计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):35−45 [doi: 10.15888/j.cnki.csa.008595]

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

深度神经网络图像目标检测算法综述①

付苗苗[1,2], 邓淼磊[1,2], 张德贤[1,2]

[1](河南工业大学 信息科学与工程学院, 郑州 450001)
[2](河南省粮食信息处理国际联合实验室, 郑州 450001)
通信作者: 张德贤, E-mail: zdx@haut.edu.cn

摘　要: 随着深度卷积神经网络优异的特征提取能力被发掘, 目标检测的进程开始以一种势不可挡的姿态向前推进, 同时, 和深度学习结合的目标检测技术取得了显著的成果, 在自动驾驶、智能化交通系统、无人机场景、军事目标检测和医学导航等现实场景中得到了广泛的应用. 本文回顾了传统目标检测算法的缺点, 介绍了常用的检测数据集以及性能评估指标, 综述了基于深度学习的目标检测经典算法, 阐述了当前目标检测的以及存在的困难与挑战, 对目标检测的未来可行的研究方向进行了展望.

关键词: 卷积神经网络; 特征提取; 深度学习; 目标检测; 计算机视觉

引用格式: 付苗苗,邓淼磊,张德贤.深度神经网络图像目标检测算法综述.计算机系统应用,2022,31(7):35–45. http://www.c-s-a.org.cn/1003-3254/8595.html

Survey on Deep Neural Network Image Target Detection Algorithms

FU Miao-Miao[1,2], DENG Miao-Lei[1,2], ZHANG De-Xian[1,2]

[1](College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China)
[2](Henan International Joint Laboratory of Grain Information Processing, Zhengzhou 450001, China)

Abstract: With the exploration of the excellent feature extraction capabilities of deep convolutional neural networks, target detection has made a great stride. At the same time, the target detection technology combined with deep learning has achieved remarkable results. It has been widely used in such real scenarios as automatic driving, intelligent transportation systems, drone scenarios, military target detection, and medical navigation. The study reviews the shortcomings of traditional target detection algorithms and introduces commonly used detection data sets and performance evaluation indicators. It also summarizes classic target detection algorithms based on deep learning and elaborates on current target detection and existing difficulties and challenges. The feasible research directions in the future are prospected.

Key words: convolutional neural network (CNN); feature extraction; deep learning; object detection; computer vision

目标检测是图像分类的自然延伸, 其目的只是识别图像中的目标. 目标检测的目的是检测预定义类的所有实例, 并通过轴对齐的框提供其在图像中的粗略定位. 检测器除了能够识别对象类的所有实例, 还要在其周围绘制边界框. 它通常被视为一个有监督的学习问题. 现代目标检测模型可以访问大量标记图像进行训练, 并在各种标准基准上进行评估. 目标检测主要是确定图像中哪些位置有对象, 这些对象分属什么类别. 对于人来说, 实现这样的任务轻而易举, 就算是几个月大的小孩也可以识别一些常见的物体, 然而对于计算机来说这几乎是一项不可能完成的任务. 尽管教计算机定位和识别视野内的实例对象很艰难, 但是在过去 20 年里, 我们还是通过一些方法教会了计算机对图像里的物体进行定位和识别, 在很多大型的图像数据集

① 收稿时间: 2021-10-14; 修改时间: 2021-11-08; 采用时间: 2021-11-30; csa 在线出版时间: 2022-05-31

上也取得了一些不错的结果.

1 传统目标检测

早期对目标检测进行研究时, 由于算力资源的紧张, 通常对数据进行复杂的特征设计, 使用各种机器学习算法对图像处理进行分功能设计, 然后进行联合训练. 传统算法对于物体的检测通常分为输入图像、区域选取、提取特征、分类器、后处理以及最终检测结果这 6 个阶段, 操作流程图如图 1 所示.

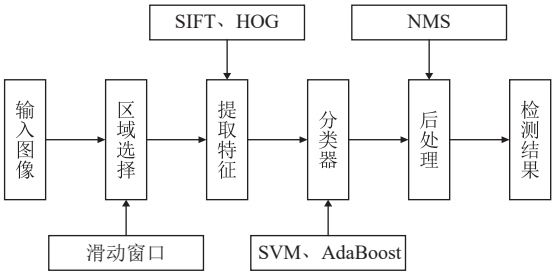

图 1 传统目标检测流程

首先对输入的图像进行区域选择, 主要使用滑动窗口 (sliding windows) 等方式对整个图像进行搜索, 目的是选择出可能含有目标的位置. 其次需要对挑选出来的目标可能存在区域进行特征提取, 主要通过 HOG (histogram of oriented gradient)、LBP (local binary pattern) 及 SIFT (scale-invariant feature transform) 等手工设计特征的方式, 目的是提取图像的纹理、尺度和空间变换特征. 然后使用支持向量机 SVM (support vector machine)[1]、AdaBoost 等分类算法对图像中提取的特征向量进行训练. 最后, 通过 NMS (non-maximum suppression)[2] 后处理算法基于指定阈值进行预测框筛选, 得到最终的检测结果.

传统目标检测的研究早在上世纪就已经开始展开, 1998 年 Papageorgiou 等人[3] 提出了针对静态图像处理从而推广到目标检测通用处理结构, 该结构直接从图像中学习 Haar 特征, 然后使用 SVM 对提取特征进行分类学习. 2001 年 Viola 等人提出 VJ 检测器[4], 首先对图像区域进行滑动窗口选择感兴趣区域, 之后提取每个窗口其中的 Haar 特征, 以积分图计算进行特征选择, 最后用级联方式训练分类器, 可以得到具有鲁棒性的实时人脸检测算法. 第一次在人脸检测上达到了实时检测效果 (没有任何约束条件), 在同等精度下其速度比任何其他算法都要快几十倍甚至几百倍.

2004 年 Lowe 等人设计了尺度不变特征交换 SIFT[5], 通过对图像区域进行梯度信息描述来得到特征, 该提取特征在面对光线变化、角度偏移及图像噪声等现象可以保持检测的鲁棒性. 2005 年 Dalal 等人设计了定向梯度直方图方法 HOG[6], 他们认为图像局部区域内物体可以被梯度信息来很好描述, 这样提取的特征可以得到不错的效果. 2010 年 Felzenszwalb 等人提出 DPM (deformable part model) 算法[7], 其将 HOG 和 SVM 相结合, 利用了多种优秀的手工特征组件进行联合训练的目标检测方法, 是传统机器学习目标检测算法的集大成之作.

上述传统的基于机器学习的目标检测算法有重大的缺陷, 第一, 需要对整个区域滑动窗口的方式暴力搜索会带来大量的无用边框, 其计算量大而且效率低; 第二, 复杂的特征提取设计, 针对性太明显, 局限性太大, 得到的特征不具有鲁棒性; 第三, SVM 等分类器多分类效果太低, 时间消耗太大, 不利于实际应用.

随着手工提取特征造成的检测算法性能饱和, 目标检测在此后一直处于“平原期”, 直到 2012 年的 AlexNet 网络[8] 的出现使得深度学习重新成为主流, 其在 ImageNet 图像分类任务中以“碾压”第二名算法的姿态取得了冠军. 由于深度卷积网络优异的特征学习能力, 能够抽取出图像中更高级别的语义特征, 于是人们思考能否将这种学习能力用到目标检测任务中. 直到 2014 年 Girshick 等人[9] 率先打破僵局, 提出了基于区域提议的卷积神经网络算法, 自此, 基于深度学习的目标检测开始以前所未有的速度向前推进. 并且随着 VGGNet[10]、ResNet[11] 和 GoogLeNet[12] 等优秀网络的接连问世, 在分类、物体检测、图像分割等领域渐渐地展现出深度学习的统治力, 大大超过了传统算法的水平. 图 2 梳理了目标检测算法发展的时间线, 深度学习时代的检测算法主要介绍两阶段和单阶段比较经典的一些模型.

2 模型评价指标

2.1 数据集

建立大型且具有较少偏差的数据集对于开发先进的目标检测算法至关重要, 近 10 年已经发布了许多目标检测相关的数据集, 如 Pascal VOC 07/12、ImageNet、COCO、Open-Image 等挑战赛中使用的数据集.

Pascal VOC 数据集[13] 始于 2005 年, 最初使用

4 个类别的图像数据来做分类和检测任务, 包含 1 578 张训练和测试图片, 这些图像标注了 2 209 个对象实例; 2007 年, 检测的类别增加到了 20 个类别, 自此类别的数量不再改变, 包含了 5k 的训练图像和 12k 的标记物体, 同时增加了分割和人体关键点检测等视觉任务; 2012 年将数据集扩大到 11k 个训练图像和超过 27k 的标记物体, 同时扩大了分割以及人体关键点检测两个任务的数据集大小. Pascal VOC 引入了平均精度 mAP (设置 IoU=0.5) 来评估模型的性能. 目前, VOC 07 和 VOC 12 主要用作新检测器的测试平台.

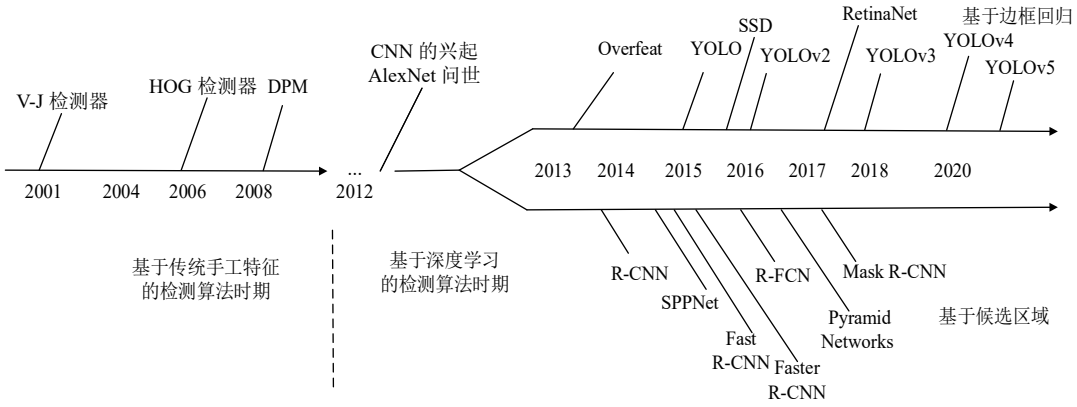

图 2 目标检测主要发展历程

ImageNet 是 ILSVRC 挑战赛[14] 的数据集, 该比赛每年都会从 ImageNet 数据集中抽出部分样本作为本次比赛的数据集, ImageNet 有 1 400 多万幅图像, 涵盖了 2 万多个类别, 可用于分类、检测、定位以及场景分类等视觉任务. 其中检测任务的数据集包含了 200 类视觉对象, 构成了超过 500k 的图像, 图像中所有类别的边界框都被标记了. ImageNet 数据集[15] 作为目前世界上最大的数据集, 推动了通用目标检测技术的发展; 另一方面, 过于庞大的数据集和类别数目增加了训练的计算量和检测难度.

COCO 数据集[16] 是目前可用最具有挑战性的数据集, COCO 数据集中的图片都是从复杂的日常生活中获取的, 具有小目标对象多、单个图像目标多等特点, 自 2015 年起每年都会有基于 COCO 数据集的比赛. 虽然该数据集的物体类别数比 ImageNet 少, 只有 80 个类别, 但是每幅图像拥有更多的目标实例, 如 COCO-17 包含 164k 的图像以及 897k 的标记样本 (来自 80 个类别). 和 VOC、ILSVRC 相比, COCO 最大的进展就是对每个目标采用 per-instance segmentation 方法来帮助更精确的定位. 此外, COCO 还包含了更多的小目标 (其面积可以不到整个图像的 1%), 定位目标更加密集, 目标分布更加接近于现实世界. 目前, COCO 数据集是检测领域的基准数据集.

Open Image[17] 是 OID (open images detection) 挑战赛使用的数据集, 该数据集主要用于两个任务: (1) 标准的目标检测; (2) 视觉关系检测, 即检测具有特定关系的成对的目标. 对于目标检测来说, 该数据集包含了 600 个物体类别以及 1 910k 张图像 (共有 15 540k 个标记边界框), 这使得它成为目前最大的目标定位数据集.

数据集是计算机视觉任务中必不可少的要素, 所有的任务都需要使用数据集训练模型, 数据集的优劣也对于模型的性能有着直接的影响, 因此我们在训练模型时对于数据集的选择也是至关重要的. 表 1 对不同的检测数据集进行了总结.

2.2 性能评价指标

目标检测使用多个准则来测量算法的性能: 检测速度 (frames per second, FPS)、精确度 (precision, P)、召回率 (recall, R)、交并比 (intersection over union, IoU)、平均精度 (average precision, AP)、平均召回率 (average recall, AR) 和平均精度均值 (mean average precision, mAP). 其中, mAP 一般作为分类准确率的度量, IoU 作为目标定位的度量值.

检测模型的性能主要是精度和速度两方面, mAP 值越大, 证明模型检测的精度越高; 速度则体现了模型的计算性能, 其 FPS 值越大, 说明模型检测速度越快, 更具有实时性. 一般情况下, 评估模型性能要结合平均精度均值 mAP 和检测速度 FPS 来判断[18], 未来的检测模型要朝着速度精度平衡这个目标努力.

表 1 不同检测数据集的比较

数据集	类别数	训练集	验证集	测试集	特点
PASCAL VOC 07	20	2 501	2 510	4 952	涵盖了生活中常见的20类物体, 每张图像中包含多个对象实例; 数据集中person类别的图像较多; 图像接近真实世界
PASCAL VOC 12	20	5 717	5 823	10 991	是PASCAL VOC 07的升级版, 拥有更多的图像数量, 标记了更多的实例对象
ImageNet	200	456 567	20 121	40 152	数据集类别数多, 涵盖了2万多个类别, 超过1 400多万幅图片, 是目前世界上最大的数据集; 数据集数量过于庞大, 增加了训练计算量
MS-COCO	80	118 287	5 000	40 670	图像类别数仅有80类; 主要来源于复杂的日常生活中; 图像中目标对象小、单个图像目标实例多; 使用像素级的实例标注; 目标分布更接近于真实世界
Open Image	600	1 743 042	41 620	125 436	约 900 万幅图像组成的数据集, 图像已经用图像级标签和目标边界框进行了注释; 是当前最大的带物体标注信息的数据集; 图像多样化, 通常包含有多个对象的复杂场景

3 基于深度学习的目标检测

深度神经网络的目标检测算法主要有两种类型, 即两阶段检测和单阶段检测, 这主要取决于是否存在候选框生成过程. 两阶段目标检测算法先使用多个固定大小的滑动窗口方法扫描整个图像, 产生一系列的候选框, 这个候选框集合同时筛选并消除掉大部分负样本集合, 然后对这些候选区域进行二次修正回归得到最终检测结果, 检测准确度较高, 但检测时效性一般; 单阶段检测算法则是将图像中的实例对象分配到提前划分好的单元格中, 每个单元格都有固定数量大小不一的锚框, 然后直接对这些锚框进行分类和位置修正, 不用单独生成 window 了. 检测速度基本可以满足实时性, 但没有起到筛选候选框的作用导致模型的精度十分低下. 因此当前目标检测致力于研究对这些主流算法的优化和改进, 以实现检测精度和检测速度的最佳平衡.

3.1 两阶段检测算法

3.1.1 R-CNN

作为 region with CNN (convolutional neural networks, CNN) 系列的开篇之作, R-CNN[9] 是 Girshick 等人在 2014 年提出的检测模型, 它使用与类无关的区域提议模块和 CNN 一起将检测任务转换为分类和定位问题. 该算法主要包含 3 个模块: (1) 使用 SS (selective research)[19] 在输入图像上提取 2 000 个候选区域; (2) 提取的候选区域经过裁剪送入 CNN产生固定长度的特征向量; (3) 使用 AlexNet 作为检测器的主干架构, 将提取的特征向量传送到经过训练的线性分类器 SVM, 为每个区域产生分数, 同时利用 IoU 计算和非极大抑制 NMS 来排除那些重叠位置的区域.

R-CNN 算法在 VOC 2012 上获得了 53.3% 的 mAP, 相比于 VOC 2012 上的最佳结果提升了 30% 的改进; 在 VOC 2007 上也有着很大改进, mAP 从 DPM 的 33.7% 提高到了 58.5%; 在 ILSVRC 2013 数据集上从 OverFeat[20] 的 24.3% 大幅提升至 31.4%.

尽管 R-CNN 在 mAP 方面有着明显提高, 但它的缺点也很明显: (1) 首先训练时对大量的候选区域分别做特征提取, 计算繁琐复杂, 且候选区域之间有着许多重叠部分, 产生的冗余计算更加导致图片的检测速度极慢 (47 fps); (2) 每张图片提取到的特征向量有成百上千张, 它们都需要被存入磁盘中. 以上这两点对于深度卷积神经网络 (deep convolutional neural networks, DCNN) 来说很不友好, 需要的时空代价过于昂贵.

3.1.2 SPP-Net

针对 R-CNN 分别对候选区域做特征提取而导致的冗余计算问题, 2015 年 He 等人提出的 SPP-Net[21] (特征空间金字塔网络) 采用了一次卷积方法来解决, 即一次性对整个图像做卷积操作提取特征图, 避免了 R-CNN 中重复计算卷积的操作, 大大减少了计算量, 提高了检测速度. 同时, SPP-Net 模型还解决了传统 CNN 提取特征图时要求输入图像大小固定的问题. 传统卷积网络调整图像尺寸的方式一般有 crop 和 wrap 这两种, 但是这样容易导致几何失真以及图像残缺, SPP-Net 模型通过在网络的最后一个卷积层和全连接层之间增加一个空间金字塔池化层 SPP (spatial pyramid pooling) 来输出固定尺寸的特征向量用于检测器的训练, 避免了对输入图像的裁剪和扭曲, 可以处理任意大小/纵横比的图像, 提高了模型对形变物体的鲁棒性.

SPP-Net 模型在 ILSVRC 2014 年的挑战赛目标检测任务中获得了第 2 名; 在 VOC 2017 上的 mAP 达到了 59.2%, 和 R-CNN 的精确度相当, 但速度却是 R-CNN 的 24-102 倍. 尽管 SPP-Net 的速度提高了很多, 但是仍然不能满足实时性的需求. 此外, 由于模型体系

结构类似于 R-CNN, SPP-Net 也有着 R-CNN 的缺点, 如多级管道训练、存储空间大、计算成本高、训练时间长, 且 SPP-Net 在微调时只更新了 SPP 层后的全连接层, 对深度网络的负面影响较大.

3.1.3 Fast R-CNN

为了进一步提高检测算法的性能, Girshick 等人在 2015 年提出了 Fast R-CNN[22] 模型, 该模型: (1) 提出了一个单段训练算法, 将 SVM 分类和 BBox (bounding box) 回归联合起来在 CNN 阶段训练, 解决了 R-CNN 和 SPP-Net 中存在的多级管道训练的问题; (2) 将 SPP 层简化成 RoI pooling 层, 利用 RoI 对特征图进行归一化操作, 然后分别输入到 Softmax 分类和 BBox 回归两个分支得到分类分数和回归偏移量; (3) 提出了 multi-task 损失, 训练时将分类和回归损失结合起来更新参数; (4) 在全连接层引入了截断奇异值分解 SVD (singular value decomposition), 将大的全连接层压缩, 从而减少在全连接层计算前向传播的时间.

Fast R-CNN 采用了 VGG16 模型, 在同样的环境配置下, 训练速度比 R-CNN 快 9 倍, 比 SPP-Net 快 3 倍, 测试速度比 R-CNN 快 213 倍, 比 SPP-Net 大约快 10 倍, 同时在 VOC 2012 数据集上可以达到 66% 的 mAP, 在 VOC 2007 上 mAP 可以达到 70%. 可以看出, Fast R-CNN 在提高速度的同时保证了精度.

虽然 Fast R-CNN 成功的集成了 R-CNN 和 SPP-Net 的优势, 检测速度也大幅度提高, 但是由于使用 Selective Search 生成区域提议, 其速度不可避免地受到了限制, 同年提出的 Faster R-CNN 解决了该问题.

3.1.4 Faster R-CNN

在 Fast R-CNN 问世不久后, Faster R-CNN[23] 被提出用于解决先前检测算法依赖区域提议算法生成候选框的问题. 之前的 Fast R-CNN 获得实时检测速度的前提是忽略在区域提议上的花费时间, 所以生成候选框集合的计算一直是两阶段检测算法速度不能大幅提升的瓶颈. Faster R-CNN 引入区域提议网络 RPN (region proposal network) 代替 Selective search 方法, 同时预测每个候选框的分类得分和回归偏移, 经过端到端的训练, 可以生成高质量、零代价的区域提议, 大大加快了模型的计算速度. 之前的算法都采用图像金字塔来解决输入图像大小变化的问题, RPN 引入了 anchor 的概念, 通过在 anchor 上使用多个尺度和多个宽高比来定义边界框, 然后对它们进行回归以便定位对象. 通过和检测阶段共享全局图像卷积特征, Faster R-CNN 将 RPN 和 Fast R-CNN 融合成一个统一的网络, 像特征提取、候选区域生成、分类以及定位等操作被集成到了一个学习框架中.

Faster R-CNN 使用深度网络模型 VGG16[10] 获得了 5fps 的速度 (包括所有步骤), 同时在 VOC 2012 上可以获得 70.4% 的 mAP, 在 VOC 2007 上获得 73.2% 的 mAP. 尽管 Faster R-CNN 突破了 Fast R-CNN 的速度瓶颈, 获得了接近实时检测速度, 但是仍存在一些问题: (1) 使用 anchor 机制设置的候选框有着固定的尺度和比例, 并不适用于所有目标, 特别是小目标的检测效果很不好; (2) 仍然使用 RoI Pooling 使得后续网络特征失去平移不变性, 影响最终定位精度[18]; (3) 在 RoI Pooling 中使用整数量化操作会产生一些误差, 这些误差在锚框映射回原图时被放大很多倍, 从而影响定位任务.

3.1.5 Mask R-CNN

Mask R-CNN[24] 是 He 等人在 2017 年提出的一个简单、灵活、通用的对象实例分割的框架, 并且很容易扩展到其他的视觉领域上, 如目标检测、人体关键点检测以及人体姿态估计等高级视觉任务. 该算法通过增加一个与现有的边界框回归和分类并行的 mask 分支来扩展 Faster R-CNN, 并将 3 个分支的损失联合起来训练. Faster R-CNN 的 RoI Pooling 中的两次整数量化操作引入了误差, 造成了特征图和原图区域不匹配, 该缺陷对于分类影响不大, 但是对检测和分割这样需要准确定位的视觉任务有着很大的负面影响. 因此 Mask R-CNN 使用了 RoIAlign 来改进这一缺陷. RoIAlign 取消了量化操作, 采用双线性差值方法计算非整数位置的像素值, 实现像素级对齐, 将掩码的准确率提升了 10% 到 50%.

Mask R-CNN 虽然在比 Faster R-CNN 增加了一点开销, 但在同样的环境下, 其运行速度可以达到 5fps, 且精度相比于 Faster R-CNN 提高了 2 个点. 但是其检测速度仍然不能满足实时性要求, 且实例分割标注的代价过于昂贵.

双阶段目标检测使用了级联结构, 采用“region proposal+CNN feature+SVM”的思路思想, 结合 CNN 网络, 大大提高了检测的精度; 后续的 SPP-Net、Fast R-CNN、Faster R-CNN 等大都沿用了这一思路, 在检测效率上进行改进[25], 但它们速度最多达到 5 fps, 就实时性而言略有不足. 表 2 总结了基于候选区域的检测算法的性能以及优缺点.

表 2 两阶段算法的性能比较

模型	骨干网络	数据集	FPS	Map (%)	优点	缺点
R-CNN	AlexNet	VOC 2007	0.03	58.5	提出region proposal模块, 将其与CNN结合, 性能比传统算法显著提高	特征提取的计算繁琐复杂; 固定的输入图像大小导致图像残缺; 时空开销大
SPP-Net	ZF-5	VOC 2007	2	59.2	采用一次卷积减少计算量; 增加SPP层避免候选区域归一化	多级管道训练、计算成本高、训练时间长、检测速度慢
Fast R-CNN	VGG 16	VOC 2007	3	70.0	同时分类和定位, 减少训练时间和特征存储空间	使用SS进行区域提取, 检测速度仍然受到限制
Faster R-CNN	VGG 16	VOC 2007	5	73.2	引入RPN网络产生区域提议; 真正的端到端训练; 加快运算速度	模型复杂; 空间量化粗糙, 对定位影响较大
Mask R-CNN	ResNeXt-101	VOC 2007	11	78.2	取消量化操作, 实现像素级对齐, 提升掩码准确率, 实例分割准确、检测精度更高	增加开销, 速度不能满足实时性, 分割标注的代价昂贵

3.2 单阶段检测算法

3.2.1 YOLO 系列

前面所述的两级检测器将目标检测作为一个分类问题来解决, 使用某个模块给出一些候选对象, 网络将其分类为目标或背景. 然而, YOLO (you only look once) 将其视为一个基于回归的检测问题, 在一个单独的网络中直接完成从图像输入到物体的定位和分类, 没有显示的提取候选框过程.

在 YOLO[26] 中, 输入图像被分成一个 $S\times S$ 网格, 网络将每个对象分配给该对象中心所在的单元. 一个网格单元预测多个边界框, 每个预测数组由 5 个元素组成: 边界框的中心 x 和 y, 框的维度 w 和 h, 以及置信度得分 (该边框包含目标的可能性). YOLO 的灵感来自图像分类的 GoogLeNet 模型, 该模型使用较小卷积网络的级联模块. 它在 ImageNet 数据上进行预训练, 直到模型达到高精度, 然后通过添加随机初始化的卷积和完全连接的层进行修改. 在训练时, 每个网格单元只预测一个类, 因为它收敛得更好, 但推理时间会增加. 提出多任务损失用于优化模型, 使用后处理 NMS 删除特定类别的多重检测. YOLO 在精确度和速度上都大大超过了当代的单阶段实时模型, 其处理速度可以达到 45 fps, 在较小模型的 Fast YOLO 甚至可以达到 155 fps. 然而, 它也有显著的缺点, 主要是对于较小对象和密集对象的定位精度较差, 以及每个单元所能预测对象数量的限制. 这些问题在 YOLO 的后续版本中得到了解决.

YOLOv2[27] 是对 YOLO 的改进, 在速度和精度之间提供了一个简单的平衡, 而 YOLOv2 又称为 YOLO9000, 闻名知意, 该模型可以实时预测 9000 个对象类. YOLOv2 用 Darknet-19 取代了 GoogLeNet 的主干架构, 同时结合了许多令人印象深刻的技术, 如: 批处理标准化 (batch normalization, BN)[28] 来提高收敛性, 分类和检测系统的联合训练来提高检测种类数量 (即用 COCO 中的物体检测标注数据进行定位训练, 用 ImageNet 中的数据学习对象分类), 去掉完全连接层以提高速度, 并使用 Fast R-CNN 中的 anchor 机制来提供更好的先验框和提高召回率. YOLOv2 在速度和精度上提供了更好的选择模型的灵活性, 新架构的参数更少, 正如论文标题所示, 它“更好、更快、更强”.

YOLOv3[29] 比之前的 YOLO 版本有“增量改进”. Redmon 等人用更大的 Darknet-53 网络取代了特征提取器网络. 同时还结合了各种技术, 如数据扩充、多尺度训练、批量标准化等. 分类器层的 Softmax 被二元交叉熵分类器取代. 尽管 YOLOv3 比 YOLOv2 快, 但与它的前身相比, 它缺乏任何突破性的变化. 它甚至比一个一年前最先进的检测器精度更低.

YOLOv4[30] 设计了一种快速且易于训练的物体探测器, 可以在现有的简单设备上工作. 它利用“免费包”, 即只增加训练时间而不影响推理时间的方法. YOLOv4 利用数据增强技术、正则化方法、类标签平滑、CIoU-loss[31]、交叉小批量归一化 CmBN (cross mini-batch normalization)、自对抗训练等技巧来改进训练. 网络中还加入了只影响推理时间的方法, 称为“特殊包”, 包括 Mish 激活[32]、跨阶段部分连接网络 CSPNet (cross-stage partial connections)[33]、SPP-Block[21]、PAN (path aggregation network)[34] 路径聚合块、多输入加权残差连接 MiWRC (multi-input weighted residual connections) 等. 它有一个 ImageNet 预先训练的 CSPNet-Darknet-53 骨干, SPP 和 PAN 块颈部和 YOLOv3 作为检测头. 大多数现有的检测算法需要多个 GPU 来训练模型, 但是 YOLOv4 可以很容易地在单个 GPU 上训练. 它的速度是同等性能的高效检测器的两倍. 这是最

先进的实时单级探测器.

3.2.2 SSD

SSD (single shot multibox detector)[35] 是第一个在保持实时速度的同时能与两级探测器 (如 Faster R-CNN) 的精度相匹配的单级检测器. SSD 是建立在 VGG-16 的基础上, 增加了辅助结构来提高性能. 当图像特征不太粗糙时, SSD 在网络中较早地检测到较小的对象, 而较深的层负责默认框和纵横比的偏移. 在训练过程中, SSD 将每个真实框与具有最佳 Jaccard 重叠的默认框相匹配, 并相应地训练网络, 类似于 Multibox[36]. 同时还使用硬负挖掘和大量数据扩充. 类似于 DPM, 它利用局部化和置信度损失的加权和来训练模型, 通过执行非最大抑制获得最终输出. 尽管 SSD 比 YOLO 和 R-CNN 等最先进的网络要快得多, 也更准确, 但它很难检测到小物体.

3.2.3 RetinaNet

鉴于单级检测器和两级检测器的精度之间的差异, Lin 等人认为单级检测器滞后的原因是“正负样本极端的不平衡”. 他们提出了一个重构的交叉熵损失 (focal loss), 称为焦点损失, 作为补救类不平衡的手段. 焦点损失的参数减少了来自 easy example 的损失贡献, 主要关注、集中训练 hard example. 2017 年 Lin 等人在一个简单的单级探测器中证明了该损失函数的有效性, 该检测器被称为 RetinaNet[37], 它通过对输入图像的位置、比例和纵横比进行密集采样来预测物体. 使用增加了特征金字塔网络 FPN[38] 的 ResNet 作为 backbone 以及两个相似的子网—分类和边框回归. FPN 的每一层都被传递给子网, 使其能够以不同的比例检测对象. 分类子网预测每个位置的对象分数, 而边框回归子网将每个锚点的偏移量回归到真实边框. 两个子网都是小型的全卷积网络 FCN, 同时两个子网共享参数. 与之前的检测模型不同, 该模型使用了一个与类无关的边界框回归器, 并发现该方法同样有效.

RetinaNet 训练简单, 收敛速度快, 易于实现. 与两级检测器相比, 它在精度和运行时间方面取得了更好的性能. 通过引入一个新的损失函数, RetinaNet 在推进物体探测器的优化方面也突破了极限.

单阶段算法的出现晚于两阶段, 其简单的结构、高效的计算优势可以满足实时性的需求, 收敛过程十分迅速, 其精度相比于双阶段算法较低一些, 但是随着视觉技术的发展, 单阶段检测框架的在精度方面有着显著的提升. 表 3 总结了单阶段算法的性能及其优缺点.

表 3 单阶段算法的性能比较

模型	骨干网络	数据集	FPS	Map (%)	优点	缺点
YOLO	VGG 16	VOC 2012	45	57.9	网络简单, 将图像划分成单元格预测对象, 处理速度快	小对象和密集对象的定位精度差, 单元格预测对象数量受限
SSD	VGG 16	VOC 2012	19.3	78.5	结合浅层和深层的特征图, 提升算法对多尺度目标的检测能力	小目标的检测效果依然不理想; 区域回归难度大, 模型难以收敛
YOLOv2	Darknet-19	VOC 2012	40	73.5	使用聚类方法产生先验框, 采用联合训练方法, 提高分类精度	对于重叠度高和小尺度目标检测仍需要改进; 使用预训练, 难迁移
YOLOv3	Darknet-53	MS-COCO	51	57.9	通过引入FPN进行多尺度预测; 使用大量残差模块的分类网络	缺乏突破性的变化, 模型的预测精度低
YOLOv4	CSPDarknet-53	MS-COCO	23	43.5	增加了许多优化技巧; 可以在单个GPU上训练; 速度和精度之间的平衡很优	检测精度有待提高
RetinaNet	ResNeXt-101+FPN	MS-COCO	5.4	40.8	引入Focal loss解决前背景类不平衡的问题	在密集样本训练时易造成样本不平衡

4 发展趋势及展望

目标检测是机器视觉领域的一个重要组成部分, 基于深度神经网络学习的目标检测依然存在一些困难和挑战. 如目标检测在行人检测方面, 存在着小目标 (图像中的行人占据的像素非常少, 如目标占整张图像的 5% 等)、行人密集以及行人遮挡等问题; 在人脸检测方面, 存在人脸拥有不同表情、目标被另一个目标遮挡、待检测目标拥有各种尺度等问题; 在文本检测方面, 存在字体和语言的多样化、目标文本损坏模糊、密集的文本排列等问题. 文献 [39] 也从目标尺度变化大、样本不均衡、弱监督检测以及实时检测等 4 个检测难点综述了最新的目标检测研究方法, 分析这些算法之间的关系, 并比较了其精度、优缺点以及适用场景.

尽管近年来目标检测研究领域取得了极大的进步, 目标检测距离实现达到人类检测识别水平仍然需要开展大量研究工作[40], 目标检测的未来可以关注但不仅

限于以下几个方面:

(1) 轻量级检测: 事实上, 许多应用程序都需要通过提高算法的速度才能在移动设备上持续稳定的执行, 如无人驾驶、智能相机、人脸识别、智能机器人等. 尽管近年来目标检测在模型性能上做出了极大的提升, 但是大多数模型网络的参数量巨大且目前嵌入式设备的计算能力有限, 因此难以在嵌入式设备上平稳运行[41]. 近几年也提出了许多轻量级的检测网络, 如实现模型压缩的 SqueezeNet[42]、使用深度可卷积分离的单阶段模型 MobileNet[43]、针对移动设备的 ShuffleNet[44] 等, 也出现了许多具有针对性的改进方法, 如文献 [45] 提出了基于特征融合的轻量级 SSD 检测方法, 文献 [46] 提出了基于轻量级注意力机制的人脸检测算法, 文献 [47] 提出了轻量级网络改进的实时人体关键点检测算法等等. 未来可以继续朝着轻量级模型优化的方向改进, 平衡模型的速度和精度, 使其在内存有限的移动设备上更加平稳快速的运行.

(2) 满足自动机器学习 (AutoML) 的检测: 现在基于深度神经网络的检测模型变得越来越复杂繁琐, 其建模与应用面临很大瓶颈和制约, 存在着严重依赖人为设计、建模周期长等问题. 国际上出现的自动化机器学习 (AutoML) 技术获得了国内外学术领域和工业领域的广泛关注, 它利用机器代替人工来自动化地完成模型选择和超参数调优, 让模型设计自动化[48], 如文献 [49] 介绍了 AutoML 使用工程化思维可以实现语义分割中的超参数优化、迁移学习与神经架构搜索等方法的自动化学习. 目标检测未来可能的方向是深度研究神经网络结构来设计智能检测模型从而减少人工对模型的干预, 例如: 如何根据图像自动选取合适的锚框和如何为模型选取好的优化方案. AutoML 会成为未来重要的对象检测技术.

(3) 弱监督学习的目标检测: 目前的主流检测算法都是基于强监督学习的, 严重依赖那些有着丰富标注的图像, 标注工作都是由人们手工完成, 但是复杂的背景和目标的多样性都使得标注阶段不仅消耗时间, 且效率十分低下, 而且像军事、航天这样的特殊领域难以获得大规模数据集, 使用基于强监督学习实现目标检测任务很难训练出检测效果好的模型. 目前, 数据集标注的成本越来越高, 如何在低成本标注的数据集上取得良好的检测结果已经成为研究的热点, 学者们研究了许多从不同角度实现弱监督检测, 如文献 [50] 中从特征处理方式的不同, 介绍了基于多示例学习、基于类激活图、基于注意力机制以及基于伪标签的弱监督目标检测算法; 文献 [51] 介绍了一些基于图像级别标签的弱监督算法, 从图像分割、多示例学习以及卷积神经网络等 3 个方面分析了目前的弱监督目标检测算法, 并比较了弱监督和强监督算法的性能. 总体来说, 近年来弱监督学习取得了不错的成绩, 但是相对于强监督学习依然有很大的发展空间. 就像文献 [50] 所展望的那样, 未来的弱监督学习可以继续朝着强弱监督学习相结合的检测算法进行研究.

(4) 小目标检测: 小目标检测是目前计算机视觉领域中的一个重难点问题, 有限的分辨率和特征信息使得小目标检测任务成为现阶段计算机视觉领域的一项巨大挑战[52]. 目前针对小目标检测算法出现了许多综述型文章, 如文献 [53] 简介了常见的两阶段、一阶段检测算法以及针对小目标检测时需要的专门数据集, 然后介绍了从多尺度方面改进的算法; 文献 [54] 综述了当前的一些小目标检测算法, 按照改进原理不同将其分成了 5 大类, 介绍了每种方法的典型模型并进行比较, 最后介绍了小目标检测常用的数据集. 从这些文章中看出来目前的小目标检测还是存在着一些问题, 对比大、中尺度目标的检测性能还存在着较大差距, 故如何实现小目标的精准识别、平衡其精度和速度仍然是一个极大的挑战, 小目标检测在自动驾驶、反恐刑侦、缺陷检测、智慧医疗等领域有着广泛的应用前景[55], 是未来研究的重要方向. 该研究方向的一些潜在应用还包括利用遥感图像进行军事状况的侦察以及通过无人机拍摄野生动物进行数量统计等任务, 未来可以关注视觉注意力机制的整合以及高分辨率轻量级网络的设计.

(5) 视频目标检测: 与静态的图像相比, 视频具有高度的冗余性, 包含了大量的时空局部信息, 其动态性给视频检测任务带来了极大的困难. 传统的检测模型大多都是为了检测静态的图像, 恰好忽视了动态视频中帧与帧之间的相关性, 无法完全利用视频的全局信息. 针对视频特有的性质, 提出了许多视频检测的算法, 文献 [56] 从视频检测面临的 3 个技术方面的挑战 (改进与优化、保持时空序列一致性以及模型轻量化) 出发, 介绍了当前的视频检测算法, 包含了基于运动信息、基于检测和跟踪结合、轻量化视频检测以及跨界模型的使用 (如将自然语言处理领域的 Transformer 和

视频检测相结合) 等 4 种类型; 文献 [57] 研究了基于时序特性的检测方法, 结合特征融合和双模型对视频进行逐帧检测, 通过前一个帧的结果反馈修正当前帧的检测结果从而提高帧之间的连续性, 提升了检测准确率和视频连续性. 现如今, 视频检测有着非常广泛的应用, 比如高清摄像头中的实时目标检测和跟踪对于视频监控、自动驾驶以及车载视频中的障碍物检测都有着重大意义. 未来视频检测可以朝着泛化能力方向进行研究, 比如让模型更适应真实场景的检测要求, 也可以尝试和弱监督学习结合起来, 研究如何在少样本或者零样本的条件下实现高精度的视频检测.

参考文献

1 Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121–167. [doi: 10.1023/A:1009715923555]

2 Neubeck A, Van Gool L. Efficient non-maximum suppression. 18th International Conference on Pattern Recognition (ICPR'06). Hong Kong: IEEE, 2006. 850–855.

3 Papageorgiou CP, Oren M, Poggio T. A general framework for object detection. 6th International Conference on Computer Vision (IEEE Cat. No. 98CH36271). Bombay: IEEE, 1998. 555–562.

4 Viola P, Jones MJ. Robust real-time object detection. International Journal of Computer Vision, 2001, 57: 137–154.

5 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

6 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.

7 Felzenszwalb PF, Girshick RB, McAllester D, *et al*. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645. [doi: 10.1109/TPAMI.2009.167]

8 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, 2012. 1097–1105.

9 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.

10 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

11 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

12 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.

13 Everingham M, van Gool L, Williams CKI, *et al*. The Pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

14 Russakovsky O, Deng J, Su H, *et al*. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

15 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.

16 Lin TY, Maire M, Belongie S, *et al*. Microsoft coco: Common objects in context. 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

17 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]

18 许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述. 计算机工程与应用, 2021, 57(8): 10–25. [doi: 10.3778/j.issn.1002-8331.2012-0449]

19 Uijlings JRR, Van De Sande KEA, Gevers T, *et al*. Selective search for object recognition. International Journal of Computer Vision, 2013, 104(2): 154–171. [doi: 10.1007/s11263-013-0620-5]

20 Sermanet P, Eigen D, Zhang X, *et al*. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv: 1312.6229, 2013.

21 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]

22 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.
23 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015. 91–99.
24 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2980–2988.
25 方路平, 何杭江, 周国民. 目标检测算法研究综述. 计算机工程与应用, 2018, 54(13): 11–18, 33. [doi: 10.3778/j.issn.1002-8331.1804-0167]
26 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.
27 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6517–6525.
28 He KM, Zhang XY, Ren SQ, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1026–1034.
29 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.
30 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.
31 Zheng ZH, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993–13000. [doi: 10.1609/aaai.v34i07.6999]
32 Misra D. Mish: A self regularized non-monotonic activation function. arXiv: 1908.08681, 2019.
33 Wang CY, Liao HYM, Wu YH, *et al*. CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 390–391.
34 Liu S, Qi L, Qin HF, *et al*. Path aggregation network for instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768.
35 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multiBox detector. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.
36 Erhan D, Szegedy C, Toshev A, *et al*. Scalable object detection using deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 2155–2162.
37 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.
38 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.
39 罗会兰, 彭珊, 陈鸿坤. 目标检测难点问题最新研究进展综述. 计算机工程与应用, 2021, 57(5): 36–46. [doi: 10.3778/j.issn.1002-8331.2011-0205]
40 张泽苗, 霍欢, 赵逢禹. 深层卷积神经网络的目标检测算法综述. 小型微型计算机系统, 2019, 40(9): 1825–1831. [doi: 10.3969/j.issn.1000-1220.2019.09.004]
41 齐榕, 贾瑞生, 徐志峰, 等. 基于 YOLOv3 的轻量级目标检测网络. 计算机应用与软件, 2020, 37(10): 208–213. [doi: 10.3969/j.issn.1000-386x.2020.10.033]
42 Iandola FN, Han S, Moskewicz MW, *et al*. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv: 1602.07360, 2016.
43 Howard AG, Zhu ML, Chen B, *et al*. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.
44 Zhang XY, Zhou XY, Lin MX, *et al*. Shufflenet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.
45 吴天成, 王晓荃, 蔡艺军, 等. 基于特征融合的轻量级 SSD 目标检测方法. 液晶与显示, 2021, 36(10): 1437–1444.
46 高刘雅, 孙冬, 卢一相. 基于轻量级注意机制的人脸检测算法. 激光与光电子学进展, 2021, 58(2): 130–138.
47 胡江颢, 王红雨, 乔文超, 等. 基于轻量级网络的实时人体关键点检测算法. 计算机工程, 2021, 47(4): 218–225.
48 方鑫. 面向典型场景的自动化机器学习算法研究及系统实现 [硕士学位论文]. 南京: 南京大学, 2020.
49 刘桂雄, 黄坚, 刘思洋, 等. 面向语义分割机器视觉的 AutoML 方法. 激光杂志, 2019, 40(6): 1–9.
50 杨辉, 权冀川, 梁新宇, 等. 基于弱监督学习的目标检测

研究进展. 计算机工程与应用, 2021, 57(16): 40–49. [doi: 10.3778/j.issn.1002-8331.2103-0306]
51 周小龙, 陈小佳, 陈胜勇, 等. 弱监督学习下的目标检测算法综述. 计算机科学, 2019, 46(11): 49–57. [doi: 10.11896/jsjkx.181001899]
52 刘洋, 战荫伟. 基于深度学习的小目标检测算法综述. 计算机工程与应用, 2021, 57(2): 37–48. [doi: 10.3778/j.issn.1002-8331.2009-0047]
53 刘晓楠, 王正平, 贺云涛, 等. 基于深度学习的小目标检测研究综述. 战术导弹技术, 2019, (1): 100–107.
54 员娇娇, 胡永利, 孙艳丰, 等. 基于深度学习的小目标检测方法综述. 北京工业大学学报, 2021, 47(3): 293–302.
55 刘颖, 刘红燕, 范九伦, 等. 基于深度学习的小目标检测研究与应用综述. 电子学报, 2020, 48(3): 590–601. [doi: 10.3969/j.issn.0372-2112.2020.03.024]
56 王迪聪, 白晨帅, 邬开俊. 基于深度学习的视频目标检测综述. 计算机科学与探索, 2021, 15(9): 1563–1577. [doi: 10.3778/j.issn.1673-9418.2103107]
57 周念. 基于时序特性的视频目标检测研究. 中国电子科学研究院学报, 2021, 16(2): 157–164. [doi: 10.3969/j.issn.1673-5692.2021.02.009]

(校对责编: 牛欣悦)