基于深度网络的图像语义分割综述

罗会兰,张 云
(江西理工大学信息工程学院,江西赣州 341000)

摘 要: 图像语义分割不仅预测一幅图像中的不同类别,同时还定位不同语义类别的位置,具有重要的研究意义和应用价值. 本文阐述了图像语义分割最新的研究成果和方法,从三个角度综述了基于深度卷积神经网络的图像语义分割模型,分别是基于候选区域模型、基于全卷积网络模型和基于弱监督学习的语义分割模型,对这三类模型的方法和结构进行了详细的研究和分析. 并在 PASCAL VOC 2012 数据集上对一些代表性的语义分割算法的性能进行了比较分析.

关键词: 图像语义分割; 深度卷积神经网络; 候选区域; 全卷积网络; 弱监督学习; PASCAL VOC 2012 数据集

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)10-2211-10

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.10.024

A Survey of Image Semantic Segmentation Based on Deep Network

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

Abstract: Image semantic segmentation not only predicts different categories in an image, but also orientates different semantic categories locations, which has important research significance and application value. This paper expounds the latest research results and methods of image semantic segmentation, and from three perspectives: based candidate region models、based full convolutional network models and based weakly-supervised learning models. The image semantic segmentation model methods and structure based on deep convolutional neural network are deeply overviewed in this paper. This paper compares the performance of some representative semantic segmentation algorithms on the PASCAL VOC 2012 dataset.

Key words: image semantic segmentation; deep convolutional neural network; candidate region; full convolution network; weakly-supervised learning; PASCAL VOC 2012 dataset

1 引言

语义分割结合了图像分类、目标检测和图像分割,通过一定的方法将一幅图像中的像素进行分类. 语义分割将图像分割成具有一定语义含义的区域块,并识别出每个区域块的语义类别,实现从底层到高层的语义推理过程,最终得到具有逐像素语义标注的分割图像. 语义分割是室内导航、地理信息系统、人机交互、自动驾驶、虚拟增强现实系统、场景理解、医学图像处理以及目标分类等视觉分析的基础. 图像语义分割是一个非常具有挑战性的问题,其难点主要体现在以下两个方面:一是类别层面上所面临的难点,即类内实例间的相异性和类间物体的相似性;二是复杂的背景,实际场景中的背景往往是错综复杂的,这种复杂性大大提升了图像语义分割的难度.

图像语义分割方法有传统方法和基于卷积神经网络的方法,其中传统的语义分割方法又可以分为基于统计的方法[1,2]和基于几何的方法[3,4]. 大多数统计方法[1,2]是基于多个简单的特征,这类似于图像分割方法,利用人工设计的特征提取方法得到图像底层特征,没有训练过程,分割效果不理想. 基于几何的方法[3,4]是通过 2D 图像来推断 3D 图像的空间布局,利用在物理上的有效结构假设,找到线段的最佳拟合模型,将其转换为完整的 3D 模型,很好地解决了语义分割中的目

收稿日期:2018-11-23;修回日期:2019-05-13;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61462035,No.61862031);江西省青年科学家培养项目(No.20153BCB23010);江西省自然科学资金项目(No.20171BAB202014)

标遮挡问题[5]. 基于卷积神经网络的语义分割方法[6~15]与传统的语义分割方法最大不同是,网络可以自动学习图像的特征,进行端到端的分类学习,极大提升语义分割的精确度.

2015 年,Zhu 等人[16]综述了传统的语义分割方法,分别介绍了无监督方法,弱监督和全监督方法. 2016 年,Thoma 等人[17]对传统的语义分割和 Alexnet 网络相关的分割技术进行了综述. 2017 年,Guo 等人[18]对语义分割领域期刊或来源进行综述,并总结了它们的优势、劣势和主要挑战. 之后,Alberto 等人[19]对各种应用场景下利用深度学习技术解决语义分割的方法进行了综述. 2018 年,Geng 等人[20]研究了卷积神经网络(Convolution Neural Network,CNN)新颖复杂的网络层,结构和策略,以及那些已经在 PASCAL VOC 2012 数据集[21]中取得先进成果的语义分割方法.

与以上图像语义分割研究综述不同的是,本文从一个新颖的视角阐述了图像语义分割研究近年来的经典模型,将语义分割深度模型分为三大类,基于候选区域模型、基于全卷积网络模型、基于弱监督语义分割模型,且在最常用的 PASCAL VOC 2012[21]数据集上对模型进行了对比和评估.

2 基于深度卷积神经网络的语义分割模型

基于深度卷积神经网络(Deep Convolution Neural Network,DCNN)的语义分割方法,将图像通过卷积神经网络提取特征,然后进行像素分类,得到语义分割图像. 图 1 是基于深度网络的图像语义分割模型分类图,下面分别对这三类模型进行具体的分析.

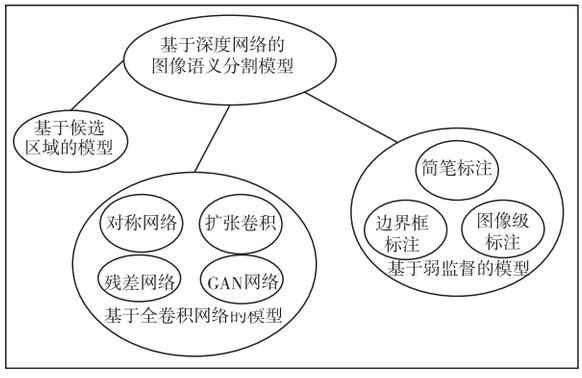


图1 基于深度网络的图像语义分割模型分类图

2.1 基于候选区域的语义分割模型

基于候选区域的语义分割方法首先从图像中提取自由形式的区域并对它们的特征进行描述,然后再基于区域进行分类,最后将基于区域的预测转换为像素级预测,使用包含像素最高得分的区域来标记像素. 基于候选区域的语义分割模型最早由 Carreira 等人[22]于 2012 年提出,采用 CPMC(Constrained Parametric Min-Cut)算法[22]产生候选区域,对这些区域进行打分排序,计算出这些区域属于特定类别的概率大小,直到得出最终的分割图像. 在此基础上,Carreira 等人[23]在特征描述阶段采用 SOP(Second-Order-Pooling)方法,利用 SOP 聚合区域的局部特征,利用黎曼流形的几何特性来描述任意区域的特征,取得了更高的分割精度.

2014 年,Girshick 等人[24]提出 R-CNN(Region-based Convolutional Neural Networks)模型,用于目标检测和语义分割. 首先利用 Selective Search 算法[25]在图像上提取大量的候选区域,它将输入图像的大小固定为 227 × 227,再用 CNN 提取每个候选区域的特征,最后使用支持向量机对每个区域进行分类,得到最终的语义分割图像. Hariharan 等人[26]在 R-CNN 的基础上加入了额外网络,利用 MCG[27](Multiscale Combinatorial Grouping)算法提取候选区域,同时获得矩形区域和区域中的目标候选区域,对输入图像的大小没做具体要求. 鉴于生成候选区域集需要大量的时间,且候选区域的质量直接影响最终的语义分割结果,2015 年 Ren 等人[28]提出区域生成网络来提高候选区域的生成速度,采用滑动窗口直接生成候选区域,这些候选区域与检测网络共享卷积特征,有效地提高了候选区域的速度和准确度. 但是该方法不能准确的聚焦于候选区域中感兴趣的区域,Caesar 等人[29]在 Fast R-CNN[30]的基础上,提出基于区域的端到端的语义分割方法,采用自由形式感兴趣区域的池化层来获得候选区域的前景特征,结合了语境信息的优点和自由形式区域表示的优点. 但该方法对于人体分割不太准确,在人机交互以及自动驾驶汽车的场景应用中存在局限性. 针对这个问题,Jiang 等人[31]提出按区域的端到端人体分割网络模型,将人体检测网络与全卷积网络进行融合以估计人的感兴趣区域,通过优化人体边界框坐标轻量级检测网络以实现实时分割,实验表明所提出的模型在运行时间和分割精度上更具竞争性.

2.2 基于全卷积网络的语义分割模型

基于候选区域的模型方法虽然为语义分割的发展带来很大的进步,但候选区域中缺乏空间信息,尤其是对于小物体的信息损失严重,直接影响了最终的语义分割效果. 在此基础上,基于全卷积网络(Fully Convolution Network,FCN)的语义分割模型应运而生[6],它不需要生成候选区域,能够输入任意大小的图像,可以直接实现端到端的像素级预测. 图 2 是基于全卷积网络的语义分割模型结构图.

如图 2 所示,首先将一幅 RGB 图像输入到卷积神经网络,经过多次卷积及池化过程得到一系列的特征图,然后利用反卷积层对最后一个卷积层得到的特征

图进行上采样,使得上采样后特征图与原图像的大小一样,最后对上采样特征图进行逐像素分类. 但 FCN 采用八倍上采样的结果比较模糊和平滑,对图像中的细节不够敏感,没有充分的考虑像素与像素之间的关系,使得分割结果不够精细. 之后,基于全卷积衍生的语义分割模型越来越多,效果也越来越好. 下面主要阐述几类经典的基于全卷积网络衍生的语义分割模型.

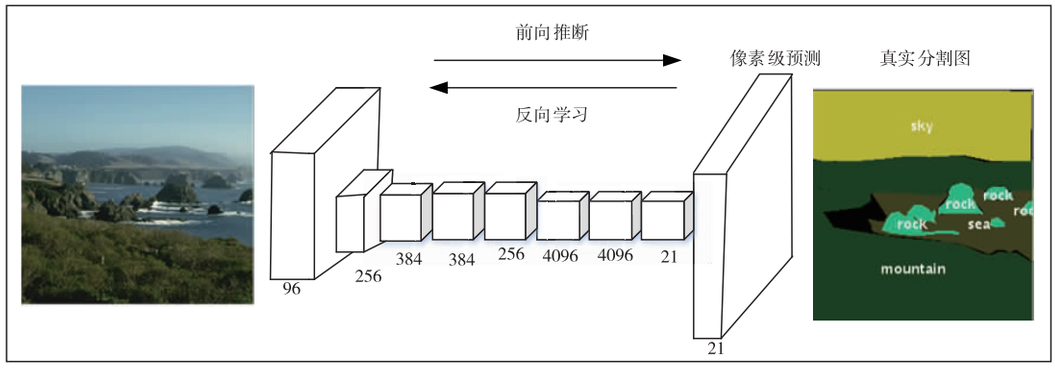


图2　基于全卷积网络的语义分割结构图

2.2.1　基于全卷积的对称语义分割模型

针对 FCN 在语义分割分割物体细节容易丢失或被平滑的问题,Noh 等人[32]提出对称语义分割模型 DeconvNet,在 VGG16[33]的基础上将 Softmax 层移除,添加对称上池化和反卷积层,上池化用来实现目标精确定位,使得特征图恢复到和池化前同样的大小,得到稀疏特征图[34]. 然后,反卷积操作将得到的稀疏特征图变为稠密特征图,有效地解决了大于或小于感受野的物体被碎片化分割或被贴上错误标签以及对于较小物体被疏忽或归类为背景的问题,然而 DeconvNet 模型的参数量太大,Vijay 等人[35]提出 SegNet 对称语义分割模型,但 SegNet 网络相对于 DeconvNet 需要的存储空间和参数量减少,计算效率高,在上采样过程中,SegNet 使用的是卷积,可以减少像素信息的损失,能够提高分辨率和准确定位图像的分割边界. 图 3 是 SegNet 网络结构示图意[35].

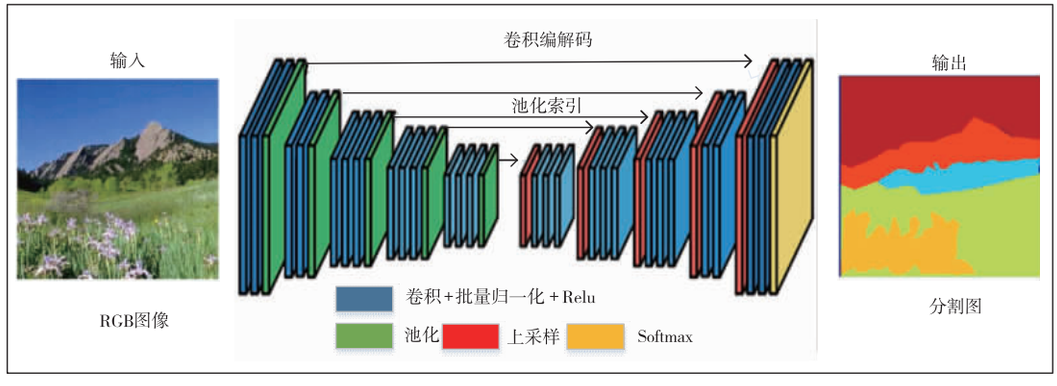


图3　SegNet网络结构示意图

由图 3 可知,SegNet 网络结构包括编码-解码两个部分,编码部分主要由 VGG16[33]网络的前 13 个卷积层和 5 个池化层组成,解码部分同样也由 13 个卷积层和 5 个上采样层组成,同时采用池化索引来保存图像的轮廓信息,降低了参数数量. 最后一个解码器输出的高维特征被送到可训练的 Softmax 分类器中对像素进行分类.

Ronneberger 等人[36]提出了 U-net 对称语义分割模型,该网络模型主要由收缩路径和对称扩张路径组成,收缩路径用来获得上下文信息,对称扩张路径用来精确定位分割边界. U-net 使用图像切块进行训练,所以训练数据量远远大于训练图像的数量,这使得网络在少量样本的情况下也能获得不变性和鲁棒性.

尽管 U-net[36]语义分割模型实现了很好的分割效果,但只能处理 2D 图像,而临床实践中使用的大多数医疗数据都包含 3D 体积. 于是,Milletari 等人[37]提出了新的 V-net 对称语义分割模型,该模型是一种将 3D 体积、全卷积与神经网络相结合的三维分割模型,引入新的目标函数来处理前景和背景数量之间的不均衡问题,解决了训练标注数据集不足的问题,该模型与其它

分割模型相比具有计算优势.

2.2.2 基于全卷积的扩张卷积语义分割模型

基于全卷积对称语义分割模型得到分割结果较粗糙,忽略了像素与像素之间的空间一致性关系. 2014年,Google 提出了一种新的扩张卷积语义分割模型,可以在不增加参数量的情况下增加感受野,目前扩张卷积语义分割模型已有四种版本,分别是 DeepLab v1[38]、DeepLab v2[39]、DeepLab v3[14],DeepLab v3 +[40].

Chen 等人[38]提出的 DeepLab v1 是由深度卷积神经网络和概率图模型级联而成的语义分割模型,由于深度卷积神经网络在重复最大池化和下采样的过程中会丢失很多的细节信息,所以采用扩张卷积算法增加感受野以获得更多上下文信息. 在深度卷积神经网络的最后一层采用全连接条件随机场(Connected Conditional Random Field,CCRF)[41],将不同类别的分类得分和低层次信息进行融合以提高模型捕获边界细节的能力. 在 PASCAL VOC 2012 数据集上的分割准确度为 71.6%.

2016 年,受文献[38]中 CRF 后处理方法的启发,Chandra 等人[42]提出深度学习结构化预测的高斯条件随机场(Gaussian Conditional Random Field,G-CRF)语义分割模型. 该模型结合了深度学习和 G-CRF 的优势,能够和任意损失函数在端到端的结构中联合训练,从而有效地提高语义分割效果. 在 PASCAL VOC 2012 数据集上的分割准确度为 75.46%.

2017 年,受文献[38]启发,Chen 等人[39]提出了 DeepLab v2 语义分割模型. 该模型增加了 ASPP(Atrous Spatial Pyramid Pooling)结构,利用多个不同采样率的扩张卷积提取特征,再将特征融合以捕获不同大小的上下文信息,该模型在 PASCAL VOC 2012 数据集上的分割准确度为 79.7%.

随后,Chen 等人[14]提出了改进版的 DeepLab v3 语义分割模型. 在 ASPP 中加入了全局平均池化,同时在平行扩张卷积后添加批量归一化,有效地捕获了全局语境信息. 该模型在 PASCAL VOC 2012 数据集上的分割准确度提升为 85.7%.

2018 年,Chen 等人[40]进一步提出了 DeepLab v3 + 语义分割模型. 该模型在 DeepLab v3 的基础上增加了编 - 解码模块和 Xception 主干网络,增加编解码模块主要是为了恢复原始的像素信息,使得分割的细节信息能够更好地保留,以及得到更加丰富的上下文信息. 增加 Xception 主干网络是为了采用深度卷积进一步提高算法的精度和速度. 该模型在 PASCAL VOC 2012 数据集上的分割准确度达到 89.0%.

2.2.3 基于全卷积的残差网络语义分割模型

深度卷积网络的每一层特征对语义分割都有影响,将高层特征的语义信息与低层识别的边界与轮廓信息结合有利于提升分割准确度. Pohlen 等人[43]提出了全分辨率残差网络语义分割模型,它对目标具有很强的定位和识别功能. 该模型使用残差流和池化流将多尺度上下文信息和像素级精度结合起来,残差流携带全分辨率信息以实现精确的分割边界,池化流用于获取高层特征. 但该模型对于分割图像的边界处理不够精细,使得一些图像的分割边界模糊或被忽略. 在此基础上,Peng 等人[13]提出大卷积核 - 全局卷积网络语义分割模型(Global Convolutional Network,GCN),该模型使用 ResNet[44] 作为特征网络,FCN4[6] 作为分割框架,能够同时解决语义分割的分类和定位问题. 在目标定位时采用堆叠思想,利用基于残差的边界细化进一步精调目标边界,将每一层的定位结果上采样后,与上一层的定位结果融合相加,得到最后的像素级预测结果. 该模型在 PASCAL VOC 2012 数据集上的分割准确度为 82.2%. 借鉴文献[13]的思想,Lin 等人[45]提出了 RefineNet 语义分割模型,将粗糙的高层语义特征与细粒度的低层特征进行融合,能够充分利用每一层的特征进行语义分割,采用长距离残差连接有效地将下采样过程中损失的像素信息融合进来,从而产生高分辨率的预测图像. 该模型在 PASCAL VOC 2012 数据集上的分割准确度为 83.4%.

Zhao 等人[46]提出金字塔场景稀疏网络语义分割模型(Pyramid Scene Parsing Network,PSP),该模型首先结合预训练网络 ResNet[44] 和扩张网络来提取图像的特征,得到原图像 1/8 大小的特征图,然后,采用金字塔池化模块将特征图同时通过四个并行的池化层得到四个不同大小的输出,将四个不同大小的输出分别进行上采样,还原到原特征图大小,最后与之前的特征图进行连接后经过卷积层得到最后的预测分割图像,该模型在 PASCAL VOC 2012 数据集上的分割准确度达到了 85.4%.

受文献[46]的启发,Luo 等人[47]提出双流图像分割(Dual Image Segmentation,DIS)模型,采用 ResNet101[39] 网络结构提取特征,双流结构中的一个流用来预测图像的像素标签图,另一个使用预测标签图重建图像. 该模型在 PASCAL VOC 2012 数据集上的分割准确度为 86.8%.

2.2.4 基于全卷积的 GAN 语义分割模型

Goodfellow 等人[48]于 2014 年提出生成对抗网络模型(Generative Adversarial Nets,GAN),能够同时训练生成器和判别器,判别器用来预测给定样本是来自于真实数据还是来自于生成模型. 在此基础上,Luc 等人[49]于 2016 年利用对抗训练方法训练语义分割模型,将传统的多类交叉熵损失与对抗网络相结合,使用对抗性损失来微调分割网络. 实验表明该方法能够提高语义

分割的准确度.

2017 年,Metzen 等人[50]提出生成对抗性扰动语义分割模型,通过对输入图像增加一个扰动,使分类器分类错误,从而实现对深度卷积网络的攻击,利用对抗训练提高语义分割的准确度. 受文献[50]的启发,Xie 等人[51]提出密集对抗生成(Dense Adversary Generation, DAG)语义分割模型,对抗性扰动可以通过网络由不同的训练数据传输,结合各种不同的扰动会带来更好的传输性能,它提供了一种有效的"黑盒"对抗攻击训练方法,有效地提高了网络的分割能力.

此外,生成对抗网络在生物医学方面也得到了广泛的应用,Zhu 等人[52]提出一种用于乳房 X 线照片肿块分割的端到端语义分割模型,乳房 X 线照片数据集的尺寸较小,使用对抗训练来控制过拟合. 该模型结合了 FCN 模拟势函数和条件随机场 CRF 结构化学习的优势,将 FCN 与先验位置相结合来确定肿块的位置,获得了很好的分割结果. Rezaei 等人[53]通过条件对抗训练,提出用于端到端训练的脑肿瘤语义分割模型,该模型利用条件生成对抗网络(Conditional Generative Adversarial Nets,CGAN)来训练语义分割卷积神经网络和对抗网络,实验表明所提出的模型在脑肿瘤分割任务中具有很强的优越性.

2.3　基于弱监督学习的语义分割模型

基于全卷积的语义分割模型需要大量的像素级标注训练样本,然而,要获得具有像素级标签的训练样本成本很大. 所以有些研究者提出了一些弱监督方法,通过使用边界框标注、简笔标注和图像级标注来实现语义分割.

2.3.1　边界框标注

为了扩展可用数据集,Dai 等人[54]使用易获取的边界框标注数据集来训练分割模型,该模型在自动生成候选区域与训练卷积网络之间交替进行,通过 MCG[27]来选取带有语义标注的区域,采用卷积网络生成候选区域的分割掩膜. 在网络迭代时,由于边界框可以增强网络识别目标的能力,通过更新卷积网络中的参数来校正分割掩膜以提升语义分割效率. 在此基础上,Papandreou 等人[12]提出采用期望最大化方法,使用弱注释数据来训练语义分割模型,实验结果表明,仅使用具有边界框注释的数据集进行弱监督训练的精确度与全监督训练所达到的分割水平相当.

最近,Khoreva 等人[55]提出一种不需要修改训练过程的弱监督语义分割模型,将弱监督问题视为输入标签噪声问题,利用递归训练作为去除噪声策略,仅训练一次就足以改善先前的弱监督结果,实验表明,在相同的训练条件下,该模型可以达到全监督模型 95% 的精确度.

2.3.2　简笔标注

Lin 等人[56]提出基于用户交互的图像语义分割方法,使用简笔对图像进行注释,利用图模型训练卷积网络,用来对简笔标注的图像进行语义分割,基于图模型将简笔标注的信息结合空间约束、外观及语义内容,传播到未标记的像素上并学习网络参数. 图 4 是简笔标注与像素级标注对比示意图[56],简笔标注无需仔细勾勒图像边界和形状,只需对每类语义画一条线作为标记,有利于注释没有明确定义形状的物体(例如,天空,草),该方法在 PASCAL VOC 2012[21]数据集上显示出优异的分割结果.

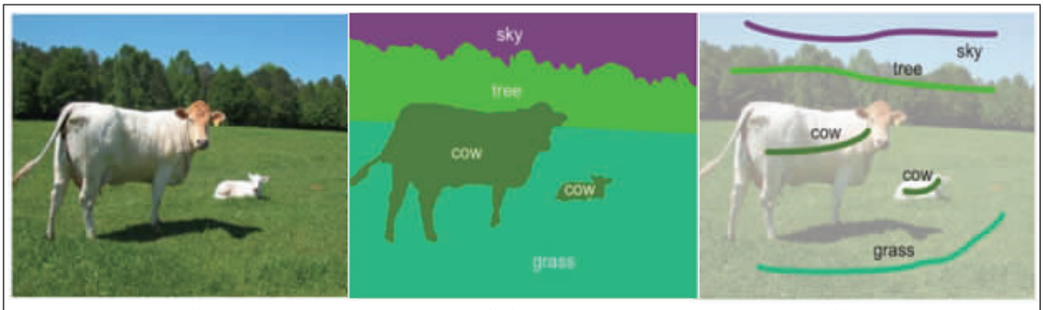
受文献[56]的启发,Vernaza 等人[57]提出一种基于 CNN 的弱监督语义分割模型. 通过传播稀疏图像标签来推断密集标签,由于通过随机行走命中概率来传播稀疏标签会产生不确定性,文中将其与分割预测器联合学习,降低传播标签不正确的可能性,能有效学习语义边缘信息,该方法与先前的分割网络相比具有一定的优势.

由于标准损失函数不能将部分简笔标注得到的种子像素与潜在错误标记的像素区分开,使得训练结果变差. 于是,Tang 等人[58]提出标准 normalized cut"浅"分割损失函数来评估网络输出,它可以评估所有像素间的一致性,达到了与全监督训练相当的效果.

2.3.3　图像级标注

除了使用边界框标注和简笔标注作为弱监督学习之外,还可以利用图像级标注来进行弱监督学习. Pinheiro 等人[44]采用多示例学习模型构建图像标签与像素之间的关联性,首先使用 ImageNet[59]图像级标签对模型进行训练,利用 CNN 生成特征平面,然后将这些特征平面通过聚合层对模型进行约束,得到较好的分割结果. 在此基础上,Pathak 等人[60]提出了一个受约束的 CNN 自训练语义分割模型,利用新的损失函数来加强每个图像标签和预测分割掩膜之间的一致性.

针对图像级监督的语义分割模型易忽视目标位置问题,Wei 等人[61]提出了 STC(simple to complex)语义分割模型,仅利用图像级标签来学习深度卷积神经网络. 首先训练一个名为 Initial-DCNN 的初始分割网络,把它当作是一个有显著性检测功能的 CNN;然后使用 Enhanced-DCNN 来精调每一个物体的分割模块,确定目标位置;最后通过使用 Powerful-DCNN 对复杂的多目标图像进行分割. 实验结果证明了 STC 模型的有效性. 借鉴文献[61]的思想,Jin 等人[62]提出了利用图像标注进行监督的弱监督语义分割模型,首先训练和细化特定类别的浅层神经网络,以获得每个类的分割掩膜,然后将所有类的浅层神经网络组装成一个深度卷积神经网络,最后用于端到端的训练和测试. 该方法在 PASCAL VOC 2012[21]上的分割效果明显优于以前的弱监督语义分割方法.



(*a*)原图像 (*b*)像素级标注 (*c*)简笔标注

图4 简笔标注与像素级标注对比示意图

在利用视频进行弱监督训练方面,Hong 等人[63]提出 Web-Crawled 视频弱监督语义分割模型. 首先,根据一组弱标注图像来训练一个用于目标物体分类和检测的模型;再将这个模型用于对网上抓取的视频进行过滤,得到所需类别的图像帧;然后根据视频中的时空信息使用优化算法得到该图像帧的运动分割结果;最后将这个分割的图像作为语义分割的训练样本. 该模型优于现有的弱监督方法[12],甚至具有与依赖额外注释方法[54]一样的竞争力. 在 Hong 等人[63]的基础上,Ahn 等人[64]提出一种新的 AffinityNet 弱监督语义分割模型,其可以预测相邻图像坐标之间的语义相关性. 它由 CAM(Class Activation Maps),AffinityNet 和分割模型组成,用于生成训练图像的分割标签和进行实际的语义分割. 整个框架仅依赖于图像级标签,该模型比依赖强监督的模型更具竞争力.

3 深度卷积语义分割模型性能对比分析

本节介绍用于语义分割最具代表性的数据集,简要介绍常用的衡量语义分割效果的评价指标,根据该指标对有代表性的深度语义分割模型的性能进行对比和分析.

3.1 语义分割数据集

用于语义分割研究的数据集有 PASCAL VOC 2012[21]、Cityscapes[65]、CamVid[66]、SYNTHIA[67]、Microsoft COCO[68]等.

PASCAL VOC 2012[21]数据集共有 20 个类别,训练数据和验证数据有 11530 张图片,包含 27450 ROI 注释物体和 6929 个分割物体,给出了图像中物体的边界框和类别标签.

Cityscapes[65]数据集由 5000 张精标注图像和 20000 张粗标注图像组成,共有 50 种城市街道场景,主要用于自动驾驶领域,用于评估视觉算法在城区场景语义理解方面的性能.

CamVid[66]数据集是道路和驾驶场景数据集,通过仪表盘上的摄像机采样出 701 帧图像,将它们标注为 32 个类别.

SYNTHIA[67]数据集有 11 个类别,提供细粒度的像素级标注,包含 13407 个渲染视频流训练图像. 反映了场景、动态物体、季节和天气方面的多样性.

Microsoft COCO[68]数据集共包含 80 个类别,包括 82783 张训练图片,40504 张验证图片和 80000 多张测试图片. 测试图片被分为四个相同大小的测试集:20000 张 test-dev 用于额外的验证及调试;20000 张 test-standard 是默认测试数据集,用于与其它方法进行对比;20000 张 test-challenge 是竞赛专用;20000 张 test-reserve 是保留测试数据集.

3.2 性能评价指标

基于像素语义分割的性能评价指标有四种[6]:像素准确率(Pixel Accuracy,PA)、平均准确率(Mean Pixel Accuracy,MPA)、平均交并比(Mean Intersection over Union,MIoU),频率加权交并比(Frequency Weighted Intersection over Union,FWIoU),常使用 MIoU 来衡量语义分割模型的性能.

像素准确率(PA)是语义分割中最简单的像素级评价指标,仅需计算图像中正确分类的像素占图像中总像素数比值,定义如式(1)所示. 其中 p_{ii} 表示正确分类的像素个数,p_{ij} 表示本应属于第 i 类却被分到第 j 类的像素数量,n 是类别数.

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

平均准确率(MPA)表示图像中所有物体类别像素准确率的平均值,定义如式(2)所示.

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

平均交并比(MIoU)是分割结果真值的交集与其并集的比值,按类计算后取平均值,定义如式(3)所示.

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

频率加权交并比(FWIoU)是对 MIoU 改进后的新的评价指标,旨在对每个像素的类别按照其出现的频率进行加权,定义如式(4)所示.

$$\mathrm{FWIoU}=\frac{1}{\sum_{i=0}^{n}\sum_{j=0}^{n}p_{ij}}\sum_{i=0}^{n}\frac{\sum_{j=0}^{n}p_{ij}p_{ii}}{\sum_{j=0}^{n}p_{ij}+\sum_{j=0}^{n}p_{ji}-p_{ii}} \tag{4}$$

3.3　语义分割模型分析与比较

表 1 在经典数据集 PASCAL VOC 2012[21]上对各类代表性语义分割模型进行了性能对比分析,PASCAL VOC 2012 [21]数据集上的模型方法都采用了 CNN 提取特征,没有用到全连接层,所以对于输入图像的大小没有要求,可以是任意大小的图像. 通过对经典语义分割模型方法的 MIOU 比较发现,基于全卷积的语义分割模型的分割效果要最优.

基于候选区域的语义分割方法,有使用特征提取和分类分开学习的,如 Hariharan 等人[26]首先采用 MCG 算法提取候选区域,其次采用 CNN 进行特征提取,然后采用支持向量机对每个候选区域分类,在 PASCAL VOC 2012 上分割精度为 51.6%. 也有区域特征和分类联合进行端到端训练的模型,如 Jiang 等人[31]提出基于区域的端到端人体分割网络模型,通过构造卷积神经网络和反卷积神经网络模型来实现逐像素分割,在 PASCAL VOC 2012 上分割精度为 56.8%. 这类方法在训练时没有考虑相邻像素之间的相关性以及图像空间位置信息缺失问题,分割效果不理想.

相较于基于候选区域的模型方法,基于全卷积的语义分割模型的效果不再依赖于候选区域的生成,直接利用完整的图像信息进行像素分类,获得了更佳的效果,是现在语义分割方法的主流研究方向. 如 Chen 等人[40]采用编-解码模块和 Xception 主干网络,在 PASCAL VOC 2012 上分割精度已经达到了 89.0%,使得大物体分割达到和真实标注在视觉上相同的效果,而对于远距离物体和极小物体也可以准确分割出来.

考虑到前两类模型在训练时需要大量的像素级标注,有一些研究者开始致力于弱监督模型语义分割. 如 Khoreva 等人[55]提出的弱监督语义分割模型只使用了图像类别标签,在 PASCAL VOC 2012 上分割精度最高达到 69.1%,对于复杂场景中的小物体的分割有较好的分割效果. 弱监督语义分割方法的分割精度虽然不及全卷积的像素级标注的分割精度,但是解决了语义分割对大量图像进行像素级标注的高成本问题.

表 1　PASCAL VOC **2012** 数据集上语义分割模型方法的比较

模型类型	文献	实现方法	年份	MIOU(%)
基于候选区域的语义分割模型	[31]	卷积神经网络 + SVM + 非极大值抑制	2014	51.6
	[36]	人体检测网络 + 人体分割网络	2019	56.8
基于全卷积的语义分割模型	[5]	全卷积网络	2015	62.2
	[43]	深度卷积神经网络 + 条件随机场	2015	71.6
	[53]	分割器 + 对抗网络	2016	73.1
	[8]	分割网络 + 辨别网络	2018	74.9
	[44]	扩张卷积 + ASPP + DenseCRF	2017	79.7
	[17]	大卷积核 + 全卷积网络	2017	82.2
	[9]	RefineNet 模型	2016	83.4
	[50]	采用金字塔池化模块将特征图同时通过四个并行的池化层	2017	85.4
	[18]	使用了 ASPP 结构,且不带有级联模块	2017	85.7
	[51]	双流图像分割网络	2017	86.8
	[45]	添加了编码 - 解码模块 + Xception 网络	2018	89.0
基于弱监督的语义分割模型	[67]	利用 Web-Crawled 视频弱监督语义分割方法	2018	58.7
	[61]	稀疏标签传播特定概率模型 + 基于梯度的方法	2017	61.1
	[58]	MCG + 采用卷积网络	2015	62.0
	[16]	采用期望最大化方法	2015	62.2
	[60]	简笔注释 + 图模型	2016	63.1
	[68]	AffinityNet 模型	2018	63.7
	[62]	利用 normalized cut 分割损失函数来评估网络输出	2018	65.1
	[59]	利用递归训练作为除噪策略	2017	69.1

4　结论

本文对基于候选区域的语义分割模型,基于全卷积的语义分割模型和基于弱监督学习的语义分割模型中的经典算法进行了研究和分析,并指出其特点. 尽管这三类模型在实际生活中得到广泛的应用,但是依然面临如下挑战:

第一个是如何构建训练图像中图像标签和像素之间的关联,自动的推断出物体在图像中的位置,进而实现弱监督的物体定位是值得探索的一个方向;第二个是如何探索出新的网络结构以及不同网络模型之间的融合学习方法也是值得研究的一个方向;第三个是如何准确得到小物体的位置信息以及上下文信息是语义分割亟待解决的问题.

参考文献

[1] Gould S,Fulton R,Koller D. Decomposing a scene into geometric and semantically consistent regions[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Kyoto,Japan:IEEE,2009. 1 – 8.

[2] Gupta A,Efros A A,Hebert M. Blocks world revisited:image understanding using qualitative geometry and mechanics[A]. Proceedings of the European Conference on Computer Vision[C]. Crete,Greece:ACM,2010. 482 – 496.

[3] Yu S X,Zhang H,Malik J. Inferring spatial layout from a single image via depth-ordered grouping[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA:IEEE,2008. 1 – 7.

[4] Lee D C,Hebert M,Kanade T. Geometric reasoning for single image structure recovery[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA:IEEE,2009. 2136 – 2143.

[5] 李亚峰. 一种基于多字典学习的图像分割模糊方法[J]. 电子学报,2018,46(7):1700 – 1709.
Li Ya-feng. An image segmentation fuzzy method based on multi-dictionary learning[J]. Acta Electronica Sinica,2018,46(7):1700 – 1709. (in Chinese)

[6] Shelhamer E,Long J,Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2015,39(4):640 – 651.

[7] Hung W C,Tsai Y H,Liou Y T,et al. Adversarial learning for semi-supervised semantic segmentation[J]. CORR,2018,57(8):7540 – 7551.

[8] Souly N,Spampinato C,Shah M. Semi supervised semantic segmentation using generative adversarial network[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. New York,USA:IEEE,2017. 5689 – 5697.

[9] Ghiasi G,Fowlkes C C. Laplacian Pyramid reconstruction and refinement for semantic segmentation[A]. Proceedings of the European Conference on Computer Vision[C]. USA:IEEE,2016. 519 – 534.

[10] Lin G,Shen C,Hengel A V D,et al. Efficient piecewise training of deep structured models for semantic segmentation[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 3194 – 3203.

[11] Liu Z,Li X,Luo P,et al. Semantic image segmentation via deep parsing network[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Santiago,Chile:IEEE,2015. 1377 – 1385.

[12] Papandreou G,Chen L C,Murphy K P,et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Santiago,Chile:IEEE,2015. 1742 – 1750.

[13] Peng C,Zhang X,Yu G,et al. Large kernel matters — improve semantic segmentation by global convolutional network[J]. IEEE Conference on Computer Vision and Pattern Recognition,2017,189(1):1743 – 1751.

[14] Ning Q,Zhu J,Chen C. Very fast semantic image segmentation using hierarchical dilation and feature refining[J]. Cognitive Computation,2017,10(2):1 – 11.

[15] Jiang Z,Yuan Y,Wang Q. Contour-aware network for semantic segmentation via adaptive depth[J]. Neurocomputing,2018,284(1):27 – 35.

[16] Zhu H,Mng F,Cai J,et al. Beyond pixels:A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation[J]. Journal of Visual Communication & Image Representation,2015,34(2):12 – 27.

[17] Thoma M. A Survey of semantic segmentation[J]. CORR,2016,65(41):1 – 15.

[18] Guo Y,Liu Y,Georgiou T,et al. A review of semantic segmentation using deep neural networks[J]. International Journal of Multimedia Information Retrieval,2017,7(2):87 – 93.

[19] Garcia-garcia A,Orts-Escolano S,Oprea S,et al. A review on deep learning techniques applied to semantic segmentation[J]. CORR,2017,75(9):41 – 65.

[20] Geng Q,Zhou Z,Cao X. Survey of recent progress in semantic image segmentation with CNNs[J]. Science China (Information Sciences),2018,61(5):1101 – 1118.

[21] Shetty S. Application of convolutional neural network for image classification on pascal voc challenge 2012 dataset[J]. CORR,2016,21(3):1 – 6.

[22] Carreira J,Sminchisescu C. CPMC:Automatic object segmentation using constrained parametric min-cuts[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2012,34(7):1312 – 1328.

[23] Carreira J,Rui C,Batista J,et al. Semantic segmentation with second-order pooling[J]. European Conference on Computer Vision,2012,7578(1):430 – 443.

[24] Girshick R,Donahue J,Darrell T,et al. Rich feature hierarchies for accurate object detection and semanticsegmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,

2014. 580 – 587.
[25] Uijlings J R R, Van DE Sande K E A, Gevers T, et al. Selective search for object recognition[J]. IJCV, 2013, 104(2): 154 – 171.
[26] Hariharan B, Arbel EZ P, Girshick R, et al. Simultaneous detection and segmentation[A]. Proceedings of the European Conference on Computer Vision[C]. USA: IEEE, 2014. 297 – 312.
[27] Arbel EZ P, Pont-Tuset J, Barron J, et al. Multiscale combinatorial grouping[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2014. 328 – 335.
[28] Ren S, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137 – 1149.
[29] Caesar H, Uijlings J, Ferrari V. Region-based semantic segmentation with end-to-end training[A]. Proceedings of the European Conference on Computer Vision[C]. USA: IEEE, 2016. 381 – 397.
[30] Girshick R. Fast R-CNN[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Santiago, Chile: IEEE, 2015. 1 – 9.
[31] X J, Y G, Z F, et al. An end-to-end human segmentation by region proposed fully convolutional network[A]. IEEE Access[C]. USA: IEEE, 2019. 16395 – 16405.
[32] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Santiago, Chile: IEEE, 2015. 1520 – 1528.
[33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[A]. International Conference on Learing Representation[C]. USA: IEEE, 2014. 1 – 14
[34] 李宝奇,贺昱曜,何灵蛟,等. 基于全卷积神经网络的非对称并行语义分割模型[J]. 电子学报, 2019, 47(5): 1058 – 1064.
Li Bao-qi, HE Yu-yao, HE ling-jiao, et al. Asymmetric parallel semantic segmentation model based on full convolutional neural network[J]. Acta Electronica Sinica, 2019, 47(5): 1058 – 1064. (in Chinese)
[35] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(1): 2481 – 2495.
[36] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[J]. Medical Image Computing and Computer-Assisted Intervention, 2015, 56(9): 234 – 241.
[37] MIilletar F, Navab N, Ahmadi S A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation[A]. Proceedings of the International Conference on 3d Vision[C]. USA: IEEE, 2016. 565 – 571.
[38] 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.
[39] 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 & Machine Intelligence, 2017, 40(4): 834 – 848.
[40] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[A]. Proceedings of the European Conference on Computer Vision[C]. Germany: IEEE, 2018. 833 – 851.
[41] Kr H H P, Koltun V. Efficient inference in fully connected crfs with Gaussian edge potentials[J]. CORR, 2012, 34(2): 1 – 9.
[42] Chandra S, Kokkinos I. Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian crfs[A]. Proceedings of the European Conference on Computer Vision[C]. USA: IEEE, 2016. 402 – 418.
[43] Pohlen T, Hermans A, Mathias M, et al. Full-fesolution residual networks for semantic segmentation in street scenes[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 3309 – 3318.
[44] Pinheiro P O, Collobert R. From image-level to pixel-level labeling with convolutional networks[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA: IEEE, 2015. 1713 – 1721.
[45] Lin G, Milan A, Shen C, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation[A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 5168 – 5177.
[46] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 6230 – 6239.
[47] Luo P, Wang G, Lin L, et al. Deep dual learning for semantic image segmentation[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice, Italy: IEEE, 2017. 2737 – 2745.
[48] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[A]. Proceedings of the International Conference on Neural Information Processing Systems[C]. USA: ACM, 2014. 2672 – 2680.
[49] Luc P, Couprie C, Chintala S, et al. Semantic segmentation using adversarial networks[J]. CORR, 2016, 56(7): 1 – 12.

[50] Metzen J H, Kumar M C, Brox T, et al. Universal adversarial perturbations against semantic image segmentation[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Venice, Italy: IEEE, 2017. 2774 – 2783.
[51] Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. USA: IEEE, 2017. 1378 – 1387.
[52] Zhu W, Xiang X, Tran T D, et al. Adversarial deep structural networks for mammographic mass segmentation[J]. CORR, 2017, 43(6): 1 – 9.
[53] Rezaei M, Harmuth K, Gierke W, et al. A conditional adversarial network for semantic segmentation of brain tumor [A]. Proceedings of the International MICCAI Brainlesion Workshop[C]. USA: IEEE, 2017. 241 – 252.
[54] Dai J, He K, Sun J. BoxSup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation[A]. IEEE International Conference on Computer Vision[C]. USA: IEEE, 2015. 1635 – 1643.
[55] Khoreva A, Benenson R, Hosang J, et al. Simple does it: weakly supervised instance and semantic segmentation [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. USA: IEEE, 2017. 1665 – 1674.
[56] Lin D, Dai J, Jia J, et al. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation [A]. Proceedings of the Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 3159 – 3167.
[57] Vernaza P, Chandraker M. Learning random-walk label propagation for weakly-supervised semantic segmentation [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. USA: IEEE, 2017. 2953 – 2961.
[58] Tang M, Djelouah A, Perazzi F, et al. Normalized cut loss for weakly-supervised cnn segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 1818 – 1827.
[59] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211 – 252.
[60] Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation [A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Santiago, Chile: IEEE, 2015. 1796 – 1804.
[61] Wei Y, Liang X, Chen Y, et al. STC: A simple to complex framework for weakly-supervised semantic segmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 39(11): 2314 – 2320.
[62] Jin B, Segovia M V O, Susstrunk S. Webly supervised semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2017. 1705 – 1714.
[63] Hong S, Yeo D, Kwak S, et al. Weakly supervised semantic segmentation using web-crawled videos[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 2224 – 2232.
[64] Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. USA: IEEE, 2018. 4981 – 4990.
[65] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2015. 3213 – 3223.
[66] 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.
[67] Ros G, Sellart L, Materzynska J, et al. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes[A]. Proceedings of the Computer Vision and Pattern Recognition [C]. USA: IEEE, 2016. 3234 – 3243.
[68] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[A]. Proceedings of the European Conference on Computer Vision [C]. USA: IEEE, 2014. 740 – 755.

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

张 云 女, 1992 年 1 月生于河南信阳. 2016 年进入江西理工大学, 在读硕士研究生. 研究方向为语义分割.
E-mail: 1040344705@qq.com