基于深度学习的图像语义分割技术研究综述

邝辉宇,吴俊君
佛山科学技术学院 机电工程学院,广东 佛山 528225

摘 要:图像语义分割技术是智能系统理解自然场景的关键技术之一,作为视觉智能领域的重要研究方向,该技术在移动机器人、无人机、智能驾驶以及智慧安防等领域具有广阔的应用前景。对于图像语义分割技术的研究发展历程进行了详细评述,包括从传统的语义分割方法到当前主流的基于深度学习的图像语义分割理论及其方法,重点阐述了基于深度学习的图像语义分割技术的框架及其实现过程,进而对当前具有代表性的典型算法的效果以及优缺点进行了分析,然后归纳了算法评价指标,最后对该技术的发展进行了总结与展望。该研究对于从事图像语义分割技术的研究人员和工程技术人员均具有很好的参考意义。

关键词:智能系统;图像语义分割;深度学习;视觉智能
文献标志码:A **中图分类号:**TP391.41 doi:10.3778/j.issn.1002-8331.1905-0325

邝辉宇,吴俊君.基于深度学习的图像语义分割技术研究综述.计算机工程与应用,2019,55(19):12-21.
KUANG Huiyu, WU Junjun. Survey of image semantic segmentation based on deep learning. Computer Engineering and Applications, 2019, 55(19):12-21.

Survey of Image Semantic Segmentation Based on Deep Learning

KUANG Huiyu, WU Junjun
School of Mechatronics Engineering, Foshan University, Foshan, Guangdong 528225, China

Abstract: Image semantic segmentation technology is one of the key technologies for intelligent systems to understand natural scenes. As an important research direction in the field of visual intelligence, this technology has broad application prospects in the fields of mobile robots, drones, intelligent driving and smart security. This paper gives a detailed review on the research and development of image semantic segmentation technology, including the traditional semantic segmentation method and the current mainstream image semantic segmentation theory based on deep learning, and the method of image semantic segmentation based on deep learning. It describes the framework and its implementation process, analyzes the effects, advantages and disadvantages of the typical representative algorithms, and then summarizes the algorithm evaluation indicators. Finally, the development of the technology is summarized and forecasted. The paper has a good reference for researchers and engineers who are engaged in image semantic segmentation technology.

Key words: intelligent system; image semantic segmentation; deep learning; visual intelligence

1 引言

图像语义分割(Image Semantic Segmentation)是一种可以使计算机能够对图像实现自动分割并识别出图像内容的技术。在计算机视觉领域中图像语义分割技术指的是对图片中不同类型的对象以不同颜色标注分割,而图像语义分割技术中的语义指的是图像中所包含的各类别的物体所特有的类别名称,这种类别名称就称作图片的语义信息。应用语义分割技术就是利用计

基金项目:国家重点研发计划(No.2018YFB1308000);国家自然科学基金(No.61603103,No.61673125);广东省自然科学基金(No.2016A030310293);广州市科技计划科学研究专项(No.201707010013);佛山科学技术学院高层次人才科研启动项目(No.Gg07176)。

作者简介:邝辉宇(1996—),男,硕士研究生,研究领域为计算机视觉、视觉智能;吴俊君(1981—),通讯作者,男,博士,副教授,硕士生导师,CCF会员,研究领域为移动机器人和视觉智能,E-mail:jjunwu@fosu.edu.cn。

收稿日期:2019-05-22 **修回日期:**2019-07-26 **文章编号:**1002-8331(2019)19-0012-10
CNKI**网络出版:**2019-08-12, http://kns.cnki.net/kcms/detail/11.2127.TP.20190812.0949.002.html

算机对于一幅图像的像素按照图像中表达的语义信息的不同进行分类。

图像语义分割技术在现阶段主要是基于深度学习神经网络进行研究,而在深度学习技术的支持下,图像语义分割技术主要被应用在移动机器人、无人机、智能驾驶以及智慧安防的领域中。

现有很多综述文章都对目前的基于深度学习的语义分割技术进行了总结,比如文献[1-2]中,对现有的基于深度学习神经网络的方法做出了综述,文献[3]介绍了网络模型的概念。这一类型的工作为后来的研究者提供了非常好的研究思路。但是这一类型的文章主要集中于基于深度学习的图像语义分割技术所使用到的数据集部分、网络模型部分又或者是对深度学习网络部分单独进行综述性讲解,缺乏对于图像语义分割技术的历史做出整体综合性的评述。

本文撰写的目的是填补关于图像语义分割技术的综合性文章的空白。主要内容是回顾图像语义分割技术的研究和开发过程,从传统的语义分割方法到当前主流的方法,重点阐述了基于深度学习的图像语义分割技术理论及其方法,进而对当前具有代表性的典型算法的效果以及优缺点进行了分析,然后归纳了算法评价指标,最后对该技术的发展进行了总结与展望,同时对于以上内容进行多角度多方面的评测总结。笔者认为这一份工作对于有志了解基于深度学习的图像语义分割技术的研究者非常有必要且有着很好的参考意义。

2 本文贡献和结构安排

本文的主要核心贡献在于:(1)对图像语义分割技术的发展历史做出总结归纳、对传统的语义分割方法到目前的基于深度学习技术的图像语义分割方法和理论进行了综述。(2)对于图像分割领域中有着重要影响地位的语义分割网络模型的分割精度和优缺点做出了对比和分析归纳。(3)对于现有的语义分割技术的算法评价进行了综合归纳。(4)总结了许多基于深度学习方法的语义分割技术研究方法,并对语义分割技术的前景进行了展望。

本文第3章介绍从传统的语义分割方法到目前语义分割技术与目标检测结合的理论与方法,同时引入了语义分割技术中的常用名词解释,重点介绍了基于深度学习的图像语义分割的理论与方法;第4章对基于深度学习的图像语义分割的处理方法进行了介绍;第5章将目前最常用的网络模型框架特点进行了总结;第6章则侧重于语义分割算法指标评估和对于现有的数据集的介绍,以及这些数据集适用范围的界定;第7章总结全文,并对语义分割领域下一步的研究提出了几种优势的思路和方法。

3 图像语义分割发展史

图像语义分割技术大概可分为3个时期。

(1)基于传统方法的图像语义分割技术时期

受限于计算机的硬件设备限制,图像分割技术仅能对灰度图像进行处理,后期才逐渐发展到可以对RGB图像进行处理的阶段。在这一时期主要是通过图像的低级特征进行分割,经此技术处理之后所输出的图像无法达到实现语义标注的效果。简而言之,这时期的图像分割技术只能被称为图像分割,无法达到语义的概念。

(2)深度学习与传统方法结合的图像语义分割技术时期

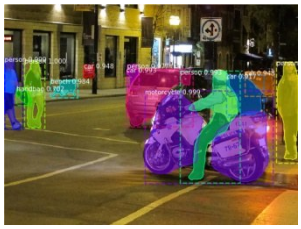
在这一个阶段主要是利用卷积神经网络算法(Convolutional Neural Networks,CNN)[4]实现语义分割效果,先利用传统的分割算法获得图像级的处理效果,然后利用CNN神经网络模型训练像素的特征分类器实现语义分割效果,这种方法准确性受到传统语义分割方法诸多不足的限制,因此准确性普遍较低。

(3)基于全卷积神经网络的图像语义分割技术时期

2015年IEEE国际计算机视觉与模式识别会议(IEEE Conference on Computer Vision and Pattern Recognition),由Long[5]等人提出了全卷积神经网络(Fully Convolutional Networks for semantic segmentation,FCN),至此图像语义分割技术进入到了全卷积神经网络时期。全卷积神经网络在深度学习中表现出了强大的潜力,计算机在图片通过深度学习网络进行深度学习后能够清楚地归纳出输入图片中的具有相同语义含义的像素点。深度学习方法成为了现今解决语义分割问题的主流。对比前两个时期,基于全卷积神经网络深度学习的语义分割技术能够获得更高的精度以及更好的运算效率,因此这一时期的语义分割技术方法介绍将会是整篇文章的讨论重点。

2018年,Michaelis[6]等人根据He[7]等人的研究工作对于语义分割提出了最新的成果,将语义分割技术与目标检测技术进行结合,对于图片中的目标达到了实例分割的效果,这意味着可以对于同一类别的不同物体进行不同的语义信息标注的效果,从最新的CVPR会议上华中科技大学推出的实例分割模型提高了检测的分数的效果来看[8],实例分割所得到的效果也非常不错,值得关注。从图1中的图(a)和(b)的对比中,能够非常明显地看出两者之间的区别。

如图1(a)和(b)是分别对同一张图片单纯利用语义分割和实例分割进行的对比,鲜明地揭示出未来的研究方向语义分割会结合目标检测往实例分割发展。


(a)Instance Segmentation


(b)FCN

图1　同一图片单纯用语义分割和实例分割的对比

3.1　基于传统方法的语义分割技术时期

在计算机硬件设备还不足以支持深度学习神经网络时,研究者想到了各种不同的方法用以实现图像语义分割技术。例如:利用像素级别的阈值分割法[9]、基于像素的聚类分割法[10],到“图割”的图像分割方法[11]、基于像素级的决策树分类法[12]等,而在其中最常用的是利用图割法实现对图像的语义分割技术。

基于图割法的图像语义分割技术,最常用的就是Normalized cut[13]和Grab cut[14]方法,N-cut提出了一种考虑全局信息的方法来进行图割(graph partitioning)用以改变经典的min-cut[15]算法操作中的不足,创新点在于将两个分割部分与全图节点的连接权重也考虑进算法之中,根据图像中的像素给出的阈值将图像一分为二。缺点在于这种分割方式比较简单直接,只能利用图像的像素进行分割,对于整体物体的影响考虑不周。为了改进这一缺点,Grab cut[16]的创新在于预先将图片中需要进行分割处理的部分进行人工标定,在计算机处理的时候也需要人工进行干预,对图像进行标注,指导辅助计算机进行判断分割。

3.2　深度学习与传统方法结合的图像语义分割技术时期

在2012年的目标识别挑战大赛(ImageNet Large Scale Visual Recognition Challenge ILSVRC)中,Krizhevsky等人获得了引人注目的成果,突出贡献在于Krizhevsky提出了基于AlexNet[17]的图像语义分割方法,为深度技术在语义分割领域打下了基础。

之后,涌现出很多基于卷积神经网络实现图像语义分割技术的网络结构模型,例如利用卷积神经网络的深且小的卷积核来对于图片进行深层处理的VGG(Very Deep Convolutional Networks)[18]神经网络结构,还有与VGG有着相似神经网络结构的卷积深入法(GoogleNet)[19],创新点在于使用了数量更多更深的层次来得到更好的结构,对比VGG,图像在经过GoogleNet网络结构的处理之后语义分割精准度得到了提高,但是GoogleNet的缺点在于计算次数太多,效率不足。

Szegedy等人在GoogleNet的基础上提出了深度残差网络(Deep Residual Network,ResNet)[20],其优点在于在VGG网络结构模型的基础上进行了修改,将源自于CNN网络结构的池化层替换为全连接层,从而解决了深度神经网络所存在的深度退化问题,使得可以实现训练更加深层次的神经网络。根据CNN网络结构所改进的网络结构对比总结,如表1所示。

表1　根据CNN改进的几个主要的网络结构

网络结构名称	网络结构的创新点	ImageNet中精度对比/%
AlexNet	在CNN中应用了ReLU、Dropout和LRN	84.6
ResNet	引入残差模块使得网络的上下两层可以直连	96.4
VGGNet	在CNN网络结构中反复利用3×3的卷积核	92.7
GoogleNet	提出了Inception模块,主要是利用了1×1卷积核	93.3

CNN神经网络模型开创了在图像语义分割领域中应用深度学习的基础,使得图像语义分割技术在多个应用领域中使用。例如在生物医学中Jiang[21]等人对VGG-16架构进行修改,去掉了最后一个卷积层,使用ReNet结构作为最大池化层将输出图的尺寸恢复至原输入图的大小,通过使用小型数据集就实现了医学图像分割的任务,在实际应用中可以用于提取心脏和肺部的边界。

3.3　基于全卷积神经网络的图像语义分割技术时期

在卷积神经网络结构中全连接层会将原来图片的二维矩阵信息压缩,导致图像的空间信息丢失,而图像语义分割技术需要输出的是一幅尺寸大小与原图相同的二维图像。全卷积神经网络的优点在于丢弃了卷积神经网络中的全连接层,并将其换成了全卷积层。FCN将卷积神经网络对于图像的识别精度从图像级的识别提升为全卷积神经网络中的像素级的识别。根据这一个特性,研究人员将上文提到的网络模型中的全连接层替换为了全卷积层,替换后的这些神经网络模型在应用于语义分割技术时所得到的效果也颇为惊人[22]。

FCN网络结构为图像语义分割技术提供了能够达到像素级语义分割的基础,更加为后来的研究人员提供了一种全新的思路和探索的方向,使得语义分割的精度得到极大的提高。研究人员以全卷积神经为基础提出了U-Net[23]、SegNet[24]、DeepLab[25]、BiSeNet[26]、ERFNet[27]、RefineNet[28]网络结构等模型。

4　基于深度学习的语义分割处理方法

4.1　卷积层

全卷积神经网络结构可分为卷积层[29]、池化层和上采样(反卷积层)三部分,其工作原理是每当有图像输入

卷积层后,卷积层中的卷积核会对图像进行卷积处理,完成卷积处理后将会输出一幅特征图,输出的特征图片将会经过池化层[30],池化层会对输入的特征图进行压缩处理,提取主要特征,在提取出了主要特征后再经过上采样操作可以将图像的尺寸还原为输入的图片尺寸的大小,从而使深度学习网络结构模型能够达到像素级别的语义分割效果。

4.1.1 空洞卷积

在图像语义分割领域中FCN有着两个关键的突破点,池化层增大感受野减少图像尺寸,上采样扩大图像尺寸。在通过池化层的减少再增大的这一过程中必然会丢失一些信息,如何不通过池化层也能够达到增大感受野的效果,利用空洞卷积层(Dilated Convolutions)[31]是一个很好的解决方案。如图2所示,以3×3卷积核为例,对于传统3×3 patch隔开一定像素进行卷积核运算,例如在池化层中利用一个dilation等于2的Dilated Convolution进行处理,对特征图做采样扩大感受野,其得到的效果就如图2中的第二幅图像所示,空洞卷积的原理就是将卷积核叠加dilate为1的变化后进行扩大,然后利用生成的Patch进行卷积核运算,就能够在保证感受野不变化的前提下,同时又能够大幅度提高对于图像语义分割的辨识度以及运算速度,其中感受野尺寸的扩充公式为:

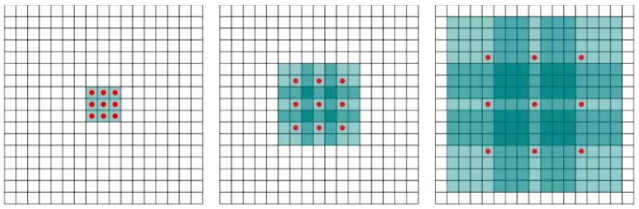


图2 空洞卷积的示意图

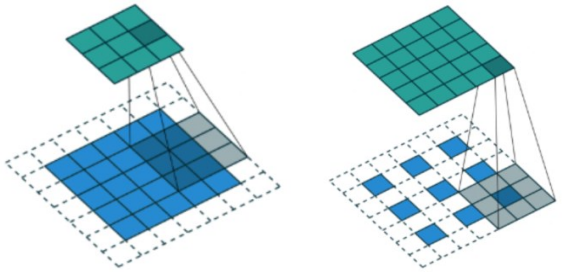
$$3+(dilate-1)\times 4=dilate\ convolution$$

空洞卷积也存在着两个问题,Wang等人[32]提出使用空洞卷积多次叠加后,内核(kernel)不连续会导致图片中的像素不能全部用于计算中,这意味着所得到的信息是不连续的,该缺陷会导致在达到像素级预测分类的深度学习任务中出现显著的误差。第二个问题是Chen[33]等人提出,当在图片中分割相对大的像素图像应用空洞卷积运算确实在速度和精度上有很大的改进,但是对于图像中比较小的像素图像,空洞卷积就会成为降低分割精度的原因。

4.1.2 转置卷积

在卷积神经网络的结构中会使用到名为反卷积层(Deconvolution)的操作过程,在数学意义上反卷积层应该是作为卷积操作的逆过程而存在的,但Noh[34]等人在对卷积操作的逆过程进行研究后发现,在深度学习的领域中应用数学意义上的反卷积操作并不会太过常见。

在编码器-解码器网络结构的解码器网络结构中使用到转置卷积[35]进行上采样,可以将卷积神经网络中的特征值(Feature Map)值还原到输出的像素空间之中,将转置卷积应用于卷积神经网络中就可以观察到feature map对网络中的哪些pattern的响应度是最大的,因此对于卷积神经网络,使用转置卷积可以实现可视化的效果,并知道通过卷积运算提取了什么特征。而应用于解码器网络结构中,转置卷积可以将经过解码器提取特征后的图片还原回输入图片的尺寸大小,如图3所示。



(a)卷积 (b)转置卷积

图3 卷积核为3×3:no padding,步长为2时卷积操作与转置卷积对比图

在深度神经网络中池化是不可逆的,研究人员们常常可以通过记录池化过程中的最大激活值得到坐标位置,这可以通过转置卷积来激活同时将其他位置值会置0,但实际神经网络结构在池化过程中除了最大的值之外,其他位置的值是不为0的,因此这种操作只能实现近似的效果。

4.2 编码器与解码器结构

编码器和解码器网络结构是现阶段语义分割的常用网络结构模型中非常常见的结构,同时编码器与解码器网络结构也是普遍应用了转置卷积处理的一个非常典型的网络结构。

例如,编码器和解码器结构U-Net的提出就是融合了顶层和底层的分辨率特征,有助于使解码器生成高分辨率的语义特征[36],而U-Net也被应用于遥感图像的语义分割中[37],对于遥感图像中的每一类别的目标训练出一个二分类模型,利用生成的预测子图生成最终的语义分割图像。该网络有着较高的准确率和较快的训练速度,在遥感领域应用上也需要尝试提高不同的类别和尺度的图像分割精度。

2018年,牛津大学的Bilinski等人推出了一种单通道的编解码器网络[38],将解码器模块化,生成语义特征,且允许解码器使用特征映射的时候连接所有更高级别的解码器从而获得特征地图,一定程度上有助于解决编解码器网络在某些特定使用场景的数据集进行耗时较长的多尺度平均才能达到较好的分割精度的问题。

2019年,杜星悦[39]等人提出将编解码器网络应用到人脸区域分割中,在编码器部分使用了CNN+RELU层,

解码器部分将上池化和最大池化层交替使用,对比常用的人脸分割算法的鲁棒性和精度都得到很大的提高,应用编解码器网络可以小样本数据集通过数据增加技术,实现良好的人脸区域分割效果。

在最新的CVPR会议上,同样有值得关注的利用编解码器网络创新的方法,旷视科技的孙剑博士推出了一种实时语义分割模型DFANet,在相同精度下[40]对比先前常用的语义分割算法减少了7倍的计算量,其通过深度多层的聚合结构利用高层特征信息,编码器结构使用了DFA轻量级特征聚合结构提高运算速度

如图4所示,编码器-解码器网络SegNet主要分为两个部分,其中编码器(Encoder)部分称为卷积提取特征部分,提取主要特征后,图片的尺寸将会逐渐减少。解码器(Decoder)部分则主要是包括了反卷积池化与上采样操作(Unsampling)解码器中的池化层使用index功能,能够在反卷积的过程中保存通过max选出卷积核的相对位置,在上采样(unsampling)过程中随着图片尺寸的变大filter会丢失权重位置,而利用index功能就可以将数据摆放入原矩阵的位置中。在图片经过编码器部分和解码器部分后的输出图像还要再经过softmax层,求出每一个像素在所有的类别中的最大概率,最后为解码器输出的图片添加上label标签,从而得到语义分割的效果,SegNet对物体边缘分割的精度较差,导致分割的效果远不能在实际中应用。

文献[41]对于解码器网络进行了创新研究,采用了不同于一般的双线性插值的上采样方法,采用了Data-dependent Up-sampling(DUpsample)减少编解码器网络的计算量,通过将解码器的下采样和解码器融合,以像素的方式恢复粗略输出的预测解码器结构,但是将该解码器结构应用于整体基于DeepLabv3+的编解码器网络中,发现不同的softmax层会影响分割精度,甚至在没有使用softmax层的时候分割精度低于现阶段大多数的编解码器网络。

4.3　卷积神经网络后处理方法

为解决FCN现存的缺点,还可以对图像进行处理使图像的边缘分割精度更加准确,条件随机场模型(Conditional Random Field,CRF)[42]最为著名,利用CRF模型对FCN的输出图进行优化可以得到很好的效果[43]。

Chen[44]等人在Deeplabv1网络结构中首先提出了一个全连接条件随机场的概念,这种网络结构可以将空洞卷积应用在DCNN网络结构中,最后采用CRF后处理方法对分割后的图像进行细节增强。但对精细的细节容易忽略。

2016年,Deeplabv2[44]结构中采用了多尺度金字塔池化的方法提高了图像分割的鲁棒性,但是在2017年更新的Deeplab版本中,Chen等人却选择放弃了CRF模块,对于原Deeplab网络结构中的空洞卷积层做出了修改,而在最新的版本Deeplabv3+中选择了编解码器网络结构加快运行速度[45],由此可以看出在网络结构模型中添加CRF模块已经逐渐被抛弃。究其原因在于CRF存在着学习速度慢,优化较低且完全取决于手动设置的高斯特征值的缺点。

当添加CRF模块的方法由于其自身缺点的缘故,已经少有新的算法使用时,Techmann[46]等人提出了ConvCRF算法模型,将条件随机场添加进卷积层中,对于图片中的对象局部特征处理能够取得比较好的效果。ConvCRF模型使用类似于CNN的卷积层实现,将信号传递步骤设置为带截断高斯核的卷积核,并应用于现有的卷积神经网络中处理卷积层,处理后GPU的运算时间大大加快。ConvCRF缺陷在于需要建立一个本地的底层替换常规的卷积操作的话,否则数据需要在GPU上进行重新组织,导致大量的运算资源和时间耗费在数据重组上。如何简便化,也成为了目前ConvCRF所急需处理的一个问题。

5　基于深度学习的语义分割方法常用网络框架

网络模型在深度学习中应用需要深度学习网络框架来支持,在视觉领域中Caffe框架[47]最广为人知。其优势就在于简洁快速,虽然内核是C++编写的却提供了Python和Matlab接口,但灵活性较差,代码不够精简,特别是在应用大型网络进行处理时过于繁琐。

TensorFlow[48]框架是现在最流行的深度学习框架,TensorFlow框架可以不仅在PC端上应用,也可以在移动设备上训练模型,支持各种CNN网络结构,缺点在于底层代码过于复杂,虽然有很多接口快速迭代,但向后兼容性较差,且运行的速度比其他的框架慢。

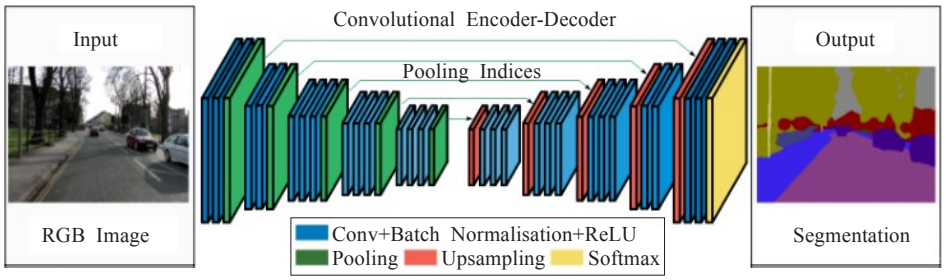


图4　以SegNet为例的编码器与解码器网络

Keras[49]就是一个对于用户来说较为友好的API,意味着能够将所有东西表示为模块,用户可以对其进行自由组合,也可以写尽可能少的代码就能测试一个神经网络,但是也存在着处理速度较慢,性能不足的缺点。

2017年,Facebook公司根据Python开发出了Pytorch[50]框架,在处理图像方面使用了Python版本的Torch库,好处在于提供动态计算图,意味着图像在运行时生成,对于在GPU上运行更加简单,但由于开发时间较短,缺乏参考资料目前尚待开发。各类型的网络框架,如表2所示。

表2 语义分割最热门的网络框架

语义分割常用框架	优势特点	开发日期	热门程度
Caffe	模块化,纯粹的C++/CUDA框架	2016年	★★★★
TensorFlow	基于Python运行速度快,能够产生网络拓扑图	2015年	★★★★★
Keras	简单快速的设计底层的库提供很大的弹性,模块化,允许自定义网络结构参数	2015年	★★★
Pytorch	利用诸如Numpy的抽象方法来表征多维数组	2017年	★★★★

6 语义分割算法指标

本章将当前诸多图像语义分割算法进行性能评估的归纳,整理出比较常用的数据集,对经典的图像语义分割算法进行归纳。

6.1 语义分割评估指标

语义分割技术经过多年的发展,对于现今种种的网络结构和模型需要一个公认的算法评估指标和统一的标准,在2017年的CVPR会议上Garcia-Garcia[51]等人专门撰写了一篇文章对诸多的现行数据集和网络模型进行了评估。

语义分割技术的标准评估点主要在以下的几个方面:时间复杂度、内存占用率和精确度,其中精确度包括了像素精度(Pixel Accuracy,PA)、均像素精度(Mean Pixel Accuracy,MPA)、均交并比(Mean Intersection over Union,MIOU)。

首先可以做一个假设,共有 $K+1$ 个类别(从 L_0 到 L_k,其中会包含一个空类或者背景),p_{ij} 表示的则是本来属于类 i 但是会被预测为类 j 的像素数量。简单来说就是 p_{ii} 表示的就是真正的数量,而 p_{ij}、p_{ji} 就会分别被解释成为假正和假负,尽管这两者其实都是假正与假负之和。

PA即像素精度,这其实是最简单的度量,用以标记正确的像素占图片总像素的整体比例。

$$PA=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{1}$$

MPA即均像素精度,这是对PA一种提升的测量精度,用以计算每个类内被正确分类的像素数的比例,之后再求出所有类的平均。

$$MPA=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{2}$$

MIOU即均交并比,成为了语义分割的标准度量。MIOU用于计算两个集合的交集和并集,具体在语义分割中指的就是真实值(ground truth)和观测值(predicted segmentation)这两个集合,而在其中所描述的这个比例可以变形为正真数(intersection)比并集(正真、假负、假正)的所得之和。在每个类上计算IOU,之后进行平均处理。

$$MIOU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{3}$$

6.2 图像语义分割数据集

在测试两种不同的网络结构或者模型的时候需要存在着一个统一的标准,同样的在研发出新的基于深度学习图像语义分割技术的网络模型时候也需要以一个共同的标准来进行判断,上节提到了图像语义分割的算法评价指标,但是对于不同的图像不同的处理方式之间可能存在着互有优劣的情况,这个时候就需要以同一个图片数据集来进行测试从而得到评估的指标。常用的语义分割数据集优劣性对比如表3所示。

早在2008年,剑桥大学的研究人员就发布了第一个道路、驾驶场景的图片数据集CamVid[52-53],该数据集来自于一个车载960×720分辨率的摄像头所拍摄的5个视频序列,在数据集中包括了32类的物体和701张图片。

PASCAL VOC[54]数据集经过从2005年发展到2012年的发展,其含有20种类别,道路场景图片数据有着11 520张图片,包含着27 450个注释对象,这是目前最常用的数据集。2014年,以PASCAL VOC为基础,PASCAL CONTEXT[55]对于物体的类别以及数据集中的图片数量进行了大量的扩充,数据类别已经扩充到33种类别。

MS COCO训练集[56]包括了200 000个图像和80个图像实例,已经公开了5 000 000个对象实例,数据集中主要包括了室内场景和室外场景。

SUN RGB-D[57]数据集由4个RGB-D传感器获取而得,其中包含了10 000个RGB-D图像,比例类似于

表3　常用语义分割数据集汇总表

时间	数据集名称	说明	主要应用场景
2009年	CamVid	包括了146 617个二维多边形和58 657个具有精确对象方向的3D边界框,数据集中包含了NYU Depth V2、Berkeley B3DO,适合于场景理解任务	道路场景
2011年	SiftFlow	收集了731个包含着102 206帧的视频作为实验数据库是LabelMe的数据集子集,图像主要包含着8种不同的户外场景	自然景观
2012年	PASCAL VOC	含有20种类别,道路场景数据有着11 520张图片,包含着27 450个注释对象	道路行人车辆
	NYU Depth V2	主要提供了1 449个RGBD图像的新数据集,其中捕获了464个不同的室内场景,并附有详细的标注,能够验证3D场景的提示和推断,实现更好的对象分割	室内物体
2014年	PASCAL-CONTEXT	包含了10 103张训练图像的像素级别的标注,共540类	道路行人车辆
	Microsoft COCO	包括了200 000个图像和80个图像实例,已经公开了5 000 000个对象实例,数据集中主要包括了室内场景和室外场景	室内室外的常用场景
2015年	Cityscape	是一个提供了城市道路场景的的数据集,来自50个不同的城市街景记录的立体视频序列,包括了20 000张弱注释图片和5 000张的高质量的强注释的图片,涵盖了各种时间及天气变化下的街道动态物体	道路车辆、行人、街景
	SUN RGB-D	是由4个RGB-D传感器获取而得,其中包含了10 000个RGB-D图像,比例类似于PASCAL VOC,整个数据集包括了146 617个二维多边形和58 657个具有精确对象方向的3D边界框,数据集中包含了NYU Depth V2、Berkeley B3DO,适合于场景理解任务	室内物体3D模型
2017年	ADE20K	包含了SUN和Places数据集的场景范畴,可视化目标目前只展示了超过250个带有注解示例的目标,以及带有超过10个注解示例的部件	室内室外景观
2019年	CityFlow	从10个路口提取的40个摄像头收集到的视频,是目前都市环境中最大规模的数据集,包含超过20万个目标框	道路场景

PASCAL VOC,整个数据集包括了146 617个二维多边形和58 657个具有精确对象方向的3D边界框,数据集中包含了NYU Depth V2、Berkeley B3DO,适合于场景理解任务。

ADE20k[58-59]数据集提供了多次注释的64个图像集,拥有者20 000张训练数据,若干张test数据,其麻烦在于需要上传到服务器才能得到结果,有着150类图片包括了室内和室外的场景。

Cityscape[60]则是一个提供了城市道路场景的数据集,来自50个不同的城市街景记录的立体视频序列,包括了20 000张弱注释图片和5 000张的高质量的强注释的图片,涵盖了各种时间及天气变化下的街道动态物体。

NYU Depth V2[61]是一个提供了1 449个RGBD图像的新数据集,其中捕获了464个不同的室内场景,并附有详细的标注,能够验证3D场景的提示和推断,实现更好的对象分割。

SiftFlow[62]数据集收集了731个包含着102 206帧的视频作为实验数据库,是LabelMe[63]的数据集子集,图像主要包含着8种不同的户外场景。

CityFlow[64]数据集包括了从10个路口提取的40个摄像头采集的超过3个小时的同步高清视频,有着多样的场景、视角和车辆模型的信息。

6.3　算法模型性能对比

深度学习在语义分割领域上取得了巨大的成功,本文将大部分对于语义分割领域有着重大推动作用的文章进行了列举综述,在上节将常用的基于深度学习的语义分割方法的数据集进行了评述列举,结合前文的内容在表4中对所提的图像语义分割方法进行了分析比较,主要包括算法分类、算法名称、各种模型算法的优缺点,以及上述各种算法模型在不同数据集上的性能对比。

7　结束语

本文主要对于图像语义分割技术的研究发展历程进行了详细评述,对于传统的语义分割方法到当前主流的基于深度学习的图像语义分割理论及其方法做出了综合性的评估,对基于深度学习语义分割技术需要用到的网络模型、网络框架、分割流程进行了详细的评估。在深入该领域后发现该领域仍然存在着非常多的未知问题值得深入探究。

基于以上分析,提出今后的研究方向:

(1)实时语义分割技术。现阶段评价应用于语义分割的网络模型主要着重点在精确率上,但是随着应用于现实场景的要求越来越高,需要更短的响应时间,因此在维持高精确率的基础上,尽量缩短响应时间应是今后工作的方向。

(2)弱监督或无监督语义分割技术。针对需要大量的标注数据集才能提高网络模型的精度这个问题,弱监督或无监督的语义分割技术将会是未来发展的趋势。

(3)三维场景的语义分割技术。目前的诸多基于深度学习的语义分割技术所用以训练的数据主要是二维的图片数据,同时测试的对象往往也是二维的图片,但

表4 基于深度学习的语义分割方法总结

分类	时间	网络结构模型	算法特点	算法待改进处	网络框架	数据集	分割精度MIOU/%
基于传统方法的语义分割技术	2000年	Normalized cut	将图分为 K 个子图,并且保证 k 个子图的割最小	随着分割图尺寸增大,计算复杂度增大	—	—	—
	2004年	Grab cut	利用了图像中的边界信息,依靠人工标注得到较好的分割结果	需要大量的人工标注数据	—	—	—
基于深度学习的语义分割技术	2012年	FCN	以CNN网络为基础架构,引入全卷积层	对于图片上下文信息考虑不足,分割精度较低	Caffe	PASCAL VOC/PASCAL-Context	62.20/53.50
	2014年	Deeplsb v1	将空洞卷积和DCNN网络结合,同时采用全连接条件随机场进行优化	DCNN网络会导致空间分辨率下降,需要大量的存储空间	Caffe	PASCAL VOC	71.60
	2015年	CRFasRNN	将CRF与RNN结合成端对端的网络,提高FCN的分割精度	使用更多的RNN网络才能更好地提高分割精度,缺乏对上下文信息的利用	Caffe	PASCAL VOC/PASCAL Context	74.70/39.28
	2015年	SegNet	利用了编码器-解码器网络结构,利用上采样方式恢复图像尺寸,提高图像分割精度,内存效率较高	分割精度不高,不能满足可实际使用的需要	Caffe	CamVid/SUNRGB-D	60.10/31.84
	2015年	DeconvnNet	针对FCN网络进行改善,引入深度反卷积网络	在光照对比强烈的场景中分割效果不如FCN	Caffe	PASCAL VOC	69.60
	2015年	Deeplab v2	使用空洞卷积层代替上采样方式,使用多尺度空间金字塔池化	未能捕获细微的物体边界,无法恢复细节处理	Caffe	PASCAL VOC/Cityscapes	79.70/70.40
	2016年	G-CRF	结合高斯特征条件随机场和深度学习,应用结构化模型得到全局最优解	过于密集的CRF处理虽然改善了性能但是会忽视了图像中精细的细节	Caffe	PASCAL VOC	80.20
	2017年	RefineNet	对解码器结构进行改进,形成long-range残差连接,能通过上采样方式融合底层和高层语义特征	网络容量较大,需要更长的训练时间	Pytorch	NYUDv2/PASCAL VOC/PASCAL-Context/SUN-RGBD/ADE20K	46.50/83.40/47.30/45.90/40.70
	2017年	Deeplab v3	对空洞金字塔层进行了改进,在并行空洞卷积模型中加入了两个1×1的卷积层	输出图的放大效果较差	TensorFlow	PASCAL VOC	85.70
	2017年	ERFNet	引入新的残差层模块,卷积层通过1Dfilter改进模型的紧凑性	牺牲了一定的精度降低运算量和功耗	Torch7	Cityscapes	87.30
	2018年	DeepLabv3+	采用了编码器-解码器网络结构改善对物体边缘的分割效果	还可以在解码器模型结构和深度可分离卷积上进行创新,进一步提高模型速度和性能	TensorFlow	PASCAL VOC/Cityscapes	89.00/82.10
	2018年	ConvCRF	将CRF引入卷积层中,利用卷积CRF能够比全连接CRF获得更高的运算速度	对于精细的细节容易忽视	Pytorch	PASCAL VOC	72.18
	2019年	DFANet	利用深度多层聚合结构利用网络中的高层特征,应用轻量级编码器将信息聚合,降低运算量	—	Pytorch	Cityscapes/CamVid	71.30/64.70

是在实际应用时所面对的环境是一个三维环境,将语义分割技术应用至实际中,未来需要针对三维数据的语义分割技术进行研究。

参考文献:

[1] 张新明,祝晓斌,蔡强,等.图像语义分割深度学习模型综述[J].高技术通讯,2017(9):808-815.

[2] 田萱,王亮,丁琪.基于深度学习的图像语义分割方法综述[J].软件学报,2019,30(2):440-468.

[3] 王宇,张焕君,黄海新.基于深度学习的图像语义分割算法综述[J].电子技术应用,2019,45(6):23-27.

[4] Sharif Razavian A,Azizpour H,Sullivan J,et al.CNN features off-the-shelf:An astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2014:806-813.

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

[6] Michaelis C,Ustyuzhaninov I,Bethge M,et al.Ecker.:One-shot instance segmentation[J].arXiv:1811.11507,2018.

[7] He K,Gkioxari G,Dollár P,et al.Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:2961-2969.

[8] Huang Z,Huang L,Gong Y,et al.Mask scoring R-CNN[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019:6409-6418.

[9] Wang Haiyang,Pan Dehua,Xia Deshen.Fast implementation of two-dimensional Otsu adaptive threshold selection algorithm[J].Acta Automatica Sinica,2007(9):968-971.

[10] Coates A,Ng A Y.Learning feature representations with *k*-means[M]//Neural networks:Tricks of the trade.Berlin,Heidelberg:Springer,2012:561-580.

[11] Liu Songtao,Yin Fuliang.Image segmentation method based on graph cut and its new development[J].Acta Automatica Sinica,2012,38(6):911-922.

[12] Shotton J,Johnson M,Cipolla R.Semantic texton forests for image categorization and segmentation[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition,2008:1-8.

[13] Shi J,Malik J.Normalized cuts and image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,22(8):888-905.

[14] Rother C,Kolmogorov V,Blake A.Grabcut:Interactive foreground extraction using iterated graph cuts[J].ACM Transactions on Graphics(TOG),2004,23(3):309-314.

[15] Golovinskiy A,Funkhouser T.Min-cut based segmentation of point clouds[C]//Proceedings of 2009 IEEE 12th International Conference on Computer Vision,2009:39-46.

[16] Han S,Tao W,Wang D,et al.Image segmentation based on GrabCut framework integrating multiscale nonlinear structure tensor[J].IEEE Transactions on Image Processing,2009,18(10):2289-2302.

[17] Krizhevsky A,Sutskever I,Hinton G E.Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems,2012:1097-1105.

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

[19] 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.

[20] Szegedy C,Ioffe S,Vanhoucke V,et al.Inception-v4,inception-resnet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,2017.

[21] Jiang F,Grigorev A,Rho S,et al.Medical image semantic segmentation based on deep learning[J].Neural Computing and Applications,2018,29(5):1257-1265.

[22] Zhong Z,Jin L,Xie Z.High performance offline handwritten chinese character recognition using googlenet and directional feature maps[C]//Proceedings of 2015 13th International Conference on Document Analysis and Recognition,2015:846-850.

[23] Ronneberger O,Fischer P,Brox T.U-net:Convolutional networks for biomedical image segmentation[C]//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention,2015:234-241.

[24] Badrinarayanan V,Kendall A,Cipolla R.Segnet:A deep convolutional encoder-decoder architecture for image segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(12):2481-2495.

[25] Chen L C,Papandreou G,Kokkinos I,et al.Semantic image segmentation with deep convolutional nets and fully connected CRFs[J].Computer Science,2014(4):357-361.

[26] Yu C,Wang J,Peng C,et al.Bisenet:Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision,2018:325-341.

[27] Romera E,Alvarez J M,Bergasa L M,et al.Erfnet:Efficient residual factorized convnet for real-time semantic segmentation[J].IEEE Transactions on Intelligent Transportation Systems,2017,19(1):263-272.

[28] Lin G,Milan A,Shen C,et al.Refinenet:Multi-path

refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:1925-1934.

[29] Sainath T N,Mohamed A,Kingsbury B,et al.Deep convolutional neural networks for LVCSR[C]//Proceedings of 2013 IEEE International Conference on Acoustics,Speech and Signal Processing,2013:8614-8618.

[30] Scherer D,Müller A,Behnke S.Evaluation of pooling operations in convolutional architectures for object recognition[C]//Proceedings of International Conference on Artificial Neural Networks,2010:92-101.

[31] Yu F,Koltun V.Multi-scale context aggregation by dilated convolutions[J].arXiv:1511.07122,2015.

[32] Wang P,Chen P,Yuan Y,et al.Understanding convolution for semantic segmentation[C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision,2018:1451-1460.

[33] Chen L C,Papandreou G,Schroff F,et al.Rethinking atrous convolution for semantic image segmentation[J].arXiv:1706.05587,2017.

[34] Noh H,Hong S,Han B.Learning deconvolution network for semantic segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:1520-1528.

[35] Dumoulin V,Visin F.A guide to convolution arithmetic for deep learning[J].arXiv:1603.07285,2016.

[36] 张永宏,何静,阚希,等.遥感图像道路提取方法综述[J].计算机工程与应用,2018,54(13):1-10.

[37] 苏健民,杨岚心,景维鹏.基于U-Net的高分辨率遥感图像语义分割方法[J].计算机工程与应用,2019,55(7):207-213.

[38] Bilinski P,Prisacariu V.Dense decoder shortcut connections for single-pass semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:6596-6605.

[39] 杜星悦,董洪伟,杨振.基于深度网络的人脸区域分割方法[J].计算机工程与应用,2019,55(8):171-174.

[40] Li H,Xiong P,Fan H,et al.Dfanet:Deep feature aggregation for real-time semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019:9522-9531.

[41] Tian Z,He T,Shen C,et al.Decoders matter for semantic segmentation:Data-dependent decoding enables flexible feature aggregation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019:3126-3135.

[42] Lafferty J,Mccallum A,Pereira F C N.Conditional random fields:Probabilistic models for segmenting and labeling sequence data[J].Proceedings of ICML,2001,3(2):282-289.

[43] Zheng S,Jayasumana S,Romera-Paredes B,et al.Conditional random fields as recurrent neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:1529-1537.

[44] Chen L C,Papandreou G,Kokkinos I,et al.DeepLab:Semantic image segmentation with deep convolutional nets,atrous convolution,and fully connected CRFs[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,40(4):834-848.

[45] Chen L C,Zhu Y,Papandreou G,et al.Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision,2018:801-818.

[46] Teichmann M T T,Cipolla R.Convolutional CRFs for semantic segmentation[J].arXiv:1805.04777,2018.

[47] Jia Y,Shelhamer E,Donahue J,et al.Caffe:Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM International Conference on Multimedia,2014:675-678.

[48] Girija S S.Tensorflow:Large-scale machine learning on heterogeneous distributed systems[J].arXiv:1603.04467,2016.

[49] Gulli A,Pal S.Deep learning with Keras[M].[S.l.]:Packt Publishing Ltd,2017.

[50] Ketkar N.Introduction to pytorch[M]//Deep learning with Python.Berkeley,CA:Apress,2017:195-208.

[51] Garcia-Garcia A,Orts-Escolano S,Oprea S,et al.A review on deep learning techniques applied to semantic segmentation[J].arXiv:1704.06857,2017.

[52] Brostow G J,Shotton J,Fauqueur J,et al.Segmentation and recognition using structure from motion point clouds[C]//Proceedings of the European Conference on Computer Vision,2008:44-57.

[53] Brostow G J,Fauqueur J,Cipolla R.Semantic object classes in video:A high-definition ground truth database[J].Pattern Recognition Letters,2009,30(2):88-97.

[54] Everingham M,Eslami S M A,Gool L J V,et al.The pascal visual object classes challenge:A retrospective[J].International Journal of Computer Vision,2015,111(1):98-136.

[55] Mottaghi R,Chen X,Liu X,et al.The role of context for object detection and semantic segmentation in the wild[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2014:891-898.

[56] Lin T Y,Maire M,Belongie S,et al.Microsoft coco:Common objects in context[C]//Proceedings of the European Conference on Computer Vision,2014:740-755.

(下转第42页)

scale datasets and predictive methods for 3D human sensing in natural environments[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,36(7):1325-1339.

[53] Sigal L,Balan A O,Black M J.Humaneva:Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J].International Journal of Computer Vision,2010,87(1/2):4.

[54] Mehta D,Rhodin H,Casas D,et al.Monocular 3D human pose estimation in the wild using improved CNN supervision[C]//Proceedings of the International Conference on 3D Vision,2017:506-516.

[55] Trumble M,Gilbert A,Malleson C,et al.Total capture:3D human pose estimation fusing video and inertial sensors[C]//Proceedings of 28th British Machine Vision Conference,2017:1-13.

[56] Fabbri M,Lanzi F,Calderara S,et al.Learning to detect and track visible and occluded body joints in a virtual world[C]//Proceedings of the European Conference on Computer Vision,2018:430-446.

[57] Varol G,Romero J,Martin X,et al.Learning from synthetic humans[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:109-117.

[58] von Marcard T,Henschel R,Black M J,et al.Recovering accurate 3D human pose in the wild using IMUs and a moving camera[C]//Proceedings of the European Conference on Computer Vision,2018:601-617.

[59] Lassner C,Romero J,Kiefel M,et al.Unite the people:Closing the loop between 3D and 2D human representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:6050-6059.

[60] Pavllo D,Feichtenhofer C,Grangier D,et al.3D human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019:7753-7762.

[61] 韩贵金,朱虹.一种基于图结构模型的人体姿态估计算法[J].计算机工程与应用,2013,49(14):30-33.

[62] Witoonchart P,Chongstitvatana P.Structured SVM back-propagation to convolutional neural network applying to human pose estimation[J].Neural Networks,2017,92:39-46.

[63] 蔡雅薇,谭晓阳.弱监督任意姿态人体检测[J].计算机科学与探索,2017,11(4):587-598.

[64] Cao Z,Hidalgo G,Simon T,et al.OpenPose:Realtime multi-person 2D pose estimation using part affinity fields[J].arXiv:1812.08008,2018.

[65] Howard A G,Zhu M,Chen B,et al.Mobilenets:Efficient convolutional neural networks for mobile vision applications[J].arXiv:1704.04861,2017.

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

[67] Osokin D.Real-time 2D multi-person pose estimation on CPU:Lightweight OpenPose[J].arXiv:1811.12004,2018.

(上接第21页)

[57] Song S,Lichtenberg S P,Xiao J.SUN RGB-D:A RGB-D scene understanding benchmark suite[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015:567-576.

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

[59] Zhou B,Zhao H,Puig X,et al.Semantic understanding of scenes through the ade20k dataset[J].International Journal of Computer Vision,2019,127(3):302-321.

[60] Cordts M,Omran M,Ramos S,et al.The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:3213-3223.

[61] Silberman N,Hoiem D,Kohli P,et al.Indoor segmentation and support inference from RGBD images[C]//Proceedings of European Conference on Computer Vision,2012:746-760.

[62] Tighe J,Lazebnik S.Superparsing-scalable nonparametric image parsing with superpixels[J].International Journal of Computer Vision,2013,101(2):329-349.

[63] 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(1/3):157-173.

[64] Tang Z,Naphade M,Liu M Y,et al.CityFlow:A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification[J].arXiv:1903.09254,2019.