https://crad.ict.ac.cn
62(12): 2895−2928, 2025

目标检测模型综述

李承烨[1,2,3,4] 张 震[1,2,3,4] 梁哲恒[5,6] 姚潮生[5,6] 张金波[5,6] 晏荣杰[2,3,4] 吴 鹏[2,3,4]

[1](中国科学院大学杭州高等研究院 杭州 310024)

[2](中国科学院基础软件与系统重点实验室 北京 100190)

[3](计算机科学国家重点实验室(中国科学院软件研究所) 北京 100190)

[4](中国科学院大学 北京 100049)

[5](网络空间安全联合实验室(中国南方电网有限公司)广州 510620)

[6](广东电网有限责任公司 广州 510620)

(licy@ios.ac.cn)

Survey on Object Detection Models

Li Chengye[1,2,3,4], Zhang Zhen[1,2,3,4], Liang Zheheng[5,6], Yao Chaosheng[5,6], Zhang Jinbo[5,6], Yan Rongjie[2,3,4], and Wu Peng[2,3,4]

[1](*Hangzhou Institute for Advanced Study*, *University of Chinese Academy of Sciences*, *Hangzhou* 310024)

[2](*CAS Key Laboratory of System Software*, *Beijing* 100190)

[3](*State Key Laboratory of Computer Science* (*Institute of Software*, *Chinese Academy of Sciences*), *Beijing* 100190)

[4](*University of Chinese Academy of Sciences*, *Beijing* 100049)

[5](*Joint Laboratory on Cyberspace Security* (*China Southern Power Grid Co., Ltd.*), *Guangzhou* 510620)

[6](*Guangdong Power Grid Co., Ltd.*, *Guangzhou* 510620)

Abstract Object detection technology, as a pivotal component in computer vision, plays a vital role in diverse practical applications. Over decades of evolution, the field has progressed from early methods relying on handcrafted feature extraction to the widespread adoption of deep learning models. Currently, there remains a lack of systematic reviews tracing the developmental trajectory of object detection through improvements in deep learning foundation models. Addressing this gap, this paper organizes the technological evolution around the progression of foundation models in artificial intelligence. We systematically survey detection models built upon various foundation models, compare their strengths and weaknesses, and analyze improvement strategies. The paper also surveys evaluation metrics and technological advancements across different eras, with particular emphasis on how deep learning has driven remarkable performance gains. We discuss persistent challenges in handling diverse scenarios, improving real-time efficiency, and enhancing accuracy. Furthermore, we explore prospective research directions, including model generalization capabilities, computational efficiency, and integration with complex tasks, proposing potential enhancement strategies. This work aims to provide a clear perspective on technological evolution to facilitate further research and applications in object detection.

Key words object detection; deep learning; model architecture; artificial intelligence; computer vision

收稿日期: 2024-05-06; **修回日期**: 2025-04-09

基金项目: 国家重点研发计划项目(2022YFA1005100, 2022YFA1005101, 2022YFA1005104); 中国科学院软件研究所创新基金重大重点项目(ISCAS-ZD-202302); 国家自然科学基金项目(62132020); 中国科学院稳定支持基础研究领域青年团队计划项目(YSBR-040); 中国南方电网有限公司项目(037800KK52220005)

This work was supported by the National Key Research and Development Program of China (2022YFA1005100, 2022YFA1005101, 2022YFA1005104), the Major Project of ISCAS (ISCAS-ZD-202302), the National Natural Science Foundation of China (62132020), the CAS Project for Young Scientists in Basic Research (YSBR-040), and the Project of China Southern Power Grid Company Limited (037800KK52220005).

通信作者: 晏荣杰(yrj@ios.ac.cn)

摘 要 目标检测技术是计算机视觉领域的关键组成部分,它在各种实际应用中扮演着至关重要的角色.目标检测技术经历了几十年的发展,从早期依赖于手工特征提取的方法,到当前深度学习模型的广泛应用.目前在目标检测领域缺少以深度学习基础模型技术的改进为发展脉络的总结研究下,以人工智能领域基础模型的发展过程为线索,围绕不同种类基础模型概述了基于这些模型的不同目标检测模型的发展,同时对这些基于不同模型的目标检测算法进行了比较,并分析不同模型的优缺点以及制定不同模型的改进策略.概述了目标检测技术的评估指标以及不同阶段的技术,特别强调了深度学习如何推动目标检测性能的显著提升,讨论了目标检测在处理多样化场景以及提高实时性和准确性方面的挑战,并对未来可能的研究方向进行了深度探讨,包括但不限于模型的泛化能力、计算效率以及与更复杂任务的结合,为多个未来研究方向提出了可能的提高策略.旨在提供一个清晰的技术演进视角,以促进目标检测领域的进一步研究和应用.

关键词 目标检测;深度学习;模型架构;人工智能;计算机视觉

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440315 **CSTR:** 32373.14.issn1000-1239.202440315

计算机视觉是人工智能的一大基础任务,其目标是要求人工智能算法或系统可以接受各种各样的图像进行处理然后获得反馈.如今,计算机视觉技术已经应用到了各种环境之中,如在工业中辅助人类对零部件缺陷进行检测、在医疗领域为医生提供快速的肿瘤分割参考、在道路环境中赋予车辆更全面且精细的环境感知能力等.在计算机视觉这一大的框架下,又可以按照侧重点的不同将其分为一些子任务,如图像分类、目标检测、实例分割、计算摄影以及3维视觉等研究方向.图像识别与分类是所有计算机视觉任务的基础,图像分类任务要求人工智能系统对输入的图像进行理解,并且对图像的类别进行判断.伴随着深度学习在计算机视觉上的崛起,人们逐渐不再满足于简单的分类任务,开始研究以前受到技术限制的其他视觉任务,如目标检测.

目标检测是继图像分类后的另一个计算机视觉基石任务.目标检测的核心在于“是什么”和“在哪里”.分类问题只需要评估图片中的物体是什么,且图像中仅包含单一的待识别物体,而目标检测的输入图像往往不止包含1个目标,同时不仅需要知道输入图像中包含了哪些种类的物体,而且需要对它们进行准确的定位.所以,对于目标检测模型性能的评估不仅需要传统的类别准确度判断,还要对定位的偏差进行评估.同时,目标检测也是很多其他重要的视觉任务的基础,如实例分割、工业缺陷检测、目标跟踪乃至视频理解等.随着计算机视觉的发展,目标检测算法也应用到了很多实际的问题上,如自动驾驶、机器人、工业保障等.大量的现实使用场景要求目标检测不仅定位要准确,同时也具有较快的识别速度.因为目标检测任务具有广泛的使用场景,目标检测算法已经得到了学术界和工业界的广泛研究.如图1所示,学术界对目标检测任务的研究呈明显的蓬勃发展趋势.

如今,目标检测算法已经被大量使用在了安全攸关的场景之中,如自动驾驶场景中的目标识别和环境感知,安防领域内的车牌识别、人脸检测、X-ray识别,工业场景下的部件缺陷检测以及医疗领域的细胞病理诊断检测和肿瘤分割与检测等.在当前快速发展的目标检测领域,虽然已有诸多综述涉及该领域的各个方面,但鉴于技术的不断进步和应用的日益扩大,对这一领域进行全面而深入的梳理仍显得尤为重要.本文旨在填补现有综述文献的空白,提供一种全新的视角来回顾目标检测领域的发展,并且对目标检测领域当前所面临的挑战以及可能的未来发展方向进行了总结和展望,提出了可能的解决办法.我们不仅详尽地介绍了评估目标检测性能的关键指标和多样化的重要数据集,还全面覆盖了从传统手工算法到2024年的最新网络结构.与现有的综述相比,我们的工作具有独特优势:它依托于人工智能(尤其是深度学习)的发展历程,按照不同的基础模型,包括卷积神经网络(convolutional neural network, CNN)、Transformer、多层感知机(multi-layer perceptron, MLP)以及最新的扩散模型和大规模模型,进行分类介绍,这不仅为读者提供了一个清晰的历史和技术脉络,如图2所示,也有助于更好地理解各种方法之间的关联和区别,同时我们还对基于不同基础模型的目标检测模型的性能进行了横向对比,并且分析不同基础模型之间的差距以及导致不同目标检测性

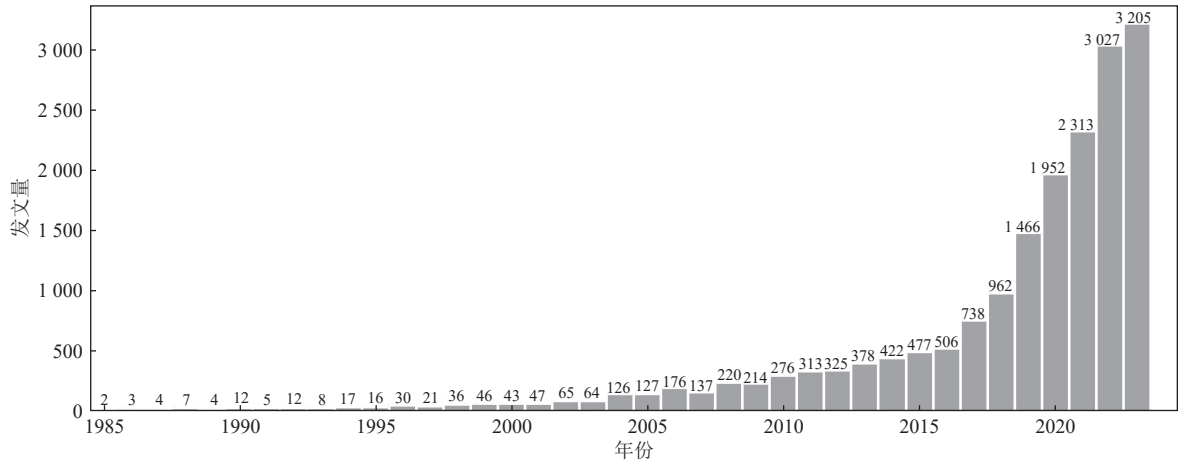


Fig. 1 The number of publications on object detection

图 1 关于目标检测文献的发表数量

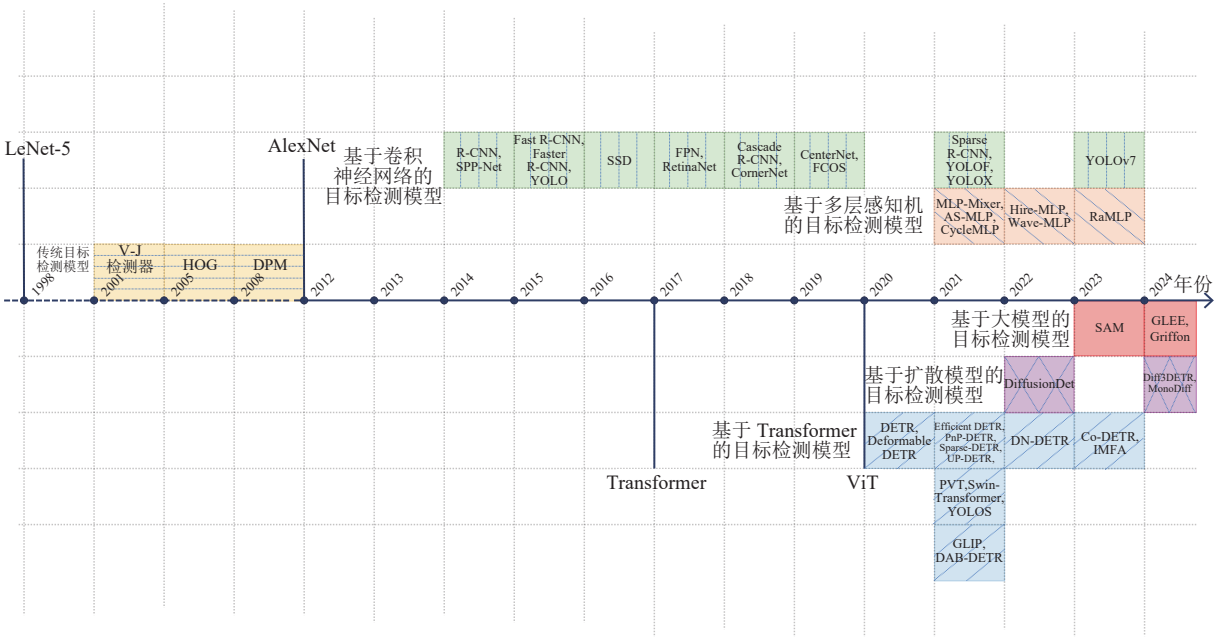


Fig. 2 Development context diagram of the target detection model introduced in this paper

图 2 本文介绍的目标检测模型的发展脉络图

能的原因. 最后, 我们展望了目标检测领域的未来发展趋势, 并且为不同的未来发展方向提供了见解. 通过这种细致的分析和对比, 我们期望本综述能够为目标检测领域的研究人员提供有价值的指导和启发, 推动该领域的进一步发展.

1 目标检测技术的评估指标和常用数据集

1.1 精度指标

混淆矩阵(confusion matrix), 通常用于分类任务, 包含以下 4 个指标:

1)真阳性(true positive, *TP*), 实际为正例且被预测为正例的样本数.

2)假阳性(false positive, *FP*), 实际为负例但被预测为正例的样本数.

3)假阴性(false negative, *FN*), 实际为正例但被预测为负例的样本数.

4)真阴性(true negative, *TN*), 实际为负例且被预测为负例的样本数.

以上 4 个指标可以用来评估目标检测技术中分类任务的精确度. 此外, 还有以下指标:

1)准确率(*Accuracy*). 准确率是评估模型分类性能的指标之一, 它衡量了模型将每个样本正确分类的能力. 准确率越高, 说明模型越能准确地识别目标物体.

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}. \tag{1}$$

2)精确率(*Precision*). 精确率是衡量模型在预测为正例的样本中真正为正例的样本的比例. 精确率越高, 说明模型预测为正例的样本中真正的正例样本越多.

$$Precision=\frac{TP}{TP+FP}. \tag{2}$$

3)召回率(*Recall*). 召回率是衡量模型在所有实际为正例的样本中被正确预测为正例的样本的比例. 召回率越高, 说明模型能够将更多的实际正例样本检测出来.

$$Recall=\frac{TP}{TP+FN}. \tag{3}$$

4)平均精确率(average precision, *AP*). *AP* 是指在一个特定的类别中, 随着预测的置信度增加, 模型正确预测的样本数占总样本数的比例. 它衡量了模型在该类别中的预测精度. 对于每个类别, 按照预测的置信度从低到高排序所有样本, 计算不同阈值下的 *Precision-Recall* 曲线, 并根据曲线下的面积计算 *AP*.

$$AP=\int_0^1 P(r)\mathrm{d}r, \tag{4}$$

其中$P(r)$是 *Precision-Recall* 曲线.

5)平均准确度均值(mean average precision, *mAP*). *mAP* 是指多个类别 *AP* 的平均值, 它用于衡量模型在所有类别中的总体性能. 将所有类别的 *AP* 进行平均, 得到 *mAP*.

$$mAP=\frac{\sum_{i=1}^{n}AP_i}{n}, \tag{5}$$

其中n是类别数量.

AP 和 *mAP* 指标均为位于 0~1 的百分数, 但是在不同的文献中可能会有省略百分号的记法. 并且在一些文献中会省略 *mAP* 的记法, 以 *AP* 来指代在指定数据集上的 *mAP*. 此外, 还有描述骨干网络性能的指标, 如下所示.

1)Top-1 准确率(Top-1 accuracy). 指对于多分类任务, 选取模型对每个测试样本生成的概率最高的类别作为模型的预测类别, 如果模型的预测类别与真实类别相同, 则该预测被视为正确. 对整个测试集中所有正确的预测进行计数, 将正确预测的总数除以测试集中样本的总数即为 Top-1 准确率.

2)交并比(intersection over union, *IoU*). 用于衡量 2 个检测框(bounding box)的相似度或重叠程度. 它是 2 个检测框的交集区域与并集区域之比. 如果 2 个检测框完全重叠, 则 *IoU*=1; 如果 2 个检测框没有重叠, 则 *IoU*=0.

1.2 计算量和速度指标

浮点操作数(floating-point operations, FLOP)是衡量深度学习模型计算复杂度的核心指标, 定义为模型在推理过程中执行的所有浮点运算操作总数, 包括乘法、加法、除法和减法等基本运算. FLOP 直接表征模型完成单次前向传播所需的计算量. 在目标检测领域, 典型模型的 FLOP 量级通常在 10^9~10^{11}(即 1~100 GFLOP), 而大规模模型可达 10^{12}(1 TFLOP)以上. FLOP 的核心用途在于量化模型对计算资源的需求: 较高的 FLOP 值意味着模型需要更多计算硬件(如 GPU/TPU)的支持, 且推理延迟可能增加. FLOP 与 FLOPS(floating-point operations per second)存在的本质区别是前者是算法复杂度的静态指标, 后者是硬件性能的动态指标.

每秒传输帧数 fps(frame per second)在目标检测算法中通常被定义为每秒内可以处理的图片数量. fps 反映了算法在指定硬件平台上的运行速度, 对于需要实时性的应用场景, 例如自动驾驶非常关键. 当精度相同时, fps 越大说明算法的运行速度越快, 能够在更短的时间完成目标检测任务, 以及可以以更细的时间粒度来对目标环境进行感知.

1.3 常见数据集

ImageNet[1] 是一个用于计算机视觉识别、检测等研究的超大型图像数据库, 其中的类别标签和图片涵盖了非常多在现实生活中出现的对象, 目前该数据集一共包含了属于 21 841 个类别的 14 197 122 张图像 URL 和它们的根据词汇网络层次结构来人工注释的图像级二进制标签, 同时, 在这 1 400 多万张图片中, 还有 1 034 908 张图片具有其中对象实例的边界框和类别标签注释. 至今绝大多数针对视觉任务的模型都使用 ImageNet 进行训练和评估.

PASCAL VOC(pattern analysis, statistical modelling and computational learning visual object classes)是另一个世界级的计算机视觉挑战赛, 从 2005 年到 2012 年每年举办. PASCAL VOC 数据集为很多著名的目标检测算法提供了数据基础. VOC 每年都会推出一个修改后的数据集, 目前最完整的是 PASCAL VOC 2012, 而大多数研究者普遍使用的是 PASCAL VOC 2007 和 PASCAL VOC 2012 这 2 个数据集, 因为它们两者并不互相包含. PASCAL VOC 2007 包含了 20 个类别的标签, 有人、动物、车辆和室内对象(如椅子、沙发和电视等), 训练集、验证集和测试集一共包含 9 963

张图像和 24 640 个带有标签的目标注释. PASCAL VOC 2012 数据集包括相同的 20 个类别标签，训练集和验证集一共包含 13 841 张图片和 27 450 个带有标签的目标注释. PASCAL VOC 作为早年间的目标检测数据集，为很多著名的目标检测模型提供了数据支持.

MS COCO(microsoft common objects in context)，数据集[2]是一个大规模的对象检测、分割、关键点检测的图-文数据集. 该数据集由 328 000 张图像组成. MS COCO 数据集包含了 80 个不同的物体类别，涵盖了大量现实世界中的各种场景和目标，如人、动物、交通工具、食物等. 因此，MS COCO 更有利于算法对现实运行环境的泛化. MS COCO 的标签中，有针对目标检测和实例、全景分割任务的边界框和分割掩码，还有包含超过 200 000 张图像和 250 000 个标有关键点的人实例(17 个可能的关键点，如左眼、鼻子、右臀部、右脚踝)，以及针对图-文多模态任务的图像和其对应的自然语言描述的图像文本对等. 与 PASCAL VOC 等相对较早的数据集相比，MS COCO 具有更丰富更全面的场景环境和标注信息，这些更复杂多样的信息为计算机视觉算法的发展提供了重要的帮助.

常见数据集的总结与对比如表 1 所示.

Table 1 Overview of the Common Datasets for Object Detection

表 1 目标检测常用数据集概览

数据集	类别	训练集	验证集	训练-验证集	测试集
PASCAL VOC 2007	20 类：人、鸟、猫、牛、狗、马、羊、飞机、自行车、船、巴士、汽车、摩托车、火车、瓶子、椅子、餐桌、盆栽、沙发和电视/显示器	2 501(6 301)	2 510(6 307)	5 011(12 608)	4 952(12 032)
PASCAL VOC 2012		5 717(13 609)	5 823(13 841)	11 540(27 450)	
MS COCO 2014	80 类：人、自行车、汽车、摩托车、飞机、巴士、火车、卡车、船、交通灯、消防栓、停车标志、停车表、长凳、鸟、猫、狗、马、羊、牛、大象、熊、斑马、长颈鹿、背包、雨伞、手提包、领带、行李箱、飞盘、滑雪板、滑雪板、运动球、风筝、棒球棍、棒球手套、滑板、冲浪板、网球拍、瓶子、红酒杯、杯子、叉子、刀、勺、碗、香蕉、苹果、三明治、橙子、西兰花、胡萝卜、热狗、披萨、甜甜圈、蛋糕、椅子、沙发、盆栽、床、餐桌、厕所、电视、笔记本电脑、鼠标、遥控器、键盘、手机、微波炉、烤箱、烤面包机、水槽、冰箱、书、时钟、花瓶、剪刀、泰迪熊、吹风机和牙刷	82 783	40 504	123 287	40 775
MS COCO 2015		82 783	40 504	123 287	81 434
MS COCO 2017		118 287	5 000	123 287	40 670

注：括号外数值表示数据集包含的图像数量，括号内数值表示对应数据集包含的目标实例数量.

2 目标检测技术发展概述

在深度学习广泛应用于人工智能之前，研究者们开发了许多基于手工特征的目标检测算法. 尽管这些早期算法性能不及现代技术，它们的设计理念对现代目标检测算法产生了重要影响. 随着深度学习在计算机视觉中的成功应用，诞生了众多基于神经网络的检测方法. 本文不按传统的算法分类(如 1 阶段和 2 阶段检测器)介绍，而是依据深度学习的发展及其在特征提取骨干网络模型中的应用来梳理关键的目标检测算法. 我们分析了这些模型的结构及其优劣，并概述了它们在各数据集上的速度与精度表现.

2.1 传统算法

2.1.1 模 型

1)VJ(Viola-Jones)检测器. Viola 等人[3-4]为人脸检测开发了 VJ 检测器，这是实现实时性能和有效检测率的目标检测框架. 在 700 MHz Pentium III CPU 上，VJ 检测器每秒可处理约 15 帧调整后的图像，速度是其他同等精度算法的数十倍. 它通过滑动窗口产生候选框，使用 Haar-like 特征进行特征提取，判断窗口是否含人脸. 由于软硬件限制，对不同大小滑动窗口尺寸的判断以及遍历每个候选窗口的计算量仍然是不可接受的，因此 VJ 检测器引入了积分图来加速 Haar-like 特征计算，并用 Ada Boost 算法[5]筛选特征，提高了特征选择能力. 此外，通过检测级联加速，提高准确率和减少计算. 然而，VJ 检测器也有改进空间：其简单的 Haar-like 特征对复杂人脸情况(如遮挡、姿态变化)处理效果欠佳；并且检测级联方法的特征传递都会被每一层检测器视为全新的特征，缺乏鲁棒性.

2)方向梯度直方图(histogram of oriented gradients, HOG). 2005 年，Dalal 等人[6]引入了 HOG 特征描述符用于行人检测. 不同于 VJ 检测器，HOG 采用固定窗口大小，并通过图像金字塔对图像进行缩放以处理不同尺寸目标. 具体而言，HOG 将图像分区，计算每个区域的梯度大小和方向，按梯度方向构建直方图来形成特征向量. 将候选框内所有特征连接以形成完整特征向量，利用线性支持向量机(support vector

machine, SVM)[7]分类, 判断是否为行人. 为降低光照影响, 还引入了图像归一化. HOG特征的优势在于对几何和光学变化的不变性, 能较好捕捉局部形状特征. 然而, HOG对人体姿势变化的鲁棒性较弱, 在姿势变化大时检测效果下降, 且HOG对噪声敏感, 易受图像噪声干扰. 整个HOG的特征维度相对较大, 而且因为是固定窗口大小, 需要多次调整输入图像大小, 所以计算更复杂.

3)可变形部件模型(deformable part-based model, DPM). DPM是一种基于部件(part)的检测模型, 在2008年作为HOG特征的扩展由Felzenszwalb等人[8]提出. 相较于HOG检测, DPM为了进一步提高精度, 在提取HOG特征的时候选择无符号梯度和有符号梯度2种特征共同构建特征向量. DPM仍然沿用部分HOG检测的步骤, 但是为了提高对姿态变化的鲁棒性, DPM模型提出了对目标的部件进行检测. 例如, 将目标"人"分解为头、身体、手、脚等部件的组合, 使用根滤波器来捕获人的整体特征响应图, 使用组件滤波器获取部分特征的响应图, 将2个响应图采样到同样分辨率后加权平均得到最终相应图, 然后使用滑动窗口和隐式SVM分类器进行最终的目标检测. DPM模型通过引入部件概念, 较之原始的HOG检测方法(通常只能处理直立姿态的人体目标), 可以更好地检测执行不同动作的人体. 但是由于组件滤波器仍然是靠手工设计, 导致DPM的可迁移性非常差, 当待检测目标为其他种类的对象时就要重新人为设计组件特征, 工作量较大.

2.1.2 小 结

早期的传统目标检测算法在目标检测领域发挥了重要作用. 然而, 这些算法的设计主要依赖于研究人员精心构造手工特征, 通常只在特定任务中有效, 且一般只能提取较为浅层的特征, 从而限制了它们对目标的表征能力. 随着算法的发展, 手工设计特征的性能逐渐进入瓶颈期. 随着检测任务复杂性的增加, 手工设计特征的性能逐渐遇到瓶颈. 这些特征对于复杂场景和多样化目标的适应性相对有限, 对传统目标检测算法的鲁棒性和迁移性提出了挑战, 光照、姿态、尺度等方面的变化会显著影响它们的检测性能, 限制了它们在真实世界中的应用. 尽管如此, 这些早期算法也为后续目标检测的发展奠定了基础.

2.2 基于卷积神经网络(CNN)的方法

2.2.1 模 型

1989年, LeCun等人[9]首次用反向传播算法训练卷积网络, 成功实现手写数字识别, 而其在1998年提出的LeNet-5[10]作为第1个用于视觉任务(手写字体识别)的CNN结构成为了里程碑式的工作, 不仅证明了卷积结构的有效性, 还为日后的CNN模型提供了基础的组成结构设计. 但是受到当时计算资源不足、设备计算能力弱的限制, 人们无法训练更深、更复杂的神经网络. 而且在当时缺乏大规模的数据集, 远远不能满足训练更深层网络的需要. 因此对卷积神经网络的研究沉寂了近20年.

2009年ImageNet的提出构建了一个超大型的图像数据库, 1 400多万张图像和2万余种类别标签为神经网络的研究提供了充足的数据, 人们可以在此基础上训练更大的网络. 2012年AlexNet[11]的成功证明了更深层次的网络可以在大规模数据上训练后能够学习到图像中更高级的特征, 超越手工特征设计表现出其强大的表征能力. 自此, CNN由于其优秀的归纳偏置成为了计算机视觉中不可或缺的结构. 这些归纳偏置包括卷积层通过使用局部连接和权值共享, 使得神经元只与输入数据的局部区域相连接, 并且共享相同的权值, 这模拟了图像数据的局部相关性, 有助于网络捕捉图像中的局部结构和特征, 而且这种共享权值的卷积层使得网络对于图像的平移变换具有不变性, 提高了网络对于目标在图像中位置的鲁棒性, 使得模型能够更好地适应不同位置的目标, 并减少了参数数量, 提高模型的泛化能力. CNN通常采用多层次的结构, 通过卷积和池化层逐渐提取图像的层次化特征, 这有助于网络理解图像的层次结构和语义信息, 从边缘、纹理到更高级的语义特征. 卷积层的局部感受野使得网络能够专注于图像中的局部区域, 捕捉更为详细和抽象的特征. CNN中的池化层能够在减小数据维度的同时保留主要信息, 提高了网络的计算效率和内存利用率.

此后, 2015年提出的VGG结构[12]进一步加深了CNN模型, 但是人们逐渐发现, 单纯加深网络不仅不会获得更优秀的效果, 还会造成梯度消失或梯度爆炸等问题, 使得网络难以训练. He等人[13]在2016年提出了ResNet, 通过引入捷径连接(shortcut connection)缓解了深度神经网络训练中的梯度消失问题, 使用恒等映射(identity mapping)避免了模型退化. ResNet的成功应用在图像分类、目标检测、语义分割等任务上, 成为深度学习领域的重要里程碑之一, 其引入的残差学习思想为设计更深层次的网络提供了有效的方法, 自此绝大部分深度神经网络都会采取残差结构, 这一思想对之后许多深度神经网络的设计产生了深远的影响.

深度学习和CNN的兴起为目标检测带来了新的思路，使算法能够从数据中学习更高层次、更具表征能力的特征. 此后基于深度学习的架构被引入目标检测领域以提高目标检测算法在更广泛应用场景中的适应性和性能，这一转变标志着目标检测领域朝着更为智能、高效的方向发展.

1）R-CNN（regions with CNN features）[14]. R-CNN是首个成功将CNN集成到目标检测流程中的模型. R-CNN首先通过选择性搜索算法自下而上提取约2 000个区域建议作为候选框，然后将这些框缩放到统一尺寸以便输入预训练的CNN模型（如AlexNet）计算特征，最终用一组训练好的线性支持向量机对每个区域进行分类. R-CNN作为第一个成功将CNN融入目标网络流程的模型取得了显著的性能提升，其在PASCAL VOC 2007和PASCAL VOC 2010数据集上的平均精度分别从DPM-v5的33.7%和33.4%提升至58.5%和53.7%. 但R-CNN也有缺点，它从每张图提取的大量区域会发生重叠，需对每个区域统一尺寸并提取特征，这导致了大量冗余计算和低检测速度，GPU上处理1张图需13 s，CPU上则需53 s.

2）空间金字塔池化网络（spatial pyramid pooling net, SPP-Net）. 深度卷积网络的输入需要固定大小的图像，而为了满足这一要求，需要对一些不符合要求的图片进行裁切或放缩. 研究人员们认为，这种操作会降低模型对图像的识别精度. 于是2014年，He等人[15]提出了SPP-Net，通过空间金字塔池化层，使CNN能生成固定长度的表示，无需调整图像尺寸. 具体来说，SPP-Net在最后1个卷积层执行不同尺度的池化，生成固定长度特征向量. 在目标检测中，SPP-Net仅需计算1次特征图，减少了R-CNN中的重复卷积计算，显著提升了速度. 在最后的分类阶段，仍然为每个类别训练了1个线性SVM分类器. SPP-Net在保持与R-CNN相似精度的同时，整体速度可以提升24~64倍，提取卷积特征可以加速30~170倍. 尽管SPP-Net在精度和速度上取得显著进步，但仍有不足：仍然没有摆脱分类阶段的SVM，需要多阶段的训练，且SPP-Net只微调了卷积层后的全连接层，仍然有较大的进步空间.

3）Fast R-CNN. 为改善R-CNN和SPP-Net的局限性，Girshick[16]于2015年提出了Fast R-CNN. 该模型首先使用深度卷积网络提取图像特征图，然后通过选择性搜索确定感兴趣区域（region of interest, RoI）. 每个RoI被重新映射到特征图的对应位置，并通过感兴趣区域池化（region of interest pooling, RoI Pooling）处理，使不同大小区域转换为统一尺寸特征. Fast R-CNN设计了多任务损失函数，结合Softmax分类和边界框回归进行联合训练，同时采用非极大值抑制（non-maximum suppression, NMS）[17]获得最终检测结果. 该网络通过共享特征图加速处理，并将分类和定位任务整合到一个框架中，实现一次性模型更新，简化训练和减少开销. Fast R-CNN训练速度比R-CNN快9.5倍，测试速度提升213倍，精度也有显著提升. 尽管取得显著成果，Fast R-CNN的速度仍受生成候选区域步骤的限制.

4）Faster R-CNN. 2015年，Ren等人[18]提出的Faster R-CNN集成了区域建议网络（region proposal network, RPN）到目标检测框架中，显著提高了算法速度，能在NVIDIA Tesla K40 GPU上达到17 fps，成为满足一定实时性要求的目标检测模型. 具体来说，Faster R-CNN使用CNN作为骨干网络提取图像特征，特征图在RPN网络和RoI池化中共享. RPN通过滑动窗口和先验锚框在特征图上生成候选框和得分，然后通过RoI池化获得特征向量，经过全连接层进行分类和边界框回归. Faster R-CNN的端到端学习框架整合了特征提取、候选区域生成、分类和边界框回归，通过RPN提高了区域提议质量，进一步提升了目标检测精度.

尽管Faster R-CNN是端到端的网络，但它因包含候选建议生成步骤而被归类为2阶段目标检测算法. 这种算法通过分离生成候选建议和目标分类回归步骤，提高了检测框召回率和检测精度. 然而，这也导致2阶段算法速度较慢，整体结构复杂. 与此相对，1阶段目标检测算法在单次推理中完成所有目标的位置估计和分类，特点是实时性高、易于部署，但在小目标检测上表现不佳.

5）YOLO（you only look once）. 2015年，Redmon等人[19]提出首个不基于区域的目标检测算法YOLO模型. YOLO不预先生成感兴趣区域，而是一次性直接在整张输入图像上进行目标检测. 它将图像划分为网格，每个网格预测多个检测框和类别概率，利用置信度和交并比进行非极大值抑制以去除冗余检测框. YOLO的速度非常快，在PASCAL VOC 2007上能够以45 fps运行并且平均精确率达到63.4%. 更小型的YOLO版本在同样数据集上以155 fps运行，平均精确率为52.7%. YOLO的局限在于每个网格仅预测1个物体，且对小目标检测能力较弱，易产生定位误差.

YOLO作为一种新颖范式，开辟了目标检测研究

的新方向. 后续众多算法是基于 YOLO 设计思想提出. YOLO 的后续版本包括 YOLO9000[20], 使用更有效的 DarkNet-19 骨干网络, 引入批归一化[21] 加速网络收敛, 使用统计生成的先验锚框优化查询能力, 预测边界框偏移量而非坐标. YOLOv3[22] 使用更大的 DarkNet-53 骨干网络, 引入多尺度预测. YOLOv4[23] 则提出了马赛克数据增强和引入 Mish 激活函数[24], 进一步提高性能.

6)SSD. Liu 等人[25] 为了解决 YOLOv1 中的缺点, 提出了另一种 1 阶段目标检测算法 SSD, 使用一个 VGG-16[12] 骨干网络提取图像特征, 并从不同阶段提取的不同尺寸特征图来实现多尺度对象检测. SSD 借鉴 Faster R-CNN 的先验锚框概念, 并为不同尺寸的特征图设计了不同数量的默认框(default boxes), 以捕捉各位置的潜在目标. 与 YOLO 不同, SSD 完全使用卷积层, 以不同的卷积核为类别和检测框生成输出. 训练阶段中, 首先按交并比匹配 ground-truth, 再为其他可能的检测结果匹配 ground-truth, 匹配失败的定义为背景类. 为平衡正负样本, 采用 Top-k 筛选负样本. 推理阶段使用 NMS 过滤输出. SSD 在 MS COCO 和 PASCAL VOC 数据集上表现卓越, VOC 2007 test 上的平均精确率最高达 81.6%, VOC 2012 test 上的平均精确率为 80.0%. SSD 虽然相比 YOLO 有明显加速, 但其手工设计默认框仍非常依赖经验.

7)特征金字塔网络(feature pyramid network, FPN). FPN 由 Lin 等人[26] 提出, 旨在解决典型神经网络中低层特征图分辨率高但语义特征弱, 以及高层特征图语义丰富但分辨率低的问题, 这在目标检测任务中尤为明显(如 YOLOv1, Fast R-CNN 等). 虽然 SSD 通过引入多尺度特征图来探索解决方案, 但仍受到底层特征图语义弱的限制. FPN 使用类似 U-Net[27] 的高层-底层语义融合方法, 先通过逐步降采样获得不同尺度特征图, 然后从最小特征图开始逐步上采样并与对应层融合, 分别进行目标预测. 通过这种方式, FPN 为各层特征图提供了丰富的语义信息, 显著提高了网络性能. 因为其未引入额外复杂计算, FPN 以 6 fps 在 GPU 上运行, 满足部分实时检测需求.

8)RetinaNet.1 阶段检测器速度快但精度不足, 而 2 阶段检测器尽管精度高但速度慢. 在 1 阶段与 2 阶段检测器各自发展的背景下, Lin 等人[28] 针对 1 阶段检测器中存在的训练样本类别不平衡问题提出了 Focal Loss. 在 SSD 中, 尽管采用 Top-k 筛选负样本来缓解类别不均衡, 但仍有大量负样本干扰模型学习. Lin 等人[28] 观察到, 当负样本过多时, 其损失在总损失中占主导, 导致模型偏向多样本类别. 为解决此问题, 提出 Focal Loss, 通过正负样本比例因子和自适应动态缩放的交叉熵损失调节模型关注难分类样本. 为验证 Focal Loss 的有效性, Lin 等人[28] 设计了 RetinaNet. 它使用 ResNet 作为特征提取网络, 以及使用 FPN 进行特征融合, 并设有 2 个卷积分支生成检测标签和框. RetinaNet 在速度上与 1 阶段检测器相似, 但在 Focal Loss 训练下超越所有 2 阶段检测器. 基于 ResNet-101-FPN 的 RetinaNet 在 MS COCO test-dev 上的平均精确率能够达到 39.1%, 运行速度为 5 fps. 其检测方法的简洁性和有效性使其成为后续目标检测任务中骨干网络应用的标准框架.

9)Cascade R-CNN. Cai 等人[29] 同样关注到了正负样本不均衡问题, 但主要关注目标检测中的交并比. 他们观察到, 交并比阈值设定对检测性能有显著影响: 太低会产生干扰检测框, 太高则抑制正样本生成, 容易导致过拟合. 因此提出了一个级联的 R-CNN 算法, 即 Cascade R-CNN. 通过分析不同交并比阈值下的训练过程, 他们发现用某个交并比阈值训练的检测器能为下一个更高交并比阈值的检测器提供良好的结果分布. 级联 R-CNN 通过每个阶段找到良好结果, 逐步调整边界框来训练后续阶段, 减轻过拟合压力, 提升训练效果. 在推理中, 逐步修正的检测结果与每个阶段的检测器更好匹配, 提高了精度. 即便使用基础的 Cascade R-CNN 模型, 也能显著超越当时所有先进的单模型目标检测器. 与 RetinaNet 使用相同骨干网络时, Cascade R-CNN 在 MS COCO test-dev 数据集上的平均精确率达到了 42.8%, 表现更优.

10)CornerNet. 传统基于锚框(anchorbox-based)的目标检测方法[16,18,20,25,28] 通常需要手动设计固定数量和尺寸的先验锚框, 不仅需要先验知识, 而且对不同数据分布缺乏泛化能力. 而且, 基于锚框的方法对于锚框的设计高度敏感, 需要较多人工来对先验框设计进行探索. 且通常需要大量先验框来满足更高的检测召回率, 计算量巨大, 且这样会造成更多的锚框只能覆盖负类样本(即不包含物体类的样本), 导致更多的正负样本不均衡. YOLO 第 1 次对无锚框(anchorbox-free)进行探索, 在 YOLO 中没有显式定义锚框的尺寸, 在训练中直接回归生成检测框, 但 YOLO 仍然指定了锚框的个数. Law 等人[30] 提出了一个完全无锚框的模型 CornerNet, 将目标检测转换为检测框的左上角点和右下角点的关键点检测问题. CornerNet 使用 Hourglass Networks[31] 作为骨干网络, 利用 2 个独立并行分支预测左上角点和右下角点. 引

入的角点池化机制通过计算特征图每个点右侧和下方的最大值并相加，生成最终预测的热力图、嵌入向量（判断角点是否属于同一物体）和位置偏移量. 尽管 CornerNet 在无锚框目标检测领域取得显著进步，但仍存在局限. 由于仅使用像素点的右侧和下方进行池化，在物体密集场景中角点匹配可能失败. 同时，CornerNet 仅用 2 个角点描述目标，未考虑物体本身特征，可能导致误检，因此需要进一步改进.

11）CenterNet. 为了进一步解决 CornerNet 的限制，Zhou 等人[32] 提出了 CenterNet，将目标实例建模为 1 个点，直接预测物体的中心点，再从中心点回归生成 2 维检测框的宽和高来完成检测任务. CenterNet 首先使用 1 个骨干网络进行特征提取来生成热力图，每一个峰值点（peak）对应 1 个物体，免去了非极大值抑制等后处理，最后使用 peaks 和周围值进行回归生成检测结果. 另一个基于 CenterNet[33] 的工作中，将目标检测定义为左上角、右下角和物体中心的 3 元组. 设计了级联角池化和中心池化模块，丰富角点信息，并在中心区域提供更多可识别的信息来提高检测性能. 在 MS COCO test-dev 上的平均精确率达到 47.0%，与当时最先进的 2 阶段检测器拥有相当的精度.

12）FCOS. Anchorbox-free 类算法不需要设置大量的 anchor 相关超参数，受到研究人员们关注. 受全卷积模型[34] 在分割领域成功的影响，Tian 等人[35] 提出了 FCOS，一种基于完全卷积（fully convolutional）的 1 阶段目标检测算法. 在 FCOS 中，首先使用预训练的 ResNet-50 网络和 FPN 结构进行特征提取. 在训练过程中 FCOS 将特征图上的每一个点都视为一个样本，如果这个点处于任何一个 ground truth 中，则视为正样本，反之为负样本. 处理重叠 ground truth 框时，点被视为模糊样本，选取最小的 ground truth 框作为回归目标. 与 CenterNet[32-33] 寻找物体中心点不同，FCOS 允许任何在 ground truth 框内的点回归检测框. 并且因为使用了特征金字塔，FCOS 可以将不同尺寸的物体分配给不同大小的特征级来缓解模糊点过多的问题. FCOS 还在分类分支的尾端引入一个额外的 center-ness 分支，预测当前点与要预测的物体中心点之间的归一化距离，减少低质量框的生成.

13）Sparse R-CNN. Sun 等人[36] 提出 Sparse R-CNN，旨在解决当前目标检测网络对密集候选对象的依赖问题. 传统方法通常需要非极大值抑制等后处理，依赖于图像上预定义的大量锚点或参考点，以及需要抑制的冗余检测框. 受 DETR[37] 的启发，该研究团队[36] 探索将目标检测转换为稀疏集合预测问题，仅依赖于一组（100 个）可学习的对象查询来生成最终检测结果. 然而，DETR 中每个对象查询需与全局图像交互计算注意力，不是纯粹使用稀疏方法. 而 Sparse R-CNN 只需 100 个稀疏的起始框（每个框由左上和右下点的坐标组成）即可预测图像中的所有对象. 每个框只与稀疏特征进行交互，通过独特的动态头为每个特征生成对象特征，完成分类和定位，无需非极大值抑制后处理. Sparse R-CNN 避免了所有密集步骤，实现了准确度和运行时间的平衡，达到优异性能.

14）YOLOF（you only look one-level feature）. 由 Chen 等人[38] 提出，是对特征金字塔结构（FPN）的一种简化和创新. FPN 在 1 阶段和 2 阶段检测器中普遍使用，其成功被认为是多尺度特征融合和不同尺寸特征图处理不同大小目标的分治策略. 然而，Chen 等人[38] 通过实验证明，FPN 的成功主要归因于不同尺度感受野对不同规模目标的处理能力. YOLOF 基于这一发现，提出无需复杂的多尺度融合，只使用骨干网络的最后一层特征图就能完成检测任务. YOLOF 使用 ResNet[13] 或 ResNeXt[39] 进行特征提取，并引入扩张卷积以获取不同尺度的感受野. 此外，YOLOF 挑战了深层特征图只适用于大物体检测的常规观点，证明其仍包含小物体信息. 为此，提出了 Uniform matching 规则，为每个 ground-truth 找到 k 个最近邻的锚框，并用交并比阈值进行筛选，实现正负样本均衡. YOLOF 的提出表明，通过合理的特征提取和锚框匹配策略，可以在不牺牲性能的情况下大幅提升目标检测任务的运行效率.

15）YOLOX. YOLO 系列检测模型[19-20,22-23] 由于其优秀的精度-速度平衡，以及 C 语言的实现便于在不同平台上部署，一直受到工业界的青睐. 但 YOLO 仍是基于锚框的算法，存在先验框设计的泛化性不足和复杂度高的问题. 因此，Ge 等人[40] 提出 YOLOX，它是 YOLO 系列的一个重大升级，融合了无锚框思想和其他先进的检测技术，如解耦头、更优秀的标签分配等. YOLOX 证明解耦头（分别处理分类、检测框、前景/背景预测）比耦合检测头（一个输出分支负责所有预测）有更好的性能. 在各种模型大小下，YOLOX 都能超越同类模型实现更好的精度-速度权衡. 小模型版本的 YOLOX-Tiny 能够在 MS COCO val 数据集上以 5.06×10^6 的参数量和 6.45 GFLOP 实现的平均精确率为 32.8%，优于当时同等大小的模型 PPYOLO-Tiny. 而使用 CSPNet[41] 为骨干网络的 YOLOX-L 能在 NVIDIA Tesla V100 GPU 上以 69.0 fps 的推理速度使平均精确率达到 50.0%. YOLO 系列网络通过改进仍

然是工业界最受欢迎的目标检测技术.

16)YOLOv7. YOLOv7 由 Wang 等人[42]提出,旨在进一步提升 YOLO 系列的性能,同时保持实时目标检测的特性,特别适合于需要快速检测的应用如多目标跟踪、自动驾驶等. YOLOv7 着重于优化梯度传播路径和动态标签分配策略,以提高深层网络的学习效率和收敛速度,同时减少参数和计算量. YOLOv7 在不同层引入模型重参数化技术[43],控制网络中的最短和最长梯度路径,使得更深的网络也能有效学习和收敛. 此外,采用了由粗到细的动态标签分配策略,基于聚类为每个 ground-truth 分配更多正样本,并使其周围网格也负责预测. 通过使用分类头的粗输出指导辅助头和分类头共同训练,并在推理阶段移除辅助头以减少计算量,YOLOv7 有效降低了参数数量和计算需求. 在性能上,YOLOv7 在 5~160 fps 的范围内超越所有已知的目标检测器,实现了速度与准确性的优秀平衡.

YOLOv7 在保持较低计算需求的同时,有效提升了检测精度和速度,非常适合在资源受限的实时检测场景中应用.

2.2.2 小 结

表 2 总结了基于 CNN 的不同模型的各自特点. 同时,表 3 给出不同模型的优点及不足. 为了进一步比较不同模型的性能,我们在表 4 中给出不同模型中公开数据集上的平均精确率的实验结果. 最后,我们通过图 3 展示了不同模型在原有 CNN 上的改进.

CNN 的核心优势在于其归纳偏置,如卷积层的局部连接和权值共享,使得网络能够捕捉图像中的局部结构和特征,并提高对目标位置的鲁棒性. 多层次结构的设计,包括卷积和池化层,有助于网络逐层提取图像的特征,从基本的边缘和纹理到更高级的语义信息. VGG[16]结构进一步加深了 CNN 模型,但随之而来的是梯度消失或爆炸问题. ResNet[13]通过引入残差学习和捷径连接(shortcut connection)来解决这一问题,成为深度学习领域的一个重要里程碑,并深刻影响了后续的网络设计. 而在目标检测领域,CNN 的引入标志着算法朝着更智能、高效的方向发展.

CNN 结构在 20 世纪 90 年代后的 30 年经历了短暂了发展低谷期,又在 2012 年随着深度学习的发展再度崛起. 而自 Transformer 结构出现,其在自然语言处理任务中的强大性能,以及 ViT(vision Transformer)[44]提出证明 Transformer 可以作为新的骨干网络有效迁移到视觉任务后,类似的基于 Transformer 结构在各种视觉任务上都展现出了超越基于 CNN 结构的表现. CNN 结构发展至今,人们开始思考,是否 Transformer 结构本身的能力超越了卷积,仅通过 Transformer

Table 2 Technical Points of Target Detection Model Based on CNN

表 2 基于 CNN 的目标检测模型技术点

模型	骨干网络	端到端	阶段	区域建议	多尺度特征	NMS	输出层
R-CNN	AlexNet	✗	2 阶段	SS	✗	✗	SVM
SPP-Net	AlexNet	✗	2 阶段	SS	✓	✓	SVM
Fast R-CNN	VGG	(训练)✓,(推理)✗	2 阶段	SS	✗	✓	FC
Faster R-CNN	VGG	✓	2 阶段	RPN	✗	✓	FC
YOLO	DarkNet	✓	1 阶段	✗(网络)	✗	✓	FC
YOLO9000	DarkNet-19	✓	1 阶段	✗	✗	✓	Conv
YOLOv3	DarkNet-53	✓	1 阶段	✗	✓	✓	Conv
SSD	VGG-16	✓	1 阶段	✗	✓	✓	Conv
FPN	ResNet-101	✗	2 阶段	RPN	✓	✓	FC
RetinaNet	ResNeXt-101	✓	1 阶段	✗	✓	✓	Conv
Cascade R-CNN	ResNet-101	✓	2 阶段	RPN	✓	✓	FC
CornerNet	Hourglass-104	✓	1 阶段	✗	✓	✓	Conv
CenterNet	Hourglass-104	✓	1 阶段	✗	✓	✗	Conv
FCOS	ResNeXt	✓	1 阶段	✗	✓	✓	Conv
Sparse R-CNN	ResNeXt-101	✓		可学习的提议目标框	✗	✗	FC
YOLOF	ResNet	✓	1 阶段	✗	✓	✓	Conv
YOLOX	Modified CSP	✓	1 阶段	✗		✓	Conv

注:✗为不支持的特性;✓为支持的特性;SS 为选择性搜索;SVM 为支持向量机;FC 为全连接层;Conv 为卷积层.

Table 3 Advantages and Disadvantages of CNN-Based Target Detection Models

表 3 基于 CNN 的目标检测模型优缺点

模型	出版物	优点	缺点
R-CNN	文献 [14]	第 1 个使用 CNN 的目标检测模型	需要大量冗余的特征计算
SPP-Net	文献 [15]	避免重复计算特征图	仍然需要多个二分类支持向量机
Fast R-CNN	文献 [16]	使用全连接层来输出结果	速度受选择性搜索算法影响
Faster R-CNN	文献 [18]	引入区域建议网络生成候选区域	2 阶段网络速度慢，算法较复杂
YOLO	文献 [19]	第 1 个 1 阶段算法，运行速度快	对小尺寸目标识别能力差
YOLO9000	文献 [20]	可识别种类更多，速度更快	仍未完全解决小目标检测能力差的问题
YOLOv3	文献 [22]	引入多尺度特征图	检测能力仍然较 1 阶段检测器差
SSD	文献 [25]	使用纯卷积结构和多尺度特征图	手工设计先验框非常依赖经验
FPN	文献 [26]	引入里程碑式的特征金字塔结构	略微增加了模型的复杂度和计算开销
RetinaNet	文献 [28]	解决正负样本不均衡问题	训练过程可能更复杂
Cascade R-CNN	文献 [29]	通过级联的方式显著提升检测精度	增加了计算成本且可能导致运行速度变慢
CornerNet	文献 [30]	以 2 个角点来建模目标检测	角点匹配较难，没有考虑物体本身的特征
CenterNet	文献 [32]	直接预测目标的中心点	无法处理中心点接近重叠的情况
FCOS	文献 [35]	完全由卷积结构组成，结构简单	重叠目标多的场景可能需要更复杂的后处理
Sparse R-CNN	文献 [36]	以完全稀疏化的操作来建模目标检测	学习高质量候选区域需要更多训练过程
YOLOF	文献 [38]	取得更好的正负样本均衡	仅使用最高层特征会限制多尺度检测能力
YOLOX	文献 [40]	YOLO 结构的进一步优化	对非常小的目标性能不佳
YOLOv7	文献 [42]	具有更强大的实时性	最高精度可能不如其他检测算法

Table 4 Experimental Results of CNN-Based Object Detection Model on Public Datasets

表 4 基于CNN的目标检测模型在公开数据集上的实验结果 %

模型	VOC 2007 test	VOC 2010 test	VOC 2012 test	COCO val		COCO test-dev	
	AP	*AP*	*AP*	*AP*	AP_{50}	*AP*	AP_{50}
R-CNN	58.5	53.7		-	-	-	-
SPP-Net	59.2	-		-	-	-	-
Fast R-CNN	70.0	68.8	68.4	-	-	-	-
Faster R-CNN	69.9	-	70.4	21.2	41.5	21.9	42.7
YOLO	63.4	-	57.9	-	-	-	-
YOLO9000	78.6	-	73.4	-	-	21.6	44.0
YOLOv3	-	-	-	-	-	33.0	57.9
SSD	76.8	-	80.0	-	-	26.8	46.5
FPN	-	-	-	-	-	36.2	59.1
RetinaNet	-	-	-	-	-	40.8	61.1
Cascade R-CNN	-	-	-	-	-	42.8	61.1
CornerNet	-	-	-	-	-	42.2	57.8
CenterNet(Zhou)	80.7	-	-	45.1	63.5	45.1	63.9
CenterNet(Duan)	-	-	-	41.3	59.2	47.0	64.5
FCOS	-	-	-	-	-	44.7	64.1
Sparse R-CNN	-	-	-	46.4 🖱	64.6 🖱	51.5 🖱	71.7 🖱
YOLOF	-	-	-	47.1 🖱	66.4 🖱	44.3 🖱	62.9 🖱
YOLOX	-	-	-	47.3 🖱	-	51.2 🖱	69.6 🖱
YOLOv7	-	-	-	55.9 🖱	73.5 🖱	56.0 🖱	73.5 🖱

注: “-”表示在原始论文中没有进行实验或未公开的实验结果，🖱表示 2017 年版本的 MS COCO 数据集实验结果，AP_{50} 表示在交并比大于 50% 时的 *AP*.

这一种网络是否能够实现文本和视觉模态的统一. 但是 ConvNeXt[45], RepLKNet[46], SlaK[47], UniRepLKNet[48] 等工作依然证明了, 尽管是在“Transformer 时代”, 卷积结构仍有很多潜力值得探索. 通过巧妙地分析卷积结构的作用, 以及增大卷积核来赋予传统 CNN 结构类似 Transformer 的全局建模能力等, 人们再次证明了 CNN 结构仍然可用, 并且依旧能在图像、视频、音频、点云领域达到 SOTA(state-of-the-art)表现, 甚至这种为图像任务设计的骨干网络还能在时序数据上超越以序列建模为优点的 Transformer. 而在移动端和嵌入式系统等边缘环境中, CNN 便于实现和优化部署的特点依旧使得其能够被大量应用, 而且在这些边缘设备中, 大部分情况下都需要神经网络能够实时地完成任务. 因此人们需要轻量化的骨干模型来实现各种视觉任务, 比如工业安防场景中的异常检测、自动驾驶场景中的实时感知等. 比如 MobileNet[49-52], ShuffleNet[53-54], EfficientNet[55-56] 和其目标检测任务优化模型 EfficientDet[57], GhostNet[58-59] 以及最新的 MobileOne[60]. 轻量级视觉骨干网络至今已经能够实现与普通的 ResNet 相媲美的精度以及在移动设备上毫秒级的推理时间. 最新的研究趋势[46-48] 表明, 尽管 Transformer 架构的引入为深度学习带来了全新的思路和可能性, 但是对 CNN 结构的深入研究仍然是不可忽视的, 卷积架构和 Transformer 之间并没有哪一种具有本质上的优越性, 也不是 2 条绝对独立的发展方向, 而是 2 种相互交叉发展、相辅相成的设计思路. 基于 CNN 的目标检测模型的概述如图 3 所示.

2.3 基于 Transformer 架构的方法

2.3.1 模 型

“注意力机制”并不是全新的概念, 在人类对图

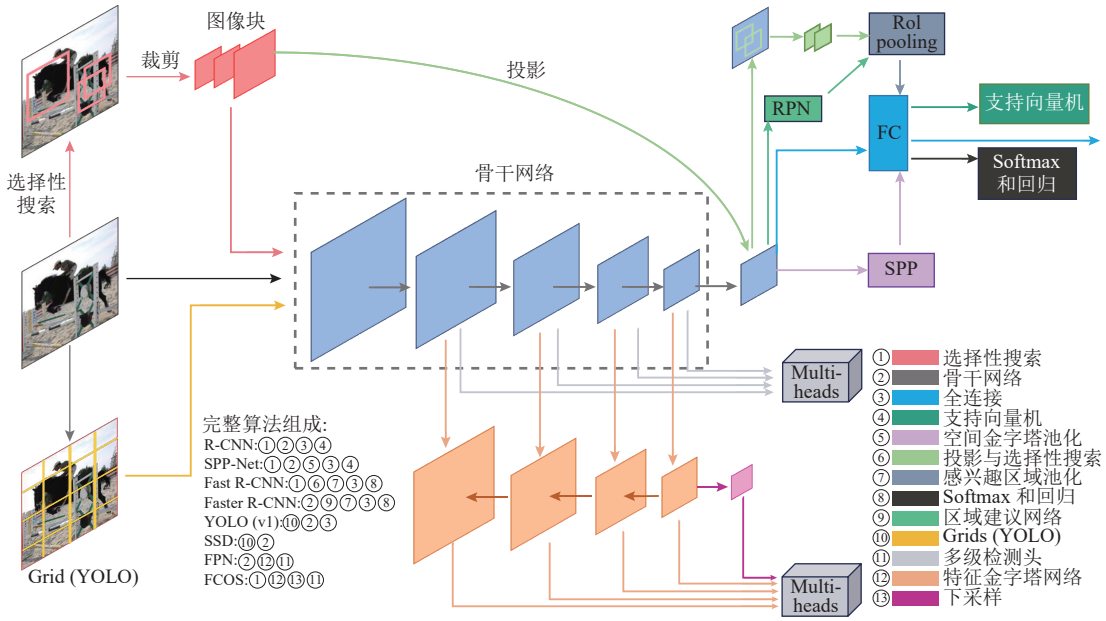


Fig. 3 Overview of CNN-based object detection models

图 3 基于 CNN 的目标检测模型的概述

像的观察中，会下意识地选择更重要的空间位置进而为其分配更多的注意力. 一些传统的计算机视觉算法使用的手工特征也可以被视为一种类注意力机制，通过人工筛选更有效的特征使得视觉算法更有效地关注对完成任务更重要的信号.

在深度学习时代，2014 年 Mnih 等人[61]将注意力机制应用于计算机视觉，此后大量工作[62-69]都关注在 CNN 上研究注意力机制的作用. 2017 年，Google 公司的研究人员提出了 Transformer[70]网络结构，这是一个不含卷积和循环，完全基于多头自注意力(multi-headed self-attention, MHSA)机制的模型，用于自然语言处理任务. 这种编码器-解码器(encoder-decoder)结构在多个数据集上展现了出色性能. 同时，BERT 模型[71]已经展示了基于 Transformer 的网络在大规模数据集上的预训练和对下游任务的微调效果. 得益于其高效的计算性能和卓越的可扩展性，基于 Transformer 的模型能够处理大规模数据且性能未显饱和. 虽然 Transformer 在自然语言处理领域已成为标准模型，但其在计算机视觉领域的应用还在初步探索阶段.

1)DETR(detection transformer). DETR[37]提出了全新目标检测范式，通过将目标检测视为集合预测任务，巧妙地融合了 Transformer 的编码器-解码器架构. 传统目标检测算法依赖于生成大量区域建议或锚框，而 DETR 消除了这些手工设计的需求，DETR 首先使用 CNN(例如 ResNet)提取特征，然后通过 Transformer 编码器处理这些特征以提取全局自注意力. 接着，使用解码器模块结合随机初始化的对象查询来生成目标检测结果，这个过程不再需要复杂的后处理，如非极大值抑制(NMS). 在训练阶段，DETR 使用二部图匹配和匈牙利算法来计算损失，从而避免了重复预测问题. 虽然在 MS COCO 数据集上的性能还无法匹敌当时的最佳模型，但 DETR 因其创新性引起了广泛关注，并催生了许多基于 DETR 的改进模型. DETR 也有局限，如训练速度慢，尤其是对小目标检测的性能较差，原因是其在高分辨率特征图上的自注意力计算复杂度较高.

2)Deformable DETR. 受到可变性卷积网络(deformable convolutional networks, DCN)[63-64]启发，Zhu 等人[72]结合可变形卷积的稀疏空间采样和 Transformer 的关系建模能力的优点，提出了一个改进的框架 Deformable DETR 用以解决 DETR 的上述缺点. 具体来说，先提取 CNN 中最后 3 个不同大小的特征图作为特征输入进 Transformer 结构. 此外，它引入了可变形注意力模块，该模块只关注参考点周围少量关键采样点，而非整个特征图. 这种方法通过在每个参考点周围仅选取有限的点来计算注意力，引入了稀疏化操作，从而减少了计算量. 这种机制使 Deformable DETR 在小物体检测上表现更佳，并且训练时间相比 DETR 减少到其 1/10. 在 MS COCO 2017 val 数据集上，仅需 50 个训练周期和 40×10^6 的参数量，Deformable DETR 的平均精确率就能达到 46.2%，同

时保持 19 fps 的推理速度，表明了其效率和效果的显著提升.

3）Efficient DETR, PnP-DETR, Sparse-DETR. 受制于 DETR 的高计算开销，更多的研究人员们开始探究更高效地训练 DETR 和推理. Yao 等人[73] 设计了大量实验来分析 DETR 中每一个子结构对模型效能的影响，并且受 Deformable DETR 中参考点的启发，提出了 Efficient DETR，通过进一步优化目标查询的生成，有效减少不必要的模块堆叠来提高运行效率. Wang 等人[74] 认为 DETR 中注意力计算量大是受到图像中冗余的空间特征影响，使用特征图中所有点来计算注意力是不必要的，因此提出了 PnP-DETR，引入了一种 poll-and-pool 采样模块将图像特征图抽象为细前景对象特征向量和少量粗背景上下文特征向量，自适应地在空间上分配计算来减少计算量. Roh 等人[75] 发现在 Deformable DETR 中，编码器中的 token 过多使得编码器的注意力计算成本仍然存在瓶颈，而在训练中即使只更新一部分的编码器 token 也不会使得网络性能过多恶化，因此提出了 Sparse-DETR，有选择性地更新编码器中预期会使用的 tokens，通过引入编码器 token 稀疏化算法，即使在仅使用 10% 的编码器 tokens 时，Sparse-DETR 架构也优于 Deformable DETR，并将整体计算量减少了 38%，同时检测速度增加了 42%.

4）ViT. Dosovitskiy 等人[44] 观察到以往视觉算法中的注意力机制常与卷积操作结合或部分替换卷积网络. 基于此，他们提出了 ViT，这是一个完全基于自注意力机制的视觉架构，展示了即便不依赖于 CNN，也能在图像分类任务中取得良好表现. ViT 通过将图像分割成固定大小的 patches 并进行线性嵌入，结合位置编码，这样就能将其作为 Transformer 编码器的输入. ViT 的成功说明只要预训练数据集充足，Transformer 结构就可以超越其本身缺乏 CNN 架构那样的归纳偏置的限制，在计算机视觉任务上取得优异成绩.

Transformer 架构的高度可扩展性使其不受性能瓶颈的限制，能支持更大规模的模型，并且作为一种创新的基础网络框架，成功地迁移到了计算机视觉领域. 这一重大转变不仅体现了 Transformer 在处理视觉任务上的巨大潜力，也预示着它可能成为未来计算机视觉研究和应用的新范式.

5）UP-DETR. 既有的基于图像实例或聚类[76-78] 的对比学习方法并不适用于 DETR 这种空间定位学习任务. 为此，Dai 等人[79] 提出了 UP-DETR. UP-DETR 是一种通过无监督方式预训练的 DETR 模型，其灵感来源于先前在自然语言处理[71,80-81] 和计算机视觉[82-85] 任务中通过代理任务进行无监督大规模预训练的成功案例. UP-DETR 使用一种全新的代理任务——随机查询区域检测，来进行无需人工标注的无监督训练. UP-DETR 的核心是在输入图像上进行随机裁剪以获得多个 patch 区域，然后结合目标查询来训练 DETR 中的 Transformer 模块. 为了避免这种 patch 检测任务破坏已学习的分类特征，UP-DETR 不仅冻结了 CNN 模块，还引入了一种特征重建损失函数以保留这些特征. 在解码器部分，该网络训练以定位随机裁剪的 patch 位置. 通过这种无监督预训练，UP-DETR 在各种下游任务上表现出色，这表明，UP-DETR 有效地将无监督预训练应用于 DETR，提高了其在下游任务中的性能.

6）PVT（pyramid vision Transformer）. Wang 等人[86] 提出 PVT 模型，解决了 ViT 结构在处理高分辨率图像时计算负担过大的问题. PVT 通过在 Transformer 中引入特征金字塔，并使用多尺度的 Transformer 解码器来提取特征，辅以空间归约注意力（spatial reduction attention），有效降低计算量的同时保持特征图分辨率和全局感受野，这使 PVT 能够适用于密集预测任务. PVT 虽然改进了特征计算过程，但是当输入图片分辨率增大的时候，资源消耗仍然会比基于卷积网络架构的网络增长得更快. 而且当特征大小调整时，位置嵌入仅通过简单的插值算法来调整，仍然还有优化的空间.

7）Swin Transformer. Swin Transformer[87] 是为解决 ViT 在目标检测等视觉任务中的挑战而设计的，这些挑战主要源自于视觉和语言领域的差异，尤其是元素规模的差异和图像分辨率高于文本的问题. Swin Transformer 通过引入基于滑动窗口（shifted windows）的架构，局部窗口内进行自注意力计算，引入局部感受野. 此外，移动窗口机制允许跨窗口的连接，促进信息交互，从而提高效率. 这种分层架构可以灵活地在不同尺度上建模，并保持相对于图像大小的线性计算复杂度. Swin Transformer 的这些特性不仅使其在对象检测和语义分割等多种视觉任务中应用广泛，还大幅提升了性能.

8）YOLOS（you only look at one sequence）. 尽管基于 DETR 的架构在目标检测领域取得了显著成果，但大多数框架仍依赖于基于 CNN 的特征提取网络. 为了探究纯基于 ViT 的网络能否在最少的 2 维空间结构先验知识下，以纯序列到序列的角度进行目标检测，Fang 等人[88] 提出了 YOLOS 架构，这一架构尽量

保持 ViT 模型的基础结构不变,并添加了尽可能少的额外归纳偏置. 它使用一组随机初始化的、可学习的“检测”标记(DET tokens)代替了原始 ViT 结构中的“分类”标记(CLS tokens),作为对象代理来生成模型的最终预测. Transformer 产生的对应“检测”标记的输出通过多层感知机来产生最终的标签分类和边界框坐标. 参照 DETR, YOLOS 也使用最佳二分匹配损失进行训练. 最终,在 ImageNet-1k 上预训练的 YOLOS 能够在 MS COCO 上微调后的平均精确率最高为 42.0%.

9)GLIP(grounded language-image pre-training). GLIP 由 Li 等人[89]提出,旨在解决 OpenAI 先前提出的跨文本-图像多模态模型 CLIP[90]在精细视觉任务中的局限性. CLIP 通过对比学习从大量的图像-文本对中学习图像级视觉表征,尽管在分类任务中表现出色,但在目标检测或实例分割等需要更精细的对象级语义表示的任务中表现不足. GLIP 通过使用 BERT 提取文本特征编码和 Swin Transformer 提取图像特征编码,结合一个深度融合模块进行跨模态特征融合,解决了这一问题. 它将目标检测描述为“phrase grounding”,即将文本提示词中代表目标物体的单词与图像中相应区域关联起来. 此外,为了提高效果,Li 等人[89]先训练了一个 GLIP-T 教师模型,在手工注释的数据集上生成基础数据,并使用这些数据以及其他标注好的数据一起训练 GLIP 学生模型. 这种教师-学生模型的训练方法使得学生模型在多个数据集上超过了教师模型的表现. 在 MS COCO val 2017 上,未经微调的 GLIP 模型的平均精确率可以达到 49.8% 的 zero-shot 检测效果,而经过 COCO 上微调后,在验证集上的平均精确率可以达到最高 60.8%,在 test-dev 上平均精确率达到最高 61.5%. GLIP 的成功表明大规模无监督预训练的多模态模型能够有效迁移到对象级别的细粒度视觉任务.

10)DAB-DETR, DN-DETR. Liu 等人[91]在 DAB-DETR 中提出,由于原始 DETR 的对象查询是随机初始化的,会导致训练速度慢且难以收敛,而且原有模型中的位置先验信息只存在于特征图中,不涉及解码器的对象查询. 为此,他们引入了动态锚框(dynamic anchor boxes)作为 Transformer 解码器的查询,并在各层中进行动态更新,从而为交叉注意力模块提供更好的空间先验知识. 这一改进增强了查询特征的相似性,加快了 DETR 的收敛速度. 在此基础上,Li 等人[92]研究了其他可能影响 DETR 的因素,指出匈牙利匹配的使用影响了训练稳定性. 由于匈牙利算法和训练的随机性导致解码器输出与 ground-truth 匹配不稳定,他们提出了基于查询去噪(query denoising)的 DN-DETR. 这一方法在训练过程中引入去噪任务,即输入是 ground-truth 叠加噪声,标签经随机变化,检测框进行中心点移动或大小缩放,而输出保持原始 ground-truth. 模型的训练任务是从带噪声的输入恢复原始的 ground-truth,通过一对一的去噪消除了原始匈牙利匹配算法中匹配不稳定问题.

11)Co-DETR. Zong 等人[93]发现 DETR 中查询(query)和 ground-truth 的一对一会导致分配为正样本的查询太少,引发对 Encoder 输出的稀疏监督,这损害了 Encoder 对判别特征的学习. 为了缓解这种情况,Zong 等人[93]使用协同混合分配训练,提出了 Co-DETR. 通过添加 1 对多标签分配的并行辅助头来监督增强编码器的学习能力,并且通过这些辅助头来添加额外的正查询来提高训练效率. 在推理阶段,可以直接丢弃这些辅助头. 所以在推理中不会为原始的检测算法添加更多的参数提高计算成本.

12)GCViT. Hatamizadeh 等人[94]分析 Swin Transformer 的结构发现,尽管这种基于滑动窗口的多分辨率结构试图找到区域注意力和全局注意力之间的平衡,但这种局部窗口注意力结构仍然对自注意力的远程信息捕获造成挑战. 因此,Hatamizadeh 等人[94]提出了一种基于全局上下文的 ViT 结构——GCViT. 具体来说,引入了一种由局部自注意力和全局自注意力机制组成的分层 ViT 结构,提出一种新的融合倒置残差模块来计算全局查询 tokens,其中包含了来自图像不同区域的全局上下文信息,这些信息以局部键-值对的表示在所有全局自注意力模块中共享. 通过这种操作,为类 ViT 架构引入了缺乏的归纳偏置,增强了通道间依赖关系的建模.

13)迭代多尺度特征聚合(iterative multi-scale feature aggregation, IMFA). 尽管多尺度特征被认为是解决大尺度差异物体检测的有效手段,但是在基于 DETR 架构的算法中,简单叠加多尺度特征会导致计算负担加重. Zhang 等人[95]提出了 IMFA,为基于 Transformer 的检测器引入通用的多尺度特征计算方法,Zhang 等人[95]观察到,在图像中由于背景通常占据了图像中的大部分空间,因此在高分辨率下计算特征是高度冗余的,而且相较于 CNN, Transformer 在计算注意力的时候不需要网格状的特征图,因此使得只从一些包含感兴趣对象的特定区域计算多尺度特征成为可能. 在 IMFA 中,将每个编码器立即连接到对应的解码器,使得可以迭代地更新编码器的图

像特征以及更精细化的预测. 此外, IMFA 只关注有更高概率出现物体的区域, 先在前景区域中搜索关键点, 然后自适应地选择尺度来对这些关键点周围的特征进行采样.

2.3.2 小 结

基于 Transformer 的目标检测架构的性能已经能够媲美甚至超越基于卷积的框架, 成为一个新的研究热点. 表 5 中概括性地总结了部分基于 Transformer 的目标检测算法的优缺点, 表 6 总结了这些算法在广泛使用的目标检测数据集 COCO 2011 val 上的实验结果. 人们在这些算法的准确性以及实时性上进行了大量的探索. 基于 Transformer 的骨干架构, 以及对 DETR 的分析和改进工作已成为当下的研究热点, 受限于篇幅, 我们无法介绍所有代表性的网络结构, 如果想多了解关于 Transformer 结构在目标检测或其他更多视觉任务乃至更多模态任务上的应用, 可以参考其他更有针对性的综述性文章[96-107].

Table 5 Advantages and Disadvantages of the Target Detection Model Based on Transformer

表 5 基于 Transformer 的目标检测模型优缺点

模型	出版物	优点	缺点
DETR	文献 [37]	将 Transformer 结构引入目标检测流程	计算复杂度高，训练难以收敛
Deformable DETR	文献 [72]	通过稀疏化的采样来选取参考点来加速	查询数增加至 300，有较高的浮点操作次数
Efficient DETR	文献 [73]	减少了 Transformer 块数，加速网络收敛	浮点操作次数较 DETR 有显著增加
PnP-DETR	文献 [74]	自适应地在空间上分配计算	模型实现较标准 DETR 更复杂
Sparse-DETR	文献 [75]	有选择地更新编码器中的 tokens	对超参数的选择可能更敏感
ViT	文献 [44]	使用纯粹 Transformer 的视觉骨干结构	计算复杂度高且难以直接迁移
UP-DETR	文献 [79]	为 DETR 引入无监督预训练方法	比优化后的 DETR 需要更多的训练轮次
PVT	文献 [86]	为 ViT 引入特征金字塔结构	网络消耗资源快速增长
Swin-Transformer	文献 [87]	引入滑动窗口赋予模型局部感受野	训练和推理速度较慢，需要更多计算资源
YOLOS	文献 [88]	以接近于原始 ViT 的结构实现目标检测	需要更多训练时间
GLIP	文献 [89]	使用文本-图像多模态预训练模型的算法	需要多种模态的训练数据，推理速度较慢
DAB-DETR	文献 [91]	引入更好的空间先验知识	引入了额外的复杂性
DN-DETR	文献 [92]	将网络的训练过程定义为一个去噪任务	只考虑操作正样本，增加模型计算负担
Co-DETR	文献 [93]	通过 1 对多方式分配正样本，增强能力	增加了训练过程的计算负担
GCViT	文献 [94]	全局自注意力和局部自注意力结合	计算成本更高，不便用于资源受限的环境
IMFA	文献 [95]	只对潜在的感兴趣区域计算多尺度特征	检测精度可能下降

Table 6 Experimental Results of the Transformer-Based Object Detection Model on Public Datasets

表 6 基于 Transformer 的目标检测模型在公开数据集上的实验结果

模型	*AP*/%	AP_{50}/%	浮点运算量/GFLOP	检测速度/fps	参数量
DETR	44.9	64.7	253	10	60×10^6
Deformable DETR	46.2	65.2	173	19	40×10^6
Efficient DETR	45.7	64.1	289		54×10^6
PnP-DETR	43.1	63.4			
Sparse DETR	49.3	69.5	144	17.2	41×10^6
UP-DETR	42.8	63.0			41.3×10^6
PVT(RetinaNet)	41.9	63.1			53.9×10^6
YOLOS	42.0		538	2.7	127×10^6
DAB-DETR	46.6		296		63×10^6
DN-DETR	48.6	67.4	195		48×10^6
Co-DETR	65.9				304×10^6
IMFA	45.5	65.0	108		53×10^6

2.4 基于 MLP 的方法

2.4.1 模 型

2021 年, Tolstikhin 等人[108] 发现尽管卷积和注意力机制是视觉模型优秀性能的充分条件, 但并不是必要条件. 他们提出了一种完全基于 MLP 的视觉架构 MLP-Mixer, 与 ViT[44] 类似, 将输入图像分割为一定数目对 patches, 通过逐 patch 全连接层投影成对应的嵌入(embedding)表示, 又引入了 2 种 MLP 层: 令牌混合(token-mixing) MLP 层和通道混合(channel-mixing) MLP 层, 分别进行位置特征混合和不同通道之间的通道信息交流, 2 种 MLP 层交替堆叠, 最后通过全局平均池化和 1 个全连接层实现图像分类. 通过使用一些正则化手段和大规模数据集的训练, MLP-Mixer 可以在 ImageNet 上获得与最新的基于 CNN 和基于 Transformer 的模型相媲美的 87.94% 的 Top-1 准确率. 但是 MLP-Mixer 的 patch 数量随着输入大小的变化而变化, 因而不能直接使用预训练并在其他分辨率上直接微调, 这使得 MLP-Mixer 无法被转移到检测和分割等下游视觉任务中. 而且 MLP-Mixer 只关注全局语义信息, 而局部感受野已经在 CNN 和一些基于 Transformer 的架构(如 Swin Transformer 等)中证明了是一种有力的归纳偏置.

MLP-Mixer 展示了纯 MLP 架构在计算机视觉任务中的潜力, 激发了研究人员对更多 MLP 结构的探索. 随后改进的基于 MLP 的视觉骨干网络 ResMLP[109], RepMLP[110], Permute-MLP(也称 ViP)[111], gMLP[112], S^2-

MLP[113], DynaMixer[114] 被提出. 因此基于 MLP 架构的模型是否能够应用于视觉下游任务(如目标检测)开始成为一些研究人员们的研究重点.

1)AS-MLP(axial shifted MLP). 受到 CNN 的启发, Lian 等人[115] 认为赋予 MLP 架构局部感受野能使得网络可以有更好的特征提取能力和良好的下游任务迁移能力, 提出了第 1 个应用于下游任务(例如, 对象检测和语义分割)的 MLP 架构 AS-MLP, 在水平和垂直方向上在空间上移动特征, 这种方法不仅聚合了不同位置的特征, 而且随后使用通道混合 MLP 结合了这些特征, 使模型能够捕获局部依赖关系. 还可以仿照卷积核的思想为 AS-MLP 模块设计感受野的大小乃至空洞膨胀等. AS-MLP 在 ImageNet1K 数据集中以 88×10^6 的参数量和 15.2 GFLOP 获得了 83.3% 的 Top-1 准确率. 并且由于轴向位移的设计, AS-MLP 架构可以转移到下游任务(例如对象检测).

2)CycleMLP. 针对 MLP-Mixer 的限制, Chen 等人[116] 提出了 CycleMLP, 引入了一个新的循环全连接(cycle fully-connected, Cycle FC)层沿通道维度循环采样进行上下文聚合, 允许处理可变的输入尺度, 能够处理各种大小的输入图像, 扩大了感受野. 同时计算复杂度与图像大小成线性, 与通道全连接层相同, 比原始的空间全连接层低. 由于能够处理可变尺度, CycleMLP 也能够用于目标检测等视觉下游任务, 其在 MS COCO val 2017 上能够以相对更少的参数量取得最高平均精确率为 44.1%.

3)Hire-MLP. Guo 等人[117] 同样注意到了 MLP-Mixer 的问题, 提出了 Hire-MLP, 引入了分层重新排列(hierarchical rearrangement). Hire-MLP 保留了通道混合 MLP 部分, 但是将 Token Mixing MLP 替换成了层次模块(hire-module block). 具体来说, 每个层次模块包括 1 个宽度方向的重排和 1 个高度方向的重排以及 1 个通道方向的全连接映射. 每个方向上的重排包括 2 类: 内部区域恢复(inner-region restore)和交叉区域恢复(cross-region restore). 通过内部区域重排捕获空间区域内的局部信息, 通过跨区域重排实现不同区域之间的信息通信, 并且通过沿空间方向循环移动所有 tokens 来捕获全局上下文. 这些重排列操作不仅交换区域之间的信息, 而且保留了相对位置, 因此 Hire-MLP 在各种视觉任务中取得了显著的性能改进.

4)Wave-MLP. Tang 等人[118] 受量子力学的启发提出了 Wave-MLP, 将 MLP 架构中的 patch 描述成一个具有振幅和相位 2 部分的波函数, 令波的振幅是原始特征, 相位是根据输入图像的语义内容变化的复数值, 通过相位的影响调节 token 和权重之间的关系来解决 MLP-Mixer 中只能通过固定权重的全连接层来融合不同 tokens, 而被迫忽视来自不同图像的不同语义信息的问题. 通过把 tokens 描述成波的表示, 他们构建了 Wave-MLP 的架构, 在 ImageNet 上能够以 63×10^6 的参数量和 10.2 GFLOP 实现最高 Top-1 准确率 83.6%, 在 MS COCO val 2017 上的平均精确率达到 45.7%.

5)RaMLP. Lai 等人[119] 考虑到 CycleMLP 和 Hire-MLP 等架构的缺陷, 为了解决远程依赖和视觉线索被忽略的情况, 设计了一种区域感知混合(region-aware mixing)方法, 即 RaMLP 架构. 并且引入了可学习池化(learnable pooling, LP)和扩张全连接(dilated fully-connection, DFC), 实现根据空间特征自适应地确定聚合权重, 可以更鲁棒地捕获空间视觉线索, 实现更鲁棒的空间特征提取. 具体而言, DFC 通过将输入的特征图进行扩张划分, 并使用空间全连接层得到全局增强特征, 再通过倒置扩张重整(inverted dilated reshaped)调整回原来大小, 实现了自适应确定权重, 从而更好地提取区域级别的特征. 最终, RaMLP 得到了截止目前最强的视觉 MLP 架构表现, 以 58×10^6 的参数量和 12.0 GFLOP 在 ImageNet 上达到 84.1% 的 Top-1 准确率, 在 MS COCO val 2017 上的平均精确率达到 46.4%.

2.4.2 小 结

在表 7 中我们简要总结了基于 MLP 的目标检测算法的优缺点, 以及在表 8 对比了这些算法在数据集上 COCO 2017 val 的实验结果. 我们通过图 4 展示了不同模型在原有 MLP-Mixer 模型上的改进.

MLP 的成功证明了, 即使网络参数几乎全由全

Table 7 Advantages and Disadvantages of the Target Detection Model Based on MLP

表 7 基于 MLP 的目标检测模型优缺点

模型	出版物	优点	缺点
MLP-Mixer	文献 [108]	在计算机视觉领域重新提出纯 MLP 结构	固定输入大小, 不易有效转移到下游任务
AS-MLP	文献 [115]	轴向移动特征为模型提供局部感受野	特征提取能力不如 SOTA 方法
CycleMLP	文献 [116]	使用循环采样进行上下文聚合	结构复杂, 可能丢失重要信息
Hire-MLP	文献 [117]	通过区域重排捕获更多局部信息	重排操作可能破坏位置先验信息
Wave-MLP	文献 [118]	赋予神经网络复数操作的优势特性	实现相对复杂
RaMLP	文献 [119]	通过区域感知混合操作增强局部感受野	混合然后恢复可能破坏原始信息

Table 8 Experimental Results of the Target Detection Model Based on MLP on Public Datasets

表 8 基于 MLP 的目标检测模型在公开数据集上的实验结果

模型	*AP*/%	AP_{50}/%	参数量	浮点运算量/GFLOP
AS-MLP	51.5	70.0	145.0×10^6	961.0
Cycle-MLP	42.7	63.3	85.9×10^6	
HireMLP	44.9		105.8×10^6	424.5
Wave-MLP	44.2	65.1	66.1×10^6	333.9
RaMLP	46.4	67.7	70.0×10^6	

连接层进行学习，也可以通过优秀的设计获得与最先进的基于其他架构（如 CNN, Transformer 等）一样好的实验结果，这便会引出更多的问题，比如各种复杂的网络设计是否仍有必要. 这些人工引入的归纳偏置是否真的产生了他们预期的效果. 为什么相对老的 MLP 模型仍然能够超越精心设计的 CNN 和 Transformer. MLP 对特征的学习与 CNN 和 Transformer 架构学习到的特征（如果有）的差异究竟在哪里. 在 MLP 网络中，是否隐藏了人们从未发现的归纳偏置，并且在泛化中起到了重要的作用. 当 MLP 在视觉领域中取得如此成绩的时候，是否也能同样对自然语言处理领域产生影响. MLP 的发展是否也能促进对 CNN 和 Transformer 的本质研究同样也是值得关注的方向.

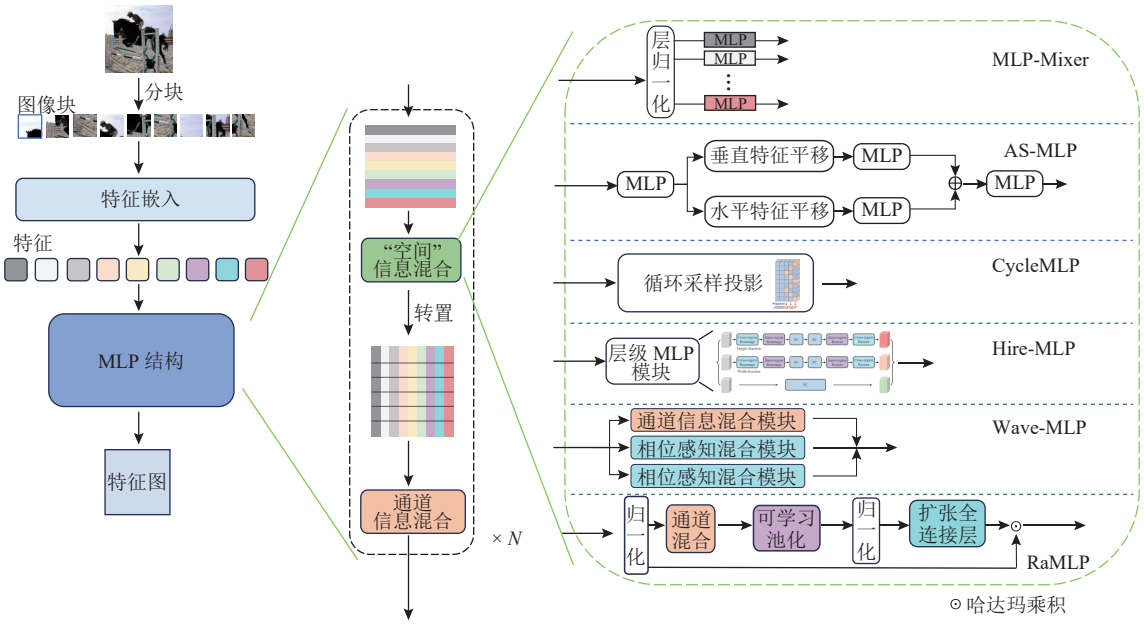


Fig. 4 Overview of MLP-based object detection models

图 4 基于 MLP 的目标检测模型的概述

2.5 基于扩散模型的方法

研究人员们认为，机器学习的核心问题是用一个高度灵活的概率分布族来建模一个复杂的数据集. 扩散模型的概念最早在 2015 年由 Sohl-Dickstein 等人[120]提出，通过迭代的正向扩散过程系统地、缓慢地破坏数据分布中的结构. 然后，学习一个反向扩散过程来恢复数据中的结构. 2020 年 Ho 等人[121]提出了去噪概率扩散模型（denoising diffusion probabilistic model, DDPM），实现了高质量图像的生成. 具体而言，通过对给定图像进行迭代地添加随机采样的高斯噪声，最终将图片还原为一个各向同性的高斯噪声，再训练一个去噪模型（通常是 U-Net）来进行反向去噪步骤，还原出初始图像. 大量研究证明了扩散模型可以成功用于各种生成任务中，并且取得了很好的结果.

1）DiffusionDet. 2023 年 Luo 等人[122]提出了 DiffusionDet 框架，将目标检测定义为图像中边界框的位置（中心坐标）和大小（宽度和高度）空间上的生成任务，定义从各种随机的噪声边界框到目标框的去噪扩散过程，第 1 次将扩散模型运用到了目标检测领域. 在 DiffusionDet 的训练阶段，向 ground truth 边界框逐渐添加噪声来生成噪声框. 使用一个预训练好的骨干网络来获取输入图像的特征图. 将得到的噪声框视为感兴趣区域投影到特征图上进行裁剪来获取对应特征，这些特征使用一个解码器网络来反向预测对应的 ground truth 边界框和对应的类别. 在推理阶段，DiffusionDet 使用去噪扩散隐式模型（denois-

ing diffusion implicit models, DDIM)[123]来从噪声框中恢复得到检测框,将噪声分布逐步调整为边界框的可学习分布. DiffusionDet在MS COCO 2017上的最高平均精确率达到52.5%,但基于扩散模型的目标检测算法参数量庞大,需要大量的计算.

2)Diff3DETR. Diff3DETR[124]是一种基于代理的扩散模型,用于半监督3维目标检测. 该模型通过结合代理对象查询生成器和框感知去噪模块,在DETR框架内实现了对动态场景的有效适应和边界框的精确细化. 代理对象查询生成器负责生成能够平衡采样位置和内容嵌入的对象查询,而框感知去噪模块则利用去噪扩散隐式模型(DDIM)的去噪过程和变换器解码器中的长距离注意力机制,逐步优化预测边界框. Diff3DETR为减少对大量标注数据的依赖提供了有效的解决方案,并在3维目标检测任务上超越了现有的最先进方法,展示了其在生成高质量伪标签和提高检测精度方面的显著优势.

3)MonoDiff. MonoDiff[125]是一个单目3维目标检测框架,其应用扩散模型来提升从单视图图像中检测3维对象的能力. 该框架采用逆扩散过程对3维边界框进行精确估计,并通过高斯混合模型对前向扩散过程中的噪声进行采样和初始化,解决了传统方法在不同维度上边界框尺寸变化的不确定性问题. MonoDiff还结合2维检测信息,通过3维/2维投影的一致性提供额外的监督信号,增强了模型对3维空间中对象的检测能力.

2.6 大模型时代

1)SAM. 随着人类技术水平的发展,在有限类别上的视觉感知已经逐渐不能满足人们的需要. 按照已有的技术路线,仅通过使用更大量的丰富标注数据来实现面向真实开放世界的视觉识别任务的工作量是不可接受的. 研究人员们开始探索其他方法. Hu等人[126]在2018年做出了第1次探索,用小部分掩码注释和大量的目标框注释来训练实力分割模型. 近年来,大模型在自然语言处理上实现了前所未有的突破. 自然而然地,人们开始探索大规模预训练方法在视觉上是否仍然可以表现出极强的泛化能力. Kirillov等人[127]提出了第1个视觉大模型SAM(segment anything model),通过构造迄今为止最大的分割数据集,即1 100万个图像和其上面共计10亿的图像掩码来对模型基于提示词的模式进行设计和训练,使得模型可以以零样本的方式迁移到新的图像分布和任务. 通过在大量任务上评估SAM模型的能力,发现在zero-shot的情况下仍然能表现出超过或具有与监督学习模型同样性能的能力. SAM的提出证明了在计算机视觉这种高分辨率和高变化性的任务上大模型仍然有极强的潜力. 尽管SAM是一个面向分割任务的模型,但是目前已经有不少的工作对SAM的迁移效果进行探索,试图发掘其在其他任务上的卓越能力,如目标检测[128]、伪装目标检测[129]、3维目标检测[130]、显著目标检测[131]、医学图像检测[132]等.

2)GLEE. GLEE[133]是一个面向大规模图像和视频处理的对象级基础模型. 它通过集成化的框架,实现了对象的检测、分割、跟踪、定位和识别等多维度感知任务. GLEE模型采用多模态输入处理方式,结合图像编码器、文本编码器和视觉提示器,以统一的学习策略从不同数据源中学习,形成具有泛化能力的对象表示. GLEE在零样本学习方面表现出色,经过在超过500万张图像上的广泛训练,展现了卓越的多功能性和泛化能力,能够无需特定任务的微调即可适应新数据和任务. 此外,GLEE的设计允许其扩展训练数据规模,通过整合大量自动标注的数据,进一步提升了模型的泛化性能. GLEE模型的架构和训练方法使其成为一个强大的通用视觉模型,适用于广泛的下游任务,为构建高效的视觉基础模型和推动人工通用智能系统的发展做出了重要贡献.

3)Griffon. Griffon[134]是一种基于大型视觉语言模型(large vision-language model, LVLM)的先进目标检测框架,它通过2个阶段的训练流程强化了对图像和视频中所有对象的细粒度识别和定位能力. 在第1阶段的基础场景预训练中,Griffon利用大量预训练数据来构建一个能够准确识别图像中所有对象的基础模型. 第2阶段的全场景指令调整进一步细化了模型对用户意图的理解. Griffon的设计理念包括一个统一的输入输出表示,允许模型接受多种形式的自由文本输入,并以统一格式输出对象的类别和坐标. 它不依赖于特定的标记、专家模型或附加的检测模块,而是直接利用LVLMs的内在能力进行对象的细粒度感知和空间定位. 此外,Griffon引入了一个无需额外训练的置信度评分机制,增强了模型对预测的置信度,提升了检测质量. 在MS COCO数据集上,Griffon展现出了卓越的性能,接近或达到了专家模型Faster R-CNN的水平,证明了其在目标检测任务上的有效性和鲁棒性.

在当今人工智能领域,大型视觉模型在目标检测任务上的应用已成为研究的热点,这些模型通过其庞大的参数量和深度学习能力,正在不断推动目标检测技术的发展和创新. 这些模型通过统一的框

架处理多模态输入，能够有效地从图像和视频中检测和识别各种对象. 它们利用大规模预训练数据集来学习丰富的视觉和语言特征，从而在零样本或少样本的设置下展现出强大的泛化能力. 这些模型不仅在传统的对象检测任务上取得了显著的成果，也在更具挑战性的开放词汇量检测、视频对象分割和交互式分割等任务上取得了突破. 未来的研究将继续推动这一领域的边界，实现更加智能、高效和可靠的计算机视觉系统.

3 总结与展望

随着人工智能领域的发展，涌现了不同的基础模型，这些模型分别有其不同的特点. 尽管当前目标检测领域的基础模型已有多种，但是研究人员们更多的是基于 CNN 和 Transformer 来进行研究，并且在各种数据集上的评估指标仍然是基于这两者模型的能够取得更好的结果. 但是基于 CNN 的模型和基于 Transformer 的模型的结构和特点是有显著差异的，这也导致了它们的具体性能表现有差异. CNN 的主要优势是其局部连通性和权重共享，基于 CNN 的目标检测模型通常有分层架构，这也使得模型能够从基本语义到高级语义特征进行渐进的特征提取，卷积结构精细的局部感受野赋予了 CNN 更加优秀的对于相对较小的物体的检测能力. Transformer 结构则完全基于注意力机制，更强调图像特征之间的全局依赖关系，例如 DETR 模型利用 Transformer 的编码器-解码器架构来分析整个图像特征，使用全局查询向量直接预测目标的类别和位置. 通过自注意力机制，Transformer 可以捕获图像中不同位置之间的长程依赖关系，这种全局注意力意味着网络在处理每个输入元素时考虑来自所有其他元素的信息，能够增强模型提取全局上下文信息的能力. 全局注意力机制可以为网络提供更好的全局建模能力，从而提高网络在检测较大目标时的性能. 然而，缺乏精确的局部感受野（通常在 CNN 中实现）使得网络在检测小目标方面效果不佳，此外与 Transformer 相比，卷积结构的稀疏连接能够为模型提供了更好的泛化性能[135]. 因此，如何更好地融合 2 种大小的感受野以实现检测精度的全面提升仍然是目标检测领域的研究重点，一个有效的方法是扩大卷积核，为卷积网络提供类似 Transformer 的全局建模能力[46-48]. 基于 MLP 的 MLP-Mixer 结构包含了 CNN 和 Transformer 两者的特点，它同样采取了 Transformer 的对输入图像进行分块并提取初步特征的结构，并且类似于 CNN，在空间层面和通道层面都进行特征提取和特征融合. 但是因为基于 MLP 的模型全都由全连接层的矩阵乘法构成，导致其需要更大的参数量且需要更多设计才能处理可变输入尺寸. 扩散模型作为一种新的生成式算法，在图像生成领域已经取得显著成功，因此研究人员们开始探索其在其他领域的潜力. 基于扩散的目标检测模型现在仍处于初步研究阶段，其参数量庞大、运行速度慢等缺点还有待解决.

3.1 实验对比

为了针对不同的模型提供更加细致的对比和分析，我们在图 5 中展示了不同模型的检测结果可视化. 并且如表 9 所示，我们选取了部分目标检测模型在 MS COCO 2017 test 数据集上进行测试，并记录了对应的实验结果指标，实验结果来自目标检测开源工具箱 MMDetection. 首先我们按照模型训练时所需显存大小（4 GB/8 GB/10 GB）来进行分类，选择有代表性的算法来进行比较，探究在相近的显存占用率下哪些模型具有更好的表现. 可以看出，CenterNet 作为一个简单的无锚框算法，只会受到使用目标中心点来回归检测结果的影响，其对于小目标（小于 32×32 像素点大小）的检测性能显著更差（mAP_{small} 仅为 9.1%）；作为 1 阶段目标检测器的代表 YOLOv3，YOLOX 等对于小目标的检测性能也比 2 阶段检测器差（mAP_{small} =14.4%，mAP_{small} =12.4%）；Cascade R-CNN 作为集联检测器，在开销相近的情况下检测性能会优于其他检测器. 而 DETR 作为 Transformer 架构在目标检测任务上的代表模型，由于其强大的全局注意力建模能力，导致 DETR 对与大目标的检测性能显著优于同等开销下的其他 1 阶段检测模型（mAP_{large} =59.4%）.

此外，如表 10 所示，我们还按照相近的 mAP 指标来分类，探究不同模型在面对不同尺寸物体时的检测性能. 在均使用 ResNet-50 为特征提取网络的情况下，由于 YOLOF 使用最高层的特征图，所以获得了相对更强大的大目标检测性能（mAP_{large} =53.2%），与强调全局稀疏查询的 Sparse R-CNN 性能相近. 同时，DAB-DETR 和 Deformable DETR 的实验结果同样证明了在使用更简单的骨干网络（ResNet-50）下，因为全局注意力的影响，大目标检测性能比使用更复杂的骨干网络的基于 CNN 的目标检测器更高.

3.2 未来发展方向与挑战

从 20 世纪 90 年代到 21 世纪初期，自基于手工特征和传统机器学习的目标检测算法到如今各种基

Fig. 5 Visualization of detection results for different object detection models

图 5 不同目标检测模型的检测结果可视化

Table 9 Comprehensive Experimental Comparison on COCO 2017 test Dataset (Similar Training Costs)

表 9 COCO 2017 test 数据集综合实验对比（训练开销相近）

模型	训练显存/GB	骨干网络	*mAP*/%	mAP_{50}/%	mAP_{75}/%	mAP_{small}/%	mAP_{medium}/%	mAP_{large}/%
CenterNet	3.45	ResNet-18	25.9	42.6	27.1	9.1	30.1	40.3
YOLOX-tiny	3.5		31.8	49.1	33.8	12.4	34.9	47.3
FCOS	3.6	ResNet-50	36.6	56.0	38.8	21.1	40.7	47.1
Faster R-CNN	3.8	ResNet-50	37.8	58.6	41.0	21.6	41.5	49.3
YOLO v3	3.8	DarkNet-53	30.8	52.8	32.0	14.4	33.4	44.7
RetinaNet	3.8	ResNet-50	36.5	55.4	39.1	20.4	40.3	48.1
Cascade R-CNN	4.2	ResNet-50	40.4	58.9	44.1	22.8	43.7	54.0
YOLO v3	7.4	DarkNet-54	33.7	56.6	35.3	19.4	36.8	44.3
Cascade R-CNN	7.6	ResNeXt-101	43.7	62.3	47.7	25.1	47.6	57.3
YOLOX-S	7.6		40.3	59.1	43.4	23.5	44.5	59.4
DETR	7.9	ResNet-50	39.9	60.4	41.7	17.6	43.5	59.4
YOLOF	8.3	ResNet-50	37.5	57.0	40.4	19.0	44.2	53.2
CornerNet	9.5	Hourglass-104	39.6	55.0	42.2	20.7	41.8	54.4
RetinaNet	10.0	ResNeXt-101	40.8	60.5	43.7	22.9	44.4	54.6
FCOS	10.0	ResNeXt-101	42.7	62.5	45.7	26.0	46.5	54.7
Faster R-CNN	10.3	ResNeXt-101	42.1	63.0	46.3	24.8	46.2	55.2
Cascade R-CNN	10.7	ResNeXt-101	44.7	63.6	48.9	26.1	48.6	58.6
CornerNet	13.9	Hourglass-104	40.4	55.9	43.2	20.2	42.7	58.4
RetinaNet	14.5	PVT-s	40.4	61.3	43.1	24.8	43.2	54.8
Deformable DETR		ResNet-50	44.3	63.2	48.6	26.8	47.7	58.8
Conditional DETR		ResNet	41.0	61.9	43.5	20.4	44.5	59.9
DAB-DETR		ResNet-50	42.3	62.9	45.2	21.6	46.1	61.3
DiffusionDet		ResNet-50	45.4	65.1	48.7	28.3	47.9	61.5
ViTDet		ViT-B	51.5	72.1	56.6	35.3	55.5	66.3
Co-DETR	19.2	ResNet-50	52.1	69.4	57.1	35.4	55.4	65.8
Co-DETR		Swin-L	58.9	76.9	64.8	42.5	62.7	75.1

Table 10 Comprehensive Experimental Comparison on COCO 2017 test Dataset (Similar Overall Performance)

表 10 COCO 2017 test 数据集综合实验对比（总体性能相近） %

模型	骨干网络	*mAP*	mAP_{50}	mAP_{75}	mAP_{small}	mAP_{medium}	mAP_{large}
RetinaNet	ResNet-50	36.5	55.4	39.1	20.4	40.3	48.1
FOCS	ResNet-50	36.6	56.0	38.8	21.1	40.7	47.1
YOLOF	ResNet-50	37.5	57.0	40.4	19.0	44.2	53.2
Faster R-CNN	ResNet-50	37.8	58.6	41.0	21.6	41.5	49.3
Sparse R-CNN	ResNet-50	37.9	56.0	40.5	20.7	40.0	53.5
CycleMLP	CycleMLP-B1	38.6	59.1	40.8	21.9	41.8	50.7
DETR	ResNet-50	39.9	60.4	41.7	17.6	43.5	59.4
DAB-DETR	ResNet-50	42.3	62.9	45.2	21.6	46.1	61.3
Cascade R-CNN	ResNet-101	42.4	60.9	46.1	23.8	46.2	56.4
FCOS	ResNeXt-101	42.7	62.5	45.7	26.0	46.5	54.7
Cascade R-CNN	ResNeXt-101	43.7	62.3	47.7	25.1	47.6	57.3
Wave-MLP	Wave-MLP-B	44.2	65.1	47.1	27.1	47.8	58.9
Deformable DETR	ResNet-50	44.3	63.2	48.6	26.8	47.7	58.8
DiffusionDet	ResNet-50	45.4	65.1	48.7	28.3	47.9	61.5

于深度学习方法的算法,目标检测领域经历了从传统方法到深度学习的革命性转变.到如今,最新的深度学习模型架构以及为不同领域的检测特殊设计的针对性方法已经能够在一定程度上满足人们对目标检测算法性能的需求.然而如3.1节中实验数据展示,现有的目标检测算法仍不能很好解决所有检测问题,因此,我们总结了一些目标检测领域内面临的挑战和未来的发展方向;还回顾在30年中目标检测算法的发展历程;总结了目标检测受到的主要挑战,列举了目标检测领域未来的研究方向,并且通过统计DBLP数据库中相应子研究领域的论文数量进行佐证.

1)多样性.多样性是目标检测领域中的一个重要难题,尤其随着数据集的不断扩展和应用场景的多样化,早期传统基于机器学习的目标检测任务由于计算资源和数据限制,主要集中在人脸检测[3-4]和行人检测[6]等单一场景.随着技术的发展,数据集的规模和类别迅速增加,PASCAL VOC包含了20类目标,MS COCO[2]扩展至80类,最新的数据集如LVIS[136],更是涵盖了超过1 000个类别.此外,不同领域的数据集也相继涌现,如红外图像检测数据集CTIR[137]、工业缺陷检测数据集[138]等.这些数据集不仅在类别数量上有显著差异,且在数据分布、图像分辨率、目标尺寸等方面也存在巨大差异.模型的泛化能力受到严峻挑战,原有模型如YOLO, Faster R-CNN在面对这些多样化的数据时表现并不理想,难以实现跨领域、跨场景的良好迁移.为解决目标种类的多样性挑战,Li等人[139]提出了平衡组Softmax(balanced group Softmax, BAGS)模块,用以改进Faster R-CNN模型,调节模型训练过程,克服类别不平衡问题使得检测性能提高.热红外图像的分辨率低,且往往含有更多的噪声,为了克服普通目标检测算法在红外数据上检测准确率低的问题,郭伟等人[140]在YOLOv7模型中引入了坐标注意力机制,通过强化模型对坐标位置的感知来增强特征表达,最终使得模型在红外图像数据集上有更优秀的表现.印刷电路板(PCB)本身种类繁多设计复杂、缺陷多样,因此普通的目标检测效果很差.尹嘉超等人[141]使用通道注意力机制对特征融合网络FPN进行优化,提高了模型的细节信息提取能力.这些研究表明,面对目标检测中的多样性挑战,未来的发展趋势将集中在模型的适应性和对不同数据模式的适应性上.

2)准确性和实时性.提高准确性作为目标检测的首要目标,一直是研究人员们探索的核心主题.不仅涉及算法的设计和优化[18-19,37,57],还包括数据质量的提升[1-2,136,142]、模型泛化能力的增强[143-145]以及对复杂场景的适应性改进[146].通过更精准的特征提取[11-12,44,87,115]和更强大的学习算法[28,147]在更加丰富的数据上进行训练,使得目标检测的准确度在特定域上有极大的提升.实时性在目标检测算法中是一个非常重要的方面,尤其是在那些对响应速度有严格要求的应用场景中,如自动驾驶、视频监控、机器人导航等.如端到端的1阶段检测器YOLO系列[19-20,22-23,40,42],通过使用1次卷积来直接生成检测结果、使用更轻量级的模型骨干架构,以及优化边界框的生成过程来提高模型的实时性.然而虽然深度学习技术极大提升了目标检测的准确性,但是计算量大、响应时间长的问题仍然存在,人们很难同时取得高准确性和高实时性[19,49,57,60].在需要实时处理的应用场景中,如何平衡速度与准确性仍是一大挑战.

3)小目标检测.小目标检测是计算机视觉领域的一个重要子领域,专注于从图像或视频中识别和定位尺寸较小的目标,特别是在遥感图像分析、医学影像诊断以及工业检测和军事领域等[148-153].该方向的发展如图6所示.小目标是指在图像中占据较少像别的特征较少、细节不足且小、易于在复杂的背景中被遮挡或混淆等特点,给检测算法带来了特殊的挑战.一些可能的改善方法有为网络增强特征提取能力[154]设计更有针对性的损失函数[155]来强化模型对于细节的学习,以及基于超分辨率重建方法来增强小目标的可分辨性[156]等.

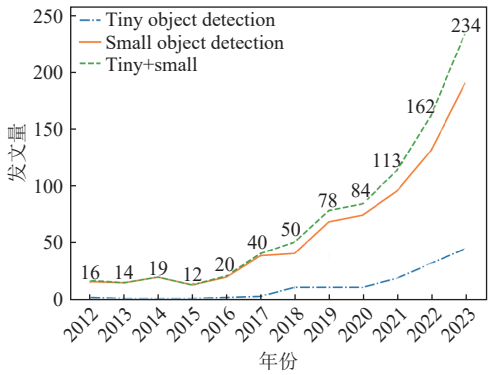


Fig. 6 Small object detection trends

图6 小目标检测趋势

4)弱监督、无监督目标检测.弱监督目标检测指的是在有限或不完整标注信息的情况下进行目标检测[157-161].例如,训练数据可能只有图像级的标签(图中是否存在某个目标)而没有精确的目标位置(边界框)信息,从而可以极大地缓解对图像进行精确注释的手工成本.其发展如图7和图8所示.但由于缺乏

精确的边界框信息，模型难以精准定位目标，且容易造成目标与背景的混淆. 更进一步的改进方法有提高伪标签的质量[162]和增强目标检测模型对于已有图像分类模型的知识的学习能力[159].

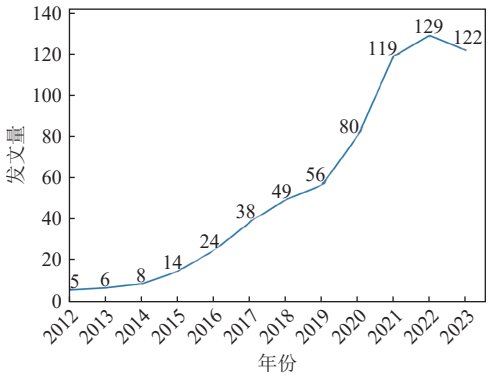


Fig. 7 Weakly-supervised object detection trend

图 7 弱监督目标检测趋势

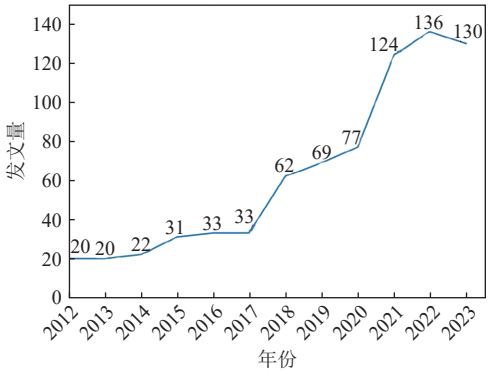


Fig. 8 Unsupervised object detection trend

图 8 无监督目标检测趋势

无监督目标检测[163-165]旨在没有任何标注信息的情况下进行目标检测，即模型需要自主学习识别并定位图像中的目标，在没有明确地标注图像中包含哪些类别的目标的情况下，模型图像中前景和背景的特征学习更难，并且难以精确定位目标. 一个潜在的解决方案是引入对比学习[165].

5)域适应、域泛化目标检测. 域适应是指在源域上训练模型，并对其进行调整使其能在一个特定的目标域上有效工作. 目标域的数据在训练过程中是可用的，但是往往没有标签数据. 域适应目标检测[143,166-172]通常发生在训练数据(源域)和实际应用数据(目标域)之间存在显著差异的情况下. 这 2 方面的发展如图 9 所示. 目标域往往缺乏足够的标注数据进行有效训练，且源域和目标域之间的差异会导致模型直接迁移的性能下降. 并且变化多样的环境和条件使得域适应变得更加复杂. 如何让模型在多个不同的目标域中都表现良好是域适应目标检测的一大挑战. 可能的解决方法有结合知识蒸馏和对比学习来增强目标检测模型的域迁移检测能力[171].

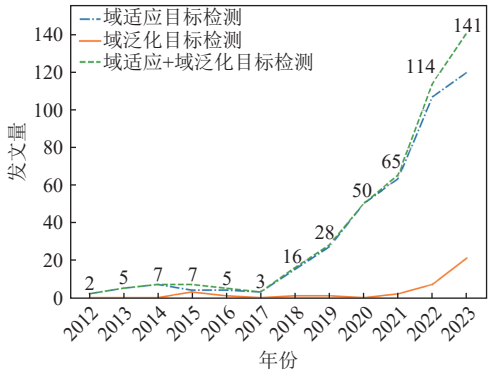


Fig. 9 Domain adaptive and generalized object detection trends

图 9 域适应与域泛化目标检测趋势

与域适应不同的是，在域泛化问题中，需要模型仅在源域上进行训练，也就意味着在训练期间，模型不可以获取任何非源域(即泛化的目标域)的数据. 因此，域泛化目标检测[173-175]需要提高模型对于未知域的泛化能力，并且要求人们开发更鲁棒的模型，使得它们能够在面对新的目标域时能够保持稳定的性能. 一个未来改进的方法是设计更有效的对齐方式来弥合不同域之间的差异[176].

6)少样本目标检测. 少样本目标检测也称 Few-shot，旨在强调训练目标检测模型时仅使用非常有限的注释样本(比如每一类仅有几个被完全标注的实例)仍然能有效地识别和定位新的目标，因此要求模型在极少数据的支持下也能够快速适应[144,177-181]. 其发展如图 10 所示. 少样本任务中极少的训练样本使得模型学习有效的特征表示变得非常困难，并且很容易出现在少量数据上的过拟合. 在很多实际应用场景中，获得大量标注数据时昂贵且需要极长时间，并且在一些特定专业或稀有对象的场景中，样本数

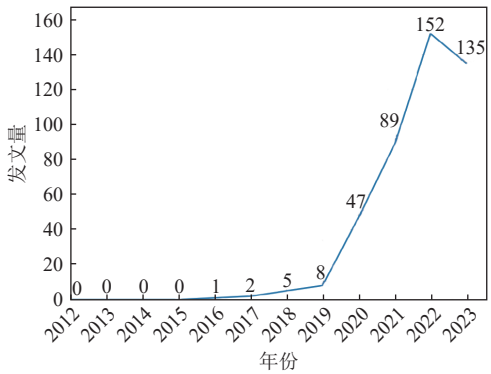


Fig. 10 Few-shot object detection trend

图 10 少样本目标检测趋势

据本身就很少，所以少样本学习尤为重要. 面对这些问题，潜在的解决方案有：提出更有效的方法提取已有的少量样本的特征并保存，并将它们与新的无标签的样本进行特征比对来进行学习[182].

7)3 维目标检测. 3 维目标检测是指在 3 维空间中识别和定位对象的技术[183-186]. 与传统的 2 维目标检测(仅在图像平面上识别和定位对象)不同，3 维目标检测考虑了对象的深度信息，提供了更全面的场景理解. 通常涉及如激光雷达、立体摄像头或深度传感器等技术来捕获多模态 3 维数据的使用. 在自动驾驶和机器人技术中 3 维目标检测可以提供更准确的环境感知，而在增强现实和虚拟现实中 3 维目标检测提供了与真实世界交互的必要信息. 其发展如图 11 所示. 3 维数据通常以点云的形式存在且一般来自多种昂贵且复杂的传感器，其处理方式比 2 维图像复杂得多. 融合更多传感器提供的数据能够有效地对 3 维目标进行感知和建模，但是需要更昂贵的计算成本和标注成本. 因此研究基于纯视觉的 3 维目标检测[187](即输入模态只有 2 维图像)可能是未来的研究重点.

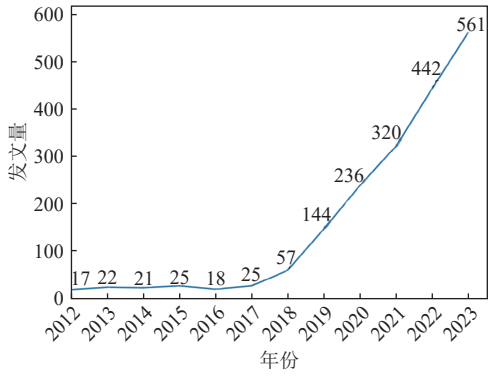


Fig. 11 3D object detection trend

图 11 3 维目标检测趋势

8)增量学习. 增量学习也称为连续学习或终生学习，是机器学习中的一个概念，指的是模型在学习新任务或新数据时，能够保留以前学到的知识，并且具有利用过去的知识来帮助学习新任务的能力. 增量学习目标检测[188-192]是指在已有的目标检测模型基础上，不断添加新类别的能力，同时保持对已学习类别的识别能力. 在实际应用中，经常会出现新的目标类别，增量学习使模型能够适应这种变化. 其发展如图 12 所示. 增量学习不需要每次都重新训练整个模型，节省了大量的时间和计算资源，并且增量学习赋予了模型持续学习和适应新任务的能力，如何克服灾难性遗忘以及实现真正的智能，需要对增量学习进行探索，符合人工智能的长期目标.

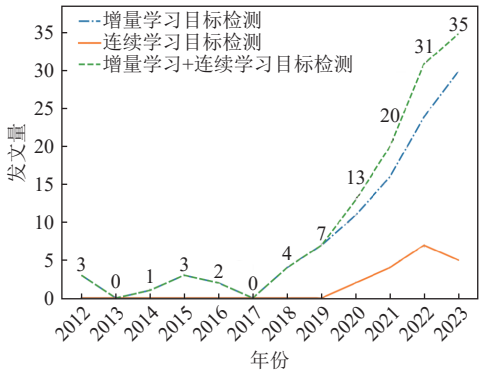


Fig. 12 Incremental learning object detection trend

图 12 增量学习目标检测趋势

9)开放世界和开放词汇目标检测. 现有的目标检测器主要依赖于大规模基准数据集进行训练，但在不同领域之间的性能差异巨大. 特别是在开放世界环境中，当遇到与已知类别差异很大的未知类别时，依赖基准数据集训练的检测模型往往无法有效识别. 这种无法识别未知目标的缺陷严重限制了目标检测器在实际场景中的泛化能力. 例如，在自动驾驶场景中，如果检测器无法识别未知物体，就可能对行车安全造成影响. 因此迫切需要开放世界目标检测算法，能够有效地处理未知类别和新领域的数据[193-199]. 为了解决这一任务，往往需要类似增量学习的方法，并且要求模型在潜在无限的类别空间中保持良好的性能，面对较少的新类别仍然能有良好的持续学习效果. 其发展趋势如图 13 所示.

开放词汇目标检测[200-206]指的是目标检测系统通常通过链接视觉数据与外部丰富的语义信息(如文本描述)能够识别训练集之外的类别. 基本与 zero-shot 目标检测的定义相同，即在训练过程中，对于某些类别不提供任何训练样本，而是通过引入辅助信息来解决未知类别的无标记样本问题. 每个未知类

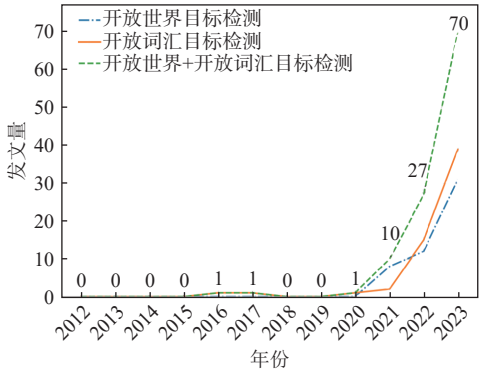


Fig. 13 Open-world and open-vocabulary object detection trend

图 13 开放世界和开放词汇目标检测趋势

别都需要有与其特征相关的辅助信息. 引入辅助信息的方法借鉴了人类的认知过程, 人类可以在语义背景知识的帮助下完成零样本学习. 常见的辅助信息是未知类的语义信息, 比如对某个类别的文字描述. 显然, 现实世界中的对象类别远远超过任何一个数据集, 通过开放词汇目标检测, 能够使得目标检测系统适应更广泛的类别, 更好地应对现实世界的多样性. Zero-shot 目标检测发展趋势如图 14 所示.

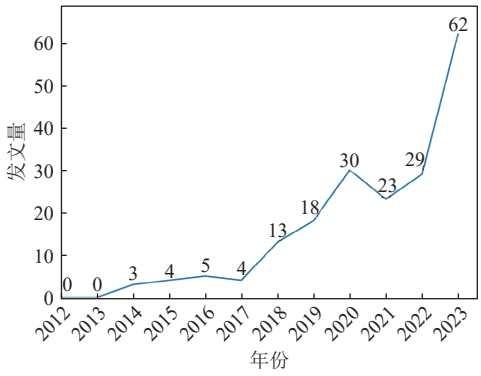


Fig. 14 zero-shot object detection trend

图 14 zero-shot 目标检测趋势

10)通用人工智能(artificial general intelligence, AGI). AGI 通常被认为是人工智能领域研究和发展的终极目标. AGI 是指一种具备广泛认知能力的人工智能, 能够像人类一样在各种环境和任务中理解、学习和应用知识. 与当前的狭义人工智能(narrow AI), 也称弱人工智能(weak AI)不同, AGI 不仅能在特定任务上表现出色, 还能进行跨领域的学习和决策, 具有自主意识和自适应能力. AGI 能够处理更复杂、多变的任务和环境, 而不仅限于特定的窄领域问题, 具有自我学习和根据新信息作出决策的能力, 这在不断变化的现实世界中极为重要, 并且 AGI 可以在各种领域进行创新和解决问题, 类似于人类的智能, 因此 AGI 被视为人工智能发展的终极目标, 它代表了智能机器的最高形态. 当 AGI 真正实现, 即可简单地解决当前目标检测领域中面临的绝大部分问题, 比如跨域适应能力、持续性学习能力以及对更复杂环境的高度理解能力等. 尽管目前 AGI 还是一个理论概念, 但其对于多种任务的潜在影响是巨大的, 需要人们持续进行探索. 受到 GPT 等自然语言处理领域的大模型的启发, 大型视觉模型也成为人们的探索重点[127,207]. 这些大规模、多功能的视觉模型在处理复杂的视觉任务时展现出了显著的能力. 然而, 这些进步仅是通向 AGI 这一终极目标的初步探索.

目标检测领域的发展已远超过几十年前的初步阶段, 不再局限于在 2 维图像中定位训练数据中已标注类别的简单任务, 而是逐步从基本的图像识别任务向更高层次的认知功能演进, 特别是在近年来, 随着深度学习和计算能力的飞速发展, 目标检测任务不再是简单地定位已有标签类别的物体, 而是融入更为复杂和动态的应用场景中, 这些场景要求模型不仅能够识别和定位对象, 还能理解这些对象在更广阔环境中的上下文和相互关系. 当前, 这一领域正面临着多维度的挑战和转变, 涵盖了更广泛的应用场景、更复杂的空间处理需求以及更高级的检测技术. 目前, 基于目标检测的自动驾驶感知系统已经大量应用于辅助驾驶和自动驾驶车辆. 然而在这种使用环境中, 由于车辆的高速行驶以及道路环境的快速变化性, 要求这些目标检测系统能快速且准确地识别出道路上可能发生的各种情况. 然而, 2024 年 7 月 31 日《华尔街日报》列举了 222 起特斯拉自动驾驶事故, 其中有 31 起事故是由于自动驾驶系统未能识别出障碍物做出反应而导致的, 这些由于检测失败而造成的事故导致了重大的安全危害和财产损失. 如图 15 所示, 在夜晚, 自动驾驶车辆的目标检测系统无法正确检测到发生在路上的车祸, 将车祸现场识别成了正常的车辆, 导致事故的发生. 在光照条件不足、环境变化快速的条件下完成对各种类别目标的检测是当前自动驾驶领域的一个重要挑战. 此外基于目标检测的系统还广泛应用于医学领域. 乳腺超声人工智能模型 DeepBC 软件[208] 已经有上万次使用, 通过手机拍摄彩超报告并上传, 即可在线实时获得智能诊断结果, 辅助临床诊疗. 虽然取得了阶段性的成果, 但是医学目标检测领域也面临着大量挑战, 如: 医疗数据设计涉及严格的隐私审查, 难以大量获

Fig. 15 Result diagram of the target detection system when the Tesla autonomous vehicle accident occurs

图 15 Tesla 自动驾驶车辆事故发生时目标检测系统结果图

取, 且数据集的制作和标注需要更高级别的专业知识, 难度更高. 此外, 由于直接涉及到人身健康安全, 医学领域的目标检测系统需要更精准的结果, 如果错误定位了人体内的病变区域则可能会造成严重的医疗事故.

11)跨领域适应性. 面对多样化和不断变化的实际应用环境, 如何使目标检测模型具有良好的跨领域适应性成为一个重要挑战. 这涉及到模型在不同光照、天气条件、场景背景下都能保持稳定和准确的检测能力.

12)实时处理与资源优化. 在需要快速反应的应用中, 如自动驾驶车辆或实时监控系统, 如何在保证高准确度的同时实现高效的实时处理, 尤其是在计算资源有限的设备上, 是一个技术难题.

13)与高级认知任务的结合. 目标检测正在逐步与更高级的认知任务结合, 如在场景理解、行为分析和人机交互中的应用. 这要求模型不仅要检测对象, 还要理解这些对象的功能、目的和它们之间的相互作用.

14)数据隐私和安全性. 随着目标检测技术在安全敏感领域的应用增多, 如何在提升性能的同时保障数据隐私和安全, 避免滥用技术, 成为了一个迫切需要解决的社会伦理问题.

15)小样本和零样本学习. 在许多实际应用场景中, 对于某些罕见或新出现的类别, 可能缺乏足够的训练样本. 因此, 如何使目标检测模型在小样本或零样本的情况下也能有效学习和适应, 是当前研究的一个热点.

这些挑战和转变不仅推动了目标检测技术的发展, 也为通向 AGI 铺平了道路. AGI 的实现将使目标检测技术不仅限于识别和定位, 而且进一步向更加智能化、自适应和多功能的方向发展, 最终实现对现实世界的全面理解和智能互动.

4 结 论

我们希望通过这种以深度学习发展历程为线索的目标检测综述, 能够让人们更全面地了解目标检测的历史演进及其与深度学习技术的紧密联系. 从目标检测领域的早期阶段, 即传统算法主导的时期, 到深度学习的兴起, 目标检测技术经历了翻天覆地的变化. 这一转变不仅体现在检测精度和效率的显著提升上, 也表现在模型对复杂场景的处理能力上. 尽管目标检测领域的快速发展已经能够解决部分问题, 但仍然面临诸多挑战. 本文旨在为未来的研究者们提供一个清晰的历史视角和当前技术挑战的概览, 从而促进该领域的持续创新与发展. 通过对过去的成就和未解决的问题的深入分析, 我们期望激发新的研究思路, 推动目标检测技术在未来的突破和进步.

作者贡献声明: 李承烨、张震、梁哲恒负责资料收集、文献归纳整理、实验设计以及论文撰写; 姚潮生、张金波提出论文指导意见; 晏荣杰、吴鹏负责指导论文写作、审阅和修改.

参 考 文 献

[1] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255

[2] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 740−755

[3] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proc of the IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2001: 511−518

[4] Viola P, Jones M J. Robust real-time object detection[J]. International Journal of Computer Vision, 2001, 57(2): 137−154

[5] Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting[C]//Proc of the 2nd European Conf on Computational Learning Theory. Berlin: Springer, 1995: 23−37

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Proc of the IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 886−893

[7] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20: 273−297

[8] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2008: 1−8

[9] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541−551

[10] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324

[11] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proc of the 26th Advances in Neural Information Processing Systems. Cambridge, MA: MIT,

2012: 1106- 1114

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 20134

[13] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[14] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580−587

[15] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904−1916

[16] Girshick R. Fast R-CNN[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1440−1448

[17] Neubeck A, Van G L. Efficient non-maximum suppression[C]//Proc of the 18th Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2006: 850−855

[18] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Proc of the 29th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 91−99

[19] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779−788

[20] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7263−7271

[21] Loffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 448−456

[22] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint, arXiv: 1804.02767, 2018

[23] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint, arXiv: 2004.10934, 2020

[24] Misra D. Mish: A self regularized non-monotonic activation function[J]. arXiv preprint, arXiv: 1908.08681, 2019

[25] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 21−37

[26] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2117−2125

[27] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Proc of the 18th Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234−241

[28] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2980−2988

[29] Cai Zhaowei, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6154−6162

[30] Law H, Deng J. CornerNet: Detecting objects as paired keypoints[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2018: 734−750

[31] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]//Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 483−499

[32] Zhou Xingyi, Wang Dequan, Krähenbühl P. Objects as points[J]. arXiv preprint, arXiv: 1904.07850, 2019

[33] Duan Kaiwen, Bai Song, Xie Lingxi, et al. CenterNet: Keypoint triplets for object detection[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6569−6578

[34] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3431−3440

[35] Tian Zhi, Shen Chunhua, Chen Hao, et al. FCOS: Fully convolutional one-stage object detection[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 9627−9636

[36] Sun Peize, Zhang Rufen, Jiang Yi, et al. Sparse R-CNN: End-to-end object detection with learnable proposals[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14454−14463

[37] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 213−229

[38] Chen Qiang, Wang Yingming, Yang Tong, et al. You only look one-level feature[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13039−13048

[39] Xie Saining, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1492−1500

[40] Ge Zheng, Liu Songtao, Wang Feng, et al. YOLOx: Exceeding YOLO series in 2021[J]. arXiv preprint, arXiv: 2107.08430, 2021

[41] Wang Chien-Yao, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 390−391

[42] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7464−7475

[43] Ding Xiaohan, Zhang Xiangyu, Ma Ningning, et al. RepVGG:

Making VGG-style convnets great again[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13733−13742

[44] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C/OL]// Proc of the 8th Int Conf on Learning Representations. Washington: ICLR, 2020[2023-11-01]. https://openreview.net/pdf?id=YicbFdNTTy

[45] Liu Zhuang, Mao Hanzi, Wu Chaoyuan, et al. A convnet for the 2020s[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11976−11986

[46] Ding Xiaohan, Zhang Xiangyu, Han Jungong, et al. Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11963−11975

[47] Liu Shiwei, Chen Tianlong, Chen Xiaohan, et al. More convNets in the 2020s: Scaling up kernels beyond 51×51 using sparsity[C/OL]// Proc of the 11th Int Conf on Learning Representations. Washington: ICLR, 2023[2023-11-02]. https://openreview.net/pdf?id=bXNl-myZkJl

[48] Ding Xiaohan, Zhang Yiyuan, Ge Yixiao, et al. UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition[J]. arXiv preprint, arXiv: 2311.15599, 2023

[49] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, arXiv: 1704.04861, 2017

[50] Sandler M, Howard A, Zhu Menglong, et al. MobileNetv2: Inverted residuals and linear bottlenecks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4510−4520

[51] Howard A, Sandler M, Chu G, et al. Searching for mobileNetv3[C]// Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1314−1324

[52] Zhou Daquan, Hou Qibin, Chen Yunpeng, et al. Rethinking bottleneck structure for efficient mobile network design[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 680−697

[53] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6848−6856

[54] Ma Ningning, Zhang Xiangyu, Zheng Haitao, et al. ShuffleNet v2: Practical guidelines for efficient cnn architecture design[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 116−131

[55] Tan Mingxing, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 6105−6114

[56] Tan Mingxing, Le Q. EfficientNetv2: Smaller models and faster training[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 10096−10106

[57] Tan Mingxing, Pang Ruoming, Le Q V. EfficientDet: Scalable and efficient object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10781−10790

[58] Han Kai, Wang Yunhe, Tian Qi, et al. GhostNet: More features from cheap operations[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1580−1589

[59] Tang Yehui, Han Kai, Guo Jianyuan, et al. GhostNetv2: Enhance cheap operation with long-range attention[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 9969−9982

[60] Vasu P K A, Gabriel J, Zhu J, et al. MobileOne: An improved one millisecond mobile backbone[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7907−7917

[61] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C]//Proc of the 28th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2204−2212

[62] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]//Proc of the 29th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 2017−2025

[63] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable convolutional networks[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 764−773

[64] Zhu Xizhou, Hu Han, Lin S, et al. Deformable ConvNets v2: More deformable, better results[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 9308−9316

[65] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks[C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132−7141

[66] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 3−19

[67] Gao Zilin, Xie Jiangtao, Wang Qilong, et al. Global second-order pooling convolutional networks[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3024−3033

[68] Wang Qilong, Wu Banggu, Zhu Pengfei, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11534−11542

[69] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7794−7803

[70] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5998−6008

[71] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep

bidirectional transformers for language understanding[C]//Proc of the Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[72] Zhu Xizhou, Su Weijie, Lu Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C/OL]// Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-11-01]. https://openreview.net/pdf?id=gZ9hCDWe6ke

[73] Yao Zhuyu, Ai Jiangbo, Li Boxun, et al. Efficient DETR: Improving end-to-end object detector with dense prior[J]. arXiv preprint, arXiv: 2104.01318, 2021

[74] Wang Tao, Yuan Li, Chen Yunpeng, et al. Pnp-DETR: Towards efficient visual analysis with transformers[C]//Proc of the IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 4661−4670

[75] Roh B, Shin J W, Shin W, et al. Sparse DETR: Efficient end-to-end object detection with learnable sparsity[C/OL]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-10-04]. https://openreview.net/pdf?id=RRGVCN8kjim

[76] Caron M, Bojanowski P, Joulin A, et al. Deep clustering for unsupervised learning of visual features[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 132−149

[77] Asano Y M, Rupprecht C, Vedaldi A. Self-labelling via simultaneous clustering and representation learning[J]. arXiv preprint, arXiv: 1911.05371, 2019

[78] Cao Yue, Xie Zhenda, Liu Bin, et al. Parametric instance classification for unsupervised visual feature learning[C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 15614−15624

[79] Dai Zhigang, Cai Bolun, Lin Yugeng, et al. UP-DETR: Unsupervised pre-training for object detection with transformers[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1601−1610

[80] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[R/OL]. OpenAI, 2018 [2024-02-01]. https://openai.com/research/language-unsupervised

[81] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[R/OL]. OpenAI blog, 2019[2024-01-01]. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[82] He Kaiming, Fan Haoqi, Wu Yuxin, et al. Momentum contrast for unsupervised visual representation learning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9729−9738

[83] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 1597−1607

[84] Chen Xinlei, Fan Haoqi, Girshick R, et al. Improved baselines with momentum contrastive learning[J]. arXiv preprint, arXiv: 2003.04297, 2020

[85] Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments[C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 9912−9924

[86] Wang Wenhai, Xie Enze, Li Xiang, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 568−578

[87] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012−10022

[88] Fang Yuxin, Liao Bencheng, Wang Xinggang, et al. You only look at one sequence: Rethinking transformer in vision through object detection[J]. Proc of the 35th Advances in Neural Information Processing Systems. Cambrige, MA: MIT, 2021: 26183−26197

[89] Li L H, Zhang Pengchuan, Zhang Haotian, et al. Grounded language-image pre-training[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10965−10975

[90] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[91] Liu Shilong, Li Feng, Zhang Hao, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR[C/OL]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-10-02]. https://openreview.net/pdf?id=oMI9PjOb9Jl

[92] Li Feng, Zhang Hao, Liu Shilong, et al. Dn-DETR: Accelerate DETR training by introducing query denoising[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13619−13627

[93] Zong Zhuofan, Song Guanglu, Liu Yu. DETRs with collaborative hybrid assignments training[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6748−6758

[94] Hatamizadeh A, Yin H, Heinrich G, et al. Global context vision transformers[C]//Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 12633−12646

[95] Zhang Gongjie, Luo Zhipeng, Tian Zichen, et al. Towards efficient use of multi-scale features in transformer-based object detectors[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 6206−6216

[96] Han Kai, Wang Yunhe, Chen Hanting, et al. A survey on vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87−110

[97] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM Computing Surveys, 2022, 54(10s): 1−41

[98] Liu Yang, Zhang Yao, Wang Yixin, et al. A survey of visual transformers[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 36(6): 7478−7498

[99] Shehzadi T, Hashmi K A, Stricker D, et al. 2D object detection with transformers: A review[J]. arXiv preprint, arXiv: 2306.04670, 2023

[100] Zuo Shuangquan, Xiao Yun, Chang Xiaojun, et al. Vision

transformers for dense prediction: A survey[J]. Knowledge-Based Systems, 2022, 253: 109552

[101] Xu Peng, Zhu Xiatian, Clifton D A. Multimodal learning with transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113−12132

[102] Lahoud J, Cao Jiale, Khan F S, et al. 3D vision with transformers: A survey[J]. arXiv preprint, arXiv: 2208.04309, 2022

[103] Ali A M, Benjdira B, Koubaa A, et al. Vision transformers in image restoration: A survey[J]. Sensors, 2023, 23(5): 2385

[104] Thisanke H, Deshan C, Chamith K, et al. Semantic segmentation using vision transformers: A survey[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 106669

[105] Aleissaee A A, Kumar A, Anwer R M, et al. Transformers in remote sensing: A survey[J]. Remote Sensing, 2023, 15(7): 1860

[106] Xu Hongming, Xu Qi, Cong Fengyu, et al. Vision transformers for computational histopathology[J]. IEEE Reviews in Biomedical Engineering, 2023, 17: 63−79

[107] Shamshad F, Khan S, Zamir S W, et al. Transformers in medical imaging: A survey[J]. Medical Image Analysis, 2023, 88: 102802

[108] Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-Mixer: An all-MLP architecture for vision[C]//Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 24261−24272

[109] Touvron H, Bojanowski P, Caron M, et al. ResMLP: Feedforward networks for image classification with data-efficient training[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 5314−5321

[110] Ding Xiaohan, Xia Chunlong, Zhang Xiangyu, et al. RepMLP: Re-parameterizing convolutions into fully-connected layers for image recognition[J]. arXiv preprint, arXiv: 2105.01883, 2021: 24261−24272

[111] Hou Qibin, Jiang Zihang, Yuan Li, et al. Vision Permutator: A permutable MLP-like architecture for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 1328−1334

[112] Liu Hanxiao, Dai Zihang, So D, et al. Pay attention to MLPs[C]//Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 9204−9215

[113] Yu Tan, Li Xu, Cai Yunfeng, et al. S^2-MLP: Spatial-shift MLP architecture for vision[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2022: 297−306

[114] Wang Ziyu, Jiang Wenhao, Zhu Yiming, et al. DynaMixer: A vision MLP architecture with dynamic mixing[C]//Proc of the 39th Int Conf on Machine Learning. New York: ACM, 2022: 22691−22701

[115] Lian Dongze, Yu Zehao, Sun Xing, et al. As-MLP: An axial shifted MLP architecture for vision[C/OL]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-10-02]. https://openreview.net/pdf?id=fvLLcIYmXb

[116] Chen Shoufa , Xie Enze, Ge Chongjian, et al. CycleMLP: A MLP-like architecture for dense prediction[C/OL]//Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-10-02]. https://openreview.net/pdf?id=NMEceG4v69Y

[117] Guo Jianyuan, Tang Yehui, Han Kai, et al. Hire-MLP: Vision MLP via hierarchical rearrangement[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 826−836

[118] Tang Yehui, Han Kai, Guo Jianyuan, et al. An image patch is a wave: Phase-aware vision MLP[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10935−10944

[119] Lai Shenqi, Du Xi, Guo Jia, et al. RaMLP: Vision MLP via region-aware mixing[C]//Proc of the 32nd Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2023: 999−1007

[120] Sohl-Dickstein J, Weiss E, aheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2256−2265

[121] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2020: 6840−6851

[122] Chen Shoufa, Sun Peize, Song Yibing, et al. DiffusionDet: Diffusion model for object detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 19830−19843

[123] Song Jiaming, Meng Chenlin, Ermon S. Denoising diffusion implicit models[C/OL]//Proc of the 8th Int Conf on Learning Representations. Washington: ICLR, 2020[2023-10-02]. https://openreview.net/pdf?id=St1giarCHLP

[124] Deng Jiacheng, Lu Jiahao, Zhang Tianzhu. Diff3DETR: Agent-based diffusion model for semi-supervised 3D object detection[C]///Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 57−73

[125] Ranasinghe Y, Hegde D, Patel V M. MonoDiff: Monocular 3D object detection and pose estimation with diffusion models[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 10659−10670

[126] Hu Ronghang, Dollár P, He K, et al. Learning to segment every thing[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4233−4241

[127] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 4015−4026

[128] Wang Rongsheng, Duan Yaofei, Li Yukun. Segment anything also detect anything[R/OL]. EasyChair, 2023[2024-02-01]. https://easychair.org/publications/preprint/T1rc

[129] Tang Lv, Xiao Haoke, Li Bo. Can SAM segment anything? When SAM meets camouflaged object detection[J]. arXiv preprint, arXiv: 2304.04709, 2023

[130] Zhang Dingyuan, Liang Dingkang, Yang Hongcheng, et al. SAM3D: Zero-shot 3D object detection via segment anything model[J]. arXiv preprint, arXiv: 2306.02245, 2023

[131] Cui Ruikai, He Siyuan, Qiu Shi. Adaptive low rank adaptation of segment anything to salient object detection[J]. arXiv preprint, arXiv: 2308.05426, 2023

[132] Deng Ruining, Cui Can, Liu Quan, et al. Segment anything model

(SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging[J]. arXiv preprint, arXiv: 2304.04155, 2023

[133] Wu Junfeng, Jiang Yi, Liu Qihao, et al. General object foundation model for images and videos at scale[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 3783−3795

[134] Zhan Yufei, Zhu Yousong, Chen Zhiyang, et al. Griffon: Spelling out all object locations at any granularity with large language models[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 405−422

[135] Zhao Yucheng, Wang Guangting, Tang Chuanxin, et al. A battle of network structures: An empirical study of CNN, transformer, and MLP[J]. arXiv preprint, arXiv: 2108.13002, 2021

[136] Gupta A, Dollár P, Girshick R. LVIS: A dataset for large vocabulary instance segmentation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5356−5364

[137] Dai Xuerui, Yuan Xue, Wei Xueye. TIRNet: Object detection in thermal infrared images for autonomous driving[J]. Applied Intelligence, 2021, 51(3): 1244−1261

[138] Bergmann P, Fauser M, Sattlegger D, et al. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 9592−9600

[139] Li Yu, Wang Tao, Kang Bingyi, et al. Overcoming classifier imbalance for long-tail object detection with balanced group softmax[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10991−11000

[140] Guo Wei, Tang Sitao, Wang Chunyan. Object detection algorithm of road traffic thermal infrared image based on YOLOv7[J/OL]. Computer Technology and Development, 2014[2024-08-19]. https://doi.org/10.20165/j.cnki.ISSN1673-629X.2024.0223 (in Chinese)
(郭伟, 唐思涛, 王春艳. 基于 YOLOv7 道路交通热红外图像目标检测算法[J/OL]. 计算机技术与发展, 2024[2024-08-19]. https://doi.org/10.20165/j.cnki.ISSN1673-629X.2024.0223)

[141] Yin Jiachao, Lü Chaowen, Suo Ke, et al. PCB defect detection algorithm based on EfficientNetV2[J/OL]. Journal of Computer-Aided Design & Computer Graphics, 2024, 7: 1260−1269 (in Chinese)
(尹嘉超, 吕耀文, 索科, 等. 基于 EfficientNetV2 的 PCB 缺陷检测算法[J/OL]. 计算机辅助设计与图形学学报, 2024, 7: 1260−1269)

[142] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2020, 128(7): 1956−1981

[143] Chen Yuhua, Li Wen, Sakaridis C, et al. Domain adaptive faster R-CNN for object detection in the wild[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 3339−3348

[144] Kang Bingyi, Liu Zhuang, Wang Xin, et al. Few-shot object detection via feature reweighting[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 8420−8429

[145] Tseng H Y, Lee H Y, Huang J B, et al. Cross-domain few-shot classification via learned feature-wise transformation[C/OL]//Proc of the 8th Int Conf on Learning Representations. Washington: ICLR, 2020[2023-10-02]. https://openreview.net/pdf?id=SJl5Np4tPr

[146] Qiao Siyuan, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 10213−10224

[147] Wu Yuxin, He Kaiming. Group normalization[C]//Proc of the European Conf on Computer Vision. Berlin: Springer, 2018: 3−19

[148] Bai Yancheng, Zhang Yongqiang, Ding Mingli, et al. Sod-mtgan: Small object detection via multi-task generative adversarial network[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 206−221

[149] Noh J, Bae W, Lee W, et al. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 9725−9734

[150] Deng Deng, Guo Jia, Ververas E, et al. Retinaface: Single-shot multi-level face localisation in the wild[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 5203−5212

[151] Gong Yuqi, Yu Xuehui, Ding Yao, et al. Effective fusion factor in FPN for tiny object detection[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 1160−1168

[152] Ying Xinyi, Liu Li, Wang Yingqian, et al. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15528−15538

[153] Yuan Xiang, Cheng Gong, Yan Kebing, et al. Small object detection via coarse-to-fine proposal generation and imitation learning[C]// Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6317−6327

[154] Chen Tianxiang, Tan Zhentao, Gong Tao, et al. Mim-iSTd: Mamba-in-Mamba for efficient infrared small target detection[J]. arXiv preprint, arXiv: 2403.02148, 2024

[155] Liu Qiankun, Liu Rui, Zheng Bolun, et al. Infrared small target detection with scale and location sensitivity[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 17490−17499

[156] Xu Jiayi. The research on the aerial small object detection algorithm based on super-resolution[D/OL]. Changsha: Hunan University, 2023. DOI:10.27135/d.cnki.ghudu.2023.000566 (in Chinese)
(许嘉怡. 基于超分辨率重建的空中小目标检测算法研究[D/OL]. 长沙: 湖南大学, 2023. DOI: 10.27135/d.cnki.ghudu.2023.000566)

[157] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2846−2854

[158] Li Dong, Huang Jia Bin, Li Yali, et al. Weakly supervised object localization with progressive domain adaptation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 3512−3520

[159] Yin Yufei, Deng Jiajun, Zhou Wengang, et al. Cyclic-bootstrap labeling for weakly supervised object detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 7008−7018

[160] Gao Shuyong, Xing Haozhe, Zhang Wei, et al. Weakly supervised video salient object detection via point supervision[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 3656−3665

[161] Tang Zongheng, Sun Yifan, Liu Si, et al. DETR with additional global aggregation for cross-domain weakly supervised object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11422−11432

[162] Liu Chang, Zhang Weiming, Lin Xiangru, et al. Ambiguity-resistant semi-supervised learning for dense object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15579−15588

[163] Yang Yanchao, Loquercio A, Scaramuzza D, et al. Unsupervised moving object detection via contextual information separation[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 879−888

[164] Wang Yangtao, Shen Xi, Hu S X, et al. Self-supervised transformers for unsupervised object discovery using normalized cut[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14543−14553

[165] Xie Enze, Ding Jian, Wang Wenhai, et al. DetCo: Unsupervised contrastive learning for object detection[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 8392−8401

[166] Inoue N, Furuta R, Yamasaki T, et al. Cross-domain weakly-supervised object detection through progressive domain adaptation[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5001−5009

[167] Kim T, Jeong M, Kim S, et al. Diversify and match: A domain adaptive representation learning paradigm for object detection[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 12456−12465

[168] Zhu Xinge, Pang Jiangmiao, Yang Ceyuan, et al. Adapting object detectors via selective cross-domain alignment[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 687−696

[169] Vs V, Gupta V, Oza P, et al. Mega-cda: Memory guided attention for category-aware unsupervised domain adaptive object detection[C]// Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4516−4526

[170] He Zhenwei, Zhang Lei. Domain adaptive object detection via asymmetric tri-way faster-RCNN[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 309−324

[171] Cao Shengcao, Joshi D, Gui L Y, et al. Contrastive mean teacher for domain adaptive object detectors[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23839−23848

[172] Zhang Wenyu, Shen Li, Foo C S. Rethinking the role of pre-trained networks in source-free domain adaptation[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 18841−18851

[173] Zhang Xingxuan, Xu Zekai, Xu Renzhe, et al. Towards domain generalization in object detection[J]. arXiv preprint, arXiv: 2203.14387, 2022

[174] Vidit V, Engilberge M, Salzmann M. Clip the gap: A single domain generalization approach for object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 3219−3229

[175] Lehner A, Gasperini S, Marcos-Ramiro A, et al. 3D-VField: Adversarial augmentation of point clouds for domain generalization in 3D object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 17295−17304

[176] Long Shaocong, Zhou Qianyu, Ying Chenhao, et al. Rethinking domain generalization: Discriminability and generalizability[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 11783−11797

[177] Fan Qi, Zhuo Wei, Tang C K, et al. Few-shot object detection with attention-RPN and multi-relation detector[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4013−4022

[178] Qiao Limeng, Zhao Yuxuan, Li Zhiyuan, et al. DeFRCN: Decoupled Faster R-CNN for few-shot object detection[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 8681−8690

[179] Xiao Yang, Lepetit V, Marlet R. Few-shot object detection and viewpoint estimation for objects in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3090−3106

[180] Jiang Xinyu, Li Zhengjia, Tian Maoqing, et al. Few-shot object detection via improved classification features[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2023: 5386−5395

[181] Dong Na, Zhang Yongqiang, Ding Mingli, et al. Incremental-DETR: Incremental few-shot object detection via self-supervised learning[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 543−551

[182] Li Yaohui, Zhou Qifeng, Chen Haoxing, et al. The Devil is in the few shots: Iterative visual knowledge completion for few-shot learning[J]. arXiv preprint, arXiv: 2404.09778, 2024

[183] Chen Xiaozhi, Ma Huimin, Wan Ji, et al. Multi-view 3D object detection network for autonomous driving[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 1907−1915

[184] Yang Bin, Luo Wenjie, Urtasun R. Pixor: Real-time 3D object detection from point clouds[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7652−7660

[185] Pan Xuran, Xia Zhuofan, Song Shiji, et al. 3D object detection with pointformer[C]//Proc of the IEEE/CVF Conf on Computer Vision

and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7463−7472

[186] Chen Yukang, Liu Jianhui, Zhang Xiangyu, et al. Voxelnext: Fully sparse voxelnet for 3D object detection and tracking[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 21674−21683

[187] Jiang Xiaohui, Li Shuailin, Liu Yingfei, et al. Far3D: Expanding the horizon for surround-view 3D object detection[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 38(3): 2561−2569

[188] Shmelkov K, Schmid C, Alahari K. Incremental learning of object detectors without catastrophic forgetting[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3400−3409

[189] Perez-Rua J M, Zhu Xiatian, Hospedales T M, et al. Incremental few-shot object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 13846−13855

[190] Zhang Junting, Zhang Jie, Ghosh S, et al. Class-incremental learning via deep model consolidation[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 1131−1140

[191] Joseph K, Rajasegaran J, Khan S, et al. Incremental object detection via meta-learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9209−9216

[192] Liu Yaoyao, Schiele B, Vedaldi A, et al. Continual detection transformer for incremental object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23799−23808

[193] Joseph K, Khan S, Khan F S, et al. Towards open world object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5830−5840

[194] Wang Zhenyu, Li Yali, Chen Xi, et al. Detecting everything in the open world: Towards universal object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11433−11443

[195] Wu Zhiheng, Lu Yue, Chen Xingyu, et al. UC-OWOD: Unknown-classified open world object detection[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2022: 193−210

[196] Zohar O, Wang K C, Yeung S. Prob: Probabilistic objectness for open world object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11444−11453

[197] Ma Zeyu, Yang Yang, Wang Guoqing, et al. Rethinking open-world object detection in autonomous driving scenarios[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 1279−1288

[198] Wang Yanghao, Yue Zhongqi, Hua X S, et al. Random boxes are open-world object detectors[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6233−6243

[199] Kim D, Lin T Y, Angelova A, et al. Learning open-world object proposals without learning to classify[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5453−5460

[200] Zareian A, Rosa K D, Hu D H, et al. Open-vocabulary object detection using captions[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14393−14402

[201] Gu Xiuye, Lin T Y, Kuo Weicheng, et al. Open-vocabulary object detection via vision and language knowledge distillation[C/OL]// Proc of the 9th Int Conf on Learning Representations. Washington: ICLR, 2021[2023-10-02]. https://arxiv.org/abs/2104.13921

[202] Du Yu, Wei Fangyun, Zhang Zihe, et al. Learning to prompt for open-vocabulary object detection with vision-language model[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 14084−14093

[203] Minderer M, Gritsenko A, Stone A, et al. Simple open-vocabulary object detection[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 728−755

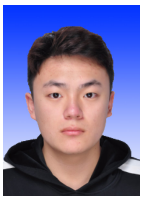
[204] Kim D, Angelova A, Kuo Weicheng. Region-aware pretraining for open-vocabulary object detection with vision transformers[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 11144−11154

[205] Shi Cheng, Yang Sibei. Edadet: Open-vocabulary object detection using early dense alignment[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 15724−15734

[206] Wang Tao. Learning to detect and segment for open vocabulary object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway. NJ: IEEE, 2023: 7051−7060

[207] Bai Yutong, Geng Xinyang, Mangalam K, et al. Sequential modeling enables scalable learning for large vision models[J]. arXiv preprint, arXiv: 2312.00785, 2023

[208] Chen Yao, Lü Qing, Qi Xiaofeng. Develop mobile phone terminal application software DeepBC based on breast ultrasound artificial intelligence[J]. Chinese Journal of Bases and Clinics in General Surgery, 2022, 29(1): 46−50 (in Chinese)
(陈瑶, 吕青, 戚晓峰. 开发基于乳腺超声人工智能的移动终端应用软件: DeepBC[J]. 中国普外基础与临床杂志, 2022, 29(1): 46−50)



Li Chengye, born in 1999. Master candidate. Student member of CCF. His main research interests include trustworthy AI and deep learning.
李承烨, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为可信人工智能、深度学习.

Zhang Zhen, born in 1995. PhD candidate. Student member of CCF. His main research interests include anomaly detection, deep learning, computer vision, and natural language processing.
张　震, 1995 年生. 博士研究生. CCF 学生会员. 主要研究方向为异常检测、深度学习、计算机视觉、自然语言处理.

Liang Zheheng, born in 1986. Master, senior engineer. Member of CCF. His main research interest includes digital evaluation technology.

梁哲恒, 1986 年生. 硕士, 高级工程师. CCF 会员. 主要研究方向为数字化评测技术.

Yao Chaosheng, born in 1989. Bachelor, engineer. Member of CCF. His main research interest includes project quality control.

姚潮生, 1989 年生. 学士, 工程师. CCF 会员. 主要研究方向为项目质量管控.

Zhang Jinbo, born in 1979. Master, senior engineer. His main research interest includes information system testing.

张金波, 1979 年生. 硕士, 高级工程师. 主要研究方向为信息系统测试.

Yan Rongjie, born in 1977. PhD, associate professor. Senior member of CCF. Her main research interest includes testing and validation of intelligent software and autonomous unmanned systems.

晏荣杰, 1977 年生. 博士, 副研究员. CCF 高级会员. 主要研究方向为智能软件、自主无人系统的测试与验证.

Wu Peng, born in 1977. PhD, associate professor. Senior member of CCF. His main research interests include formal method, concurrent testing, and machine learning.

吴　鹏, 1977 年生. 博士, 副研究员. CCF 高级会员. 主要研究方向为形式化方法、并发测试、机器学习.