基于深度学习的目标检测模型综述

翁成康*，张龙信

湖南工业大学计算机学院，湖南株洲 412007

摘要： 目标检测通过在图像或视频中对物体出现的具体位置进行准确高效的定位以识别出待检测的实例，是计算机视觉领域中需要解决的经典问题。其应用范围广泛，包括人脸检测、智能驾驶辅助、卫星遥感检测等重要领域。本综述旨在帮助研究人员快速了解目标检测，从目标检测算法的概念出发，对目标检测算法的发展经历进行分析论述，详细地阐述了目标检测算法从独自发展到与深度学习技术相结合的进化历程。本文根据目标检测由检测过程是否产生锚点框进行划分，并从基于锚点框、无锚点框类型出发，分析了单阶段、两阶段的目标检测算法的研究现状，详细归纳了目标检测算法发展历程中的经典模型结构，并介绍了目标检测中的难点，即解决在训练过程中不带有详细标注的少样本情况下目标检测问题的思路。最后，本文总结对比各种经典模型的优缺点，对主流数据集及评价指标等进行全面归纳，并对目标检测领域现今所面临的挑战和未来的发展方向进行了展望。

关键词： 计算机视觉；深度学习；目标检测

DOI: 10.57237/j.cst.2023.02.006

A Survey of Object Detection Models Based on Deep Learning

Weng Chengkang*, Zhang Longxin

College of Computer Science, Hunan University of Technology, Zhuzhou 412007, China

Abstract: Object detection is a crucial problem in computer vision that involves accurately locating objects in images or videos to identify instances to be detected. It has various applications, including facial detection, intelligent driving assistance, and satellite remote sensing detection. This review aims to aid researchers in quickly comprehending object detection. It covers the concept of object detection algorithms, analyzes the development history of object detection algorithms, and elaborates on the evolution process of object detection algorithms from independent development to combination with deep learning technology. The article divides object detection based on whether anchor boxes are generated during the detection process. It discusses the types of anchor boxes and non-anchor boxes and analyzes the current research status of single-stage and two-stage object detection algorithms. It also summarizes classic model structures in the development process of object detection algorithms and introduces difficulties in object detection. The article aims to solve the problem of object detection with a small number of samples and without detailed annotations during the training process. It summarizes and compares the advantages and disadvantages of various classic models, mainstream datasets, and evaluation indicators. Additionally, it looks forward to current challenges and future development directions in the field of object detection.

Keywords: Computer Vision; Deep Learning; Object Detection

*通信作者：翁成康, chengkang_weng@163.com

收稿日期: 2023-03-02；接受日期: 2023-04-12；在线出版日期: 2023-04-25

1 引言

在人工智能领域的发展道路上，使用计算机来模拟生物视觉感知的计算机视觉是目前非常活跃的研究方向，它落地应用时间早，成果广，迭代快，可以说，计算机视觉已经融入人类社会的方方面面，改变了我们的生活方式。而目标检测问题可以说是计算机视觉研究金字塔中最底层的基础领域，在信息化的今天具有巨大的实用价值和应用前景。因此，目标检测技术受到了学术研究、工业生产等领域的广泛关注。在学术界，每年有大量的相关论文发表。在工业生产方面，目标检测也广泛应用于提高生产效率，减轻人力负担等方面。在此背景下，本文对基于深度学习的目标检测技术进行了发展梳理和基本总结。

2 概念与发展历程

计算机视觉（Computer Vision，CV）将生物对客观世界的感知、识别和理解等用计算机实现，其主要任务包括分类，定位，分割及检测。分类用于解决“What”的问题，即给定一张图或一段视频，对其进行一个整体的判断，关注在其中的物体包括哪些类别。定位用于解决“Where”的问题，即对图片中待检测物体进行定位，以边界框（Bounding Box）将物体的几何位置确定出来。检测则是对分类和定位做一个总体上的解决，即在图像上定位出目标坐标位置的同时判断识别目标的具体种类。分割的目的比检测更进一步，需要从像素层面逐个分辨每一个像素属于哪个目标，将图片按像素级别进行区分，对不同的语义进行分组、分割，根据任务的不同又可以细化分为语义分割和实例分割。

目标检测算法在 1991 年由 Turk M A 等[1]提出的人脸空间识别算法就出现了原型，如图 1 所示，主要使用手动构造特征的计算方法，选择输入图像的区域，使用滑动窗口方法（设置不同大小的窗口，通过窗口搜索整个图像）遍历穷举出所有可能的位置，对目标可能存在的区域进行视觉特征提取，该步骤主要通过是 SIFT [2]、HOG [3]等特征提取算子提取图像的纹理、尺度和空间变换等特定的特征。再通过支持向量机[4] （Support Vector Machine, SVM）等常用的分类器对图像中提取的特征向量进行分类，这其中 SVM 是使用较为广泛的分类器之一，它在图像分类上有着比较优越的性能。最后选取一个合适的阈值，使用非极大值抑制算法[5]（Non-maximum Suppression, NMS）筛选得到结果并输出。

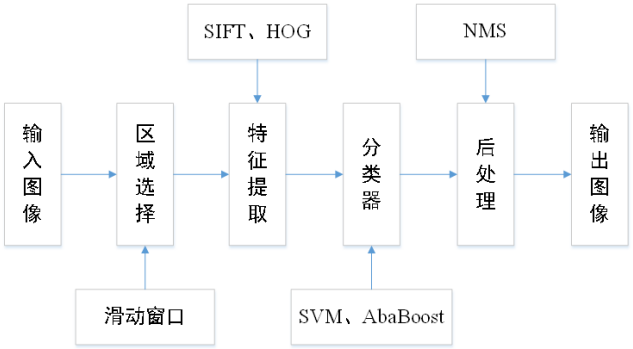

图 1 传统目标检测方法流程

基于手动设计特征的传统检测算法存在许多缺陷。第一，在区域选择阶段，通过滑动窗口对全图进行搜索会带来巨大数量的边框，这些边框中只有少数是真正有用的，大部分都是浪费系统计算量资源的产物，极大的拉低了整体识别的效率，影响了整个检测过程。第二，特征提取设计针对性明显，泛化性差。第三，SVM、AdaBoost [6]等分类器多分类的准确率低、耗时大、不利于实际应用。因此，传统方法已经难以满足高性能目标检测需求。

从 2012 年 Krizhevsky 等[7]提出 AlexNet 网络并在 ImageNet 大赛中取得冠军开始，将目标检测通过卷积神经网络（Convolutional Neural Networks，CNN）[8]实现就是应用十分广泛的热点领域，大量论文相继发表。基于深度学习的目标检测算法，解决了传统方法的弊端。以前所未有的速度向前推进，在短时间内超过了传统方法的水平。

3 基于深度学习的目标检测

基于深度学习的目标检测技术极大推进了目标检测的落地应用，其根据检测过程是否产生 Anchor（锚框），以及训练过程中带详细标注的基类是否充分可以被分为三类，即 Anchor-based 目标检测算法、Anchor-free 目标检测算法与少样本目标检测算法。接下来本文将从这几个方面分别来介绍基于深度学习的目标视觉算法的研究现状。

3.1 Anchor-based 检测算法

Anchor-based 目标检测是一种基于固定形状锚框的检测方法，通过在图像上放置多个锚框，用深度卷积神经网络学习这些锚框和物体的关系，从而实现检测。它的优点是可以快速检测物体，缺点是固定大小

的锚框会影响到检测效果。其又可以细分为基于区域与基于回归两类。

基于区域的目标检测算法在实现分类与回归时包括两个阶段，一阶段通过得到建议框来找出目标出现的可能位置，负责保证足够的准确率和召回率；二阶段将之前产生建议框分类，使预测位置更加精准。故又称两阶段目标检测（Two-stage）。该方案通常精度较高，但要求更多的计算资源，速度较慢。

基于回归的算法整合了特征提取、目标分类和位置回归，在一个阶段中完成目标的分类与定位，直接产生目标类别概率和位置，故又称单阶段目标检测（One-stage）。速度一般比两阶段检测算法更快，但精度有所损失。

3.1.1 基于区域的目标检测

Girshick 等首次将 CNN 应用于目标检测领域，提出了开创性的 R-CNN [9]。R-CNN 模型如图 2 所示，首先输入图片，通过选择性搜索算法[10]（Selective Search）提取建议区域，这些区域可能包含要检测的目标。然后裁剪所有建议区域，固定其大小，再使用 CNN 提取特征，经过卷积层（Conv）和全连接层（Fc），得到特征向量。类别信息需要使用支持向量机线性分类特征向量得到。由于存在一些高度重叠的建议区域，作者采用 IoU 计算和非极大值抑制算法来舍弃那些非目标部分。最后目标预测的结果通常采用边界框回归[11]（Bounding Box Regression）完成。

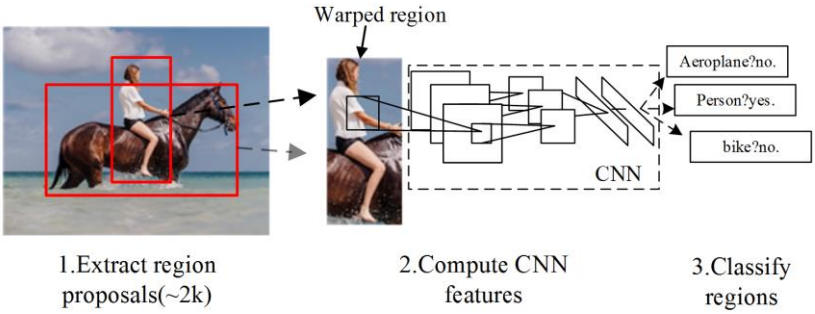

图 2 R-CNN 模型流程图

R-CNN 算法设计固然十分精巧，但是在提取候选区域时将多个尺寸的候选框直接裁切为同尺寸，丢失了大量原图信息。将检测过程分为多个阶段，冗余了大量的建议框与特征，导致了算力浪费，检测时间过长，效率低下。Girshick 注意到这些缺点，于 2015 年提出了改进的端到端联合训练算法，即 Fast R-CNN [12]算法，模型结构如图 3 所示。

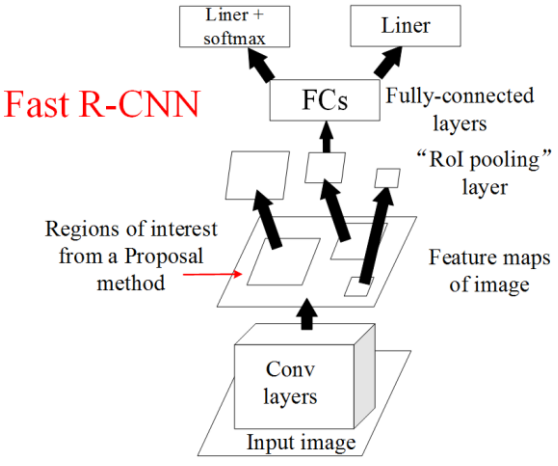

图 3 Fast R-CNN 模型结构图

输入图像后首先通过选择性搜索算法提取感兴趣区域（Regions of Interest, RoI）并送入卷积层，RoI 池化层（RoI Pooling）的作用是将每个 RoI 分成均匀的若干块，并将该区域的最大值赋给每个小块；池化后的特征送入全连接层，输出结果两个向量。输出结果的其中一个进行 Softmax [13]分类，另一个则负责回归边框。

改进后的 Fast R-CNN 算法通过分类和边框回归的同步训练降低了复杂度，进而提高速度。但由于 Fast R-CNN 使用 Selective Search 算法会消耗大量计算资源，Ren S 等[14]的新算法 Faster R-CNN 选择使用区域建议网络（Region proposal network, RPN）来生成建议区域。

新算法极大的提高了区域建议速度，其结构如图 4 所示。主要由卷积层，RPN 网络，RoI 池化层以及分类回归四个部分组成。卷积层用于提取图片特征。RPN 网络用于推荐候选区域，代替之前的选择性搜索算法，输入图片，输出多个候选区域。池化层将大小各异的 RoI 规整化输出。最终由分类和回归层的输出目标所属的类和精确位置。

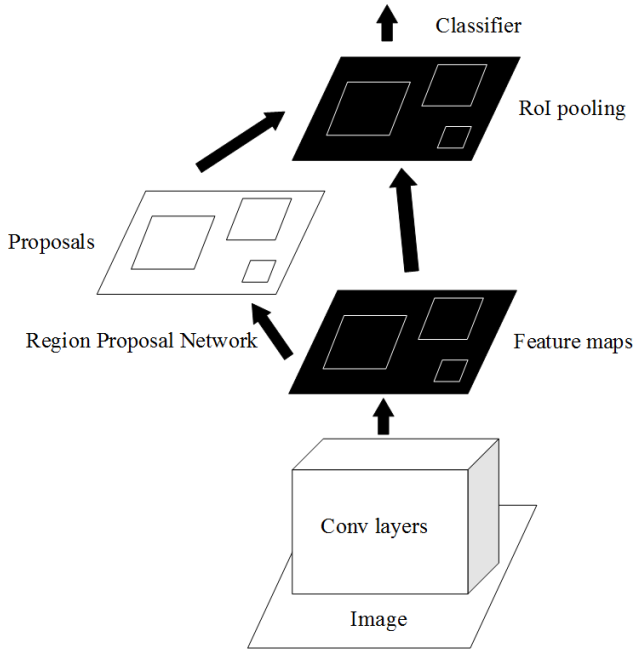

图 4 Faster R-CNN 模型结构图

Faster R-CNN 虽然具有较高的检测精度，但由于池化层存在的取整近似计算操作丢失了网络需要的平移不变性，从而导致检测结构和被提取的特征不统一，且由于 Anchor 机制有着固定的大小比例，其预测功能只在工作在单个尺度，因此相对来说不利于检测较小的目标。

3.1.2 基于回归的目标检测

针对上一小节中提到的各种问题，Redmon 等[15]于 2015 年提出了 YOLO 算法。它极大的简化了网络结构，舍去了算法中的候选框提取过程，将特征提取、边框分类和回归统一为一个无分支结构。如图 5 所示。

首先输入图像，将其缩放为固定大小，并按照 S×S 的标准划分为网格。这其中每个网格负责预测 B 个边框，除了要回归自身的位置外，还要预测一个置信度，即包含目标的可信程度，它同时也反映了预测的准确度。

除此之外每个网格还要预测 C 个类型的条件概率。作者将预测结果编码为一个 S×S×(B×5+C)维的张量。其中，S 等于 7，B 等于 2，5 代表边界框的中心坐标 x 和 y、框的宽高 w 和 h、置信度得分，C 是数据集的类别个数，即 20。经过数次卷积与池化以及两个全连接层，将输出目标分类和边界框。最后通过极大值抑制算法剔除重叠边界框，实现检测。

与上一小节提到的传统手动特征方法或区域建议方法不同，YOLO 在进行训练和测试期间工作在整个图像层面，进行全局推理。与此相对的是，Fast R-CNN 则会将图像中的背景补丁误认为是对象，因为它无法看到更广阔的上下文信息。可以说，YOLO 是目标检测算法引入深度学习技术后诞生的第一个单阶段算法，自此目标检测算法首次有了单阶段、两阶段之分。

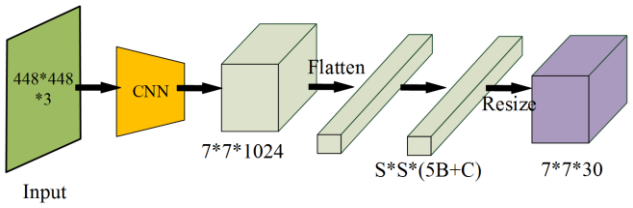

图 5 YOLO 工作流程图

虽然在速度上 YOLO 取得了很大的优势，但由于网格的长宽是根据训练集手动设定的，因此对于小目标的检测性能较差。并且由于网格数量恒定，每个单元格能预测的数量受到限制，因此对密集目标的检测性能也不太好。针对上述问题，Redmon 等提出了速度与精度之间更加平衡的算法，即 YOLO9000 [16]。

在骨干网络上，YOLOV2 用 Darknet-19 取代了 GoogleNet [17]，使用高分辨率的 ImageNet [36]数据集训练分类网络，再结合 COCO [37]数据集进行定位训练，以此来提高检测的种类数与模型训练的稳定性。为了保证数据的分布相对稳定，作者在每一个卷积层后引入了批标准化层[18]（Batch Normalization, BN），并引入了 Anchor 机制，通过 k-means [19]算法聚类得到 Anchor 的宽高，提高模型的泛化性能，在保证高分类准确率的同时提高了召回率和准确度。

YOLOV3 [20]是 YOLOV2 的增量改进模型，骨干网络采用 Darknet-53，变深的骨干网络可以更好的拟合特征，更好的学习，每一层要做的事情也更简单，但也会造成过拟合以及梯度不稳定，最终导致网络退化。这可以缓解，但无法消除。

常用的缓解手段包括随机丢弃神经元、批归一化等。但是在层数过高的情况下，这些优化的能力有限，于是 Redmon 等在骨干网络中添加了可以不通过卷积，转而直接从前一个特征层映射到后面的特征层的残差连接结构，网络结构如图 6 所示。

V3 在延用 V2 一些操作的同时还结合了许多创新点，包括采用 K-means 聚类算法对锚框进行聚类操作；分类使用 Logistic 激活函数替代 Softmax 分类层，更有效地进行数据拟合等。尽管 V3 做出了许多改进，但仍缺乏突破。

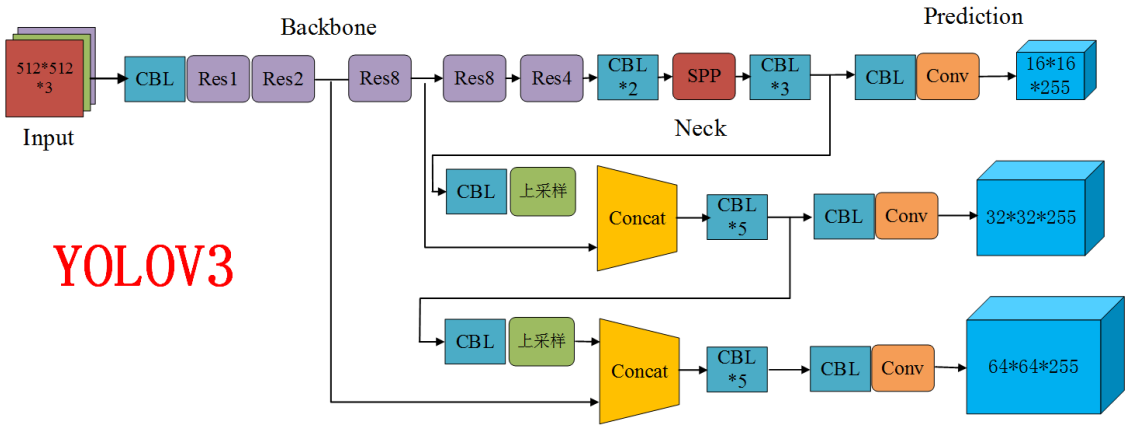

图 6 YOLOV3 网络结构图

于是 Bochkovskiy 等[21]结合了当时的一些目标检测算法中的优点技巧，在骨干网络上，将 YOLOV3 选用的骨干网络 Darknet53 修改为 CSP-Darknet53 来进行特征提取。激活函数也由 Leaky_relu 换成了 Mish。在特征金字塔部分，YOLOV4 使用了 SPP [22]与 PANet [23]的结构替代了 YOLOV3 中的 FPN [24]，进行特征融合。SPP 结构参杂在对 CSPdarknet53 的最后一个特征层的卷积里，在卷积后分别利用四个不同尺度的最大池化进行处理，池化核大小分别为 13x13、9x9、5x5、1x1。这个结构既发挥了深层网络的语义，特化，抽象特征，也充分利用了浅层网络的细粒度底层特征，实现多尺度的特征融合与物体预测。这些修改令新模型在提高检测精确度和速度的同时仍然保持着较低的训练成本。

3.2 Anchor-free 检测算法

Anchor-free 目标检测算法是一种基于深度学习的无锚框的检测方法，它不需要使用预先设定好的固定的锚框，而是通过学习图像中物体的形状特征，直接计算物体的边界框，从而达到物体检测的目的。它的优点是解决了传统目标检测算法锚框设定不合理导致的检测效果不佳的问题，也适合检测不同大小的物体。缺点是训练时间较长，而且检测的速度较慢。下面介绍 Anchor-free 目标检测领域的几个优秀模型。

Ge 等[25]提出了以 YOLOV3 作为基线改进的 YOLOX，新框架剔除了 Anchor 操作，降低模型计算量，缓解正负样本不平衡的问题。同时引入预测分支解耦头（Decoupled Head），如图 7 所示[26]。

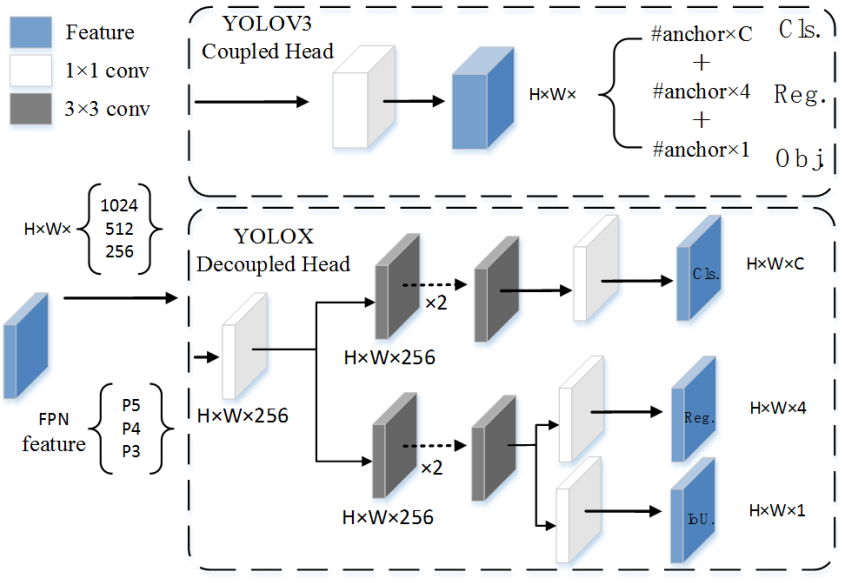

图 7 解耦探测头示意图

新的解耦头在很大程度上改善了训练的收敛速度，对于每个级别的特征，首先通过 1×1 卷积将降维特征图，然后添加两个具有 3×3 个卷积层的并行分支，分别用于分类和回归任务，同时回归分支里还添加了 IoU 分支，丰富检测头的表达。

2018 年，Law 等[27]提出了 CornerNet，它是一种端到端的目标检测算法，旨在将目标检测中检测定位任务转化为检测框的角点，即利用一对关键点（边界框的左上角和右下角）实现目标的定位。算法框架如图 8 所示。沙漏网络[28]（Hourglass Network）是 CornerNet 选取的骨干网络，从图片可以看出，该骨干网络是对称的，其特点为通过卷积和池化将下采样特征图，此时的特征图尺度缩小很多，再用最近邻上采样方法上采样小尺度的特征图，将其还原为开始时的尺度。

并且在每个上采样层都有一个 Skip Connection [29]，其上是一个残差模块。使用这种沙漏结构的目的是为了反复获取高低分辨率下所包含的多尺度特征信息。最后通过上采样组合局部和全局信息。在其后紧跟两个预测模块，其中一个负责检测上左角，另一个负责下右角。通过对检测到的两组角进行筛选、修正得到物体的真实对角，从而定位目标。

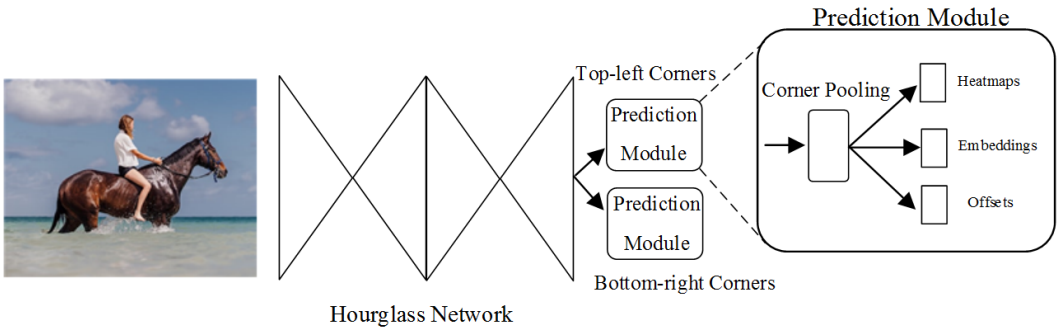

图 8 CornerNet 算法模型图

CornerNet 算法摒弃了对 Anchor 的需求，降低了网络的复杂度，同时通过全新的角点池化（Corner Pooling）方法，更好地对物体对角进行定位，提高了检测精度。但其只利用一对对角生成物体边框，没有充分利用好物体区域的内部信息。于是 Duan 等[30]对其进行了改良，在 2019 年提出了 CenterNet 检测算法。

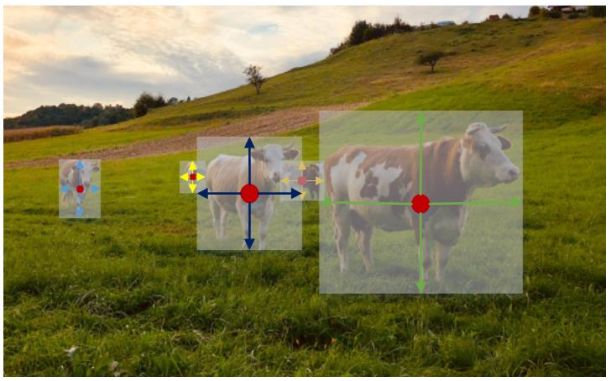

图 9 CenterNet 检测示意图

CenterNet 将目标看作一个点，一个目标由一个特征点确定，如图 9 所示。模型将输入进来的图片划分成若干个区域，每个区域都存在一个特征点。当物体的中心落在这个区域，则保留该预测框，由这个区域的特征点来确定这个物体的种类，如果预测框与真实框的 IOU 较大，则判断该框质量较高，应该包含中心关键点。同时还会对这个特征点进行调整，获得物体中心的坐标，回归预测出物体的宽高。

3.3 少样本目标检测算法

少样本目标检测（FSOD）是目前目标检测领域中的一个重难点问题，主要关注在数据集中有限的特征信息下，目标的样本数量远远少于其他类别样本时，如何提高对该类型目标的检测能力的问题。

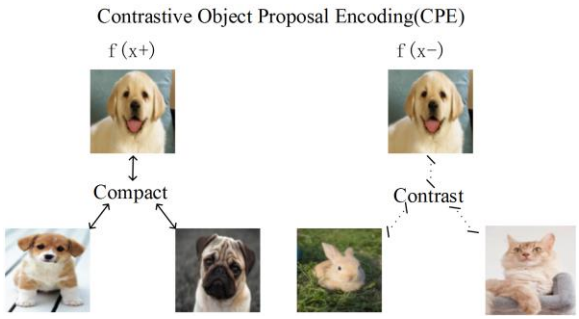

图 10 得分函数示意图

Bo Sun 等[31]的研究展现了一个基于对比方案编码的少样本目标检测（FSCE）模型，作者引入了一个得分函数来度量区域建议之间的语义相似度。如图 10 所示。积极提案（X+）是指来自同一类别或同一对象的区域提案。否定提案（X-）是指来自不同类别的提案。对“正对”和“负对”施加了对比嵌入损失，使“正对”的得分远大于“负对”的得分。这样对比学习对象建议就能具有较小的类内方差和较大的类间差异。

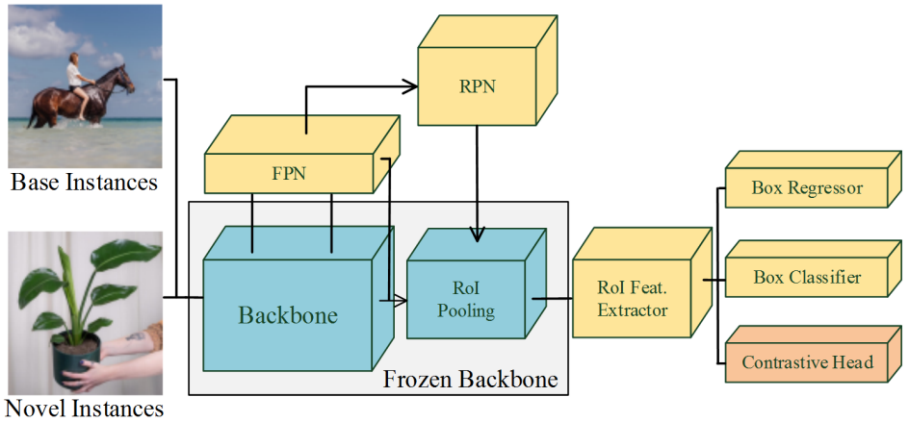

图 11 FSCE 结构示意图

FSCE 模型的结构如图 11 所示，其包含一个两阶段训练。首先在基类上训练基础的 Faster R-CNN，然后在基类和新类上训练微调后的检测器。在模型的 Head 增加了一个对比分支，这个分支附带了一个损失函数，用来将同一类别的实例凑的更“近”，而不同类别尽可能“远”。

Hu 等[32]的研究提出了一种基于上下文感知聚合的密集关系提取目标检测模型，如图 12 所示。以往的小样本目标检测工作没有充分利用支持特征和查询特征之间的关系，一般是利用支持特征全局池化生成的分类向量来调整查询特征，从整体角度指导特征学习。然而，由于外观变化或遮挡在自然图像中是常见的，当同一类对象在查询和支撑样本之间变化很大时，整体特征可能会产生误导，而以往的方法完全忽略了这一点。同时多尺度特征提取器中存在基类和新类过拟合的问题。为了解决上述问题，提出了基于上下文感知聚合的密集关系蒸馏算法（DCNet），包括两个新的模块，即 DRD 密集关系蒸馏模块和 CFA 上下文感知聚合模块。

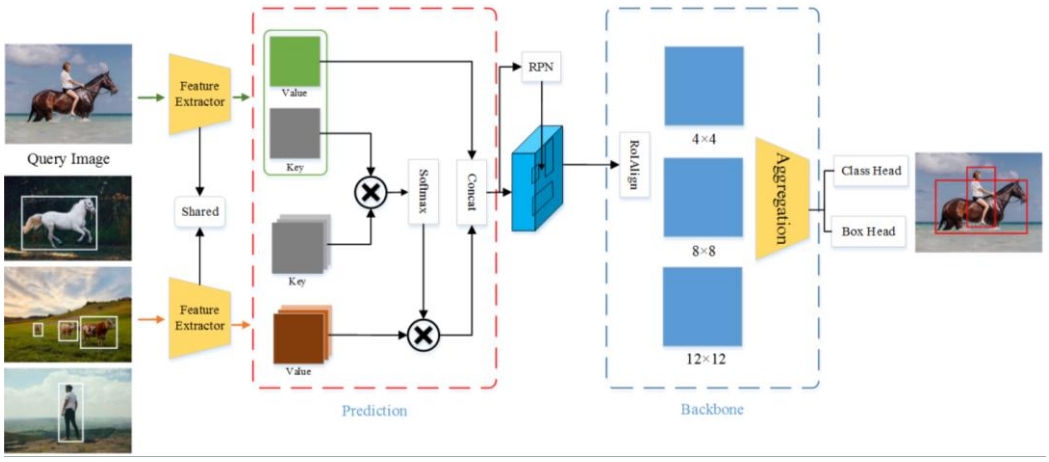

图 12 DCNet 模型示意图

Zhang 等[33]的工作将相关聚合方法引入到 DETR 检测框架中，实现了纯图像级元学习。它没有任何区域提议，绕过了普遍的少镜头检测框架中不准确提议的限制，消除了由于新类的区域建议不准确而造成的约束。此外，Meta-DETR 可以在一个前馈中同时处理多个支持类。这种独特的设计可以捕获不同类之间的

类间相关性同时有效利用，增强了知识向新类的泛化。

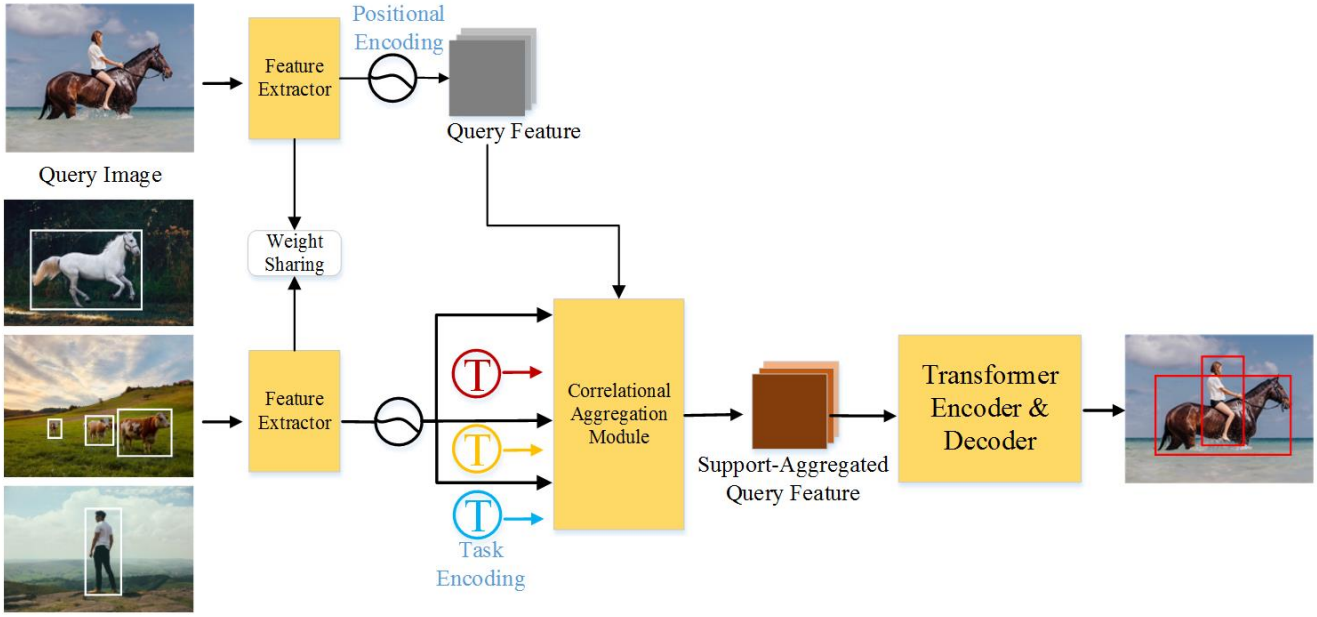

图 13 Meta-DETR 框架图

Meta-DETR 的工作流程如图 13 所示，给定一个查询图像和一组带有实例注释的支持图像，通过权重共享特征提取器处理查询和支持图像，生成查询和支持特征。为了利用元学习中的类间相关性，相关聚合模块（CAM）首先将查询特性与多个支持类同时匹配，然后将支持类集映射到一组预定义的任务编码，这些编码以一种与类无关的方式区分这些支持类。最后，通过预测目标的位置和相应的任务编码的 Transformer [34]结构得到结果。

表 1 基于深度学习的目标检测算法总结

模型	骨干网络	优缺点
RCNN	AlexNet	优于手动设计特征的传统方法，但计算量大，空间复杂度高
Fast R-CNN	VGG-16	实现了端到端的检测，但时间复杂度高
Faster R-CNN	VGG-16	可以实现实时检测，但准确度低，小目标检测效果差
YOLOV1	VGG-16	网络简单，速度快，但定位精度低
YOLOV2	DarkNet-19	减少分类错误，但准确度不高
YOLOV3	DarkNet-53	速度相较于 V2 有提升，缺乏突破性的创新
YOLOX	DarkNet-53	检测速度快，可以进行实时视频检测，但架构简单，检测精度低，易受噪声影响
CornerNet	Hourglass	检测速度较基于 Anchor 的算法高，但小目标检测精度低，未考虑边界框的内部信息
CenterNet	Resnet/DLA/Hourglass	考虑了边界框的内部信息，提高关键预测精度，但多目标检测精度较差

Anchor-free 目标检测算法概念虽然提出的时间较早，但是很长一段时间内都没有得到研究人员的关注，发展缓慢，一直到 Focal Loss 损失函数的完善才逐渐受到学界的关注，快速取得了许多发展成就，与 Anchor-based 目标检测算法可以说各有优劣，少样本目标检测算法则是目标检测领域一直以来的难点和重点。因此，当前的研究方向不仅侧重于对这些主流算法的优化和改进，还力求达到检测精度和检测速度表现的平衡。表 1 列举了本篇综述中提到的目标检测算法，对其做了一个简单的总结，文中提到的 FSOD 大多以 Fast R-CNN 作为基线，因此表格中不再赘述。

4 数据集与性能评价指标

4.1 常用数据集

数据集即 DataSet，是根据不同需求，由多种数据类型（表，文字，数字，图像等）打包的数据对象的

集合，通常基于数据存储中的文件，也可以根据不同的使用场景基于网络或其他源。常用数据集包括PASCAL VOC [35], ImageNet [36], MS-COCO [37]等。

PASCAL VOC 是一个用于识别和分类任务的数据集，包括图像分类、对象检测、语义分割等多个任务，最初只有 4 个类别（Bicycles, Cars, Motorbike, People）的标注数据用于分类检测任务。在 2007 年，初步建立成一个完善的数据集，检测的类别扩充到了 20 个，如表 2 所示。目前使用上主要有两个版本，分别是 VOC07 与 VOC12。

ImageNet 由英特尔公司的研究人员于 2009 年收集而成，目前由斯坦福大学的计算机视觉实验室和网络图像应用研究组维护，它包含超过 l400 万幅图像，涵盖了超过 2 万个不同的类别，每个类别都有至少 1000 张图像，它的结构基本上是金字塔型，即目录->子目录->图片集。ImageNet 可用于分类、检测、定位以及场景分类等视觉任务。

MS-COCO 是由 Microsoft 发布的一个大型计算机视觉数据集，用于图像分类、语义分割、目标检测等任务，提供了超过 330 万张图像，涵盖了 80 个不同的类别，是当今最具挑战性的目标检测数据集。

表 2 VOC 数据集类别

Vehicles	Animal	Indoor	Person
Aeroplane	Bird	Bottle	Person
Bicycle	Cat	Chair	
Boat	Cow	Dining Table	
Bus	Dog	Potted Plant	
Car	Horse	Sofa	
Motorbike	Sheep	TV/Monitor	
Train			

数据集是处理大规模数据的推荐方式，使用大型且具准确的数据集训练对于设计检验一个优秀的目标检测算法至关重要，所有的目标检测算法都需要通过使用数据集来训练优化模型，也因此在机器学习与深度学习中被广泛使用。表 3 对上述的几种数据集进行了简单总结。

表 3 常用数据集

Dataset	Categories	Train	Validation	Test
VOC 2007	20	2501	2510	4952
VOC 2012	21	5717	5823	10991
ImageNet	200	456567	20121	40152
MS-COCO	80	118287	5000	40670

4.2 性能评价指标

目标检测的目的是检测和识别图像中的物体，为了保证检测的精度、效率、准确性、评估目标检测算法的性能，需要使用性能评估指标。这些指标可以帮助作者评估算法模型设计的准确性，不同环境场景中的表现，以及不同算法在同一任务上的优劣，以便选择最优模型。

通常，目标检测的性能评估指标主要包括准确度（Accuracy）、召回率（Recall）、精确率（Precision）、AP（Average Precision）、mAP（mean Average Precision）、平均召回率（mean Average Recall）、交并比（IoU，Intersection-over Union）、FPS（Frames Per Second）等。

若将检测结果设定为正例（Positive）和负例（Negative）两种情况，则最终存在的结果将包括以下四种：正确识别成正例的正例 TP（True Positive）、错误识别成正例的负例 FP（False Positive）、正确识别成负例的负例 TN（True Negative）以及错误识别成负例的正例 FN（False Negative）[38]，如表 4 所示。

表 4 混淆矩阵

预测值	真实值	
	Positive	Negative
True	TP	TN
False	FP	FN

准确度，是指被检测出来的所有目标之中，真正的正样本所占的比例，即检测出来的目标中检测正确的比例，用如下公式表示：

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

精确率，是指在所有检测出的目标（包括误检）中检测正确的概率，用如下公式表示：

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

召回率，指的是所有真实的正样本中，被检测出来的样本占所有真正的正样本的比例，用如下公式表示：

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

AP，指的是检测出来的框的平均准确率，衡量模型在检测每一个类别的目标时的精确度，值越大即性能越好，可以反映模型在不同置信度下的表现，用置信度大于某阈值的框的准确率的平均值来表示，即如下公式：

$$AP = \int_0^1 P(R)d(R) \tag{4}$$

mAP，是多类别目标检测中的一种度量标准，表示模型在多类别检测任务上的平均性能，它可以反映模型在不同类别的表现，用所有类别的 AP 的平均值来表示，即如下公式：

$$mAP = \frac{1}{classes}\sum_{i=1}^{classes}\int_0^1 P(R)d(R) \tag{5}$$

AP 与 mAP 是两个极易混淆的概念，简单来说，AP 在单一类别目标检测任务中使用，表示模型在检测出每一个类别的目标时的精确度；而 mAP 在多类别目标检测任务中使用，表示模型在多类别目标检测任务上的平均性能。

交并比作为目标定位的度量值，用于评估模型定位的准确性。指模型产生的“预测边框”与原图片标注的“真值”的交叠比值，取值范围从 0 到 1，数值越大表示重叠程度越大。一般来说，如果 IoU 大于某一阈值（如 0.5），则认为正确地检测到了目标。

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

FPS 表示每秒处理图片的帧数，主要用于判断模型的计算性能，数值越大代表检测速度越快，在大部分情况下超过 30 帧就可以满足实时性的要求。FPS 的计算公式为：FPS=1/（每一帧检测时间+每一帧处理时间）。FPS 与算法的处理速度有关，一般来说，算法越复杂，FPS 越低，反之，算法越简单，FPS 越高。

综上所述，目标检测模型的性能主要是精度和速度两方面。精确率和召回率是最重要的指标，它们可以帮助比较不同模型的性能，判断模型是否可以满足需求。mAP 值可以帮助作者评估模型在多类别分类任务中的性能，它可以更好地反映模型的整体表现，值越大，代表精度越高。FPS 体现了模型的计算性能，值越大说明检测速度越快。

5 总结与展望

发展至今，目标检测技术已取得显著成就。本文从概念出发，回顾了目标检测中的一些评估指标、数据集、关键技术等，也思考了目标检测目前面临的挑战和未来的发展方向。经过论述，可以看到深度学习已经成为目标检测领域的重要研究方向，它改变了传统的基于手动特征提取的机器学习方法，为目标检测提供了更好的精度，使模型更加紧凑，能够以更快的速度完成检测。此外还可以提高模型的泛化能力，使模型能够更好地检测复杂的目标。这些改进都极大地提高了检测准确率和运行效率。

目前，主要的深度学习模型有卷积神经网络（CNN）、深度卷积神经网络（DCNN）等。此外，深度学习技术也被用于目标检测的一些其他方面，如目标检测的可解释性、数据增强等。虽然在过去的几十年间取得了一些成果，但是目前仍存在着一些限制。

5.1 数据集的多样性

目标检测模型需要在数据集上进行大量的训练，因此检测性能的优劣很大程度上取决于数据集质量。一般来说，可以通过增大数据的规模来提高模型的泛化能力，并以此避免过拟合。但是由于采样与标注的困难，目前缺乏通用的大规模多样性数据集，导致模型只在某单个领域内可用。同时，在军事，测绘等特殊领域，很难获得有详细标注的清晰图像。因此创建多元化的数据集，针对不同任务都有多样性的大规模数据集可用，那么检测效果的提高则指日可待。

5.2 轻量化的网络模型

现有模型架构复杂，参数多，在搭载于一些迷你嵌入式移动设备，或在一些对实时性要求比较高的场景下应用时比较困难。轻量级的模型可以提供更快的训练时间，从而满足实时性要求。目前深度学习在目标检测中的应用仍面临运算复杂度及成本偏高的问题，难以快速验证改善模型。因此，降低计算开销，提供更轻量、快速的模型将成为努力的方向之一。

5.3 更具可解释性的方法

可解释性是指人类能够理解决策原因的程度。尽管新兴的基于深度学习的检测方法拥有卓越的性能，但是大部分研究人员无法从人类角度完全理解现在的深度神经网络模型决策。许多方法还依赖设计者的经验假设。这也是现在几乎所有的模型都没法完全部署到运输，医疗，法律，财经等关键领域的原因。

从解释性的角度对模型展开研究不仅有助于设计者理解模型运行原理，还能反向推动新模型的开发。

因此，更具解释性的方法和模型可视化等模型理解的研究不只在目标检测领域，也在整个计算机视觉领域极具意义[39]。

未来目标检测领域的发展无疑会伴随着深度学习的脚步前进，期待能够早日出现更多新的突破，将目标检测技术的应用范围变得更加广泛。

参考文献

[1] Turk M A, Pentland A P. Face recognition using eigenfaces [C] // Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 1991: 586-591.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60: 91-110.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[4] Balcazar J L, Dai Y, Watanabe O. Provably fast training algorithms for support vector machines [C] // Proceedings 2001 IEEE International Conference on Data Mining. IEEE, 2001: 43-50.

[5] Neubeck A, Van Gool L. Efficient non-maximum suppression [C] // 18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 3: 850-855.

[6] Wang P, Shen C, Barnes N, et al. Fast and robust object detection using asymmetric totally corrective boosting [J]. IEEE Transactions on Neural Networks and Learning Systems, 2011, 23 (1): 33-46.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60 (6): 84-90.

[8] Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning [J]. IEEE Transactions on Medical Imaging, 2016, 35 (5): 1285-1298.

[9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[10] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104: 154-171.

[11] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 658-666.

[12] Girshick R. Fast r-cnn [C] // Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[13] Luce R D. Individual choice behavior: A theoretical analysis [Z]. Courier Corporation, 2012.

[14] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 28.

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

[16] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[17] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[18] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C] // International Conference on Machine Learning. pmlr, 2015: 448-456.

[19] MacQueen J. Some methods for classification and analysis of multivariate observations [C] // Proc. 5th Berkeley Symposium on Math., Stat., and Prob. 1965: 281.

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

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

[22] He K, Zhang X, Ren S, 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.

[23] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8759-8768.

[24] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[25] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021 [J]. arXiv preprint arXiv: 2107.08430, 2021.

[26] 董文轩, 梁宏涛, 刘国柱, 等. 深度卷积应用于目标检测算法综述 [J]. 计算机科学与探索, 2022, 16 (5): 1025.

[27] Law H, Deng J. Cornernet: Detecting objects as paired keypoints [C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.

[28] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation [C] // Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing, 2016: 483-499.

[29] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[30] Zhou X, Wang D, Krähenbühl P. Objects as points [J]. arXiv preprint arXiv: 1904.07850, 2019.

[31] Sun B, Li B, Cai S, et al. Fsce: Few-shot object detection via contrastive proposal encoding [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7352-7362.

[32] Hu H, Bai S, Li A, et al. Dense relation distillation with context-aware aggregation for few-shot object detection [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10185-10194.

[33] Zhang G, Luo Z, Cui K, et al. Meta-detr: Few-shot object detection via unified image-level meta-learning [J]. arXiv preprint arXiv: 2103.11731, 2021, 2 (6).

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30.

[35] Shetty S. Application of convolutional neural network for image classification on Pascal VOC challenge 2012 dataset [J]. arXiv preprint arXiv: 1607.03785, 2016.

[36] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database [C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, 2009: 248-255.

[37] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context [C] // Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[38] 刘洋, 战荫伟. 基于深度学习的小目标检测算法综述 [J]. 计算机工程与应用, 2021, 57 (2): 37-48.

[39] 罗东亮, 蔡雨萱, 杨子豪, 等. 工业缺陷检测深度学习方法综述 [J]. 中国科学: 信息科学, 2022 (052-006).