综述

基于深度学习的图像目标检测算法综述

张婷婷[1]，章坚武[1]，郭春生[1]，陈华华[1]，周迪[2]，王延松[3]，徐爱华[2]
（1. 杭州电子科技大学，浙江 杭州 310018；2. 浙江宇视科技有限公司，浙江 杭州 310051；
3. 之江实验室，浙江 杭州 311121）

摘　要：图像目标检测是找出图像中感兴趣的目标，并确定他们的类别和位置，是当前计算机视觉领域的研究热点。近年来，由于深度学习在图像分类方面的准确度明显提高，基于深度学习的图像目标检测模型逐渐成为主流。首先介绍了图像目标检测模型中常用的卷积神经网络；然后，重点从候选区域、回归和 anchor-free 方法的角度对现有经典的图像目标检测模型进行综述；最后，根据在公共数据集上的检测结果分析模型的优势和缺点，总结了图像目标检测研究中存在的问题并对未来发展做出展望。

关键词：计算机视觉；图像目标检测；深度学习；图像分类
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020199

A survey of image object detection algorithm based on deep learning

ZHANG Tingting[1], ZHANG Jianwu[1], GUO Chunsheng[1], CHEN Huahua[1], ZHOU Di[2],
WANG Yansong[3], XU Aihua[2]
1. Hangzhou Dianzi University, Hangzhou 310018, China
2. Zhejiang Uniview Technologies Co., Ltd., Hangzhou 310051, China
3. Zhijiang Lab, Hangzhou 311121, China

Abstract: Image object detection is to find out the objects of interest in the image and determine their classifications and locations. It is a research hotspot in the field of computer vision. In recent years, due to the significant improvement in the accuracy of image classification with deep learning, image object detection models based on deep learning have gradually became mainstream. Firstly, the convolutional neural networks commonly used in image object detection were briefly introduced. Then, the existing classical image object detection models were reviewed from the perspective of candidate regions, regression and anchor-free methods. Finally, according to the detection results on the public dataset, the advantages and disadvantages of the models were analyzed, the problems in the image object detection research were summarized and the future development was forecasted.

Key words: computer vision, image object detection, deep learning, image classification

收稿日期：2020−03−23；**修回日期**：2020−06−30
通信作者：章坚武，jwzhang@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.U1866209，No.61772162）；国家重点研发计划基金资助项目（No.2018YFC0831503）；浙江省自然科学基金资助项目（No.LYl6F020016）；浙江省重点研发计划基金资助项目（No.2018C01059，No.2019C01062）
Foundation Items: The National Natural Science Foundation of China (No.U1866209, No.61772162), The National Key Research Development Program of China (No.2018YFC0831503), The Natural Science Foundation of Zhejiang Province of China (No.LYl6F020016), The Key Research Development Program of Zhejiang Province of China (No.2018C01059, No.2019C01062)

1　引言

计算机视觉（computer vision）是人工智能（artificial intelligence，AI）的关键领域之一，是一门研究如何使机器“看”的科学。图像目标检测又是计算机视觉的关键任务，主要对图像或视频中的物体进行识别和定位，是 AI 后续应用的基础。因此，检测性能的好坏直接影响到后续目标追踪[1-2]、动作识别[3-4]的性能。

传统图像目标检测的滑窗法虽然简单易于理解，但随目标大小而变化的窗口对图像进行从左至右、从上至下的全局搜索导致效率低下。为了在滑动窗口检测器的基础上提高搜索速度，Uijlings 等[5]提出了选择性搜索方法（selective search method），该方法的主要观点是图像中的目标存在的区域具有相似性和连续性，基于这一想法采用子区域合并的方式进行候选区域的提取从而确定目标。Girshick 等[6]提出的基于区域的卷积神经网络（region-based convolutional neural network，R-CNN）就是采用了选择性搜索方法提取候选区域，进而越来越多的学者在不断改进确定目标的方法的基础上提出新的检测模型。

本文首先介绍了图像目标检测模型中常用的卷积神经网络；然后，重点从候选区域、回归和 anchor-free 方法等角度对现有的图像目标检测模型进行综述；最后，根据在公共数据集上的检测结果分析模型的优势和缺点，总结了现有图像目标检测研究中存在的问题并对未来发展做出展望。

2　相关基础

卷积神经网络（convolutional neural network，CNN）是一种类似人工神经网络多层感知器的深度学习模型，在计算机视觉方面贡献巨大。它以原始数据作为输入，通过卷积、池化和非线性激活函数映射等一系列操作，将原始数据逐层抽象为目标任务的特征表示[7]。一个标准的 CNN 主要由卷积层、池化层和全连接层等核心层次构成，其中卷积层是通过卷积核与输入的单通道或多通道图像进行卷积进而提取特征的操作；池化层是一种提取输入数据核心特征的操作，该操作通过压缩原始数据减少了 CNN 中的模型参数，还在一定程度上提升了计算效率；全连接层的主要作用是对经过卷积和池化之后的数据进行压缩，并且根据压缩的特征完成模型的分类功能。

LeNet-5 是在 1998 年由 Lecun 等[8]多次研究后最终提出的卷积神经网络结构，是最早的卷积神经网络之一，有效地推动了深度学习领域的发展。该网络共有 7 层，利用卷积、参数共享和池化等操作提取特征，最后经过全连接层和 softmax 函数完成目标分类，网络结构如图 1（a）所示。LeNet-5 为卷积神经网络的发展奠定了基础，现在的许多卷积神经网络都是以此为雏形进行优化的。

Alexnet 是 Krizhevsky 等[9]发明的一个深度学习模型，该模型 2012 年在计算机视觉竞赛 ILSVRC 中一举夺冠，使得 CNN 成为图像分类的核心算法，引来了深度学习大爆发，具有深远意义。Alexnet 的网络结构如图 1（b）所示，该网络总共有 8 层，5 个卷积层和 3 个全连接层，在每一个卷积层中包含激励函数 Relu 以及局部响应归一化处理，Alexnet 的最大创新是提出 Dropout 随机忽略一部分的神经元防止训练过程中出现过拟合。

ZFNet[10]是在 Alexnet 中进行微调后的结果，该模型通过反卷积操作可视化了特征图，证明了浅层网络学习得到的是图像的边缘和颜色的纹理特征，而深层网络的学习得到的则是图像的抽象特征。ZFNet 的网络结构如图 1（c）所示，与 Alexnet 相比，ZFNet 在第一个卷积层中采用了更小的卷积核和步长，使得图像特征更好地被保留。该模型在 2013 年的 ILSVRC 竞赛中获得了第一名。

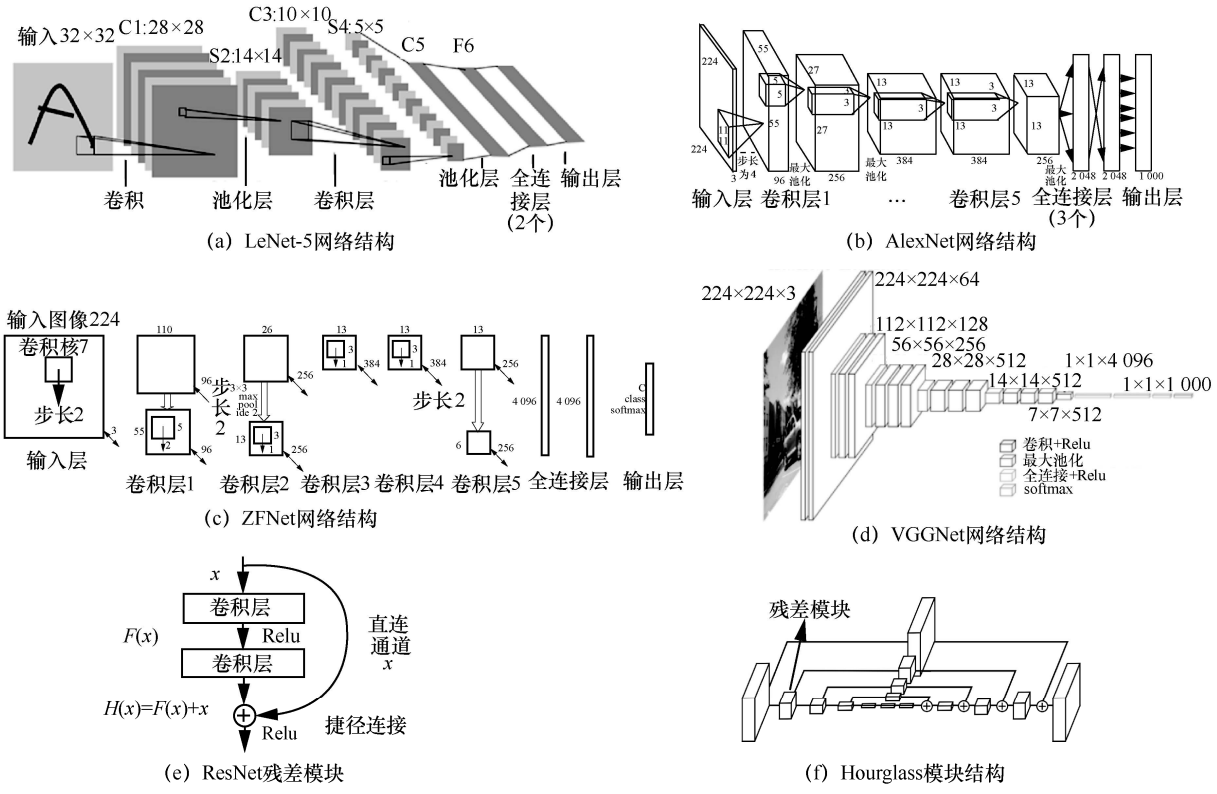


图 1　卷积神经网络结构对比

VGGNet 是 Simonyan 等[11]提出的深度学习模型，该模型与 Alexnet 结构相似，都由卷积层和全连接层两部分构成，是 Alexnet 的加深版本。VGGNet 的网络结构如图 1（d）所示，在网络中 VGGNet 用小的卷积核组替代大的卷积核，在具有相同感受视野[12]的同时减少了参数，并且引入了更多的非线性因素，增强了模型的表达；另外 VGGNet 通过使用小池化核增加通道数，与 Alexnet 相比获得了更细节的信息。VGGNet 构建了 16~19 层深的卷积神经网络，并成功地证明了适当地增加网络的深度可以在一定程度上影响网络的最终性能。

2015 年 He 等[13]提出具有残差学习思想的深度残差网络 ResNet，利用残差模块成功地训练出 152 层的卷积神经网络，并在 ILSVRC2015 比赛中获得冠军，残差模块如图 1（e）所示。在 ResNet 残差模块中直接将输入信息绕道传到输出，然后整个网络对输入、输出差别的那一部分进行学习，通过这种增加直连通道的做法不但保护了信息的完整性，而且还简化了学习目标和难度，并且解决了由直接增加网络深度所带来的信息丢失、损耗、梯度消失[14]、梯度爆炸[15]等问题。ResNet 用残差模块堆叠成不同的网络层数，常用的有 ResNet-50、ResNet-101、ResNet-152。

堆叠沙漏网络（stacked hourglass network）是 Newell 等[16]提出应用于人体姿态估计的网络，由多个堆叠沙漏模块组成，对多尺度特征进行处理并加以合并，很好地捕获了与身体相关的各种空间关系。单个沙漏（hourglass）模块结构如图 1（f）所示，其特点是特征由高分辨率到低分辨率，再由低分辨率到高分辨率的分布是高度对称的，并且借鉴了 ResNet 的残差思想，即图 1 中的每个方块均对应一个残差模块，最后基于池和上采样的连续操作生成用于预测人体关键点的 Heatmaps 集。

3　基于深度学习的图像目标检测模型

本节将介绍近几年提出的基于候选区域、回归和 anchor-free 的图像目标检测模型，总结各模型相比之前模型的改进策略以及自身的创新点和不足，并在 PASCAL VOC2007[17]、PASCAL VOC2012[17]和 MS COCO[18]等常用公共数据集上做出比较。

3.1　基于候选区域的图像目标检测模型

R-CNN 图像目标检测模型是 Girshick 等[6]于 2013 年提出的，它是候选区域和卷积神经网络这一框架的开山之作，也是第一个可以真正应用于工业级图像目标检测的解决方案，为基于 CNN 图像目标检测的发展奠定了基础。网络结构如图 2 所示。R-CNN 首先使用选择性搜索方法从输入的图像中提取出 2 000 个候选区域，使用剪裁[9]和变形[19]的方法将候选区域的尺寸固定为 277×277 以适应全连接层的输入，通过 CNN 前向传播对每个候选区域进行特征计算；然后将每个候选区域的特征向量送入特定线性分类器中进行分类和预测概率值；最后使用非极大值抑制（non-maximum suppression，NMS）[20]算法消除多余的目标框，找到目标的最佳预测位置。

R-CNN 图像目标检测模型虽然将 mAP（mean average precision，平均精度值）[17]在 VOC2007 和 VOC2012 数据集上分别达到了 58.5%和 53.3%，在基于深度学习的图像目标检测领域取得了重大突破，但由于其输入图像经过剪裁和变形后会导致信息丢失和位置信息扭曲，从而影响识别精度，并且 R-CNN 需要对每张图片中的上千个变形后的区域反复调用 CNN，所以特征计算非常耗时，速度较慢。

基于 R-CNN 需固定输入图像尺寸以及检测速度较慢的缺点，2014 年 He 等[21]提出了 SPP-Net，该模型先是计算整个输入图像的卷积特征图，根据选择性搜索方法提取候选区域，通过对特征图上与候选区域相对应位置的窗口使用金字塔池化（spatial pyramid pooling，SPP）可以得到一个固定大小的输出，即全连接层的输入。与 R-CNN 相比，SPP-Net 避免了反复使用 CNN 计算卷积特征，在无须对输入图像进行剪裁和变形的情况下实现了多尺度输入卷积计算，保留了图像的底层信息，在 VOC2007 数据集上测试时 mAP 达到了 59.2%，在达到相同或更好的性能前提下，比 R-CNN 模型快 24~102 倍。

虽然 R-CNN 和 SPP-Net 在 VOC2007 数据集上都获得了很高的精度，但两者将分类和回归分为多阶段进行，使得网络占用了较多的硬件资源。2015 年 Girshick 等[22]提出了一种快速的基于区域的卷积网络模型（fast R-CNN）。该网络首先用选择性搜索方法提取候选区域，将归一化到统一格式的图片输入 CNN 进行卷积计算，然后借鉴了 SPP-Net 中金字塔池化的思想，用最大值池化层 ROI pooling 将卷积特征变成固定大小的 ROI 特征输入全连接层进行目标分类和位置回归。该网络采用多任务训练模式，用 softmax 替代 SVM（support vector machine，支持向量机）[23]进行分类，将分类和回归加入网络同时训练，在末尾采

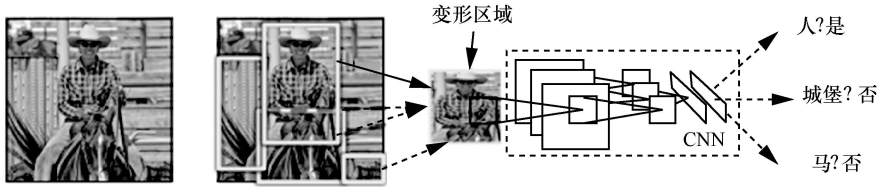


图 2　R-CNN 网络结构

用可同时输出分类和回归结果的并行全连接层。fast R-CNN 减少了硬件缓存，提高了检测速度，初步实现了端对端的图像目标检测，并且在 VOC2007 和 VOC2012 数据集上的 mAP 分别为 66.9%和 66.0%。

由于 fast R-CNN 无法满足实时检测的需求，Ren 等[24]提出了改进模型 faster R-CNN。该网络的最大创新就是提出了区域提议网络（region proposal network，RPN），即在基础卷积网络提取输入图像特征的基础上用 RPN 代替 fast R-CNN 中的选择性搜索方法进行候选区域的提取。RPN 是一个全卷积网络，网络结构如图 3 所示，该网络可以同时在每个位置上预测出目标边界和目标概率并产生高质量候选区域，然后通过 ROI pooling 将卷积特征变成固定大小的 ROI 特征输入全连接层进行目标分类和位置回归。RPN 和 fast R-CNN 通过四步交替训练法使两个网络共享卷积特征合并为单一网络，解决了区域计算的瓶颈问题，在实现真正端对端训练模式的基础上满足了实时应用的需求[23]。

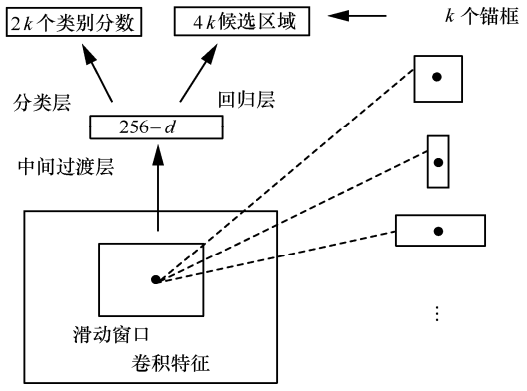


图 3　区域提议网络结构（RPN）

2017 年 He 等[25]提出了 mask R-CNN 目标检测模型，该模型以 faster R-CNN 为原型，即在 faster R-CNN 中生成的候选区域中融入 FCN（fully convolutional network，全卷积神经网络）[26]作为新的支路用于生成每个候选区域的掩膜，同时把 faster R-CNN 中 RoI pooling 修改成为了 ROI align 用于处理掩膜与原图中物体不对齐的问题。Mask R-CNN 在训练时可以同时生成目标边界、目标概率和掩膜，但在预测时通过将目标边界和目标概率的结果输入掩膜预测中以生成最后的掩膜，该方法减弱了类别间的竞争优势，从而达到了更好的效果，在 MS COCO 数据集上的 mAP 测试结果达到 35.7%。

3.2　基于回归的图像目标检测模型

3.2.1　YOLO 及扩展模型

检测精度和检测速度是评判图像目标检测模型好坏的重要标准[27]。基于候选区域的图像目标检测模型，虽然在检测精度方面首屈一指，但是它检测图像的效率低是其主要弊端。2016 年 Redmon 等[28]提出 YOLO（you only look once）检测模型，该模型将图像目标检测抽象为回归问题，通过对完整图片的一次检测就直接预测出感兴趣目标的边界框和类别，避免了 R-CNN 系列中将检测任务分两步进行的烦琐操作，解决了之前图像目标检测模型检测效率低的问题。检测网络将输入的图片分成 $s\times s$ 个网格，如图 4 所示，各网格只负责检测中心落在该网格的目标，预测出网格的类别信息以及多个边界框和各个边界框的置信度，通过设定阈值过滤掉置信度较低的边界框，然后对保留的边界框进行 NMS 处理以确定最终的检测结果。YOLO 以回归替代了之前图像目标检测模型的候选区域方法，在满足实时需求的基础上检测速度达到 45 f/s，但由于 YOLO 在检测过程中仅选择置信度最高的边界框作为最终的输出，即每个网格最多只检测出一个物体，因此 YOLO 在检测紧邻群体目标或小目标时效果不佳，在 VOC2007 上的 mAP 也仅有 66.4%。

针对 YOLO 在目标定位方面不够准确的问题，2017 年 Redmon 等[29]提出了 YOLO 的扩展模型 YOLOv2 和 YOLO9000。YOLOv2 首先在卷积层中添加批量归一化（batch normalization，BN）[30]

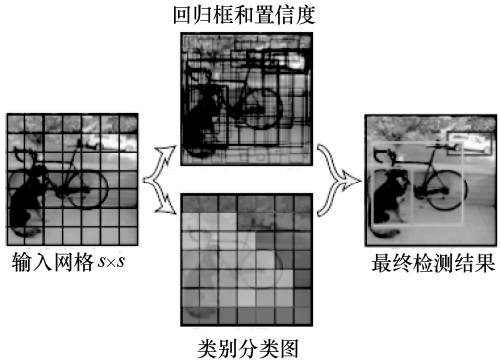


图 4　YOLO 检测模型

技术使得模型的收敛性有显著的提升，然后借鉴 faster R-CNN 的思想用聚类方法产生的锚框替代了 YOLO 中预测出的边界框，最后通过输入更高的分辨率图像并对其进行迁移学习[31]从而提升网络对高分辨率图像的响应能力，训练过程中无须固定图像的尺寸，因此在一定程度上提升了网络的泛化能力。除此之外 YOLOv2 还提出将一个由 19 个卷积层和 5 个 MaxPooling 层构成的 Darknet-19[28]网络作为骨干网进一步提升检测速度。而 YOLO9000 则是在 YOLOv2 的基础上提出了目标分类和检测的联合训练方法，使 YOLOv2 的检测种类扩充到 9 000 种。2017 年 Redmon 等[32]提出了 YOLOv3 检测模型，它借鉴了残差网络结构，形成网络层次更深的 Darknet-53，通过特征融合的方式采用 3 个不同尺度的特征图进行目标检测，并且用 logistic 代替 softmax 进行类别预测实现了多标签目标检测，该网络不仅提升了小目标检测效果，在边界框预测不严格并且检测精度相当的情况下检测速度是其他模型的 3~4 倍。

3.2.2　SSD 及扩展模型

2016 年 Liu 等[33]提出 SSD 图像目标检测模型，该模型彻底淘汰了生成候选区域和特征重采样阶段，选择将所有计算封装在单个深层神经网络中，网络结构如图 5 所示。SSD 网络继承了 YOLO 中将目标检测问题抽象为回归问题的思想，采用特征金字塔的方式进行检测，即利用不同卷积层产生不同的特征图，使用一个小的卷积滤波器来预测特征图上一组固定的默认边界框类别和位置偏移量。为了实现较高的检测精度，在不同尺度的特征图中进行不同尺度的预测，并设置不同长宽比的边界框进行分离预测。由于图像中的目标具有随机性，大小不一，所以小目标的检测是由 SSD 使用底层特征图来实现的，大目标的检测是由 SSD 使用高层特征图来实现的，相对于 YOLO 精确度大幅度提高，并且效率也有所提升。

2017 年 Fu 等[34]提出 DSSD 检测模型，即将 Resnet-101 作为 SSD 的骨干网，在分类回归之前引入残差模块，并且在原本 SSD 添加的辅助卷积之后又添加了反卷积层，与 SSD 相比，DSSD 在小目标的检测精度上有了很大的提升，但 Resnet-101 网络太深导致 DSSD 的检测速度相比 SSD 较慢。2017 年 Jisoo 等[35]在未改动 SSD 主干

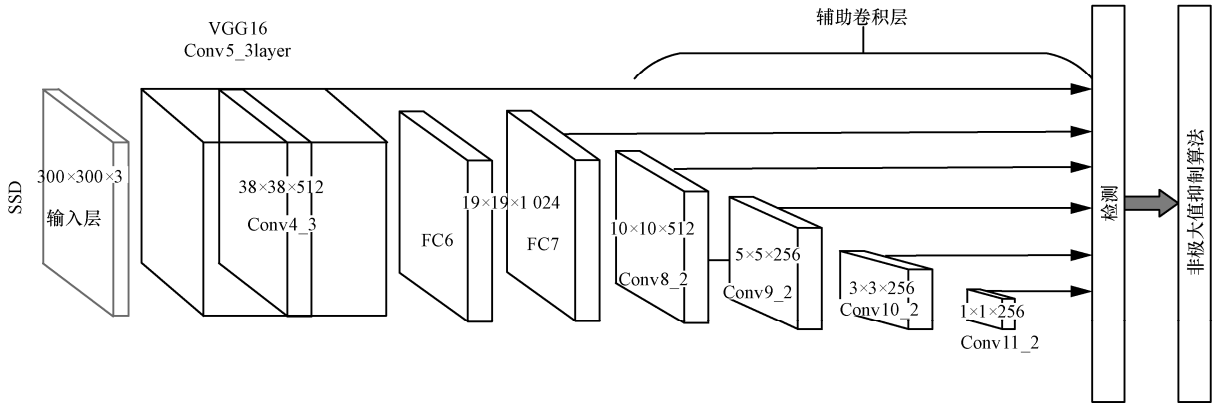


图 5　SSD 网络结构

网络的基础上提出了 RSSD（rainbow SSD）检测模型，该网络同时采用池化和反卷积的方式进行特征融合，不仅增强了不同特征层之间的关系，由于融合后的特征大小相同，还一定程度上增加了不同层的特征个数。这种特征融合方式解决了 SSD 存在的重复框的问题，同时提升了对小目标的检测效果，但与 SSD 相比检测速度较慢。2017 年 Li 等[36]提出了 FSSD，该模型通过重构一组金字塔特征图充分融合了不同层不同尺度的特征，在保证检测速度与 SSD 相当的同时使得检测精度有了明显的提升。2019 年 Yi 等[37]借鉴注意力机制[38]的思想在 SSD 检测模型中设计了一个注意力模块，该注意力模块基于全局特征关系可以分析出不同位置特征的重要性，从而达到在网络中突出有用信息和抑制无用信息的效果，ASSD[37]检测精度提高，但与 SSD 相比，检测速度较慢。

3.2.3 RetinaNet 模型

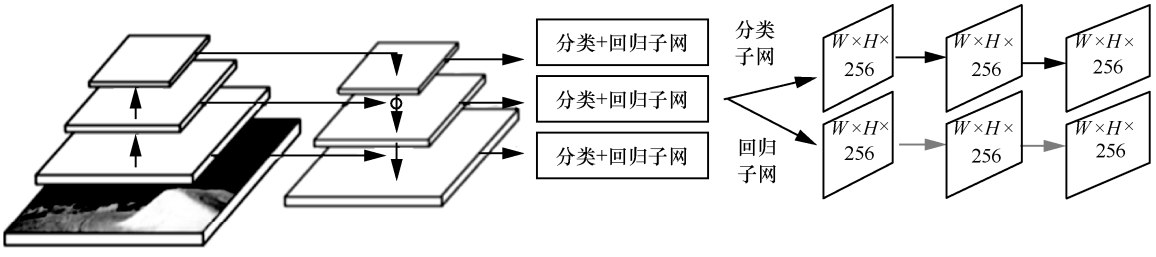
2017 年 Lin 等[39]借鉴了参考文献[24,40]中的想法设计并训练了一个简单的高密度 RetinaNet 检测模型，该模型的主要思想是通过重塑标准交叉熵损失来解决之前检测模型在训练过程中出现的类不平衡问题。RetinaNet 模型包含一个骨干网 ResNet-101-FPN 和两个 FCN 子网，网络结构如图 6 所示。骨干网是负责计算整个输入图像特征图的自卷积网络，两个子网分别实现了目标分类和边界框回归，特征金字塔网络（feature pyramid network，FPN）[41]的主要创新是通过自上而下的横向连接使得标准卷积网络得以增强，因此在 ResNet[13]的基础上使用 FPN 作为骨干网可以有效地构建丰富的多尺度特征金字塔。RetinaNet 高效准确，在 MS COCO 数据集上以 5 f/s 运行的同时达到 39.1%的 mAP，但相比 SSD 检测速度较慢。

3.3 基于 anchor-free 的图像目标检测模型

图像目标检测发展日新月异，越来越多优秀目标检测模型陆续被提出，基于候选区域和回归方法的检测模型目前发展稳定并且成熟，而基于 anchor-free 的检测模型是当下目标检测领域中新的热门研究方向，anchor-free 检测模型有两种，分别为基于关键点的检测和基于分类和回归进行改进的检测。

3.3.1 基于关键点的检测

2018 年 Law[42]受到 Newell 等在姿态估计[43-46]中的关联嵌入的启发提出了 CornerNet，这是一种新型的图像目标检测方法。CornerNet 将一个目标检测为一对关键点，即目标边界框的左上角点和右下角点，是第一个将图像目标检测任务表述为利用嵌入角点进行分组和检测任务的模型，开启了基于关键点的目标检测方法的大门。CornerNet 首先使用沙漏网络[15]作为其骨干网络输出最后一层卷积特征，骨干网后接两个分支模块，分别进行左上角点预测和右下角点预测，每个分支模块包含一个 Corner pooling（角池化）和 3 个输出，网络结构如图 7 所示。heatmaps（热图）输出的是预测角点的位置信息，当图像中出现多个目标



(a) ResNet (b) FPN (c) 分类子网（上） (d) 回归子网（下）

图 6 RetinaNet 网络结构

时，embeddings（嵌入）根据左上角点和右下角点嵌入向量之间的距离对属于同一目标的一对角点进行分组；offsets（误差）是输出从图像到特征图的量化误差，用来对预测框进行微调。

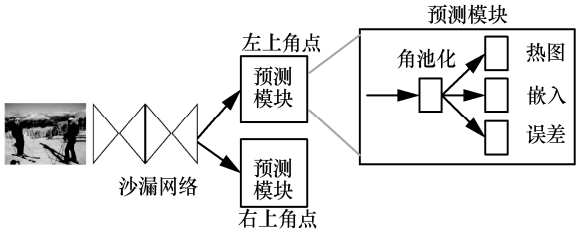


图 7　CornerNet 网络结构

当角点在目标之外时，此时获取的信息不足以进行当前的定位，为了能够更好地定位边界框的角点，Law 等[42]介绍了一种新型池化层—角池化层，该池化层包含两个特征图，在每个像素位置，它将第一个特征图下侧的所有特征向量和第二个特征图右方的所有特征向量最大化，然后将两个合并后的结果相加输出最后的角点。CornerNet 极大地简化了网络的输出，彻底消除了图像目标检测对候选区域和候选框的需要，在 MS COCO 上实现了 42.1%的 mAP，但当 CornerNet 将边界框的角点定位在物体之外时目标的局部特征表现不强烈，并且在判断两个角点是否属于同一目标时，由于缺乏全局信息的辅助导致匹配角点时产生错误目标框，因此存在一定的误检率。

2019 年 Zhou 等[47]借鉴 CornerNet 的思想提出一种新的检测思路，即通过关键点估计[48-50]网络对每个目标预测出 4 个极值点和 1 个中心点，然后提取极值点的峰值，暴力枚举所有的组合并计算出每个组合的几何中心点，若几何中心点与预测的中心点匹配度高于设定阈值，则接受该组合，并将这 5 个极值点的得分平均值作为组合的置信度。ExtremeNet[47]将目标检测问题转化成单纯的基于外观信息的关键点估计问题，避免了对目标隐含特征的学习，相对于 CornerNet 更好地反映了物体的信息，检测效果更好。

2019 年 Duan 等[51]在 CornerNet 基础上提出了改进模型 CenterNet，该模型通过 Corner pooling 和 Cascade Corner pooling 将每个目标检测为 2 个角点和 1 个中心点，有效地获取了目标的边界信息和内部信息，然后通过判断中心区域是否包含中心点来决定目标检测框是否正确。为了减小边界框中心区域的大小对检测结果的影响，该网络还提出了一种多尺度中心区域算法以适用预测框的动态变化，算法计算式如下：

$$\begin{cases} \mathrm{cul}_x = \dfrac{(t+1)\mathrm{ul}_x + (t-1)\mathrm{dr}_x}{2t} \\ \mathrm{cul}_y = \dfrac{(t+1)\mathrm{ul}_y + (t-1)\mathrm{dr}_y}{2t} \\ \mathrm{cdr}_x = \dfrac{(t-1)\mathrm{ul}_x + (t+1)\mathrm{dr}_x}{2t} \\ \mathrm{cdr}_y = \dfrac{(t-1)\mathrm{ul}_y + (t+1)\mathrm{dr}_y}{2t} \end{cases} \tag{1}$$

其中，ul_x、ul_y 是边界框的左上角坐标，dr_x、dr_y 是边界框的右下角坐标，cul_x、cul_y 是中心区域的左上角坐标，cdr_x、cdr_y 是中心区域的右下角坐标，t 是中心区域的尺度，由式（1）可以看出中心区域可以随着边界框大小动态变化。

CenterNet 通过融入中心点的语义信息有效地解决了 CornerNet 中容易产生错误目标框的问题，在 MS COCO 数据集上的 mAP 达到了 47%。2019 年 Zhou 等[52]提出同名 CenterNet 检测网络，该网络提供了一种更简洁的思路，即通过判断一个像素点是否为目标的中心点从而检测目标的边界。与 Duan 等[51]提出的 CenterNet 相比检测过程更简单，并且该思路具有很大的普遍性。

3.3.2　基于分类和回归进行改进的检测

自 2018 年 CornerNet 提出以来，基于 anchor-free 的目标检测模型在分类和回归的方法上又有了新的创新，如 2019 年 Zhu 等[53]提出一种基于 anchor-free 的动态选择特征层的方法，该方法主要是在 RetinaNet 的基础上建立一个 FSAF

（feature selective anchor-free）模块，即对每个层次的特征都建立一个可以将目标分配到合适特性级别的 anchor-free 分支，使得目标框能够在任意特征层通过 anchor-free 分支进行编解码操作。FSAF 可以和基于锚的分支并行工作平行的输出预测结果，有效地提升了 RetinaNet 的稳健性，解决了传统基于锚框检测的网络根据候选框选择特征层的局限性，并在 MS COCO 上实现了 42.8%的 mAP。

传统基于锚框的检测网络面对变化较大的目标时需要根据检测任务预定义锚框尺寸，通过手工设置锚框提高召回率这一操作不仅占用较大的计算和内存资源，还在一定程度上深化了正负样本不平衡问题。2019 年 Tian 等[54]提出一种全卷积目标检测网络 FCOS，类似语义分割中[55]利用逐像素点预测的方式解决目标检测问题。为了提高检测效果，FCOS 引入 center-ness 分支用于降低检测效果不理想的目标框权重，然后通过 NMS 算法确定最终检测结果。基于 anchor-free 的 FCOS 检测网络极大地降低了参数计算，可以与其他视觉任务相结合，并且尽可能多地使用正样本参与训练，解决了之前检测模型中出现的正负样本不平衡问题，但在检测时由于目标真实框重叠，可能会出现语义模糊情况。

2019 年 Kong 等[59]提出了 FoveaBox 目标检测网络，结合人类视觉系统是通过眼球中对物体感应最敏锐的中央凹（Fovea）结构确定物体位置的原理对目标真实框进行位置变换，更具体地说是通过目标真实框找到目标对应在特征图中的中心位置，然后设定两个缩放因子分别对目标真实框向中心点进行收缩和扩展，将收缩边框的内部点作为正样本，扩展边框外部点作为负样本。这种通过位置变化忽略两个边框中间点的方法不仅增加了正负样本之间的识别度、解决了样本不平衡问题，还有效提升了检测性能，但与其他 anchor-free 模型相比检测精度略低，在 MS COCO 上实现的 mAP 仅有 40.6%。

3.4 图像目标检测模型对比

本文对现有经典图像目标检测模型的创新点及优缺点做出对比，见表 1。无论是候选区域法、回归法还是 anchor-free 法，提出模型的主要目的都是为了能够高精度、高速率地识别并检测出目标。由表 1 可以看出，基于候选区域法模型的提出开启了用 CNN 提取特征的大门使图像目标检测进入深度学习时代，回归法则解决了候选区域法的速度瓶颈问题，实现了端对端的图像目标检测。而基于 anchor-free 的算法消除了候选区域法和回归法中候选框的设计，生成高质量的目标框并在未来形成了一个有前途的方向。

对本文中提到的图像目标检测模型在公共数据集上的检测结果做出对比，见表 2。“—”表示此数据集没有该模型的测试结果，2007 表示数据集 VOC 2007，2012 表示数据集 VOC 2012；AP@0.5 表示该模型在 MS COCO 数据集上是取阈值为 0.5 计算精度的，AP@[0.5,0.95]表示该模型在 MSCOCO 数据集上是取 10 个阈值（间隔 0.05）计算精度的，即 mAP，表 2 中所有的数据集精确率检测结果均以百分比为单位。FPS 表示该模型每秒处理图片的数量。

4 今后发展与技术动向

基于深度学习的图像目标检测算法应用广泛，目前在行人检测[60]、车辆检测[61]、工地安全[62]、无人驾驶[63]等领域都有涉及，并且已经取得了很大的成就。但对于复杂场景而言仍然面临诸多挑战，例如，数据集不足、未训练目标对数据的依赖以及用户需求的多样化等。

4.1 数据集标注方法

基于深度学习的目标检测方法对于解决图像处理和计算机视觉问题是非常有利的，而这种功能强大的方法取决于训练数据的数量和质量，因此对大量有效标注数据的需求是非常迫切的。常用的数据集标注方式主要有两种：全人工方式和

表 1　基于深度学习的图像目标检测模型创新点及优缺点对比

检测方法	模型	创新点	优缺点
基于候选区域	R-CNN[6]	用 CNN 提取特征	用卷积网络提取特征；但特征提取复杂，耗时长
	SPP-Net[20]	输入整张图片提取特征，并共享特征图	实现多尺度卷积计算；但占用较大硬件资源
	fast R-CNN[22]	用 ROI pooling 层提取特征	初步实现端对端检测；但依赖传统方法生成候选区域
	faster R-CNN[24]	提出区域生成网络（RPN）	实现实时检测；但对小目标检测效果不佳
	mask R-CNN[25]	增加了用于分割任务的分支	减弱类别间的竞争优势；但与 faster R-CNN 相比，检测速度较慢
基于回归	YOLO[28]	将图像空间划分为网格单元	检测速度快；但定位不准确，对密集物体检测效果不佳
	YOLOv2[29]	使用聚类产生锚框	减少定位错误，提高分类精度；但准确率还不够高
	YOLO9000[29]	提出了目标分类和检测的联合训练方法	为跨数据集训练提供思路；但准确率还不够高
	YOLOv3[32]	借鉴残差学习思想，并进行多尺度检测	检测精度高并且速度是其他模型的 3~4 倍；但对边界框预测严格的情况下检测精度略低
	SSD[33]	生成多尺度的锚框对边界框空间离散化	检测速度快；但准确率低，对小目标检测效果不佳
	DSSD[34]	将 ResNet-101 作为骨干网，并添加反卷积层	提高小目标检测效果；但与 SSD 相比检测速度较慢
	RSSD[35]	改进特征融合方式	检测精度提高；但与 SSD 相比检测速度较慢
	FSSD[36]	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；但与 SSD 相比检测速度较慢
	ASSD[37]	融合了注意力机制的思想	检测精度提高；但与 SSD 相比检测速度较慢
	RetinaNet[39]	重塑交叉熵损失，用焦点损失解决类不平衡问题	检测精度提高；但检测速度一般
基于 anchor-free	CornerNet[42]	将物体检测为两个角点，无须生成锚框	检测精度提高；但容易产生错误目标框，检测速度一般
	ExtremeNet[47]	通过预测极值点和中心点来检测目标	于 CornerNet 相比，对小目标检测效果更好；但对大目标检测效果略差
	CenterNet[51]	将物体检测为两个角点和一个中心点	解决容易产生错误目标框问题；但检测速度一般
	CenterNet[52]	直接判断像素点是否为目标的中心点	算法更简洁；但与同名 CenterNet 相比，检测精度有所降低
	FSAF[53]	提出动态选择特征的方式	摒弃了传统基于锚框选择合适特征层；但在 anchor-free 分支中引入了较多超参数
	FCOS[54]	针对每个像素点进行预测	可以与其他视觉任务相结合；但在检测时由于真实框重叠，可能会出现语义模糊情况
	FoveaBox[59]	对真实框进行位置变换，增加正负样本的识别度	有效地解决了正负样本不平衡问题；与其他 anchor-free 模型相比，检测精度略低

半自动方式，例如 LabelMe[64]是一个基于网络的全人工在线标注系统，它具有两种标记方式，即多边形和掩膜，标记结果导出为 XML 格式。Labelbox 也是一种全人工在线标注系统，但与 LabelMe 相比，具有更多的标记类型和导出格式，可以生成与 Mask R-CNN[25]兼容的掩膜，具有最佳用户体验。

虽然全人工方式目前仍是获得有效标注数据的最佳方式，但是在标注大量数据时工作复杂并且需要大量时间，2019 年 Hidayatullah 等[65]提出一种半自动数据集标注方法，主要思想是利用现有的目标检测方法来自动检测和标注训练数据，并且附加了人工标注方式以标注无法实现自动标注和修改不能正确标注的数据。半自动数据标注方式有效地提升了标注的效率，但是如果未来能

表 2 基于深度学习的图像目标检测模型性能对比

检测方法	模型	骨干网	2007	2012	MS COCO		FPS
					AP@0.5	AP@[0.5,0.95]	
基于候选区域	R-CNN[6]	AlexNet[9]	58.5%	53.3%	—	—	0.02
	SPP-Net[21]	ZF-5[9]	59.2%	—	—	—	0.5
	fast R-CNN[22]	VGG-16[11]	66.9%	66.0%	35.9%	19.7%	0.5
	faster R-CNN[24]	VGG-16	73.2%	70.4%	42.7%	21.9%	7
	mask R-CNN[25]	ResNet-101-FPN[38]	—	—	58.0%	35.7%	5
基于回归	YOLO[28]	VGG-16	66.4%	57.9%	—	—	45
	YOLOv2[28]	DarkNet-19[28]	78.6%	73.5%	44.0%	21.6%	40
	YOLOv3[32]	DarkNet-53[32]	—	—	57.9%	33.0%	19.6
	SSD[33]	VGG-16	79.8%	78.5%	48.5%	28.8%	19
	DSSD[34]	ResNet-101[12]	81.5%	80.0%	53.3%	33.2%	5.5
	RSSD[35]	VGG-16	80.8%	—	—	—	16.6
	FSSD[36]	VGG-16	80.9%	84.2%	52.8%	31.8%	35.7
	ASSD[37]	ResNet-101	83%	81.3%	55.5%	34.5%	2.7
	RetinaNet[39]	ResNet-101-FPN	—	—	59.1%	39.1%	8.2
基于 anchor-free	CornerNet[42]	Hourglass[15]	—	—	57.8%	42.1%	4.1
	ExtremeNet[47]	Hourglass	—	—	60.5%	43.7%	3.1
	CenterNet[51]	Hourglass	—	—	64.5%	47.0%	3.7
	CenterNet[52]	Hourglass	—	—	63.9%	45.1%	7.8
	FSAF[53]	ResNet-101	—	—	63.1%	42.8%	—
	FCOS[54]	ResNet-101-FPN	—	—	60.7%	41.5%	—
	FoveaBox[59]	ResNet-101	—	—	60.1%	40.6%	—

够在半自动标注方式的基础上实现全自动数据标注方式，不仅节省更多数据标注的时间成本和人力成本，还解决了数据集不足的问题，更是促进了图像目标检测的发展。

4.2 多源数据融合

图像目标检测的场景比较复杂，目标周围的信息对物体的检测具有十分重要的作用，为了更快速有效地检测出感兴趣目标，综合利用图像、视频、音频、传感器等多源数据构建多模态、多尺度、多颗粒的目标检测模型是未来的难点。Yu 等[66]提出了一种基于多源异构数据融合的危险化学品安全状态评估模型，诸如射频识别（radio frequency identification，RFID）、传感器、摄像机、警报器、卫星定位和环境监测之类的传感设备被用于构建面向危险化学品的物联网。Liu 等[67]提出一种基于熵特征轻量级神经网络的多源特征融合的异构虹膜识别方法，该方法从 JLU 虹膜库中的 3 个不同设备收集的虹膜数据，基于统计学习思想和多源特征融合机制设计虹膜特征标签的信息熵用于设置虹膜熵特征类别标签，并根据类别标签设计识别功能以获得识别结果。

4.3 优化损失函数

在 One-Stage 检测网络中主要是通过同时优化分类损失和定位损失来训练检测模型，由于锚框数量巨大而导致正负样本不均衡的问题比较严重。为了解决此问题，很多检测网络如 YOLO[28]

和 RetinaNet[39]均引入了新的分类损失函数，通过对样本重新赋权值来解决样本不平衡的问题，但是并未充分考虑不同样本之间的联系，无法很好地适应不同类型的数据集，因此对于设计损失函数还有待创新。2019 年 Chen 等[68]提出一种用排序来代替分类损失的方法，通过 AP-loss（average-precision loss）解决排序问题，并且提出一种在感知学习中可以将误差驱动更新机制和深度网络中反向传播机制进行巧妙结合的优化方法，该方法可以显著提高 One-Stage 网络的检测性能。

4.4　减少数据依赖

常用的图像目标检测模型均为有监督学习，即仅能对训练过的目标进行检测和识别，无法检测未训练过的目标，对数据有着极大的依赖性。随着互联网技术的发展，目前有许多基于弱监督的算法与框架陆续被提出，例如 Lin 等[69]提出了一种基于弱监督的端到端的目标实例挖掘（object instance mining，OIM）框架，OIM 无须任何标注，仅通过在空间图和外观图上引入信息传播来检测图像中所有可能存在的目标。Wang 等[70]提出了一种学习成对关系的迭代算法，该算法基于图像中的像素之间存在很强的成对关系，可以将稀疏图传播到更密集的关系，然后将成对网络的改进结果用作监督来训练 Unary 网络，并反复进行该过程以获得更好的分割结果。如果能够在弱监督的基础上实现无监督学习，将增强网络的泛化能力，实现对未训练目标的检测，提升模型在场景中的应用能力。

4.5　模型轻量化

深度神经网络已经可以有效地解决诸多领域的图像问题，但是随着移动互联网技术的不断发展，用户对便携性设备的需求越来越迫切，因此在保证检测精度和检测速度不降低的情况下，设计一种高性能、轻量化的检测模型是非常有必要的。目前构建轻量级的神经网络比较流行的方法有模型压缩算法和自动化神经网络架构设计，知识蒸馏是模型压缩的最有效方法之一，关键思想是从深层 teacher 模型（T）将知识转移到浅层 student（S）。Gao 等[71]为了从 T 中获得更好的知识，提出一种残差知识蒸馏，该方法通过引入 assistant（A）进一步蒸馏知识，具体来说，训练 S 来模仿 T 的特征图，而 A 则是通过学习它们之间的残差来辅助此过程。Yang 等[72]提出了一种新的知识蒸馏方法，具体来说是通过一种基于自适应实例归一化的新损失来有效地传递特征统计量，主要思想是通过自适应实例归一化（以 student 为条件）将学习到的统计信息回传给 teacher，并让 teacher 网络通过损失评估 student 学习到的统计信息是否可靠。

5　结束语

本文首先对基于深度学习的图像目标检测模型中常用的卷积神经网络进行了简单的介绍；其次从候选区域、回归和 anchor-free 方法的角度对现有的经典图像目标检测模型进行较为详细的综述，总结了各模型相比之前模型的改进策略，以及自身的创新点和优缺点；然后基于当前发展形势分析了图像目标检测的今后发展与技术动向。

目前，图像目标检测技术已经比较成熟，并且实验中检测精度与检测速度也已达到较高水准，但在真实场景中的检测结果相比实验数据仍有一些差距。因此，图像目标检测在未来仍具有一定的挑战难度，特别是对于不同场景的适用性、小样本训练模型和轻量化模型设计等方面仍有很大的提升空间和发展潜力。

参考文献：

[1] 刘芳, 杨安喆, 吴志威. 基于自适应 Siamese 网络的无人机目标跟踪算法[J]. 航空学报, 2020, 41(1): 248-260.
LIU F, YANG A Z, WU Z W. Adaptive siamese network based UAV target tracking algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(1): 248-260.

[2] 陈莹莹, 房胜, 李哲. 加权多特征外观表示的实时目标追踪[J]. 中国图象图形学报, 2019, 24(2): 291-301.

CHEN Y Y, FANG S, LI Z. Real-time visual tracking via weighted multi-feature fusion on an appearance model[J]. Journal of Image and Graphics, 2019, 24(2): 291-301.

[3] 何冰倩, 魏维, 张斌. 基于深度学习的轻量型的人体动作识别模型[J]. 计算机应用研究, 2020, 37(8): 1-6.
HE B Q, WEI W, ZHANG B. Lightweight human action recognition model based on deep learning[J]. Application Research of Computers, 2020, 37(8): 1-6.

[4] 罗会兰, 童康. 时空压缩激励残差乘法网络的视频动作识别[J]. 通信学报, 2019, 40(10): 189-198.
LUO H L, TONG K. Spatiotemporal squeeze-and-excitation residual multiplier network for video action recognition[J]. Journal on Communications, 2019, 40(10): 189-198.

[5] UIJLINGS J, VAN D S K, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[6] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 27th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[7] 苏赋, 吕沁, 罗仁泽. 基于深度学习的图像分类研究综述[J]. 电信科学, 2019, 35(11): 58-74.
SU F, LV Q, LUO R Z. Review of image classification based on deep learning[J]. Telecommunications Science, 2019, 35(11): 58-74.

[8] 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.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[10] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of 13th European Conference on Computer Vision. Berlin: Springer-Verlag, 2014: 818-833.

[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of 3rd International Conference on Learning Representations. [S.l.:s.n.], 2015.

[12] LUO W J, LI Y J, URTASUN R, et al. Understanding the effective receptive field in deep convolutional neural networks[J]. arXiv: 1701.04128, 2017.

[13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 770-778.

[14] SQUARTINI S, PAOLINELLI S, PIAZZA F. Comparing different recurrent neural architectures on a specific task from vanishing gradient effect perspective[C]//Proceedings of 2006 IEEE International Conference on Networking, Sensing and Control. Piscataway: IEEE Press, 2006: 380-385.

[15] PASCANU R, MIKOLOV T, BENGIO Y. Understanding the exploding gradient problem[J]. arXiv:1211.5063, 2012.

[16] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]//Proceedings of 21st ACM Conference on Computer and Communications Security. Berlin: Springer-Verlag, 2016: 483-499.

[17] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010: 3485-3492.

[18] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of 13th European Conference on Computer Vision. Berlin: Springer-Verlag, 2014: 740-755.

[19] DONAHUE J, JIA Y, VINYALS O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition[C]//Proceedings of 31st International Conference on Machine Learning. New York: ACM Press, 2014: 988-996.

[20] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS—improving object detection with one line of code[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5562-5570.

[21] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-1916.

[22] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision. Washington: IEEE Computer Society Press, 2015: 1440-1448.

[23] YING Z, LI B, LU H, et al. Sample-specific SVM learning for person re-identification[C]//Proceedings of IEEE Conference on Computer Vision & Pattern Recognition. Washington: IEEE Computer Society Press, 2016: 1278-187.

[24] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6): 1137-1149.

[25] HE K, GEORGIA G, PIOTR D, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 1.

[26] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[27] 阮激扬. 基于 YOLO 的目标检测算法设计与实现[D]. 北京: 北京邮电大学, 2019.
RUAN J Y. Design and implementation of object detection al-

gorithm based on YOLO[D]. Bejing: Beijing University of Posts and Telecommunications, 2019.

[28] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society Press, 2016: 429-442.

[29] REDMON J, FARAFADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recongnition. Piscataway: IEEE Press, 2017: 6517-6525.

[30] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of International Conference on Machine Learning. [S.l.:s.n.], 2015: 448-456.

[31] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks[J]. arXiv:1608.06019, 2016.

[32] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv: 1608.06019, 2018.

[33] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of Computer Vision-ECCV. Springer: International Publishing, 2016: 21-37.

[34] FU C Y, LIU W, RANGA A, et al. DSSD: deconvolutional single shot detector[J]. arXiv: 1701.06659, 2017.

[35] JISOO J, HYOJIN P, NOJUN K. Enhancement of SSD by concatenating feature maps for object detection[J]. arXiv: 1705.09587, 2017.

[36] LI Z, ZHOU F Q. FSSD: feature fusion single shot multibox detector[J]. arXiv: 1512.02325, 2017.

[37] YI J, WU P, METAXAS D N. ASSD: attentive single shot multibox detector[J]. Computer Vision and Image Understanding, arXiv: 1909.12456, 2019.

[38] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[39] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017(99): 2999-3007.

[40] ERHAN D, SZEGEDY C, TOSHEV A, et al. Scalable object detection using deep neural networks[J]. arXiv: 1312.2249, 2013.

[41] LIN T Y, PIOTR D, GIRSHICK R, et al. Feature pyramid networks for object detection[J]. arXiv: 1612.03144, 2016.

[42] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2018: 734-750.

[43] NEWELL A, HUANG Z, DENG J, et al. Associative embedding: end-to-end learning for joint detection and grouping[C]//Proceedings of Neural Information Processing Systems. Cambridge: MIT Press, 2017: 2277-2287.

[44] 唐心宇, 宋爱国. 人体姿态估计及在康复训练情景交互中的应用[J]. 仪器仪表学报, 2018, 39(11): 198-206.
TANG X Y, SONG A G. Human pose estimation and its implementation in scenario interaction system of rehabilitation training[J]. Chinese Journal of Scientific Instrument, 2018, 39(11): 198-206.

[45] GATTUPALLI S. Human motion analysis and vision-based articulated pose estimation[C]//Proceedings of International Conference on Healthcare Informatics. Piscataway: IEEE Press, 2015: 470-470.

[46] HUANG Z, LIU Y, FANG Y, et al. Video-based fall detection for seniors with human pose estimation[C]//Proceedings of 4th IEEE International Conference on Universal Village 2018. Piscataway: IEEE Press, 2018: 1-4.

[47] ZHOU X Y, ZHOU J C, KRHENBUHL P. Bottom-up object detection by grouping extreme and center points[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 850-859.

[48] CAO Z, SIMON T, WEI S, et al. Realtime multi person 2d pose estimation using part affinity fields[J]. arXiv: 1611.08050, 2017.

[49] CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7103-7112.

[50] XIAO B, WU H P, WEI Y C. Simple baselines for human pose estimation and tracking[J]. arXiv:1804.06208, 2018.

[51] DUAN K, BAI S, XIE L, et al. CenterNet: keypoint triplets for object detection[C]//Proceedings of International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6568-6577.

[52] ZHOU X Y, WANG D Q, KRHENBUHL P. Objects as points[J]. arXiv: 1904.07850, 2019.

[53] ZHU C C, HE Y H, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 840-849.

[54] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9626-9635.

[55] HE T, SHEN C H, TIAN Z, et al. Knowledge adaptation for efficient semantic segmentation[J]. arXiv: 1903.04688, 2019.

[56] LIU Y F, CHEN K, LIU C, et al. Structured knowledge distillation for semantic segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2599-2608.

[57] LONG J, SHELHAMER E, DARRELL T. Fully convolutional

networks for semantic segmentation[C]//Proceedings of IEEE Conference on Computer Vision & Pattern Recognition. Piscataway: IEEE Press, 2015.
[58] TIAN Z, HE T, SHEN C H, et al. Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation[J]. arXiv: 1903.02120, 2019.
[59] KONG T, SUN F C, LIU H P, et al. FoveaBox: beyond anchor-based object detector[J]. arXiv:1904.03797, 2019.
[60] 邢惠钧, 昌硕. 基于移动小车的行人监控系统[J]. 电信科学, 2017, 33(2): 120-127.
XING H J, CHANG S. Pedestrian surveillance system based on mobile vehicle[J]. Telecommunications Science, 2017, 33(2): 120-127.
[61] 杨恩泽. 基于深度学习的交通车辆检测与识别算法研究[D]. 北京: 北京交通大学, 2019.
YANG E Z. Vehicle detection and recognition in traffic scenes based on deep learning[D]. Beijing: Beijing Jiaotong University, 2019.
[62] 王忠玉. 智能视频监控下的安全帽佩戴检测系统的设计与实现[D]. 北京: 北京邮电大学, 2018.
WANG Z Y. Design and implementation of detection system of wearing helmets based on intelligent video surveillance[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.
[63] 陈虹, 郭露露, 宫洵, 等. 智能时代的汽车控制[J]. 自动化学报, 2019, 45(x): 1-21.
CHEN H, GUO L L, GONG X, et al. Automotive control in intelligent era[J]. Acta Automatica Sinica, 2019, 45(x): 1-21.
[64] RUSSELL B C, TORRALBA A, MURPHY K P, et al. LabelMe: a database and Web-based tool for image annotation[J]. International Journal of Computer Vision, 2008, 77(1): 157-173.
[65] HIDAYATULLAH P, MENGKO T E R, MUNIR R, et al. A semiautomatic sperm cell data annotator for convolutional neural network[C]//Proceedings of 5th International Conference on Science in Information Technology. [S.l.:s.n.], 2019: 211-216.
[66] YU J, MA Z H, WU D, et al. The safety state control of hazardous chemicals based on multi-source heterogeneous data fusion[C]//Proceedings of 7th International Conference on Computer Science and Network Technology. Piscataway: IEEE Press, 2019: 156-159.
[67] LIU S, LIU Y, ZHU X, et al. Multi-source feature fusion and entropy feature lightweight neural network for constrained multi-state heterogeneous iris recognition[J]. IEEE Access, 2020: 1.
[68] CHEN K, LI J, LIN W, et al. Towards accurate one-stage object detection with AP-loss[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5114-5122.
[69] LIN C H, WANG S, XU D, et al. Object instance mining for weakly supervised object detection[C]//Proceedings of 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020.
[70] WANG X, LIU S F, MA H M, et al. Weakly-supervised semantic segmentation by iterative affinity learning[J]. International Journal of Computer Vision, 2020: 1-14.
[71] GAO M, SHEN Y J, LI Q Q, et al. Residual knowledge distillation[J]. arXiv: 2002.09168, 2020.
[72] YANG J, MARTINEZ B, BULAT A, et al. Knowledge distillation via adaptive instance normalization[J]. arXiv: 2003.04289, 2020.

[作者简介]

张婷婷（1995– ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为计算机视觉与人工智能等。

章坚武（1961– ），男，博士，杭州电子科技大学通信工程学院教授、博士生导师，中国电子学会高级会员，浙江省通信学会常务理事，主要研究方向为移动通信、多媒体信号处理与人工智能、通信网络与信息安全。

郭春生（1971– ），男，博士，杭州电子科技大学通信工程学院副教授、硕士生导师，主要研究方向为视频分析与模式识别。

陈华华（1975– ），男，博士，杭州电子科技大学通信工程学院副教授、硕士生导师，主要研究方向为视频分析与模式识别。

周迪（1975– ），男 ，浙江宇视科技有限公司教授级高级工程师、宇视研究院院长，主要研究方向为视频安全、人工智能等。

王延松（1970– ），男 ，之江实验室研究员，教授级高工，科技部“宽带通信与新型网络”领域总体组专家、指南编制组专家，工信部“网络通信技术”领域咨询专家、中国通信学会委员、中国通信标准化协会工业互联网 ST8 组副组长等职务。主要研究方向为工业互联网、SDN/NFV、网络安全等。

徐爱华（1989– ），女，浙江宇视科技有限公司工程师，主要研究方向为视频安全、人工智能等。