计算机视觉 / 综述译本

目标检测二十年：一篇综述

Object Detection in 20 Years: A Survey

Zhengxia Zou⋆、Keyan Chen、Zhenwei Shi（IEEE 会员）、Yuhong Guo、Jieping Ye⋆（IEEE 会士）

摘要

目标检测是计算机视觉中最基础、最具挑战性的问题之一，近年来备受关注。在过去二十年中，目标检测技术迅速演进，对整个计算机视觉领域产生了深远影响。如果说今天的目标检测技术经历了一场由深度学习驱动的变革，那么回看 20 世纪 90 年代，则能看到早期计算机视觉研究中巧妙的构思和有远见的设计。本文以技术演进为线索，广泛回顾这一快速发展的研究领域，时间跨度超过四分之一个世纪（从 20 世纪 90 年代到 2022 年）。文章涵盖历史上的里程碑检测器、检测数据集与评价指标、检测系统的基本模块、加速技术，以及近期达到领先水平的检测方法。

关键词：目标检测，计算机视觉，深度学习，卷积神经网络，技术演进。

目录

I. 引言 | II. 目标检测二十年 | III. 检测加速
IV. 目标检测的近期进展 | V. 结论与未来方向 | 参考文献

I. 引言

目标检测是计算机视觉中的一项重要任务，即从数字图像中找出某一类别的可见目标实例，如人、动物、汽车等。目标检测旨在开发计算模型和技术，回答计算机视觉应用所需的一个最基本的问题：有哪些目标，它们分别在哪里？衡量目标检测的两个最重要的指标是准确性（包括分类准确性与定位准确性）和速度。

目标检测是许多其他计算机视觉任务的基础，如实例分割 [1–4]、图像描述生成 [5–7]、目标跟踪 [8] 等。近年来，深度学习技术 [9] 的迅速发展极大推动了目标检测的进步，带来显著突破，并使其成为受到空前关注的研究热点。如今，目标检测已广泛用于自动驾驶、机器人视觉、视频监控等实际场景。图 1 展示了过去二十年与“目标检测”相关的出版物数量的增长。

不同检测任务的目标与约束可能截然不同，因而难度也不尽相同。除其他计算机视觉任务也会遇到的常见难题，如观测视角不同、光照不同及类内差异，目标检测的挑战还包括但不限于：目标旋转及尺度变化（如小目标）、精确定位、密集与遮挡目标的检测，以及检测加速等。第 IV 节将更详细地分析这些问题。

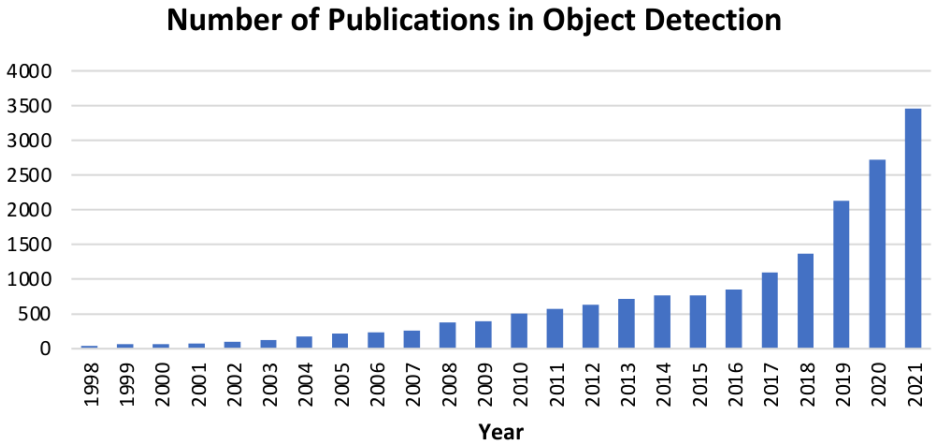


图 1：1998 至 2021 年目标检测相关出版物数量的增长。 (数据来自 Google 学术高级搜索：allintitle: “object detection” OR “detecting objects”。)

横轴为年份（Year），纵轴为相应检索条件下的出版物数量。

本综述希望帮助初学者从多个角度完整把握目标检测技术，重点着眼于它的演进。本文有三个特点：以技术演进为线索进行全面回顾，深入探讨关键技术和近期领先方法，并全面分析检测加速技术。全文以过去、现在和未来为主线，辅以数据集、评价指标和加速技术等目标检测的必要内容。本综述沿着这条技术主线呈现相关技术的演进，帮助读者在不纠缠于技术细节的情况下掌握核心概念，并找到潜在的未来方向。

本文其余部分安排如下：第 II 节回顾目标检测二十年来的演进；第 III 节回顾目标检测的加速技术；第 IV 节介绍最近三年的领先检测方法；第 V 节总结全文，并深入分析进一步的研究方向。

II. 目标检测二十年

本节从多个角度回顾目标检测的历史，介绍里程碑检测器、数据集、评价指标以及关键技术的演进。

A. 目标检测发展路线图

一般认为，过去二十年目标检测的发展大体经历了两个历史时期：“传统目标检测时期（2014 年以前）”和“基于深度学习的检测时期（2014 年以后）”，如图 2 所示。下文将以这些检测器的问世时间和性能为主要线索，总结这一时期的里程碑检测器，突出背后的推动性技术，参见图 3。

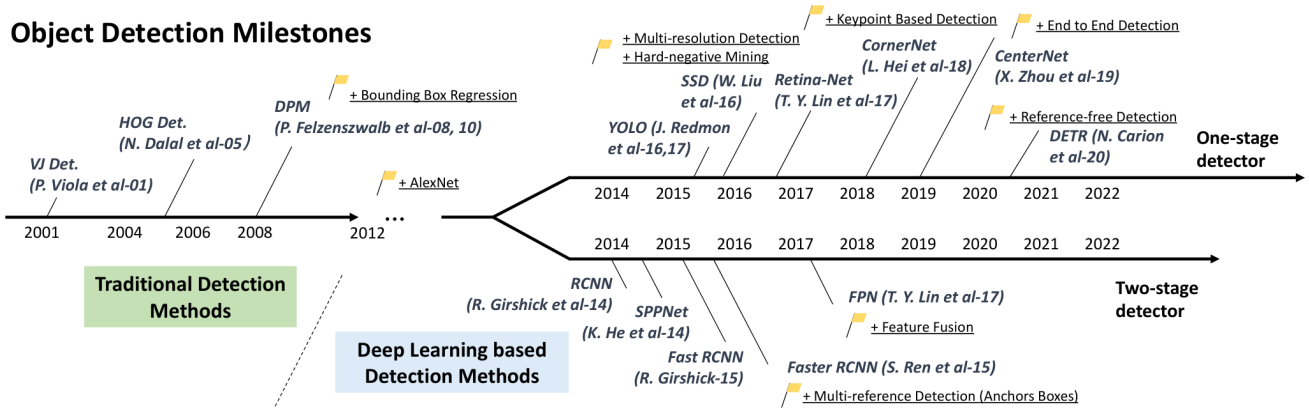


图 2：目标检测路线图。图中的里程碑检测器：VJ Det. [10, 11]、HOG Det. [12]、DPM [13–15]、RCNN [16]、SPPNet [17]、Fast RCNN [18]、Faster RCNN [19]、YOLO [20–22]、SSD [23]、FPN [24]、Retina-Net [25]、CornerNet [26]、CenterNet [27]、DETR [28]。

路线图区分传统与基于深度学习的检测方法，并按单阶段（one-stage）和双阶段（two-stage）分列深度学习方法。

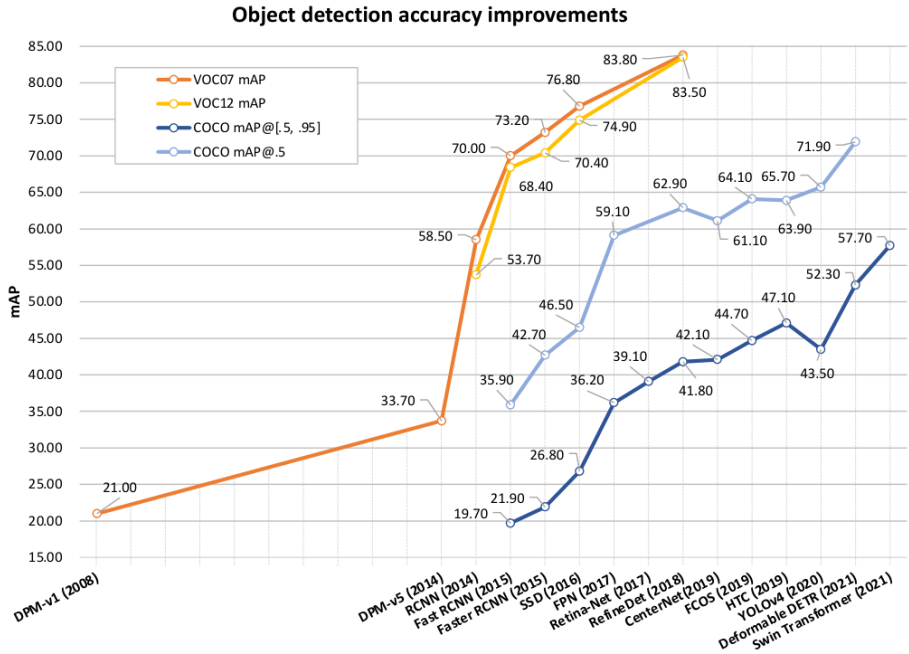


图 3：VOC07、VOC12 和 MS-COCO 数据集上目标检测精度的提升。图中的检测器：DPM-v1 [13]、DPM-v5 [37]、RCNN [16]、SPPNet [17]、Fast RCNN [18]、Faster RCNN [19]、SSD [23]、FPN [24]、Retina-Net [25]、RefineDet [38]、TridentNet [39]、CenterNet [40]、FCOS [41]、HTC [42]、YOLOv4 [22]、Deformable DETR [43]、Swin Transformer [44]。

纵轴为 mAP；VOC07、VOC12，及 COCO mAP@.5、mAP@[.5, .95] 使用不同曲线。

1) 里程碑：传统检测器：如果说当今的目标检测技术经历了一场由深度学习驱动的变革，那么回到 20 世纪 90 年代，我们会看到早期计算机视觉技术设计得何其巧妙，又富有远见。多数早期目标检测算法建立在手工设计的特征之上。由于当时缺少有效的图像表征，研究人员不得不设计复杂的特征表示，并采用多种加速技巧。

Viola–Jones 检测器：2001 年，P. Viola 和 M. Jones 首次在不附加任何约束（如先做肤色分割）的条件下实现了实时人脸检测 [10, 11]。这款检测器运行在 700 MHz 的 Pentium III CPU 上，在准确度相当的情况下，比同时期的其他算法快几十倍，甚至上百倍。VJ 检测器采用最直接的检测方

式——滑动窗口：遍历图像中所有可能的位置和尺度，检查各窗口内是否有人脸。这个过程看似简单，背后的计算量却远超当时计算机的能力。VJ 检测器通过结合三项关键技术大幅提升了检测速度："积分图像""特征选择"和"级联检测"（见第 III 节）。

HOG 检测器：2005 年，N. Dalal 和 B. Triggs 提出了方向梯度直方图（Histogram of Oriented Gradients，HOG）特征描述符 [12]。HOG 可视为对当时尺度不变特征变换 [29, 30] 和形状上下文 [31] 的重要改进。为了在特征的不变性（包括对平移、尺度、光照等变化的不变性）与非线性之间取得平衡，HOG 在由等间距单元格构成的密集网格上计算描述符，并在相互重叠的"块"内做局部对比度归一化。HOG 可用于检测多类目标，但最初主要是为解决行人检测问题而提出的。为了检测不同大小的目标，HOG 检测器保持检测窗口的尺寸不变，对输入图像进行多次缩放。多年来，HOG 检测器一直是许多目标检测器 [13, 14, 32] 和多种计算机视觉应用的重要基础。

可变形部件模型（DPM）：DPM 赢得了 VOC-07、-08 和 -09 检测挑战赛，是传统目标检测方法的代表。P. Felzenszwalb 于 2008 年最早提出 DPM [13]，作为对 HOG 检测器的扩展。DPM 采用"分而治之"的检测思想：训练可简单理解为学习怎样合理拆分目标，而推理可理解为汇集对该目标各部件的检测结果。例如，"汽车"检测可拆解为车窗、车身和车轮的检测。P. Felzenszwalb 等人将这项工作称为"星形模型" [13]。后来，R. Girshick 将星形模型进一步扩展为"混合模型"，以应对现实中变化更大的目标，并作了一系列其他改进 [14, 15, 33, 34]。

尽管现在的目标检测器在检测精度上早已超过 DPM，其中许多仍深受其重要思想的影响，包括混合模型、难负样本挖掘、边界框回归和上下文引导等。2010 年，PASCAL VOC 授予 P. Felzenszwalb 和 R. Girshick"终身成就"奖。

2) 里程碑：基于 CNN 的双阶段检测器：手工特征的性能趋于饱和，目标检测研究在 2010 年后进入平台期。2012 年，卷积神经网络（CNN）重新崛起 [35]。深层卷积网络能够学习图像的稳健高层特征表示，由此自然产生一个问题：能否将它用于目标检测？2014 年，R. Girshick 等人率先打破僵局，提出了利用 CNN 特征的区域检测方法（Regions with CNN features，RCNN）[16, 36]。自此，目标检测以前所未有的速度演进。深度学习时代有两类检测器："双阶段检测器"和"单阶段检测器"。前者把检测视为"由粗到细"的过程，后者则希望"一步完成"。

RCNN：RCNN 的思路很简单：首先用选择性搜索 [45] 提取一组候选区域（目标候选框）；接着把每个候选区域缩放成固定大小的图像，送入在 ImageNet 上预训练的 CNN（如 AlexNet [35]）以提取特征；最后用线性支持向量机（SVM）判断各区域是否包含目标，并识别其类别。RCNN 在 VOC07 上的表现明显提升，平均精度均值（mAP）从 DPM-v5 [46] 的 33.7% 大幅提高到 58.5%。尽管 RCNN 有很大进步，缺点也很明显：对大量重叠的候选区域（一张图像有 2,000 多个候选框）重复计算特征，导致检测非常慢（使用 GPU 时每张图像需 14 秒）。同年晚些时候提出的 SPPNet [17] 解决了这一问题。

SPPNet：2014 年，K. He 等人提出空间金字塔池化网络（SPPNet）[17]。此前的 CNN 要求输入尺寸固定，例如 AlexNet [35] 要求图像为 224×224。SPPNet 的主要贡献是引入空间金字塔池化（SPP）层，使 CNN 无须重新缩放图像或感兴趣区域，就能从任意尺寸的输入获得固定长度的表示。用 SPPNet 进行目标检测时，只需对整张图像计算一次特征图，随后可生成任意区域的定长

表示来训练检测器，从而避免重复计算卷积特征。SPPNet 的速度是 R-CNN 的 20 倍以上，同时没有牺牲检测准确性（VOC07 mAP = 59.2%）。SPPNet 虽然有效加快了检测，但仍有两个缺点：一是训练仍需多个阶段，二是只微调全连接层，忽略此前各层。次年提出的 Fast RCNN [18] 解决了这些问题。

Fast RCNN：2015 年，R. Girshick 提出了 Fast RCNN 检测器 [18]，进一步改进了 R-CNN 和 SPPNet [16, 17]。Fast RCNN 允许在相同的网络配置下同时训练检测器和边界框回归器。在 VOC07 数据集上，Fast RCNN 将 mAP 从 RCNN 的 58.5% 提高至 70.0%，检测速度是 R-CNN 的 200 倍以上。Fast RCNN 成功结合了 R-CNN 和 SPPNet 的优势，但其检测速度仍受候选区域生成环节的限制（详见第 II-C1 节）。接下来的问题自然是："我们能用 CNN 生成候选区域吗？"随后提出的 Faster R-CNN [19] 回答了这个问题。

Faster RCNN：2015 年，Fast RCNN 问世不久，S. Ren 等人提出了 Faster RCNN 检测器 [19, 47]。Faster RCNN 是首个接近实时的深度学习检测器（COCO mAP@.5 = 42.7%，VOC07 mAP = 73.2%，采用 ZF-Net [48] 时为 17 fps）。Faster RCNN 的主要贡献是引入候选区域网络（RPN），使生成候选区域的计算开销几乎可以忽略。从 R-CNN 到 Faster RCNN，目标检测系统的大多数独立模块——如候选区域生成、特征提取与边界框回归——逐步整合到统一的端到端学习框架中。Faster RCNN 虽然突破了 Fast RCNN 的速度瓶颈，之后的检测阶段仍有冗余计算。后来，研究者提出了 R-FCN [49] 和 Light-head RCNN [50] 等改进（详见第 III 节）。

特征金字塔网络（FPN）：2017 年，T.-Y. Lin 等人提出了 FPN [24]。在此之前，大多数基于深度学习的检测器只在网络顶层的特征图上进行检测。CNN 中较深层的特征有利于类别识别，却不利于目标定位。为此，FPN 构建了带横向连接的自顶向下的架构，在所有尺度上生成高层语义特征。CNN 在前向传播时自然会形成一个特征金字塔，FPN 因而能更好地检测尺度差异很大的目标。只需把 FPN 用于基础版 Faster R-CNN，不加其他技巧，就能在 COCO 数据集上取得当时领先的单模型检测成绩（COCO mAP@.5 = 59.1%）。如今，FPN 已成为许多新检测器的基本模块。

3) 里程碑：基于 CNN 的单阶段检测器：多数双阶段检测器遵循由粗到细的处理方式。粗检力求提高召回能力，精检则在粗检的基础上完善定位，更重视辨别能力。这类检测器无须额外技巧就能达到较高精度，但速度较慢、复杂度很高，因而很少用于工程实践。相反，单阶段检测器只需一次推理就能检出所有目标。它们易于部署且可实时运行，因而很受移动端青睐，但在检测密集目标和小目标时，性能会明显下降。

You Only Look Once（YOLO）：R. Joseph 等人于 2015 年提出 YOLO，这是深度学习时代首个单阶段检测器 [20]。YOLO 速度极快：快速版的速度为 155 fps，VOC07 mAP 为 52.7%；增强版的速度为 45 fps，VOC07 mAP 为 63.4%。YOLO 采用与双阶段检测器截然不同的思路：用一个神经网络处理整张图像。该网络把图像划分为若干区域，同时预测各区域的边界框和概率。尽管检测速度大幅提高，YOLO 的定位准确性仍不及双阶段检测器，尤其是在部分小目标上。YOLO 的后续版本 [21, 22, 51] 和随后提出的 SSD [23] 更加重视这一问题。近期，YOLOv4 团队进一步提出 YOLOv7 [52]；它引入动态标签分配与模型结构重参数化等优化设计，在速度和准确性上超越多数已有目标检测器（速度范围为 5 FPS 至 160 FPS）。

单次多框检测器（SSD）： W. Liu 等人于 2015 年提出 SSD [23]。SSD 的主要贡献是引入多参考和多分辨率检测技术（见第 II-C1 节），显著提升单阶段检测器的检测精度，尤其是对部分小目标的检测。SSD 兼具速度和精度优势（COCO mAP@.5 = 46.5%，快速版可达 59 fps）。SSD 与先前检测器的主要区别在于，它在网络的不同层上检测不同尺度的目标，先前的检测器则只在顶层进行检测。

RetinaNet： 单阶段检测器虽然快速、简单，但多年来检测精度一直落后于双阶段检测器。T.-Y. Lin 等人探究了其中的原因，并于 2017 年提出 RetinaNet [25]。他们发现，密集检测器训练时极端的前景与背景类别不平衡是核心原因。为此，RetinaNet 改造了标准交叉熵损失，引入一种名为“焦点损失”的新损失函数，以使检测器在训练时更关注容易误判的困难样本。焦点损失使单阶段检测器在保持很高检测速度的同时，达到与双阶段检测器相当的精度（COCO mAP@.5 = 59.1%）。

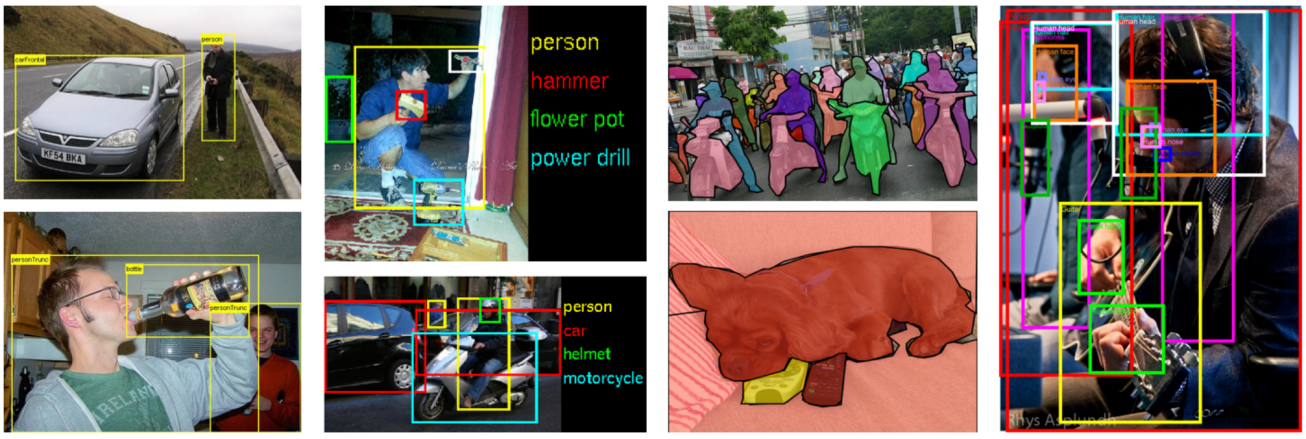
CornerNet： 此前的方法主要用锚框作为分类和回归的参考。目标的数量、位置、尺度和宽高比等经常变化。这些方法为了获得较好的性能，不得不设置大量参考框，以更好地匹配真实标注。然而，这会加剧类别不平衡，引入大量手工设置的超参数，并延长网络收敛时间。为解决这些问题，H. Law 等人 [26] 放弃此前的检测范式，把任务视为对边界框角点的预测。得到关键点后，CornerNet 利用额外的嵌入信息将分开预测的角点重新配对，组成边界框。CornerNet 的表现优于当时大多数单阶段检测器（COCO mAP@.5 = 57.8%）。

CenterNet： X. Zhou 等人于 2019 年提出 CenterNet [40]。它同样采用基于关键点的检测范式，但省去了开销较大的后处理，例如 CornerNet [26]、ExtremeNet [53] 等方法中的关键点分组分配，以及非极大值抑制（NMS），从而得到完全端到端的检测网络。CenterNet 把一个目标视为单点（目标中心），以中心点为参考回归该目标的各项属性，例如尺寸、方向、位置和姿态。该模型简洁，还能在同一框架内整合三维目标检测、人体姿态估计、光流学习、深度估计等任务。尽管检测思路如此简洁，CenterNet 的检测结果仍能达到相当水平（COCO mAP@.5 = 61.1%）。

DETR： 近年来，Transformer 对整个深度学习领域产生了深远影响，计算机视觉尤为如此。为克服 CNN 的局限并获得全局感受野，Transformer 放弃传统卷积算子，只使用注意力计算。2020 年，N. Carion 等人提出 DETR [28]，将目标检测视为集合预测问题，并借助 Transformer 构建端到端的检测网络。至此，目标检测进入了无需锚框或锚点即可检测目标的新阶段。后来，X. Zhu 等人针对 DETR 收敛慢、小目标检测性能有限的问题，提出 Deformable DETR [43]。该方法在 MS-COCO 数据集上取得了当时的领先成绩（COCO mAP@.5 = 71.9%）。

B. 目标检测数据集和评价指标

1) 数据集： 构建规模更大、偏差更小的数据集，对开发先进的检测算法至关重要。过去十年发布了多种知名检测数据集，包括 PASCAL VOC 挑战赛数据集 [54, 55]（如 VOC2007、VOC2012）、ImageNet 大规模视觉识别挑战赛数据集（如 ILSVRC2014）[56]、MS-COCO 检测挑战赛数据集 [57]、Open Images 数据集 [58, 59]、Objects365 [60] 等。表 I 给出了这些数据集的统计数据。图 4 展示了其中一些示例图像。图 3 显示了 2008 至 2021 年间目标检测在 VOC07、VOC12 和 MS-COCO 数据集上取得的精度提升。



(a) (b) (c) (d)

图 4：(a) PASCAL-VOC07、(b) ILSVRC、(c) MS-COCO 和 (d) Open Images 的示例图像及标注。

从左到右为 (a) 至 (d)；图中保留各数据集的原始英文标注。

Dataset	train		validation		trainval		test	
	images	objects	images	objects	images	objects	images	objects
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：若干知名目标检测数据集及其统计数据。

train = 训练集；validation = 验证集；trainval = 训练与验证合并集；test = 测试集；images / objects = 图像数 / 目标数。Objects365-2019 的测试目标数原表写作“1,700,00”。

Pascal VOC：PASCAL 视觉目标类别（VOC）挑战赛[1]（2005 至 2012 年）[54, 55] 是早期计算机视觉界最重要的竞赛之一。目标检测最常用的是 Pascal VOC 的两个版本：VOC07 和 VOC12。前者包含约 5,000 张训练图像和 12,000 个标注目标，后者包含约 11,000 张训练图像和 27,000 个标注目标。这两个数据集对日常生活中常见的 20 类目标进行了标注，如“人”“猫”“自行车”“沙发”等。

ILSVRC：ImageNet 大规模视觉识别挑战赛（ILSVRC）[2] [56] 推动了通用目标检测的技术前沿。该比赛于 2010 至 2017 年每年举办一次，其中包括一项使用 ImageNet 图像 [61] 的检测挑战。ILSVRC 检测数据集包含 200 类视觉目标，其图像数量与目标实例数量均比 VOC 大两个数量级。

MS-COCO：MS-COCO[3] [57] 是当时最具挑战性的目标检测数据集之一。自 2015 年起，基于 MS-COCO 的竞赛每年举办。与 ILSVRC 相比，MS-COCO 的目标类别更少，但目标实例更多。例如，MS-COCO-17 包含约 164,000 张图像，在 80 个类别中标注了约 897,000 个目标。相比 VOC 和 ILSVRC，MS-COCO 最大的进步是：除边界框标注外，还对每个目标做逐实例的分割标注，以便精确定位。此外，MS-COCO 包含更多小目标（面积不足图像的 1%），目标排列也更密集。正如 ImageNet 当年一样，MS-COCO 已成为目标检测领域的事实标准。

Open Images：2018 年推出的 Open Images 检测挑战赛（OID）[4] [62] 继 MS-COCO 之后出现，且规模空前。Open Images 包含两项任务：1）标准目标检测；2）视觉关系检测，即检测具有特定关系的成对目标。用于标准检测任务的数据集包含约 1,910,000 张图像、约 15,440,000 个标注边界框，覆盖 600 个目标类别。

2) 评价指标：如何衡量检测器的准确性？不同时期的答案可能不同。目标检测研究早期，并无广泛认可的准确性评价指标。例如，早期行人检测研究 [12] 通常使用“漏检率与每窗口假阳性数（FPPW）的关系”作为评价指标。然而，按窗口度量存在缺陷，无法预测整幅图像上的表现 [63]。2009 年，Caltech 行人检测基准问世 [63, 64]；此后，指标从 FPPW 改为每图像假阳性数（FPPI）。

近些年最常用的检测评价指标是“平均精度（Average Precision，AP）”，它最初由 VOC2007 引入。论文将 AP 定义为不同召回率下检测精度的平均值，通常按类别分别计算。对所有类别的 AP 取均值，即平均精度均值（mAP），通常作为最终性能指标。为度量目标定位的准确性，要检验预测框与真实框的交并比（IoU）是否超过预设阈值，例如 0.5。超过则将目标判为“检出”，否则判为“漏检”。以 0.5 为 IoU 阈值的 mAP 随后成为目标检测的事实评价标准。

2014 年后，随着 MS-COCO 数据集的引入，研究人员开始更重视目标定位的准确性。MS-COCO 的 AP 不采用单一固定的 IoU 阈值，而是在 0.5 至 0.95 之间的多个 IoU 阈值上取平均。这有助于推动更精确的目标定位，对一些实际应用可能很重要，例如机器人要抓取一把扳手。

C. 目标检测技术的演进

本节介绍检测系统的一些关键模块及其技术演进。先介绍模型设计中的多尺度检测与上下文引导，再介绍训练过程中的样本选择策略与损失函数设计，最后介绍推理中的非极大值抑制。图表和正文中的时间以论文发表时间为准。图中所示的演进顺序主要是为便于理解，各阶段在时间上可能重叠。

1) 多尺度检测的技术演进：对“不同尺寸”和“不同宽高比”的目标进行多尺度检测，是目标检测的主要技术挑战之一。过去二十年，多尺度检测经历了多个历史阶段，如图 5 所示。

特征金字塔 + 滑动窗口：VJ 检测器之后，研究者开始更重视一种更直观的检测方式，即构建“特征金字塔 + 滑动窗口”。从 2004 年起，多种里程碑检测器采用了这种范式，包括 HOG 检测器、DPM 等。这些方法通常在图像上滑动一个固定大小的检测窗口，不太考虑目标“宽高比不同”的问题。为检测外观更复杂的目标，R. Girshick 等人开始寻找特征金字塔之外的更好方案。一种当时的解法是“混合模型” [15]，即训练多个检测器，分别检测不同宽高比的目标。此外，基于样例的检测 [32, 70] 提供了另一种解法：为每个目标实例（样例）分别训练模型。

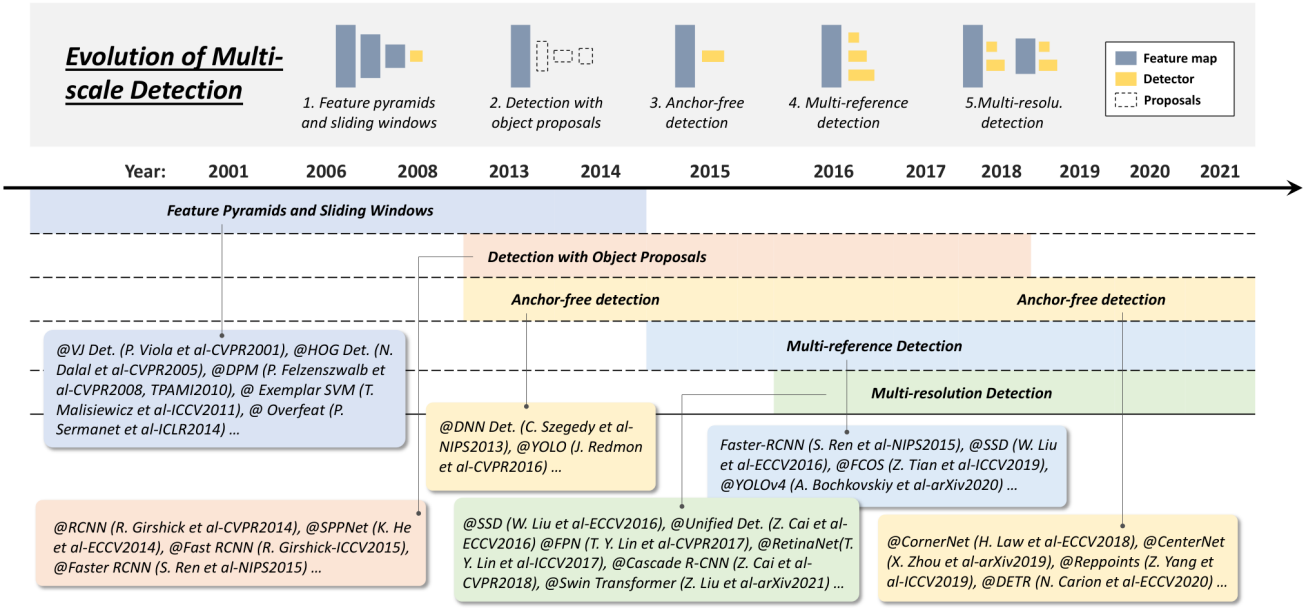


图 5：目标检测中多尺度检测技术的演进。图中的检测器：VJ Det. [10]、HOG Det. [12]、DPM [13]、Exemplar SVM [32]、Overfeat [65]、RCNN [16]、SPPNet [17]、Fast RCNN [18]、Faster RCNN [19]、DNN Det. [66]、YOLO [20]、SSD [23]、Unified Det. [67]、FPN [24]、RetinaNet [25]、RefineDet [38]、Cascade R-CNN [68]、Swin Transformer [44]、FCOS [41]、YOLOv4 [22]、CornerNet [26]、CenterNet [40]、Reppoints [69]、DETR [28]。

图上方依次为特征金字塔与滑动窗口、基于候选区域的检测、无锚框检测、多参考检测和多分辨率检测。

基于候选区域的检测：候选区域指一组不区分类别、可能包含目标的参考框。利用候选区域进行检测，可避免在图像上用滑动窗口穷举搜索。关于这一主题的完整综述，参见文献 [71, 72]。早期的候选区域生成方法采用自底向上的思路 [73, 74]。2014 年以后，随着深层 CNN 在视觉识别中的普及，自顶向上、基于学习的方法在这一问题上开始展现更多优势 [19, 75, 76]。单阶段检测器兴起后，候选区域生成逐渐淡出了研究视野。

深度回归与无锚框检测：近年来，随着 GPU 算力的提升，多尺度检测逐渐变得更直接，处理方式也更简单粗放。利用深度回归解决多尺度问题的思路很简单：直接根据深度学习特征预测边界框的坐标 [20, 66]。2018 年以后，研究人员开始从关键点检测的角度思考目标检测。这些方法通常遵循两种思路：一类依赖分组，先检测关键点（角点、中心点或代表点），再按目标将关键点分组 [26, 53, 69, 77]；另一类无须分组，将目标视为一个或多个点，再以这些点为参照回归目标的属性（大小、宽高比等）[40, 41]。

多参考检测与多分辨率检测：多参考检测是当时最常用的多尺度检测方法 [19, 22, 23, 41, 47, 51]。其主要思路 [19, 22, 23, 41, 47, 51] 是先在图像的每个位置定义一组参考（也称锚，包括锚框和锚点），然后基于这些参考预测检测框。另一种常用技术是多分辨率检测 [23, 24, 44, 67, 68]，即在网络的不同层上检测不同尺度的目标。多参考检测与多分辨率检测已成为当时先进目标检测系统的两个基本模块。

2) 上下文引导的技术演进：视觉目标通常处于由周围环境构成的典型上下文中。人脑利用目标与环境的关联来促进视觉感知与认知 [96]。目标检测也早已利用上下文引导来改善表现。图 6 展示了目标检测中上下文引导技术的演进。

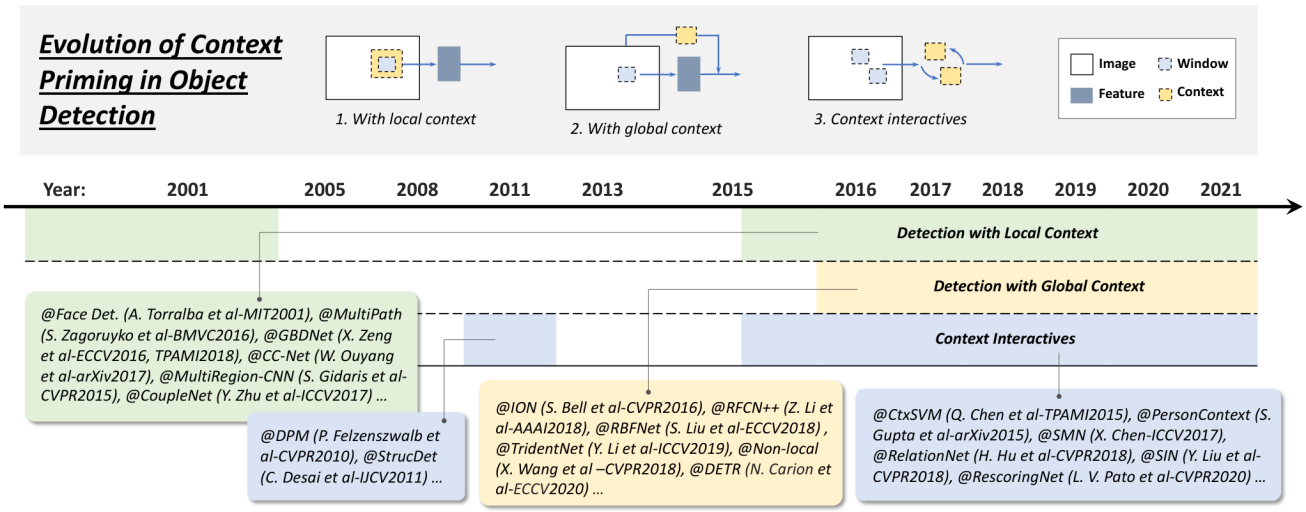


图 6：目标检测中上下文引导方法的演进。图中的检测器：Face Det. [78]、MultiPath [79]、GBDNet [80, 81]、CC-Net [82]、MultiRegion-CNN [83]、CoupleNet [84]、DPM [14, 15]、StructDet [85]、ION [86]、RFCN++ [87]、RBFNet [88]、TridentNet [39]、Non-local [89]、DETR [28]、CtxSVM [90]、PersonContext [91]、SMN [92]、RelationNet [93]、SIN [94]、RescoringNet [95]。

图区分局部上下文（local context）、全局上下文（global context）和上下文交互（context interactives）。

利用局部上下文进行检测：局部上下文是指待检测目标周围区域的视觉信息。人们早已意识到，局部上下文有助于改进目标检测。21 世纪初，Sinha 和 Torralba [78] 发现，加入脸部外轮廓等局部上下文区域能够显著改善人脸检测表现。Dalal 和 Triggs 也发现，纳入少量背景信息可提高行人检测的精度 [12]。对近期基于深度学习的检测器，只要扩大网络的感受野或增大候选区域，也能利用局部上下文改善检测效果 [79–84, 97]。

利用全局上下文进行检测：全局上下文方法将场景构成作为目标检测的额外信息来源。早期检测器通常通过整合场景组成元素的统计摘要来引入全局上下文，例如 Gist [96]。较新的检测器有两种整合全局上下文的方式。第一种是利用深层卷积、空洞卷积、可变形卷积和池化操作 [39, 87, 88] 获得更大的感受野（甚至超过输入图像）。后来，研究者尝试使用注意力机制（如非局部模块、Transformer 等）来获得覆盖整幅图像的感受野，并取得了很好的效果 [28, 89]。第二种方法是将全局上下文视为一种序列信息，用循环神经网络来学习 [86, 98]。

上下文交互：上下文交互指视觉元素之间的约束和依赖关系。近期研究表明，考虑上下文交互能够改进现代检测器。近期的一些改进可归为两类：一类探索各个目标之间的关系 [15, 85, 90, 92, 93, 95]，另一类探索目标与场景之间的依赖关系 [91, 94]。

3) 难负样本挖掘的技术演进：训练检测器本质上是一个不平衡学习问题。对基于滑动窗口的检测器，背景与目标的比例甚至可能高达 10^7:1 [71]。此时，使用全部背景样本反而会损害训练，因为海量易分类的负样本将淹没学习过程。难负样本挖掘（HNM）旨在解决这个问题。图 7 展示了 HNM 的技术演进。

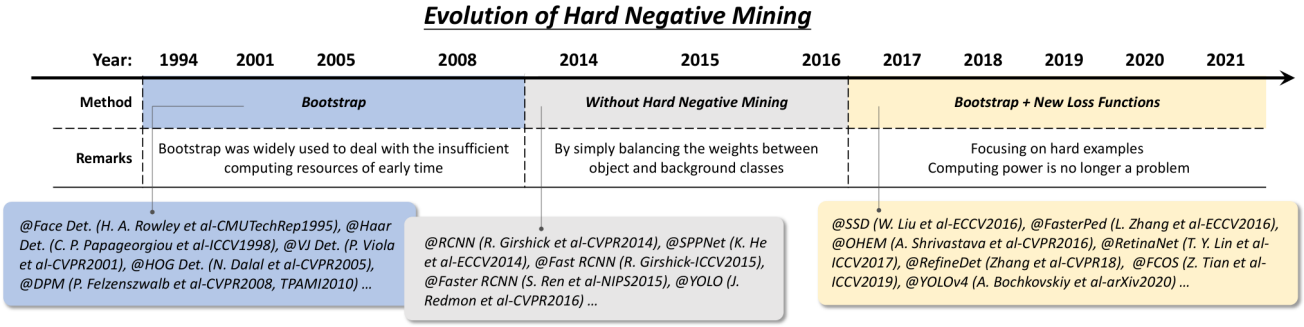


图 7：目标检测中难负样本挖掘技术的演进。图中的检测器：Face Det. [99]、Haar Det. [100]、VJ Det. [10]、HOG Det. [12]、DPM [13, 15]、RCNN [16]、SPPNet [17]、Fast RCNN [18]、Faster RCNN [19]、YOLO [20]、SSD [23]、FasterPed [101]、OHEM [102]、RetinaNet [25]、RefineDet [38]、FCOS [41]、YOLOv4 [22]。

时间轴上的方法分组依次为自举、不使用难负样本挖掘、自举结合新损失函数。

自举：目标检测中的自举指一类训练技术：训练先从一小部分背景样本开始，再迭代加入新出现的误分类样本。早期检测器通常用自举来降低处理数百万背景样本的训练开销 [10, 99, 100]。后来，自举成为 DPM 和 HOG 检测器解决数据不平衡问题的标准技术 [12, 13]。

基于深度学习的检测器中的难负样本挖掘：在深度学习时代，由于算力提升，2014 至 2016 年间自举一度在目标检测中被弃用 [16–20]。为缓解训练中的数据不平衡，Faster RCNN、YOLO 等检测器仅对正、负窗口的权重进行平衡。后来研究人员发现，这无法彻底解决不平衡问题 [25]。为此，2016 年以后自举再次被引入目标检测 [23, 38, 101, 102]。另一种改进是设计新的损失函数 [25]：改造标准交叉熵损失，使其更关注被错误分类的困难样本 [25]。

4) 损失函数的技术演进：损失函数衡量模型与数据的吻合程度，即预测与真实标签之间的偏差。计算损失可以得到模型权重的梯度，随后用反向传播更新权重，使模型更好地拟合数据。分类损失和定位损失共同构成目标检测的监督信号，见式（1）。损失函数的一般形式可写为：

$$L(p, p^*, t, t^*) = L_{cls.}(p, p^*) + \beta I(t) L_{loc.}(t, t^*)$$
$$I(t) = \begin{cases} 1 & \text{IoU}\{a, a^*\} > \eta \\ 0 & \text{else} \end{cases} \qquad (1)$$

式 (1)：若参考的 IoU 超过阈值 η，则 I(t) = 1，否则 I(t) = 0。

其中，t 和 t∗ 分别为预测框和真实框的位置，p 和 p∗ 分别是其类别概率。IoU{a, a∗} 是参考框或参考点 a 与其真实标注 a∗ 之间的交并比。η 是 IoU 阈值，例如 0.5。如果某个锚框或锚点不匹配任何目标，其定位损失就不计入最终损失。

分类损失：分类损失用于衡量预测类别与真实类别之间的偏差。在采用 MSE/L2 损失（均方误差）的 YOLOv1 [20] 和 YOLOv2 [51] 等早期工作中，分类损失未得到充分研究。后来通常采用 CE 损失（交叉熵）[21, 23, 47]。L2 损失衡量欧氏空间中的差异，而 CE 损失可以衡量分布之间的差异（可视为一种似然形式）。由于分类的预测值是概率，CE 损失比 L2 损失更适合这一问题，因为它对误分类的惩罚更大，梯度消失的影响也更小。为改善分类效果，研究者提出标签平滑来增强模型泛化能力，缓解含噪标签造成的过度自信 [103, 104]；焦点损失则旨在解决类别不平衡和分类难度不一的问题 [25]。

定位损失：定位损失用于优化位置和大小的偏差。L2 损失在早期工作中较为常见 [16, 20, 51]，但它容易受异常值影响，并容易导致梯度爆炸。结合 L1 损失与 L2 损失的优点，研究者提出了平滑 L1 损失（Smooth L1）[18]，如下式所示：

$$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{else} \end{cases} \qquad (2)$$

式 (2)：|x| < 1 时取第一行，否则取第二行。

其中，x 表示目标值与预测值之差。计算误差时，上述损失将表示边界框的四个数值 (x, y, w, h) 视为独立变量，但它们之间实际存在相关性。此外，评价时用 IoU 判断预测框是否与真实框相匹配。相同的平滑 L1 损失值可能对应截然不同的 IoU 值，因此引入如下的 IoU 损失 [105]：

$$\text{IoU loss} = -\log(\text{IoU}) \qquad (3)$$

式 (3)：IoU 损失。

此后，有多种算法改进了 IoU 损失。G-IoU（广义交并比）[106] 改善了在边界框不重叠（即 IoU = 0）时 IoU 损失无法优化的问题。根据 Distance-IoU [107]，有效的检测回归损失应当考虑三个几何量：重叠面积、中心点距离和宽高比。因此，在 IoU 损失和 G-IoU 损失的基础上，DIoU（距离交并比）引入预测框与真实框中心点之间的距离；CIoU（完全交并比）[107] 则在 DIoU 的基础上考虑了宽高比的差异。

5) 非极大值抑制的技术演进：由于相邻窗口的检测得分通常相近，非极大值抑制被用作后处理步骤，去除重复的边界框以获得最终检测结果。目标检测早期不一定包含 NMS [121]，因为当时对目标检测系统应输出什么还没有完全明确的认识。图 8 展示了过去二十年间 NMS 的演进。

贪心选择：贪心选择是一种历史较早、但最常用的 NMS 方法。其思路简单直观：对一组重叠的检测结果，保留检测得分最高的边界框，再按预设的重叠阈值删除其相邻框。虽然贪心选择已成为 NMS 的事实标准方法，但仍有改进空间。首先，得分最高的框未必与目标最吻合；其次，它可能抑制临近的其他目标；最后，它不能抑制假阳性 [116]。针对上述问题，已有多项研究提出了解法 [107, 112, 114, 115]。

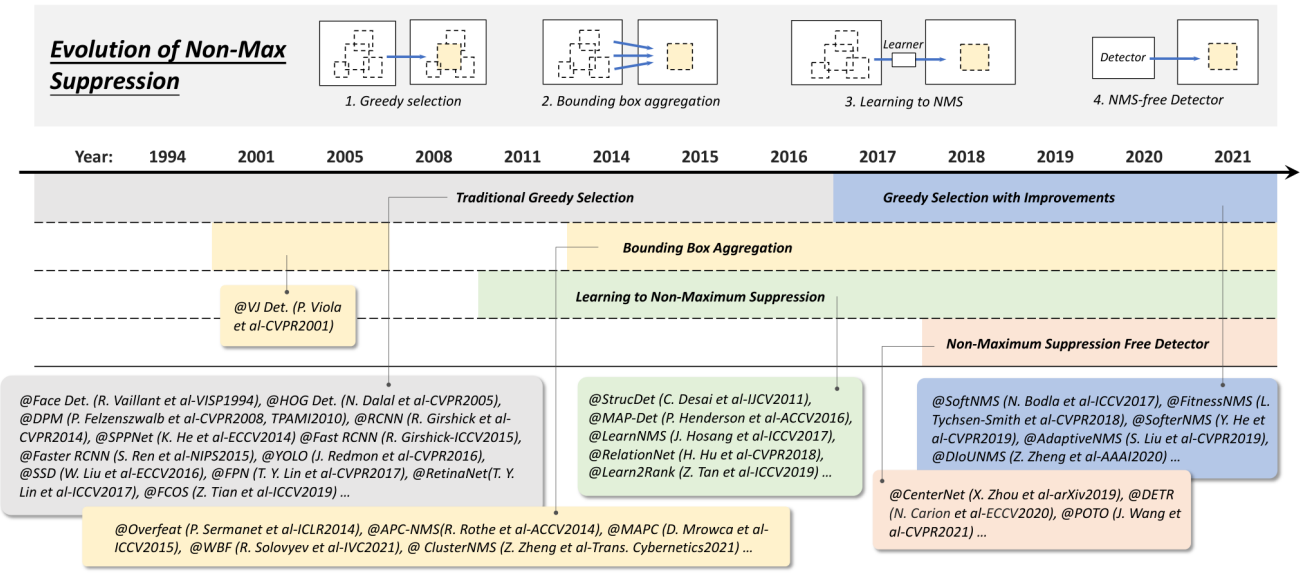


图 8：1994 至 2021 年间，目标检测中非极大值抑制（NMS）技术的演进：1）贪心选择；2）边界框聚合；3）基于学习的 NMS；4）无 NMS 检测。图中的检测器：Face Det. [108]、HOG Det. [12]、DPM [13, 15]、RCNN [16]、SPPNet [17]、Fast RCNN [18]、Faster RCNN [19]、YOLO [20]、SSD [23]、FPN [24]、RetinaNet [25]、FCOS [41]、StrucDet [85]、MAP-Det [109]、LearnNMS [110]、RelationNet [93]、Learn2Rank [111]、SoftNMS [112]、FitnessNMS [113]、SofterNMS [114]、AdaptiveNMS [115]、DIoUNMS [107]、Overfeat [65]、APC-NMS [116]、MAPC [117]、WBF [118]、ClusterNMS [119]、CenterNet [40]、DETR [28]、POTO [120]。

可沿时间轴对照图注中的四组方法及其代表检测器。

边界框聚合：边界框聚合是另一类 NMS 技术 [10, 65, 116, 117]。它将多个重叠的边界框合并或聚类为一个最终检测结果。这类方法的优点是充分考虑目标间的关系及其空间分布 [118, 119]。一些知名的检测器采用这一方法，如 VJ 检测器 [10] 和 Overfeat（ILSVRC-13 定位任务冠军）[65]。

基于学习的 NMS：近来，基于学习的 NMS 是备受关注的一类 NMS 改进 [85, 93, 109–111, 122]。其主要思路是把 NMS 当作一个为所有原始检测结果重新评分的过滤器，在网络中对其进行端到端训练，或另训练一个网络来模仿 NMS 的行为。与传统手工设计的 NMS 方法相比，这类方法在改善遮挡目标和密集目标的检测方面取得了有前景的结果。

无 NMS 检测器：为摆脱 NMS，获得可完全端到端训练的目标检测网络，研究人员开发了一系列方法来实现一对一标签分配（一个目标只对应一个预测框）[28, 40, 120]。这类方法通常遵循一个准则：使用质量最高的框来训练，从而实现无 NMS 检测。无 NMS 的检测器更接近人类视觉感知系统，也是目标检测未来的一个可能方向。

III. 检测加速

如何加速检测器，长期以来都是一个难题。图 9 将目标检测的加速技术分为三个层次：加速“检测流程”“检测器的骨干网络”以及“数值计算”。更详细的内容见文献 [123]。

A. 特征图共享计算

在检测器的各个计算环节中，特征提取通常占据大部分计算量。减少特征计算冗余最常用的办法，是对整张图像只计算一次特征图 [18, 19, 124]，由此实现数十倍甚至数百倍的加速。

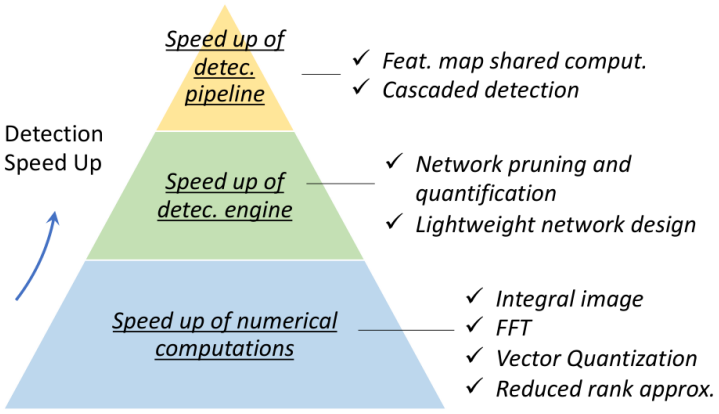


图 9： 目标检测加速技术概览。

由上至下为检测流程、检测引擎和数值计算的加速；各自对应右侧的方法。

B. 级联检测

级联检测是常用技术之一 [10, 125]。它采用由粗到细的检测思路：先用简单计算滤除大部分易判别的背景窗口，再用复杂计算处理较难判别的窗口。近些年，级联检测尤其常用于“大场景中的小目标”检测任务，如人脸检测 [126, 127]、行人检测 [101, 124, 128] 等。

C. 网络剪枝与量化

“网络剪枝”和“网络量化”是两种常见的 CNN 加速方法。前者剪去部分网络结构或权重，后者缩短编码长度。“网络剪枝”的研究最早可追溯至 20 世纪 80 年代 [129]。近期的网络剪枝通常交替进行训练和剪枝，即每个训练阶段后只去掉一小部分不重要的权重，然后重复这一过程 [130]。近期的网络量化工作主要关注网络二值化，即将激活值或权重量化为二进制变量（例如 0/1）来压缩网络，从而将浮点运算转换成逻辑运算。

D. 轻量化网络设计

加速基于 CNN 的检测器的最后一类方法，是直接设计轻量化网络。除“更少的通道、更多的层”之类的一般设计原则 [131] 外，近年来还提出了其他方法 [132–136]。

1) 卷积分解： 卷积分解是构建轻量化 CNN 最直接的方式。分解方法有两类。第一类是将一个大卷积核分解为若干小卷积核 [50, 87, 137]，如图 10(b) 所示。例如，可把一个 7×7 的卷积核分解为三个 3×3 的卷积核；两种方式的感受野相同，后者却更高效。第二类是在通道维度上分解卷积 [138, 139]，见图 10(c)。

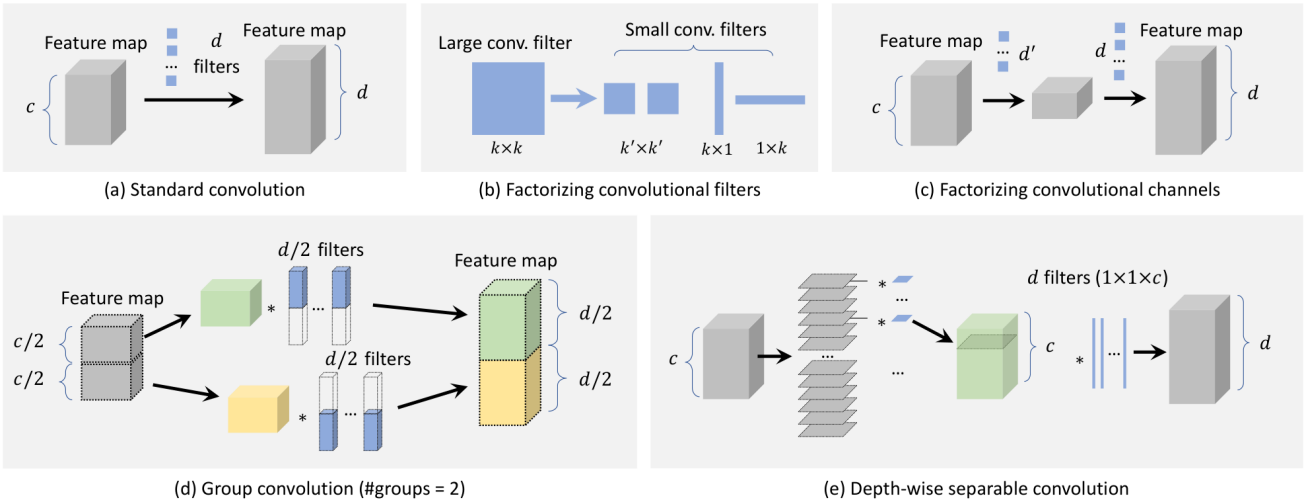


图 10：CNN 卷积层加速方法及其计算复杂度比较。(a) 标准卷积：$O(dk^2c)$；(b) 卷积核分解（k×k → (k′×k′)²，或 1×k、k×1）：$O(dk'^2c)$ 或 $O(dkc)$；(c) 卷积通道分解：$O(d'k^2c)+O(dk^2d')$；(d) 分组卷积（分组数 = m）：$O(dk^2c/m)$；(e) 深度可分离卷积：$O(ck^2)+O(dc)$。

示意图中 c、d 分别表示输入、输出的通道数；(a)–(e) 对应图注中的五种计算结构。

2) 分组卷积：分组卷积旨在减少卷积层的参数数量。它把特征通道分成若干组，然后分别对每组独立进行卷积 [140, 141]，如图 10(d) 所示。如果将特征均分为 m 组，且其他配置保持不变，计算量理论上会降至原来的 1/m。

3) 深度可分离卷积：如图 10(e) 所示，深度可分离卷积 [142] 可视为分组卷积的一个特例，此时分组数等于通道数。通常再用若干 1×1 卷积核来变换维度，使最终输出具有所需的通道数。使用深度可分离卷积，可将计算量从 $O(dk^2c)$ 降至 $O(ck^2)+O(dc)$。近年来，这一思路已用于目标检测和细粒度分类 [143–145]。

4) 瓶颈设计：与前面的层相比，神经网络中的瓶颈层结点较少。近年来，瓶颈设计广泛用于设计轻量化网络 [50, 133, 146–148]。其中一种做法是压缩检测器的输入层，从检测一开始就减少计算量 [133, 146, 147]。也可以将特征图压缩得更薄，以加快后续检测 [50, 148]。

5) 采用 NAS 的检测：基于深度学习的检测器日益复杂，高度依赖人工设计的网络架构和训练参数。神经架构搜索（NAS）主要关注如何定义合理的候选网络空间，如何快速、准确地搜索，以及如何以较低成本验证搜索结果。在设计检测模型时，NAS 可以减少在骨干网络和锚框设计中对人工干预的需求 [149–155]。

E. 数值计算加速

数值计算加速旨在从实现的底层加快目标检测器的运行。

1) 用积分图像加速：积分图像是图像处理中的一种重要方法，可快速计算图像子区域内的求和。积分图像的原理是信号处理中卷积的积分与微分可分离性：

$$f(x) * g(x) = \left(\int f(x)dx\right) * \left(\frac{dg(x)}{dx}\right), \qquad (4)$$

如果 dg(x)/dx 是稀疏信号，就可以用等式右侧的形式加速卷积 [10, 156]。积分图像还能加快目标检测中更一般的特征计算，例如颜色直方图、梯度直方图 [124, 157–159] 等。一个典型例子是通过计算积分 HOG 特征图来加速 HOG [124, 157]，见图 11。积分 HOG 特征图已用于行人检测，并在准确性不受损的前提下实现了数十倍加速 [124]。

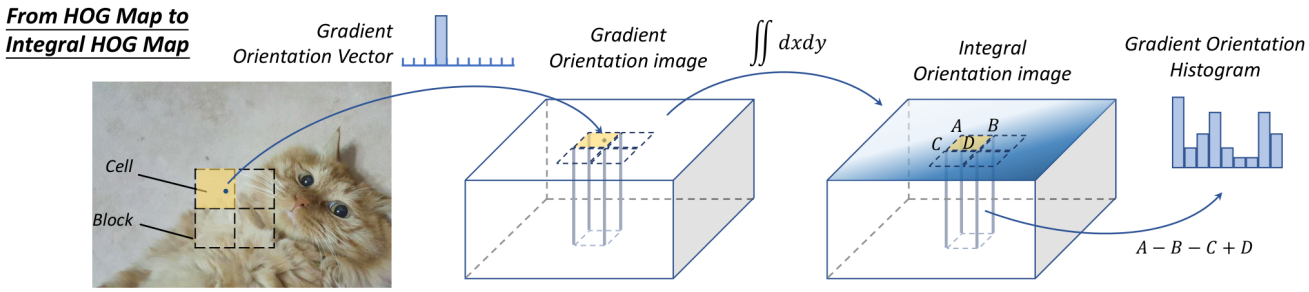


图 11：“积分 HOG 特征图”的计算示意 [124]。借助积分图像技术，可以用常数计算复杂度高效计算任意位置、任意尺寸区域的直方图特征。

cell = 单元格，block = 块。右侧示例用 A−B−C+D 从积分图计算区域直方图。

2) 频域加速：卷积是目标检测中的一类重要数值运算。线性检测器的检测可视为在每个窗口内对特征图和检测器权重做内积，而这可通过卷积实现。傅里叶变换是加快卷积的一种实用方法，其理论基础是信号处理中的卷积定理：在适当条件下，两个信号的卷积 $I * W$ 的傅里叶变换 F，等于它们在频域中的逐点乘积：

$$I * W = F^{-1}(F(I) \odot F(W)) \qquad (5)$$

其中 F 为傅里叶变换，F^{-1} 为逆傅里叶变换，$\odot$ 为逐点乘积。上述计算可以用快速傅里叶变换（FFT）和逆快速傅里叶变换（IFFT）加速 [160–163]。

3) 矢量量化：矢量量化（VQ）是信号处理中的经典量化方法，旨在用一小组原型向量来近似大量数据的分布。它可用于数据压缩，也可加快目标检测中的内积运算 [164, 165]。

IV. 目标检测的近期进展

过去二十年，不断涌现的新技术对目标检测产生了相当大的影响，但其基本原理和底层逻辑并未改变。前几节在较大的时间跨度上梳理了二十年来的技术演进，帮助读者把握目标检测；本节则在较短的时间范围内聚焦近年的先进算法，帮助读者深入理解目标检测。其中一部分是前述技术的延伸（如第 IV-A 至 IV-E 节），另一部分则是不同概念相结合的新方法（如第 IV-F 至 IV-H 节）。

A. 超越滑动窗口的检测

图像中的目标可由其真实边界框的左上角和右下角唯一确定，因此检测任务可等价表述为成对关键点的定位问题。近期实现这一思路的一种方式，是预测角点的热力图 [26]。还有一些方法沿用这一思路，使用更多关键点（角点和中心点 [77]、极值点和中心点 [53]、代表点 [69]）来获得更好的性能。另一种范式将目标视为一个或多个点，无须分组，直接预测目标的属性（如高度和宽度）。这种做法的优点是可在语义分割框架下实现，而且无须设计多尺度锚框。进一步地，DETR [28, 43] 将目标检测视为集合预测，完全摆脱了基于预设参考的框架。

B. 对旋转和尺度变化具有鲁棒性的检测

近年来，研究者致力于使检测对旋转和尺度变化保持鲁棒。

1) 旋转鲁棒检测：在人脸检测、文本检测和遥感目标检测中，目标旋转很常见。最直接的办法是进行数据增强，使增强后的数据分布能覆盖目标的各种朝向 [166]；也可以为每种朝向分别训练独立的检测器 [167, 168]。近期一种常用方法是设计旋转不变的损失函数：对检测损失加上约束，使目标旋转后特征保持不变 [169–171]。另一种近来的解法是学习候选目标的几何变换 [172–175]。在双阶段检测器中，感兴趣区域（ROI）池化旨在从任意位置、任意大小的候选区域提取定长特征。由于特征池化通常在笛卡尔坐标系中进行，因此不具备旋转不变性。一项近期改进是在极坐标系中进行 ROI 池化，使特征对旋转变化保持鲁棒 [167]。

2) 尺度鲁棒检测：近期研究从训练和检测两个阶段入手，提高检测对尺度变化的鲁棒性。

尺度自适应训练：现代检测器通常将输入图像缩放到固定尺寸，并对所有尺度目标的损失进行反向传播。这会带来“尺度不平衡”问题。在检测阶段构建图像金字塔可以缓解问题，但无法从根本上解决 [49, 178]。一项近期改进是图像金字塔尺度归一化（Scale Normalization for Image Pyramids，SNIP）[176]。它在训练和检测阶段都构建图像金字塔，并且只对选定尺度上目标的损失进行反向传播，如图 12 所示。研究人员又提出了一种更高效的训练策略：结合高效重采样的 SNIP（SNIP with Efficient Resampling，SNIPER）[177]，即从图像中裁出一组子区域并重新缩放，从而利用大批量训练的优势。

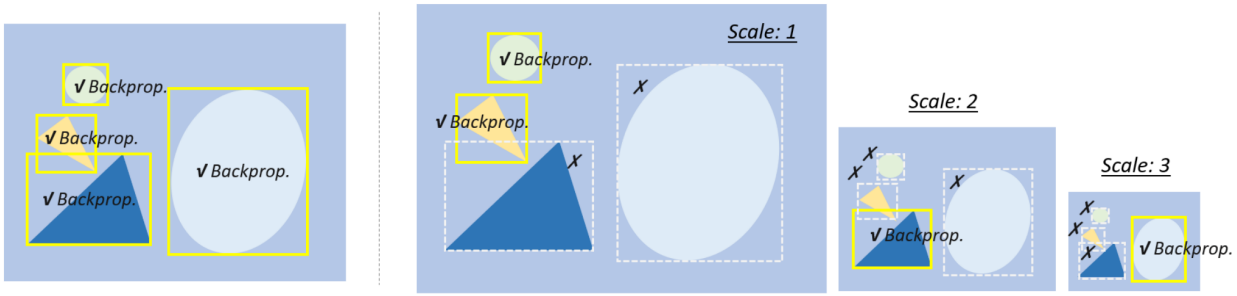


图 12：多尺度目标检测的两种训练策略。(a) 在单一分辨率图像上训练，对所有尺度目标的损失进行反向传播 [17–19, 23]。(b) 在多分辨率图像（图像金字塔）上训练，只对选定尺度的目标反向传播。若目标过大或过小，则丢弃其梯度 [39, 176, 177]。

带勾的 Backprop. 表示在当前分辨率下对该目标反向传播；叉号表示不对该目标反向传播。

尺度自适应检测：基于 CNN 的检测器通常会仔细设计锚框的尺寸和宽高比，其缺点是这些配置无法适应意料之外的尺度变化。为改善小目标检测，一些近期检测器提出了“自适应放大”技术，按需将小目标放大为“较大的目标” [179, 180]。另一项近期改进，是先预测图像内目标的尺度分布，再据此对图像进行自适应缩放 [181, 182]。

C. 采用更好的骨干网络进行检测

检测器的准确性和速度在很大程度上取决于特征提取网络，也就是骨干网络，如 ResNet [178]、CSPNet [183]、Hourglass [184] 和 Swin Transformer [44]。关于深度学习时代一些重要检测骨干网络的详细介绍，可参阅综述 [185]。图 13 展示了三种知名检测系统——Faster RCNN [19]、R-FCN [49] 和 SSD [23]——搭配不同骨干网络时的检测精度 [186]。近来，目标检测受益于 Transformer 强大的特征提取能力。在作者引用的 COCO 数据集排行榜上，前十名的检测方法都基于 Transformer[5]。Transformer 与 CNN 的性能差距逐渐扩大。

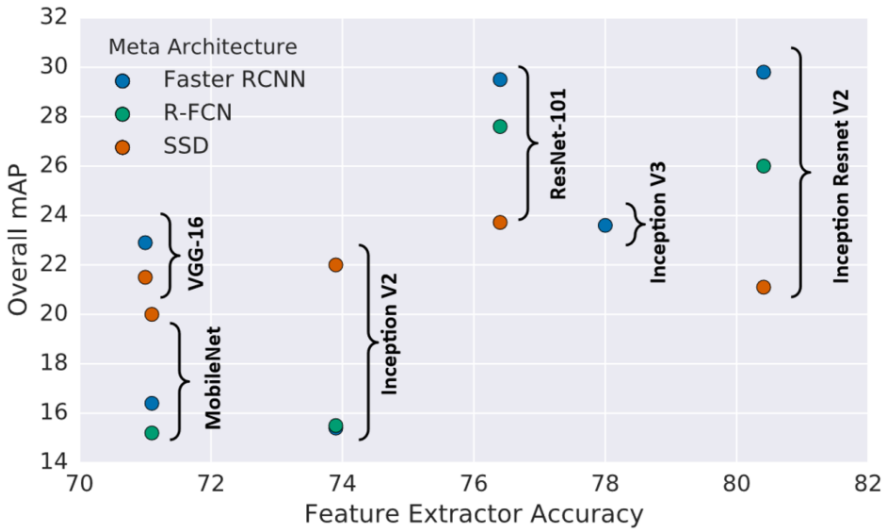


图 13：比较三种检测器在 MS-COCO 数据集上搭配不同检测骨干网络时的检测精度：Faster RCNN [19]、R-FCN [49] 和 SSD [23]。图片引自 J. Huang 等人的 CVPR 2017 论文 [186]。

横轴是特征提取网络自身的准确性，纵轴为整体检测 mAP；颜色区分三种检测框架。

D. 改进定位精度

近期检测器主要通过两类方法改善定位精度：1）边界框细化；2）面向精确定位的新损失函数。

1) 边界框细化：提高定位准确性最直观的方式是细化边界框，可视为对检测结果的后处理。近期一种做法是反复将检测结果输入边界框回归器，直到预测收敛到正确的位置和大小 [187–189]。不过，也有研究指出，这种方法无法保证定位准确性单调提高 [187]；进行多次细化反而可能使定位变差。

2) 面向精确定位的新损失函数：在大多数现代检测器中，目标定位被视为坐标回归问题，但这一范式的缺点很明显。首先，回归损失与最终的定位评价并不一致，尤其是某些宽高比很大的目标。其次，传统边界框回归不提供定位置信度。当多个边界框相互重叠时，这可能导致非极大值抑制失效。设计新的损失函数可缓解上述问题。最直观的改进是直接用 IoU 构造定位损失

[105–107, 190]。此外，一些研究者也尝试在概率推断框架下改进定位 [191]。与此前直接预测边界框坐标的方法不同，这种方法预测的是边界框位置的概率分布。

E. 利用分割损失学习

目标检测与语义分割是计算机视觉中的两项基础任务。近期研究表明，利用语义分割损失来学习，可以改进目标检测。

用分割改善检测，最简单的办法是把分割网络视为固定的特征提取器，将其输出作为辅助特征整合到检测器中 [83, 192, 193]。它的优点是容易实现，缺点是分割网络可能带来额外的计算开销。

另一种办法是在原有检测器上增加分割分支，并用多任务损失（分割 + 检测）训练模型 [4, 42, 192]。优点是在推理阶段移除分割分支，因此不会影响检测速度。缺点是训练需要像素级图像标注。

F. 对抗训练

A. Goodfellow 等人于 2014 年提出生成对抗网络（GAN）[194]。这种网络在图像生成 [194, 195]、图像风格迁移 [196] 和图像超分辨率 [197] 等多种任务中受到广泛关注。

近期，对抗训练也被用于目标检测，尤其是改善小目标及遮挡目标的检测。对于小目标，可用 GAN 缩小小目标与大目标的特征表示差距，增强小目标的特征 [198, 199]。为了改善遮挡目标检测，近期一种思路是用对抗训练生成遮挡掩码 [200]。对抗网络不是在像素空间中生成样本，而是直接修改特征以模拟遮挡。

G. 弱监督目标检测

训练基于深度学习的目标检测器，通常需要大量人工标注的数据。弱监督目标检测（WSOD）的目的是只用图像级标注而不用边界框来训练检测器，降低对数据标注的依赖 [201]。

多实例学习是一类监督学习算法，在 WSOD 中得到广泛应用 [202–209]。多实例学习模型不从逐一标注的实例集学习，而是接收一组有标签的“包”，每个包包含多个实例。若将一张图像中的候选目标视为一个包，并将图像级标注作为包的标签，WSOD 就可以表述为一个多实例学习过程。

类激活映射是另一类近期提出的 WSOD 方法 [210, 211]。关于 CNN 可视化的研究表明，即使没有对目标位置进行监督，CNN 的卷积层也会表现出类似目标检测器的行为。类激活映射让人们看到，如何使仅用图像级标签训练的 CNN 也具备定位能力 [212]。

除上述方法外，一些研究人员将 WSOD 视为候选区域排序过程：先选出信息量最大的区域，再用图像级标注对这些区域进行训练 [213]。另一些研究人员提出遮蔽图像的不同部分。如果检测得分急剧下降，被遮蔽的区域就很可能包含目标 [214]。再近一些的研究也将生成对抗训练用于 WSOD [215]。

H. 基于域适应的检测

大多数目标检测器的训练过程，本质上可视为在数据独立同分布（i.i.d.）假设下进行似然估计。非独立同分布数据上的目标检测仍是一项挑战，某些实际应用尤其如此。除了收集更多数据或采用合适的数据增强，域适应还提供了缩小域间差距的可能。为获得域不变的特征表示，已有研究在图像级、类别级或目标级探索基于特征正则化与对抗训练的方法 [216–221]。循环一致变换 [222] 也被用于弥合源域与目标域的差距 [223, 224]。还有些方法同时采用两种思路 [225]，以取得更好的性能。

V. 结论与未来方向

过去二十年，目标检测取得了显著成就。本文广泛回顾了这一领域二十年间的一些里程碑检测器、关键技术、加速方法、数据集和评价指标。为帮助读者从上述框架之外获得更多启发，下面列出了一些有前景的未来方向，但不限于这些方向。

轻量化目标检测：轻量化目标检测旨在加快检测推理，使其能在低功耗边缘设备上运行。重要应用包括移动增强现实、自动驾驶、智慧城市、智能摄像头和人脸核验等。尽管近些年已为此付出许多努力，机器与人眼之间仍有很大的速度差距，尤其是在检测某些小目标或利用多源信息进行检测时 [226, 227]。

端到端目标检测：尽管已有方法采用一对一标签分配进行训练，完全端到端地检测目标（由一个网络把图像直接转为边界框），多数方法仍采用一对多标签分配，将非极大值抑制操作另行设计。未来研究可着眼于设计兼具检测准确性与效率的端到端流程 [228]。

小目标检测：在大场景中检测小目标长期以来是一大难题。潜在应用包括统计人群中的人数或露天环境中的动物数量，以及从卫星图像中检测军事目标。可进一步探索的方向包括融合视觉注意力机制，以及设计高分辨率轻量化网络 [229, 230]。

三维目标检测：尽管二维目标检测近来有所进展，自动驾驶等应用仍依赖于获取目标在三维世界中的位置与姿态。未来，三维目标检测及多源、多视角数据的利用将得到更多关注，例如利用多个传感器采集的 RGB 图像和三维激光雷达点 [231, 232]。

视频中的检测：在高清视频中实时检测和跟踪目标，对视频监控和自动驾驶十分重要。传统目标检测器通常按逐图像检测设计，忽略了视频帧之间的相关性。如何在计算资源有限的条件下利用时空相关性来改善检测，是一个重要的研究方向 [233, 234]。

跨模态检测：用多种来源或模态的数据进行目标检测，例如 RGB-D 图像、激光雷达、光流、声音、文本和视频等，对于建立更准确、接近人类感知方式的检测系统十分重要。尚待解决的问题包括：如何将训练良好的检测器迁移到不同模态的数据上，以及如何融合信息以提升检测效果等 [235, 236]。

迈向开放世界检测：域外泛化、零样本检测和增量检测是目标检测中的新兴主题。相关工作大多设法减轻灾难性遗忘，或者利用补充信息。人类能够本能地发现环境中类别未知的目标；一旦获得相应的知识（标签），又能从中学习新知识，记住所发现的模式。然而，现有目标检测算法很

难具备检测未知类别目标的能力。开放世界目标检测旨在没有明确监督信号、或只提供部分监督信号时发现未知类别的目标，在机器人和自动驾驶等应用中很有前景 [237, 238]。

沿着技术演进的脉络回顾，我们相信，本文将帮助读者建立完整的目标检测路线图，并找到这个快速发展领域的未来方向。

资助、作者单位与脚注

本研究得到中国国家自然科学基金（项目编号 62125102）、中国国家重点研发计划（项目名称“脑启发的通用视觉模型及应用”），以及中央高校基本科研业务费的资助。通信作者：Zhengxia Zou（zhengxiazou@buaa.edu.cn）和 Jieping Ye（jieping@gmail.com）。

Zhengxia Zou 任职于北京航空航天大学宇航学院制导、导航与控制系（中国北京 100191），同时也任职于上海人工智能实验室（中国上海 200232）。

Keyan Chen 与 Zhenwei Shi 任职于北京航空航天大学宇航学院图像处理中心、数字媒体北京市重点实验室、虚拟现实技术与系统国家重点实验室（中国北京 100191），同时也任职于上海人工智能实验室（中国上海 200232）。

Yuhong Guo 任职于卡尔顿大学计算机科学学院（加拿大安大略省渥太华 K1S 5B6）。

Jieping Ye 任职于阿里巴巴集团（中国杭州 310030）。

[1] PASCAL VOC 网站：http://host.robots.ox.ac.uk/pascal/VOC/

[2] ILSVRC 网站：http://image-net.org/challenges/LSVRC/

[3] MS-COCO 网站：http://cocodataset.org/

[4] Open Images 网站：https://storage.googleapis.com/openimages/web/index.html

[5] Papers with Code 的 COCO 目标检测榜单：https://paperswithcode.com/sota/object-detection-on-coco

参考文献

原文条目与编号 [1]–[238]

[1] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in ECCV. Springer, 2014, pp. 297–312.

[2] ——, "Hypercolumns for object segmentation and fine-grained localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.

[3] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in CVPR, 2016, pp. 3150–3158.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in ICCV. IEEE, 2017, pp. 2980–2988.

[5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in CVPR, 2015, pp. 3128–3137.

[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015, pp. 2048–2057.

[7] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1367–1381, 2018.

[8] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in CVPR, vol. 1. IEEE, 2001, pp. I–I.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137–154, 2004.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in CVPR, vol. 1. IEEE, 2005, pp. 886–893.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in CVPR. IEEE, 2008, pp. 1–8.

[14] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in CVPR. IEEE, 2010, pp. 2241–2248.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014, pp. 580–587.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in ECCV. Springer, 2014, pp. 346–361.

[18] R. Girshick, "Fast r-cnn," in ICCV, 2015, pp. 1440–1448.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in CVPR, 2016, pp. 779–788.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[22] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in ECCV. Springer, 2016, pp. 21–37.

[24] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection." in CVPR, vol. 1, no. 2, 2017, p. 4.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," IEEE transactions on pattern analysis and machine intelligence, 2018.

[26] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 734–750.

[27] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision. Springer, 2020, pp. 213–229.

[29] D. G. Lowe, "Object recognition from local scale-invariant features," in ICCV, vol. 2. Ieee, 1999, pp. 1150–1157.

[30] ——, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[31] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” CALIFORNIA UNIV SAN DIEGO LA JOLLA DEPT OF COMPUTER SCIENCE AND ENGINEERING, Tech. Rep., 2002.

[32] T. Malisiewicz, A. Gupta, and A. A. Efros, “Ensemble of exemplar-svms for object detection and beyond,” in ICCV. IEEE, 2011, pp. 89–96.

[33] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester, “Object detection with grammar models,” in Advances in Neural Information Processing Systems, 2011, pp. 442–450.

[34] R. B. Girshick, From rigid templates to grammars: Object detection with structured models. Citeseer, 2012.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Region-based convolutional networks for accurate object detection and segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, pp. 142–158, 2016.

[37] M. A. Sadeghi and D. Forsyth, “30hz object detection with dpm v5,” in ECCV. Springer, 2014, pp. 65–79.

[38] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, “Single-shot refinement neural network for object detection,” in CVPR, 2018.

[39] Y. Li, Y. Chen, N. Wang, and Z. Zhang, “Scale-aware trident networks for object detection,” arXiv preprint arXiv:1901.01892, 2019.

[40] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” arXiv preprint arXiv:1904.07850, 2019.

[41] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627–9636.

[42] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang et al., “Hybrid task cascade for instance segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4974–4983.

[43] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” arXiv preprint arXiv:2103.14030, 2021.

[45] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” International journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.

[46] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, “Discriminatively trained deformable part models, release 5,” http://people.cs.uchicago.edu/rbg/latentrelease5/.

[47] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 6, pp. 1137–1149, 2017.

[48] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in ECCV. Springer, 2014, pp. 818–833.

[49] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, 2016, pp. 379–387.

[50] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, “Light-head r-cnn: In defense of two-stage object detector,” arXiv preprint arXiv:1711.07264, 2017.

[51] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” arXiv preprint, 2017.

[52] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” arXiv preprint arXiv:2207.02696, 2022.

[53] X. Zhou, J. Zhuo, and P. Krahenbuhl, “Bottom-up object detection by grouping extreme and center points,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 850–859.

[54] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” International journal of computer vision, vol. 88, no. 2, pp. 303–338, 2010.

[55] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” International journal of computer vision, vol. 111, no. 1, pp. 98–136, 2015.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., “Imagenet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in ECCV. Springer, 2014, pp. 740–755.

[58] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,” IJCV, 2020.

[59] R. Benenson, S. Popov, and V. Ferrari, “Large-scale interactive object segmentation with human annotators,” in CVPR, 2019.

[60] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8430–8439.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in CVPR. Ieee, 2009, pp. 248–255.

[62] I. Krasin and T. e. a. Duerig, "Openimages: A public dataset for large-scale multi-label and multi-class image classification." Dataset available from https://storage.googleapis.com/openimages/web/index.html, 2017.

[63] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in CVPR. IEEE, 2009, pp. 304–311.

[64] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 4, pp. 743–761, 2012.

[65] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229, 2013.

[66] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Advances in neural information processing systems, 2013, pp. 2553–2561.

[67] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in ECCV. Springer, 2016, pp. 354–370.

[68] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.

[69] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9657–9666.

[70] T. Malisiewicz, Exemplar-based representations for object detection, association and beyond. Carnegie Mellon University, 2011.

[71] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 4, pp. 814–830, 2016.

[72] J. Hosang, R. Benenson, and B. Schiele, "How good are detection proposals, really?" arXiv preprint arXiv:1406.6962, 2014.

[73] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in CVPR. IEEE, 2010, pp. 73–80.

[74] ——, "Measuring the objectness of image windows," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 11, pp. 2189–2202, 2012.

[75] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in CVPR, 2014, pp. 3286–3293.

[76] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in CVPR, 2014, pp. 2147–2154.

[77] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6569–6578.

[78] A. Torralba and P. Sinha, "Detecting faces in impoverished images," MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, Tech. Rep., 2001.

[79] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár, "A multipath network for object detection," arXiv preprint arXiv:1604.02135, 2016.

[80] X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang, "Gated bi-directional cnn for object detection," in ECCV. Springer, 2016, pp. 354–369.

[81] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang et al., "Crafting gbdnet for object detection," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 9, pp. 2109–2123, 2018.

[82] W. Ouyang, K. Wang, X. Zhu, and X. Wang, "Learning chained deep features and classifiers for cascade in object detection," arXiv preprint arXiv:1702.07054, 2017.

[83] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in ICCV, 2015, pp. 1134–1142.

[84] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu et al., "Couplenet: Coupling global structure with local parts for object detection," in ICCV, vol. 2, 2017.

[85] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," International journal of computer vision, vol. 95, no. 1, pp. 1–12, 2011.

[86] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in CVPR, 2016, pp. 2874–2883.

[87] Z. Li, Y. Chen, G. Yu, and Y. Deng, "R-fcn++: Towards accurate region-based fully convolutional networks for object detection." in AAAI, 2018.

[88] S. Liu, D. Huang et al., "Receptive field block net for accurate and fast object detection," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 385–400.

[89] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[90] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 1, pp. 13–27, 2015.

[91] S. Gupta, B. Hariharan, and J. Malik, "Exploring person context and local scene context for object detection," arXiv preprint arXiv:1511.08177, 2015.

[92] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4086–4096.

[93] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3588–3597.

[94] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in CVPR, 2018, pp. 6985–6994.

[95] L. V. Pato, R. Negrinho, and P. M. Q. Aguiar, "Seeing without looking: Contextual rescoring of object detections for ap maximization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[96] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in CVPR. IEEE, 2009, pp. 1271–1278.

[97] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-cnn for small object detection," in Asian conference on computer vision. Springer, 2016, pp. 214–230.

[98] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 944–954, 2017.

[99] H. A. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," in Advances in Neural Information Processing Systems, 1996, pp. 875–881.

[100] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in ICCV. IEEE, 1998, pp. 555–562.

[101] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in ECCV. Springer, 2016, pp. 443–457.

[102] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in CVPR, 2016, pp. 761–769.

[103] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[104] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" Advances in neural information processing systems, vol. 32, 2019.

[105] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016, pp. 516–520.

[106] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 658–666.

[107] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[108] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," IEE Proceedings-Vision, Image and Signal Processing, vol. 141, no. 4, pp. 245–250, 1994.

[109] P. Henderson and V. Ferrari, "End-to-end training of object class detectors for mean average precision," in Asian Conference on Computer Vision. Springer, 2016, pp. 198–213.

[110] J. H. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression." in CVPR, 2017, pp. 6469–6477.

[111] Z. Tan, X. Nie, Q. Qian, N. Li, and H. Li, "Learning to rank proposals for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8273–8281.

[112] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms—improving object detection with one line of code," in ICCV. IEEE, 2017, pp. 5562–5570.

[113] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," arXiv preprint arXiv:1711.00164, 2017.

[114] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2888–2897.

[115] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6459–6468.

[116] R. Rothe, M. Guillaumin, and L. Van Gool, "Non-maximum suppression for object detection by passing messages between windows," in Asian Conference on Computer Vision. Springer, 2014, pp. 290–306.

[117] D. Mrowca, M. Rohrbach, J. Hoffman, R. Hu, K. Saenko, and T. Darrell, "Spatial semantic regularisation for large scale object detection," in ICCV, 2015, pp. 2003–2011.

[118] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," Image and Vision Computing, vol. 107, p. 104117, 2021.

[119] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," IEEE Transactions on Cybernetics, 2021.

[120] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng, "End-to-end object detection with fully convolutional network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 849–15 858.

[121] C. Papageorgiou and T. Poggio, "A trainable system for object detection," International journal of computer vision, vol. 38, no. 1, pp. 15–33, 2000.

[122] L. Wan, D. Eigen, and R. Fergus, "End-to-end integration of a convolution network, deformable parts model and non-maximum suppression," in CVPR, 2015, pp. 851–859.

[123] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," arXiv preprint arXiv:1905.05055, 2019.

[124] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in CVPR, vol. 2. IEEE, 2006, pp. 1491–1498.

[125] F. Fleuret and D. Geman, "Coarse-to-fine face detection," International Journal of computer vision, vol. 41, no. 1-2, pp. 85–107, 2001.

[126] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in CVPR, 2015, pp. 5325–5334.

[127] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[128] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in ICCV, 2015, pp. 3361–3369.

[129] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in Advances in neural information processing systems, 1990, pp. 598–605.

[130] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," arXiv preprint arXiv:1510.00149, 2015.

[131] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in CVPR, 2015, pp. 5353–5360.

[132] Z. Qin, Z. Li, Z. Zhang, Y. Bao, G. Yu, Y. Peng, and J. Sun, "Thundernet: Towards real-time generic object detection on mobile devices," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6718–6727.

[133] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 1967–1976.

[134] R. Huang, J. Pedoeem, and C. Chen, "Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers," in 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 2503–2510.

[135] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint based object detection," arXiv preprint arXiv:1904.08900, 2019.

[136] G. Yu, Q. Chang, W. Lv, C. Xu, C. Cui, W. Ji, Q. Dang, K. Deng, G. Wang, Y. Du et al., "Pp-picodet: A better real-time object detector on mobile devices," arXiv preprint arXiv:2111.00902, 2021.

[137] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in CVPR, 2016, pp. 2818–2826.

[138] X. Zhang, J. Zou, X. Ming, K. He, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in CVPR, 2015, pp. 1984–1992.

[139] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 1943–1955, 2016.

[140] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017.

[141] G. Huang, S. Liu, L. van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," group, vol. 3, no. 12, p. 11, 2017.

[142] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," arXiv preprint, pp. 1610–02 357, 2017.

[143] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[144] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in CVPR. IEEE, 2018, pp. 4510–4520.

[145] Y. Li, J. Li, W. Lin, and J. Li, "Tiny-dsod: Lightweight object detection for resource-restricted usages," arXiv preprint arXiv:1807.11013, 2018.

[146] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[147] B. Wu, F. N. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving." in CVPR Workshops, 2017, pp. 446–454.

[148] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in CVPR, 2016, pp. 845–853.

[149] Y. Chen, T. Yang, X. Zhang, G. Meng, C. Pan, and J. Sun, "Detnas: Neural architecture search on object detection," arXiv preprint arXiv:1903.10979, 2019.

[150] H. Xu, L. Yao, W. Zhang, X. Liang, and Z. Li, "Autofpn: Automatic network architecture adaptation for object detection beyond classification," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6649–6658.

[151] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7036–7045.

[152] J. Guo, K. Han, Y. Wang, C. Zhang, Z. Yang, H. Wu, X. Chen, and C. Xu, "Hit-detector: Hierarchical trinity architecture search for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 405–11 414.

[153] N. Wang, Y. Gao, H. Chen, P. Wang, Z. Tian, C. Shen, and Y. Zhang, "Nas-fcos: Fast neural architecture search for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 943–11 951.

[154] L. Yao, H. Xu, W. Zhang, X. Liang, and Z. Li, "Smnas: structural-to-modular neural architecture search for object detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12 661–12 668.

[155] C. Jiang, H. Xu, W. Zhang, X. Liang, and Z. Li, "Sp-nas: Serial-to-parallel backbone search for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 863–11 872.

[156] P. Simard, L. Bottou, P. Haffner, and Y. LeCun, "Boxlets: a fast convolution algorithm for signal processing and neural networks," in Advances in Neural Information Processing Systems, 1999, pp. 571–577.

[157] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in ICCV. IEEE, 2009, pp. 32–39.

[158] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in CVPR, vol. 1. IEEE, 2005, pp. 829–836.

[159] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

[160] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," arXiv preprint arXiv:1312.5851, 2013.

[161] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "Fcnn: Fourier convolutional neural networks," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017, pp. 786–798.

[162] N. Vasilache, J. Johnson, M. Mathieu, S. Chintala, S. Piantino, and Y. LeCun, "Fast convolutional nets with fbfft: A gpu performance evaluation," arXiv preprint arXiv:1412.7580, 2014.

[163] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in Advances in neural information processing systems, 2015, pp. 2449–2457.

[164] M. A. Sadeghi and D. Forsyth, "Fast template evaluation with vector quantization," in Advances in neural information processing systems, 2013, pp. 2949–2957.

[165] I. Kokkinos, "Bounding part scores for rapid detection with deformable part models," in ECCV. Springer, 2012, pp. 41–50.

[166] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in ICIP. IEEE, 2015, pp. 3735–3739.

[167] B. Cai, Z. Jiang, H. Zhang, Y. Yao, and S. Nie, "Online exemplar-based fully convolutional network for aircraft detection in remote sensing images," IEEE Geoscience and Remote Sensing Letters, no. 99, pp. 1–5, 2018.

[168] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pp. 119–132, 2014.

[169] G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in CVPR, 2016, pp. 2884–2893.

[170] ——, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, pp. 7405–7415, 2016.

[171] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 265–278, 2018.

[172] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in CVPR, 2018, pp. 2295–2303.

[173] M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, pp. 2017–2025.

[174] D. Chen, G. Hua, F. Wen, and J. Sun, "Supervised transformer network for efficient face detection," in ECCV. Springer, 2016, pp. 122–138.

[175] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849–2858.

[176] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–snip," in CVPR, 2018, pp. 3578–3587.

[177] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," arXiv preprint arXiv:1805.09300, 2018.

[178] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[179] M. Gao, R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Dynamic zoom-in network for fast object detection in large images," in CVPR, 2018.

[180] Y. Lu, T. Javidi, and S. Lazebnik, "Adaptive object detection using adjacency and zoom prediction," in CVPR, 2016, pp. 2351–2359.

[181] S. Qiao, W. Shen, W. Qiu, C. Liu, and A. L. Yuille, "Scalenet: Guiding object proposal generation in supermarkets and beyond." in ICCV, 2017, pp. 1809–1818.

[182] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in CVPR, vol. 3, 2017.

[183] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390–391.

[184] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in European conference on computer vision. Springer, 2016, pp. 483–499.

[185] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang et al., "Recent advances in convolutional neural networks," arXiv preprint arXiv:1512.07108, 2015.

[186] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in CVPR, vol. 4, 2017.

[187] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in CVPR, vol. 1, no. 2, 2018, p. 10.

[188] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "Refinenet: Iterative refinement for accurate object localization," in ITSC. IEEE, 2016, pp. 1528–1533.

[189] M.-C. Roh and J.-y. Lee, "Refining faster-rcnn for accurate object detection," in Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on. IEEE, 2017, pp. 514–517.

[190] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in Proceedings of the ECCV, Munich, Germany, 2018, pp. 8–14.

[191] S. Gidaris and N. Komodakis, "Locnet: Improving localization accuracy for object detection," in CVPR, 2016, pp. 789–798.

[192] S. Brahmbhatt, H. I. Christensen, and J. Hays, "Stuffnet: Using 'stuff'to improve object detection," in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE, 2017, pp. 934–943.

[193] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster r-cnn," in ECCV. Springer, 2016, pp. 330–348.

[194] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680.

[195] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[196] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," arXiv preprint, 2017.

[197] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network." in CVPR, vol. 2, no. 3, 2017, p. 4.

[198] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in CVPR, 2017.

[199] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sodmtgan: Small object detection via multi-task generative adversarial network," Computer Vision-ECCV, pp. 8–14, 2018.

[200] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in CVPR, 2017.

[201] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 9, pp. 5866–5885, 2021.

[202] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axisparallel rectangles," Artificial intelligence, vol. 89, no. 1-2, pp. 31–71, 1997.

[203] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in Advances in neural information processing systems, 2003, pp. 577–584.

[204] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pp. 189–203, 2017.

[205] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "We don't need no bounding-boxes: Training object class detectors using only human verification," in CVPR, 2016, pp. 854–863.

[206] D. Zhang, W. Zeng, J. Yao, and J. Han, "Weakly supervised object detection using proposal-and semanticlevel relationships," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[207] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "Pcl: Proposal cluster learning for weakly supervised object detection," IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 1, pp. 176–191, 2018.

[208] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 3, pp. 712–725, 2018.

[209] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," International Journal of Computer Vision, vol. 127, no. 4, pp. 363–380, 2019.

[210] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in ICCV, 2017, pp. 1841–1850.

[211] A. Diba, V. Sharma, A. M. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks." in CVPR, vol. 1, no. 2, 2017, p. 8.

[212] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in CVPR, 2016, pp. 2921–2929.

[213] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in CVPR, 2016, pp. 2846–2854.

[214] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," in Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016, pp. 1–9.

[215] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, "Generative adversarial learning towards fast weakly supervised detection," in CVPR, 2018, pp. 5764–5773.

[216] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3339–3348.

[217] Y. Wang, R. Zhang, S. Zhang, M. Li, Y. Xia, X. Zhang, and S. Liu, "Domain-specific suppression for adaptive object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9603–9612.

[218] L. Hou, Y. Zhang, K. Fu, and J. Li, "Informative and consistent correspondence mining for cross-domain weakly supervised object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9929–9938.

[219] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 687–696.

[220] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strongweak distribution alignment for adaptive object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6956–6965.

[221] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, "Exploring categorical regularization for domain adaptive object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 724–11 733.

[222] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[223] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12 456–12 465.

[224] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5001–5009.

[225] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, and M.-H. Yang, "Progressive domain adaptation for object detection," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 749–757.

[226] B. Bosquet, M. Mucientes, and V. M. Brea, “Stdnetst: Spatio-temporal convnet for small object detection,” Pattern Recognition, vol. 116, p. 107929, 2021.

[227] C. Yang, Z. Huang, and N. Wang, “Querydet: Cascaded sparse query for accelerating high-resolution small object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 668–13 677.

[228] P. Sun, Y. Jiang, E. Xie, W. Shao, Z. Yuan, C. Wang, and P. Luo, “What makes for end-to-end object detection?” in International Conference on Machine Learning. PMLR, 2021, pp. 9934–9944.

[229] X. Zhou, X. Xu, W. Liang, Z. Zeng, S. Shimizu, L. T. Yang, and Q. Jin, “Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems,” IEEE Transactions on Industrial Informatics, vol. 18, no. 2, pp. 1377–1386, 2021.

[230] G. Cheng, X. Yuan, X. Yao, K. Yan, Q. Zeng, and J. Han, “Towards large-scale small object detection: Survey and benchmarks,” arXiv preprint arXiv:2207.14096, 2022.

[231] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, “Detr3d: 3d object detection from multi-view images via 3d-to-2d queries,” in Conference on Robot Learning. PMLR, 2022, pp. 180–191.

[232] Y. Wang, T. Ye, L. Cao, W. Huang, F. Sun, F. He, and D. Tao, “Bridged transformer for vision and point cloud 3d object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 114–12 123.

[233] X. Cheng, H. Xiong, D.-P. Fan, Y. Zhong, M. Harandi, T. Drummond, and Z. Ge, “Implicit motion handling for video camouflaged object detection,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 864–13 873.

[234] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, “Transvod: End-to-end video object detection with spatial-temporal transformers,” arXiv preprint arXiv:2201.05047, 2022.

[235] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, “Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection,” IEEE Transactions on Image Processing, 2022.

[236] Y. Wang, L. Zhu, S. Huang, T. Hui, X. Li, F. Wang, and S. Liu, “Cross-modality domain adaptation for freespace detection: A simple yet effective baseline,” in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4031–4042.

[237] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, “Promptdet: Expand your detector vocabulary with uncurated images,” arXiv preprint arXiv:2203.16513, 2022.

[238] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li et al., “Regionclip: Region-based language-image pretraining,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16 793–16 803.