计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(2):1−12 [doi: 10.15888/j.cnki.csa.008303]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

深度学习目标检测方法综述①

谢　富, 朱定局

(华南师范大学 计算机学院, 广州 510631)
通信作者: 朱定局, E-mail: zhudingju@m.scnu.edu.cn

摘　要: 随着深度学习在目标检测领域的大规模应用, 目标检测技术的精度和速度得到迅速提高, 已被广泛应用于行人检测、人脸检测、文字检测、交通标志及信号灯检测和遥感图像检测等领域. 本文在基于调研国内外相关文献的基础上对目标检测方法进行了综述. 首先介绍了目标检测领域的研究现状以及对目标检测算法进行检验的数据集和性能指标. 对两类不同架构的目标检测算法, 基于区域建议的双阶段目标检测算法和基于回归分析的单阶段目标检测算法的一些典型算法的流程架构、性能效果、优缺点进行了详细的阐述, 还补充了一些近几年来新出现的目标检测算法, 并列出了各种算法在主流数据集上的实验结果和优缺点对比. 最后对目标检测的一些常见应用场景进行说明, 并结合当前的研究热点分析了未来发展趋势.

关键词: 深度学习; 目标检测; 区域建议; 回归分析

引用格式: 谢富,朱定局.深度学习目标检测方法综述.计算机系统应用,2022,31(2):1–12. http://www.c-s-a.org.cn/1003-3254/8303.html

Survey on Deep Learning Object Detection

XIE Fu, ZHU Ding-Ju

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: With the large-scale application of deep learning in the field of object detection, the accuracy and speed of object detection technology have been rapidly improved, and it has been widely used in many fields, including pedestrian detection, face detection, text detection, traffic sign and signal light detection, and remote sensing image detection. This study reviews object detection technology based on the investigation of relevant domestic and foreign literature. First, the research status of object detection as well as the datasets and performance indicators for object detection algorithm tests are introduced. In this paper, two kinds of typical object detection algorithms with different architectures, namely two-stage object detection algorithms based on region proposals and one-stage object detection algorithms based on regression analysis, are described elaborately in their process architectures, performance effect, advantages, and disadvantages. In addition, some new object detection algorithms developed in recent years have been supplemented, and the experimental results and advantages and disadvantages of various algorithms on mainstream datasets are listed. Finally, some common application scenarios of object detection are specified, and future development trends are analyzed considering current research hotspots.

Key words: deep learning; object detection; region proposal; regression analysis

1　引言

目标检测的基本任务是需要判别图片中被检测的目标类别, 同时需要使用矩形边界框来确立目标的所在位置及大小, 并给出相应的置信度. 作为计算机视觉

① 基金项目: 广东省高校人工智能重点领域专项 (2019KZDZX1027); 广东省公益研究与能力建设 (2018B070714018)
收稿时间: 2021-04-16; 修改时间: 2021-05-11; 采用时间: 2021-05-19; csa 在线出版时间: 2022-01-17

领域的一个基本问题, 目标检测也是许多计算机视觉任务如图像分割、目标追踪、图像描述的基础. 在过去的 10 年里, 目标检测在计算机视觉领域受到了热烈的关注, 出现了越来越多的有关目标检测的论文发表(如图 1), 其中包含了目标检测方法的理论创新, 和对已有目标检测模型的改进和推广应用. 由于在目标检测过程中各类目标的大小, 形状, 姿态等各有不同, 同时还受到外部条件如光线, 遮挡等原因[1] 影响, 给目标检测带来了一系列困难, 国内外许多学者都对此进行了系统性的研究.

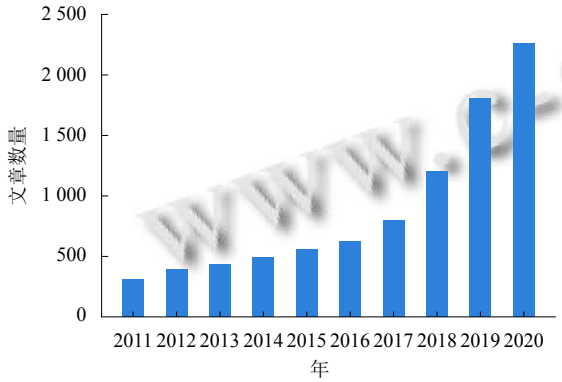


图 1 2011–2020 年目标检测相关论文的数量
(数据来源自 Google 学术检索关键字“object detection”和 “detecting objects”)

深度学习方法应用到目标检测领域之前, 目标检测领域发展平缓. 在 2012 年的 ImageNet[2] 分类任务中, 卷积神经网络的应用使得图像分类任务的效果大大提高, 在此推动下, Girshick 等[3] 首次在目标检测领域中使用了区域卷积网络 (regions with CNN features, R-CNN), 在检测效果上取得了非常巨大的提升. 此后, 深度学习与目标检测任务的结合使得目标检测领域开始迅速发展, 并在实践中得到广泛应用. 目标检测具有广阔的应用前景, 已经在监控安防[4]、自动驾驶[5]、遥感侦测[6]、医学图像病灶检测[7] 等领域取得了很好的效果.

2 目标检测数据集与评价指标

2.1 数据集

目标检测领域常用数据集有 PASCAL VOC[8]、ImageNet[9]、MS-COCO[10]、Open Images[11]、DOTA[12]. 常用数据集的样本与标注示例如图 2 所示[12,13].

2.1.1 PASCAL VOC

PASCAL VOC (the PASCAL visual object classification) 数据集最早于 2005 年发布, 最初只有 4 个类别, 2005–2012 年每年更新一次, 主要用于图像分类、目标检测任务. 目前广泛使用的是 PASCAL VOC 2007 和 PASCAL VOC 2012 两个版本的数据集, 其中, PASCAL VOC 2007 包含 9 963 张标注过的图片, 标注出 24 640 个目标物体; PASCAL VOC 2012 包含 11 530 张图片, 标注出 27 450 个目标物体. 这两个数据集都包含了 20 个类别的数据, 主要有人、动物、交通工具、室内物品等, 并且数据集中的图像都有对应的 XML 文件对目标的位置和类别进行标注.

(a) PASCAL VOC (b) ImageNet (c) MS-COCO (d) Open Images (e) DOTA

图 2 常用数据集的样本与标注

2.1.2 ImageNet

ImageNet 是由斯坦福大学和普林斯顿大学根据 WordNet 层次结构合作组织建立起来的用于视觉对象识别软件研究的大型可视化数据库, 其中层次结构的每个节点都是由成百上千张图像组成的. ImageNet 由计算机视觉领域的专业人员维护, 文档详细, 应用广泛, 已经成为计算机视觉领域图像算法性能检验的标准数据集. 数据集包含了 1 400 多万张图片, 2 万多个类别. 其中使用最多的子数据集是 ILSVRC (ImageNet large scale visual recognition challenge), 涵盖 1 000 个对象类别, 包含 1 281 167 张训练图像, 50 000 张验证图像和 100 000 张测试图像.

2.1.3 MS-COCO

MS-COCO (Microsoft common objects in context) 数据集首次发布于 2015 年, 是由微软公司开发维护的大型图像数据集, 主要用于目标检测, 图像分割, 图像标题生成任务. 一共包含了 32.8 万张图片, 其中有超过 20 万张图片有详细标注, 包含了 91 个物体类别, 具有场景复杂、单张图片目标多、小目标物体多等特点, 是目前图像分割领域最大的数据集.

2.1.4 Open Images

Open Images 是谷歌团队发布的用于图像分类、

目标检测、视觉关系检测、图像分割和图像描述的数据集. 2020 年最新发布的 Open Images V6 包含 900 万张图片, 600 种对象, 1 600 万个 bounding-box 标注, 是目前最大的带图像位置标注的数据集. Open Images 图像库中的 bounding-box 大部分都是由专业人员手工绘制的, 确保了标注的准确性与一致性. 图像场景复杂, 通常包含多个目标 (平均每张图片 8.3 个).

2.1.5 DOTA

航空遥感图像不同于传统的图像数据, 具有尺度变化大、目标小且密集、检测目标形态多样等特点. DOTA 是航空遥感图像检测的常用数据集, 包含了 2 806 张各种尺度大小图像, 图像尺寸从 800×800 到 4000×4000 不等, 数据集划分为 1/6 验证集, 1/3 测试集, 1/2 训练集. DOTA 数据集的图像全部是由领域内人士标注的, 总计 15 个类别 188 282 个目标对象.

2.2 评价指标

2.2.1 FFPW/FFPI

FPPW (false positives per-window) 最早是用于 INRIA 行人数据集[14] 评估性能, 在早期的行人检测中应用较广, 但是由于 FPPW 存在缺陷, 对某些实例不能很好的预测整张图片. 到 2009 年, Caltech 行人数据集[15] 出现后, 评估标准就由针对窗口的 FPPW 过渡为适用于整张图片的 FPPI (false positives per-image).

2.2.2 AP/mAP

在目标检测研究中, 常用于评价检测效果的一个标准是 AP (average precision), 最初在 PASCAL VOC 2007 被引入, 由 P-R 曲线和坐标围起来的面积组成, 用于表示不同召回率下检测的平均正确性, 是对一个特定类别下目标检测器效果的评估. mAP (mean average precision) 为各类别 AP 的平均值, 用于对所有目标类别检测的效果取平均值, 是检测性能的最终度量.

2.2.3 IoU

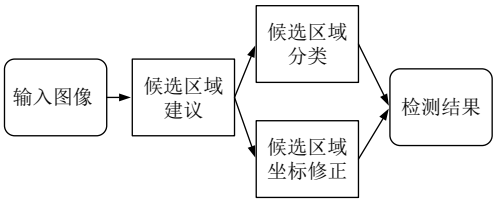
交并比 (intersection over union, IoU) 在目标检测的性能评价时用的非常多, 表示的是预测的边框和原图片标注的真实边框的交叠率, 是两者交集与并集的比值. 当比值为 1 的时候则说明预测的效果达到最佳.

2.2.4 FPS/FLOPs

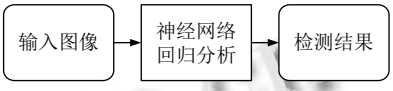
检测速度代表目标检测算法与模型的计算性能, 需要在同一硬件条件下进行比较. 目标检测技术在准确度上已经有了很大的提高, 但是如果不考虑计算性能, 使用复杂的模型会对硬件的计算能力和内存要求较高, 导致部署成本大大增加. 通常目标检测的速度性能评价指标有 FPS (frame per second), 代表检测器每秒可以处理的图片帧数, 数值越大代表检测速度越快. 浮点运算数 (floating point operations, FLOPs) 可以理解为计算量, 用来衡量算法与模型的复杂度. 模型的 FLOPs 与许多因素有关, 比如参数量、网络层数、选用的激活函数等. 一般情况下, 参数量低的网络运算量会比较小, 使用的内存也小, 更利于嵌入式端的部署.

3 基于深度学习的目标检测算法

目前主流的深度学习目标检测算法有两类 (如图 3), 基于区域建议的双阶段目标检测算法, 如 R-CNN、SPP-Net、Fast R-CNN、Faster R-CNN、FPN、Mask R-CNN; 基于回归分析的单阶段目标检测算法, 如 YOLO 系列、SSD 系列、RetinaNet. 最近几年, 还出现了 NAS-FPN、EfficientDet、YOLOF 等新算法.



(a) 双阶段目标检测算法



(b) 单阶段目标检测算法

图 3 基于深度学习的目标检测算法

3.1 双阶段目标检测算法

双阶段目标检测算法一般先使用算法 (选择性搜索或者区域建议网络等) 对图像提取候选框, 然后对候选框目标进行二次修正得到检测结果. 代表算法有: R-CNN、SPP-Net、Fast R-CNN、Faster R-CNN、feature pyramid networks (FPN)、Mask R-CNN.

3.1.1 R-CNN

2014 年, 伯克利大学的 Girshick 等提出了 R-CNN[3], 在 PASCAL VOC 2007 数据集中取得出色的效果, 比之前其他方法有近 50% 的性能提升, mAP 达到了 58.5%. R-CNN 的模型如图 4 所示, 首先是通过选择性搜索提取可能的目标区域, 统一大小后使用 CNN 在这些候选区域上提取特征, 最后将这些特征输入支持向量机分类器对区域进行分类, 位置信息则由全连接神经网络回归来得到.

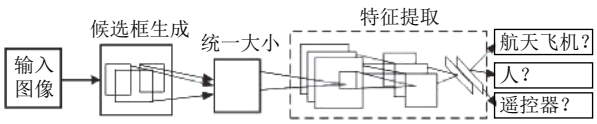


图 4 R-CNN 模型

R-CNN 缺点:

(1) 多阶段训练过程. 各阶段相对独立, 训练繁琐复杂.

(2) 图像易失真. 候选区域需要放缩到固定大小会导致不期望看到的几何形变.

(3) 计算开销大, 检测速度慢. 尤其对于高密度的图片, 使用选择性搜索找到的每个区域都要使用卷积神经网络提取特征.

3.1.2 SPP-Net

针对 R-CNN 对图像进行缩放导致图像失真的问题, He 等[16] 在 2014 年提出了 SPP-Net, 将一个空间金字塔池化 (spatial pyramid pooling, SPP) 层添加在卷积层和全连接层之间, 从而可以不用对候选区域进行缩放就能进行任意比例区域的特征提取. 同时由于 SPP-Net 是将整个图片送入卷积神经网络提取特征, 减少了候选区域的重复计算. 这些改进使得 SPP-Net 算法比 R-CNN 算法检测速度提高 24–102 倍, 同时在 PASCAL VOC 2007 数据集上的 mAP 提高到 59.2%.

SPP-Net 仍然存在以下问题:

1) 训练过程仍然是多阶段的, 步骤繁杂 (微调网络+训练 SVM+训练边框回归器).

2) 微调算法不更新 SPP 层之前的卷积层参数, 不能有效地调整参数权重, 限制了准确率.

3) 分类器使用 SVM, 无法实现端到端训练.

3.1.3 Fast R-CNN

结合了 R-CNN 和 SPP-Net 各自的特点, Girshick 等于 2015 年又提出了 Fast R-CNN[17]. Fast R-CNN 的网络结构如图 5 所示, Fast R-CNN 将整幅图像和候选区域作为输入, 经过卷积层提取到特征图, 用感兴趣池化层 (region of interest, RoI) 代替了 SPP-Net 的空间金字塔池化层来输出特征图, 然后输入到全连接层. Fast R-CNN 使用 Softmax 代替了 SVM 的二分类, 通过多任务的方式去训练整个网络, 骨干网络则使用 VGG-16 代替 AlexNet, 在一个模型中将特征提取、目标分类和位置回归都整合到一起, 方便进行训练, 同时提高了检测精度和速度. 在 PASCAL VOC 2007 数据集上, Fast R-CNN 的 mAP 达到 68%. 但是由于 Fast R-CNN 的候选区域建议上选用的算法仍是只能利用 CPU 运行的选择性搜索算法, 这限制了检测速度, 仍旧无法实现实时检测.

3.1.4 Faster R-CNN

Fast R-CNN 的论文发布不久, 2015 年 Ren 等[18] 针对选择性搜索算法提取候选区域较慢的问题, 提出了 Faster R-CNN, 在该论文中首次提出了区域建议网络 (region proposal network, RPN) 来取代选择性搜索算法. Faster R-CNN 目标检测框架如图 6 所示, 模型中的 RPN 中设计了多参考窗口机制, 使得 Faster R-CNN 可以在同一网络中, 完成候选区域推荐、特性提取和定位并分类, 大大提高了训练效率. Faster R-CNN 是第一个兼顾端到端训练和 GPU 上实时性的目标检测算法, 在 PASCAL VOC 2007 数据集上的 mAP 提升至 78%, 同时速度由 R-CNN 的 0.025 FPS 提高到 17 FPS (640×480 像素). Faster R-CNN 仍然存在一些缺点, 由于 anchor 机制的存在, 其对小目标的检测效果并不理想.

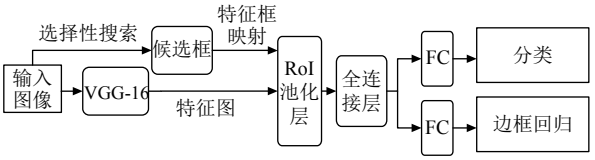


图 5 Fast R-CNN 网络结构图

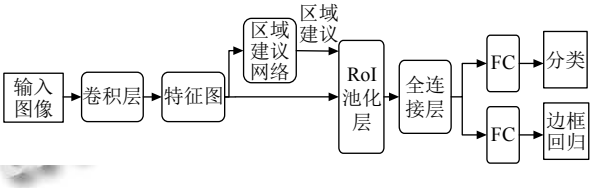


图 6 Faster R-CNN 目标检测框架

3.1.5 FPN

在 2017 年, Lin 等[19] 又在 Faster R-CNN 的基础上提出了特征金字塔网络 (FPN) 检测算法. FPN 的模型结构如图 7 所示, 其主要创新点是加入了多层特征和特征融合, 原先的目标检测算法只对语义信息丰富的顶层特征进行检测, 但是顶层特征的目标位置信息较少; 而底层特征恰好与之相反, 包含了明确的位置信息, 但是语义信息较少. FPN的做法是在网络前馈结束后, 从最顶层开始逐层与下层的特征图融合, 通过分治的方式从而可以在网络的不同深度引出检测端以便对不同尺度的目标进行检测, 大幅提升了小目标物体的检测效果. 由于 FPN 只是做了一些网络连接的优化, 基

本不增加网络的计算量, 因此在规模更大的 MS-COCO 数据集上取得了当时最佳的检测效果.

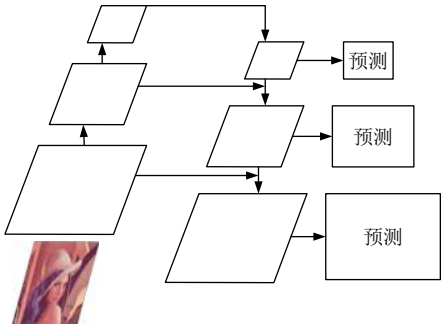


图 7 FPN 模型

3.1.6 Mask R-CNN

2017 年, He 等提出 Mask R-CNN[20], 添加了 Mask 分支, 是一个结合了图像语义分割和目标检测的通用网络. Mask R-CNN 的网络结构如图 8 所示, 通过使用 RoI Align 层替换 Faster R-CNN 的 RoI Pooling 层, 加入线性插值算法避免了特征图和原始图像由于 RoI 池的整数量化导致的偏差问题, 让每个感受野取得的特征能更好地与原图感受野区域对齐, 从而提高了检测精度. Mask R-CNN 在 MS-COCO 数据集上的 mAP 达到了 39.8%. 但是由于 Mask R-CNN 加入了分割分支, 因此计算开销比 Faster R-CNN 大.

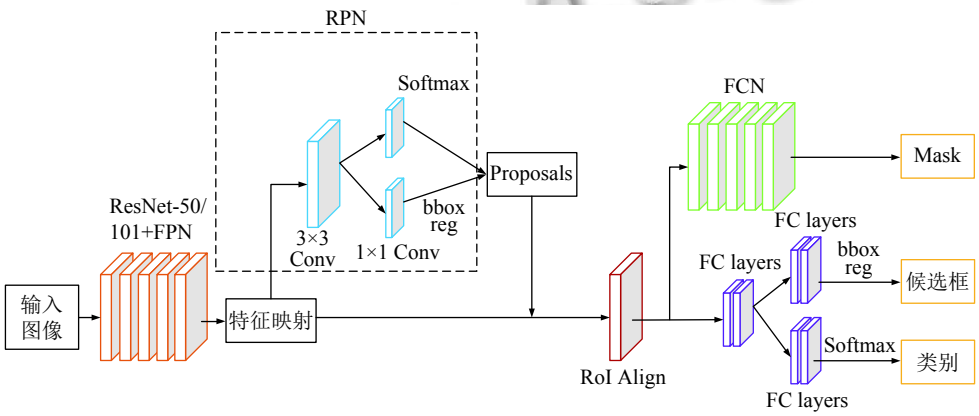


图 8 Mask R-CNN 网络结构图

3.1.7 双阶段目标检测算法对比

双阶段目标检测算法发展迅速, 检测精度也在不断提高, 但是自身体系结构的问题限制了检测速度. 双阶段目标检测算法的骨干网络以及在主流数据集上的实验结果如表 1 所示, 算法的优点/创新点和缺点以及适用场景如表 2 所示. 使用不同的输入图像、骨干网络、硬件设施会对实验结果有一定影响, 但是总体来说算法性能的对比还是符合预期的.

表 1 双阶段目标检测算法性能对比

算法	骨干网络	测试数据集	mAP (%)
R-CNN	AlexNet	PASCAL VOC 2007	58.5
SPP-Net	ZF-5	PASCAL VOC 2007	59.2
Fast R-CNN	VGG-16	PASCAL VOC 2007	68.0
Faster R-CNN	ResNet-101	PASCAL VOC 2007	78.0
FPN	FPN	MS-COCO	34.4
Mask R-CNN	ResNeXt-101	MS-COCO	39.8

3.2 单阶段目标检测算法

单阶段目标检测算法与双阶段目标检测算法最大的不同之处在于前者没有候选区域推荐阶段, 训练过程也相对简单, 可以在一个阶段直接确定目标类别并得到位置检测框. 代表算法有 YOLO 系列、SSD 系列、RetinaNet.

3.2.1 YOLO 系列

YOLO (you only look once) 是由 Redmon 等于 2015 年提出的, 是深度学习领域第一个单级检测器[21]. YOLO 算法的最大优势就是处理速度快, 其增强版本在 GPU 上的速度达到了 45 FPS, 快速版本甚至可以达到 155 FPS. 从 YOLO 的名字 (你只需要看一次) 可以看出其有着完全不同的检测方式, YOLO 的网络流程如图 9 所示, 通过使用单个神经网络直接将整张图片划分为成 $S\times S$ 的网格单元, 判定预测目标的中心是否落在网格中, 让网格来决定预测对象类别并给出相应的置信度, 使用阈值筛选去除目标出现概率较低的目标窗口, 最后使用 NMS 去除冗余窗口即可.

YOLO 算法虽然速度较快, 但也存在以下几个明

显的缺点:

(1) YOLO 划分的网格最后只会选择 IoU 最高的边界框作为输出, 因此划分的网格最多只会检测出一个目标, 如果网格中包含多个小型目标 (如鸟群这类目标), YOLO 只能检测出一个.

(2) YOLO 没有解决多尺度窗口的问题, 相比 Faster R-CNN, 其小目标检测效果较差, 定位准确度也不够优秀.

表 2 双阶段目标检测算法的优缺点及适用场景

算法	优点/创新点	缺点	适用场景
R-CNN	引入卷积神经网络结合候选区域建议	训练复杂, 耗时, 候选区域放缩易失真	目标检测
SPP-Net	整张图片作为输入, 实现了任意比例区域的特征提取, 减少计算量	SPP层之前的参数无法更新, 限制了准确率, 且空间开销大	目标检测
Fast R-CNN	引入RoI Pooling进行特征提取, 节省了检测时间和空间开销	候选区域提取的选择性搜索算法只能在CPU中运行, 限制了检测速度	目标检测
Faster R-CNN	提出了区域建议网络来提取候选区域, 提高了效率	小目标物体检测效果不好	目标检测
FPN	加入多层特征和特征融合, 提高了小物体的检测精度	多层特征融合增加了计算量	目标检测
Mask R-CNN	使用RoI Align层减少了特征图与原始图的偏差	Mask分支增加了计算开销	目标检测, 图像分割

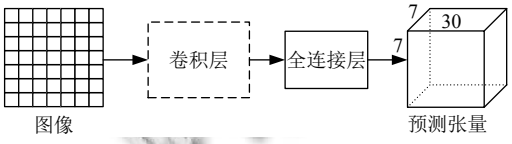


图 9 YOLO 网络流程图

YOLOv2[22] 相比前一版本, 主要改进点是提出了联合训练算法, 其基本思想是使用两种数据集同时对检测器进行训练, 检测数据集和分类数据集, 在检测数据集上来定位物体的位置, 而分类数据集则用来增加检测器的识别的物体种类. YOLOv2 在保持 YOLO 处理速度的同时, 定位更精准, 且可以识别 9 000 种不同对象, 因此又被称为 YOLO9000.

YOLOv3[23] 的特色是引入了 FPN 来实现多尺度预测, 同时还使用了更加优秀的基础网络 Darknet-53 和二值交叉熵损失函数 (binary cross-entropy loss), 并且可以通过改变模型的网络结构来实现速度与精度的平衡.

YOLOv4[24] 是 YOLO 系列的一个重大里程碑, 在 MS-COCO 数据集上的 mAP 达到了 43.5%, 速度也达到了惊人的 65 FPS. 如此大的性能提升得益于全方位的改进. YOLOv4 引入 CSPDarknet-53 提取特征, 加入了 SPP 网络来提高图像提取效果, 使用了 Mish 激活函数, 还采用 Mosaic 做数据增强, 标签平滑防止过拟合等措施, 这些改进也让 YOLOv4 成为一个极其高效强大的目标检测器.

3.2.2 SSD 系列

Liu 等[25] 于 2015 年提出的 SSD 算法结合了 YOLO 检测速度快和 Faster R-CNN 定位精准的优势. SSD 的网络结构如图 10 所示, 其主要创新是引入了多参考和多分辨率检测技术, 不同层的网络检测尺度不同的对象, 对于小目标的检测效果有了大大的提升. 除此之外, 在训练 SSD 的过程中, Liu 等[25] 为了解决难样本聚焦问题引入了难样本挖掘技术. SSD 在 PASCAL VOC 2007 上的 mAP 为 79.8%, PASCAL VOC 2012 上的 mAP 为 78.5%, MS-COCO 上的 mAP 为 28.8%, 检测速度和精度方面取得了很好的平衡.

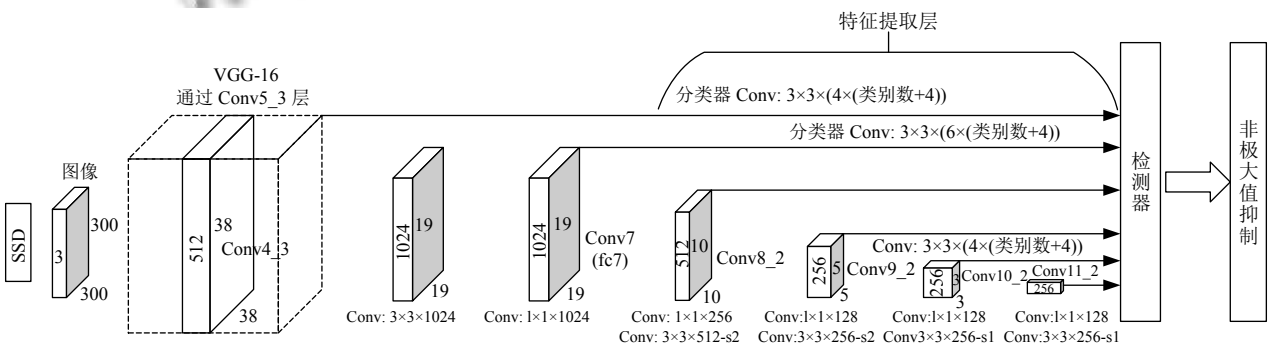


图 10 SSD300 网络结构图

DSSD[26] 使用 ResNet101 作为骨干网络以便提取更深层次的特征, 同时增加了反卷积模块 (deconvolutional module) 将原始特征图与上采样后的特征图进行融合, 并且在预测阶段引入残差单元优化分类和候选框回归. 通过这些优化, DSSD 在对小目标物体的检测效果大大提升, 但速度上则逊于 SSD 算法.

FSSD[27] 是基于 SSD 和 FPN 思想的结合, 为了解决定位和识别的语义矛盾, 需要将浅层的细节特征和高层的语义特征结合起来, 其基本做法是把各个水平的特征进行连接, 然后融合特征生成特征金字塔. FSSD 在 MS-COCO 数据集上的 mAP 达到了 31.8%, 稍弱于使用更优秀骨干网络 ResNet101 的 DSSD, 但明显优于同样使用 VGGNet 的 Faster R-CNN, 且在小目标上的检测效果是最优的.

3.2.3 RetinaNet

单阶段目标检测器虽然在速度上明显快于基于候选区域推荐的双阶段目标检测器, 但是其精度上却一直无法媲美双阶段目标检测器. Lin 等认为导致一体化卷积神经网络精度不够高的真正原因在于图像中的目标和背景层次的不匹配不均衡, 于是在 2017 年提出了 RetinaNet[28] 来解决这一问题. RetinaNet 的模型结构如图 11 所示, RetinaNet 通过引入一个聚焦损失 (focal loss) 函数, 重构了标准交叉熵损失函数, 使得检测器在训练过程中会更加注重分类困难的样本. Focal loss 的引入解决了实例样本不平衡的问题, 实现了一个精度可以媲美双阶段目标检测器的检测框架.

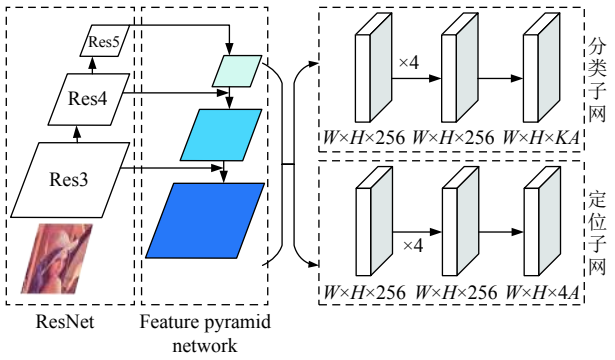


图 11 RetinaNet 模型

3.2.4 单阶段目标检测算法对比

单阶段目标检测算法提出虽然晚于双阶段目标检测算法, 但是由于其结构相对简单、检测速度优越, 因此同样受到了许多研究人员的关注. 一些单阶段目标检测算法通过引入双阶段目标检测算法的方法如 FPN、改变骨干网络、引入损失函数如 focal loss 等措施提高了检测效果, 使检测精度逐渐可以媲美双阶段目标检测算法. 单阶段目标检测算法的骨干网络以及在主流数据集上的实验结果如表 3 所示, 算法的优点/创新点和缺点以及适用场景如表 4 所示.

表 3 单阶段目标检测算法性能对比

算法	骨干网络	测试数据集	mAP (%)
YOLO	VGG-16	PASCAL VOC 2007	63.4
		PASCAL VOC 2012	57.9
YOLOv2	Darknet-19	PASCAL VOC 2007	78.6
YOLOv3	Darknet-53	MS-COCO	31.0
YOLOv4	CSPDarknet-53	MS-COCO	43.5
SSD	VGG-16	PASCAL VOC 2007	79.8
		PASCAL VOC 2012	78.5
		MS-COCO	28.8
DSSD	ResNet-101	PASCAL VOC 2007	81.5
		PASCAL VOC 2012	80.0
		MS-COCO	33.2
FSSD	VGGNet	PASCAL VOC 2007	82.7
		PASCAL VOC 2012	82.0
		MS-COCO	31.8
RetinaNet	ResNet-101	MS-COCO	34.4

3.3 最新出现的改进算法

3.3.1 NAS-FPN

2019 年, 谷歌大脑团队提出了 NAS-FPN (neural architecture search feature pyramid network)[29]. NAS-FPN 是采用神经网络结构搜索发现的一种特征金字塔结构, 同过自顶向下和自下向上的连接来进行不同尺度的特征融合. 搜索的过程中, 架构的扩展是通过 FPN 重复 N 次后连接到一起形成的. 传统的 FPN 连接都是手工设计的, 而谷歌大脑团队通过使用强化学习的方法在给定的搜索空间内不断训练网络, 控制器使用搜索空间中的子模型的准确度作为更新网络参数的奖励信号, 经过不断的反馈调整, 最终找到给定搜索空间中最好的模型架构. 在 MS-COCO 测试集上的结果显示, 使用 AmoebaNet[30] 骨干网络的 NAS-FPN 达到 48.3% mAP, 检测速度和精度都超越了 Mask R-CNN.

3.3.2 EfficientDet

随着目标检测技术的不断进步, 先进的目标检测器所需的硬件资源也越来越昂贵. 针对不同场景下的资源约束, 谷歌团队在 2019 年 11 月发表的论文中提出了 EfficientDet[31]. EfficientDet 是一系列目标检测算法的总称, 包含 8 个算法 D0–D7, 在通常的资源约束下可

以达到当时最好的检测结果. EfficientDet 的主要创新有两点, 首先提出了一种可以简单、快速地进行多尺度特征融合的加权双向特征金字塔网络 (BiFPN). 其次, 通过一种复合特征金字塔网络缩放方法, 统一缩放所有模型的分辨率、深度和宽度、特征网络和定位与分类预测网络, 使得 EfficientDet 系列参数量比之前传统的目标检测算法减少了 4–9 倍, FLOPs 缩小了 12–42 倍. 在单模型和单尺度的情况下, EfficientDet-D7 在 MS-COCO 数据集上 AP 达到了最先进的 55.1%.

3.3.3 YOLOF

之前的研究一般认为 FPN 的主要功能是可以进行多级特征的融合, 因此大多数学者的研究重点都在于实现可以更加高效地进行特征融合的网络, 如 NAS-FPN 和 EfficientDet 中的 BiFPN, 而忽视了 FPN 的另一个重要特性: 分治策略. 基于分治优化的思想, Chen 等[32] 使用了单层特征图来替代复杂的特征金字塔, 设计出了 YOLOF (you only look one-level feature) 检测框架. YOLOF 通过设计了两个核心组件, 膨胀编码器 (dilated encoder) 和均衡匹配策略 (uniform matching), 大大地提高了检测性能. 在 MS-COCO 数据集上的结果表明, 在检测精度相当的情况下, YOLOF 的检测速度比 RetinaNet 快了 2.5 倍, 比 YOLOv4 快了 13%. 同时由于 YOLOF 没有 Transformer 层, YOLOF 的训练次数也比同种类型的单层特征图算法 DETR[33] 少了 7 倍.

3.3.4 最新出现的目标检测算法对比

除了一些经典的目标检测算法, 近年来通过应用深度学习领域的新方法新技术, 出现了一些检测精度和速度都较高的目标检测算法. 这些检测算法的骨干网络以及在主流数据集上的实验结果如表 5 所示, 优点/创新点和缺点以及适用场景如表 6 所示.

表 4 单阶段目标检测算法的优缺点及适用场景

算法	优点/创新点	缺点	适用场景
YOLO	使用网格预测, 检测速度非常快	对密集和小目标检测效果不理想	目标检测
YOLOv2	使用聚类的方式生成锚框, 分类精度高	预训练的方式导致难以迁移	目标检测
YOLOv3	通过残差网络解决多尺度问题, 提高了小目标物体检测精度	模型复杂度高, 对中、大尺度物体检测效果有所降低	多尺度目标检测
YOLOv4	检测速度和精度达到了很好的平衡	模型复杂度高	高精度实时目标检测
SSD	引入了多参考和多分辨率检测技术	模型难以收敛	多尺度目标检测
DSSD	骨干网络使用了ResNet-101, 增加了反卷积模块, 提升了小目标检测效果	与SSD相比检测速度较慢	目标检测
FSSD	重构金字塔特征图以融合不同尺度特征, 有利于小目标检测	与SSD相比计算开销大, 检测速度较慢	多尺度目标检测
RetinaNet	Focal Loss的引入解决了实例样本不平衡的问题	无法适应密集样本训练	轻量级目标检测

表 5 最新出现的目标检测算法性能对比

算法	骨干网络	测试数据集	mAP (%)
NAS-FPN	ResNet-50	MS-COCO	48.3
EfficientDet	AmoebaNet	MS-COCO	55.1
YOLOF	ResNet-50	MS-COCO	44.3

4 应用领域

目标检测技术经过长时间的发展, 已经在许多领域得到了广泛的应用, 主要包括行人检测、人脸检测、文本检测、交通标志及信号灯检测和遥感图像检测等重要领域.

表 6 最新出现的目标检测算法的优缺点及适用场景

算法	优点/创新点	缺点	适用场景
NAS-FPN	使用神经网络结构调整特征金字塔结构	训练模型较为繁琐, 耗时较长	目标检测
EfficientDet	使用加权双向特征金字塔网络进行特征融合. 通过复合特征金字塔网络缩放模型, 减少了计算量	模型难以理解	目标检测, 图像分割
YOLOF	设计了膨胀编码器和均衡匹配策略提高了检测性能. 没有Transformer层, 减少了训练次数	设置的anchor比较稀疏, 推理阶段不够灵活	目标检测

4.1 行人检测

行人检测 (pedestrian detection) 研究具有悠久的历史, 早在 20 世纪 90 年代就有学者开始研究这一问题. 行人检测的难点主要在于检测目标同时具有动态和静

态的特点, 同时也受到外界环境如背景、光照的影响, 导致许多目标检测算法在应用到行人检测领域的效果并不理想. 目前主流的行人检测算法主要分为基于全局特征、基于人体部位和基于立体视觉的方法. 基于全局特征的典型算法如 Dalal 等[34] 提出的 HOG, 在当时的 MIT 行人数据集上表现非常突出. 基于人体部位的方法如 Tian 等[35] 提出了 deep parts, 其基本思想是把人体的各个部位进行分割, 分别检测后再合并, 有效解决了遮挡的问题. 基于立体视觉的检测方法如 Chen 等[36] 提出通过多光谱相机采集图像并提取图像中目标的三维信息来定位行人.

4.2 人脸检测

目标检测的另一个非常常见的应用领域是人脸检测. 人脸检测的需求最初来源于人脸识别, 逐步扩展到视频处理、图像检索、生物验证等方面. 人脸检测一直以来都受到人们热切的关注, 重要的计算机视觉领域会议 ICIP、CVPR 等每年都会有大量有关人脸检测的论文发表. 人脸检测的主要难点在于两个方面: 一是人脸自身存在的变化, 如肤色、脸型、表情等; 二是外在条件的差异如拍摄角度、光照、图像的成像条件等. Liang 等[37] 提出了通过反复曝光生成模块 (recurrent exposure generation, REG) 和多重曝光检测模块 (multi-exposure detection, MED) 结合来解决非均匀光照和图像噪声问题, 改善了人脸在弱光条件下的检测问题. 在人脸检测性能上, Zhang 等[38] 参考了特征金字塔网络的思想提出了特征集聚网络 (feature agglomeration networks, FANet), 实现了在 GPU 上实时检测 VGA 分辨率的人脸图像.

4.3 文本检测

对于文本检测, 主要包含两个过程: 文本定位和文本识别. 文本检测的挑战在于文本有不同的字体、颜色、语言等, 除此之外文本的透视失真以及模糊离散化也增加了文本识别的难度. 目前的文本检测方法主要有步进检测和综合检测两种. 步进检测是按照一定顺序逐步推进的检测方法, 由分割字符, 候选区域提取验证, 字符组划分, 单词识别等步骤组成. 步进检测过程中可以进行背景滤波, 从而降低后续处理的难度. 但是其缺点是需要确定较适宜的参数, 否则会引起误差累积. 相比较而言, 综合检测是在统一的框架下进行文本定位、分组和识别, 因此降低了累积误差, 易于集成, 其缺点是计算开销大, 因为要对大量字符类和候选框进行运算推理. 针对文本因角度变换导致的歧义问题, Zhu 等[39] 提出了 TextMountain 来确立文本中心边界概率和定位文本中心方向, 检测精度和效率都有了很大的提升. Liao 等[40] 提出了可微分二值化 (differentiable binarization, DB) 用于基于分割的场景文本检测, 该方法可以自动适应场景设置阈值进行二值化, 简化了后期处理, 同时提高了检测性能.

4.4 交通标志及信号灯检测

近些年来, 随着自动驾驶技术的火热, 交通标志及信号灯的检测也引起了许多学者的研究兴趣. 交通标志及信号灯检测的主要困难包括:

(1) 强光或夜间光照的影响;

(2) 天气如雨雪带来的干扰;

(3) 交通环境场景复杂;

(4) 车载摄像头由于运动导致拍下的画面模糊.

交通标志与灯光检测技术可以划分为两大类, 传统的检测方法和基于深度学习的检测方法. 传统的检测方法通常基于颜色, 显著性目标检测, 形态滤波, 边缘与轮廓分析, 这些方法在复杂的条件下往往会失效[41]. 基于深度学习的方法如 Faster R-CNN 和 SSD 已经具有相当高的精度, 同时也出现了一些新的技术, 如 Fan 等[42] 提出了将注意力机制用于多尺度交通标志检测, 检测效果达到了一个更高的水准.

4.5 遥感图像检测

遥感目标检测技术在城市规划、军事侦察、农业生产和航空航天等领域都有着广泛的应用. 主要检测目标包括道路、机场、港口、湖泊、飞机、船舶等.

遥感图像由于其特殊性质, 存在以下困难.

(1) 视角多样. 遥感图像只能通过俯拍得到, 目标旋转方向各有不同.

(2) 尺度变化. 同一类目标由于海拔高度等原因大小可能存在差异.

(3) 背景复杂. 遥感图像背景比较多样化, 比如城市、丛林、沙漠、山地等.

最近几年, 基于深度学习的遥感目标检测也正在逐步解决这些困难. 针对遥感图像中目标较小的问题, Long 等[43] 提出了一种使用非极大值抑制与无监督评分边框回归 (unsupervised score-based bounding box regression, USB-BBR) 相结合的方法来精准的定位小目标物体. 针对大规模遥感图像中的多尺度和任意方向的遥感目标检测, Fu 等[44] 在 Faster R-CNN 的基础

上, 使用定向边界框替代轴对齐边界框, 提出了确定方向的区域建议网络 (oriented region proposal network, RPN-O) 实现了一种旋转感知的目标检测器.

5 发展趋势

5.1 视频目标检测

视频目标检测需要对视频中每一帧图片中的可能存在目标进行正确的定位和分类. 不同于图像目标, 视频中的目标存在着运动模糊、遮挡、场景变化等因素, 使得这项任务难以取得很好的效果. 对于信息密度大的视频来说, 视频存在的大量冗余对检测的实时性也是一个巨大挑战. 研究移动目标和结合时序定位视频数据的主体目标是未来研究的主要方向.

5.2 显著目标检测

显著性目标检测研究最早开始于 1998 年 Itti 等[45] 发表的论文之后, 随着近年来图像描述技术的兴起开始受到热烈的关注. 显著性检测指的是根据显著特征, 从输入图片中定位最受关注的物体. 深度学习用于显著性目标检测的方法主要有两类, 使用多层感知机 (MLPs) 方法, 另一类则是通过完全卷积神经网络 (FCN) 进行. 多层感知机方法通常是将输入图片划分成单独的多尺度小区域, 使用卷积神经网络提取图像中的高级特征, 最后将获取的高级特征反馈到多层感知机进行显著性确认. 值得一提的是, 由于多层感知机的使用会使卷积神经网络中的空间信息丢失. 第二类基于完全卷积神经网络的显著目标检测采取另一种做法, 完全卷积神经网络最初是为了解决语义分割问题的, 但是显著对象本质上也是一种分割任务, 因此完全卷积神经网络被引入, 且不同于多层感知机, 完全卷积神经网络具有保存空间信息的能力.

5.3 基于 GAN 的目标检测

2014 年 Goodfellow 等[46] 提出了一种无监督的模型对抗生成网络 (GAN). GAN 主要有两个组件, 生成器和判别器. 生成器用于获得原始数据的特征分布以生成新的数据分布, 而判别器则是判定生成的新数据是否是真实的, 两种组件互相学习共同进步. 2017 年, Wang 等[47] 在论文中首次将 GAN 的基本思想引入目标检测领域, 其主要目的是增加训练数据让检测器能够学习到一些在通常的数据集中很难出现的特征, 通过生成检测器难以识别的样例, 从而提高模型的泛化能力. 实验结果表明, 随着检测器的性能提升, 生成对抗网络生成数据的质量也有所提高, 两者通过互相博弈提高了检测性能.

6 结论与展望

基于深度学习的目标检测技术因其巨大的优势, 如泛化能力强、复杂场景下效果出众、应用前景广阔等已经成为一个计算机视觉领域的一个热门方向. 行人检测、人脸检测、文字检测、交通标志及信号灯检测和遥感图像检测等都是目标检测的常见应用场景. 通过对不同方式的目标检测算法的对比可以看出, 双阶段目标检测算法先使用算法提取候选区域, 然后对候选框目标进行二次修正, 精度较高且定位准确, 但是训练复杂计算量大, 难以实现实时检测; 单阶段目标检测算法没有候选区域推荐过程, 在一个阶段就能确定目标类别并定位目标, 模型简单且速度快, 但是对小目标和密集目标的检测精度有待提高. 近几年来, 视频目标检测、显著目标检测和基于 GAN 的目标检测都有良好的发展势头, 新出现的目标检测算法如 NAS-FPN、EfficientDet、YOLOF 等的提出也为目标检测领域的发展提供了新的思路. 随着人们对基于深度学习的目标检测技术的进一步深入, 相信其应用领域会更加广泛, 为人类的生存发展带来更加巨大的效益.

参考文献

1 Mane S, Mangale S. Moving object detection and tracking using convolutional neural networks. 2018 2nd International Conference on Intelligent Computing and Control Systems (ICICCS). Madurai: IEEE, 2018. 1809–1813.

2 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2012. 1097–1105.

3 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

4 Joshi KA, Thakore DG. A survey on moving object detection and tracking in video surveillances system. International Journal of Soft Computing and Engineering, 2012, 2(3): 44–48.

5 Chen XZ, Ma HM, Wan J, *et al*. Multi-view 3D object

detection network for autonomous driving. Proceedings of the 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6526–6534.
6 Cheng G, Han JW. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 117: 11–28.
7 Javed R, Rahim MSM, Saba T, *et al*. A comparative study of features selection for skin lesion detection from dermoscopic images. Network Modeling Analysis in Health Informatics and Bioinformatics, 2020, 9(1): 4. [doi: 10.1007/s13721-019-0209-1]
8 Everingham M , van Gool L, Williams CK , *et al*. The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338.
9 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.
10 Lin TY, Maire M, Belongie S, *et al*. Microsoft coco: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.
11 Kuznetsova A, Rom H, Alldrin N, *et al*. The open images dataset V4. International Journal of Computer Vision, 2020, 128(7): 1956–1981. [doi: 10.1007/s11263-020-01316-z]
12 Xia GS, Bai X, Ding J, *et al*. DOTA: A large-scale dataset for object detection in aerial images. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3974–3983.
13 Zou ZX, Shi ZW, Guo YH, *et al*. Object detection in 20 years: A survey. arXiv: 1905.05055, 2019.
14 INRIA person dataset. http://pascal.inrialpes.fr/data/human/. (2020-07-30)[2021-05-13].
15 Caltech pedestrian detection benchmark. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/. (2019-07-01)[2021-05-13].
16 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
17 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440–1448.
18 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv: 1506.01497, 2015.
19 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.
20 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2980–2988.
21 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 779–788.
22 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6517–6525.
23 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.
24 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.
25 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single Shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.
26 Fu CY, Liu W, Ranga A, *et al*. DSSD: Deconvolutional single shot detector. arXiv: 1701.06659, 2017.
27 Li ZX, Zhou FQ. FSSD: Feature fusion single shot multibox detector. arXiv: 1712.00960, 2017.
28 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2999–3007.
29 Ghiasi G, Lin TY, Le QV. NAS-FPN: Learning scalable feature pyramid architecture for object detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 7029–7038.
30 Shah SAR, Wu WJ, Lu QM, *et al*. AmoebaNet: An SDN-enabled network service for big data science. Journal of Network and Computer Applications, 2018, 119: 70–82. [doi: 10.1016/j.jnca.2018.06.015]
31 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 10778–10787.

32 Chen Q, Wang YM, Yang T, *et al*. You only look one-level feature. arXiv: 2103.09460, 2021.
33 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.
34 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego: IEEE, 2005. 886–893.
35 Tian YL, Luo P, Wang XG, *et al*. Deep learning strong parts for pedestrian detection. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1904–1912.
36 Chen ZL, Huang XM. Pedestrian detection for autonomous vehicle using multi-spectral cameras. IEEE Transactions on Intelligent Vehicles, 2019, 4(2): 211–219. [doi: 10.1109/TIV.2019.2904389]
37 Liang JX, Wang JW, Quan YH, *et al*. Recurrent exposure generation for low-light face detection. IEEE Transactions on Multimedia, 2021: 1–14.
38 Zhang JL, Wu XW, Hoi SCH, *et al*. Feature agglomeration networks for single stage face detection. Neurocomputing, 2020, 380: 180–189. [doi: 10.1016/j.neucom.2019.10.087]
39 Zhu YX, Du J. Textmountain: Accurate scene text detection via instance segmentation. Pattern Recognition, 2021, 110: 107336. [doi: 10.1016/j.patcog.2020.107336]
40 Liao MH, Wan ZY, Yao C, *et al*. Real-time scene text detection with differentiable binarization. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11474–11481. [doi: 10.1609/aaai.v34i07.6812]
41 Ai CB, Tsai YCJ. Critical assessment of an enhanced traffic sign detection method using mobile LiDAR and INS technologies. Journal of Transportation Engineering, 2015, 141(5): 04014096. [doi: 10.1061/(ASCE)TE.1943-5436.0000760]
42 Fan BB, Yang H. Multi-scale traffic sign detection model with attention. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2021, 235(2–3): 708–720. [doi: 10.1177/0954407020950054]
43 Long Y, Gong YP, Xiao ZF, *et al*. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2486–2498. [doi: 10.1109/TGRS.2016.2645610]
44 Fu K, Chang ZH, Zhang Y, *et al*. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 161: 294–308. [doi: 10.1016/j.isprsjprs.2020.01.025]
45 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254–1259. [doi: 10.1109/34.730558]
46 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27: 1–9.
47 Wang XL, Shrivastava A, Gupta A. A-Fast-RCNN: Hard positive generation via adversary for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3039–3048.