列出评估中使用的非困难对象（所有注释对象）。对于COCO挑战赛，在2017年之前，测试集有四个分割（开发、标准、保留和挑战），每个分割约有20K图像。从2017年开始，测试集只有Dev和Challenge拆分，其他两个拆分被删除。从2017年开始，train和val集的排列方式不同，测试集被分为两个大小大致相同的部分，每个部分大约有20000个图像：测试开发和测试挑战。请注意，2017年测试开发/挑战拆分包含与2015年测试开发/挑战拆分相同的图像，因此多年的结果具有直接可比性

数据集中的类被详尽地注释，而对于开放图像V4，分类器被应用于每个图像，并且只有那些分数足够高的标签被发送给人类验证。因此，在OICOD中，仅对人类确认阳性标签的对象实例进行注释。**4.2评价标准**

评估检测算法性能有三个标准：检测速度（以每秒帧数（FPS）为单位）、精度和召回率。最常用的指标是平均准确度（AP），由准确度和召回率得出。AP通常以特定于类别的方式进行评估，即分别为每个对象类别计算。为了比较所有对象类别的性能，采用所有对象类别的平均AP（mAP）作为性能的最终度量。关于这些指标的更多详情，请参见Everingham等人（2010年）、Everingham等人（2015年）、Russakovsky等人（2015年）、Hoiem等人（2012年）。[[8]](" \l "_ftn8" \o ")

应用于测试图像I的检测器的标准输出为预测检测{（，）}，由对象j索引，边界盒（BB）bj，预测类别cj和置信度pj。如果*北京希杰pjJBCP*

•预测的c类等于地面真相标签cg。

•重叠率IOU（联合体上的交叉点）（Everingham等人，2010年；Russakovsky等人，2015年）

*G*= (*B*∩∪ *背景*) *地区*

        借据（，），（4）*BB*

*地区*(*b背景*)

预测的BB b和地面真实值bg之间不小于预定义的阈值ε，其中∩ 和



脚注3续

这是合理的，因为根据mAP对团队进行的排名始终与根据赢得的对象类别数量进行的排名相同（Russakovsky et al.2015）。

**表4**每个检测挑战最常见的对象类



每个单词的大小与训练数据集中该类的频率成正比

*杯*分别表示相交和负。ε的非典型值为0.5。

否则，它将被视为假阳性（FP）。置信水平p通常与某个阈值β进行比较，以确定预测的类别标签c是否被接受。

根据精度和召回率，分别为每个对象类计算AP。对于给定的对象类c和

检测器返回的测试图像，rankedbyconfidenceIi，让{（，）}=1表示检测*比吉皮吉MJ皮吉*不屈不挠的

秩序。每个检测（，）是TP或FP，其可以通过图10中的算法来确定。基于TP和FP检测，精度P（β）和召回率R（β）（Everingham et al.2010）可以作为置信阈值β的函数计算，因此通过改变置信阈值可以获得不同的对（，），原则上允许将精度视为召回的函数，即P（），从中可以找到平均精度（AP）（Everingham et al.2010；Russakovsky et al.2015）。*比吉皮吉*[[9]](" \l "_ftn9" \o ") *PRR*

自从引入MS COCO以来，更多的注意力放在边界框位置的准确性上。MSCO引入了一些指标（总结在表5中）来表征对象检测器的性能，而不是采用固定阈值。例如，与在单个IoU为0.5时计算的传统贴图不同，APcoco在所有对象类别和从0.5到0.95的多个IoU值之间以0.05的步长进行平均。由于MS COCO中41%的对象是小对象，24%的对象是大对象，因此

*美联社可可购物中心*此外，还介绍了。最后*椰壳共极化*

表5总结了PASCAL、ILSVRC和MS COCO目标检测挑战中使用的主要指标，以及Kuznetsova等人（2018）提出的开放图像挑战的指标修改。

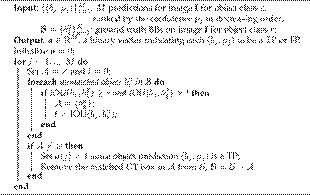
# 5检测框架

从手工制作的特征（Viola和Jones，2001年；Dalal和Triggs，2005年；Felzenszwalb等人，2008年；Harzallah等人，2009年；Vedaldi等人，2009年）到学习的DCNN特征的巨大变化证明，用于识别的对象特征表示和分类器取得了稳步进展（Girshick et al.2014；Ouyang et al.2015；Girshick 2015；Ren et al.2015；Dai et al.2016c）。相比之下，就本地化而言，基本的“滑动窗口”策略（Dalal和Triggs 2005；Felzenszwalb et al.2010b，2008）仍然是主流，尽管做出了一些努力来避免彻底搜索（Lampert et al.2008；Uijlings et al.2013）但是，窗口的数量很大，并且随着图像像素的数量呈二次增长，需要在多个尺度和纵横比上搜索进一步增加了搜索空间。因此，设计高效的检测框架在降低计算成本方面起着关键作用。常用的策略包括级联、共享特征计算和减少每窗口计算。

本节回顾了图11和表11所列的检测框架，以及自深入学习进入该领域以来出现的里程碑式方法，这些方法被组织为女性主要类别：

（a） 两阶段检测框架，包括用于生成对象建议的预处理步骤；

（b） 单阶段检测框架，或区域建议框架，具有单个建议方法，该方法不分离检测建议的过程。



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表5**评估对象检测器的常用指标摘要   |  |  |  |  | | --- | --- | --- | --- | | 米制的 | 意思 | 定义和说明 | | | 总磷 | 真阳性 | 真阳性检测，如图10所示 | | | 计划生育 | 假阳性 | 假阳性检测，如图10所示 | | | β | 置信阈值 | 计算P（β）和R（β）的置信阈值 | | | ε | 借据阈值 | 挥发性有机化合物通常约为0.5  ILSVRC min×是燃气轮机箱体的尺寸*H*  MS COCO十个借据阈值ε∈ {0.5 : 0.05 : 0.95} | | | *P*(β) | 精确 | 置信度至少为β的检测器返回的总检测中正确检测的分数 | | | *R*(β) | 回忆起 | 探测器检测到的所有Nc对象的分数，置信度至少为β | | | 美联社 | 平均精度 | 通过改变置信度β计算不同召回水平 | | | 地图 | 平均精度 | 挥发性有机化合物 | 单个IOU的AP和所有类的平均值 | |  |  | ILSVRC | 在修改的IOU上的AP，并在所有类别上取平均值 | |  |  | 可可女士 | *APcoco*：mAP平均值超过10条：{0.5:0.05:0.95}；  *APcoco*借据=IOU=0.50时的映射（帕斯卡VOC度量）；0.[[10]](" \l "_ftn10" \o ")  *APcoco*借据=IOU处的映射=0.75（严格度量）；0.75  *APcoco*小的：面积小于322的小型物体的地图；  *APcoco*中等的：322和962之间区域对象的地图；  大的  *APcoco*：面积大于962的大型对象的地图； | | 应收账 | 平均召回率 | 最大阈值 | m recall给定每个图像的固定检测次数，在所有类别和IOU上取平均值 | | 应收账 | 平均召回率 | 可可女士 | *阿科科*最大值=:AR每幅图像有1次检测；ARcocomax=:AR每幅图像有10次检测；110  *阿科科*最大值=:AR，每幅图像100次检测；100  *阿科科*小的：AR表示面积小于322的小物体；  *阿科科*中等的：AR表示面积介于322和962之间的物体；  大的  *阿科科*：AR，适用于面积大于962的大型物体； | |

**图10**通过将目标检测结果与地面真值盒贪婪匹配确定TPs和FPs的算法

第6-9节将更详细地讨论检测框架中涉及的基本子问题，包括DCNN功能、检测方案和上下文建模。

## 5.1基于区域的（两阶段）框架

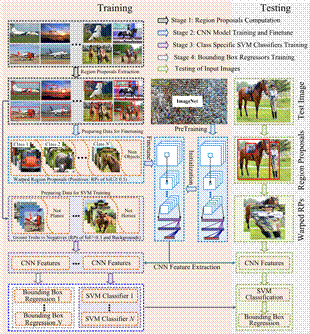
在基于区域的框架中，从图像生成类别无关的区域建议，从这些区域提取CNN（Krizhevsky et al.2012a）特征，然后使用类别特定分类器确定建议的类别标签。如图11所示，DetectorNet（Szegedy et al.2013），OverFeat（Sermanet al.2014）、MultiBox（Erhan et al.2014）和RCNN（Girshick et al.2014）独立且几乎同时提出使用CNN进行一般目标检测。5

RCNN（Girshick et al.2014）：受CNN获得的突破性图像分类结果和人工特征选择搜索区域提案的成功（Uijlings et al.2013）的启发，Girshick et al.（2014、2016）是最早探索CNN用于一般目标检测的人之一，并开发了RCNN，该RCNN集成了

|  |
| --- |
|  |
|  |  |

（Sermanet al.）

**图11**通用目标检测中的里程碑



**图12**RCNN检测框架示意图（Girshick等人，2014年、2016年）

AlexNet（Krizhevsky等人，2012a）和区域建议选择性搜索（Uijlings等人，2013）。如图12所示，训练RCNN框架由多级管道组成：

1.    *区域建议计算*类不可知区域建议（可能包含对象的候选区域）是通过选择性搜索获得的（Uijlings et al。2013).

2.    *CNN模型微调*从图像裁剪并扭曲成相同大小的区域建议用作微调CNN模型预处理的输入-

使用大型数据集（如ImageNet.At）进行培训

在这一阶段，所有与地面真值框重叠0.5 IOU的区域提案均被定义为该地面真值框类的积极提案，其余为消极提案。[[11]](" \l "_ftn11" \o ")

3.    *分类支持向量机分类器训练*使用CNN提取的固定长度特征来训练一组类别特定的线性SVM分类器，取代通过微调学习的softmax分类器。对于训练SVM分类器，每个类别的正样本都要细化到基本真值框。一个IOU小于0.3的区域建议与cla的所有基本真值实例重叠ss对于该类是否定的。请注意，为训练SVM分类器而定义的肯定和否定示例与微调CNN的示例不同。

4.    *特定于类的边界框回归器训练*为每个具有CNN特征的对象类学习边界框回归。

尽管实现了较高的目标检测质量，但RCNN存在明显的缺陷（Girshick 2015）：

1.    培训是一个多阶段的过程，缓慢且难以优化，因为每个阶段都必须单独进行培训。

2.    对于SVM分类器和边界盒回归器训练，在磁盘空间和时间上都很昂贵，因为需要从每个图像中的每个对象建议中提取CNN特征，这对大规模检测提出了巨大挑战，特别是对于非常深的网络，如VGG16（Simonyan和Zisserman 2015）。

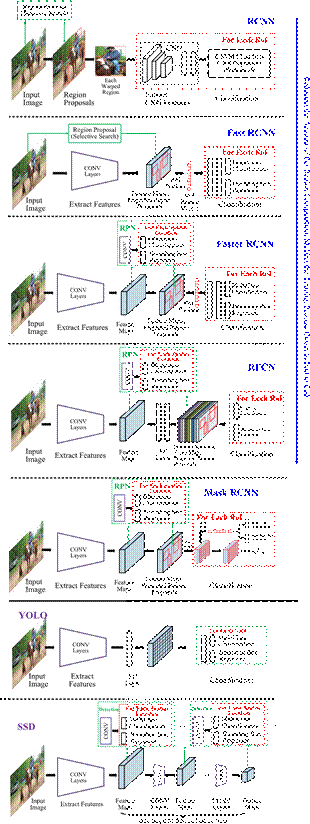
3.    测试很慢，因为CNN特征是通过每个测试图像中的对象建议提取的，没有共享计算。

所有这些缺点推动了后续创新，导致了许多改进的检测框架，如SPPNet、Fast RCNN、Fast RCNN等，如下所示。

SPPNet（He et al.2014）在测试期间，CNN特征提取是RCNN检测管道的主要瓶颈，需要从每幅图像的数千个扭曲区域建议中提取CNN特征。因此，He et al.（2014）引入了传统的空间金字塔池（SPP）（Grauman and Darrell 2005；Lazebnik et al.2006）CNN体系结构。由于卷积层接受任意大小的输入，CNN中固定大小图像的要求仅是由于完全连接（FC）层，因此他等人在最后一个卷积层（CONV）的顶部添加了SPP层层以获得FC层的固定长度特征。使用此SPPNet，RCNN在不牺牲任何检测质量的情况下获得显著的加速，因为它只需在整个测试图像上运行一次卷积层，即可为任意大小的区域建议生成固定长度特征。而SPPNet则加速RCNN评估通过数量级的评估，它不会导致检测器训练的可比加速。此外，SPPNet中的微调（He et al.2014）无法更新SPP层之前的卷积层，这限制了非常深的网络的精度。

Fast RCNN（Girshick 2015）Girshick提议的Fast RCNN（Girshick 2015）这解决了RCNN和SPPNet的一些缺点，同时提高了它们的检测速度和质量。如图13所示，Fast RCNN通过开发一个简化的训练过程来实现端到端检测器训练，该过程同时学习softmax分类器和特定于类的边界盒回归，而不是单独学习在RCNN/SPPNet中培训softmax分类器、支持向量机和边界盒回归器（BBR）。Fast RCNN采用跨区域建议共享卷积计算的思想，并添加感兴趣区域（RoI）最后一个CONV层和第一个FC层之间的池层，用于为每个区域方案提取固定长度的特征。本质上，RoI池使用特征级别的扭曲来近似图像级别的扭曲。RoI池层之后的特征被馈送到FC层序列中，最终分支为两个同级输出层：用于对象类别预测的softmax概率，以及用于提案细化的特定于类的边界框回归偏移。与RCNN/SPPNet相比，快速RCNN提高了效率，通常在训练中快3倍，在测试中快10倍。因此检测质量更高，单个训练过程hat更新所有网络层，功能缓存不需要存储。

Faster RCNN（Ren等人，2015年，2017年）虽然Fast RCNN显著加快了检测过程，但它仍然依赖于外部区域方案，其计算被暴露为Fast RCNN中新的速度瓶颈。最近的工作表明，CNNSHAVE具有显著的能力，可以在CONV层中定位对象（Zhou等人2015年、2016a；Cinbis等人2017年；Oquab等人2015年；Hariharan等人2016年），一份



**图13**通用目标检测的主要框架的高级图表。表11总结了这些方法的特性

在FC层中被削弱的能力。因此，在生成区域提案时，可使用CNN代替选择性搜索。Ren等人（2015年、2017年）提出的更快的RCNN框架提供了高效准确的区域提案网络（RPN）用于生成区域建议。它们使用相同的主干网络，使用来自最后一个共享卷积层的特征来完成RPN用于区域建议和快速RCNN用于区域分类的任务，如图13所示。

RPN首先初始化k个参考框（即所谓的锚）在每个CONV特征映射位置具有不同的比例和纵横比。定位点位置独立于图像内容，但从定位点提取的特征向量本身依赖于图像内容。每个定位点映射到一个低维向量，该向量被馈送到两个同级FClayers AnObjectCategoryClassificationLayers和一个盒回归层。与快速RCNN中的检测相比，RPN中用于回归的特征与锚盒的形状相同，因此k个锚导致k个回归器。RPN与快速RCNN共享NV特征，从而实现高效的区域建议计算。RPN实际上是一种完全卷积网络（FCN）（Long等人，2015年；Shelhamer等人，2017年）；因此，更快的RCNN是一个纯粹基于CNN的框架，不使用手工制作的功能。

对于VGG16模型（Simonyan和Zisserman 2015），更快的RCNN可以在GPU上以5 FPS（包括所有阶段）进行测试，同时在PASCAL VOC 2007上实现最先进的目标检测精度，每幅图像使用300个方案。Ren等人（2015）中最初的更快的RCNN包含几个交替训练阶段，后来在Ren等人中进行了简化。(2017).

同时开发更快的RCNN、Lenc和Vedaldi（2015年）挑战了区域建议生成方法（如选择性搜索）的作用，研究了基于CNN的检测器中区域建议生成的作用，发现CNN包含足够的几何信息，以便在CONV而不是FC层中进行准确的对象检测。它们显示了构建集成的、更简单的仅依赖CNN的ster对象检测器，消除区域建议生成方法，如选择性搜索。

*RFCN*（基于区域的完全卷积网络）虽然快速RCNN比快速RCNN快一个数量级，但每个RoI（每个图像数百个RoI）仍需要应用区域子网络这一事实导致Dai等人（2016c）提出了完全卷积（无隐藏FC层）的RFCN检测器几乎所有计算都在整个图像上共享。如图13所示，RFCN与快速RCNN的区别仅在于RoI子网络。在快速RCNN中，RoI池后的计算不能共享，soDaietal。（2016c）建议使用所有ConvLayers构建共享Roisub网络，并在预测前从CONV功能的最后一层获取RoI作物。然而，Dailetal.（2016c）发现这种幼稚的设计结果具有相当低的检测精度，推测更深的CONV层对类别语义更敏感，对翻译更不敏感，而目标检测需要尊重翻译不变性的本地化表示。基于这一观察，Dai等人（2016c）通过使用一组专门的CONV层ASTHEFCNOUTPUT，ONTOPOF构建了一组位置敏感分数图，其中添加了位置敏感RoI池层。他们表明，使用ResNet101的RFCN（He等人，2016）可以达到与更快的RCNN相当的精度，通常运行时间更快。

*掩码RCNN*He等人（2017）提出了Mask RCNN，通过扩展更快的RCNN来解决像素级对象实例分割问题。Mask RCNN采用相同的两阶段管道，具有相同的第一阶段（RPN），但在第二阶段，与预测类和盒偏移量平行，Mask RCNN添加了一个分支，该分支为每个RoI输出一个二进制掩码。新分支是CNNfeaturemap之上的一个完全卷积网络（FCN）（Long等人2015；Shelhamer等人2017）。为了避免原始RoI池（RoIPool）使用的对齐错误层，建议使用ROIALLIGN层来保留像素级的空间对应关系。使用主干网络ResNeXt101 FPN（Xie et al.2017；Lin et al.2017a），Mask RCNN在COCO对象实例分割和包围盒对象检测方面取得了最好的结果。它训练简单，通用性好，并且只为运行速度为5 FPS的更快的RCNN增加了一小部分开销（He et al.2017）。

*链式级联网络与级联RCNN*cascade（Felzenszwalb et al.2010a；Bourdev and Brandt 2005；Li and Zhang 2004）的本质是通过使用多级分类器学习更多的区分性分类器，以便早期阶段丢弃大量容易否定的样本，以便后期阶段可以专注于处理更难的示例。两阶段目标检测可以看作是一个级联，第一阶段检测器去除大量背景，第二阶段对剩余区域进行分类。最近，链式级联网络（Ouyang et.2017a）中提出了两个以上级联分类器和DCNN的端到端学习，扩展了CascaderCNN（CaiandVasconcelos 2018），最近应用于同时对象检测和设备分割（Chenetal.2019a），赢得COCO 2018检测挑战赛。

*光头*为了进一步提高RFCN的检测速度（Dai等人，2016c），Li等人（2018c）提出了光头RCNN，使检测网络的光头尽可能轻，以减少RoI计算。特别是，Li et al.（2018c）应用卷积生成具有小通道数（例如，COCO为490个通道）和廉价RCNN子网络的薄特征图，从而在速度和精度方面实现了极好的权衡。

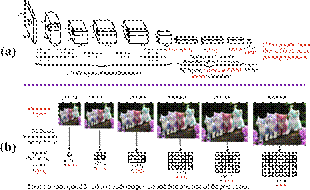
## 5.2统一（一阶段）框架

Sect基于区域的管道策略。5.1自RCNN（Girshick et al.2014）以来一直占据主导地位，因此流行基准数据集的领先结果都基于更快的RCNN（Ren et al.2015）。然而，对于当前存储和计算能力有限的移动/可穿戴设备而言，基于区域的方法在计算上非常昂贵，因此研究人员开始开发统一的检测策略，而不是试图优化复杂的基于区域的管道的各个组件。

统一管道是指在不涉及区域建议生成或分类后/特征重采样的单片设置中，使用单个前馈CNN直接预测类概率和从完整图像的边界框偏移的架构，将所有计算封装在单个网络中。由于整个以太网是单个网络，因此可以直接对端到端检测性能进行优化。

*探测网*（Szegedy et al.2013）是最早探索CNN用于目标检测的人之一。DetectorNet将对象检测问题表述为对象边界框遮罩的回归问题。他们使用AlexNet（Krizhevsky et al.2012a）并用回归层替换最终的softmax分类器层。给定一个图像窗口，他们使用一个网络预测粗网格上的前景像素，以及四个额外的网络预测对象的上、下、左、右半部分。然后，分组过程将预测的遮罩转换为检测到的边界框。网络需要按照对象类型和遮罩类型进行训练，并且不能扩展到多个类。DetectorNet必须对图像进行多次裁剪，并对每个裁剪的每个部分运行多个网络，从而使其速度变慢。

*壮举*由Sermanet et al.（2014）提出，如图14所示，可被视为第一个基于完全卷积的单级目标检测器之一



**图14**OverFeat（Sermanet al.2014）检测框架示意图

深度网络。它是最有影响力的目标检测框架之一，赢得了ILSVRC2013本地化和检测竞赛。OverFeat通过网络中的完全卷积层（即“特征提取器”，如图14a所示）的单个前向传递执行目标检测。测试时目标检测的关键步骤可总结如下：

1.    *通过在多尺度图像上通过滑动窗口方式执行对象分类来生成候选对象*OverFeat使用类似于AlexNet的CNN（Krizhevsky等人。2012a)，这将需要固定大小的输入图像，因为它的层是完全连接的，为了使滑动窗口方法计算效率高，OverFeat对网络进行了投射（如图所示）。14a） 通过将完全连接的层视为具有大小为1的内核的卷积，将其转换为完全卷积网络，获取任何大小的输入× 1.OverFeat通过网络将原始图像放大六倍，利用多尺度特性提高整体性能（如图。14b） ，从而大大增加了已评估上下文视图的数量。对于每个多尺度输入，分类器输出预测网格（类别和置信度）。

2.    *通过偏移量最大池增加预测数*为了提高分辨率，OverFeat在最后一个CONV层之后应用offset max pooling，即在每个偏移处执行子采样操作，产生更多的投票视图，提高稳健性，同时保持效率。

3.    *边界盒回归*识别对象后，将应用单个边界框回归器。分类器和回归器共享相同的特征提取（CONV）层，只有在计算分类网络后需要计算的特征提取层。

4.    *联合用药*OverFeatures使用合并策略，在所有位置和比例上合并单个边界框预测。

OverFeat具有显著的速度优势，但精确度不如RCNN（Girshicketal.2014），因为当时很难训练完全卷积网络。速度优势来自于在完全卷积网络中重叠窗口之间共享卷积计算。OverFeat类似于后来的框架，如YOLO（Redmon et al.2016）和SSD（Liu et al.2016），只是OverFeat中的分类器和回归器是按顺序训练的。

*约洛*Redmon等人（2016）提出了YOLO（你只看一次），这是一种统一的检测器，将对象检测作为从图像像素到空间分离的边界框和相关类别概率的回归问题，如图13所示。由于区域建议生成阶段已完全取消，因此YOLO直接使用一小组候选区域预测检测。与基于区域的方法（例如更快的RCNN）不同，基于区域的方法基于局部区域的特征预测检测，YOLO全局使用整个图像的特征。特别是，YOLO将图像划分为S×网格，每个网格预测C类概率、B边界框位置和置信度分数。通过完全放弃区域提案生成步骤，YOLO的设计速度很快，实时运行速度为45 FPS，fast YOLO（Redmon et al.2016）为155 FPS。由于YOLO在进行预测时看到了整个图像，因此它隐式编码了有关对象类的上下文信息，并且在Background中不太可能预测FalsePositivesInTheBackground。YOLO比Fast RCNN产生更多的定位错误，这是由于边界框位置、比例和纵横比的粗略划分造成的。如Redmon等人（2016）所述，YOLO可能无法定位某些对象，尤其是小对象，这可能是因为网格划分粗糙，并且每个网格单元只能包含一个对象。目前尚不清楚YOLO在多大程度上可以在每个图像包含多个对象的数据集上转化为良好的性能，例如MS COCO。[[12]](" \l "_ftn12" \o ")*s*

*YOLOv2和YOLO9000*Redmon和Farhadi（2017）提出了YOLOv2，这是YOLO的一个改进版本，其中自定义GoogLeNet（Szegedy et al.2015）网络被更简单的DarkNet19取代，加上批量标准化（He et al.2015），移除了完全连接的层，并使用通过kmeans和multiscaletraining学习的良好锚定框。Yolov2完成了最先进的标准检测任务。Redmon和Farhadi（2017）也推出了YOLO9000，通过提出一种联合优化方法，在ImageNet分类数据集和COCO检测数据集上同时进行训练，并结合WordTree将来自多个来源的数据进行组合，可以实时检测9000多个对象类别。例如，点训练允许WSYOLO9000执行弱监督检测，即，检测没有边界框注释的对象类。[[13]](" \l "_ftn13" \o ")

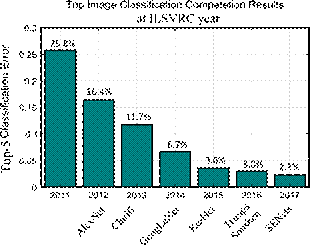
*固态硬盘*为了在不牺牲太多检测精度的情况下保持实时速度，Liu等人（2016年）提出了SSD（单次激发检测器），比YOLO（Redmon等人，2016年）更快，并且精度与基于区域的检测器（如更快的RCNN）相竞争（Ren等人，2015年）。SSD有效地结合了快速RCNN（Ren等人2015）、YOLO（Redmon等人2016）和多尺度CONV功能（Hariharan等人2016）中RPN的思想，以实现快速检测速度，同时保持高检测质量。与YOLO一样，SSD预测固定数量的边界框和分数，然后通过NMS步骤生成最终检测。SSD中的CNN网络是完全卷积的，其早期层基于标准体系结构，如VGG（Simonyan和Zisserman 2015），然后是几个辅助网络，其规模逐渐减小。最后一层中的信息在空间上可能过于粗糙，无法允许精确定位，因此通过在多个CONV特征图上操作，在多个尺度上执行检测，每个CONV特征图预测适当大小边界框的类别分数和框偏移。对于300×300输入，SSD在VOC2007测试中以59 FPS的速度达到74.3%的mAP，而更快的RCNN 7 FPS/mAP为73.2%或YLO 45 FPS/mAP为63.4%。

*角落网*最近，LawandDeng（2018）质疑锚箱在SoA对象检测框架中的主导作用（Girshick 2015；He等人2017；Redmon等人2016；Liu等人2016）。Law和Deng（2018）认为，锚箱的使用，特别是在一级探测器中（Fu等人2017；Lin等人2017b；Liu等人2016；Redmon等人2016）存在缺陷（Law和Deng 2018；Lin等人2017b），例如造成正面和负面示例之间的巨大不平衡，减缓训练和引入额外的超高参数。借鉴多人姿势估计中的关联嵌入工作（Newell等人，2017年）的思想，Law和Deng（2018年）通过将边界框对象检测公式化为检测成对的左上和右下关键点，提出了角网。在CornerNet中，主干网络由两个堆叠的沙漏网络组成（Newell et al.2016），通过一种简单的角点池方法可以更好地定位角点。CornerNet在MS COCO上实现了42.1%的AP，优于之前所有的单级探测器；然而，Titan X GPU上的平均推断时间约为4FPS，明显低于SSD（Liu等人2016）和YOLO（Redmon等人2016）。CornerNet生成不正确的边界框，因为很难确定哪些关键点对应分组到相同的对象中。为了进一步改进CornerNet，Duan等人（2019）提出了CenterNet，通过在提案中心引入一个额外的关键点，将MS COCO AP提高到[[14]](" \l "_ftn14" \o ")

47.0%，但推理速度比CornerNet慢。

# 6对象表示法

作为任何探测器的主要组件之一，良好的特征表示在目标检测中具有首要重要性（Dickinson等人，2009年；Girshick等人，2014年；Gidaris和Komodakis 2015年；Zhu等人，2016a）。过去，大量工作致力于设计局部描述符[例如，SIFT（Lowe 1999）和HOG（Dalal和Triggs 2005）]，并探索方法[例如，Bagowords（SivicandZisserman 2003）和Fisher向量（Perronnin et al.2010）]以进行分组和分类



**图15**2011年至2017年ILSVRC竞赛获奖作品在图像分类任务中的表现

将描述符抽象为更高层次的表示，以允许区分部分出现；然而，这些特征表示方法需要仔细的工程设计和相当多的领域专业知识。

相比之下，深度学习方法（尤其是深度CNN）可以直接从原始图像中学习具有多个抽象层次的强大特征表示（Bengio et al.2013；LeCun et al.2015）。由于学习过程减少了对特定领域知识和传统特征工程所需复杂过程的依赖（Bengio等人，2013年；LeCun等人，2015年），特征表示的负担已转移到设计更好的网络体系结构和培训程序上。

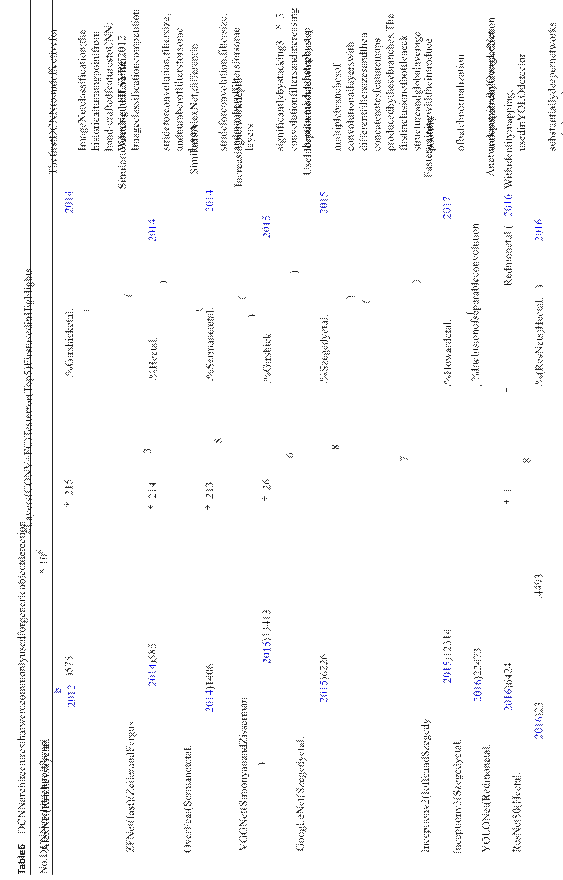
第节中回顾的主要框架。5[RCNN（Girshicketal.2014）、FastRCNN（Girshicketal.2015）、FasterRCNN（Renetal.2015）、YOLO（Redmonetal.2016）、SSD（Liuetal.2016）]一直在提高检测精度和速度，人们普遍认为CNN架构（第6.1节和图15）起着关键作用。因此，最近在检测精度方面的大多数改进都是通过研究新型网络的发展来实现的。因此，我们首先回顾了通用对象检测中使用的流行CNN体系结构，然后回顾了致力于改进对象特征表示的工作，例如开发不变特征以适应对象比例、姿势、视点和形状的几何变化，零件变形和执行多尺度分析，以改进大范围尺度上的对象检测。

## 6.1流行的CNN架构

CNNarchitectures（Sect.3）服务于ECT.5检测框架中使用的网络骨干。代表性的网络包括AlexNet（Krizhevsky等人2012b）、ZFNet（ZeilerandFergus2014）VGGNet（SimonyanandZisserman 2015）、GoogLeNet（Szegdy等人2015）、Inception series（Ioffeand Szegdy 2015；Szegdy等人2016、2017）、ResNet（He等人2016），DenseNet（Huang et al.2017a）和SENet（Hu et al.2018b）总结在表6中，图15显示了随时间的改善。Gu等人（2018年）对CNN最新进展进行了进一步回顾。

架构演进的趋势是更深入：AlexNet有8层，VGGNet有16层，最近ResNet和DenseNet都超过了100层，VGGNet（Simonyan和Zisserman 2015）和GoogLeNet（Szegedyetal.2015）表明，增加深度可以提高代表性。从表6可以看出，AlexNet、OverFeat、ZFNet和VGGNet等网络的参数数量巨大，尽管只有几层深，因为大部分参数来自FC层。较新的网络，如Inception、ResNet和DenseNet，虽然有很大的深度，但实际上通过避免使用FC层，参数要少得多。

通过在精心设计的拓扑中使用感应模块（Szegdy etal.2015），与AlexNet、ZFNet或VGNet相比，GoogLeNet的参数数量显著减少。同样，ResNet展示了SkipConnections学习具有数百层的极端深度网络的有效性，赢得了ILSVRC 2015分类任务。受ResNet（He et al.2016）的启发，InceptionResnet（Szegedy et al.2017）将初始网络与快捷连接相结合，因为快捷连接可以显著加快网络培训。Huang等人（2017a）对Resnet进行了扩展，提出了DenseNets，该DenseNets由密集区块构建，以前馈方式将每一层连接到其他每一层，从而带来了引人注目的优势，如参数效率、隐含的深度监控和功能重用。最近，他等人（2016）提出了挤压和激励（SE）块，可与现有的深层结构结合，以最小的额外计算成本提高其性能，通过明确建模卷积特征通道之间的相互依赖性，自适应地重新校准通道特征响应，从而赢得了ILSVRC 2017年分类任务。CNN架构的研究仍然活跃，新兴网络如沙漏（Law and Deng 2018）、扩张剩余网络（Yu等人2017）、Exception（Chollet 2017）、DetNet（Li等人2018b）、DualPathNetworks（DPN）（Chenetal.2017b）、FishNet（Sun等人2018）和GLoRe（Chen等人2019b）。[[15]](" \l "_ftn15" \o ")





CNN的训练需要具有组内多样性的大规模标记数据集。与图像分类不同，检测需要对罗马图像中的（可能有许多）对象进行定位。已经证明（Ouyang et al.2017b）使用具有对象级注释（如ImageNet）的大规模数据预训练深度模型，而不仅仅是图像级注释，可以提高检测性能。然而，收集边界框标签的成本很高，尤其是对于成千上万的类别。一个常见的场景是，CNN在带有图像级标签的大型数据集（通常具有大量视觉类别）上进行预训练；然后，预训练的CNN可以直接应用于一个小数据集，作为通用特征提取器（Razavian等人2014；Azizpour等人2016；Donahue等人2014；Yosinski等人2014），它可以支持更广泛的视觉识别任务。对于检测，预先训练的网络通常在给定的检测数据集上进行微调（Donahue等人2014；Girshick等人2014，2016）。CNN预训练使用了多个大规模图像分类数据集，其中ImageNet1000（Deng等人2009；Russakovsky等人2015）具有120万张1000个对象类别、位置的图像（Zhou等人2017a），比ImageNet1000大得多，但类数较少，是最近的Places Imagenet混合（Zhou等人2017a），或JFT3亿（Hinton等人，2015年；Sun等人，2017年）。[[16]](" \l "_ftn16" \o ")

Donahue et al.（2014）、Girshick et al.（2016）、Agrawal et al.（2014）对未经微调的预训练CNN进行了探索，发现从不同层面提取的特征的检测精度不同；例如，对于ImageNet上预训练的LexNet，FC6/FC7/Pool5的检测精度按降序排列（Donahue等人2014；Girshick等人2016）。微调预先训练的网络可以显著提高检测性能（Girshick et al.2014，2016），尽管在AlexNet的情况下，FC6/FC7的线调谐性能提升比Pool5大得多，这表明Pool5的功能更通用。此外，源数据集和目标数据集之间的关系显示出关键性的作用，例如，基于图像的CNN功能在目标检测方面比在人类行为方面表现出更好的性能（Zhou等人2015；Azizpour等人2016）。

## 6.2改进对象表示的方法

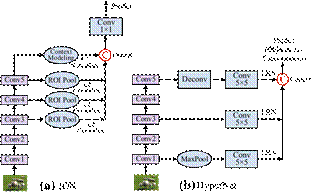
基于深度CNN的探测器，如RCNN（Girshick et al.2014）、Fast RCNN（Girshick 2015）、Faster RCNN（Ren et al.2015）和YOLO（Redmonetal.2016），通常将表6中列出的YUSETHEDEEP CNN架构作为主干网络，并使用CNN顶层的功能作为对象表示；然而，在大范围内检测物体是一个基本的挑战。解决这一问题的经典策略是在大量缩放输入图像（如图像金字塔）上运行检测器（Felzenszwalb等人2010b；Girshick等人2014；He等人2014），这通常会产生更准确的检测，但推理时间和内存明显受限。**6.2.1物体比例变化的处理**

由于CNN逐层计算其特征层次结构，特征层次结构中的子采样层已经形成固有的多尺度金字塔，以不同的空间分辨率生成特征地图，但面临挑战（Hariharan等人2016；Long等人2015；Shrivastava等人2017）。特别是，更高层具有更大的接受域和更强的语义，对物体姿态、照明和零件变形等变化最为鲁棒，但解决方案较低，几何细节丢失。相比之下，较低的层有一个小的感受野和丰富的几何细节，但分辨率很高，对语义的敏感性要低得多。直观地说，根据对象的大小，对象的语义概念可以出现在不同的层中。因此，如果目标物体很小，则需要在早期层中获得精细的细节信息，并且很可能在后期层中消失，原则上，这使得小物体检测非常具有挑战性，为此提出了一些技巧，如扩大或“萎缩”卷积（Yu和Koltun 2015；Dai et al.2016c；Chen et al.2018b），增加特征分辨率，但增加计算复杂性。另一方面，如果目标对象很大，那么语义概念将出现在更晚的层中。已经提出了许多方法（Shrivastava等人，2017；Zhang等人，2018e；Lin等人，2017a；Kong等人，2017），通过利用多个CNN层来提高检测精度，大致分为三种类型的多尺度目标检测：

1.         结合多层特征进行检测；

2.         多层检测；3.上述两种方法的组合。

（1） 利用多层膜的组合特征进行检测许多方法，包括超柱（Hariharan等人2016）、超网（Kong等人2016）和离子（Bell等人2016）在进行预测之前，结合多层膜的特征。这种特征组合通常通过串联来完成，串联是一种经典的神经网络思想，它将来自不同层的特征串联起来，这种结构最近在语义分割中变得很流行（Long等人2015；Shelhamer等人2017；Hariharan等人2016）。如图16a所示，ION（Bell等人，2016）使用RoI池从多个项目中提取RoI特征



**图16**超网和离子的比较。LRN是局部响应标准化，通过对局部输入区域进行标准化来实现某种“侧向抑制”（Jia等人，2014年）

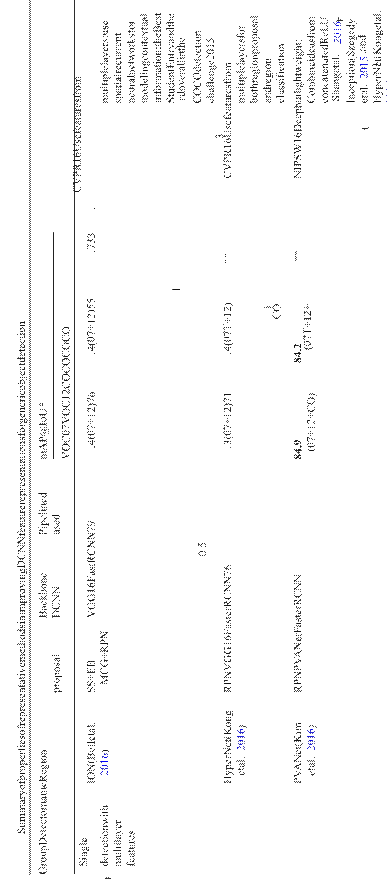
图层，然后通过选择性搜索和分类框生成的对象建议将使用相关特征进行分类。图16b所示的HyperNet（Kong et al.2016）遵循类似的思想，并集成了深度、中间和浅层特征，以生成目标建议，并通过端到端联合训练策略预测目标。组合特征更具描述性，更有利于定位和分类，但增加了计算复杂性。

（2） 最近的一些方法通过预测不同层上不同分辨率的对象，然后结合这些预测来改进检测：SSD（Liu等人2016）和MSCNN（Cai等人2016）、RBFNet（Liu等人2018b）和DSOD（Shen等人2017）。SSD（Liu et al.2016）将不同比例的默认框扩展到CNN中的多个层，并强制每个层专注于预测特定比例的对象。RFBNet（Liu等人，2018b）用感受野阻滞（RFB）代替SSD的后期卷积层，以增强特征的可辨别性和鲁棒性。RFB是一种多分支卷积块，类似于初始块（Szegedyetal.2015），但将具有不同内核和卷积层的多支路结合在一起（Chen等人，2018b）。MSCNN（Cai et al.2016）在使用层来学习区域建议和池特征之前，在CNN的多层上应用反褶积以提高特征地图分辨率。与RFBNet（Liu et al.2018b）类似，TridentNet（Li et al.2019b）构建了一个并行多分支架构，其中每个分支共享相同的转换参数，但具有不同的感受野；具有不同扩张率的扩张卷积用于适应不同尺度对象的感受野。*在多个CNN层上进行检测*

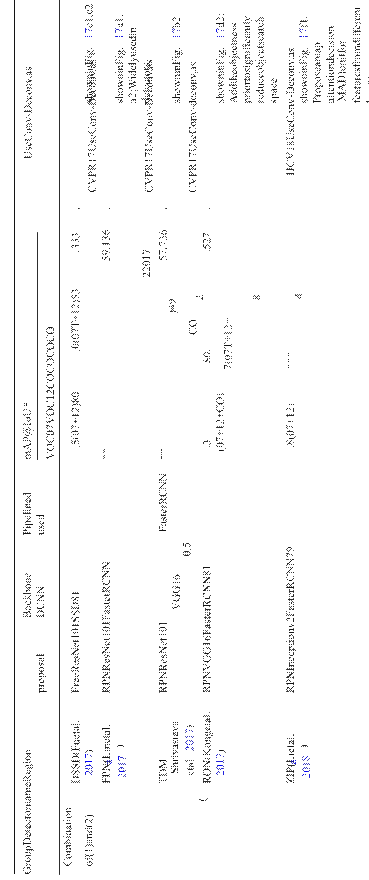
（3） 如超柱（Hariharan et al.2016）、超网（Kong et al.2016）和离子（Bell et al.2016）所示，不同层的特征相互补充，可以提高检测精度。然而，另一方面，使用大致相同大小的特征检测不同比例的对象是很自然的，这可以通过从缩小的特征地图检测大对象，同时从放大的特征地图检测小对象来实现。因此，为了将这两个世界的优点结合起来，最近的一些工作建议在多个层检测对象，并通过组合来自不同层的特征来获得结果特征。这种方法对于分割（Long等人，2015；Shelhamer等人，2017）和人体姿势估计（Newell等人，2016）是有效的，已被一级和两级检测器广泛利用，以缓解对象实例之间的尺度变化问题。代表性方法包括SharpMask（Pinheiro等人，2016年）、反褶积单次激发检测器（DSSD）（Fu等人，2017年）、特征金字塔网络（FPN）（Linetal.2017a）、上下调制（TDM）（Shrivastava等人，2017年）、与目标先验网络反向连接（RON）（Kong等人，2017年）、ZIP（Li等人，2018a），规模转移检测网络（STDN）（Zhou et al.2018b）、RefineDet（Zhang et al.2018a）、StairNet（Woo et al.2018）、路径聚合网络（PANet）（Liu et al.2018c）、特征金字塔重构（FPR）（Kong et al.2018）、DetNet（Li et al.2018b）、规模感知网络（SAN）（Kim et al.2018），多尺度位置感知内核表示（MLKP）（Wang等人，2018年）和M2Det（Zhao等人，2019年），如表7所示，并在图17中进行对比。*上述两种方法的组合*

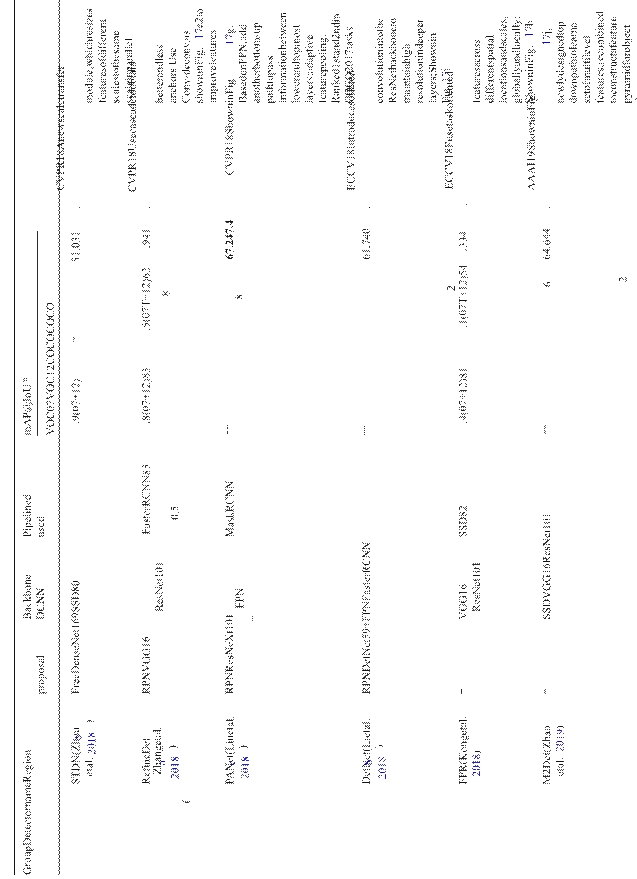
早期作品，如FPN（Lin等人2017a）、DSSD（Fu等人2017）、TDM（Shrivastava等人2017）、ZIP（Li等人2018a）、RON（Kong等人2017）和RefineDet（Zhang等人2018a），根据骨干的多尺度金字塔结构构建了天然吡虫啉，并取得了令人鼓舞的成果。从图17a1-f1可以看出，这些方法具有非常相似的检测体系结构，包括带有横向连接的自顶向下网络，以补充标准自底向上前馈网络。具体地说，在自底向上传递之后，最终的高层语义特征由自顶向下的网络传回，在横向处理后与来自中间层的自底向上特征组合，然后将组合的特征用于检测。从图17a2-e2可以看出，主要区别在于simpleFeatureFusionBlock（FFB）的设计，它处理不同层特征的选择和多层特征的组合。

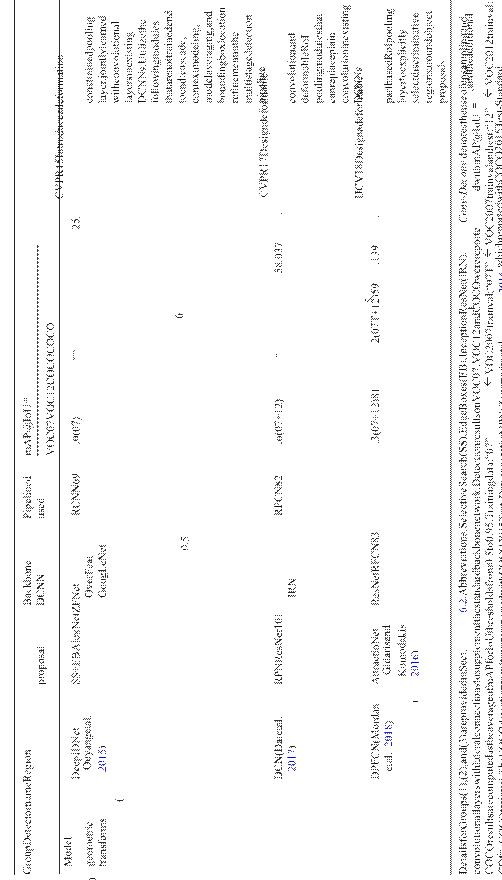
FPN（Lin et al.2017a）作为一种通用特征提取器，在包括目标检测（Lin et al.2017a，b）和实例分割（He et al.2017）在内的多个应用中表现出显著的改进。在基本的Faster RCNN系统中使用FPN在COCO检测数据集上取得了最新的结果。STDN（Zhou et al.2018b）使用DenseNet（Huang et al.2017a）组合不同层的特征，并设计比例转换模块以获得特征图











|  |
| --- |
|  |
|  |  |

**图17**沙漏架构：Conv1到Conv5是VGG或ResNet等骨干网络中的主要Conv块。该图比较了最近方法中常用的一些特征融合块（FFB）：FPN（Lin等人2017a）、TDM（Shrivastava等人2017）、DSSD（Fu等人2017）、RON（Kong等人2017）、RefineDet（Zhang等人2018a）、ZIP（Li等人2018a）、PANet（Liu等人2018c）、FPR（Kong等人2018）、DetNet（Li等人2018b）和M2Det（Zhao等人，2019）。FFM特征融合模块，TUM减薄U形模块

具有不同的分辨率。电子秤传输模块可直接嵌入DenseNet中，且无需额外成本。

如图17g-j所示，最近的工作，如PANet（Liu等人，2018c）、FPR（Kong等人，2018）、DetNet（Li等人，2018b）和M2Det（Zhao等人，2019），建议以不同的方式进一步改进类似FPN的金字塔结构。基于FPN，Liu等人设计了PANet（Liu等人，2018c）（图17g1）通过添加另一条自底向上的路径，该路径具有从低到高的干净横向连接，以缩短信息路径并增强特征金字塔。然后，提出了一种自适应特征池来聚合每个建议的所有特征级别的特征。此外，在建议子网中，一个互补分支captur为每个方案创建不同的视图，以进一步改进掩模预测。这些额外的步骤只会带来略微额外的计算开销，但非常有效，使PANet在COCO 2017挑战实例分割任务中获得第一名，在目标检测任务中获得第二名。Kong等人提出了FPR（Kong等人，2018年）通过明确重新制定特征金字塔构造过程[例如，FPN（Lin等人，2017a）]由于特征重构以一种高度非线性但高效的方式运行。如图17h1所示，FPR不是像FPN那样使用自上而下的路径从最顶层向下传播强语义特征，而是首先通过自适应级联从骨干网络的多个层提取特征，然后设计SamoreComplexFBModule（图17h2）将强语义扩展到所有尺度。Li等人（2018b）提出了DetNet（图17i1），方法是在主干网络的后几层引入扩展卷积，以保持深层的高空间分辨率。Zhao等人（2019）提出了多级特征金字塔网络（MLFPN）为了建立更有效的特征金字塔，用于检测不同比例的对象。如图17j1所示，首先将来自两个不同主干层的特征融合为基本特征，然后创建一条自上而下的路径，该路径带有来自基本特征的横向连接，以建立特征金字塔。如图17j2、j5、th所示FFB模块比FPN模块复杂得多，因为FFB包含一个减薄的U形模块（TUM）为了生成第二个金字塔结构，然后将来自多个TUM的具有等效大小的特征映射组合用于目标检测，作者通过将MLFPN集成到SSD中，提出了M2Det，实现了比其他单级检测器更好的检测性能。

## 6.3其他类内变化的处理

强大的对象表示应该结合显著性和鲁棒性。正如第6.2.1节所述，最近的大量工作致力于处理对象规模的变化。如第2.2节所述和图5所示，对象检测仍然需要对现实世界变化的鲁棒性，而不仅仅是我们分组的规模变化分为三类：

•几何变换，

•遮挡和•图像退化。

为了处理这些类内变化，最直接的方法是使用足够数量的变化来扩充训练数据集；例如，可以通过向训练数据中添加多个方向的旋转对象来实现对旋转的鲁棒性。鲁棒性通常可以通过这种方式学习，但通常以因此，研究人员提出了解决这些问题的替代方案。

*几何变换的处理*DCNN因缺乏对输入数据的几何变换保持空间不变的能力而受到严重限制（Lenc和Vedaldi，2018年；Liu等人，2017年；Chellappa，2016年）。局部最大池层的引入使DCNN具有一定的平移不变性，但中间特征映射实际上对输入数据的大型几何变换不具有不变性（Lenc和Vedaldi 2018）因此，已经提出了许多方法来增强稳健性，旨在学习不同类型转换的方差NN表示，如规模（Kim等人2014；Bruna和Mallat 2013）、旋转（Bruna和Mallat 2013；Cheng等人2016；Worrall等人2017；Zhou等人2017b）或两者（Jaderberg等人2015）.一项代表性工作是空间变换网络（STN）（Jaderberg等人，2015年），它引入了一个新的可学习模块，通过全局参数变换处理缩放、裁剪、旋转以及非刚性变形。STN现在已用于旋转文本检测（Jaderberg等人，2015年），旋转人脸检测和一般目标检测（Wang等人，2017年）。

尽管旋转不变性在某些应用中可能很有吸引力，例如场景文本检测（He等人2018；Ma等人2018）、人脸检测（Shi等人2018）和航空图像（Ding等人2018；Xia等人2018），但由于流行的基准检测数据集，通用对象检测工作聚焦于旋转不变性有限（例如PASCALVOC、ImageNet、COCO）不实际呈现旋转图像。

在深入学习之前，基于可变形零件的模型（DPM）（Felzenszwalb等人，2010b）成功地实现了通用对象检测，通过可变形配置中的组件来表示对象。虽然DPM的性能明显优于较新的对象检测器，但其精神仍然深刻地影响着许多最新的检测器。DPM建模对变换对象姿态、视点和非R不太敏感igiddeformations，激励研究人员（Dai等人2017年；Girshick等人2015年；Mordan等人2018年；Ouyang等人2015年；Wan等人2015年）明确建模对象组成，以改进基于CNN的检测。首次尝试（Girshick等人2015年；Wan等人2015年）通过使用AlexNet在基于DPM的检测中学习到的深度特征，将DPM与CNN结合起来，但没有区域建议。为了使CNN能够受益于建模对象零件变形的内置功能，提出了许多方法，包括DeepIDNet（Ouyang et al.2015）、DCN（Dai et al.2017）和DPFCN（Mordan et al.2018）（如表7所示）。尽管精神上相似，但变形的计算方式不同：DeepIDNet（Ouyang et al.2017b）设计了一个变形约束池层，以取代常规的最大池，以了解不同对象类的共享视觉模式及其变形特性；DCN（Dai et al.2017）设计了一个可变形卷积层和一个可变形RoI池层，这两个层都基于在特征地图中增加规则网格采样位置的想法；以及DPFCN（Mordan等人，2018年）提出了一种基于可变形部分的RoI池层，该层通过同时优化所有部分的潜在位移来选择对象提议周围对象的区分部分。

*闭塞的处理*在现实世界的图像中，遮挡很常见，导致对象实例的信息丢失。可变形部分的想法可用于遮挡处理，因此可变形RoI池（Dai等人2017；Mordanetal.2018；OuyangandWang2013）和可变形卷积（Dai等人2017）已提出通过赋予典型固定几何结构更大的灵活性来缓解遮挡。Wang et al.（2017）建议学习生成遮挡和变形示例的对抗性网络，而上下文可能有助于处理遮挡（Zhang et al.2018b）.尽管做出了这些努力，但遮挡问题远未得到解决；将GANs应用于该问题可能是一个有希望的研究方向。

*图像退化的处理*图像噪声是许多现实应用中的一个常见问题。它通常是由照明不足、低质量照相机、图像压缩或边缘设备和可穿戴设备上有意安装的低成本传感器引起的。虽然低图像质量可能会降低视觉识别的性能，但大多数现有方法都需要进行评估PASCAL VOC、ImageNet、MS COCO和Open Images都专注于相对高质量的图像，这一事实证明了无污染和清洁的环境。据我们所知，迄今为止，解决这一问题的工作非常有限。

# 7上下文建模

在物理世界中，视觉对象出现在特定的环境中，通常与其他相关对象共存。有强有力的心理学证据（Biederman 1972；Bar 2004）表明，语境在人类对象识别中起着至关重要的作用，人们认识到，恰当的语境建模有助于对象检测和识别（Torralba 2003；Oliva and Torralba 2007；Chen et al.2018b，2015a；Divvala et al.2009；Galleguillos and Belongie 2010），尤其是当对象外观特征因对象尺寸小、对象遮挡或图像质量差而不足时。已经讨论了许多不同类型的上下文（Divvala等人，2009年；Galleguillos和Belongie，2010年），可大致分为三类：

1.    语义语境：在某些场景中找到某个对象，但在其他场景中找不到该对象的可能性；

2.    空间背景：相对于场景中的其他对象，在某个位置找到某个对象而不是其他对象的可能性；

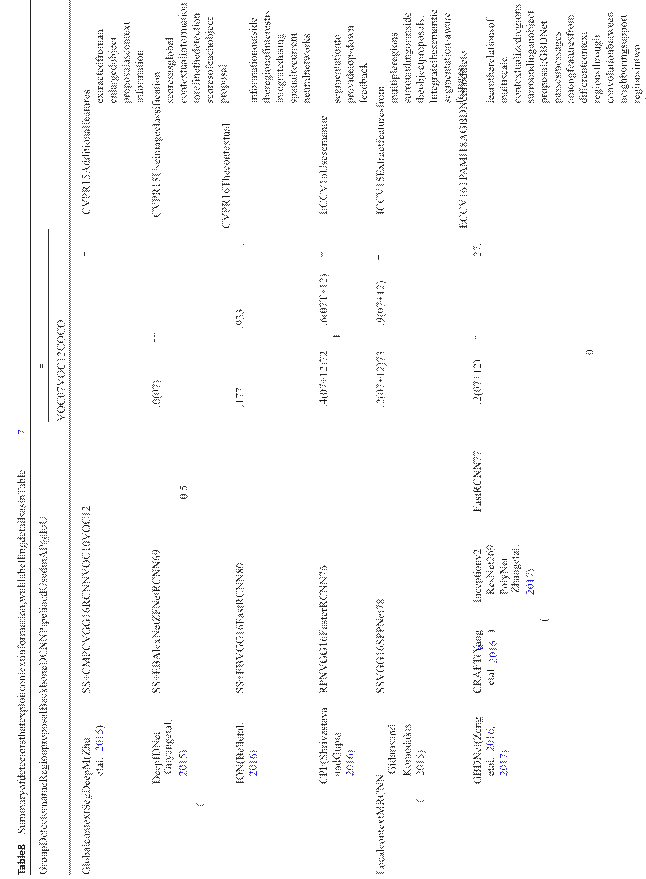
3.    缩放上下文：相对于场景中的其他对象，对象具有一组有限的大小。

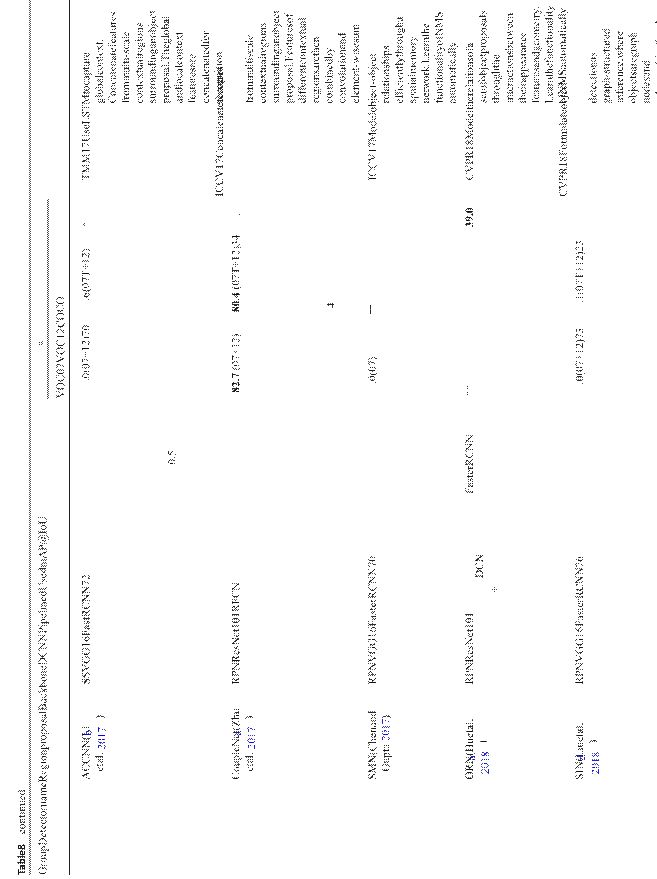
大量工作（Chen等人2015b；Divvala等人2009；Galleguillos和Belongie 2010；Malisiewicz和Efros 2009；Murphy等人2003；Rabinovich等人2007；Parikh等人2012）先于深度学习的普及，而这项工作中的大部分还没有在基于CNN的物体检测器上进行探索（Chen和Gupta 2017；Hu等人2018a）。

物体检测的当前技术水平（Ren等人2015；Liu等人2016；He等人2017）在不明确利用任何上下文信息的情况下检测物体。人们普遍认为DCNN隐含地利用上下文信息（Zeiler和Fergus 2014；Zheng等人2015）因为他们学习多层次抽象的层次表示法。然而，在基于DCNN的检测器中明确地描述上下文信息是有价值的（Hu等人2018a；Chen和Gupta 2017；Zeng等人2017）因此，以下回顾了在基于DCNN的对象检测器中利用上下文线索的最新工作，并将其分为类别FGlobalandLocalcontexts，由Zhang等人（2013）、Galleguillos和Belongie（2010）的早期工作推动。表8总结了代表性方法。

## 7.1全球环境

全局上下文（Zhang等人2013；Galleguillos和Belongie 2010）指图像或场景上下文，它们可以作为目标检测的线索（例如，卧室可以预测床的存在）。在DeepIDNet中（欧阳等人2015），图像分类分数被用作上下文特征，并与目标检测分数连接以改善检测结果。在ION（Bell等人，2016年）中，Bell等人提出使用空间递归神经网络（RNN）来探索整个图像的上下文信息。在Segdepm中（Zhu等人，2015年），Zhu等人提出了一个马尔可夫随机场模型，该模型对外观和上下文进行评分





对于每次检测，允许每个候选框从大量对象分割建议中选择一个片段，并对它们之间的一致性进行评分。在Shrivastava和Gupta（2016）中，语义分割被用作上下文启动的一种形式。

## 7.2地方背景

局部环境（Zhang等人2013；Galleguillos and Belongie 2010；Rabinovich等人2007）考虑了局部附近物体之间的关系，以及物体与其周围区域之间的相互作用。一般来说，对象关系建模具有挑战性，需要对不同类别、位置、比例等的边界框进行推理。明确建模对象关系的深入研究非常有限，代表性研究包括空间记忆网络（SMN）（Chen和Gupta 2017）、对象关系网络（Hu等人2018a），结构推理网络（SIN）（Liu等人，2018d）。在SMN中，空间内存基本上将对象实例重新组合成伪图像表示，该伪图像表示易于输入另一个CNN进行对象关系推理，从而形成一种新的顺序推理体系结构，其中图像和内存并行处理以获得检测，从而进一步更新内存。受自然语言处理中注意力模块最近取得的成功（Vaswani等人，2017）的启发，ORN通过外观特征和几何体之间的交互同时处理一组对象。它不需要额外的监督，并且易于嵌入现有网络，有效地改进了现代目标检测管道中的目标识别和重复消除步骤，从而产生了第一个完全端到端的目标检测器。SIN（Liu等人，2018d）考虑了两种上下文：场景上下文信息和单个图像中的对象关系。它将对象检测描述为一个图形推理问题，其中对象被视为图形中的节点，对象之间的关系被建模为边。

更广泛的方法用一个更简单的想法来解决上下文挑战：扩大检测窗口大小以提取某种形式的局部上下文。代表性方法包括MRCNN（Gidaris和Komodakis，2015年）、门控双向CNN（GBDNet）Zeng等人（2016年）、Zeng等人（2017年）、关注上下文CNN（ACCNN）（Li等人2017b）、对联网（Zhu等人2017a）和Sermanet等人（2013年）。在MRCNN（Gidaris和Komodakis 2015）（图18a）中，除了从主干最后一个CONV层的原始对象提案中提取的特征外，Gidaris和Komodakis还建议从对象提案的多个不同区域提取特征（半区域、边界区域、中心区域、上下文区域和语义分割区域），以获得更丰富和更健壮的对象表示。所有这些特征都通过串联组合。

自那时以来，已经提出了许多方法，都与RCNN密切相关。Zagoruyko等人（2016年）的方法仅使用了四个上下文区域，以中心凹结构组织，其中多条路径上的分类器端到端联合训练。Zeng等人（2016年），Zeng等人（2017年）提出了GBDNet（图18b）从对象周围的多尺度上下文化区域提取特征，以提高检测性能。与为每个区域分别学习CNN特征然后将其串联的有些天真的方法不同，GBDNet在不同上下文区域的特征之间传递消息。注意，消息传递Zeng等人（2016年）使用门控函数来控制消息传输。Li等人（2017b）提出了ACCNN（图18c），以利用全局和局部上下文信息：全局上下文是使用多尺度局部上下文（MLC）捕获的子网，当前生成输入图像的自然触角图，以突出有前景的上下文位置；本地上下文采用了类似于MRCNN的方法（Gidaris和Komodakis 2015）。如图18d所示，对联网（Zhu等人2017a）在概念上类似于ACCNN（Li等人2017b），但建立在RFCN（Dai等人2016c）的基础上，它通过位置敏感的RoI池捕获对象信息，耦合器添加了一个分支，用RoI池对全局上下文进行编码。

# 8检测建议方法

物体可以位于图像中的任何位置和比例。在手工特征描述符的全盛时期[SIFT（Lowe 2004）、HOG（Dalal和Triggs 2005）和LBP（Ojala et al.2002）]，最成功的物体检测方法[例如DPM（Felzenszwalb et al.2008）]使用了滑动窗口技术（Viola和Jones 2001；Dalal和Triggs 2005；Felzenszwalb等人2008；Harzallah等人2009；Vedaldi等人2009）。然而，窗口的数量是巨大的，随着图像中像素数量的增加而增加，并且需要以多个尺度和纵横比进行搜索，这进一步增加了搜索空间。因此，应用复杂分类器的计算成本太高。[[17]](" \l "_ftn17" \o ")

|  |
| --- |
| **图18**探索当地周边环境特征的代表性方法：MRCNN（Gidaris和Komodakis，2015年）、GBDNet（Zeng等人，2016年、2017年）、ACCNN（Li等人，2017b）和对联网（Zhu等人，2017a）；另见表8 |

2011年前后，研究人员提出通过使用检测方案缓解计算可处理性和高检测质量之间的紧张关系（Van de Sande et al.2011；Uijlings et al.2013）。源自Alexe et al.（2010）提出的对象性思想，对象建议是图像中可能包含对象的一组候选区域，如果可以使用少量的对象建议实现较高的对象召回率（如Undered），可以获得比滑动窗口方法更大的速度，从而允许使用更复杂的分类器。检测建议通常用作预处理步骤，限制需要检测器评估的区域数量，并且应具有以下特征：[[18]](" \l "_ftn18" \o ")

1.    高召回率，仅需几项建议即可实现；

2.    准确定位，使提案尽可能准确地匹配对象边界框；以及

3.    低计算成本。

基于探测方案的目标探测成功（Van de Sande et al.2011；Uijlings et al.2013）吸引了广泛的兴趣（Carreira和Sminchisescu 2012；Arbeláez et al.2014；Alexe et al.2012；Cheng et al.2014；Zitnick和Dollár 2014；Endres和Hoiem 2010；Krähenbühl和Koltun 2014；Manen et al.2013）.对目标提议算法的全面审查超出了本文的范围，因为目标提议的应用超出了目标检测（Arbeláez等人2012年；Guillaumin等人2014年；Zhu等人2017b）。我们请感兴趣的读者参考最近的调查（Hosangetal.2016；Chavalital.2016）其中深入分析了许多经典的对象建议算法及其对检测性能的影响。我们的兴趣在于回顾基于DCNNs、输出类不可知建议以及与通用对象检测相关的对象建议方法。

2014年，对象提案（Van de Sande et al.2011；Uijlings et al.2013）和DCNN特征（Krizhevsky et al.2012a）的整合导致里程碑式RCNN（Girshick et al.2014）在通用目标检测中。从那时起，检测方案很快成为标准的预处理步骤，因为自2014年以来，PASCAL VOC（Everingham et al.2010）、ILSVRC（Russakovsky et al.2015）和MS COCO（Lin et al.2014）目标检测挑战赛中的所有获奖项目都使用了检测方案（Girshick等人2014年；欧阳等人2015年；Girshick 2015年；任等人2015年；曾等人2017年；何等人2017年）。

在基于传统低级线索（例如颜色、纹理、边缘和梯度）的对象提议方法中，选择搜索（Uijlings等人2013年）、MCG（Arbeláez等人2014年）和EdgeBoxes（Zitnick和Dollár 2014年）更受欢迎。随着领域的快速发展，传统的对象提议方法（Uijlings et al.2013；Hosang et al.2016；Zitnick和Dollár 2014）被用作独立于探测器的外部模块，成为探测管道的速度瓶颈（Ren et al.2015）。一类新兴的目标提议算法（Erhan et al.2014；Ren et al.2015；Kuo et al.2015；Ghodrati et al.2015；Pinheiro et al.2015；Yang et al.2016a）使用DCNNs引起了广泛关注。

最近基于DCNN的对象建议方法通常分为两类：基于边界框的方法和基于对象段的方法，表9总结了具有代表性的方法。

Ren等人（2015）的RPC方法最能说明边界框建议方法，如图19所示。RPN通过在最后一个共享CONV层的特征图上滑动一个小网络来预测对象建议。在每个滑动窗口位置，使用k个锚框预测k个建议，其中每个锚框集中在图像中的某个位置，并与特定比例和纵横比关联。Ren等人（2015年）建议通过共享卷积层将RPN和Fast RCNN集成到单个网络中，从而实现Faster RCNN，这是第一条端到端检测管道。许多最先进的目标检测器广泛选择RPN作为建议方法，如表7和表8所示。[[19]](" \l "_ftn19" \o ")

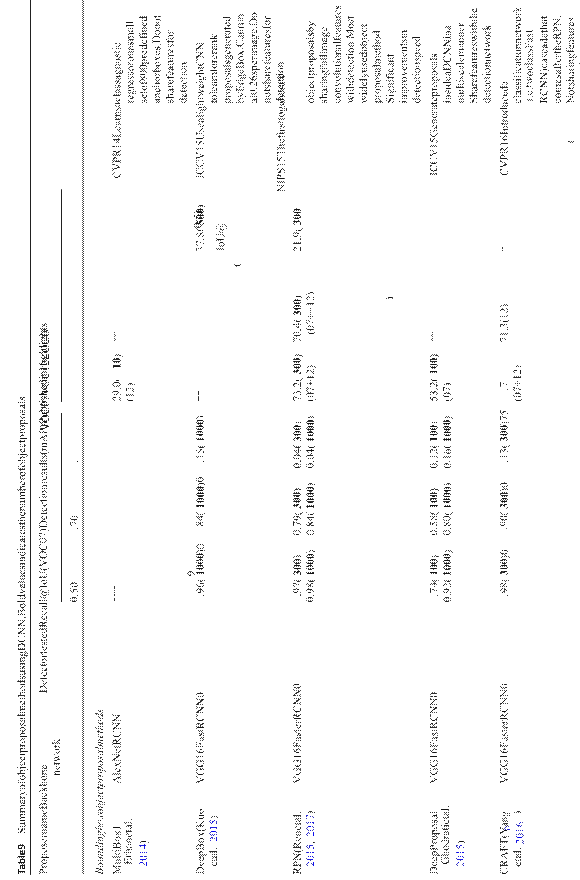
Lu等人（2016）没有将一组先验锚固定为多箱（Erhan等人2014；Szegedy等人2014）和RPN（Ren等人2015）建议使用递归搜索策略生成锚点位置，该策略可以自适应地引导计算资源关注可能包含对象的子区域。从整个图像开始，搜索过程中访问的所有区域都作为锚点。对于搜索过程中遇到的任何锚点区域，scalarzoomin指示器UsedToDecideWhether进一步划分区域，一组具有对象性分数的边界框由邻接和缩放网络（AZNet）计算，该网络通过添加分支来扩展RPN，以计算与现有分支并行的标量缩放指示器。

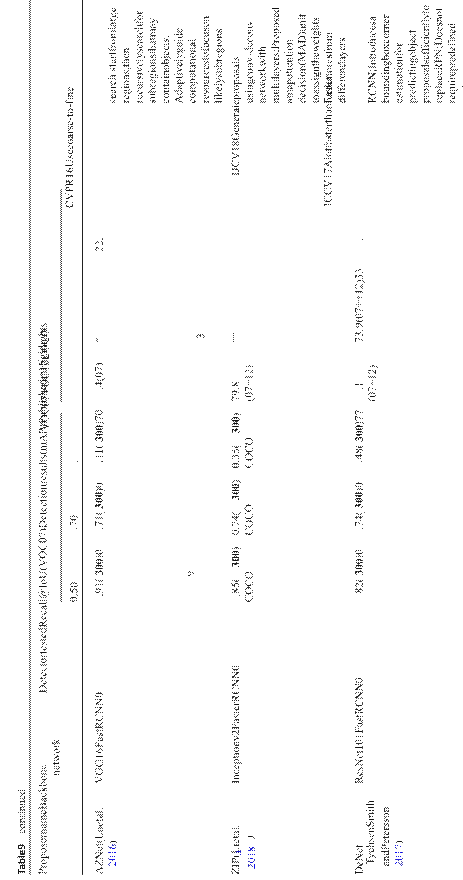
进一步的工作试图通过利用多层卷积特征生成对象建议。与RPN并行（Ren等人2015），Ghodrati等人（2015）提议的DeepProposal，通过使用多个卷积特征的级联生成对象提议，构建逆级联以选择最有希望的对象位置，并以从粗到精的方式细化其框。RPN的改进变体，HyperNet（Kong et al.2016）设计聚合多层卷积特征的超特征，并通过端到端联合训练策略在生成建议和检测对象时共享它们。Yang等人（2016a）提出的工艺也使用了级联策略，首先训练RPN网络生成目标建议，然后使用它们训练另一个二进制FastRCNN网络，以进一步区分背景中的目标。Li等人（2018a）建议使用ZIP来改进RPN，方法是在不同网络深度预测具有多个卷积特征映射的对象建议，以集成低级细节和高级语义。ZIP中使用的主干是一个“缩小和缩小”网络，其灵感来自conv和DECOV结构（Long等人，2015）。

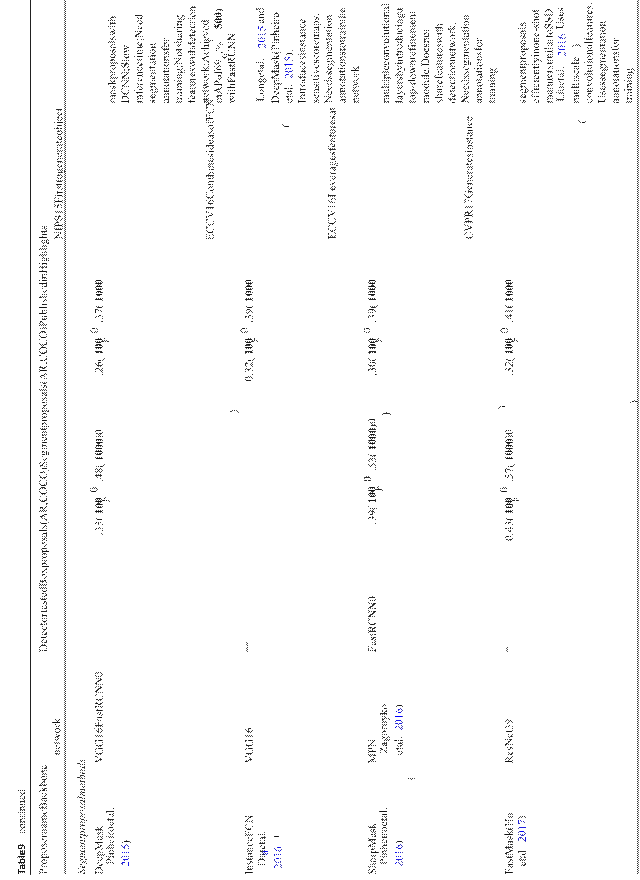
最后，值得一提的最新工作包括Deepbox（Kuo et al.2015），它提出了一种由EdgeBox生成的轻量级CNNTLearnToreRankProposals，以及DeNet（TychsensmithandPeterson2017），它引入了边界盒角点估计来有效地预测对象建议，以在更快的RCNN类型检测器中取代RPN。

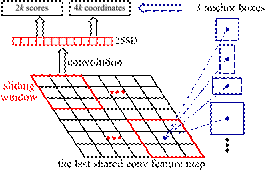
*对象段建议方法*Pinheiro等人（2015年），Pinheiro等人（2016年）的目标是生成可能与对象对应的分段方案。分段方案比边界框方案更具信息性，并朝着对象实例分段迈出了一步（Hariharan等人2014年；Dai等人2016b；Li等人2017e）此外，使用仪器分割监控可以提高包围盒目标检测的性能。Pinheiro等人提出的DeepMask的开创性工作。（2015），使用深度网络将直接从原始图像数据中学习到的建议分割。与RPN类似，在多个共享卷积层后，深度掩码将网络分割为两个分支，以预测类别不可知掩码和相关的对象性分数。也类似于OverFeat中的高效滑动窗口策略（Sermanet al.2014），经过训练的DeepMask网络在推理过程中以滑动窗口的方式应用于图像（及其重新缩放的版本）。最近，Pinheiro et al.（2016）提出了SharpMask，通过使用细化模块扩展DeepMask架构，类似于图17（b1）和（b2）中所示的架构，通过自上而下的细化过程增强前馈网络。SharpMask可以有效地集成早期特征的空间信息，这些信息具有强大的语义信息嵌入层，以生成高保真的对象掩码。

受语义分段的完全卷积网络（FCN）（Long等人2015）和DeepMask（Pinheiro等人2015；Dai等人2016a）的推动建议InstanceFCN生成实例段提案。与DeepMask类似，InstanceFCN网络被分为两个完全卷积的分支，一个用于生成实例敏感分数图，另一个用于预测对象性分数。Hu et al.（2017）提议的FastMask以一次性方式高效生成实例段提案，类似于SSD（Liu等人，2016年），以利用多尺度卷积特征。从多尺度卷积特征图中密集提取的滑动窗口被输入到一个尺度容忍注意头模块，以预测分割掩码和对象性得分。据称FastMask在800×600图像上以13 FPS的速度运行。









**图19**Ren等人（2015年）介绍的区域提案网络（RPN）说明

# 9其他问题

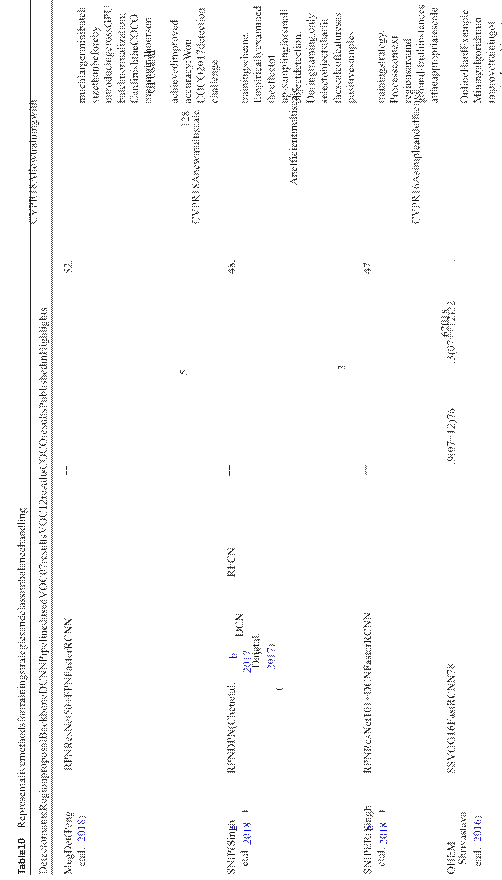
*数据扩充*为学习DCNNs执行数据扩充（Chatfield等人2014；Girshick 2015；Girshick等人2014）通常被认为对视觉识别很重要。琐碎的数据增强是指通过使基本类别保持不变的变换来扰动图像，例如裁剪、翻转、旋转、缩放、平移、颜色扰动和添加噪声。通过人为地扩大样本数，数据增强有助于ps用于减少过度拟合和提高泛化。它可以在训练时、测试时或两者同时使用。然而，它具有明显的局限性，即训练所需的时间显著增加。数据增强可能会合成全新的训练图像（Peng et al.2015；Wangetal.2017）然而，这些合成图像可以很好地概括为真实图像。一些研究人员（Dwibedi等人2017年；Gupta等人2016年）建议通过将真实分割对象粘贴到自然图像中来扩充数据集；事实上，Dvornik等人（2018年）证明了对对象周围的视觉上下文进行适当建模对于将其放置在正确的环境中至关重要，并提出了一种上下文模型来自动在图像上找到合适的位置以放置新的对象以进行数据扩充。

*新颖的培训策略*在范围广泛的尺度变化下检测物体，特别是检测非常小的物体，standsoutasakeychallenge.Ithasbeensown（Huang等人2017b；Liu等人2016）图像分辨率对检测精度有很大影响，因此缩放在数据增强中特别常用，因为更高的分辨率增加了检测小对象的可能性（Huang等人2017b）。最近，Singh等人提出了先进有效的数据论证方法SNIP（Singh和Davis 2018）和SNIPER（Singh et al.2018b）1举例说明了比例不变性问题，如表10所示。受小物体和大物体分别在较小和较大尺度下难以检测这一错误理解的激励，Snipin引入了一个水平序列-

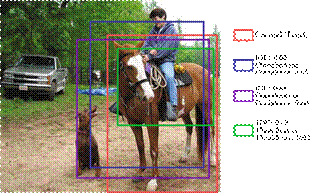
可以在训练过程中减少尺度变化，但不减少训练样本的ing方案；SNIPER允许进行有效的多尺度训练，仅以适当的尺度处理地面真实对象周围的上下文区域，而不是处理整个图像金字塔。Peng等人（2018年）研究了训练中的一个关键因素，即小批量大小，并提出了一种大型小批量对象检测器MegDet，使训练的小批量大小比以前大得多（从16到256）。为了避免收敛失败并显著加快训练过程，Peng et al.（2018）提出了学习速率策略和跨GPU批量标准化，有效利用了128个GPU，使MegDet能够在4小时内用128个GPU完成COCO培训，并赢得COCO 2017检测挑战赛。

*减少定位误差*在对象检测中，检测到的边界框与其地面真值框之间的联合上的相交（IOU）是最常用的评估指标，需要IOU阈值（例如，典型值0.5）来定义正和负。从图13可以看出，在大多数最先进的检测器中（Girshick 2015；Liu等人2016；He等人2017；Ren等人2015；Redmon等人2016）目标检测被描述为一个多任务学习问题，即，联合优化SoftMaxClassifier，它使用类标签和边界框回归器来指定目标方案，通过最大化IOU或检测结果与地面真实度之间的其他度量来定位目标。边界框只是铰接对象的粗略近似值，因此，背景像素几乎总是包含在边界框中，这会影响分类和定位的准确性显示除相似对象之间的融合外，对象定位误差是最具影响的MSOFERROR之一。定位误差可能源于重叠不足（小于要求的IOU阈值，如图20中的绿色框）或重复检测（即，一个对象实例的多个重叠检测）.通常，一些后处理步骤（如非最大值抑制（NMS）（Bodla等人2017；Hosang等人2017）用于消除重复检测。但是，由于未对准，在NMS期间可以抑制具有更好定位的边界框，导致定位质量较差（如图20所示的紫色框）因此，有很多方法旨在通过减少定位误差来提高检测性能。[[20]](" \l "_ftn20" \o ")

MRCNN（Gidaris和Komodakis 2015）引入了迭代边界盒回归，其中RCNN应用了多次。CRAFT（Yang等人2016a）和AttractioNet（Gidaris和Komodakis 2016）使用多阶段检测子网络生成准确的建议，以转发至Fast RCNN。Cai和Vascocelos（2018）提出了级联RCNN，它是RCNN的多级扩展，其中检测器序列是随着时间的增加而依次训练的







**图20**定位错误可能源于重叠或重复检测不足。定位错误是误报的常见原因（在线彩色图）

IOU阈值，基于以下观察结果：使用特定IOU训练的检测器的输出是一个很好的分布，用于训练下一个高层的检测器，以便对接近的误报具有更高的选择性。该方法可与任何基于RCNN的检测器一起构建，并证明可获得一致的增益（约2到4点）独立于基线检测器强度，计算量略有增加。最近也有工作（Jiang等人2018；Rezatofighi等人2019；Huang等人2019）将IOU直接制定为优化目标，并提出改进的NMS结果（Bodla等人，2017年；He等人，2019年；Hosang等人，2017年；TychsenSmith和Peterson，2018年），例如软NMS（Bodla等人，2017年）和学习NMS（Hosang等人，2017年）。

*类不平衡处理*与图像分类不同，对象检测还有另一个独特的问题：标记对象实例的数量与背景示例（不属于任何感兴趣对象类别的图像区域）的数量之间严重不平衡。大多数背景示例都是容易产生负面影响的，但是这种不平衡会使培训非常低效，并且大量容易产生负面影响往往会压倒培训。在过去，这个问题通常通过引导等技术来解决（Sung和Poggio，1994）。最近，这个问题也引起了一些关注（Li等人，2019a；Lin等人，2017b；Shrivastava等人，2016）。由于区域建议阶段快速过滤掉大部分背景区域，并提出少量候选对象，因此在两阶段检测器中（Girshick et al.2014；Girshick 2015；Ren et al.2015；He et al.2017）在一定程度上缓解了此类不平衡问题，尽管采用了示例挖掘方法，如在线硬示例挖掘（OHEM）（Shrivastava et al.2016）可用于维持前景和背景之间的合理平衡。对于单级目标探测器（Redmon et al.2016；Liu et al.2016），这种不平衡非常严重（例如，每个目标有100000个背景示例）。Lin et al.（2017b）建议通过校正交叉熵损失来解决这一问题，从而降低分配给正确分类示例的损失的权重。Li等人（2019a）从梯度范数分布的角度研究了这一问题，并提出了处理这一问题的梯度去谐化机制（GHM）。

# 10讨论和结论

通用目标检测是计算机视觉中一个重要且具有挑战性的问题，受到了广泛的关注。由于深度学习技术的显著发展，目标检测领域有了长足的发展。作为对通用目标检测的深度学习的一个全面综述，本文着重介绍了近年来层次，根据方法在检测中的作用，提供了方法的结构分类，总结了现有的流行数据集和评估标准，并讨论了表情表征方法的性能。我们在第10.1节中讨论了最新技术，在第10.2节中对关键问题进行了全面讨论，最后总结了本综述第10.3节提出了未来的研究方向。**10.1最先进的表演**

在过去几年中，出现了各种各样的检测器，并引入了标准基准，如PASCAL VOC（Everingham等人，2010年、2015年）、ImageNet（Russakovsky等人，2015年）和COCO（Lin等人，2014年），使比较探测器变得更容易。从我们之前在第5-9节中的讨论中可以看出，比较探测器最初报告的性能（例如精度、速度）可能会产生误导，因为它们可能在基本/上下文方面有所不同，包括以下选择：

•元检测框架，如RCNN（Girshick et al.2014）、Fast RCNN（Girshick 2015）、Fast RCNN（Ren et al.2015）、RFCN（Dai et al.2016c）、Mask RCNN（He et al.2017）、YOLO（Redmon et al.2016）和SSD（Liu et al.2016）；

•骨干网络，如表6中列出的VGG（Simonyan和Zisserman 2015）、Inception（Szegedy等人2015；Ioffe和Szegedy 2015；Szegedy等人2016）、ResNet（He等人2016）、ResNeXt（Xie等人2017）和Exception（Chollet 2017）等；

•创新，如多层特征组合（Lin et al.2017a；Shrivastava et al.2017；Fu et al.2017）、可变形卷积网络（Dai et al.2017）、可变形RoI池（Ouyang et al.2015；Dai et al.2017）、重型头部（Ren et al.2016；Peng et al.2018）和轻型头部（Li et al.2018c）；

•使用以下数据集进行预培训：ChasImageNet（Russakovsky等人，2015年）、COCO（Lin等人，2014年）、Places（Zhou等人，2017a）、JFT（Hinton等人，2015年）和Open Images（Krasin等人，2017年）；

•不同的检测建议方法和不同数量的目标建议；

•训练/测试数据扩充、新颖的多尺度训练策略（Singh和Davis 2018；Singh等人2018b）等，以及模型集成。

虽然比较每一个最近提出的探测器可能是不切实际的，但将具有代表性的和公开可用的探测器整合到一个共同的平台中并以统一的方式进行比较是有价值的。在这方面的工作非常有限，除了黄的研究（Huang et al.2017b）通过改变主干网络、图像分辨率和盒子数量，对三个主要探测器系列[Faster RCNN（Ren et al.2015）、RFCN（Dai et al.2016c）和SSD（Liu et al.2016）]进行分类。

从表7、8、9、10、11可以看出，我们总结了许多方法在三个广泛使用的标准基准上的最佳报告性能。这些方法的结果在相同的测试基准上报告，尽管它们在上述一个或多个方面存在差异。

图3和图21简要概述了最新技术，总结了PASCALVOC、ILSVRCA和MSCOCO挑战的最佳检测结果；更多结果可在检测挑战网站上找到（ILSVRC 2018；MS COCO 2018；PASCAL VOC 2018）。公开图像挑战目标检测任务的竞赛获胜者在公共黑板上获得61.71%的mAP，在私人领导板上获得58.66%的mAP，这是通过结合两个阶段的检测结果得出的，包括FAST RCNN（Girshick2015）、FasterRCNN（Renetal.2015）、FPN（Lin et al.2017a）、可变形RCNN（Dai et al.2017）总的来说，主干网、检测框架和大规模数据集的可用性是检测准确性的三个最重要因素。多个模型的集成、上下文特征的合并和数据扩充都有助于实现更高的准确性。

自AlexNet（Krizhevsky et al.2012a）被提出以来，在不到5年的时间里，1000个类别的ImageNet分类（Russakovsky et al.2015）的前5位误差从16%下降到了2%，如图15所示。然而，COCO上性能最好的检测器（Peng et al.2018）的地图（Lin et al.2014），仅能检测80类，仅为73%，即使在0.5 IoU的情况下也是如此，说明了目标检测比图像分类困难得多。最先进的检测器实现的准确性和鲁棒性远远不能满足实际应用的要求，因此仍有很大的改进空间。

## 10.2总结和讨论

在本文中讨论了数百篇参考文献和数十种方法之后，我们现在想重点讨论基于深度学习的通用目标检测中出现的关键因素。

（1） 检测框架：两阶段与一阶段

在第5节中，我们确定了两大类检测框架：基于区域的（两阶段）和统一的（一阶段）：

•当允许较大的计算成本时，两级检测器通常比一级检测器产生更高的检测精度，这一事实证明，在著名的检测挑战（如）中使用的大多数成功方法主要基于两级框架，因为它们的结构更灵活，更适合基于区域的分类最广泛使用的框架是快速RCNN（Ren等人2015）、RFCN（Dai等人2016c）和Mask RCNN（He等人2017）。

•Huang等人（2017b）的研究表明，单级SSD（Liu等人，2016年）的检测精度对主干网络质量的敏感性低于代表性的两级框架。

•一级检测器，如Eyolo（Redmonetal.2016）和SSD（Liu等人，2016）由于避免了预处理算法、使用轻量级骨干网络、使用较少的候选区域执行预测以及使分类子网络完全卷积，通常比两级检测器更快。但是，两级检测器可以通过引入类似技术实时运行。在任何情况下，无论是一个阶段还是两个阶段，最耗时的步骤是特征抽取器（主干网）（Law and Deng 2018；Ren et al.2015）。

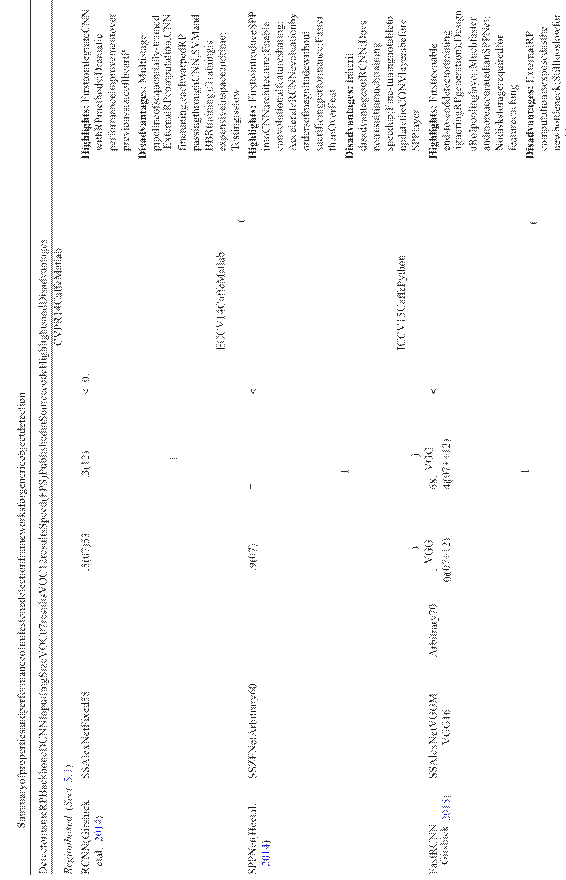
•已经证明（Huang等人2017b；Redmon等人2016；Liu等人2016），像Yoloa和SSDTypically这样的一级框架在检测小对象时比像更快的RCNN和RFCN这样的两级架构具有更高的性能，但在检测大对象时具有竞争力。

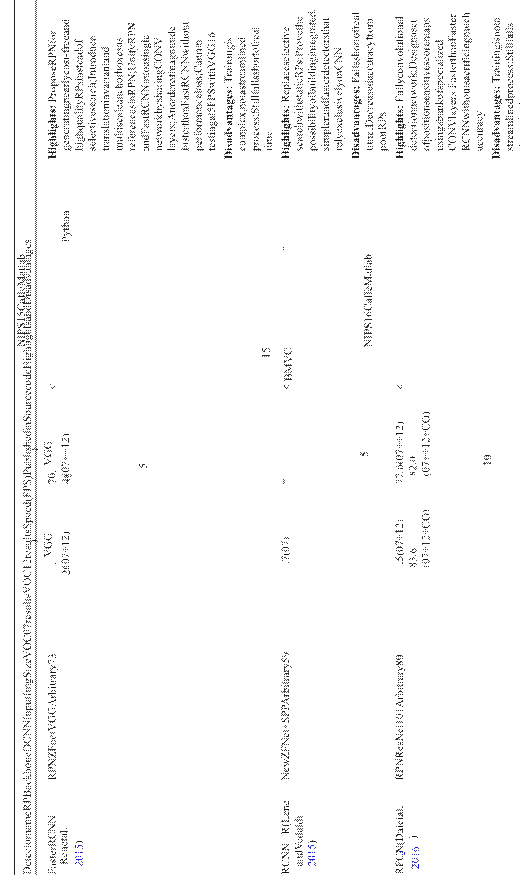
通过攻击检测框架的每个阶段，已经有许多人尝试构建更好（更快、更准确或更健壮）的检测器。无论是一个、两个还是多个阶段，检测框架的设计都向一些关键的设计选择靠拢：

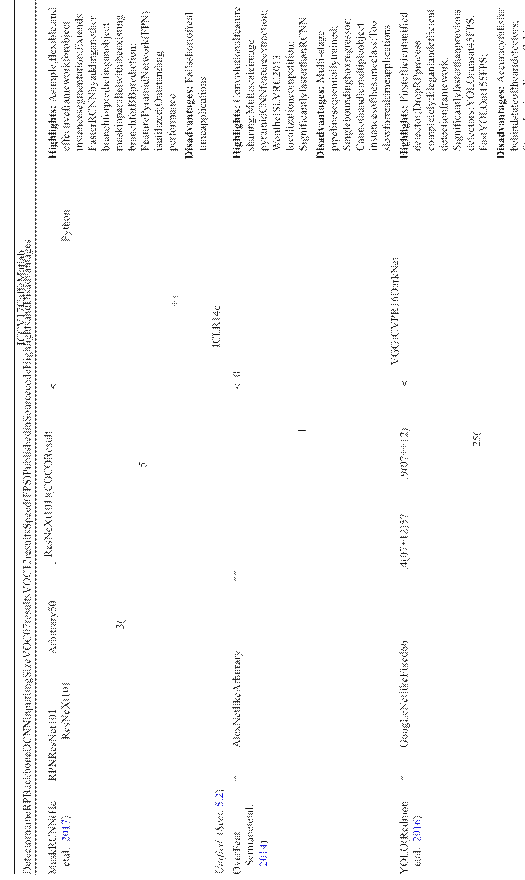
•完全卷积管道

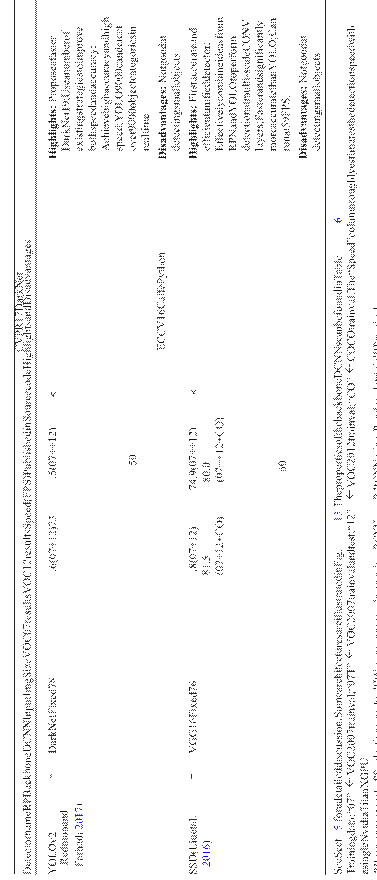
•探索来自其他相关任务的补充信息，例如面具RCNN（He等人，2017年）

•滑动窗（Ren等人，2015年）









•融合主干网不同层的信息。

最近成功的CASCADEFORObjectDetection（Cai和Vasconcelos 2018；Cheng等人，2018a，b）和COCO实例分割（Chen等人，2019a）以及其他挑战的证据表明，多级目标检测可能是速度-精度权衡的未来框架。2019年更广泛的挑战中正在进行一项挑逗性调查（Loy等人，2019年）。

（2） 骨干网

如第6.1节所述，主干网络是快速提高检测性能的主要驱动力之一，因为区分性对象特征表示发挥了关键作用。通常，较深的主干网络，如ResNet（He等人2016）、ResNeXt（Xie等人2017）、InceptionResNet（Szegedy等人2017）性能更好；但是，它们的计算成本更高，并且需要更多的数据和大量的训练计算。一些主干网（Howardetal.2017；Iandolaetal.2016；Zhang et al.2018c）被提议将重点放在速度上，例如MobileNet（Howard et al.2017）)已经证明，在ImageNet上仅通过计算成本和模型大小即可实现VGGNet16精度。随着更多的训练数据和更好的训练策略可用，从头开始骨干训练可能成为可能（Wu和He2018年；罗等人，2018年、2019年）。

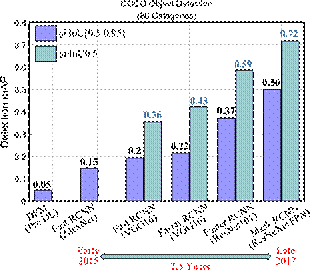
（3） 提高对象表示的鲁棒性

现实世界图像的变化给物体识别带来了挑战。这些变化包括照明、姿势、变形、背景杂波、遮挡、模糊、分辨率、噪声和相机变形。

（3.1）对象比例和小对象尺寸

物体尺度的巨大变化，特别是小物体的变化，带来了巨大的挑战。这里总结和讨论了第6.2节中确定的主要策略：•使用图像金字塔：它们简单有效，有助于放大小物体和缩小大物体。它们的计算成本很高，但仍然很复杂通常在推理过程中使用，以提高准确性。

•使用来自不同分辨率卷积层的特征：在早期的工作中，如SSD（Liu等人，2016），预测是独立进行的，没有来自其他层的信息被组合或合并。现在，组合来自不同层的特征是相当标准的，例如在FPN中（Lin等人，2017a）。



**图21**目标检测性能的演变（测试开发结果）。结果引自（Girshick 2015；He等人2017；Ren等人2017）。主干网络、检测框架的设计以及良好和大规模数据集的可用性是检测准确性的三个最重要因素

•使用扩张卷积（Li等人，2018b，2019b）：一种简单有效的方法，用于整合更广泛的背景并维持高分辨率特征图。

•使用不同比例和长宽比的锚箱：具有许多参数的缺点，锚箱的比例和长宽比通常是试探性确定的。

•放大：特别是对于小型物体的检测，可以开发高分辨率网络（Sun等人，2019a，b）。目前尚不清楚超分辨率技术是否能提高检测精度。

尽管最近取得了一些进展，但小物体的检测精度仍远低于大物体的检测精度。因此，小物体的检测仍然是物体检测的关键挑战之一。可能需要将定位要求概括为一个尺度函数，因为某些应用，例如自动驾驶，只有需要识别较大区域内小物体的存在，不需要精确定位。（3.2）变形、遮挡等因素

如第2.2节所述，处理几何变换、遮挡和变形的方法主要基于两种范式。第一种是空间变换网络，它使用回归获得变形场，然后根据变形场扭曲特征（Dai等人，2017）第二种是基于可变形零件的模型（Felzenszwalb等人2010b），该模型在考虑空间约束的情况下发现零件过滤器的最大响应（Ouyang等人2015；Girshick等人2015；Wan等人2015）。

旋转不变性在某些应用中可能很有吸引力，但由于流行的基准检测数据集（PASCALVOC、ImageNet、COCO），因此专注于旋转不变性的通用对象检测工作有限旋转变化不大。遮挡处理在人脸检测和行人检测中得到了深入的研究，但对于一般对象检测的遮挡处理研究很少。总体而言，尽管最近取得了一些进展，深度网络仍然受到对许多变化缺乏鲁棒性的限制，这些变化表明限制了它们在现实世界中的应用。

（4） 上下文推理

如第7节所述，野外物体通常与其他物体和环境共存。人们已经认识到，上下文信息（物体关系、全局场景统计）有助于物体检测和认知（Oliva和Torralba，2007年），特别是对于小物体、遮挡物体和图像质量差的物体。在深度学习之前有大量的工作（Malisiewicz和Efros 2009；Murphy等人2003；Rabinovich等人2007；Divvala等人2009；Galleguillos和Belongie 2010），在深度学习时代也有不少工作（Gidaris和Komodakis 2015；Zengetal.2016，2017；ChenandGupta2017；Huetal.2018a）。如何高效和有效地整合上下文信息有待探索，可能以人类视觉如何使用上下文为指导，基于场景图（Li等人2017d），或通过使用全景分割对对象和场景进行完整分割（Kirillov等人，2018年）。

（5） 探测建议

检测方案显著减少了搜索空间。正如Hosang等人（2016）所建议的那样，未来的检测方案肯定必须在重复性、召回率、定位精度和速度方面进行改进。自RPN成功以来（Ren等人，2015年），它将建议生成和检测集成到一个通用框架中，基于CNN的检测建议生成方法在区域建议中占主导地位。建议对新的检测建议进行对象检测评估，而不是单独评估检测建议。

（6） 其他因素

如第9节所述，影响目标检测质量的还有许多其他因素：数据增强、新颖的训练策略、主干模型的组合、多个检测框架、合并其他相关任务的信息、减少定位误差的方法、处理正负样本之间的巨大不平衡性、，挖掘硬阴性样本，改进损失函数。

## 10.3研究方向

尽管最近在目标检测领域取得了巨大的进步，但这项技术仍然比人类视觉要原始得多，还不能令人满意地解决现实世界中的挑战，如方法论。2.2.我们看到了许多长期存在的挑战：

•在开放世界中工作：对任何数量的环境变化都具有鲁棒性，能够进化或适应。•受限条件下的对象检测：从标记较弱的数据或少量边界框注释、可穿戴设备、看不见的对象类别等学习。

•其他模式下的目标检测：视频、RGBD图像、3D点云、激光雷达、遥感图像等。

基于这些挑战，我们看到了以下未来研究方向：

(1)最终目标是开发能够准确、高效地识别和定位开放世界场景中数千个或更多对象类别中的实例的对象检测，其水平与人类视觉系统相当。一般来说，对象检测算法无法识别外部的对象类别r训练数据集，但理想情况下应具有识别新对象类别的能力（Lake等人2015；Hariharan和Girshick 2017）。当前检测数据集（Everingham等人2010；Russakovsky等人2015；Lin等人2014）仅包含几十到数百个类别，远远少于人类可以识别的类别。需要开发具有更多类别的新的更大规模数据集（Hoffman等人2014；Singh等人2018a；Redmon和Farhadi 2017）。*开放世界学习*

（2） 通用目标检测成功的原因之一是开发了高级检测框架，包括基于区域的[RCNN（Girshicketal.2014）、FastRCNN（Girshick 2015）、快速RCNN（Ren et al.2015）、Mask RCNN（He et al.2017）]和一级检测器[YOLO（Redmon et al.2016）、SSD（Liu et al.2016）].基于区域的检测器具有更高的精确度，单级检测器通常更高级和更简单。对象检测器严重依赖于底层主干网络，这些主干网络已针对图像分类进行了优化，可能会导致学习偏差；从头开始学习对象检测器可能有助于新的检测框架。*更好、更高效的检测框架*

（3） CNN的深度显著增加，从几层[AlexNet（Krizhevsky et al.2012b）]到数百层[ResNet（He et al.2016），DenseNet（Huang et al.2017a）]。这些网络有数百万到数亿个参数，需要大量数据和GPU进行训练。为了减少或消除网络冗余，设计紧凑、轻量级网络的研究兴趣越来越大（Chen等人2017a；Alvarez和Salzmann 2016；Huang等人2018；Howard等人2017；Lin等人2017c；Yu等人2018）和网络加速（Cheng等人2018c；Hubara等人2016；Han等人2016；Li等人2017a，c；Wei等人2018）。*紧凑高效的CNN功能*

（4）深度学习绕过人工特征工程，这需要具有较强领域知识的人类专家，然而DCNNs需要类似的显著的专门知识。很自然地考虑检测骨干架构的自动化设计，例如最近的自动机器学习（AutoML）（QuunMin等人，2018）。，已应用于图像分类和对象检测（Caiteal.2018；Chenetal.2019c；Ghiasietal.2019；Liuetal.2018a；ZophandLe2016；Zoph等人，2018）。*自动神经结构搜索*

（5） 为了更丰富、更详细地理解图像内容，需要解决像素级对象实例分割问题（Lin等人2014；He等人2017；Hu等人2018c），这在需要单个对象精确边界的潜在应用中可以发挥重要作用。*对象实例分割*

（6） 当前最先进的探测器采用从带有对象边界框或分割遮罩的标记数据中学习的全监督模型（Everingham等人2015；Lin等人2014；Russakovsky等人2015；Lin等人2014）。然而，全监督学习具有严重的局限性，特别是在边界框注释的收集是劳动密集型的，并且图像数量很大的情况下。在没有完全标记的培训数据的情况下，完全监督学习是不可扩展的，因此有必要了解在仅提供弱/部分注释数据的情况下如何利用CNN的力量（Bilen和Vedaldi 2016；Diba等人2017；Shi等人2017）。*弱监督检测*

（7） 深度探测器的成功在很大程度上依赖于大量带注释的训练数据。当标记数据稀少时，深度探测器的性能经常恶化，无法很好地推广。相比之下，人类（甚至儿童）可以从极少的示例中快速学习视觉概念，并且通常可以很好地概括（Biederman 1987b；Lake et al.2015；FeiFei et al.2006）。因此，仅从几个示例学习的能力，如fewshotdetection，非常引人注目*少镜头/零镜头目标检测*

（Chen等人，2018a；Dong等人，2018；Finn等人，2017；Kang等人，2018；Lake等人，2015；Ren等人，2018；Schwartz等人，2019）。更受约束的零炮目标检测定位和识别以前从未见过的目标类别（Bansal等人2018；Demirel等人2018；Rahman等人2018b，a），这对于需要智能和增量发现新目标类别的终身学习机器至关重要。[[21]](" \l "_ftn21" \o ")

（8） 大多数探测器基于静止的2D图像；其他模式下的目标检测在自主车辆、无人机和机器人等领域具有高度相关性。这些模式在有效利用深度（Chen等人2015c；Pepik等人2015；Xiang等人2014；Wu等人2015）、视频（Feichtenhofer等人2017；Kang等人2016）和点云（Qi等人2017、2018）方面提出了新的挑战。*其他模式中的目标检测*

（9） 最近，人们在学习通用表示法方面的努力不断增加，这些表示法在多个图像领域中都是有效的，如自然图像、视频、航空图像和医学CT图像（Rebuffi et al.2017、2018）。大多数此类研究侧重于图像分类，很少针对目标检测（Wang et al.2019），开发的检测器通常是特定领域的。独立于图像域的目标检测和跨域目标检测是未来的重要方向。*通用目标检测：*

通用目标检测的研究领域还很不完善。然而，考虑到过去5年的突破，我们对未来的发展和机遇持乐观态度。

**致谢**奥卢大学提供的开放获取资金，包括奥卢大学医院。作者要感谢通用物体检测和其他相关领域的先驱研究人员。作者还将向副主编吉里马塔斯教授和匿名评论员表达他们的兴趣，感谢他们的评论和建议。这项工作得到了奥卢大学（芬兰）的机器视觉和信号分析中心和中国国家自然科学基金61872379号资助。

**开放存取**本文根据知识共享署名4.0国际许可的条款分发(http://creativecomm ons.org/licenses/by/4.0/），允许在任何媒体中不受限制地使用、分发和复制，前提是您给予原作者和来源适当的信任，提供到知识共享许可证的链接，并说明是否进行了更改。

# 工具书类

Agrawal，P.，Girshick，R.，和Malik，J.（2014）。分析了用于目标识别的多层神经网络的性能。在ECCV中（第329-344页）。

Alexe，B.，Deselaers，T.，和Ferrari，V.（2010）。什么是物体？在CVPR中（第73-80页）。

Alexe，B.，Deselaers，T.，和Ferrari，V.（2012）。测量图像窗口的对象性。IEEE TPAMI，34（11），2189–2202。

Alvarez，J.，和Salzmann，M.（2016）。学习深层网络中神经元的数量。在NIPS中（第2270-2278页）。

Andreopoulos，A.，和Tsotsos，J.（2013）。物体识别50年：前进方向。计算机视觉与图像理解，117（8），827–891。

阿尔贝雷斯，P.，哈里哈兰，B.，顾，C.，古普塔，S.，布尔德夫，L.，和马利克，J.（2012）。使用区域和部分进行语义分割。在CVPR中（第3378-3385页）。

Arbeláez，P.，Pont Tuset，J.，Barron，J.，Marques，F.，和Malik，J.（2014）。多尺度组合分组。在CVPR中（第328-335页）。

阿齐兹普尔，H.，拉扎维，A.，沙利文，J.，马基，A.，和卡尔森，S.（2016）。通用convnet表示的可转移性因素。IEEE TPAMI，38（9），1790-1802。

Bansal，A.，Sikka，K.，Sharma，G.，Chellappa，R.，和Divakaran，A.（2018）。零炮目标检测。在ECCV中。

Bar，M.（2004）。视觉对象背景。自然评论神经科学，5（8），617-629。

贝尔，S.，劳伦斯，Z.，巴拉，K.，和吉尔希克，R.（2016）。内外网：使用跳转池和递归神经网络在上下文中检测对象。在CVPR中（第2874-2883页）。

Belongie，S.，Malik，J.，和Puzicha，J.（2002）。使用形状上下文的形状匹配和对象识别。IEEE TPAMI，24（4），509–522。

Bengio，Y.，Courville，A.，和Vincent，P.（2013）。代表性学习：回顾和新视角。IEEE TPAMI，35（8），1798-1828。

比德曼，I.（1972年）。感知真实世界的场景。IJCV，177（7），77-80。

比德曼，I.（1987a）。组件识别：人类图像理解的理论。《心理学评论》，94（2），115。

比德曼，I.（1987b）。组件识别：人类图像理解的理论。《心理学评论》，94（2），115。

Bilen，H.，和Vedaldi，A.（2016）。弱监督深度检测网络。在CVPR中（第2846-2854页）。

博德拉，N.，辛格，B.，切拉帕，R.，戴维斯L.S.（2017）。SoftMS使用一行代码改进目标检测。在ICCV中（第5562-5570页）。

Borji，A.，Cheng，M.，Jiang，H.，和Li，J.（2014）。显著目标检测：调查，1，1-26。arXiv:1411.5878v1。

Bourdev，L.，和Brandt，J.（2005）。通过软级联实现鲁棒目标检测。CVPR，2236-243。

Bruna，J.，和Mallat，S.（2013）。不变散射卷积网络。IEEE TPAMI，35（8），1872-1886年。

蔡，Z.，和Vasconcelos，N.（2018）。级联RCNN：深入研究高质量目标检测。在CVPR中。

蔡，Z.，范，Q.，费里斯，R.，和瓦斯康塞洛斯，N.（2016）。用于快速目标检测的统一多尺度深卷积神经网络。在ECCV中（第354-370页）。

蔡，H.，杨，J.，张，W.，韩，S.，和于，Y.等。（2018）有效架构搜索的路径级网络转换。在ICML中。

Carreira，J.，和Sminchisescu，C.（2012）。CMPC：使用约束参数化最小割集的自动对象分割。IEEE TPAMI，34（7），1312–1328。

查特菲尔德，K.，西蒙扬，K.，韦达尔迪，A.，和齐瑟曼，A.（2014）。细节中的魔鬼归来：深入研究卷积网络。在BMVC中。

查瓦利，N.，阿格拉瓦尔，H.，马亨德鲁，A.，和巴特拉，D.（2016）。对象建议评估协议是可玩的。在CVPR中（第835-844页）。

Chellappa，R.（2016）。模式识别和计算机视觉不断变化的命运。图像和视觉计算，55，3-5。

陈庚，蔡文伟，余，X，韩，T和钱德拉克姆（2017a）。利用知识提取学习有效的目标检测模型。在紧要关头。

陈浩，王，Y，王，G.，和乔，Y.（2018a）。LSTD：一种用于目标检测的低激发转移检测器。在AAAI。

陈坤，庞，J.，王，J.，熊，Y.，李，X.，孙，S.，冯，W.，刘，Z.，史，J.，欧阳，W.，等（2019a）。用于实例分割的混合任务级联。在CVPR中。

Chen，L.，Papandreou，G.，Kokkinos，I.，Murphy，K.，和Yuille，A.（2015a），使用深度卷积网和全连接CRF的语义图像分割。在ICLR。

陈，L.，帕潘德里欧，G.，科基诺斯，I.，墨菲，K.，和尤耶，A.（2018b）。DeepLab：使用深度卷积网、Atrus卷积和完全连接的CRF进行语义图像分割。IEEE TPAMI，40（4），834–848。

陈秋松，宋，Z.，董，J.，黄，Z.，华，Y.，燕，S.（2015b）。背景化目标检测和分类。IEEE TPAMI，37（1），13–27。

陈克星和古普塔，A.（2017）。目标检测中上下文推理的空间记忆。在ICCV中。

Chen，X.，Kundu，K.，Zhu，Y.，Berneshawi，A.G.，Ma，H.，Fidler，S.，和Urtasun，R.（2015c）精确对象类检测的3d对象建议。在NIPS中（第424-432页）。

陈勇、李俊杰、肖浩、金某、闫南、冯杰（2017b）。双路径网络。在NIPS中（第4467-4475页）。

陈，Y.，罗巴赫，M.，燕，Z.，燕，S.，冯，J.，和Kalantis，Y.（2019b），基于图的全球推理网络。在CVPR中。

陈妍，杨婷，张，X，孟，G，潘，C.，孙，J.（2019c）。DetNAS：目标检测的神经结构搜索。

arXiv:1903.10979。

程斌、魏玉英、石海华、费里斯、熊建和黄泰泰（2018a）。解耦分类细化：用于目标检测的硬假阳性抑制。arXiv:1810.04002。

程斌、魏玉英、史海华、费里斯、熊强、黄婷（2018b）。重温RCNN：唤醒更快RCNN的分类能力。在ECCV中。

程国刚，周培平，韩建军（2016）。RIFDCNN：用于目标检测的旋转不变和fisher判别卷积神经网络。在CVPR中（第2884-2893页）。

程明明，张志忠，林文和托尔，P.（2014）。BING：用于以300fps速度估计对象性的二值化赋范梯度。在CVPR中（第3286-3293页）。

程勇，王德德，周平，张婷（2018c）。深层神经网络的模型压缩和加速：原理、进展和挑战。IEEE信号处理杂志，35（1），126–136。

Chollet，F.（2017）。例外：深度可分离卷积的深度学习。在CVPR中（第1800-1807页）。

辛比斯，R.，韦贝克，J.，和施密德，C.（2017）。基于多重多实例学习的弱监督目标定位。IEEE TPAMI，39（1），189–203。

苏尔卡，G.，舞蹈，C.，范，L.，威拉莫夫斯基，J.，和布雷，C.（2004）。使用关键点包进行视觉分类。在ECCV计算机视觉统计学习研讨会上。

戴，J.，何，K.，李，Y.，任，S.，和孙，J.（2016a）。实例敏感的完全卷积网络。在ECCV中（第534-549页）。

戴，J.，何，K.，和孙J.（2016b）。通过多任务网络级联实现实例感知语义分割。在CVPR中（第3150-3158页）。

戴，J.，李，Y.，何，K.，孙，J.（2016c）。RFCN：通过基于区域的完全卷积网络进行目标检测。在NIP中（第379-387页）。戴，J.，齐，H.，熊，Y.，李，Y.，张，G.，胡，H.，魏，Y.（2017）。可变形卷积网络。在ICCV中。

Dalal，N.，和Triggs，B.（2005）。用于人体检测的定向梯度直方图。CVPR，1886-893。

Demirel，B.，Cinbis，R.G.，和Ikizler Cinbis，N.（2018）。混合区域嵌入的零镜头目标检测。在BMVC中。

邓，J.，董，W.，索契，R.，李，L.，李，K.，和李，F.（2009）。ImageNet:Alargescalehierarchicalimagedatabase.InCVPR（第248-255页）。Diba，A.，Sharma，V.，Pazandeh，A.M.，Pirsiavash，H.，和Van Gool L.（2017）。弱监督级联卷积网络。在CVPR中（第3卷，第9页）。

狄金森，S.，莱昂纳迪斯，A.，希勒，B.，和塔尔，M.（2009）。对象分类的演变和对象分类中图像抽象的挑战：计算机和人类视觉视角。剑桥：剑桥大学出版社。

丁，J.，薛，N.，龙，Y.，夏，G.，和陆，Q.（2018）。航空图像中定向目标检测的学习。在CVPR中。

Divvala，S.，Hoiem，D.，Hays，J.，Efros，A.，和Hebert，M.（2009）。目标检测中上下文的实证研究。在CVPR中（第1271-1278页）。

Dollar，P.，Wojek，C.，Schiele，B.，和Perona，P.（2012）。行人检测：对最新技术的评估。IEEE TPAMI，34（4），743–761。

唐纳休，J.，贾，Y.，维尼亚斯，O.，霍夫曼，J.，张，N.，曾，E.，等（2014）。DECF：一种用于一般视觉识别的深度卷积激活特征。ICML，32647–655。

董，X.，郑，L.，马，F.，杨，Y.，和孟，D.（2018）。使用模型通信的少数示例对象检测。IEEE模式分析和机器智能学报，41（7），1641–1654。

段坤、白南山、谢丽玲、齐海浩、黄秋和田秋秋（2019）。CenterNet：用于对象检测的关键点三元组。arXiv:1904.08189。

德沃尼克，N.，迈拉尔，J.，和施密德，C.（2018）。可视化上下文建模是扩充目标检测数据集的关键。在ECCV中（第364-380页）。

Dwibedi，D.，Misra，I.，和Hebert，M.（2017）。剪切、粘贴和学习：非常简单的合成，例如检测。在ICCV中（第1301-1310页）。

Endres，I.，和Hoiem，D.（2010）。与类别无关的对象建议。在K.Danilidis、P.Maragos和N.Paragios（编辑）的《欧洲计算机视觉会议》（第575-588页）。柏林：斯普林格。

Enzweiler，M.，和Gavrila，D.M.（2009）。单目行人检测：调查和实验。IEEE TPAMI，31（12），2179–2195。

Erhan，D.，Szegedy，C.，Toshev，A.，和Angelov，D.（2014）。使用深度神经网络的可伸缩目标检测。在CVPR中（第2147-2154页）。

Everingham，M.，Eslami，S.，Gool，L.V.，Williams，C.，Winn，J.，和Zisserman，A.（2015）。pascal视觉对象类挑战：回顾。IJCV，111（1），98-136。

Everingham，M.，Gool，L.V.，Williams，C.，Winn，J.，和Zisserman，A.（2010）。pascal可视对象类（voc）挑战。IJCV，88（2），303-338。

Feichtenhofer，C.，Pinz，A.，和Zisserman，A.（2017）。检测到跟踪和跟踪到检测。在ICCV中（第918-927页）。

菲菲，L.，弗格斯，R.，和佩罗纳，P.（2006）。对象类别的一次性学习。IEEE TPAMI，28（4），594–611。

费尔森斯瓦尔布，P.，吉尔希克，R.，和麦卡莱斯特，D.（2010a）。使用变形部件模型进行级联对象检测。InCVPR（第2241-2248页）。

费尔森斯瓦尔布，P.，吉尔希克，R.，麦卡莱斯特，D.，和拉马南，D.（2010b）。基于区分训练的零件模型的目标检测。IEEE TPAMI，32（9），1627–1645。

Felzenszwalb，P.，McAllester，D.，和Ramanan，D.（2008）。经过鉴别训练的多尺度可变形零件模型。在CVPR中（第1-8页）。

Finn，C.，Abbeel，P.，和Levine，S.（2017年）。模型不可知元学习用于深度网络的快速适应。在ICML中（第1126-1135页）。

费施勒，M.，和埃尔施拉格，R.（1973）。图形结构的表示和匹配。计算机上的IEEE交易，100（1），67–92。

Fu，C-Y.，Liu，W.，Ranga，A.，Tyagi，A.，和Berg，A.C.（2017）。DSSD：反褶积单发探测器。arXiv:1701.06659。

Galleguillos，C.，和Belongie，S.（2010）。基于上下文的对象分类：一个重要的综述。计算机视觉和图像理解，114712–722。

Geronimo，D.，Lopez，A.M.，Sappa，A.D.，和Graf，T.（2010）。高级驾驶员辅助系统的行人检测调查。IEEE TPAMI，32（7），1239–1258。

贾斯，G.，林，T.，庞，R.，乐，Q.（2019）。NASFPN：学习用于目标检测的可伸缩特征金字塔结构。

arXiv:1904.07392。

戈德拉蒂，A.，迪巴，A.，彼得索利，M.，图伊特拉尔斯，T.，和范古尔，L.（2015）。DeepProposal：通过级联深层卷积层来搜索对象。在ICCV中（第2578-2586页）。

吉达里斯和科莫达基斯，N.（2015）。Objectdetectionviaamultiregion and SemanticSegmentationAwareCnnModel.InICCV（第1134-1142页）。

吉达里斯和科莫达基斯，N.（2016）。参与优化重复：通过输入输出本地化生成活动箱提案。在BMVC中。

Girshick，R.（2015年）。快R-CNN。在ICCV中（第1440-1448页）。

吉尔希克，R.，多纳休，J.，达雷尔，T.，和马利克，J.（2014）。丰富的特征层次结构，用于精确的对象检测和语义分割。在CVPR中（第580-587页）。

Girshick，R.，Donahue，J.，Darrell，T.，和Malik，J.（2016）。用于精确目标检测和分割的基于区域的卷积网络。IEEE TPAMI，38（1），142–158。

Girshick，R.，Iandola，F.，Darrell，T.，和Malik，J.（2015）。可变形零件模型是卷积神经网络。在CVPR中（第437-446页）。

古德费罗，I.，本吉奥，Y.，和库维尔，A.（2016）。深度学习。剑桥：麻省理工学院出版社。

古德费罗，I.，Shlens，J.，和Szegedy，C.（2015）。解释和利用对抗性例子。在ICLR。

格雷曼，K.，和达雷尔，T.（2005）。金字塔匹配核：具有图像特征集的鉴别分类。ICCV，21458-1465。

格雷曼，K.，和莱贝，B.（2011）。视觉物体识别。人工智能和机器学习综合讲座，5（2），1-181。

顾，J.，王，Z.，权，J.，马，L.，沙鲁迪，A.，帅，B.，等（2018）。进化神经网络中的新进展。模式识别，77354-377。

纪尧明，M.，库特尔，D.，和法拉利，V.（2014）。带有分段传播的Imagenet自动注释。国际计算机视觉杂志，110（3），328-348。

Gupta，A.，Vedaldi，A.，和Zisserman，A.（2016）。用于自然图像中文本定位的合成数据。在CVPR中（第2315-2324页）。

韩S.，戴利，W.J.，和毛H.（2016）。深度压缩：通过剪枝、训练量化和哈夫曼编码压缩深度神经网络。在ICLR。

Hariharan，B.，Arbeláez，P.，Girshick，R.，和Malik，J.（2014）。同时检测和分割。在ECCV中（第297-312页）。

Hariharan，B.，Arbeláez，P.，Girshick，R.，和Malik，J.（2016）。使用超列进行对象实例分割和细粒度定位。IEEE模式分析和机器智能学报，39（4），627–639。

Hariharan，B.和Girshick R.B.（2017）。通过缩小和产生幻觉的特征进行低镜头视觉识别。在ICCV中（第3037-3046页）。

Harzallah，H.，Jurie，F.，和Schmid，C.（2009）。结合物体定位和图像分类的效率。在ICCV中（第237-244页）。

He，K.，Gkioxari，G.，Dollár，P.，和Girshick，r.（2017）。屏蔽RCNN。在ICCV中。

何坤，张，X，任，S，孙，J.（2014）。用于视觉识别的深卷积网络中的空间金字塔池。在ECCV中（第346-361页）。

何坤，张，X，任，S和孙，J.（2015）。深入研究整流器：在ImageNet分类上超越人类水平的性能。在ICCV中（第1026-1034页）。

何坤，张，X，任，S，孙，J.（2016）。深度残差学习用于图像识别。在CVPR中（第770-778页）。

何，T.，田，Z.，黄，W.，沈，C.，乔，Y.，和孙，C.（2018）。一个具有明确对齐和注意力的端到端文本观察者。在CVPR中（第5020-5029页）。

何勇，朱，C.，王，J.，萨维德斯，M.，张，X.（2019）。具有不确定性的包围盒回归，用于精确的目标检测。在CVPR中。

辛顿，G.，和萨拉胡季诺夫，R.（2006）。利用神经网络降低数据的维数。《科学》，313（5786），504-507。

Hinton，G.，Vinyals，O.，和Dean，J.（2015）。在神经网络中提取知识。arXiv:1503.02531。

霍夫曼，J.，瓜达拉马，S.，曾，E.S.，胡，R.，多纳休，J.，吉尔希克，R.，达雷尔，T.，和萨恩科，K.（2014）。LSDA：通过自适应进行大规模检测。在NIPS中（第3536-3544页）。

Hoiem，D.，Chodpathumwan，Y.，和Dai，Q.（2012）。目标探测器中的错误诊断。在ECCV中（第340-353页）。

Hosang，J.，Benenson，R.，DolláR，P.，和Schiele，B.（2016）。什么是有效的检测建议？IEEE TPAMI，38（4），814–829。

Hosang，J.，Benenson，R.，和Schiele，B.（2017）。学习非最大值抑制。在ICCV中。

Hosang，J.，Omran，M.，Benenson，R.，和Schiele，B.（2015）。深入观察行人。IEEE计算机视觉和模式识别会议记录（第4073-4082页）。

Howard，A.，Zhu，M.，Chen，B.，Kalenichenko，D.，Wang，W.，Weyand，T.，Andreetto，M.，和Adam，H.（2017）。Mobilenets：用于移动视觉应用的高效卷积神经网络。在CVPR中。

胡浩，顾，J.，张，Z.，戴，J.，魏，Y.（2018a）。用于目标检测的关系网络。在CVPR中。

胡浩、兰南、蒋云、曹中、沙福（2017）。快速蒙版：在一次拍摄中分割多尺度候选对象。在CVPR中（第991-999页）。胡杰、沈磊和孙庚（2018b）。挤压和激励网络。在CVPR中。

胡，P.，和拉马南，D.（2017）。寻找微小的面孔。在CVPR中（第1522-1530页）。

胡，R.，多拉尔，P.，何，K.，达雷尔，T.，和吉尔希克，R.（2018c）。学习分割每件事。在CVPR中。

黄，G.，刘，S.，范德马腾，L.，和温伯格，K.（2018）。concedenet：一种使用学习组卷积的有效densenet。在CVPR中。

黄，G.，刘，Z.，温伯格，K.Q.，和范德马腾，L.（2017a）。密集连接的卷积网络。在CVPR中。

黄，J.，拉托德，V.，孙，C.，朱，M.，科拉提卡拉，A.，法蒂，A.，菲舍尔，I.，沃伊纳，Z.，宋，Y.，瓜达拉马，S.，和墨菲，K.（2017b）。现代卷积目标探测器的速度/精度权衡。在CVPR中。

黄，Z.，黄，L.，龚，Y.，黄，C.，和王，X.（2019）。面具评分rcnn。在CVPR中。

I.胡巴拉、M.库巴里奥、D.苏德里、R.埃尔亚尼夫和Y.本吉奥（2016年）。二值化神经网络。在NIPS中（第4107-4115页）。

伊恩多拉，F.，汉，S.，莫斯科维茨，M.，阿什拉夫，K.，达利，W.，和库策，K.（2016）。挤压网：Alexnet级精度，参数少50倍，模型尺寸少0.5 mb。arXiv:1602.07360。

ILSVRC检测挑战结果。(2018). http://www.image-net. org/challenges/LSVRC/。

Ioffe，S.，和Szegedy，C.（2015）。Batchnormalization：通过减少内部协变量转移来加速深度网络训练。在国际机器学习会议上（第448-456页）。

Jaderberg，M.，Simonyan，K.，Zisserman，A.，等人（2015年）。空间变压器网络。在NIPS中（2017-2025页）。

Jia，Y.，Shelhamer，E.，Donahue，J.，Karayev，S.，Long，J.，Girshick，R.，Guadarrama，S.，和Darrell，T.（2014）。Caffe：用于快速特征嵌入的卷积结构。在ACM-MM中（第675-678页）。

蒋，B.，罗，R.，毛，J.，肖，T.，和蒋，Y.（2018）。获取精确目标检测的定位置信度。在ECCV中（第784-799页）。

康，B.，刘，Z.，王，X.，余，F.，冯，J.，和达雷尔，T.（2018）。通过特征重加权进行少镜头目标检测。arXiv:1812.01866。

康康，欧阳，W.，李，H.，和王，X.（2016）。基于卷积神经网络的视频小结节目标检测。在CVPR中（第817-825页）。

Kim，A.，Sharma，A.，和Jacobs，D.（2014）。局部尺度不变卷积神经网络。在紧要关头。

Kim，K.，Hong，S.，Roh，B.，Cheon，Y.，和Park，M.（2016）。PVANet：用于实时目标检测的深度但轻量级的神经网络。在NIPSW。

Kim，Y，Kang，B-N.，和Kim，D.（2018）。SAN：用于多尺度目标检测的卷积特征之间的学习关系。在ECCV中（第316-331页）。

基里洛夫，A.，他，K.，吉尔希克，R.，罗瑟，C.，和多拉，P.（2018）。全景分割。arXiv:1801.00868。

孔T、孙F、谭C、刘H和黄W.（2018）。用于目标检测的深度特征金字塔重构。在ECCV中（第169-185页）。

孔，T.，孙，F.，姚，A.，刘，H.，陆，M.，陈，Y.（2017）。RON：与objectness Previor网络反向连接，用于目标检测。在CVPR中。

孔T，姚A，陈Y，孙F.（2016）。HyperNet：实现精确的区域建议生成和联合目标检测。在CVPR中（第845-853页）。

Krähenbühl，P.，和Koltun，V.（2014年），测地物体提案。在ECCV中。

I.克拉辛、T.杜里格、N.奥尔德林、V.法拉利、S.阿布海亚、A.库兹涅佐娃等（2017年）。OpenImages：用于大规模多标签和多类图像分类的公共数据集。数据集可从https://storage.googleapis.com/openimages/web/ index.html。

Krizhevsky，A.，Sutskever，I.，和Hinton，G.（2012a）。使用深度卷积神经网络进行图像网络分类。在NIPS中（第1097-1105页）。

Krizhevsky，A.，Sutskever，I.，和Hinton，G.（2012b）。使用深度卷积神经网络进行图像网络分类。在NIPS中（第1097-1105页）。

Kuo，W.，Hariharan，B.，和Malik，J.（2015）。DeepBox：卷积网络的学习对象性。在ICCV中（第2479-2487页）。

库兹涅佐娃，A.，罗姆，H.，奥尔德林，N.，尤伊林斯，J.，克拉辛，I.，蓬图塞特，J.，等人（2018年）。开放图像数据集v4：统一图像分类、对象检测和VisualRelationshipDetectionAtScale。arXiv:1811.00982。

Lake，B.，Salakhutdinov，R.，和Tenenbaum，J.（2015）。通过概率程序归纳法进行人类层面的概念学习。科学，350（6266），1332-1338。

Lampert，C.H.，Blaschko，M.B.，和Hofmann，T.（2008）。超越滑动窗口：通过有效子窗口搜索的对象定位。在CVPR中（第1-8页）。

罗，H.，邓，J.（2018）。CornerNet：将对象检测为成对的关键点。在ECCV中。

Lazebnik，S.，Schmid，C.，和Ponce，J.（2006）。功能包之外：空间金字塔匹配用于识别信号场景类别。CVPR，2169-2178。

LeCun，Y.，Bengio，Y.，和Hinton，G.（2015）。深度学习。自然杂志，521436-444。

LeCun，Y.，Bottou，L.，Bengio，Y.，和Haffner，P.（1998）。梯度学习在文档识别中的应用。IEEE会议录，86（11），2278–2324。

李春春、谢淑、加拉赫、P、张、Z、图、Z.（2015）。严密监督的渔网。《人工智能与统计》（第562-570页）。

Lenc，K.，和Vedaldi，A.（2015）。R-CNN减去BMVC15中的R。

Lenc，K.，和Vedaldi，A.（2018）。通过测量图像表示的等变性和等同性来理解图像表示。在IJCV中。

李斌、刘妍、王小新（2019a）。梯度协调单级检测器。在AAAI。

李，H.，卡达夫，A.，杜达诺维奇，I.，萨米特，H.，和格拉夫，H.P.（2017a）。修剪过滤器以获得高效的网络。在ICLR。

李浩，林中正，沈晓霞，勃兰特，J.，和华，G.（2015a）。一种用于人脸检测的卷积神经网络级联算法。在CVPR中（第5325-5334页）。

李浩、刘尧、欧阳、王伟、X.（2018a）。使用MAPATTENTIONDECISIONFOR Region Proposal对象检测缩小和缩小网络。在IJCV中。

李杰，魏勇，梁，X，董，J，徐，T，冯，J.等（2017b）。AttentionContextsObjectDetection.IEEETransactionMultimedia，19（5），944-954。

李Q，靳，S.，和闫，J.（2017c）。模拟非常有效的目标检测网络。在CVPR中（第7341-7349页）。

李世忠、张志忠（2004）。Floatboost学习和统计人脸检测。IEEE TPAMI，26（9），1112–1123。

李彦宏，陈彦宏，王彦宏，张志强（2019b）。用于目标检测的缩放感知三叉戟网络。arXiv:1901.01892。

李彦，欧阳，W.，周，B.，王，K.，和王，X.（2017d）。从对象、短语和区域标题生成场景图。在ICCV中（第1261-1270页）。

李彦宏、齐海宏、戴俊杰、季小新、魏彦宏（2017e）。完全卷积的实例感知语义分割。在CVPR中（第4438-4446页）。

李勇，王绍，田，Q.，丁，X.（2015b）。基于统计学习的目标检测的特征表示：综述。模式识别，48（11），3542-3559。

李志强、彭聪、于庚、张晓阳、邓耀阳、孙杰（2018b）。DetNet：用于目标检测的骨干网络。在ECCV中。

李志强、彭聪、于国光、张晓阳、邓云、孙杰（2018c）。光头RCNN：用于防御两级目标探测器。在CVPR中。

林，T.，多拉，P.，吉尔希克，r.，何，K.，哈里哈兰，B.，和贝隆吉，S.（2017a）。用于目标检测的特征金字塔网络。在CVPR中。

林，T.，戈亚尔，P.，吉尔希克，R.，何，K.，和多拉，P.（2017b）。密集目标检测的焦点损失。在ICCV中。

林，T.，梅尔，M.，贝隆吉，S.，海斯，J.，佩罗纳，P.，拉马南，D.，多拉尔，P.，和齐特尼克，L.（2014）。Microsoft COCO：上下文中的常见对象。在ECCV中（第740-755页）。

林，X.，赵，C.，潘，W.（2017c）。朝向精确的二元卷积神经网络。在NIPS中（第344-352页）。

Litjens，G.，Kooi，T.，Bejnordi，B.，Setio，A.，Ciompi，F.，Ghafoorian，M.，et al.（2017年）。医学图像分析深度学习研究综述。医学图像分析，42，60-88。

刘，C.，佐夫，B.，诺依曼，M.，施伦斯，J.，华，W.，李，L.，菲菲，L.，尤尔，A.，黄，J.，和墨菲，K.（2018a）。渐进式神经结构搜索。在ECCV中（第19-34页）。

刘，L.，费古思，P.，郭，Y.，王，X.，和Pietikäinen，M.（2017）。纹理分类的局部二元特征：分类和实验研究。模式识别，62135-160。

刘绍，黄，D.，和王，Y.（2018b）。接收区块网，用于准确快速的目标检测。在ECCV中。

刘绍，齐，L.，秦，H.，史，J.，贾，J.（2018c）。用于实例分段的路径聚合网络。在CVPR中（第8759-8768页）。

刘伟、安圭洛夫、二汉、塞格迪、里德、傅、C和贝格（2016）。SSD：单激发多盒探测器。在ECCV中（第21-37页）。

刘尧、王瑞山、陈新山（2018d）。结构推理网：使用场景级上下文和实例级关系进行对象检测。在CVPR中（第6985-6994页）。

Long，J.，Shelhamer，E.，和Darrell，T.（2015）。用于语义分割的完全卷积网络。IEEE计算机视觉和模式识别会议记录（第3431-3440页）。

Lowe，D.（1999年）。基于局部尺度不变特征的目标识别。ICCV，21150-1157。

Lowe，D.（2004年）。来自比例不变关键点的独特图像特征。IJCV，60（2），91-110。

洛伊，C.，林，D.，欧阳，W.，熊，Y.，杨，S.，黄，Q.，等（2019）。2018年宽脸和行人挑战：方法和结果。arXiv:1902.06854。

卢，Y.，Javidi，T.，和Lazebnik，S.（2016）。使用邻接和缩放预测的自适应目标检测。在CVPR中（第2351-2359页）。

罗平，王，X.，邵，W.，彭，Z.（2018）。了解批量标准化中的正则化。在ICLR。

罗平，张瑞仁，任，J，彭，Z，李，J.（2019）。可切换的规范化，用于学习规范化深度表示。模式分析和机器智能IEEE交易。https://doi.org/10.1109/TPAMI.2019.2932062.

Malisiewicz，T.，和Efros，A.（2009）。超越类别：用于推理对象关系的可视memex模型。在紧要关头。

马，J.，邵，W.，叶，H.，王，L.，王，H.，郑，Y.，等（2018）。通过旋转提案进行任意定向的SCenetextDetection.IEEE TMM，20（11），3111–3122。

Manen，S.，Guillaumin，M.，和VanGool，L.（2013）。随机prim算法的PrimeObjectPropositions。在CVPR中（第2536-2543页）。

Mikolajczyk，K.，和Schmid，C.（2005）。局部描述符的性能评估。IEEE TPAMI，27（10），1615–1630。

Mordan，T.，Thome，N.，Henaff，G.，和Cord，M.（2018）。用于目标检测的基于潜在可变形部分表示的端到端学习。在IJCV中（第1-21页）。

MS COCO检测排行榜。(2018). http://cocodataset.org/# 检测排行榜。

Mundy，J.（2006）。几何时代的物体识别：回顾。在J.Ponce、M.Hebert、C.Schmid和A.Zisserman（编辑）的著作《类别级对象识别》（第3-28页）。柏林：斯普林格。

Murase，H.，和Nayar，S.（1995年a）。从外观视觉学习和识别3D对象。IJCV，14（1），5-24。

Murase，H.，和Nayar，S.（1995年b）。从外观视觉学习和识别3d对象。IJCV，14（1），5-24。

墨菲，K.，托拉尔巴，A.，弗里曼，W.（2003）。使用森林查看树木：与特征、对象和场景相关的图形模型。在紧要关头。

Newell，A.，Huang，Z.，和Deng，J.（2017）。关联嵌入：用于联合检测和分组的端到端学习。在NIPS中（第2277-2287页）。

Newell，A.，Yang，K.，和Deng，J.（2016）。用于人体姿势估计的堆叠沙漏网络。在ECCV中（第483-499页）。

Ojala，T.，Pietikäinen，M.，和MaenpääT.（2002年）。具有局部二值模式的多分辨率灰度和旋转不变纹理分类。IEEE TPAMI，24（7），971–987。

Oliva，A.，和Torralba，A.（2007）。语境在物体识别中的作用。《认知科学趋势》，11（12），520-527。

奥佩特，A.，平茨，A.，福塞内格尔，M.，和奥尔，P.（2006）。具有boosting的通用对象识别。IEEE TPAMI，28（3），416–431。

Oquab，M.，Bottou，L.，Laptev，I.，和Sivic，J.（2014）。使用卷积神经网络学习和传输中级图像表示。在CVPR中（第1717-1724页）。

Oquab，M.，Bottou，L.，Laptev，I.，和Sivic，J.（2015）。对象定位是免费的吗？卷积神经网络的弱监督学习。在CVPR中（第685-694页）。

Osuna，E.，Freund，R.，和Girosit，F.（1997）。训练支持向量机：人脸检测的应用。在CVPR中（第130-136页）。

欧阳，W.，王，X.（2013）。用于行人检测的联合深度学习。在ICCV中（第2056-2063页）。

欧阳，W.，王，X.，曾，X.，邱，S.，罗，P.，田，Y.，李，H.，杨，S.，王，Z.，罗，C.-C.，等（2015）。DeepIDNet：用于目标检测的可变形深度卷积神经网络。在CVPR中（第2403-2412页）。

欧阳，W.，王，X.，张，C.，杨，X.（2016）。用于长尾分布目标检测的微调深度模型中的因素。在CVPR中（第864-873页）。

欧阳，W.，王，K.，朱，X.，和王，X.（2017a）。用于目标检测的链式级联网络。在ICCV中。

欧阳，W.，曾，X.，王，X.，邱，S.，罗，P.，田，Y.，等。（2017b）。DeepIDNet：基于可变形部分的卷积神经网络的目标检测。IEEE TPAMI，39（7），1320–1334。

Parikh，D.，Zitnick，C.，和Chen，T.（2012）。探索微小图像：外观和上下文信息在机器和人类目标识别中的作用。IEEE TPAMI，34（10），1978-1991年。

帕斯卡挥发性有机化合物检测排行榜。(2018). http://host.robots.ox.ac. 英国：8080/leadboard/main\_bootstrap.php

彭聪、肖天堂、李志强、蒋耀勇、张晓霞、贾庆康、俞国光、孙建军（2018）。MegDet：一种大型小型目标探测器。在CVPR中。

彭，X.，孙，B.，阿里，K.，和萨恩科，K.（2015）。从3d模型学习深度目标探测器。在ICCV中（第1278-1286页）。

Pepik，B.，Benenson，R.，Ritschel，T.，和Schiele，B.（2015）。什么是阻止ConvNet进行检测？在德国模式识别会议上（第517-528页）。

Perronnin，F.，Sánchez，J.，和Mensink，T.（2010）。改进fisher-kernelforlargescaleimageclassification.InECCV（第143-156页）。

Pinheiro，P.，Collobert，R.，和Dollar，P.（2015）。学习分割候选对象。在NIPS中（1990-1998页）。

Pinheiro，P.，Lin，T.，Collobert，R.，和DolláR，P.（2016）。学习细化对象分段。在ECCV中（第75-91页）。

庞塞，J.，赫伯特，M.，施密德，C.，齐瑟曼，A.（2007）。面向类别级对象识别。柏林：斯普林格。

Pouyanfar，S.，Sadiq，S.，Yan，Y.，Tian，H.，Tao，Y.，Reyes，M.P.，等人（2018年）。关于深度学习的调查：算法、技术和应用。ACM计算调查，51（5），92:1–92:36。

齐中荣、刘伟、吴春春、苏海华和吉巴斯，L.J.（2018）。用于从RGBD数据检测三维对象的平截头体点网。在CVPR中（第918-927页）。

齐中荣、苏海浩、莫国光和吉巴斯，L.J.（2017）。PointNet：深入学习3D分类和分割的点集。在CVPR中（第652-660页）。

杨全明，孟朔，W.，雨果，J.E.，伊莎贝尔，G.，伊奇，H.，玉峰，L.，等（2018）。将人类从学习应用中解放出来：关于自动机器学习的调查。arXiv:1810.13306。

拉比诺维奇，A.，韦达尔迪，A.，加列吉洛斯，C.，维尤拉，E.，和贝隆吉，S.（2007）。上下文中的对象。在ICCV中。

Rahman，S.，Khan，S.，和Barnes，N.（2018a）。零炮目标检测极性损失。arXiv:1811.08982。

Rahman，S.，Khan，S.，和Porikli，F.（2018b）。零镜头目标检测：学习同时识别和定位新概念。在ACCV。

拉扎维安，R.，阿齐兹普尔，H.，沙利文，J.，和卡尔森，S.（2014）。CNN的特色是现成的：一个令人震惊的识别基线。在CVPR研讨会上（第806-813页）。

Rebuffi，S.，Bilen，H.，和Vedaldi，A.（2017）。使用剩余适配器学习多个可视域。神经信息处理系统的进展（第506-516页）。

Rebuffi，S.，Bilen，H.，和Vedaldi A.（2018年）。多域深度神经网络的有效参数化。在CVPR中（第8119-8127页）。

雷德蒙，J.，迪瓦拉，S.，吉尔希克，R.，和法哈迪，A.（2016）。您只需查看一次：统一的实时对象检测。在CVPR中（第779-788页）。

Redmon，J.和Farhadi，A.（2017）。YOLO9000：更好、更快、更强。在CVPR中。

Ren，M.，Triantafillou，E.，Ravi，S.，Snell，J.，Swersky，K.，Tenenbaum，J.B.，Larochelle，H.，和Zemel R.S.（2018）。半监督少镜头分类的元学习。在ICLR。

任，S.，何，K.，吉尔希克，R.，孙，J.（2015）。更快的R-CNN：通过区域建议网络实现实时目标检测。在NIP中（第91-99页）。

任S.何K.吉尔希克R.和孙J.（2017）。更快的RCNN：使用区域建议网络实现实时目标检测。IEEE TPAMI，39（6），1137–1149。

任S.，何，K.，吉尔希克，R.，张，X.，和孙，J.（2016）。基于卷积特征映射的目标检测网络。IEEE模式分析和机器智能学报，39（7），1476–1481。

Rezatofighi，H.，Tsoi，N.，Gwak，J.，Sadeghian，A.，Reid，I.，和Savarese，S.（2019年）。并集上的广义交集：边界盒回归的度量和损失。在CVPR中。

Rowley，H.，Baluja，S.，和Kanade，T.（1998）。基于神经网络的人脸检测。IEEE TPAMI，20（1），23–38。

卢萨科夫斯基，O.，邓，J.，苏，H.，克劳斯，J.，萨蒂什，S.，马，S.，等人（2015年）。ImageNet大规模视觉识别挑战。IJCV，115（3），211-252。

罗素，B.，托拉尔巴，A.，墨菲，K.，弗里曼，W.（2008）。LabelMe：用于图像注释的数据库和基于web的工具。IJCV，77（1-3），157-173。

施密德，C.，和莫尔，R.（1997）。用于图像检索的局部灰度值不变量。IEEE TPAMI，19（5），530–535。

Schwartz，E.，Karlinsky，L.，Shtok，J.，Harary，S.，Marder，M.，Pankanti，S.，Feris，R.，Kumar，A.，Giries，R.，和Bronstein，A.（2019年）。RepMet：用于分类和一次性目标检测的基于代表性的度量学习。在CVPR中。

Sermanet，P.，Eigen，D.，Zhang，X.，Mathieu，M.，Fergus，R.，和LeCun，Y.（2014）。OverFeat：使用卷积网络集成识别、定位和检测。在ICLR。

Sermanet，P.，Kavukcuoglu，K.，Chintala，S.，和LeCun，Y.（2013）。无监督多阶段特征学习的行人检测。在CVPR中（第3626-3633页）。

Shang，W.，Sohn，K.，Almeida，D.，和Lee，H.（2016）。通过串联校正线性单元理解和改进卷积神经网络。在ICML中（第2217-2225页）。

Shelhamer，E.，Long，J.，和Darrell，T.（2017）。用于语义分割的完全卷积网络。IEEE TPAMI。

沈志强，刘志强，李建强，蒋耀勇，陈耀勇，薛某（2017）。DSOD：从头开始学习深度监督的对象检测器。在ICCV中。

石，X.，单，S.，坎，M.，吴，S.，和陈，X.（2018）。基于渐进校准网络的实时旋转不变人脸检测。在CVPR中。

施志强，杨玉英，霍斯佩代尔斯，T.，向T.（2017）。基于对象和属性的弱监督图像标注和分割。IEEE TPAMI，39（12），2525–2538。

Shrivastava，A.，和Gupta A.（2016），快速RCNN的情境启动和反馈。在ECCV中（第330-348页）。

Shrivastava，A.，Gupta，A.，和Girshick，R.（2016）。使用在线硬示例挖掘训练基于区域的对象检测器。在CVPR中（第761-769页）。

Shrivastava，A.，Sukthankar，R.，Malik，J.，和Gupta，A.（2017）。超越跳过连接：用于对象检测的自上而下调制。在CVPR中。

Simonyan，K.，和Zisserman，A.（2015）。用于大规模图像识别的超深卷积网络。在ICLR。

Singh，B.，和Davis，L.（2018）。目标检测SNIP中的尺度不变性分析。在CVPR中。

Singh，B.，Li，H.，Sharma，A.，和Davis，L.S.（2018a）。RFCN 3000，每秒30帧：解耦检测和分类。在CVPR中。

Singh，B.，Najibi，M.，和Davis，L.S.（2018b）。狙击手：有效的多尺度训练。arXiv:1805.09300。

西维奇，J.，和齐瑟曼，A.（2003）。视频谷歌：一种用于视频中对象匹配的文本检索方法。国际计算机视觉会议（ICCV），21470-1477。

Sun，C.，Shrivastava，A.，Singh，S.，和Gupta，A.（2017）。重温深度学习时代数据的不合理有效性。在ICCV中（第843-852页）。

孙坤、肖斌、刘德华、王健（2019a）。用于人体姿态估计的深度高分辨率表示学习。在CVPR中。

孙坤，赵，Y，蒋，B，程，T，肖，B，刘，D，等（2019b）。用于标记像素和区域的高分辨率表示。CoRR。，。arXiv:1904.04514。

孙绍，庞，J.，史，J.，易，S.，和欧阳，W.（2018）。FishNet：用于图像、区域和像素级预测的多功能主干。在NIPS中（第754-764页）。

Sun，Z.，Bebis，G.，和Miller，R.（2006）。道路车辆检测：综述。IEEE TPAMI，28（5），694–711。

Sung，K.，和Poggio，T.（1994）。对象和模式检测的学习和示例选择。麻省理工学院人工智能备忘录（1521）。

Swain，M.，和Ballard，D.（1991）。颜色索引。IJCV，7（1），11-32。

塞格迪，C.，刘，W.，贾，Y.，塞尔马内，P.，里德，S.，安圭洛夫，D.，二汉，D.，万霍克，V.，和拉比诺维奇，A.（2015）。用卷积来深入研究。在CVPR中（第1-9页）。

塞格迪，C.，约菲，S.，万霍克，V.，和阿莱米，A.（2017）。启始v4、启始resnet和剩余连接对学习的影响。在AAAI中（第4278-4284页）。

塞格迪，C.，里德，S.，二汉，D.，安圭洛夫，D.，和约菲，S.（2014）。可扩展、高质量的对象检测。arXiv:1412.1441。

塞格迪，C.，托舍夫，A.，和埃尔汉，D.（2013）。用于目标检测的深度神经网络。在NIPS中（第2553-2561页）。

塞格迪，C.，范霍克，V.，约菲，S.，斯伦斯，J.，和沃伊纳，Z.（2016）。重新思考计算机视觉的初始架构。在CVPR中（第2818-2826页）。

Torralba，A.（2003年）。用于目标检测的上下文启动。IJCV，53（2），169-191。

Turk，M.A.，和Pentland，A.（1991）。利用特征脸进行人脸识别。在CVPR中（第586-591页）。

Tuzel，O.，Porikli，F.，和Meer P.（2006）。区域协方差：一个快速的分类描述符。InECCV（第589-600页）。

TychsenSmith，L.，和Peterson，L.（2017年）。DeNet：具有定向稀疏采样的可伸缩实时对象检测。在ICCV中。

TychsenSmith，L.，和Peterson，L.（2018年）。利用适应度nms和有界iou损失改进目标定位。在CVPR中。

尤伊林斯，J.，范德桑德，K.，盖弗斯，T.，和斯穆尔德，A.（2013）。目标识别的选择性搜索。IJCV，104（2），154-171。

维兰特，R.，蒙罗克，C.，和勒昆，Y.（1994）。图像中物体定位的原始方法。IEE论文集《视觉、图像和信号处理》，141（4），245–250。

Van de Sande，K.，Uijlings，J.，Gevers，T.，和Smelders，A.（2011）。分割作为目标识别的选择性搜索。在ICCV中（第1879-1886页）。

Vaswani，A.，Shazeer，N.，Parmar，N.，Uszkoreit，J.，Jones，L.，Gomez，A.N.，Kaiser，Ł和Polosukhin，I.（2017）。注意力是你所需要的。在NIP中（第6000-6010页）。

韦达尔迪，A.，古尔山，V.，瓦尔马，M.，和齐瑟曼，A.（2009）。多核目标检测。在ICCV中（第606-613页）。

Viola，P.，和Jones，M.（2001年）。使用简单特征的增强级联进行快速目标检测。CVPR，1，1-8。

Wan，L.，Eigen，D.，和Fergus，R.（2015）。卷积网络、可变形零件模型和非最大值抑制的端到端集成。在CVPR中（第851-859页）。

王浩、王魁、高明、李沛、左伟（2018）。多尺度定位AwareKernelRepresentation for ObjectDetection.InCVPR。

王克星，蔡，Z.，高，D.，和Vasconcelos，N.（2019）。通过领域注意进行的全球目标检测。arXiv:1904.04402。

王旭，韩，T，燕S.（2009）。具有部分遮挡处理的HOG-LBP人体检测器。在国际计算机视觉会议上（第32-39页）。

王，X.，Shrivastava，A.，和Gupta，A.（2017）。一种快速RCNN：通过敌手进行目标检测的硬正生成。在CVPR中。

魏彦，潘小霞，秦，H，欧阳，W，燕，J.（2018）。量化模拟：用于目标检测的极小CNN。在ECCV中（第267-283页）。

吴，S.，黄，S.，和川，I.（2018）。StairNet：自上而下的语义聚合，用于精确的一次性检测。在WACV中（第1093-1102页）。

Worrall，D.E.，Garbin，S.J.，Turmukhambetov，D.，和Brostow，G.J.（2017）。谐波网络：深度平移和旋转等变。在CVPR（第2卷）中。

吴勇，何克（2018）。组规范化。在ECCV中（第3-19页）。

吴志强，潘南生，陈福龙，张春龙，余平生（2019）。关于图形神经网络的综合调查。arXiv:1901.00596。

吴志强、宋绍、科斯拉、A、余、F、张、L、唐、X、肖、J.（2015）。3D ShapeNet：体积形状的深度表示。在CVPR中（第1912-1920页）。

夏，G.，白，X.，丁，J.，朱，Z.，贝隆杰，S.，罗，J.，达特库，M.，佩利洛，M.，张，L.（2018）。DOTA：用于航空图像中目标检测的大型数据集。在CVPR中（第3974-3983页）。

项，Y.，莫塔吉，R.，和萨瓦雷斯，S.（2014）。Beyond PASCAL：野外三维物体检测的基准。在WACV中（第75页-

82).

肖，R.，朱，L.，张，H.（2003）。推进目标检测的链学习。在ICCV中（第709-715页）。

谢绍、吉尔希克、R、多拉尔、P、图、Z和何、K.（2017）。深度神经网络的聚合剩余变换。在CVPR中。

杨斌、颜俊杰、雷志强、李小生（2016a）。来自图像的手工对象。在CVPR中（第6043-6051页）。

杨芳、蔡文华和林妍（2016b）。利用所有层：快速准确的CNN对象检测器，具有与比例相关的池和级联拒绝分类器。在CVPR中（第2129-2137页）。

杨，M.，克里格曼，D.，和阿胡亚，N.（2002）。检测人脸图像：一项调查。IEEE TPAMI，24（1），34–58。

叶，Q.，和多尔曼，D.（2015）。图像中的文本检测和识别：综述。IEEE TPAMI，37（7），1480–1500。

Yosinski，J.，Clune，J.，Bengio，Y.，和Lipson，H.（2014）。深度神经网络中的特征有多大的可转移性？在NIPS中（第3320-3328页）。

Young，T.，Hazarika，D.，Poria，S.，和Cambria，E.（2018）。基于深度学习的自然语言处理的最新趋势。IEEE计算智能杂志，13（3），55–75。

余，F.，和科尔顿V.（2015）。通过扩展卷积进行多尺度上下文聚合。arXiv预印本arXiv:1511.07122。

余，F.，Koltun，V.，和Funkhouser，T.（2017）。扩张的剩余网络。在CVPR中（第2卷，第3页）。

于瑞丽、李安、陈春春、赖俊杰等（2018）。NISP：使用神经元重要性评分传播修剪网络。在CVPR中。

Zafeiriou，S.，Zhang，C.，和Zhang，Z.（2015）。野外人脸检测：过去、现在和未来。计算机视觉和图像理解，138，1-24。

Zagoruyko，S.，Leler，A.，Lin，T.，Pinheiro，P.，Gross，S.，Chintala，S.，和Dollár，P.（2016）。用于目标检测的多路径网络。在BMVC中。

Zeiler，M.，和Fergus，R.（2014）。可视化和理解卷积网络。在ECCV中（第818-833页）。

曾，X.，欧阳，W.，闫，J.，李，H.，肖，T.，王，K.，等（2017）。制作用于目标检测的gbd网络。IEEE模式分析和机器智能学报，40（9），2109–2123。

曾，X.，欧阳，W.，杨，B.，燕，J.，王，X.（2016）。门控双向cnn用于目标检测。在ECCV中（第354-369页）。

张克强，张子强，李子强，乔勇（2016a）。使用多任务级联卷积网络的联合人脸检测和对齐。IEEE SPL，23（10），1499–1503。

张磊，林磊，梁，X.，和何，K.（2016b）。更快的RCNN对行人检测效果好吗？在ECCV中（第443-457页）。

张思文，文，L，边，X，雷，Z，李，S.（2018a）。用于目标检测的单次激发细化神经网络。在CVPR中。

张绍，杨杰，和希勒，B.（2018b）。在CNN中通过引导注意进行遮挡行人检测。在CVPR中（第2056-2063页）。

张学贤，李子忠，张学贤，张学贤，李子忠，张学贤，林学德（2017）。波利尼特：在非常深的网络中追求结构多样性。在CVPR中（第718-726页）。

张学友，杨耀彦，韩，Z，王，H.，高，C.（2013）。对象类检测：综述。ACM计算调查，46（1），10:1–10:53。

张晓阳，周晓阳，林明明，孙杰（2018c）。ShuffleNet：一种非常有效的用于移动设备的卷积神经网络。在CVPR中。

张，Z.，盖格，J.，波哈莱宁，J.，穆萨，A.E.，金，W.，和舒勒，B.（2018d）。环境鲁棒语音识别的深度学习：最新发展概述。ACM智能系统与技术交易，9（5），49:1–49:28。

张，Z.，乔，S.，谢，C.，沈，W.，王，B.，和Yuille，A.（2018e）。具有丰富语义的单镜头目标检测。在CVPR中。

赵，Q.，盛，T.，王，Y.，唐，Z.，陈，Y.，蔡，L.，凌，H.（2019）。M2Det：一种基于多级特征金字塔网络的单镜头目标检测器。在AAAI。

郑，S.，Jayasumana，S.，Romera Paredes，B.，Vineet，V.，Su，Z.，Du，D.，Huang，C.，和Torr，P.（2015）。条件随机场作为递归神经网络。在ICCV中（第1529-1537页）。

Zhou，B.，Khosla，A.，Lapedriza，A.，Oliva，A.，和Torralba，A.（2015）。目标探测器出现在深场景CNN中。在ICLR。

Zhou，B.，Khosla，A.，Lapedriza，A.，Oliva，A.，和Torralba，A.（2016a）。学习深层特征以进行区分性定位。在CVPR中（第2921-2929页）。

Zhou，B.，Lapedriza，A.，Khosla，A.，Oliva，A.，和Torralba，A.（2017a）。地点：用于场景识别的1000万图像数据库。IEEE模式分析和机器智能学报，40（6），1452–1464。

周，J.，崔，G.，张，Z.，杨，C.，刘，Z.，孙，M.（2018a）。图神经网络：方法和应用综述。arXiv:1812.08434。

周沛，倪，B.，耿，C.，胡，J.，徐，Y.（2018b）。尺度可转移目标检测。在CVPR中。

周勇、刘磊、邵磊和梅勒（2016b）。DAVE：用于快速车辆检测和注释的统一框架。在ECCV中（第278-293页）。

周亦云，叶，Q.，邱，Q.，焦，J.（2017b）。定向反应网络。在CVPR中（第4961-4970页）。

朱晓霞，涂亚，D.，牟，L.，夏，G.，张，L.，徐，F.，等（2017）。遥感深度学习：全面回顾和资源清单。IEEE地球科学和遥感杂志，5（4），8-36。

朱，X.，冯德里克，C.，福克斯，C.，和拉马南，D.（2016a）。我们需要更多的训练数据吗？IJCV，119（1），76-92。

Zhu，Y.，Urtasun，R.，Salakhutdinov，R.，和Fidler，S.（2015）。Segdepm：利用深度神经网络中的分割和上下文进行目标检测。在CVPR中（第4703-4711页）。

朱咏，赵，C.，王，J.，赵，X.，吴，Y.，陆，H.（2017a）。耦合网：将全局结构与局部部分耦合，用于目标检测。在ICCV中。

朱尧，周尧，叶，Q，邱，Q，焦，J.（2017b）。弱监督目标定位的软建议网络。在ICCV中（第1841-1850页）。

朱子亮、梁德明、张树南、黄小十、李小白、胡树南（2016b）。野外交通标志检测与分类。在CVPR中（第2110-2118页）。

Zitnick，C.，和Dollár，P.（2014）。边缘框：从边缘定位对象建议。在ECCV中（第391-405页）。

Zoph，B.，和Le，Q.（2016）。神经结构搜索与强化学习。arXiv预印本arXiv:1611.01578。

Zoph，B.，Vasudevan，V.，Shlens，J.，和Le，Q.（2018年）。学习可伸缩图像识别的可转移架构。在CVPR中（第8697-8710页）。

**出版商说明**Springer Nature在公布的地图和机构附属机构的管辖权主张方面保持中立。

[[1]](" \l "_ftnref1" \o ")国防科技大学，长沙，中国

[[2]](" \l "_ftnref2" \o ")奥卢大学，奥卢，芬兰

[[3]](" \l "_ftnref3" \o ")悉尼大学，Camperdown，澳大利亚

[[4]](" \l "_ftnref4" \o ")中国香港大学，Sha Tin，中国

[[5]](" \l "_ftnref5" \o ")滑铁卢大学，滑铁卢，加拿大

[[6]](" \l "_ftnref6" \o ")据我们所知，关于各种视觉子任务的定义，文献中没有普遍的一致意见。诸如检测、定位、识别、分类、分类、验证、识别、注释、标记和理解等术语的定义往往不同（Andreopoulos和Tsotsos 2013）。

[[7]](" \l "_ftnref7" \o ")测试集上的注释未公开发布，PASCAL VOC2007除外。

[[8]](" \l "_ftnref8" \o ")在对象检测挑战中，如PASCAL VOC和ILSVRC，每个对象类别的获胜条目是AP得分最高的条目，而挑战的获胜者是在大多数对象类别上获胜的团队。地图也被用来衡量一个团队的能力

[[9]](" \l "_ftnref9" \o ")值得注意的是，对于给定的阈值β，图像中同一对象的多次检测并不被视为所有正确检测，只有具有最高置信水平的检测被视为TP，其余被视为FPs。

[[10]](" \l "_ftnref10" \o ")对象建议，也称为区域建议或检测建议，是图像中可能包含对象的一组候选区域或边界框（Chavali et al.2016；Hosang et al.2016）。

[[11]](" \l "_ftnref11" \o ")请参阅第节。4.2关于借据的定义。

[[12]](" \l "_ftnref12" \o ")相比之下，YOLO使用的边界框要少得多，每张图像只有98个

从选择性搜索到大约2000。

[[13]](" \l "_ftnref13" \o ")不同尺寸和光谱的盒子，可作为候选对象。

[[14]](" \l "_ftnref14" \o ")使用关键点进行目标检测的想法之前出现在DeNet（TychsenSmith和Petersson 2017）。

[[15]](" \l "_ftnref15" \o ")Densenet以隐式方式执行深度监控，即单个层通过较短的连接接收其他层的额外监控。深度监督的好处此前已在深度监督网络（DSN）中得到证明（Lee等人，2015年）。

[[16]](" \l "_ftnref16" \o ")微调是通过初始化网络来完成的，该网络的权重针对大型标记数据集（如ImageNet）进行了优化。然后使用目标任务训练集更新网络的权重。

[[17]](" \l "_ftnref17" \o ")基于滑动窗口的检测需要对每个图像进行大约104–105个窗口的分类。当考虑多个比例和纵横比时，每个图像的窗口数会显著增加到106–107个窗口。

[[18]](" \l "_ftnref18" \o ")我们交替使用术语检测建议、对象建议和区域建议。

[[19]](" \l "_ftnref19" \o ")“锚”的概念首次出现在Ren等人（2015年）中。

[[20]](" \l "_ftnref20" \o ")请参阅第节。4.2有关借据定义的更多详细信息。

[[21]](" \l "_ftnref21" \o ")尽管可以提供辅助信息，例如维基百科页面或属性向量。