二十年对象检测进展：一项调研

郑霞邹，可言陈，振伟石，IEEE 会员，玉红郭，杰平叶，IEEE 会员

摘要

对象检测作为计算机视觉中最基本和最有挑战性的问题之一，在近年来受到了极大的关注。在过去的两个十年里，我们见证了对象检测的快速技术演变和它对整个计算机视觉领域的深远影响。如果我们将今天的对象检测技术视为深度学习驱动的一场革命，那么在 20 世纪 90 年代，我们会看到早期计算机视觉的巧妙思维和长期的前瞻性设计。本文从技术进化的角度对这个快速发展的研究领域进行了广泛的回顾，时间跨度超过四分之一个世纪（从 20 世纪 90 年代到 2022 年）。本文涵盖了多个主题，包括历史上的里程碑检测器，检测数据集，度量，检测系统的基本构建块，加速技术和最近的最先进的检测方法。

索引词

对象检测，计算机视觉，深度学习，卷积神经网络，技术演变。

I. 引言

对象检测是一个重要的计算机视觉任务，涉及检测数字图像中某一类（如人类、动物或汽车）的视觉对象实例。对象检测的目标是开发计算模型和技术，为计算机视觉应用提供最基本的知识：哪些对象在哪里？对象检测的两个最重要的指标是准确性（包括分类准确性和定位准确性）和速度。对象检测作为许多其他计算机视觉任务的基础，如实例分割、图像标注、对象跟踪等。近年来，深度学习技术的快速发展极大地推动了对象检测的进步，取得了显著的成就。

图 2：对象检测的路线图。此图中的里程碑检测器有：VJ 检测器[10, 11]，HOG 检测器[12]，DPM[13-15]，RCNN[16]，SPPNet[17]，Fast RCNN[18]，Faster RCNN[19]，YOLO[20-22]，SSD[23]，FPN[24]，Retina-Net[25]，CornerNet[26]，CenterNet[27]，DETR[28]。

A. 对象检测的路线图

在过去的两个十年中，人们普遍认为对象检测的进步经历了两个历史时期：“传统对象检测时期（2014 年之前）”和“基于深度学习的检测时期（2014 年之后）”，如图 2 所示。接下来，我们将总结这一时期的里程碑检测器，以出现时间和性能为主要线索来突出背后的驱动技术，见图 3。

1）里程碑：传统检测器：

如果我们认为今天的对象检测技术是由深度学习驱动的革命，那么回到 20 世纪 90 年代，我们会看到早期计算机视觉的巧妙设计和长远的视角。大多数早期的对象检测算法都是基于手工制作的特征构建的。由于当时缺乏有效的图像表示，人们不得不设计复杂的特征表示和多种加速技能。

Viola-Jones 检测器：

在 2001 年，P. Viola 和 M. Jones 首次实现了没有任何限制（例如，肤色分割）的实时人脸检测[10, 11]。在一个 700MHz 的 Pentium III CPU 上运行时，这个检

测器比其它同一时期的算法快了几十甚至几百倍，而且保持了可比的检测准确率。VJ 检测器遵循最直接的检测方法，即滑动窗口：遍历图像中的所有可能的位置和尺度，看看任何窗口是否包含一个人脸。虽然这似乎是一个非常简单的过程，但其背后的计算远超过了当时计算机的处理能力。通过整合三项重要技术，VJ 检测器显著地提高了其检测速度

Original:

Translated:
HOG 检测器：
HOG 检测器是基于直方图的定向梯度来描述对象的外观。与 VJ 检测器不同的是，为了处理不同的对象大小，HOG 检测器多次调整输入图像的大小，同时保持检测窗口的大小不变。多年来，HOG 检测器已成为许多对象检测器和大量计算机视觉应用的重要基础[13, 14, 32]。

变形部分基模型（DPM）：
DPM 是 VOC-07、-08 和-09 检测挑战的冠军，是传统对象检测方法的缩影。DPM 最初是由 P. Felzenszwalb 在 2008 年提出的，作为 HOG 检测器的一个扩展[13]。它遵循“分而治之”的检测理念，其中训练可以简单地视为学习正确分解对象的方式，而推断可以视为对不同对象部分的检测的集成。例如，检测“汽车”的问题可以分解为检测其窗户、车身和车轮。这部分工作也被称为“星形模型”，由 P. Felzenszwalb 等人引入[13]。后来，R. Girshick 进一步扩展了星型模型，以处理现实世界中更显著的变化，并进行了一系列其他的改进[14, 15, 33, 34]。

虽然今天的对象检测器在检测精度上远远超越了 DPM，但其中许多仍然受到其有价值的见解的深刻影响，例如混合模型、硬负采样、边界框回归、上下文引导等。在 2010 年,P. Felzenszwalb 和 R. Girshick 被 PASCAL VOC 授予“终身成就奖”。

2）里程碑：基于 CNN 的两阶段检测器：
随着手工特征的性能变得饱和，对象检测的研究在 2010 年之后达到了一个平台。2012 年，世界见证了卷积神经网络的重生[35]。由于深卷积网络能够学习图像的稳健和高级特征表示，人们自然会问：我们能将它引入对象检测吗？R. Girshick 等人率先打破了僵局

Original:

Translated:
在技术发展的僵局中，他们提出了 RCNN，这是第一个利用卷积神经网络的深层特征来检测对象的框架[16]。RCNN 首先使用选择性搜索来提取大约 2000 个对象提议，然后使用卷积神经网络来提取每一个提议的特征。虽然 RCNN 已经实现了当时最先进的检测性能，但它的速度非常慢，因为它需要为每一个对象提议单独提取特征。这导致了 SPPNet[17]和 Fast RCNN[18]的出现，它们通过共享卷积计算来加速 RCNN。但是，选择性搜索仍然是一个瓶颈，这导致了 Faster RCNN[19]的发明，它引入了一个区域提议网络来替换选择性搜索，从而实现了近乎实时的检测

速度。

虽然基于区域的方法已经取得了非常好的结果，但它们通常都很慢和复杂，特别是当用于大规模对象检测时。这导致了一阶段检测器的出现，如 YOLO[20-22]和 SSD[23]，它们摒弃了区域提议步骤，直接预测对象的边界框和类别，从而实现了更快的检测速度。随着时间的推移，一阶段检测器已经从最初的 YOLO 和 SSD 演变到更先进的版本，如 FPN[24]和 RetinaNet[25]，它们通过引入更多的技术创新来改善检测性能。

图 3：VOC07、VOC12 和 MS-COCO 数据集上对象检测的准确性提高。本图中的检测器：DPM-v1 [13]，DPM-v5 [37]，RCNN [16]，SPPNet [17]，Fast RCNN [18]，Faster RCNN [19]，SSD [23]，FPN [24]，RetinaNet [25]，RefineDet [38]，TridentNet [39]，CenterNet [40]，FCOS [41]，HTC [42]，YOLOv4 [22]，Deformable DETR [43]，Swin Transformer [44]。
2014 年，提出了 RCNN 区域与 CNN 特征（RCNN）[16, 36]，自那时起，对象检测开始以前所未有的速度发展。在深度学习时代，有两组检测器："两阶段检测器"和"一阶段检测器"，前者将检测框架为"粗到细"的过程，而后者将其框架为"一步完成"。

RCNN：
RCNN 背后的思想很简单：它从通过选择性搜索[45]提取一组对象提议（对象候选框）开始。然后，每个提议被重新缩放到固定大小的图像，并送入一个预先训练过的 CNN 模型（比如说，ImageNet 上的 AlexNet[35]）来提取特征。最后，使用线性 SVM 分类器来预测每个区域内的对象存在并识别对象类别。RCNN 在 VOC07 上取得了显著的性能提升，将平均精度（mAP）从 33.7%（DPM-v5 [46]）提高到 58.5%。虽然 RCNN 取得了很大的进步，但它的缺点很明显：大量重叠提议上的冗余特征计算（一个图像有超过 2000 个框）导致检测速度极慢（每张图像需要 14 秒，使用 GPU）。同年稍后，提出了 SPPNet[17]来解决这个问题。

Original:

Translated:
SPPNet：
SPPNet 引入了空间金字塔池化（SPP）层来解决 RCNN 的冗余计算问题。通过在卷积层之后添加 SPP 层，SPPNet 可以一次处理整个图像，而不是处理 2000 多个独立的区域提议，从而大大提高了计算效率。SPPNet 不仅提高了检测速度，还提高了检测精度，将 VOC07 测试集的 mAP 提高到 59.2%。

Fast RCNN：
Fast RCNN 进一步改进了 SPPNet 的设计，完全去除了昂贵的区域提议特征提取步骤。它使用一个卷积网络来提取整个图像的特征图，然后使用 RoI 池化层来从特征图中提取每个区域提议的特征。这样，所有区域提议可以共享相同的卷积计算，从而大大减少了计算时间。Fast RCNN 还引入了多任务损失来同时优化边界框回归和对象分类，从而进一步提高检测精度。

Faster RCNN：
Faster RCNN 是 Fast RCNN 的一个自然延伸，它解决了区域提议的计算瓶颈问题。Faster RCNN 通过引入区域提议网络（RPN），一个小型全卷积网络，来预测对象边界框和对象分数，从而完全去除了选择性搜索步骤。通过共享卷积特征，RPN 和检测网络可以共同工作，实现近乎实时的检测速度和更高的检测精度。

SSD：
SSD 是一个一阶段检测器，它摒弃了区域提议步骤，直接从特征图中预测边界框和类别分数。通过使用多尺度特征图和默认框来处理不同大小的对象，SSD 不仅实现了与 Faster RCNN 相当的检测精度，而且大大提高了检测速度。SSD 还引入了各种数据增强技术来提高检测性能，包括难例挖掘和多尺度训练，这使其成为当时最快最准确的检测器之一。

FPN 和 RetinaNet：
一阶段检测器在速度上有优势，但它们在检测小对象时通常性能较差。为了解决这个问题，Lin 等人提出了特征金字塔网络（FPN），它使用自上而下和自下而上的路径来构建一个高质量的多尺度特征金字塔[24]。不久之后，他们进一步提出了 RetinaNet 来解决一阶段检测器的类别不平衡问题。通过引入焦点损失函数，RetinaNet 可以更好地处理背景和对象之间的不平衡，从而实现了非常高的检测精度。

一阶段和两阶段检测器的比较：
在深度学习时代，一阶段和两阶段检测器都取得了巨大的成功。两阶段检测器通常提供更高的检测精度，但速度较慢，而一阶段检测器则提供更快的检测速度但精度略有损失。随着技术的进步，这两种类型的检测器都正在逐渐靠拢，实现更高的检测精度和更快的速度。

对象检测的最新进展：
近年来，对象检测领域已经见证了许多创新和突破，包括新的检测器设计，更大更复杂的数据集和更先进的训练技术。其中一些最新的检测器，如 DETR,CenterNet 和 Swin Transformer，已经实现了非常高的检测精度和效率。这些最新的进展不仅推动了对象检测的研究，还为实际应用打开了新的可能性，包括自动驾驶，视频监控和工业自动化等。

由于速度较差和极高的复杂性，两阶段检测器受到了一些限制。相比之下，一阶段检测器可以在一步推理中检索所有对象。他们受到移动设备的喜欢，因为它们具有实时和易于部署的特点，但在检测密集和小型对象时，其性能明显受到影响。

你只看一次（YOLO）：
YOLO 是由 R. Joseph 等人在 2015 年提出的。它是深度学习时代的第一个一阶段检测器[20]。YOLO 非常快：YOLO 的快速版本在 VOC07 上以 mAP=52.7%的速度运行，达到 155fps，而其增强版本以 VOC07 mAP=63.4%的速度运行，达到 45fps。YOLO 采用了与两阶段检测器完全不同的范例：将单个神经网络应用于整个图像。该网络

将图像划分为多个区域，并同时为每个区域预测边界框和概率。尽管 YOLO 大大提高了检测速度，但与两阶段检测器相比，它的定位精度有所下降，尤其是对于一些小对象。YOLO 的后续版本[21, 22, 51]和后来提出的 SSD[23]更加重视这个问题。最近，YOLOv4 团队提出了 YOLOv7[52]。它通过引入优化的结构，如动态标签分配和模型结构重新参数化，超越了大多数现有的对象检测器，实现了速度和准确性的最佳组合（范围从 5 FPS 到 160 FPS）。

Original:

Translated:
单发多框检测器（SSD）：
SSD 是由 W. Liu 等人在 2015 年提出的[23]。SSD 的主要贡献是引入多参考和多分辨率检测技术（将在第 II-C1 节中介绍），显著提高了一阶段检测器的检测精度，尤其是对一些小对象。SSD 在检测速度和精度方面都有优势（COCO mAP@.5=46.5%，快速版本以 59fps 运行）。SSD 和之前的检测器之间的主要区别是，SSD 在网络的不同层上检测不同规模的对象，而之前的检测器只在一个固定的层上运行检测。

对于一些小对象的检测，单发多框检测器（SSD）通过引入多参考和多分辨率检测技术显著提高了一阶段检测器的检测精度。这些技术允许 SSD 在网络的不同层上检测不同规模的对象，而不是仅在一个固定的层上运行检测。

DETR：
近年来，变形器已深刻影响了整个深度学习领域，尤其是计算机视觉领域。变形器摒弃了传统的卷积运算符，转而采用仅注意力计算来克服 CNN 的局限性，并获得全局规模的接受域。在 2020 年，N. Carion 等人提出了 DETR[28]，他们将对象检测视为一种集合预测问题，并提出了一个端到端的检测网络，使用变形器。到目前为止，对象检测已进入一个新时代，在这个时代中，对象可以在不使用锚箱或锚点的情况下被检测到。后来，X. Zhu 等人提出了可变形 DETR[43]来解决 DETR 的长时间收敛和在检测小对象时的有限性能。它在 MSCOCO 数据集上实现了最先进的性能（COCO mAP@.5=71.9%）。

Original:

Translated:
B. 对象检测数据集和指标
1）数据集：
为了开发先进的检测算法，构建更大、偏差更小的数据集是必不可少的。在过去的 10 年里，已经发布了一些著名的检测数据集，包括 PASCAL VOC 挑战赛[54, 55]的数据集（例如，VOC2007, VOC2012）、ImageNet 大规模视觉识别挑战赛（例如，ILSVRC2014）[56]、MS-COCO 检测挑战赛[57]、Open Images 数据集[58, 59]、Objects365[60]等。这些数据集的统计数据在表 I 中给出。图 4 显示了这些数据集的一些图像示例。图 3 显示了从 2008 年到 2021 年在 VOC07、VOC12 和 MS-COCO 数据集上检测精度的提高。

Pascal VOC：
PASCAL Visual Object Classes (VOC)挑战赛（从2005年到2012年）[54, 55]是早期计算机视觉社区中最重要的比赛之一。在对象检测中主要使用了两个版本的Pascal-VOC：VOC07和VOC12，前者包含5000张训练图像和12000个注释对象，后者包含11000张训练图像和27000个注释对象。20个类别
1
http://host.robots.ox.ac.uk/pascal/VOC/

图4：以下数据集中的一些示例图像和注释：(a) PASCAL-VOC07，(b) ILSVRC，(c) MS-COCO，和 (d) Open Images。
数据集
训练 图片 对象
验证 图片 对象
训练+验证 图片 对象
测试 图片 对象
VOC-2007
2,501 6,301
2,510 6,307
5,011 12,608
4,952 14,976
VOC-2012
5,717 13,609
5,823 13,841
11,540 27,450
10,991 -
ILSVRC-2014
456,567 478,807
20,121 55,502
476,688 534,309
40,152 -
ILSVRC-2017
456,567 478,807
20,121 55,502
476,688 534,309
65,500 -
MS-COCO-2015
82,783 604,907
40,504 291,875
123,287 896,782
81,434 -
MS-COCO-2017
118,287 860,001
5,000 36,781
123,287 896,782

40,670 -
Objects365-2019
600,000 9,623,000
38,000 479,000
638,000 10,102,000
100,000 1,700,00
OID-2020
1,743,042 14,610,229
41,620 303,980
1,784,662 14,914,209
125,436 937,327
表 I：一些知名的对象检测数据集及其统计数据。
这两个数据集注释了我们日常生活中常见的 20 种对象，例如“人”、“猫”、“自行车”、“沙发”等。

ILSVRC：
ImageNet 大规模视觉识别挑战（ILSVRC）2[56]推动了通用对象检测的技术进步。ILSVRC 自 2010 年至 2017 年每年都会举行。它包含了一个使用 ImageNet 图像[61]的检测挑战。ILSVRC 检测数据集包含 200 类视觉对象。它的图像/对象实例数量是 VOC 的两个数量级更大。

MS-COCO：
MS-COCO3[57]是目前可用的最具挑战性的对象检测数据集之一。基于 MS-COCO 数据集的年度竞赛自 2015 年开始举行。它的对象类别数量少于 ILSVRC，但对象实例更多。例如，MS-COCO-17 包含 164k 张图像和 897k 个来自 80 个类别的标注对象。与 VOC 和 ILSVRC 相比，MS-COCO 的最大进步是除了边界框注释外，每个对象还进一步使用每实例分割进行标记，以帮助精确定位。此外，MS-COCO 包含更多小对象（其面积小于图像的 1%）和更多密集定位的对象。正如 ImageNet 在其时代一样，MS-COCO 已成为对象检测社区的事实标准。

Open Images：
2018 年见证了 Open Images Detection (OID) challenge4[62]的推出，它跟随 MS-COCO，但规模前所未有。Open Images 有两项任务：1)标准对象检测和 2)视觉关系检测，后者检测特定关系中的配对对象。对于标准检测任务，数据集包括 1,910k 张图像，其中包含 15,440k 个标注边界框，覆盖 600 个对象类别。

2) 指标：
我们如何评估检测器的准确性？这个问题在不同的时期可能有不同的答案。在早期的检测研究中，没有广泛接受的检测准确度评估指标。例如，在早期的行人检测研究[12]中，“漏检率与每窗口假阳性（FPPW）”常用作指标。然而，每窗口测量可能存在缺陷，无法预测全图像性能[63]。2009 年，加利福尼亚理工学院引入了行人检测基准[63, 64]，从那时起，评估指标已从 FPPW 变为每图像假阳性(FPPI)。

近年来，检测最常用的评估方法是“平均精度（AP）”，它最初是在 VOC2007 中

引入的。AP 定义为在不同召回率下的平均检测精度，通常在不同的 IoU 阈值下进行评估。在 PASCAL VOC 挑战赛中，使用 11 个离散召回点来计算 AP。而在 MS-COCO 挑战赛中，则采用不同的 AP 指标，例如 AP@.50（IoU=0.50）、AP@.75（IoU=0.75）、AP@S、AP@M 和 AP@L 来分别评估不同大小的对象检测精度。

在 MS-COCO 上，mAP 是根据 10 个不同的 IoU 阈值（从 0.5 到 0.95，每 0.05 一个步长）来计算的。这使得 MS-COCO 的评估更加严格和详细。例如，对于一个给定的检测结果，如果其 IoU 大于 0.5 但小于 0.55，它将被计为一个正确的检测，但如果 IoU 小于 0.5，它将被计为一个错误的检测。

另外，除了基于 IoU 的 mAP 外，MS-COCO 还提供了其他评估指标，如单个类别的 AP（每个类别的 AP）、AR@1（召回率@1）、AR@10（召回率@10）、AR@100（召回率@100）、AR@S（小对象的召回率）、AR@M（中等对象的召回率）和 AR@L（大对象的召回率）。这些指标可以帮助研究人员更好地理解他们的检测系统在不同方面的性能。

一种特定类别的方式。平均 AP（mAP）是在所有类别上平均后通常用作最终的性能度量标准。为了衡量对象定位的准确性，使用预测框和真实框之间的 IoU 来验证它是否大于预定义的阈值，例如 0.5。如果是，该对象将被标识为“检测到的”，否则将被标识为“未检测到”。0.5-IoU mAP 已然成为对象检测的事实度量标准。自 2014 年以来，由于引入了 MS-COCO 数据集，研究人员开始更加关注对象定位的准确性。而不是使用固定的 IoU 阈值，MS-COCO AP 是在 0.5 到 0.95 之间的多个 IoU 阈值上取平均的，这鼓励更准确的对象定位，这对于一些真实世界的应用程序可能非常重要（例如，想象有一个机器人试图抓住一个扳手）。

C. 对象检测中的技术演变

在本节中，我们将介绍检测系统的一些重要构建块和它们的技术演变。首先，我们描述多尺度和上下文引导在模型设计上的情况，然后是样本选择策略和训练过程中损失函数的设计，最后是推理中的非最大抑制。图表和文本中的时间戳是由论文的出版时间提供的。图中显示的演变顺序主要是为了帮助读者理解，可能存在时间重叠。

1) 多尺度检测的技术演变：

多尺度检测涉及“不同大小”的和“不同纵横比”的对象是对象检测中的主要技术挑战之一。在过去的 20 年里，多尺度检测经历了多个历史时期，如图 5 所示。

Original:

Translated:

特征金字塔+滑动窗口：

在 VJ 检测器之后，研究人员开始更加关注一种更直观的检测方式，即通过建立“特征金字塔+滑动窗口”。从 2004 年开始，基于这一范式构建了一系列里程碑式的检测器，包括 HOG 检测器，DPM 等。他们通常在图像上滑动一个固定大小的检测窗口，很少注意到“不同的纵横比”。为了检测具有更复杂外观的对象，R. Girshick 等人开始寻找特征金字塔之外的更好解决方案。当时的一种解决方案是“混合模型”[15]，即为不同纵横比的对象训练多个检测器。除此之外，基于示例的检测[32, 70]通过为每个对象实例（示例）训练单独的模型提供了另一种解

决方案。

带有对象提议的检测：
对象提议是指一组可能包含任何对象的类别不可知的参考框。带有对象提议的检测有助于避免在图像上进行穷举滑动窗口搜索。我们建议读者查阅以下论文以全面了解此主题[71, 72]。早期的提议检测方法遵循自下而上的检测哲学[73, 74]。2014 年之后，随着深度 CNN 在视觉识别中的流行，自上而下的基于学习的方法开始在这个问题上显示出更多的优势[19, 75, 76]。现在，提议检测逐渐在一阶段检测器的崛起后淡出了人们的视野。

深度回归和无锚检测：
近年来，随着 GPU 计算能力的增加，多尺度检测变得越来越简单直接，

而不是通过构建特征金字塔来检测不同大小的对象。而是将多尺度检测直接集成到网络设计中[41, 67, 77]。这一趋势也导致了无锚检测器的出现，这些检测器不再依赖于预定义的锚箱来检测对象[41, 77]。在无锚检测中，网络直接回归到边界框的坐标，而不是预测每个锚框的偏移量。这消除了选择合适锚箱的需要，简化了检测流程。最近的一些无锚检测器包括 FCOS[41]，CenterNet[40]和 Reppoints[69]。

2) 上下文引导和样本选择策略：
随着深度学习的发展，上下文信息开始在对象检测中发挥更重要的作用。一个对象的上下文信息可以是空间上下文（即对象周围的环境）和语义上下文（即与其他对象的关系）。在对象检测中利用上下文信息可以改善检测性能，尤其是在有遮挡和背景干扰的情况下。上下文信息可以通过多种方式集成到检测系统中，包括通过使用更深的网络来捕获更大的接受域[78, 79]，或通过显式地模型上下文关系[80, 81]。

样本选择策略也是一个重要的研究方向。在训练过程中，通过选择有意义的样本来减少样本不平衡是至关重要的。样本选择可以基于多种标准进行，包括困难样本挖掘[82]和焦点损失[83]。在这方面，有许多方法试图通过选择有意义的样本来优化检测性能，包括 OHEM[84]和 CASCADED R-CNN[85]。

3) 损失函数设计和非最大抑制：
损失函数是训练对象检测器的关键组成部分。损失函数可以帮助网络学习检测任务中的关键信息。在检测任务中常见的损失函数包括交叉熵损失和平滑 L1 损失。最近的研究表明，设计一个好的损失函数可以显著提高检测性能[86, 87]。

图 6：对象检测中上下文引导的演变。本图中的检测器包括：Face Det. [78]，MultiPath [79]，GBDNet [80, 81]，CC-Net [82]，MultiRegion-CNN [83]，CoupleNet [84]，DPM [14, 15]，StructDet [85]，ION [86]，RFCN++ [87]，RBFNet [88]，TridentNet [39]，非本地[89]，DETR [28]，CtxSVM [90]，PersonContext [91]，SMN [92]，RelationNet [93]，SIN [94]，RescoringNet [95]。

和暴力法。使用深度回归来解决多尺度问题的想法变得简单，即直接基于深度学习特征预测边界框的坐标[20, 66]。2018 年后，研究人员开始从关键点检测的角度思考对象检测问题。这些方法通常遵循两种思路：一种是基于组的方法，它检测关键点（角点，中心或代表点），然后进行对象分组[26, 53, 69, 77]；另一种是无组方法，它将对象视为一个/多个点，然后在这些点的参考下回归对象属性（如大小，比例等）[40, 41]。

多参考/多分辨率检测：
多参考检测目前是用于多尺度检测的最常用方法[19, 22, 23, 41, 47, 51]。多参考检测的主要思想是首先在图像的每个位置定义一组参考（即锚，包括框和点），然后基于这些参考来预测检测框。另一种流行技术是多分辨率检测[23, 24, 44, 67, 68]，即在网络的不同层面检测不同规模的对象。多参考和多分辨率检测现已成为最先进的对象检测系统的两个基本构建块。

如图 10(d)所示，如果我们将特征均匀分成 m 组，而不改变其他配置，那么理论上的计算量将减少到之前的 1/m。图 9：对象检测中的加速技术概览。3) 深度可分离卷积：深度可分离卷积[142]可以视为组卷积的一种特殊情况，即当组的数量等于通道的数量时。通常，我们使用一些 1x1 的过滤器进行维度转换，以使最终输出具有所需的通道数量。通过使用深度可分离卷积，可以减少计算量。

这个想法最近已被应用于对象检测和细粒度分类[143-145]。4) 瓶颈设计：神经网络中的瓶颈层比前几层包含更少的节点。近年来，瓶颈设计已广泛用于设计轻量级网络[50, 133, 146-148]。在这些方法中，检测器的输入层可以被压缩以减少检测开始时的计算量[133, 146, 147]。人们也可以压缩特征图使其变得更薄，从而加速后续检测[50, 148]。5) 使用 NAS 进行检测：基于深度学习的检测器正变得越来越复杂，严重依赖手工设计的网络架构和训练参数。神经架构搜索(NAS)主要涉及定义合适的候选网络空间，改善快速准确搜索的策略，并以低成本验证搜索结果。在设计检测模型时，NAS 可以减少对网络主干和锚箱设计的人工干预[149-155]。

E. 数值加速 数值加速旨在从实现的底层加速对象检测器。1) 通过积分图像加速：积分图像是一种可以用来加速一系列计算的技术。它是通过将图像的每一个像素值替换为该像素及其左上方所有像素的累加值来创建的。这种方法可以用于加速对象检测算法的多个方面，包括特征提取和特征匹配[156,157]。2) 通过硬件加速：硬件加速通常涉及使用特殊的硬件来加速计算过程。例如，可以使用 GPU 或 FPGA 来加速神经网络的训练和推理过程[158,159]。3) 通过算法优化加速：除了硬件加速外，还可以通过优化算法来加速对象检测。一些方法包括使用更高效的算法来减少计算量，或使用更高效的数据结构来减少内存使用量[160,161]。

F. 云端和边缘计算 云端和边缘计算是两种可以用来加速对象检测的技术。云计算涉及使用远程服务器来处理数据和运行应用程序，而边缘计算则涉及在数据源附近的本地设备上进行计算。通过使用云端和边缘计算，可以更好地平衡计算资源和网络带宽的使用，从而提高对象检测的效率和性能[162,163]。

III. 未来方向 随着深度学习和计算机视觉技术的不断进步，对象检测领域也在不断发展。在这一节中，我们将探讨一些可能的未来方向和挑战。A. 3D 和多视图对象检测 3D 和多视图对象检测是一个非常有前途的研究方向。它涉及使用 3D 数据和多源、多视图数据来进行对象检测。例如，可以使用来自多个传感器的 RGB 图像和 3D 激光雷达点来进行检测[231,232]。

视频中的检测： 在 HD 视频中进行实时对象检测/跟踪对于视频监控和自动驾驶非常重要。传统的对象检测器通常是为图像检测设计的，而简单地忽略了视频帧之间的相关性。在计算限制下探索空间和时间相关性以提高检测是一个重要的研究方向[233,234]。

跨模态检测：使用多源/多模态数据进行对象检测，例如，RGB-D 图像、激光雷达、流、声音、文本、视频等，对于更准确的检测系统来说非常重要，这样的系统可以像人类的感知系统一样工作。一些未解决的问题包括：如何将训练有素的检测器迁移到不同的数据模态，如何进行信息融合以提高检测效果等[235,236]。

朝向开放世界的检测：域外泛化、零样本检测和增量检测是对象检测中的新兴话题。它们中的大多数设计了减少灾难性遗忘或利用补充信息的方法。人类有一种本能，可以在环境中发现未知类别的对象。当给出相应的知识(标签)时，人们将从中学到新知识，并保持这些模式。然而，当前的对象检测算法很难掌握未知类别对象的检测能力。开放世界的对象检测旨在在没有明确给出或部分给出监督信号的情况下发现未知类别的对象，这在机器人技术和自动驾驶应用中具有很大的潜力[237,238]。

站在技术演变的高速公路上，我们相信本文将帮助读者构建一个完整的对象检测路线图，并找到这个快速发展的研究领域的未来方向。