基于深度学习的图像实例分割技术研究进展

梁新宇[1],林洗坤[2],权冀川[1],肖铠鸿[3]
(1. 陆军工程大学指挥控制工程学院,江苏南京 210007; 2. 华中科技大学软件学院,湖北武汉 430070;
3. 中国人民解放军 73676 部队,江苏无锡 214400)

摘 要: 随着深度学习算法在图像分割领域的成功应用,在图像实例分割方向上涌现出一大批优秀的算法架构. 这些架构在分割效果、运行速度等方面都超越了传统方法. 本文围绕图像实例分割技术的最新研究进展,对现阶段经典网络架构和前沿网络架构进行梳理总结,结合常用数据集和权威评价指标对各个架构的分割效果进行比较和分析. 最后,对目前图像实例分割技术面临的挑战以及可能的发展趋势进行了展望.

关键词: 深度学习; 图像分割; 实例分割
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112 (2020)12-2476-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.12.025

Research on the Progress of Image Instance Segmentation Based on Deep Learning

LIANG Xin-yu[1], LIN Xi-kun[2], QUAN Ji-chuan[1], XIAO Kai-hong[3]
(1. *College of Command & Control Engineering, Army Engineering University of PLA, Nanjing, Jiangsu* 210007, *China*;
2. *College of Software Engineering, Huazhong University of Science & Technology, Wuhan, Hubei* 430070, *China*;
3. *Unit* 73676 *of PLA, Wuxi, Jiangsu* 214400)

Abstract: With the successful application of deep learning algorithms in the field of image segmentation, a large number of excellent algorithm architectures have emerged in the direction of image instance segmentation. These architectures surpass the traditional methods in terms of segmentation effects and running speed. This paper focuses on the latest research progress of image instance segmentation technology, summarizes the current classic network architecture and cutting-edge network architecture, and uses common datasets and authoritative evaluation indicators to compare and analyze the segmentation effects of each architecture. Finally, the challenges and possible development trends of image instance segmentation technology are prospected.

Key words: deep learning; image segmentation; instance segmentation

1 引言

图像分割是一种将数字图像在像素级别上分割为互不交叠区域的图像处理过程. 其分割的原则是,在不同区域间呈现明显差异性,同一区域内呈现相似性. 图像分割的目的是将图像转换为更有意义、更易分析的内容表达,是计算机视觉场景理解的基础. 图像分割任务主要分为两大类:语义分割和实例分割.

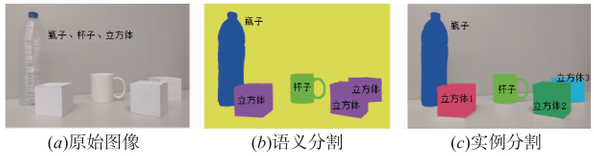
语义分割是对图像中的每个像素划分出对应的类别,实现像素级别的分类. 根据图 1(*a*)需要划分的类别标签,在图 1(*b*)中,将图中的“瓶子”、“杯子”、“立方体”实现像素级别的分类.

现阶段语义分割的架构旨在优化分割结果的精确性和提高分割效率,以便在图像语义实时处理领域进行应用. 然而,在进一步理解图像内容的层面上,语义分割只能判断类别,无法区分个体,在许多复杂的真实场景中达不到准确理解语义信息以及解析场景的能力. 而实例分割算法架构的出现很好地解决了这个问题.

实例分割是在语义分割的基础上,进一步分割已划分类别的具体对象,即分割出实例. 实例分割与语义分割的不同之处在于,不仅要进行像素级别的分类,还需在具体类别基础上区分出不同个体. 这也对分割算

收稿日期:2020-01-15;修回日期:2020-09-03;责任编辑:李勇锋

法提出了更高的要求. 如图 1(*c*)所示,在语义分割的基础之上不仅要分割出“瓶子”和“杯子”两个类,还要用不同的颜色区分同属一类的不同“立方体”实例.



(*a*)原始图像　(*b*)语义分割　(*c*)实例分割

图1　图像分割

2　实例分割的应用场景与技术需求

实例分割的重要性在于越来越多的应用需要利用图像进行推断理解,它能够使计算机获得自动理解场景的能力,在生活、工业等各个场景发挥重要作用:①在医学图像处理领域[1~7],致力于对病人器官的 CT 影像进行处理,精确定位到病灶的边界,从而自动判断病症的位置、形状和大小,辅助医生进行病灶的检测;②在辅助驾驶与自动驾驶领域[8~10],根据实时道路场景,判断道路周边环境,包括车道线走向、来往行人车辆的安全位置、交通标志等,对车辆给出正确行驶指导,保证行人的安全;③在遥感影像处理领域[11,12],对地形地貌、水纹走向、城市分布、农耕规划等地理空间信息进行高效率的勘测与规划指导. 除前文列举的应用场景外,还包括农作物监测[13~15]、文字提取[16,17]等方面. 在这些应用场景中,都需要计算机有类似于人类的感知能力,对图像采集设备获取的场景自主进行语义层面的理解,获得高层语义信息,指导下一步动作.

早期的图像分割大多是基于传统方法的分割技术,包括基于边缘的图像分割技术、基于阈值的图像分割技术、基于区域的图像分割技术、基于特定理论的图像分割技术等[18]. 传统方法的分割技术在实际应用中会产生不同的问题:在基于边缘的方法中,分割结果可能会面临无边缘、噪声严重、边界过度平滑等问题;在基于阈值的分割方法中,阈值设定不得当,会在分割边缘上出现过分割或者欠分割现象,从而导致伪边缘或者丢失边缘;在基于区域的方法中,可能会出现被分割区域大小与实际物体不相符等问题[19].

除此之外,基于传统方法的图像处理技术在分割精度和分割效率上也难以达到实际应用的要求.

在分割精度方面,由于成像采集设备性能的差异、情景环境的复杂,如光线、姿态、遮挡、背景杂波、阴影和模糊等,导致图像品质不良,最终对分割精度造成较大的影响,要求算法必须具备高鲁棒性;在分割效率方面,在当前信息爆炸的时代,指数级增长的图像和视频数量要求高效且可扩展的分割模型,同时汽车、无人机及可穿戴设备等应用场景对模型的内存和存储提出了很高的要求,尤其是在实时场景理解和图像信息处理方面.

随着对图像分割技术的深入研究和相关领域的发展,图像语义分割技术已经取得了长足进步,其网络架构也逐渐趋于成熟;而图像实例分割的技术难度更大,应用领域前瞻性更强,是目前图像分割领域最具价值的研究方向之一. 近年来,将深度学习技术引入图像分割领域,在图像实例分割方面取得了以往传统方法难以达到的分割效果. 本文对基于深度学习的图像实例分割技术的最新进展进行了梳理总结,对 2017 年以来提出的新型网络架构进行了综合分析. 同时,从图像实例分割的性能评价需求出发,分析了分割试验常用的测试数据集和性能评价指标,对主流的实例分割网络架构进行了性能对比,可为相关领域的理论研究和应用实践提供有价值的参考.

3　基于深度学习的图像实例分割技术

深度学习[20]是一个复杂的机器学习算法,也是近十年机器学习领域的研究热点之一. 深度学习利用多层神经网络结构,学习样本数据的内在规律和表示层次,将隐含在高层中的信息进行建模. 深度学习最终目标是让计算机能够像人一样具有分析学习能力,能够识别文字、图像和声音等数据. 近几年的发展,使得深度学习在语音和图像识别方面取得的效果,远远超过先前相关技术.

基于深度学习的图像实例分割技术(简称深度图像实例分割),根据实际需求设计好深层网络算法,不进行人为的特征设计,直接向深层网络输入大量原始图像数据,通过对图像数据的复杂处理,得到高级别的抽象特征,最终输出与输入图像同分辨率的实例分割图像.

3.1　实例分割的经典网络架构

深度图像实例分割的基本思路是在语义分割基础上增加目标检测,先用目标检测算法对图像中的实例进行定位,再用语义分割方法对不同定位框中的目标物体进行分割标记. 深度图像实例分割技术很好改善了传统方法在分割性能方面的不足之处. 本节阐述了深度图像实例分割的经典网络架构,并对这些架构的特性进行总结和分析.

(1)SDS

SDS[21]是首个结合了目标检测和分割的架构,如图 2 所示,架构分为四个阶段:推荐生成(Proposal Generation)阶段、特征提取(Feature Extraction)阶段、区域分类(Region Classification)阶段、区域改良(Region Refinement)阶段. 通过 MCG(Multiscale Combinatorial Grouping)算法[22]为每个图像生成候选区域;在特征提取阶段联合训练两

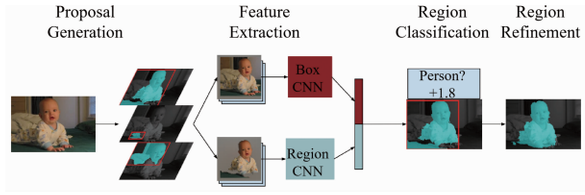


图2 SDS网络架构

个网络,从而产生一个端到端训练的特征提取器用以从候选区域和区域前景中提取并融合边界框特征和区域前景特征;使用传统卷积神经网络(Convolution Neural Network, CNN)[23]所得特征训练支持向量机(Support Vector Machine, SVM)[24]对候选区域进行分类;最后对许多重复覆盖的区域进行非最大抑制(Non-MaximumSuppression, NMS)[25]处理,使用 CNN 产生的特征直接进行掩码预测(mask prediction),并通过与原有的候选区域相结合对结果进行改善和提高.

(2) DeepMask 与 SharpMask

DeepMask[26]是一种发现和切割单张图像中每个物体的算法,它将分割看作是海量的二进制分类问题. 如图 3 所示,给定一个图像块(image patch)作为输入,DeepMask 会输出一个与类别无关的掩码(mask)和一个相关的得分(score),再通过反向传播以及联合学习的方式进行整图推断,估计这个图像块完全包含一个物体的概率. 其最大特点是不依赖于边缘、超像素或者其他任何形式的低级别分割,是首个直接从原始图像数据学习产生分割候选的算法.

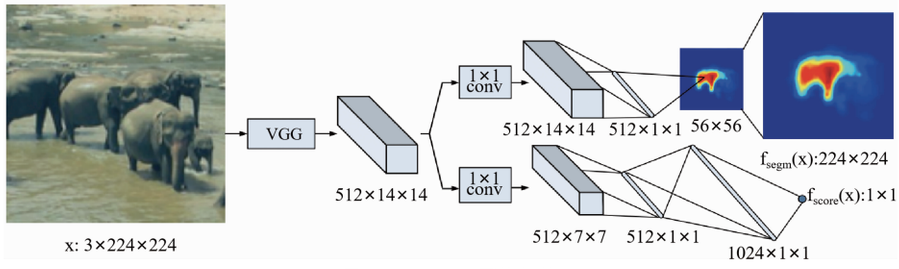


图3 DeepMask网络架构

但是,由于上层功能以相当低的空间分辨率进行计算,不是像素级别的精准分割,为掩码预测带来一个问题:掩码能捕捉一个物体的大致外形,但不能准确捕捉物体的边界.

图 4 所示的 refinement 模块能够有效应对这一问题,由卷积、ReLU、双线性上采样和级联构成. 其中,用 M^i 表示一个掩码编码(mask encoding),用 F^i 表示对应层产生的匹配特征. 通过反向传播和随机梯度下降算法训练网络,融合输入自上而下传递的掩码编码 M^i 和自下而上传递的特征图 F^i 的信息. 由于它们在结构上具有相同的维度,可以生成新的拥有两倍空间分辨率的掩码编码 M^{i+1},以生成掩码的精细输出.

SharpMask[27]在 DeepMask 算法架构基础上进行优化改进,主要思路是,首先用 DeepMask 结构生成粗略的掩码,然后把这个粗略的掩码通过贯穿网络的多个 refinement 模块进行处理,生成最终的精细掩码.

(3) InstanceFCN 与 FCIS

在全卷积网络(Fully Convolutional Networks, FCN)[28]中,输入任意尺寸的图像可以通过一系列的卷积最终为每一个像素产生关于每个类别的概率,从而实现了简单、高效、端对端的语义分割. 但是 FCN 并不能解决实例分割任务,由于卷积具有平移不变性,即同一个像素点接收的响应(类别得分)是相同的,与它在上下文中的相对位置无关,无法区分单个物体实例. 简

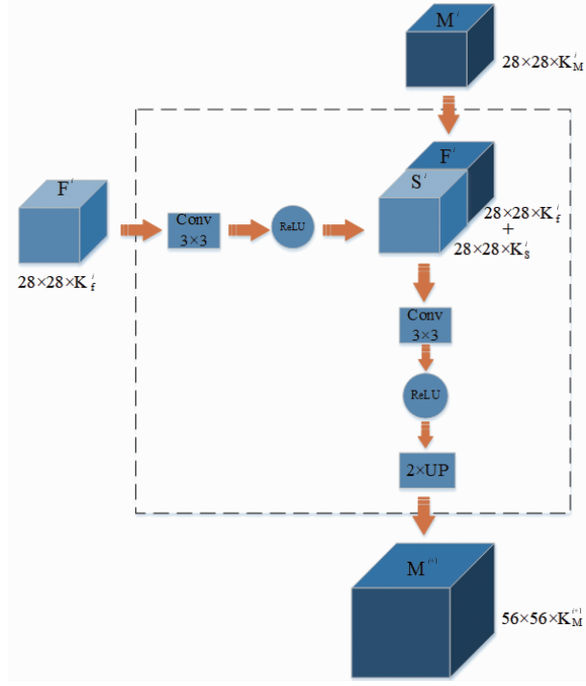


图4 refinement模块架构

单来说,因为卷积的平移不变性,图像中像素点无论在哪个位置,它所对应的类别是固定的,所以语义分割中,每个像素点只能对应一种语义. 而实例分割任务需要在区域级上操作,并且同一个像素在不同的区域中

具有不同的语义,比如在这个区域中,它可能是前景,但在另一个区域中可能就是背景. 鉴于此,本小节介绍的两种经典架构:InstanceFCN[29] 与 FCIS[30],很好解决了这个问题.

InstanceFCN 主要针对图像的局部像素进行改善的一种架构,提出了实例敏感得分图(instance-sensitive score map),每个得分表示一个像素在某个相对位置上属于某个物体实例的得分,很好解决了同一像素在不同区域中有不同响应的问题. 具体结构如图 5 所示.

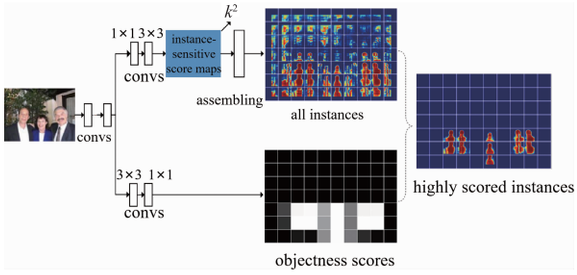


图5　InstanceFCN架构

输入图像通过特定结构的 CNN 进行特征提取,并将提取的特征作为两个全卷积分支的输入. 上支路用于预估部分实例:首先生成多个实例敏感得分图,然后经过聚合模块(assembling module)生成全部实例图(all instance map). 下支路计算了对象得分图(objectness score map),每个像素通过逻辑回归对以该像素为中心的滑动窗口进行实例或非实例分类,生成目标评分图,用来对应全实例图中像素所属实例的得分. 最后将两者融合获得最终的实例分割结果图.

但是,可以发现,InstanceFCN 在生成全实例图后需要一个得分网络进行辅助判别,它并不是一个端到端的网络架构.

FCIS 是建立在 InstanceFCN 基础上的首个全卷积端到端的实例分割模型. 如图 6 所示,架构继续采用实例敏感得分图,同时加入了区分同一像素在目标实例所属位置关系的内部/外部得分图(inside/outside score maps),与区域预测网络(Region Proposal Network, RPN)[31] 产生的兴趣区域(regions of interest, RoI)[32] 共同作用并进行聚合,完成同时进行分割和分类的任务.

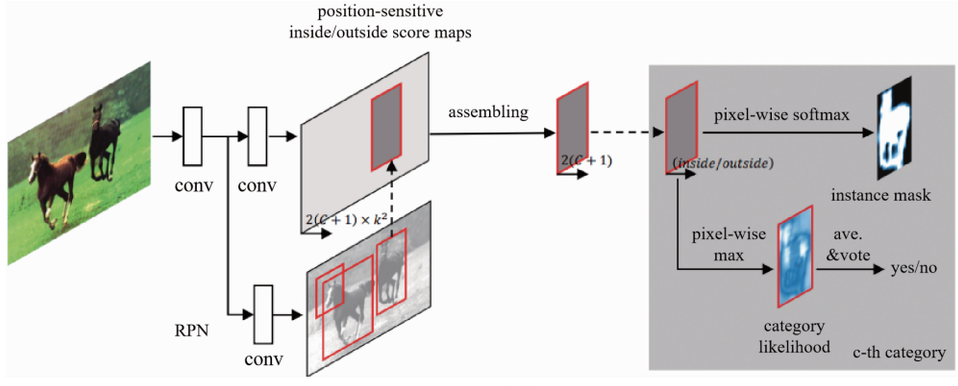


图6 FCIS架构

总的来说,FCIS 一方面很好解决了同一像素在不同区域中有不同响应的问题,并将分割和分类任务同时进行;另一方面,优化了网络结构,实现了实例分割端到端训练,大大提升了架构的分割性能.

3.2　实例分割的前沿网络架构

2017 年以来,在相关技术和研究学者的推动下,经典网络架构有了新突破. 新的设计思想和观点的引入与应用催生出新的网络架构,这些网络架构代表了实例分割的前沿方向.

(1) Mask R-CNN

Mask R-CNN[33] 是由 He 等在继承现有架构技术优点的基础之上提出的实践效果极佳的分割架构. Mask R-CNN 使用与 Faster R-CNN[31] 类似的架构. Faster R-CNN 的输出是物体的边界框(bounding box)和类别,而 Mask R-CNN 增加了一个掩码预测分支(mask prediction branch),对 Faster R-CNN 的每个推荐区域都使用 FCN 进行语义分割,并且改良了兴趣区域池化,提出了 RoI Align 算法以改善和预测物体的语义分割图. 分割任务与定位、分类任务同时进行,即架构同时学习两项任务,可以互相促进.

Mask R-CNN 成就了 ICCV 2017 的最佳论文,彰显

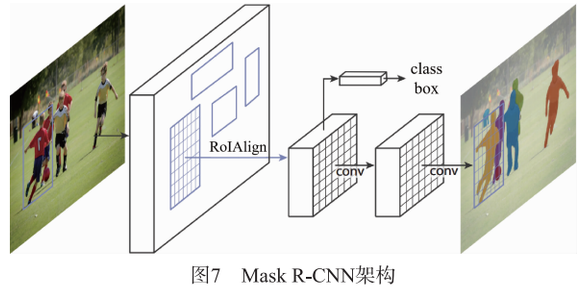


图7　Mask R-CNN架构

了当时机器学习在计算机视觉领域的最新成果.

(2)PANet

在分割任务中,低级别特征有助于大型目标的识别,但低级别特征到高级别特征的路径太长,这增加了定位信息流动的难度. PANet[34]在 Mask R-CNN 的基础上进一步聚合底层和高层特征,旨在提升基于候选区域的实例分割框架内的信息流传播. 具体来说,通过自下向上的路径增强在较低层中准确的定位信息流,建立底层特征和高层特征之间的信息路径,从而增强整个特征层次架构.

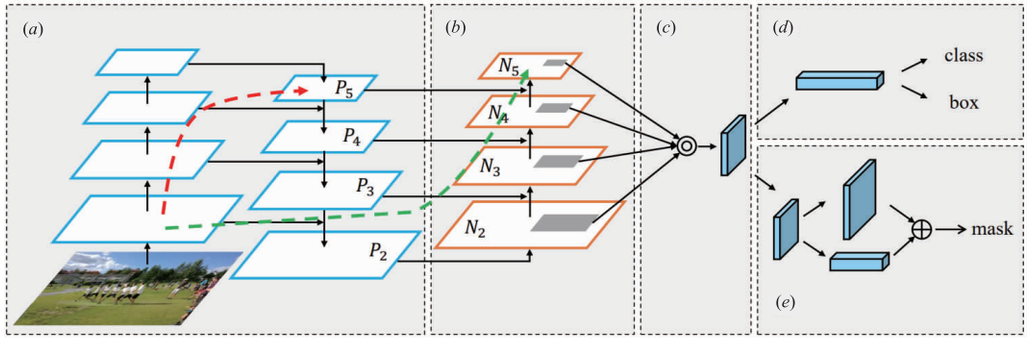


图8 PANet架构

PANet 架构如图 8 所示由五部分组成:(*a*)特征金字塔网络(Feature Pyramid Network, FPN);(*b*)自下而上的路径增强网络;(*c*)自适应特征池化(adaptive feature pooling)结构;(*d*)掩码预测分支结构. (*e*)全连接融合结构.

具体而言,自下而上的增强路径(*b*):用于缩短信息路径,利用底层特征中存储的精确定位信号,提升特征金字塔架构;自适应特征池化(*c*):用于恢复每个候选区域和所有特征层次之间被破坏的信息路径,聚合每个特征层次上的每个候选区域,让每个特征层次的有用信息直接传播到后续的候选区域子网络;使用一个小型全连接层用于补充掩码预测,这能够捕获每个候选区域不同视图,该全连接层与 Mask R-CNN 原始的特征金字塔网络有互补作用. 通过融合两个视图,信息的多样性会增加,进一步提升掩码预测的效果.

(3)TensorMask

在目标检测任务中,常采用滑动窗口方法生成目标的检测框. 而在实例分割任务中,主流方法是首先检测目标对象的边界框,然后对这些部分进行裁剪分割,如 Mask R-CNN.

TensorMask[35]是最新提出的实例分割架构,与主流的实例分割架构不同,它采用结构化的四维张量定义掩码的表示方式,通过张量双尺度金字塔(Tensor Bipyramid)采集不同尺度对象的空间位置和掩码,保持恒定的分辨率密度. TensorMask 提出的基于滑窗方法的密集实例分割框架,为密集掩码预测任务提供了一个新的理解方向,并为该领域提供了新的基础方法.

3.3 综合分析

从早期的 SDS 架构到最新的 TensorMask 架构,实例分割算法在语义分割技术基础上不断优化和改进,不仅能够划分类别,还可以较为精确地区分个体. 我们对上述网络架构进行了综合分析,并从主要思想、优缺点、关键技术和主要功能等几个方面进行了对比总结,如表 1 所示.

从表 1 可以看出,小目标、遮挡物体[36~40]成为实例分割效果进一步提升的瓶颈,也可能成为该领域下一步需要突破的方向.

4 数据集与实验分析

4.1 常用数据集

收集并创建一个规模足够大且具有代表性的应用场景数据集,对于任何基于深度学习的应用领域来说都是极为重要的. 使用一个现有的、权威的、有足够代表性的标准数据集可以使架构之间的性能对比更加公平.

如表 2 所示,本文从数据集的特点、应用场景、类别数目、发布时间以及训练集、验证集、测试集七个方面出发,对目前图像实例分割常用的几种大规模数据集进行了归纳整理.

目前,大型数据集的创建是一个工作量庞大的工程,不仅需要大量的时间、相关的软硬件设施,同时也需要专业领域的知识. 本文列举的数据集大多由专业的赛事主办方、专业团队、公司等所创建,虽然年份较早,但适用性强,是目前深度图像实例分割常用的数据集,受到该领域广大研究者的青睐. 当然,近几年也出现了不少根据特殊研究需求或场景创建的新型数据集:ADE20K[46]是 Zhou 等在 2017 年创建的一个场景理解的新数据集,由各种物体、场景共 151 个类别、20210 张场景图片组成. 2018 年,Zhang 等[47]针对四种天候

(黄昏天候,夜间天候,下雨天候和艳阳天候)下的驾驶场景,采集并构建了 UADS 数据集,整个数据集共计 6380 张图像,提供全天候道路图片以及对应的二值标签.

表 1　实例分割典型网络架构的对比总结

架构	主要思想及其优缺点			关键技术及其功能	
	主要思想	优点	缺点	关键技术	主要功能
DeepMask, SharpMask	分割掩码	挖掘多种尺度、背景图像中的隐含信息,精细化分割结果	对小目标、遮挡物体以及背景复杂物体分割准确率较低	refinement module	精细化分割掩码,降低分割过程中的信息损失,提升分割效果
				整图推断	将模型密集地用于多个位置和多个尺度,捕获图像全局信息,提升分割性能
SDS, Mask R-CNN	候选区域	用物体检测技术产生候选区域,同时完成物体检测与图像语义分割两项任务	没有充分考虑图像中的全局语义信息,图像中的小目标区域和小面积区域分类时易出错	MCG	生成候选区域
				RoI Align	改善区域特征聚集方式,优化区域量化匹配问题,提升检测模型的准确性
				NMS	消除多余(交叉重复)的窗口,找到图像中最佳的物体检测位置
				SVM	建立分割超平面,将不同类别目标完全分割,同时保证分割区间最大化
TensorMask	密集滑动窗口	用四维张量使分割掩码获得更丰富的表征,同时获取更多的图像内容	较大的计算复杂度,分割速度较慢	Tensor Bipyramid	采集不同尺度对象的空间位置和掩码
Instance FCN, FCIS	集成、多任务式	集成、多任务式处理图像的分割与分类问题,实现了实例分割端到端训练,优化网络结构,增强了架构的分割与分类性能	结构较为复杂,小目标分割效果差	instance-sensitive score map	表征相同区域不同语义,克服卷积的平移不变性
				inside/outside score map	图像分割与图像分类共享特征图,更加有效进行前景、背景分类
				RPN	在特征图上产生候选预测区域
PANet	特征融合	通过特征金字塔、路径增强等技术,建立底层特征和高层特征之间的信息路径,捕获图像中不同尺度的特征信息并将其融合从而增强整个特征层次架构,降低了计算量	可能导致图像中小目标对象的边界信息部分丢失,分割效果不理想	FPN	聚合不同层次语义,整合各层次信息
				adaptive feature pooling	池化来自所有层次的特征,恢复并聚合候选区域与特征层间的信息

4.2　实例分割架构的性能评价指标

为了使实例分割架构能够发挥其实际作用,必须对其性能进行严格评估. 同时,为了对现有架构性能进行公平比较,必须使用标准的、公认的、多维度的指标进行评估. 一般情况下,深度图像实例分割架构性能指标主要从执行时间、内存占用和准确性三个维度进行对比.

随着计算机硬件技术的发展,数据处理和存储的能力有了大幅提升. 同时,不同场景下的架构设计技术、架构功能等方面受执行时间和内存占用的影响逐渐变小. 现阶段,在实例分割的相关研究中,对实验结果的分析大都以提升架构的分割准确度为研究重点,用以横向对比不同架构的性能. 在图像实例分割领域,已经提出了许多评估架构准确度的标准. 一般来说选取如平均召回率(Average Recall, AR)[48]、平均精确度(Average Precision, AP)[48]、均值平均精确度(mean Average Precision, mAP)[48]以及交并比(Intersection over Union, IoU)[28]等几项评价指标综合分析. 为方便理解,对上述评价指标作简单介绍:

平均召回率(AR):表示每个类别中正确识别物体的个数占测试集中物体总数的百分数在所有类别上的平均数. 对于同一个架构,测试集物体总数规模越大,AR 表现效果越好. 一般情况下,在物体总数为 10、100、1000 等规模下进行实验.

平均精确度(AP):通常用于计算平均的检测精度,用于衡量检测器在每个类别上的性能优劣,通常情况下,分类器越好,AP 值越高.

均值平均精确度(mAP):计算每个类别的 AP 值后

再取所有类别的平均值. 通常用于评价多目标的检测器性能,衡量检测器在所有类别上的性能优劣. mAP 的取值范围为[0,1]区间,数值越大则表示检测器性能越好.

表 2 深度图像实例分割常用数据集

数据集	特点	文献	应用场景	类别数目	训练样本集	验证样本集	测试样本集	时间
PASCAL VOC 2012	主要是多标签型数据集,图像大小不定,包含常见生活物体	[41]	通用场景	21	1464	1449	1452	2012
NYUDv2	由一系列表示各种室内场景的视频序列组成,分为三个子数据集:RGB 图像、深度图像和 RDB-D 图像	[42]	室内场景	40	795	654	—	2012
PASCAL Part	图像包含像素级分割标注,提供丰富的细节信息,目标物体不同部位标注精确	[43]	人体解析	21	10103	10103	9637	2014
MS COCO	图像包含复杂的日常场景,规模巨大,内容丰富。图像中的物体具有精确的位置标注	[44]	通用场景	81	82783	40504	81434	2014
Cityscapes (coarse)	数据集动态信息丰富、场景布局多样和街道背景复杂,主要提供无人驾驶环境下的图像分割数据,图像涵盖不同环境、不同背景、不同季节的街道场景	[45]	街道场景	30	22973	500	—	2015
Cityscapes(fine)				30	2975	500	1525	2015

交并比(IoU):表示分割结果与原始图像真值的重合程度. 在目标检测中可以理解为系统预测的检测框与原图片中标记的检测框的重合程度. 取值范围为[0,1]区间.

4.3 实验结果对比与分析

由于实例分割同时需要完成目标检测和语义分割任务,所以,在对实验结果进行评判时除了需要考虑分割精度,还需要考虑目标检测的准确度. 下文根据对架构考察的侧重点不同,选择 AR、AP、mAP 作为评价指标,对上述实例分割典型网络架构在相应数据集上进行了测试实验和性能对比. 此处 AR 是指测试集物体总数为 1000 时的实验结果,AP 是在 IoU≥0.5 的筛选条件下所得的实验结果.

表 3 是对实例分割典型网络架构在 PASCAL VOC 2012、MS COCO、Cityscapes 等数据集上的实验结果数据.

表 3 实例分割典型网络架构的实验结果

架构	数据集	评价指标	数值(%)	时间
SDS	PASCAL VOC 2012	mAP	49.7	2014
DeepMask	PASCAL VOC 2012	AR	47	2015
	MS COCO	AR	33.1	
SharpMask	MS COCO	AR	35.5	2016
InstanceFCN	PASCAL VOC 2012	AR	52.6	2016
	MS COCO	AR	39.2	
FCIS	MS COCO	mAP	59.9	2016
Mask R-CNN	MS COCO	AP	60	2017
	Cityscapes	AP	58.1	
PANet	MS COCO	AP	65.1	2018
	Cityscapes	AP	63.1	
TensorMask	MS COCO	AP	59.3	2019

从表 3 可以看到,在数据集方面,实例分割架构主要针对较为复杂的生活场景,以便能够训练成熟架构,达到实时分割、工业化应用的目的. 因此,选用的多为 MS COCO、PASCAL VOC 2012 等规模巨大、内容丰富、情景复杂、种类繁多的数据集;或是如 Cityscapes 这类在自动驾驶领域应用的专用实时分割数据集.

在架构方面,SDS 作为早期的实例分割架构,mAP 虽然不高,但由于其提出时间早、代码开源,是首个通过候选区域思想实现实例分割的架构,为之后的领域研究提供了参考;SharpMask 对 DeepMask 进行改进,通过将 DeepMask 生成的粗略分割掩码输入精细模块,进行逐步优化后最终生成精细的分割掩码,在分割性能上有一定的提升,证明了精细化掩码模块对保留语义特征、减少信息损失的有效性;FCIS 在 InstanceFCN 的基础之上,通过集成、多任务式的思想,改进了早期分割网络的结构,在场景复杂的 MS COCO 数据集上分割效果良好,是当时实例分割架构的典型代表;Mask R-CNN 沿用候选区域的思想,由于其良好的分割表现和

稳定的架构性能,备受领域青睐,在实例分割各类应用场景中广泛存在;PANet 继承了 Mask R-CNN 的优点,采用多尺度特征融合的思想,在两类数据集上均取得了 AP 超过 50% 的表现,分割精度取得突破;TensorMask 架构提出了全新的分割思路,在 MS COCO 数据集上的 AP 为 59.3%,达到甚至超越许多成熟架构的分割效果,通过实验数据表明了"密集窗口滑动"在实例分割领域也是行之有效的途径. 但是总体来看,现有的主流架构在几类复杂数据集上的分类和分割效果还有很大的提升空间.

5　挑战与展望

随着计算机性能的提升和图像分割算法架构的不断优化,基于深度学习的图像实例分割技术在计算机视觉领域将发挥越来越大的作用,同时也面临着诸多挑战.

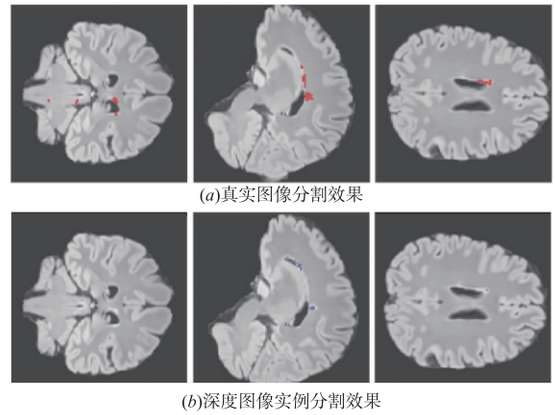
(1)小目标图像实例分割. 小目标分割,在无人机战场环境侦察、医学影像病灶检测等实际应用场景中十分常见. 正确的实例分割有利于指挥员准确判断战场态势,或者辅助医生进行有针对性的病灶诊断.

一般在特征提取阶段,CNN 会有若干池化操作以提取高级别的语义信息. 虽然可以通过上采样等方法恢复空间分辨率,但是对于小目标而言,这样的处理在池化过程中会造成细节信息丢失. 当池化层达到一定数量时,小目标的部分甚至全部信息都将丢失. 因此,针对小目标分割问题,保留低层特征图上的信息是十分必要的.

目前,针对小目标的图像实例分割算法架构分割准确率低、效果差,还不能完全满足实际应用的要求,存在明显的漏分割、错分割、分割边界模糊等问题. 如图 9 所示,在进行脑部病灶检测过程中出现漏分割. 如何对小目标图像进行高效、精准分割是目前深度图像实例分割领域的重要研究方向.

(2)低质量图像实例分割. 由于图像采集设备和采集环境的特殊性,某些特定领域内获取图像的质量存在许多问题:如灰度值分布大,往往伴随着模糊性;图像中存在一定扰动使得图像内各实例之间出现重叠效应. 除此之外,军事战场环境下获得的图像数据,往往伴随着雨天[49~51]、雾天[52~56]、阴天、烟雾[57]等复杂环境因素,使得图片质量大幅下降. 图 10 列举了部分低质量图像的具体情况. 这些图像质量问题严重影响分割效果,最终影响到相关领域专业人员的判断,导致决策失误. 因此,必须改善原始图像的质量以提升后续实例分割的效果.

(3)轻量化网络架构的需求[58]. 现有的网络架构研究和相关的竞赛更加偏向于追求准确率的提升. 但



(a)真实图像分割效果

(b)深度图像实例分割效果

图9　脑部小目标病灶分割图像

(a)雨天图像　(b)模糊图像

(c)噪点图像　(d)有雾图像

图10　低质量图像

准确性表现良好的网络架构往往结构复杂,运行时间长,忽略了实际应用中对算法架构的轻量、快速响应的需求. 随着移动端、嵌入式设备的普及应用,满足简化架构、压缩和复用计算同时又能保证准确率要求的轻量化网络架构,将是今后深度图像实例分割技术的重要发展方向.

(4)新型网络架构的实践应用. 诸如 TensorMask 和 Auto-DeepLab[59]这类最新的网络架构,为深度图像实例分割提供了新的思路,在实验中也取得了良好的分割效果. 如何将这类新的架构思想与实际工程应用相结合也是非常值得研究的方向.

6　结束语

本文从图像分割概念引入,讨论了语义分割与实例分割的联系与区别,并重点梳理总结了基于深度学习的图像实例分割网络架构的最新研究进展,对其在常用数据集下的实验性能进行了对比分析. 面对应用场景日趋丰富、要求日益严苛的实际应用需求,基于深度学习的图像实例分割技术将面临更多的挑战. 本文最后对该方向的技术难点和发展趋势进行了展望. 总

的来说,基于深度学习的图像实例分割技术虽然取得了一定突破,但该方向的研究成果与现实需求还有较大差距,应该更多地注重技术落地和场景推广,对相关技术的研究和探索仍然任重道远.

参考文献

[1] 郑冰. 面向肺部 CT 影像表征的多层语义检索[D]. 哈尔滨:哈尔滨工程大学,2013.

[2] 南洋. 基于深度学习的粗标记胃癌病理切片图像分割算法[D]. 长沙:湖南大学,2018.

[3] 杨少戈. 基于深度学习的冠脉造影图像分割[D]. 北京:北京邮电大学,2019.

[4] 杨远航. 面向深度学习的医学影像分析系统及其在胃镜视频分割中的实践[D]. 杭州:浙江大学,2018.

[5] 赵旭. 基于医学先验的多尺度乳腺超声肿瘤实例分割方法[D]. 哈尔滨:哈尔滨工业大学,2019.

[6] 张晓林. 基于卷积神经网络的腹部 CT 图像分割[D]. 镇江:江苏大学,2019.

[7] 宫进昌,赵尚义,王远军. 基于深度学习的医学图像分割研究进展[J]. 中国医学物理学杂志,2019,36(04):420-424.
GONG Jin-chang, ZHAO Shang-yi, WANG Yuan-jun. Research progress of medical image segmentation based on deep learning[J]. Chinese Journal of Medical Physics, 2019,36(04):420-424. (in Chinese)

[8] 白辰甲. 基于计算机视觉和深度学习的自动驾驶方法研究[D]. 哈尔滨:哈尔滨工业大学,2017.

[9] 邓琉元,杨明,王春香,等. 基于环视相机的无人驾驶汽车实例分割方法[J]. 华中科技大学学报(自然科学版),2018,46(12):24-29.
DENG Liu-yuan, YANG Ming, WANG Chun-xiang, et al. Method for segmentation of unmanned car instance based on surround view camera[J]. Journal of Huazhong University of Science & Technology (Natural Science Edition), 2018,46(12):24-29. (in Chinese)

[10] 姜立标,台啟龙. 基于实例分割方法的复杂场景下车道线检测[J]. 机械设计与制造工程,2019,48(05):113-118.
JIANG Li-biao, TAI Qi-long. Lane line detection in complex scenes based on instance segmentation method[J]. Mechanical Design and Manufacturing Engineering,2019, 48(05):113-118. (in Chinese)

[11] 惠健,秦其明,许伟,等. 基于多任务学习的高分辨率遥感影像建筑实例分割[J]. 北京大学学报(自然科学版),2019,55(06):1067-1077.
HUI Jian, QIN Qi-ming, XU Wei, et al. Segmentation of high-resolution remote sensing image building instance based on multi-task learning[J]. Journal of Peking University (Natural Science Edition),2019,55(06):1067-1077. (in Chinese)

[12] 李澜. 基于 Mask R-CNN 的高分辨率光学遥感影像的目标检测与实例分割[D]. 武汉:武汉大学,2018.

[13] 帅靖文. 自然场景中的文本检测研究[D]. 成都:电子科技大学,2018.

[14] 张小爽. 基于实例分割的场景图像文字检测[D]. 杭州:浙江大学,2018.

[15] 邓丹. PixelLink:基于实例分割的自然场景文本检测算法[D]. 杭州:浙江大学,2018.

[16] 谢元澄,于增源,姜海燕,等. 小麦麦穗几何表型测量的精准分割方法研究[J]. 南京农业大学学报,2019,42(05):956-966.
XIE Yuan-cheng, YU Zeng-yuan, JIANG Hai-yan, et al. Research on accurate segmentation method of wheat ear geometric phenotype measurement[J]. Journal of Nanjing Agricultural University, 2019, 42(05): 956-966. (in Chinese)

[17] 乔虹,冯全,赵兵,等. 基于 Mask R-CNN 的葡萄叶片实例分割[J]. 林业机械与木工设备,2019,47(10):15-22.
QIAO Hong, FENG Quan, ZHAO Bing, et al. Grape leaf instance segmentation based on Mask R-CNN[J]. Forestry Machinery and Woodworking Equipment, 2019, 47(10):15-22. (in Chinese)

[18] 冈萨雷斯. 数字图像处理[M]. 北京:电子工业出版社,2011.
Gonzal R. Digital Image Processing[M]. Beijing: Publishing House of Electronics Industry,2011. (in Chinese)

[19] 苏雯. 语义分割及其在图像检索中的应用[D]. 合肥:中国科学技术大学,2018.

[20] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786):504-507.

[21] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation[A]. European Conference on Computer Vision[C]. Cham: Springer, 2014. 297-312.

[22] Arbeláez P, Pont-Tusent J, Barron J T, et al. Multiscale combinatorial grouping[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2014. 328-335.

[23] 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.

[24] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. USA: John Wiley & Sons, 2012.

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

[26] Pinheiro P O, Collobert R, Dollár P. Learning to segment object candidates[A]. Advances in Neural Information Processing Systems[C]. Canada: NIPS, 2015. 1990 – 1998.

[27] Pinheiro P O, Lin T Y, Collobert R, et al. Learning to refine object segments[A]. European Conference on Computer Vision[C]. Cham: Springer, 2016. 75 – 91.

[28] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Boston: IEEE, 2015. 3431 – 3440.

[29] Dai J, He K, Li Y, et al. Instance-sensitive fully convolutional networks[A]. European Conference on Computer Vision[C]. Cham: Springer, 2016. 534 – 549.

[30] Li Y, Qi H, Dai J, et al. Fully convolutional instance-aware semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu: IEEE, 2017. 2359 – 2367.

[31] Ren S, He K, Girshick R, et al. Faster R-CNN: Towardsreal-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137 – 1149.

[32] Girshick R. Fast r-cnn[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Santiago: IEEE, 2015. 1440 – 1448.

[33] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice Italy: IEEE, 2017. 2961 – 2969.

[34] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City: IEEE, 2018. 8759 – 8768.

[35] Chen X, Girshick R, He K, et al. Tensormask: A foundation for dense object segmentation[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Seoul: IEEE, 2019. 2061 – 2069.

[36] Lazarow J, Lee K, Tu Z. Learning instance occlusion for panoptic segmentation[A]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Seattle, WA: IEEE, 2020. 10720 – 10729.

[37] 张国光. 基于神经网络的有遮挡图像分割方法[J]. 电子科技, 2015, 28(05): 132 – 135, 139.
ZHANG Guo-Guang. An occluded image segmentation method based on neural network[J]. Electronic Technology, 2015, 28(05): 132 – 135, 139. (in Chinese)

[38] 师晓利, 尚怡君, 褚玉晓. 安防监控中人员遮挡区域的有效图像分割研究[J]. 计算机仿真, 2015, 32(06): 452 – 455.
SHI Xiao-li, SHANG Yi-jun, CHU Yu-xiao. Research on effective image segmentation of people's occlusion area in security surveillance[J]. Computer Simulation, 2015, 32(06): 452 – 455. (in Chinese)

[39] 刘伟. 基于 Kinect 遮挡条件下行人的深度图像分割[J]. 重庆邮电大学学报(自然科学版), 2014, 26(02): 271 – 275.
LIU Wei. Depth image segmentation of pedestrians based on Kinect occlusion conditions[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2014, 26(02): 271 – 275. (in Chinese)

[40] 李晶晶. 局部遮挡物体的轮廓修复算法研究[D]. 南昌: 南昌航空大学, 2014.

[41] Everingham M, Eslami S A, Van G L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal on Computer Vision, 2014, 11(1): 98 – 136.

[42] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[A]. European Conference on Computer Vision[C]. Berlin, Heidelberg: Springer, 2012. 746 – 760.

[43] Chen X, Mottaghi R, Liu X, et al. Detect what you can: Detecting and representing objects using holistic models and body parts[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, OH: IEEE, 2014. 1971 – 1978.

[44] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[A]. European Conference on Computer Vision[C]. Cham: Springer, 2014. 740 – 755.

[45] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Las Vegas, NV: IEEE, 2016. 3213 – 3223.

[46] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu: IEEE, 2017. 633 – 641.

[47] Zhang Y, Chen H, He Y, et al. Road segmentation for all-day outdoor robot navigation[J]. Neurocomputing, 2018, 314: 316 – 325.

[48] Turpin A, Sholer F. User performance versus precision measures for simple search tasks[A]. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Seattle Washington: ACM, 2006. 11 – 18.

[49] 诸葛瑞彬. 基于卷积神经网络的单幅图像去雨研究[D]. 桂林: 广西师范大学, 2019.

[50] Fu X, Liang B, Huang Y, et al. Lightweight pyramid networks for image deraining[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(6): 1794 – 1807.

[51] 安鹤男,涂志伟,张昌林,等. 单张图像去雨的多流细节加强网络[J]. 计算机系统应用,2019,28(11):202-207.
AN He-nan,TU Zhi-wei,ZHANG Chang-lin,et al. Multi-stream detail enhancement network to remove rain from a single image[J]. Computer System Applications,2019,28(11):202-207. (in Chinese)

[52] 寇大磊,钱敏,权冀川,等. 基于多尺度卷积网络的快速图像去雾算法[J]. 计算机工程与应用,2020,56(20):191-198.
KOU Da-lei,QIAN Min,QUAN Ji-chuan,et al. Fast image defogging algorithm based on multi-scale convolutional network[J]. Computer Engineering and Applications,2020,56(20):191-198. (in Chinese)

[53] Gao Y,Li Q,Li J. Single image dehazing via a dual-fusion method[J]. Image and Vision Computing,2020,94:103868.

[54] Wu Y,Qin Y,Wang Z,et al. Densely pyramidal residual network for UAV-based railway images dehazing[J]. Neurocomputing,2020,371:124-136.

[55] 赵阳,王剑,曹浩男. 基于自适应改进的遥感图像去雾算法研究[J]. 电子设计工程,2019,27(19):164-169.
ZHAO Yang,WANG Jian,CAO Hao-nan. Research on remote sensing image defogging algorithm based on adaptive improvement[J]. Electronic Design Engineering,2019,27(19):164-169. (in Chinese)

[56] 丁春玲. 基于CNN网络的图像去雾霾技术研究[J]. 西安文理学院学报(自然科学版),2019,22(05):57-60.
DING Chun-ling. Research on image removal technology based on CNN network[J]. Journal of Xi'an University of Arts and Science (Natural Science Edition),2019,22(05):57-60. (in Chinese)

[57] 许骏. 面向火灾场景的图像去烟雾系统研究[D]. 上海:东华大学,2016.

[58] 寇大磊,权冀川,张仲伟. 基于深度学习的目标检测框架进展研究[J]. 计算机工程与应用,2019,55(11):25-34.
KOU Da-lei,QUAN Ji-chuan,ZHANG Zhong-wei. Research on the progress of target detection framework based on deep learning[J]. Computer Engineering and Applications,2019,55(11):25-34. (in Chinese)

[59] Liu C,Chen L C,Schroff F,et al. Auto-deeplab:Hierarchical neural architecture search for semantic image segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Long Beach:IEEE,2019. 82-92.

作者简介

梁新宇 男,1993年出生,福建顺昌人,陆军工程大学指挥控制工程学院计算机技术专业硕士研究生,主要研究方向为深度学习、图像分割技术及其应用.
E-mail:674330397@qq.com

林洗坤 男,1999年出生,陕西西安人,华中科技大学软件学院软件工程专业硕士研究生,主要研究方向为人工智能和信息安全.
E-mail:lin_haokun@hust.edu.cn

权冀川(通信作者) 男,1974年出生,河北辛集人,陆军工程大学指挥控制工程学院教授、硕士生导师,研究方向为系统效能评估、多源信息融合.
E-mail:qjch_cn@sina.com

肖铠鸿 男,1996年出生,重庆人,研究方向为深度学习、微波通信.
E-mail:1782800328@qq.com