基于深度学习的目标检测研究综述

罗会兰,陈鸿坤
(江西理工大学信息工程学院,江西赣州 341000)

摘 要: 目标检测是计算机视觉领域内的热点课题,在机器人导航、智能视频监控及航天航空等领域都有广泛的应用.本文首先综述了目标检测的研究背景、意义及难点,接着对基于深度学习目标检测算法的两大类进行综述,即基于候选区域和基于回归算法.对于第一类算法,先介绍了基于区域的卷积神经网络(Region with Convolutional Neural Network,R-CNN)系列算法,然后从四个维度综述了研究者在R-CNN系列算法基础上所做的研究:对特征提取网络的改进研究、对感兴趣区域池化层的改进研究、对区域提取网络的改进研究、对非极大值抑制算法的改进研究.对第二类算法分为YOLO(You Only Look Once)系列、SSD(Single Shot multibox Detector)算法及其改进研究进行综述.最后根据当前目标检测算法在发展更高效合理的检测框架的趋势下,展望了目标检测算法未来在无监督和未知类别物体检测方向的研究热点.

关键词: 目标检测; 深度学习; 特征提取; 计算机视觉; 视频监控; 图像处理; 卷积神经网络
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2020)06-1230-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.06.026

Survey of Object Detection Based on Deep Learning

LUO Hui-lan, CHEN Hong-kun
(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*)

Abstract: Object detection is a hot topic in the field of computer vision, and has been widely used in robot navigation, intelligent video surveillance, aerospace, and other fields. The research background, significance and challenges of object detection were introduced. Then the object detection algorithms based on deep learning were reviewed according to two categories: candidate region-based and regression-based. For the candidate region-based algorithms, we first introduced the R-CNN (Region with Convolutional Neural Network) based series of algorithms, and then the R-CNN based methods were overviewed from four dimensions: the research of feature extraction networks, the region of interesting pooling researches, improved works based on region proposal networks, and some improved approaches of non maximum suppression algorithms. Next, the regression-based algorithms were surveyed in terms of YOLO(You Only Look Once) series and SSD(Single Shot multibox Detector) series. Finally, according to the current trend of object detection algorithms that are developing more efficient and reasonable detection frameworks, the future research focuses of unsupervised and unknown category object detection directions were prospected.

Key words: object detection; deep learning; feature extraction; computer vision; video surveillance; image processing; convolutional neural network

1 引言

目标检测的主要任务是从输入图像中定位感兴趣的目标,然后准确地判断每个感兴趣目标的类别.当前目标检测技术已经广泛应用于日常生活安全、机器人导航、智能视频监控、交通场景检测及航天航空等领域.同时目标检测是行为理解、场景分类和视频内容检索等其他高级视觉问题的基础.但是,由于同一类物体的不同实例间可能存在很大的差异性,而不同类物体间可能非常相似,以及不同的成像条件和环境因素会对物体的外观产生巨大

收稿日期:2019-08-13;修回日期:2019-11-27;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61862031,No.61462035);江西省机器视觉及智能系统重点实验室(No.20181BCD40009);江西省赣州市"科技创新人才计划"项目

的影响[1],使得目标检测具有很大的挑战性.

传统的目标检测算法采用类似穷举的滑动窗口方式或图像分割技术来生成大量的候选区域,然后对每一个候选区域提取图像特征(包括 HOG[2]、SIFT[3]、Haar[4]等),并将这些特征传递给一个分类器(如 SVM[5]、Adaboost[6]和 Random Forest[7]等)用来判断该候选区域的类别. 由于传统方法提取的特征存在局限性,产生候选区域的方法需要大量的计算开销,检测的精度和速度远远达不到实际应用的要求,这使得传统目标检测技术研究陷入了瓶颈[8].

近些年基于深度学习的目标检测算法形成两大类别:基于候选区域和基于回归. 基于候选区域的目标检测算法也称为二阶段方法,将目标检测问题分成两个阶段:一是生成候选区域(region proposal),二是把候选区域放入分类器中进行分类并修正位置. 基于回归的目标检测算法只有一个阶段,直接对预测的目标物体进行回归.

Sharma 等人[9,10]仅仅综述了传统的目标检测算法,Chahal 等人[11]对基于深度学习的目标检测算法从算法实现的角度进行了综述,Kemal 等人[12]从目标检测算法中不平衡问题的角度进行了综述,Zhao 等人[13]从检测框架和检测子任务两个角度进行了综述. 与以上研究综述不同的是,本文从一个新颖的角度归类综述了近些年目标检测领域的经典算法. 在将其分为基于候选区域和基于回归两大类的前提下,对基于候选区域的目标检测算法,介绍基于区域的卷积神经网络(Region with Convolutional Neural Network,R-CNN)系列算法的发展史后,根据对不同模块的改进研究进行归类综述:特征提取网络、感兴趣区域池化(Region of Interesting Pooling,ROI Pooling)层、区域提取网络(Region Proposal Networks,RPN)、非极大值抑制(Non Maximum Suppression,NMS). 对基于回归的目标检测算法,介绍 YOLO(You Only Look Once)系列和 SSD(Single Shot Multibox Detector)算法后,对基于 SSD 算法的改进研究进行细分论述:基于 Anchor-based 的改进研究和基于 Anchor-free 的改进研究. 随后介绍目标检测领域流行的数据集. 最后展望未来目标检测研究的发展方向.

2　基于候选区域的目标检测算法综述

本节主要将近年来基于候选区域的目标检测算法分为五个部分进行综述,首先介绍了 Faster R-CNN[14]框架的发展历程,然后综述了对 Faster R-CNN 算法的四个重要组成部分(特征提取网络、ROI Pooling 层、RPN、NMS 算法)的改进研究.

2.1　R-CNN 系列基础框架的发展史

2014 年,Girshick 等人[15]成功将卷积神经网络(Convolutional Neural Networks,CNN[16])运用在目标检测领域中,提出了 R-CNN 算法,它将 AlexNet[17]与选择性搜索[18](selective search)算法相结合,把目标检测任务分解为若干个独立的步骤(如图 1 所示),首先采用选择性搜索算法提取 2000 个候选区域,然后对每个候选区域进行归一化,并逐个输入 CNN 中提取特征,最后对特征进行 SVM 分类和区域回归.

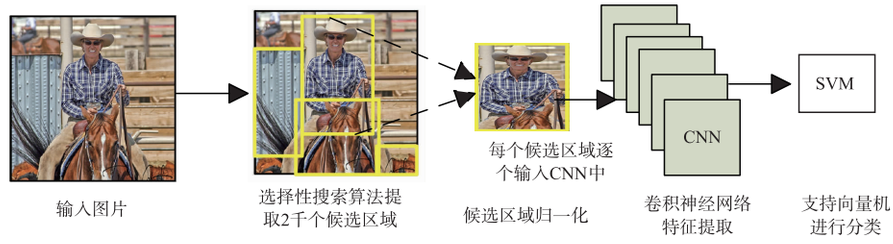

图1　R-CNN模型结构

R-CNN[15]算法在 PASCAL VOC2007[19]数据集上的检测精度达到了 58.5%,相较于传统的目标检测算法取得了跨越性的进展. 但还存在非常多的改进空间,如:对于单张图像提取的 2000 个候选区域需要逐个输入 CNN 中,导致计算开销十分巨大,严重影响了检测速度;而且候选区域输入 CNN 前,必须剪裁或缩放至固定大小,这会使候选区域发生形变且丢失较多的信息,导致网络检测精度下降.

2014 年 He 等人[20]提出了空间金字塔网络(Spatial Pyramid Pooling Network,SPP-Net)检测算法,它在 CNN 最后一层卷积层和全连接层之间加入 SPP 层(如图 2 所示),使得网络能够输入任意尺度的候选区域,从而每张输入图片只需一次 CNN 运算,就能得到所有候选区域的特征,这使得计算量大大减少. SPP-Net 的检测速率比 R-CNN 快了 24 ~ 102 倍,并打破了固定尺寸输入的束缚.

2015 年,Girshick 等人[21]提出了 Fast R-CNN 算法(如图 3 所示),他们受到 SPP-Net 算法的启发,将 SPP 层简化成单尺度的 ROI Pooling 层以统一候选区域特征的大小,而且进一步提出了多任务损失函数思想,将分类损失和边界框回归损失统一训练学习,使得分类和定位任务不仅可以共享卷积特征,还可以相互促进提升检测效果.

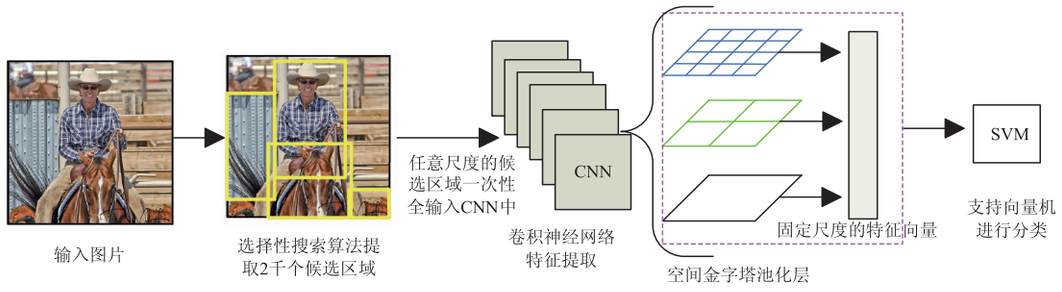

图2 空间金字塔网络结构

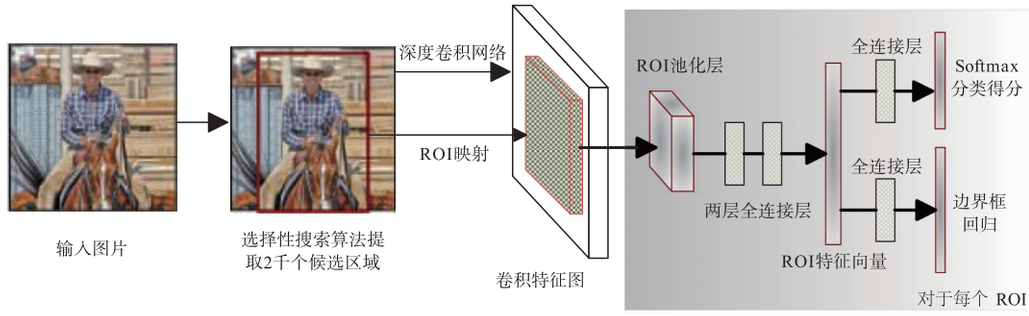

图3 Fast R-CNN结构

虽然 Fast R-CNN 有效地加快了检测速率,但仍然依赖于选择性搜索算法[18]来产生候选区域. 有研究表明,卷积神经网络的卷积层具有良好的定位目标的能力,只是这种能力在全连接层被削弱了. 因此,2015 年 Ren 等人[14]提出了 Faster R-CNN 算法框架(结构如图4所示),设计了辅助生成样本的 RPN 取代选择性搜索算法. RPN 是一种全卷积神经网络(Fully Convolutional Network,FCN[22])结构,它将任意大小的特征图作为输入,经过卷积操作后产生一系列可能包含目标的候选区域,使算法实现了端到端的训练,极大提高了检测速度.

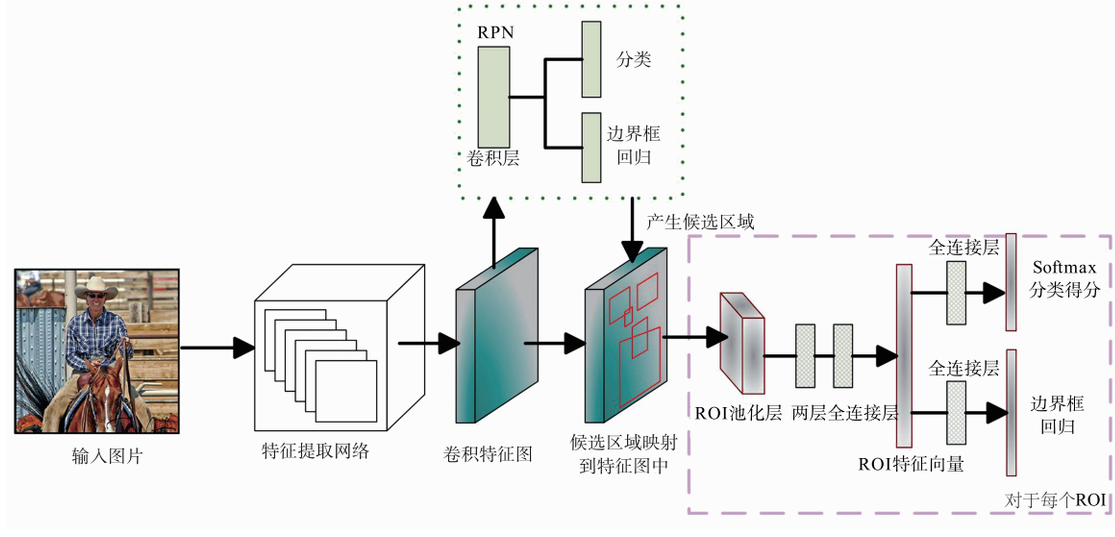

图4 Faster R-CNN结构

2.2 基于 Faster R-CNN 的改进研究

Faster R-CNN[14]算法在检测的精度和速度上都取得了不错的效果. 它主要由四个模块组成:特征提取网络用于提取图像特征;ROI Pooling 层将不同大小的候选区域特征进行归一化输出;RPN 根据图像特征生成目标的候选区域; NMS[23]算法用于去除冗余检测框. 本小节综述了在这四个功能模块上的改进研究.

2.2.1 对特征提取网络的改进研究

深度卷积神经网络的浅层特征具有丰富的几何信息,但对语义信息不敏感,不利于目标分类;而高层具有丰富的语义信息,但分辨率太低,不利于目标定位.仅利用最后一层卷积层的特征进行不同尺度目标的预测,效果显然是不理想的,所以 Faster R-CNN 算法对于小目标的检测精度较低.针对这个问题,有许多研究是通过融合多个卷积层的特征来提高小尺度目标的检测效果.

2016 年 Kong 等人[24]提出了 HyperNet 算法,通过融合多层卷积层的特征图,得到具有多尺度信息的 Hyper 特征,该特征结合了卷积层高层的强语义信息、中层的辅助信息以及浅层的几何信息.同年,Huang 等人[25]采用多尺度思想,在特征提取网络的高层和低层中提取多个不同尺度的特征分别进行预测.

2017 年 Lin 等人[26]提出了特征金字塔网络(Feature Pyramid Network,FPN),FPN 构造了一种自顶向下带有横向连接的层次结构,提取多个不同尺度特征用于检测,每个尺度特征都是高层特征与浅层特征融合所得,不仅具有较强的语义信息,还具有较丰富的几何信息.

2018 年 Bharat 等人[27]提出了图像金字塔的尺度归一化方法(Scale Normalization for Image Pyramids,SNIP).他们借鉴多尺度训练思想,使用图像金字塔网络将图像生成三种不同分辨率的输入图像,高分辨率图像只用于小目标检测,中等分辨率图像只进行中等目标检测,低分辨率图像只进行大目标检测.

2.2.2 对感兴趣区域池化层的改进研究

ROI Pooling,即感兴趣区域池化是将候选区域对应的特征图划分成固定数量的空间小块,再对每个空间小块进行最大池化或者平均池化操作,这样就实现了不同尺度的候选区域能够输出同样大小的特征图.近年来的改进研究旨在更好保留或融合空间位置信息到 ROI 池化中,以提高检测效果.

2016 年 Dai 等人[28]提出了基于区域的全卷积神经网络(Region-based Fully Convolutional Network,R-FCN),他们考虑到目标检测任务是由分类任务和定位任务组成,分类任务要求目标特征具有平移不变性,而定位任务要求目标特征具有平移敏感性.为了缓解这两者间的矛盾,提出了位置敏感 ROI 池化,可以编码每个候选区域的相对空间位置信息,使得特征具有了对位置的敏感性.在此基础上,Zhu 等人[29]提出了 CoupleNet 算法,设计了由两个分支组成的耦合模块,一个分支采用位置敏感 ROI 池化获取对象的局部信息,另一分支则使用两个 ROI 池化分别获取对象的全局信息和上下文信息,然后有效的结合候选区域的局部信息、全局信息和上下文信息进行检测.

2017 年 Dai 等人[30,31]提出了形变卷积网络(Deformation Convolution Network,DCN),设计了可形变卷积和可形变 ROI 池化层.它们的感受野不再是一成不变的正方形,而是和物体的实际形状相匹配,缓解了物体形变问题,使网络学习了更多的空间位置信息,增强了定位能力.

2017 年 He 等人[32]提出了 Mask R-CNN 算法,为了解决特征图和原始图像上的感兴趣区域出现不对准问题提出了 ROI Align 层,并且增加了 Mask 预测分支,可以并行实现像素级的语义分割任务.而 2018 年 Jiang 等人[33]进一步改进了 ROI Pooling 提出了精准的感兴趣区域池化(Precise ROI Pooling,PrROI Pooling).ROI Pooling 采用的是最近邻插值方法,它在将 ROI 映射到特征图时和将 ROI 划分池化区域时都存在取整近似运算,会丢失部分空间位置信息;ROI Align 则取消了所有的取整运算,采用双线性插值的方法计算每个空间块的值,但只考虑 N 个插值点的值,而且 N 的大小是预定义的,不能根据空间块的大小进行调整;而 PrROI Pooling 是采用二阶积分的方法对空间块进行池化操作,使感兴趣区域保持更多的空间位置信息,实现更精准定位.

2.2.3 对区域提取网络的改进研究

RPN 是 Faster R-CNN 算法的主要创新点,它主要基于 Anchor 机制来产生大量目标候选区域.近年来的改进研究旨在产生更精确的候选区域,以提高检测效果.

2017 年,Zhao 等人[34]提出了 Cascade R-CNN 算法,通过级联三个区域交并比(Intersection Over Union,IOU)阈值递增的 R-CNN[15]检测模型,对 RPN 产生的候选区域进行筛选,留下高 IOU 值的候选区域,有效提高了模型的检测精度.与此不同,2018 年 Chen 等人[35]在 RPN 阶段引入上下文信息对候选区域进行微调,使得网络定位的更加准确.

针对 RPN 中的 Anchor 机制需要人工预先设定尺度大小和长宽比等超参数的问题,2019 年,Wang 等人[36]提出了 Guided-Anchoring 方法,通过图像特征来指导 Anchor 的生成.它由 Anchor 生成模块和特征自适应模块组成,其中 Anchor 生成模块采用两个分支分别预测 Anchor 的位置和形状:位置预测分支预测出哪些区域作为中心点来生成 Anchors;形状预测分支则是根据位置预测分支得到的中心点来预测 Anchor 最佳的长和宽.特征自适应模块根据生成的 Anchor 的形状,使用一个 3×3 的可形变卷积来修正特征图,以适应 Anchor 的形状.

2.2.4 对 NMS 的改进研究

NMS 算法首先人工设定一个 IOU 阈值,将同一类的所有检测框按照分类置信度排序,选取分类置信度得分

最高的检测结果,去除那些与之 IOU 值超过阈值的相邻结果,使网络模型在召回率和精度之间取得较好的平衡.

NMS 算法采用单一的 IOU 阈值会导致漏检情况发生,为了解决这个问题,2017 年,Bodla 等人[37]提出了 Soft NMS 算法,它不是直接去除那些超过 IOU 阈值的相邻结果,而是采用线性或者高斯加权的方式衰减它的置信度值,再选取合适的置信度阈值进行检测框去重,对模型的漏检有了很好的改善. 在此基础上,He 等人[38]提出了 Softer NMS 算法,不是直接选取分类置信度得分最高的检测框作为最终检测结果,而是将与分类置信度最高的检测框的交并比值大于一定阈值的所有检测框的坐标进行加权平均,作为最终检测结果,从而能够更准确的定位物体.

2018 年,Hu 等人[39]提出目标关系模块(Relation Module,RM)替代了 NMS 算法来对目标的检测框进行去除冗余操作. RM 借鉴了文献[40]的思想对不同目标间的关系进行建模,并引入了注意力机制来优化检测效果. 同年,Jiang 等人[33]发现检测结果中存在分类置信度和定位准确度之间不匹配问题,所以提出了 IOU-guided NMS[33]方法. 他们将预测的检测框与真值间的 IOU 值作为定位置信度,每一类根据定位置信度进行排序,从而改进了 NMS 过程,保留了定位更准确的检测框.

针对常用的边界框回归损失函数(L1 范数或 L2 范数)与 IOU 没有强相关性,不能很好度量检测框准确性的问题,2019 年 Hamid 等人[41]提出了 GIOU 作为边界框回归损失函数,在计算检测框与真值框 IOU 的基础上,添加了对这两个框的最小闭包区域面积的计算,通过 IOU 减去两框非重叠区域占最小闭包区域的比重得到 GIOU,其保留了 IOU 的原始性质的同时弱化了它的缺点,对边界框的定位能力上有了大幅度的提升.

3 基于回归的目标检测算法综述

基于回归的目标检测算法不需要候选区域生成分支,对给定输入图像,直接在图像的多个位置回归出目标的候选框和类别. 本文将分成两大系列来综述基于回归的目标检测算法:YOLO[42]系列和 SSD[43]系列.

3.1 YOLO 系列目标检测算法

2015 年 Redmon 等人[42]提出了 YOLO 算法,将分类、定位、检测功能融合在一个网络当中,输入图像只需要经过一次网络计算,就可以直接得到图像中目标的边界框和类别概率. 如图 5 所示,YOLO 算法将整张输入图像划分成 $S\times S$ 的网格图,每个网格只负责物体中心落在该网格的目标物体以及只预测 B 个边界框信息,然后选择合适的置信度阈值去除那些存在目标可能性低的边界框. 虽然 YOLO 算法完全舍弃了候选区域生成步骤,极大提高了检测速率,能满足实时目标检测的速度要求,但由于其网络设计比较粗糙,远远达不到实时目标检测的精度要求,而且存在目标不能精准定位、容易漏检,小目标和多目标检测效果不好等问题.

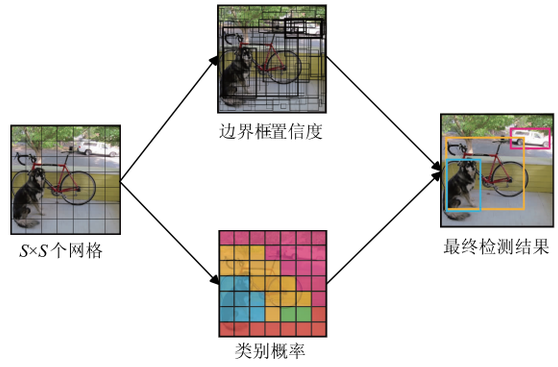

图5 YOLO算法结构

2017 年 Redmon 等人[44]提出了 YOLOv2 算法,对 YOLO 算法进行了一系列改进,重点解决召回率低和定位精度差的问题. 它借鉴了 Faster R-CNN 算法的 Anchor 机制,移除了网络中的全连接层,使用卷积层预测检测框的位置偏移量和类别信息. 而且不同于原 Anchor 机制的手工设计,它利用 K-Means 聚类方式在训练集中学习最佳的初始 Anchor 模板. 不仅如此,YOLOv2 添加了一个 pass-through 层,将浅层的特征图连接到深层的特征图,使网络具有了细粒度特征. 此外,YOLOv2 可以采用多种数据集联合优化训练的方式,利用 WordTree 方法在 ImageNet[45]分类数据集和 MS COCO[46]检测数据集上同步训练,实现超过 9000 个目标类别的实时检测任务.

2018 年 Redmon 等人[47]提出了 YOLOv3 算法,它借鉴残差网络中跳跃连接的思路,构建了名为 DarNet-53 的 53 层基准网络,该网络只采用 3×3 和 1×1 的卷积层,具有与 ResNet-152[48]相仿的分类准确率,但大大减少了计算量;为了处理多尺度目标,采用了 3 种不同尺度的特征图来进行目标检测,每个特征图都是高层与浅层特征图融合所得;在预测类别时,使用 Logistic 回归方法代替 Softmax 方法,使得每个候选框可以预测多个类别,支持检测具有多个标签的对象. YOLOv3 算法能满足实时检测任务的精度与速率的要求,成为了当前工程界首选的目标检测算法之一.

3.2 SSD 系列目标检测算法

3.2.1 SSD 算法

2016 年 Liu 等人[43]提出了 SSD 算法,在回归思想的基础上,有效结合多尺度检测的思想,提取多个不同尺度的特征图进行检测,遵循较大的特征图用来检测相对较小的目标,较小的特征图检测较大目标的策略,

显著提高了对大目标的检测效果,对小目标检测也有一定的提升. 同时借鉴 Faster R-CNN 算法的 Anchor 机制,对提取的特征图的每个位置上都预设固定数量的不同尺度和长宽比的先验框(default boxes),网络可以直接在特征图上进行密集采样提取候选框进行预测,在保持实时检测速度的同时,提高了模型的定位准确度. 如图 6 所示,SSD 网络是基于全卷积网络结构,它将基础网络 VGG16[49] 的全连接层替换为了卷积层,并在 VGG16[49] 网络末端添加了几个使特征图尺寸逐渐减小的辅助性卷积层,用于提取不同尺度的特征图,而且直接采用卷积操作对不同尺度的特征图进行检测.

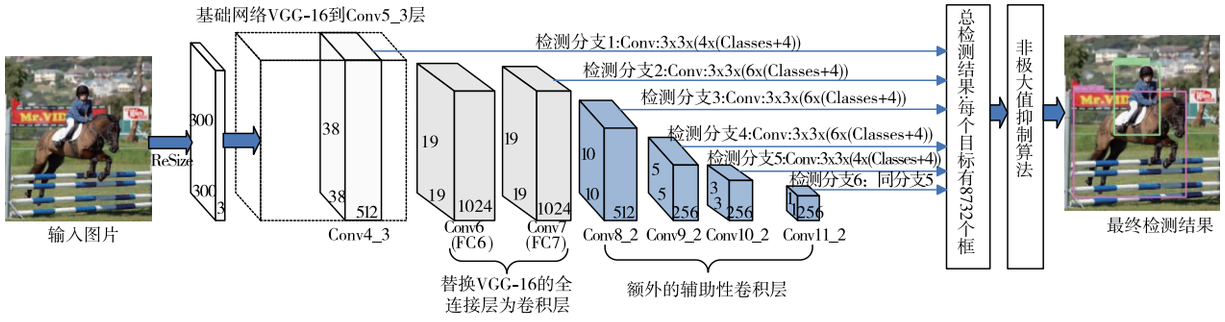

图6　SSD算法网络结构

SSD 算法在检测的速度和精度上都超越了 Faster R-CNN 算法,但 SSD 算法提取的不同卷积层特征独立输入各自的检测分支,容易出现同一个物体被不同大小的边界框同时检测出来的情况,即重复检测问题. 而且每层的检测分支仅关注自己分支上特定尺度的目标,没有考虑到不同层、不同尺度目标间的关联性,所以对小目标检测效果一般.

3.2.2　基于 Anchor-based 方式的改进

2017 年 Jisoo 等人[50] 提出了 RSSD 算法,其在 SSD[43] 算法的基础上,对提取的不同尺度的特征采用了特殊的特征融合方式:对于每个特定的尺度特征,分别将比其大的尺度特征进行池化操作,比其小的尺度特征进行反卷积操作,然后将这些特征进行串接融合形成新的特定尺度特征. 这种融合方式使得每个尺度的特征都具有其他尺度的信息,增加了不同层特征图之间的联系,避免了同一目标重复检测的问题. 同年,Cheng 等人[51] 提出了 DSSD 算法,将 VGG16[49] 替换为 ResNet101[48],增强了网络特征提取能力,并设计了两个特殊的模块:预测模块和反卷积模块. 预测模块的结构类似残差模块,通过跳跃连接实现不同层特征之间的融合,从而提高特征的表征能力. 反卷积模块则是采用反卷积操作建立了一个 Top-to-Down 路径,得到新的不同尺度的特征图,这些特征图融合了高层与浅层特征,引入了丰富的空间上下文信息,使得 DSSD 算法在检测精度上有了大幅度的提升,但检测速度有较大牺牲. 在此基础上,Lin 等人[52] 提出了 RetinaNet 算法,针对 SSD 算法因密集采样导致的难易样本严重失衡问题,提出了 Focal Loss 函数,其是在交叉熵损失函数的基础上添加了两个平衡因子,抑制了简单样本的梯度,将更多的注意力放在难分的样本上. 受 Focal Loss 的启发,Li 等人[53] 提出了梯度协调机制(Gradient Harmonizing Mechanism,GHM)来解决样本失衡问题,这种机制可以同时嵌入分类和回归损失中来平衡训练样本的梯度,不仅减少了易分样本的关注,而且避免了特别难分样本对模型的负面影响.

2018 年 Liu 等人[54] 提出了 RFB-Net 算法,通过模拟人类视觉感受野,设计了感受野模块(Receptive Field Block,RFB)增加网络的特征提取能力. RFB 结构借鉴了 Inception[55] 的思想,引入三个不同扩张率的 3×3 卷积层增大感受野,并且将这三个卷积的输出以串接方式进行特征融合. 此外,Zhang 等人[56] 提出了 RefineDet 算法,结合了一阶段和二阶段检测算法的优点,设计了两个模块:物体检测模块和 Anchor 微调模块,前者对密集的 Anchors 进行筛选去除一些不包含物体的负样本,同时粗调筛选后的 Anchors 位置和尺寸,后者对物体检测模块输出的 Anchors 进一步回归,这使得网络进行了两次回归任务,有效提升了网络定位能力,并且样本的筛选有效缓解了正负样本不均衡问题.

SSD 最新的改进研究[57~61] 更加关注于合理和高效的运用 FPN 结构,提取具有丰富上下文信息和空间信息的多尺度特征,解决目标尺度变化问题. Ghaisi 等人[58] 受到神经结构搜索(Neural Architecture Search,NAS)的启发,提出了 NAS-FPN 算法,该网络模型自动搜索设计最优的 FPN 结构,实现跨尺度的特征融合,在网络性能上超越了 Mask R-CNN,但模型的训练需要大量的 GPU 支持. 此外,Zhao 等人[59] 提出了多层特征金字塔网络(Multi-Level Feature Pyramid Network,MLFPN),通过级联多个小型的 FPN 子网络,形成不同层级

的不同尺度特征,并对特征进行充分的重利用和融合,使网络性能和小目标检测都有很大的提升.

3.2.3 基于 Anchor-free 方式的改进

虽然 SSD 算法借鉴 Anchor 机制的思想大幅度提高了网络的定位能力.但 Anchor 机制中存在两个人工设计的超参数:尺度大小和长宽比.这不仅需要较强的先验知识,而且提取的候选区域太多,增加了计算开销,还引起正负样本不均衡问题,所以有些研究者提出了 Anchor-free 的改进方法.

2018 年,Hei Law 等人[62]提出了 CornerNet 算法,借鉴了文献[63]对关键点检测的思想,采用 Hourglass104 网络[63]作为特征提取网络,直接预测物体的左上角点和右下角点来得到检测框,将目标检测问题当作关键点检测问题来解决.在此基础上,Zhou 等人[64]提出了 ExtremeNet 算法,在关键点选取和关键点组合方式上做出了创新,通过选取物体上下左右四个极值点和一个中心点作为关键点,更加直接关注物体边缘和内部信息,使得检测更加稳定.Duan 等人[65]发现 ConerNet 只使用左右角点会造成大量的误检,为了解决这个问题,提出了 CenterNet 算法,它在 CornerNet 的基础上添加了中心点预测分支,使得组成一个物体检测框的要求不仅仅是左右角点能够匹配,而且检测框的中心点也要有对应的中心点匹配.

上述的 Anchor-free 的方法都是基于人体关键点检测的思想,使用非常庞大的 Hourglass-104[63]网络作为特征提取网络,与此不同的是,Zhi 等人[66]提出了基于全卷积的一阶段目标检测器(Fully Convolutional One-Stage object detection,FCOS),借鉴语义分割任务的思想,采用逐像素预测方式解决目标检测问题,完全避免了与 Anchor 相关的复杂计算和超参数设计,同时使用 FPN 结构实现多尺度目标的预测,每个预测分支中添加了中心点损失来抑制中心点偏差大的检测框,保证每个检测框尽可能靠近目标中心,提高了模型定位能力.

4 相关数据集综述

当前通用目标检测任务中流行的数据集有:PASCAL VOC2007[19]、PASCAL VOC2012[67]、MS COCO[46]、ImageNet[45]、Open Images[68]、LIVS[69]等.

PASCAL VOC[19,67]数据集主要用于图像分类和目标检测任务,主要流行的有 PASCAL VOC2007[19]数据集和 PASCAL VOC2012[67]数据集.它们包含了 20 个常见的类别,每张图片都有与之对应的 XML 文件标注了每个待检测目标的位置和类别.

MS COCO[46]数据集用于目标检测、语义分割、人体关键点检测和字幕生成等任务,对于目标检测任务,它是挑战性最大的数据集之一.该数据集中的目标大部分来自于自然场景,包含日常复杂场景的图像,而且在进行评估时使用更加严格的评估标准,要求算法具有更精确的定位能力.该数据集使用 JSON 格式的标注文件给出每张图片中目标像素级别的分割信息,而且数据集中共包含 80 个对象类别的待检测目标,目标间的尺度变化大,具有较多的小目标物体.

ImageNet[45]数据集用于图像分类、目标检测和场景分类等任务,包含约 1420 万张图片,2.2 万个类别,其中约 103 万张图片拥有明确的类别标注和物体的位置标注.对于目标检测任务,它是具有 200 个对象类别的重要数据集,每张图片的批注都以 PASCAL VOC 数据格式保存在 XML 文件中.

Open Images[68]数据集是对图像分类、目标检测、视觉关系检测和实例分割等任务具有统一注释的单个数据集,对于目标检测任务,它总共包含 190 万张图片和针对 600 个对象类别的 1600 万个边界框,是具有对象位置注释的最大现有数据集.

LIVS[69]数据集是 2019 年提出的大型实例分割数据集,包含了 1000 多个类别,164000 张图像,220 万个高质量的实例分割掩码,这是即将应用于目标检测领域的全新数据集,而且 LIVS 数据集中每个对象类别的训练样本很少,旨在用于目标检测在低样本数量条件下的研究.

5 总结和展望

目标检测是一个十分重要的研究领域,具有广泛的应用前景.本文将近些年涌现的基于深度学习的目标检测算法分为基于候选区域和基于回归的前提下,对这两类算法从发展及不同方向的改进研究角度进行了详细的综述.并介绍了目前目标检测领域流行的数据集.虽然当前目标检测算法在实际生活中得到了广泛应用,但依然存在许多挑战,未来目标检测算法在以下几个方面值得进一步研究:

一是如何有效的结合上下文信息,解决小目标和被遮挡目标在复杂现实场景的检测;二是探索更优的或专门为检测任务设计的特征提取网络,以及更优的检测框选定方法;三是现在的目标检测算法都是基于监督学习,现实中存在海量没有标注的数据,所以研究如何采用弱监督学习的目标检测算法是非常有价值的;四是探索如何从已知类别的目标检测,结合有效语义信息,迁移到未知类别的目标检测也是一个值得研究的方向.

参考文献

[1] LIU Li, OUYANG W, WANG Xiao-gang, et al. Deep Learning for Generic Object Detection: A Survey[OL].

http://dblp.org/abs/ 1809.02165,2018.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE,2005. 886 – 893.

[3] LOWE D G. Distinctive image features from scale-invariant key points[J]. International Journal of Computer Vision, 2004,60(2):91 – 110.

[4] LIENHART R, MAYDT J. An extended set of haar-like features for rapid object detection[A]. Proceedings of the International Conference on Image Processing[C]. USA: IEEE,2002. 900 – 903.

[5] SHAWE-TAYLOR J,CRISTIANINI N. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. England:Cambridge University Press,2000.

[6] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[A]. International Conference on Machine Learning[C]. USA:IMLS,1996. 148 – 156.

[7] LIAW A, WIENER M. Classification and regression by random-forest[J]. R News,2002,2(3):18 – 22.

[8] 姜维,张重生,殷绪成. 基于深度学习的场景文字检测综述[J]. 电子学报,2019,47(5):1152 – 1161.
JIANG Wei, ZHANG Chong-sheng, YIN Xu-cheng. Deep learning based scene text detection: A survey[J]. Acta Electronica Sinica, 2019, 47(5): 1152 – 1161. (in Chinese)

[9] SHARMA K U,THAKUR N V. A review and an approach for object detection in images[J]. International Journal of Computational Vision and Robotics, 2017, 7(1 – 2): 196 – 237.

[10] 毕威,黄伟国,张永萍,等. 基于图像显著轮廓的目标检测[J]. 电子学报,2019,45(8):1902 – 1910.
BI Wei,HANG Wei-guo,ZHANG Yongping,et al. Object detection based on salient contour of image[J]. Acta Electronica Sinica, 2019, 45(8): 1902 – 1910. (in Chinese)

[11] CHAHAL KS,DEY K. A Survey of Modern Object Detection Literature Using Deep Learning[OL]. http://dblp.org/abs/1808.07265,2018.

[12] OKSUZ Kemal,CAM Baris Can,KALKAN S,et al. Imbalance Problems in Object Detection: A Review[OL]. http://dblp.org/abs/1909.00169,2019.

[13] ZHAO Z-Q, ZHENG P, XU S-T, et al. Object detection with deep learning: A review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212 – 3232.

[14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Trans Pattern Anal Mach Intell,2015,39(6):1137 – 49.

[15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2014. 580 – 587.

[16] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报,2017,40(6):1229 – 1251.
ZHOU Fei-yan,JIN Lin-peng,DONG Jun. Review of convolutional neural network[J]. Chinese Journal of Computers,2017,40(6):1229 – 1251. (in Chinese)

[17] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Neural Information Processing Systems,2012, 141(5):1097 – 1105.

[18] UIJLINGD J R,DE Sande K E,GEVERS T,et al. Selective search for object recognition[J]. International Journal of Computer Vision,2013,104(2):154 – 71.

[19] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303 – 38.

[20] HE K,ZHANG X,REN S,et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [A]. Proceedings of the European Conference on Computer Vision[C]. Switzerland:IEEE,2014. 346 – 361.

[21] GIRSHICK R. Fast R-CNN[A]. Proceedings of IEEE International Conference on Computer Vision[C]. USA: IEEE,2015. 1440 – 1448.

[22] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2015. 3431 – 3440.

[23] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[A]. Proceedings of the International Conference on Pattern Recognition[C], USA: IEEE, 2006. 850 – 855.

[24] KONG T,YAO A,CHEN Y,et al. Hypernet:towards accurate region proposal generation and joint object detection[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 845 – 853.

[25] 黄继鹏,史颖欢,高阳. 面向小目标的多尺度 Faster-RCNN 检测算法[J]. 计算机研究与发展,2019,56(2):319 – 327.
HUANG Ji-peng,SHI Ying-huan,GAO Yang. Multi-scale faster R-CNN algorithm for small object detection[J]. Journal of Computer Research and Development,2019,56(2):319 – 327. (in Chinese)

[26] LIN T Y,DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection[A]. Proceedings of

IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 936 – 944.
[27] SINGH B,DAVIS L S. An analysis of scale invariance in object detection-snip[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2018. 3578 – 3587.
[28] DAI J,LI Y,HE K,et al. R-FCN:object detection via region-based fully convolutional networks[A]. Neural Information Processing Systems[C]. Spain:NIPS Foundation,2016. 379 – 387.
[29] ZHU Y,ZHAO C,WANG J,et al. Couplenet:coupling global structure with local parts for object detection[A]. the IEEE International Conference on Computer Vision[C]. Italy:IEEE,2017. 4146 – 4154.
[30] DAI J,QI H,XIONG Y,et al. Deformable convolutional networks[A]. International Conference on Computer Vision,Venice[C]. Italy:IEEE,2017. 764 – 773.
[31] ZHU X,HU H,LIN S,et al. Deformable Convnets-v2:More Deformable,Better Results[OL]. http://dblp.org/abs/1811.11168,2018.
[32] HE K,GKIOXARI G,DOLLAR P,et al. Mask r-cnn[A]. International Conference on Computer Vision[C]. Italy:IEEE,2017. 2980 – 2988.
[33] JIANG B,LUO R,MAO J,et al. Acquisition of localization condence for accurate object detection[A]. European Conference on Computer Vision[C]. Germany:IEEE,2018. 816 – 832.
[34] CAI Z,VASCONCELOS N. Cascade R-CNN:delving into high quality object detection[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2018.
[35] CHEN Z,HUANG S,TAO D. Context refinement for object detection[A]. Proceedings of the European Conference on Computer Vision[C]. Germany:IEEE,2018. 6154 – 6162.
[36] WANG J,CHEN K,YANG S,et al. Region proposal by guided anchoring[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2019. 2965 – 2974.
[37] BODLA N,SINGH B,CHELLAPPA R,et al. Soft-NMS:improving object detection with one line of code[A]. International Conference on Computer Vision[C]. Italy:IEEE,2017. 5562 – 5570.
[38] HE Yi-hui,ZHANG Xiang-yu,MARIOS S,et al. Softer-NMS:Rethinking Bounding Box Regression for Accurate Object Detection[OL]. http://dblp.org/abs/1809.08545,2018.
[39] HU H,GU J,ZHANG Z,et al. Relation networks for object detection[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2018. 3588 – 3597.
[40] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[A]. Neural Information Processing Systems[C]. USA:NIPS Foundation,2017. 5998 – 6008.
[41] REZATOFIGHI S H,TSOI N,GWAK J,et al. Generalized Intersection over Union:A Metric and A Loss for Bounding Box Regression[OL]. http://dblp.org/abs/1902.09630,2019.
[42] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 779 – 778.
[43] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot multibox detector[A]. European Conference on Computer Vision,Amsterdam[C]. Netherlands:IEEE,2016. 21 – 37.
[44] REDMON J,FARHADI A. YOLO9000:Better,faster,stronger[A]. Computer Vision and Pattern Recognition,Hawaii[C]. USA:IEEE,2017. 6517 – 6525.
[45] RUSSAKOVSKY O,DENG J,SU H,et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision,2015,115(3):211 – 52.
[46] LIN T,MAIRE M,BELONGIE S J,et al. Microsoft coco:common objects in context[A]. European Conference on Computer Vision[C]. Switzerland:IEEE,2014. 740 – 755.
[47] REDMON J,FARHADI A. YOLOV3:An incremental improvement[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2018. 3523 – 3541.
[48] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 770 – 778.
[49] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[A]. International Conference on Learning Representations[C]. USA:IEEE,2015. 714 – 723.
[50] JEONG J,PARK H,KWAK N. Enhancement of SSD by concatenating feature maps for object detection[A]. British Machine Vision Conference[C]. UK:BMVA,2017.
[51] FU Cheng-yang,LIU Wei,RANGA A,et al. DSSD:Deconvolutional Single Shot Detector[OL]. http://arxiv.org/abs/1701.06659,2017.
[52] LIN T-Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[A]. International Conference on Computer Vision[C]. USA:IEEE,2017. 2999 – 3007.
[53] LI B,LIU Y,WANG X. Gradient harmonized single-stage detector[A]. the AAAI Conference on Artificial Intelligence[C]. USA:IEEE,2019. 8577 – 8584.
[54] LIU S,HUANG D,WANG Y. Receptive field block net for accurate and fast object detection[A]. European Conference on computer Vision[C]. Germany:IEEE,2018.

404 – 419.
[55] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2016. 2818 – 2826.
[56] ZHANG S, WEN L, BIAN X, et al. Single shot refinement neural network for object detection[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 4203 – 4212.
[57] SUN F, KONG T, HUANG W, et al. Feature pyramid reconfiguration with consistent loss for object detection [J]. IEEE Transactions on Image Processing, 2019, 28 (10): 5041 – 5051.
[58] GHIASI G, LIN T-Y, LE Q V. NAS-FPN: learning scalable feature pyramid architecture for object detection[A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2019. 7036 – 7045.
[59] ZHAO Q, SHENG T, WANG Y, et al. M2det: A single-shot object detector based on multi-level feature pyramid network[A]. The AAAI Conference on Artificial Intelligence[C]. USA: IEEE, 2019. 9259 – 9266.
[60] 裴伟,许晏铭,朱永英,等. 改进的 SSD 航拍目标检测方法[J]. 软件学报,2019,30(3):738 – 758.
FEI Wei, XU Yan-ming, ZHU Yong-ying, et al. The target detection method of aerial photography images with Improved SSD[J]. Journal of Software, 2019, 30(3): 738 – 758. (in Chinese)
[61] YI J, WU P, METAXAS D N. ASSD: Attentive single shot multibox detector[J]. Computer Vision and Image Understanding, 2019, 189(1): 102827 – 102836.
[62] LAW H, DENG J. Cornernet: detecting objects as paired keypoints[A]. European Conference on Computer Vision [C]. Germany: IEEE, 2018. 765 – 781.
[63] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[A]. European Conference on Computer Vision[C]. Netherlands: IEEE, 2016. 483 – 499.
[64] ZHOU Xing-yi, ZHUO Jia-cheng, KRAHENBUHL P. Bottom-up object detection by grouping extreme and center points[A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2019. 850 – 859.
[65] DUAN Kai-wen, BAI Song, XIE Ling-xi, et al. Centernet: Keypoint Triplets for Object Detection [OL]. http://dblp.org/abs/1904.08189, 2019.
[66] TIAN Zhi, SHEN Chun-hua, CHEN Hao, et al. FCOS: Fully Convolutional One-Stage Object Detection [OL]. http://dblp.org/abs/1904.01355, 2019.
[67] SHETTY S. Application of convolutional neural network for image classification on pascal VOC challenge 2012 dataset [A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2016.
[68] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The Open Images Dataset-v4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale[OL]. http://dblp.org/abs/1811.00982, 2018.
[69] GUPTA A, DOLLAR P, GIRSHICK R. LVIS: a dataset for large vocabulary instance segmentation[A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2019. 5356 – 5364.

作者简介

罗会兰　女,1974 年 9 月生于江西上高. 2008 年获浙江大学工学博士学位. 现为江西理工大学图像处理实验室教授、硕士生导师. 主要从事机器学习、模式识别等方面的研究.
E-mail: luohuilan@sina.com

陈鸿坤　男,1995 年 10 月生于江西赣州. 2017 年进入江西理工大学,在读硕士研究生. 研究方向为图像目标检测.
E-mail: 1174426105@qq.com