

基于深度学习的图像语义分割方法综述*

田萱, 王亮, 丁琪

(北京林业大学 信息学院, 北京 100083)

通讯作者: 田萱, E-mail: tianxuan@bjfu.edu.cn



摘要: 近年来,深度学习技术已经广泛应用到图像语义分割领域.主要对基于深度学习的图像语义分割的经典方法与研究现状进行分类、梳理和总结.根据分割特点和处理粒度的不同,将基于深度学习的图像语义分割方法分为基于区域分类的图像语义分割方法和基于像素分类的图像语义分割方法.把基于像素分类的图像语义分割方法进一步细分为全监督学习图像语义分割方法和弱监督学习图像语义分割方法.对每类方法的代表性算法进行了分析介绍,并详细总结了每类方法的基本思想和优缺点,系统地阐述了深度学习对图像语义分割领域的贡献.对图像语义分割相关实验进行了分析对比,并介绍了图像语义分割实验中常用公共数据集和性能评价指标.最后,预测并分析总结了该领域未来可能的研究方向及相应的发展趋势.

关键词: 图像语义分割;深度学习;像素分类;全监督学习;弱监督学习

中图法分类号: TP391

中文引用格式: 田萱,王亮,丁琪.基于深度学习的图像语义分割方法综述.软件学报,2019,30(2):440–468. <http://www.jos.org.cn/1000-9825/5659.htm>

英文引用格式: Tian X, Wang L, Ding Q. Review of image semantic segmentation based on deep learning. Ruan Jian Xue Bao/ Journal of Software, 2019, 30(2): 440–468 (in Chinese). <http://www.jos.org.cn/1000-9825/5659.htm>

Review of Image Semantic Segmentation Based on Deep Learning

TIAN Xuan, WANG Liang, DING Qi

(School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)

Abstract: Recent years, applying Deep Learning (DL) into Image Semantic Segmentation (ISS) has been widely used due to its state-of-the-art performances and high-quality results. This paper systematically reviews the contribution of DL to the field of ISS. Different methods of ISS based on DL (ISSbDL) are summarized. These methods are divided into ISS based on the Regional Classification (ISSbRC) and ISS based on the Pixel Classification (ISSbPC) according to the image segmentation characteristics and segmentation granularity. Then, the methods of ISSbPC are surveyed from two points of view: ISS based on Fully Supervised Learning (ISSbFSL) and ISS based on Weakly Supervised Learning (ISSbWSL). The representative algorithms of each method are introduced and analyzed, as well as the basic workflow, framework, advantages and disadvantages of these methods are detailedly analyzed and compared. In addition, the related experiments of ISS are analyzed and summarized, and the common data sets and performance evaluation indexes in ISS experiments are introduced. Finally, possible research directions and trends are given and analyzed.

Key words: image semantic segmentation; deep learning; pixel classification; fully supervised learning; weakly supervised learning

当前,计算机视觉领域的核心研究包括图像分类、物体检测和图像语义分割(image semantic segmentation, 简称 ISS)等,其中,ISS 是一门涉及计算机视觉、模式识别与人工智能等研究领域的交叉学科,是数字图像处理与机器视觉的研究重点.ISS 在虚拟现实、工业自动化、视频检测等不同领域有广泛的应用,具有重要的研究意义

* 基金项目: 中央高校基本科研业务费专项资金(TD2014-02)

Foundation item: Fundamental Research Funds for the Central Universities (TD2014-02)

收稿时间: 2018-01-24; 修改时间: 2018-03-26, 2018-05-28; 采用时间: 2018-08-30

和应用价值.ISS 由 Ohta 等人首次提出,其定义是:为图像中的每一个像素分配一个预先定义好的表示其语义类别的标签^[1].与传统的图像分割相比,ISS 在其基础上为图像中的目标或前景加上一定的语义信息,能够根据图像本身的纹理、场景和其他高层语义特征来得出图像本身需要表达的信息,更具实用价值.近年来,国内外众多科研机构相继开展了针对该领域的学术研究,人工智能、模式识别以及计算机视觉方面的国内外学术会议都会对该领域和相关研究成果做重点讨论.这些机构和组织有效地推动了 ISS 技术的发展.

近年来,深度学习(deep learning,简称 DL)技术^[2]迅猛发展,基于深度学习的图像语义分割方法(image semantic segmentation based on deep learning,简称 ISSbDL)也日新月异.鉴于目前国内还没有全面细致论述 ISSbDL 方法的综述文献^[3,4],我们总结并整理了相关研究后得到本文.如图 1 所示,按照 ISS 的方法特点和处理粒度,将 ISSbDL 方法分为基于区域分类的图像语义分割方法(ISS based on the regional classification,简称 ISSbRC)和基于像素分类的图像语义分割方法(ISS based on the pixel classification,简称 ISSbPC),对每类方法按照处理特点又细分为若干种不同的子方法.

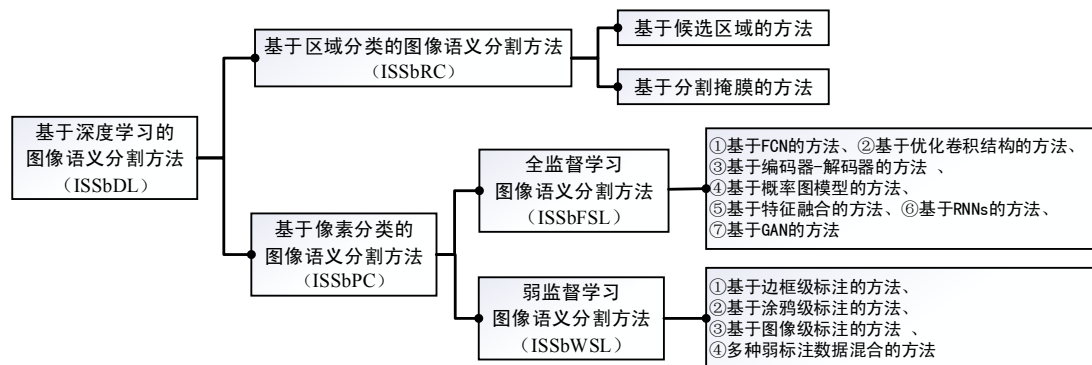


Fig.1 A taxonomy of ISSbDL

图 1 基于深度学习的图像语义分割方法分类

本文第 1 节介绍 DL 与 ISS 的相关背景及 ISSbDL 的早期研究.第 2 节对 ISSbRC 方法进行详细介绍和总结.第 3 节对 ISSbPC 方法进行分析 and 总结,并根据处理特点进一步分类,详细介绍每类子方法的基本思想和优缺点.第 4 节对 ISS 相关实验进行分析与比较,并介绍常用公共数据集和算法性能评价指标.第 5 节总结讨论 ISS 未来的发展方向和发展趋势.

1 相关背景及早期研究介绍

DL 的概念由 Hinton 等人^[2]在 2006 年首次提出,是机器学习中一种基于对数据进行表征学习的方法.DL 技术能够有效地提取图像中的低级、中级和高级语义信息,并结合分类器辅助进行像素分类,提高了 ISS 方法的分割准确率.目前,主流的 DL 模型有卷积神经网络(convolutional neural network,简称 CNN)^[5]、循环神经网络(recurrent neural network,简称 RNN)^[6]和生成对抗网络(generative adversarial network,简称 GAN)^[7]等.

其中,CNN 的基本结构由输入层、卷积层、池化层、全连接层及输出层组成.输入图像经过 CNN 多个卷积操作和池化操作进行特征提取,将低层粗糙特征逐步转变为高层精细特征,高层特征再经过全连接层和输出层后进行分类.CNN 由于其特殊的网络结构,特别适合处理图像数据,对平移、伸缩、倾斜等图像形变具有较高的不变性.RNN 由一连串重复的神经网络模块序列组成,序列中的每个元素都执行相似的任务,图像上下文(image context)之间的连续信息可合理利用.当前,具有代表性的 RNN 包括传统的 RNN 序列模型、长短期记忆神经网络(long short-term memory,简称 LSTM)^[8]以及门控递归单元(gated recurrent unit,简称 GRU)^[9].GAN 由一个生成器网络(generator network)和一个判别器网络(discriminator network)组成,其基本思想是:从训练库中获取大量训练样本进行学习,生成器网络不断产生人造样本,判别器网络不断对人造样本进行判断,训练时,让两组网络

相互对抗、相互提高.

在 ISS 领域,CNN 因其高效的学习性能和良好的应用效果,受到研究者更多的青睐,成为研究热点.除 CNN 外,RNN 因其递归处理历史信息 and 建模历史记忆的特点,特别适合处理与时空序列有关的信息,也常被一些研究者用于捕获图像上下文信息.此外,GAN 模型避免了一些传统生成模型在实际应用中的困难,具有新颖性和良好的适应性,在 ISS 研究中也正逐渐得到重视.总的来说,DL 技术利用深层模型来学习图像特征,促进了 ISS 领域中相关研究的发展,掀起了一股 ISSbDL 的研究热潮.

2013 年,文献[10]尝试使用 DL 技术对室内场景进行语义分割:利用 CNN 对 RGB-D 图像进行特征提取的同时,将 RGB 图像聚类得到超像素,然后使用分类器对超像素进行分类,完成 ISS 任务.文献[11]则在上述工作的基础上,利用深层 CNN 提取、整合不同分辨率图像的特征,并使用分割树对粗糙图像块中的超像素进行平滑预测处理.这些早期的 ISSbDL 方法^[10,11]出现在 ISS 由传统的机器学习方式向深度学习方式过渡的阶段,先使用聚类操作生成超像素,再用 CNN 等分类器对超像素进行分类.图像分割过程分阶段进行,耗时、费力,分割过程无法有效利用图像的全局语义特征,分割结果比较粗糙.

随着 DL 技术的发展,研究者又进一步改进,先将图像划分为一系列目标候选区域,再用 DL 技术对目标区域进行分类,避免生成超像素,提高了分割效率,本文第 2 节所介绍的基于区域分类的图像语义分割方法即为此类改进.另一些研究者则直接利用深度神经网络(deep neural network,简称 DNN)以像素分类的方式进行分割,将分割过程改为端到端(end-to-end)的模式,避免了预先生成图像块所带来的问题,提高了分割准确率,本文将在第 3 节中介绍这类基于像素分类的图像语义分割方法.基于这些改进,我们将 ISSbDL 的处理流程概括为特征提取、语义分割和后期处理这 3 个核心步骤,如图 2 所示,其中,实线表示一般处理步骤,虚线表示选择使用环节.

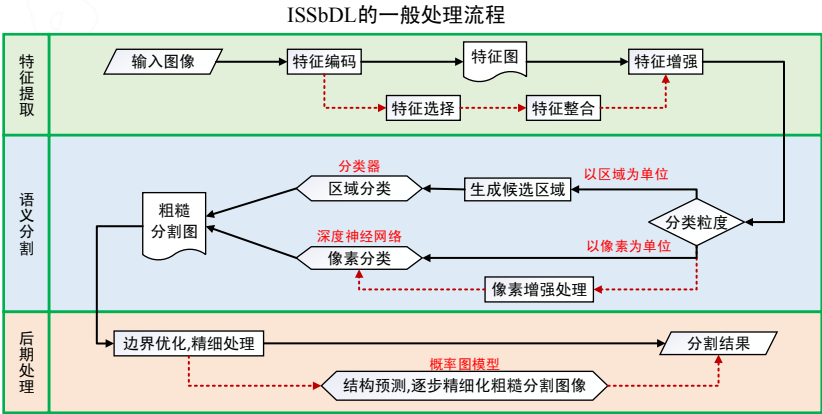


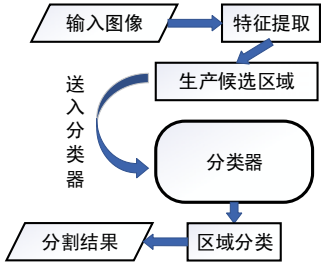
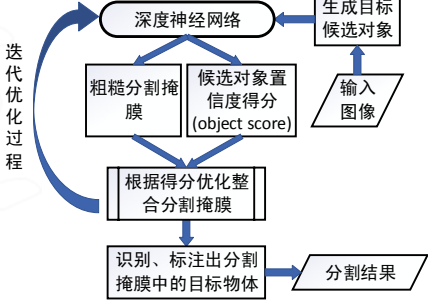
Fig.2 Workflow for ISSbDL

图 2 基于深度学习的图像语义分割方法的一般处理流程

2 基于区域分类的图像语义分割方法

在 ISSbDL 方法中,基于区域分类的图像语义分割方法把传统图像处理算法与 DNN 相结合,先将原始图像划分成不同的目标候选区域,得到一系列图像块(image patch),再利用 DNN 对图像块或图像块中的每个像素进行语义分类,最后根据分类结果对原始图像进行标注,得到最终分割结果.因为图像块的质量直接决定分割结果的好坏,ISSbRC 方法的关键在于如何从原始图像产生不同目标区域的图像块.根据区域生成算法和图像块划分标准不同,下面将 ISSbRC 方法划分为两类:基于候选区域的方法和基于分割掩膜的方法.表 1 对这两类方法从方法特点、优缺点和处理流程等几个方面进行了分析与比较.下面对其进行详细介绍.

Table 1 Comparison of ISSbRC
表 1 基于区域分类的图像语义分割方法对比

| 方法类别 | 代表算法 | 方法特点 | 优缺点总结 | 处理流程 |
|-----------|--|--|---|---|
| 基于候选区域的方法 | RCNN ^[12] , SDS ^[13] , MPA ^[14] , Mask-RCNN ^[15] | (1) 利用区域生产算法得到一系列候选区域,每个候选区域中都有可能包含潜在的目标物体,改进的区域产生算法有 SelectiveSearch、Fast-RCNN、Faster-RCNN | 优点:使用物体检测技术产生候选区域,能够同时完成物体检测与图像语义分割两项任务 |  |
| | | (2) 使用分类器对经过特征提取后的图像区域进行分类,以区域分类的方式进行像素分类,主要分类器有:SVM、VGG-16、ResNet | 缺点:没有充分考虑图像中的全局语义信息,分类图像中的小尺度物体和小面积区域时易出错 | |
| 基于分割掩膜的方法 | DeepMask ^[16] , SharpMask ^[17] , MultiPath ^[18] | (1) 通过物体检测方法识别出图像中潜在的目标候选对象,产生目标候选对象的技术主要有:RCNN、DeepMask、MultiPath模型 (2) 目标候选区域中的像素经过二分类处理后得到分割掩膜,通过对多张分割掩膜进行优化处理得到分割结果 | 优点:利用RCNN等物体检测技术生成分割掩膜,使用精炼模块对粗糙分割掩膜进行优化,可挖掘多种尺寸、背景图片中的隐含信息 缺点:针对小尺寸物体、被遮挡物体以及背景复杂物体的准确率较低 |  |

2.1 基于候选区域的方法

该类方法首先利用相应的算法生成大量候选区域并筛选出合理的候选区域,再运用 CNN 对每个候选区域提取图像特征和语义信息,接着利用分类器对候选区域中的图像块或像素进行分类,最后输出分割结果.因为每个候选区域都有可能包含图像中潜在的目标物体,候选区域的质量不但影响 CNN 捕获图像特征的能力,而且影响分类器对候选区域进行分类的精度.

2014 年,文献[12]在 CNN 的基础上提出了区域卷积神经网络(regions with CNN features,简称 RCNN). RCNN 将选择搜索(selective search,简称 SS)算法产生的候选区域与 CNN 产生的视觉特征相结合,可同时完成目标检测和 ISS 两项任务.RCNN 的处理流程如图 3 所示,首先,使用 SS 算法抽取约 2 000 个候选区域;再用 CNN 提取每个候选区域的特征;最后,根据捕获的特征,使用 SVM(support vector machine)对每个候选区域中的目标物体分类.但 RCNN 也存在严重依赖候选区域、产生图像变形、分割精确度不够高和速度不够快等缺点,其综合性能仍有待提高.

文献[13]在 RCNN 的基础上提出了 SDS(simultaneous detection and segmentation)方法.与 RCNN 方法相比,SDS 方法使用 MCG(multi-scale combinatorial grouping)^[19]算法在 CNN 中独立地从候选区域和区域前景中提取特征,再将这两部分特征进行联合训练,然后使用非极大值抑制(non-maximum suppression,简称 NMS)算法进行区域增强,分割性能有显著提升.

由于 RCNN 存在着生成的候选区域数量过多、网络运算量大且候选区域形状不规则等局限性,一些研究

者开始探索产生高质量候选区域的方法.文献[20]中的 SPPNet 网络将空间金字塔池化层(spatial pyramid pooling player,简称 SPP player)插入到 RCNN 卷积层的后面,减少了特征提取过程中的重复计算.文献[21]中的 Fast-RCNN 网络将候选区域映射到 CNN 的卷积特征图上,通过 ROI Pooling 层,将每个候选区域生成固定尺寸的特征图,提升了生成候选区域的速度.文献[22]中的 Faster-RCNN 网络在 Fast-RCNN 网络的基础上加入区域建议网络(region proposal network,简称 RPN),能够快速生成高质量的候选区域.

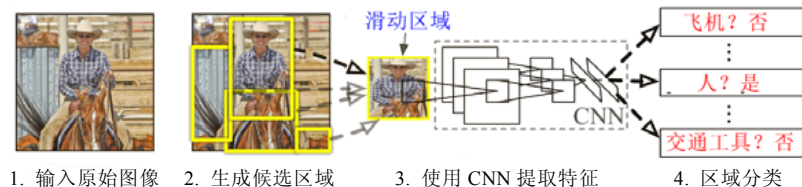


Fig.3 Workflow for RCNN^[12]

图 3 RCNN 的处理流程^[12]

2016 年,文献[14]以 SDS 方法为基础,使用不同大小的滑动窗口对原图进行卷积、池化等操作,得到多尺度特征图,再通过尺度对齐将不同尺度的特征图归一化到同样大小,并将其做并行处理,提出了能够完成定位、分类和分割这 3 个任务的 MPA(multi-scale path aggregation)方法^[14].MPA 方法通过组合不同尺度的特征图,能够综合图像中多个不同部位的局部信息,有效地避免了物体理解的歧义,使分割结果更具鲁棒性.

2017 年,文献[15]在 Faster-RCNN 基础网络中加入 ROI Align 层和分割子网,提出了能够实现目标检测和实例级图像语义分割(即实例分割)两个任务的 Mask-RCNN 网络.Mask-RCNN 由两个分支子网组成:第 1 个分支子网继承自 Faster-RCNN,用于对候选区域进行分类和回归,可有效地检测出图像中的目标物体;第 2 个分支子网使用一个微型全卷积网络进行高质量的实例分割.Mask-RCNN 能够完成分类、回归和分割这 3 项任务,并保留目标对象的空间结构信息,促进了 ISS 的发展.

2.2 基于分割掩膜的方法

基于分割掩膜(segmentation mask)的方法重点关注如何有效生成与目标候选区域相对应的分割掩膜,大致可分为两个核心步骤.

- (1) 首先,在图像中检测出所有潜在的目标候选对象,将原图像划分为一系列大小不等的图像块,每个图像块包含一个潜在的目标候选对象.
- (2) 将产生的图像块送入 CNN 进行处理,其每个像素经过“是否属于该目标候选对象”的二分类判断后得到分割掩膜,再将多张分割掩膜进行优化处理得到最终分割结果.

文献[16]基于 CNN 提出用于生成目标建议(object proposal)的 DeepMask 模型.该模型直接从原始图像中生成与目标候选对象对应的图像块,再根据这些图像块为每个目标候选对象生成分割掩膜.DeepMask 模型使用前馈网络在整张图像中生成目标对象的分割掩膜,所产生的掩膜边界粗糙,不能精准对齐.文献[17]在 DeepMask 的基础上将浅层网络中的低级特征与深层网络中的高级特征相结合,并且自下而上进行图像处理,提出了 SharpMask 模型.SharpMask 模型先通过 DeepMask 模型对每个输入产生一个粗糙的分割掩膜,再将该粗糙分割掩膜传入不同的精炼模块,然后结合不同尺寸的特征图,以自下而上的方式重新生成精细化的分割掩膜.文献[18]以 Fast-RCNN 为基础,提出了 MultiPath 模型.该模型使用跳跃连接(skip connection)、中心凹区域(foveal region)和一个积分损失函数(integral loss function)对分割掩膜中的对象再次识别并分类,能够解决物体检测过程中尺度、遮挡和集群等问题,分割掩膜生成过程中的定位精度有显著提升.

3 基于像素分类的图像语义分割方法

ISSbRC 方法取得了一定的分割效果,但也存在图像分割精度不高和分割速度不够快等问题,因此,一些研

研究者提出直接在像素级别上进行图像语义分割,产生了基于像素分类的图像语义分割方法.ISSbPC 方法利用 DNN 从带有大量标注的图像数据中提取出图像特征和语义信息,再根据这些信息来学习、推理原始图像中像素的类别,通过端到端训练的方式对每个像素进行分类,以像素分类的方式达到语义分割的目标.ISSbPC 方法无需产生目标候选区域,直接为图像中的每个像素进行分类,原始图像经过一个端到端模型后直接输出分割结果,是一种从训练数据出发,贯穿整个模型后直接输出结果的新模式.ISSbPC 方法将原始图像、标注图像以及弱标注(weak label)图像等海量数据作为训练样本,可以捕获更丰富的图像特征,不仅增加了模型的整体契合度,而且提高了学习效率,有效提升了分割准确率。

根据标注类型和学习方式不同,我们将 ISSbPC 方法主要分为两类:全监督学习图像语义分割方法(ISS based on fully supervised learning,简称 ISSbFSL)和弱监督学习图像语义分割方法(ISS based on weakly supervised learning,简称 ISSbWSL).ISSbFSL 方法使用经过人工精确加工的像素级标注作为训练样本,其分割流程为:先对图像中的每个像素预先给定一个语义标签得到标注数据,然后利用标注数据对 DNN 进行训练,再将训练好的 DNN 用于图像语义分割.ISSbWSL 方法则使用弱标注数据作为样本对 DNN 进行训练,再用训练后的 DNN 对图像进行语义分割.这两类方法按照改进特点不同又可分为若干子类方法,其分类示意如图 4 所示.下面进行具体介绍和分析。

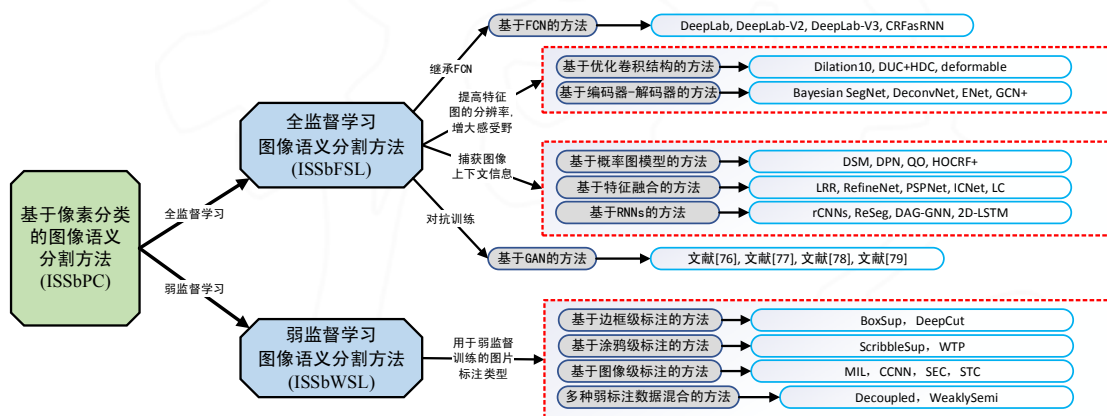


图4 基于像素分类的图像语义分割方法分类

3.1 全监督学习图像语义分割方法

经过人工精确标注的图像样本能够提供大量细节信息和局部特征,有利于提高网络训练效率和分割精确度,因此,目前的 ISSbDL 方法大多是全监督学习类型.ISSbFSL 方法在训练阶段会利用 DNN 从大量带有像素级标注的图像中提取丰富的视觉特征和语义信息,再利用这些特征和信息对图像像素进行分类。

针对早期 ISSbRC 方法存在的存储开销大、计算效率低等问题,Long 等人^[23]于 2014 年设计了一种兼容任意尺寸图像、以全监督学习方式进行图像语义分割的全卷积网络(fully convolutional network,简称 FCN)。如图 5 所示,FCN 在 VGG-16 网络^[24]基础上进行改进,使用卷积层替换传统 CNN 中的全连接层,使用跨层(skip layer)方法组合中间卷积层产生的特征图,再通过双线性插值(bilinear interpolation,简称 BI)算法进行上采样(upSample),将粗糙的分割结果转换为细密的分割结果.FCN 采用跨层方法,既同时兼顾全局语义信息和局部位置信息,又能从抽象特征中恢复出像素所属的类别,把图像级别的分类进一步延伸到了像素级别的分类,成功地将原本用于图像分类的网络转变为用于图像分割的网络。

FCN 在分割过程中能够恢复像素所属的类别,极大地推动了 ISS 的发展.然而该领域仍然存在两个问题:一是图像经过池化操作后,特征图的分辨率不断降低,部分像素的空间位置信息丢失;二是分割过程未能有效地考虑图像上下文(image context)信息,无法充分利用丰富的空间位置信息,导致局部特征和全局特征的利用率失

衡.FCN 未能有效地解决这两个问题,致使分割结果粗糙、分割边界不连续.针对这两个问题,在 FCN 的基础上,研究者又提出了一系列新方法,根据这些方法的改进特点不同,我们将其划分为 7 类:基于 FCN 的方法、基于优化卷积结构的方法、基于编码器-解码器的方法、基于概率图模型的方法、基于特征融合的方法、基于 RNN 的方法和基于 GAN 的方法.表 2 从方法特点、优缺点、关键技术和主要功能这几个方面对这 7 种方法进行了归纳总结.下面将详细介绍这 7 类方法.

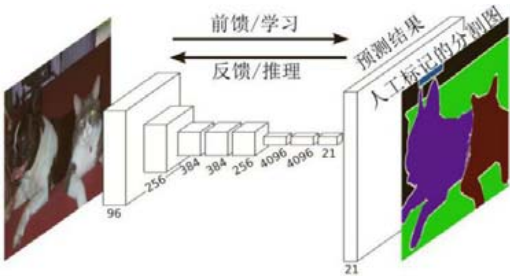


Fig.5 Framework for FCN^[23]

图 5 FCN 框架结构^[23]

Table 2 Comparison of fully-supervised image semantic segmentation algorithm based on pixel classification

表 2 基于像素分类的全监督学习图像语义分割方法对比

| 分类 | 代表算法 | 优缺点总结 | 方法特点 | 关键技术及与其对应的主要功能介绍 | |
|-------------|--|--|---|---|--------------------------------|
| | | | | 关键技术 | 主要功能 |
| 基于 FCN 的方法 | DeepLab ^[25] , DeepLab-V2 ^[26] , DeepLab-V3 ^[27] , CRFasRNN ^[28] | 优点:针对 FCN 的不足进行改进,可有效增强滤波器的视野、获取图像的多尺度表示,提高分割结果的空间精度 | 通过图像金字塔、带孔卷积、带孔空间金字塔池化等技术获得多尺度图像信息,对空间变换具有较高的不变性.对 FCN 进行了优化改进,可提取稠密的图像特征和增大感受野,并使用条件随机场进行结构预测 | 带孔卷积 ^[25] | 增大感受野,减少特征图分辨率降低的速度 |
| | | 缺点:分割速度慢,对小尺度物体的分割效果不明显 | | ASPP ^[25] , CRF ^[25] | 捕获图像上下文信息 |
| | | | | CRFasRNN ^[28] | 将 CRF 建模为 RNN 序列 |
| 基于优化卷积结构的方法 | Dilation10 ^[29] , DUC+HDC ^[30] , deformable ^[31] | 优点:增大感受野,有效减缓特征图分辨率降低的速度,保存像素的空间位置信息 | 使用经过优化的卷积结构替换传统的卷积、池化等操作,在网络中适当舍弃了 pooling 层 | 扩张卷积 ^[29] , 混合扩张卷积 ^[30] , 可变形卷积 ^[31] | 优化卷积结构,增大特征图的感受野,减少特征图分辨率降低的速度 |
| | | 缺点:像素局部信息的连续性被打断,对未知形变的适应性较差 | | 稠密上采样卷积 ^[30] | 代替 BI 算法进行上采样操作 |
| 基于编码解码器的方法 | BayesianSegNet ^[32] , DeconvNet ^[33] , ENet ^[34] , GCN+ ^[35] | 优点:还原图像的空间维度和像素的位置信息,避免池化操作后特征图分辨率降低的问题 | 通过反卷积或上池化等操作构成的解码器对低分辨率特征图进行上采样处理,其相关网络结构有:UnPooling layers, Deconvolution 和 Global Convolution Network | Deconvolution ^[32] , UnPooling ^[32] , GCN ^[35] | 特征解码,增大特征图的感受野,减少特征图分辨率降低的速度 |
| | | 缺点:与 FCN 相比,其网络训参数过多、计算量较大;大部分算法的分割速度无法满足实时处理 | | 去除池化层,填零操作 | 上采样策略 |

Table 2 Comparison of fully-supervised image semantic segmentation algorithm based on pixel classification (Continued)

表 2 基于像素分类的全监督学习图像语义分割方法对比(续)

| 分类 | 代表算法 | 优缺点总结 | 方法特点 | 关键技术及其对应的 主要功能介绍 | |
|------------|---|--|---|---|----------------------------|
| | | | | 关键技术 | 主要功能 |
| 基于概率图模型的方法 | DSM ^[36] , C&G ^[37] , DPN ^[38] ,QO ^[39] , HOCRF+ ^[40] | 优点:捕获上下文信息,并充分利用不同位置的局部特征 | 使用概率图模型来对粗糙分割图像进行后期处理,以结构化预测的方式优化粗糙分割物体的边界,逐步提高分割精度 | CRF, MRF ^[38] | 捕捉图像上下文信息,精细化分割结果 |
| | | 缺点:分割速度慢,计算量大、训练时间长、严重消耗内存 | | GCRF ^[41] , Higher Order Potentials ^[42] | 捕捉图像上下文信息,学习潜在的目标对象 |
| 基于特征融合的方法 | LRR ^[43] , RefineNet ^[44] , PSPNet ^[45] , ICNet ^[46] ,LC ^[47] | 优点:捕获上下文信息,避免使用概率图模型带来的计算量大、消耗内存等问题 | 通过跨层结构、空间金字塔池化模块、多尺度卷积和级联模型等方法,捕获图像中不同尺度、不同位置的特征信息并将其融合,逐步精细化分割结果 | Skip layer ^[43] , ASPP ^[45] , laplacian pyramid ^[43] , attention model ^[48] | 进行特征融合,捕获图像上下文信息,逐步精细化分割结果 |
| | | 缺点:分割目标的边界信息部分丢失 | | refineNet block ^[44] , 级联结构 ^[44,45] | 特征融合,逐步精细化分割结果 |
| | | | | 多尺度卷积 ^[49,50] | 多尺度特征融合,捕捉上下文信息 |
| 基于RNN的方法 | rCNN ^[51] , ReSeg ^[52] , DAG-GNN ^[53] , 2D-LSTM ^[54] | 优点:可递归处理历史信息 and 建模历史记忆,易于提取图像中的像素序列信息和捕获上下文信息 | 将 RNN 部分与卷积层相结合嵌入到 DNN 中,利用 CNN 中的卷积层提取局部空间特征,再用 RNN 层提取像素序列特征,其相关网络结构有: ReNet、LSTM 和 GRU | 传统RNN序列模型 ^[51] , ReNet ^[52] | 序列化图像像素,递归处理历史信息 |
| | | 缺点:图像像素序列过程中会发生部分像素信息丢失的问题 | | LSTM ^[54] , GRU ^[52] | 从多种途径获取图像上下文信息 |
| 基于GAN的方法 | 文献[55], 文献[56], 文献[57], 文献[58] | 优点:避免使用CRF带来的问题,在不增加模型复杂度和训练时间的前提下增加长距离空间标签的连续性,具有较强的空间一致性 缺点:优化过程不稳定,训练时容易坍塌到一个鞍点上;处理大规模数据时,可解释性和延展性不高 | 使用一个分割网络组成生成器网络并让其不断产生分割图像,判别器网络不断观察真实标签和分割图像并判断输入数据的真假;训练时两组网络相互竞争、相互提高,直至分割图像与真实标签一致 | GAN ^[55] , CGAN ^[58] | 捕获图像上下文信息,增加长距离空间标签间的连续性 |

3.1.1 基于 FCN 的方法

FCN 在进行图像语义分割时没有充分考虑像素与像素之间的关系,缺乏空间一致性,对图像中的细节不够敏感,导致分割结果不够精细.文献[25]在 FCN 的末端增加全连接条件随机场(fully connected conditional random field,简称 FCCRF),对粗糙分割图进行边界优化,并使用带孔卷积(atrous convolution)扩大特征图的感受野(receptive field,简称 RF),提出了 DeepLab 网络.DeepLab 的处理流程如图 6 所示,首先,将图像送入到结合了 Hole 算法的 FCN 中进行处理,得到粗略的特征图,再使用 BI 算法对 FCN 的输出结果进行上采样操作得到粗糙

分割图像;然后,使用 FCCRF 对粗糙分割图像进行结构化预测,并对图像中的像素点进行建模、求解,平滑处理粗糙分割图像的边缘;最后得到一个完整的图像语义分割结果。

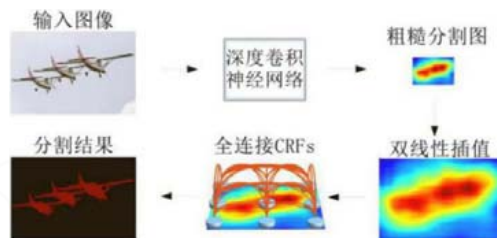


Fig.6 General framework for DeepLab^[25]

图 6 DeepLab 基本框架^[25]

2016 年,文献[26]在 DeepLab 网络的基础上提出了 DeepLab-V2 网络,对特征图分辨率变小、定位精度过低等问题进行改进.与 DeepLab 网络相比,DeepLab-V2 网络不仅使用带孔卷积作为上采样滤波器进行稠密特征提取,而且将带孔卷积与空间金字塔池化方法(spatial pyramid pooling method)^[20]相结合后,提出带孔空间金字塔池化(atrous spatial pyramid pooling,简称 ASPP),并利用 ASPP 整合多尺度特征,最后,再用 FCCRF 优化分割图像,在不增加过多参数的情况下,增大了感受野、提高了分割精度。

2017 年,文献[27]基于上述两种方法^[25,26]级联多个带孔卷积模块,并在空间维度上对 ASPP 进行改进,提出了 Deeplab-V3 网络. Deeplab-V3 网络以并行方式将 4 个不同孔洞率(atrous rate)的带孔卷积并联,组成一个改进版的 ASPP,再以串行方式将多个带孔卷积模块与改进后的 ASPP 串联在一起,构成一个端到端处理图像的网络. Deeplab-V3 结合串行方式与并行方式的带孔卷积后,能够多尺度(multiple scale)地分割物体,获取多尺度的图像信息.实验结果表明,Deeplab-V3 在无需加入 FCCRF 的情况下,分割准确率仍然优于文献[25,26]中的方法。

由于文献[25,26]只是简单地把 FCCRF 加在 FCN 的末尾,需要分别训练 FCN 和 FCCRF,导致 FCN 的粗分割操作与 FCCRF 的精分割操作没有过多的交互联系.文献[28]在文献[25]的基础上提出了 CRFasRNN 网络,训练时,运用 BP(back propagation)算法将 CRFasRNN 网络放在 FCN 的末尾,进行端到端处理. CRFasRNN 网络的基本思路是:把条件随机场(conditional random field,简称 CRF)的学习、推理、求解等过程迭代建模为 RNN 的相关运算,通过迭代 mean field 算法把该过程嵌入到 CNN 中,从而将 CNN 与 CRF 的学习过程统一在一个完整的网络中,提升了分割效果。

3.1.2 基于优化卷积结构的方法

在使用 CNN 进行图像语义分割过程中,池化操作能够增大特征图的感受野,并汇合图像的背景信息,但也带来了特征图分辨率不断降低、部分像素的空间位置信息丢失等问题.一个解决该问题的思路是对神经网络中的卷积结构进行优化,使用经过优化的卷积结构来代替传统的卷积、池化等操作。

文献[29]对普通的卷积操作进行优化,从中引入不同的扩张率(dilation rate),提出了扩张卷积(dilated convolution).扩张卷积是在正常连续的卷积操作中加入不同的间隔,间隔大小由扩张率决定,可以在不损失分辨率、不增加计算量的情况下使感受野呈指数增长,并可捕获图像的多尺度局部特征和保留大部分像素的空间位置信息,提升了分割准确率.事实上,这里的扩张卷积也是一种带孔卷积^[23],两类卷积操作的具体结构和计算方式基本一致,只是名称略有不同。

在 ISS 过程中,使用扩张卷积能够更有效地提取图像特征、增加感受野,并保留一部分像素的空间位置信息.但扩张卷积在操作过程中易产生空间间隙,会出现信息丢失、信息不相关和局部信息之间的连续性被打断等问题.2017 年,文献[30]使用混合扩张卷积(hybrid dilated convolution,简称 HDC)代替扩张卷积,并使用稠密上采样卷积(dense upsampling convolution,简称 DUC)替换 BI 算法. HDC 由一系列不同扩张率的扩张卷积模块组成,既能增加感受野,又能保持局部信息的相关性,有效地避免了上述问题。

虽然上述方法^[29,30]可以增加感受野的大小,但其固定形状的卷积核对几何变换的模拟能力不强,对图像中

一些未知形变的适应性较差,不利于提取形状不规则物体的特征.文献[31]将带有偏移量的采样操作代替原来位置固定的采样操作,在卷积处理的作用区域上加入一个可学习的偏移量,使卷积操作的卷积核具有可变的形状,提出了可变形卷积(deformable convolution).可变形卷积既能增大感受野的范围,又有利于 DNN 学习图像中的感兴趣区域,增强了对 ISS 过程中对几何变换的适应性,提高了分割准确率.

3.1.3 基于编码器-解码器的方法

在 ISS 领域,要解决“池化操作后特征图分辨率不断降低、部分像素空间位置信息丢失”等问题,除了对卷积结构进行优化外,另一类方法是使用编码器-解码器(encoder-decoder)结构.该方法是一种利用对称网络结构进行图像语义解析的机制,其本质是利用 DL 技术中的卷积、池化等操作所构成的编码器来编码被捕获的像素位置信息和图像特征,再利用反卷积(deconvolution)或上池化(unpooling)等操作所构成的解码器来对其进行解析,还原图像的空间维度和像素的位置信息.

2015 年,文献[59]利用编码器-解码器结构在编码过程中进行下采样(subsampled)操作,逐步减少特征图的分辨率,在解码过程则进行上采样(upsampled)操作,逐步恢复物体细节和图像分辨率,提出一种对生物医学图像进行语义分割的 U-net 网络.文献[60]以解决自动驾驶汽车和智能机器人的 ISS 问题为目标,提出了 SegNet-Basic 网络.SegNet-Basic 网络基于先验概率计算每个像素点的分类,是一个类似编解码过程的对称结构网络,其基本结构如图 7 所示.该网络的左边是一个由全卷积网络构成的编码器,通过卷积、池化等操作进行下采样处理;右边是一个由反卷积网络构成的解码器,利用转置卷积和上池化操作进行上采样处理.针对先验概率无法给出分类结果置信度的问题,文献[32]基于 SegNet-Basic 网提出了 Bayesian SegNet 网络.Bayesian SegNet 在每个卷积层后面增加了一个 DropOut 层,可有效防止权重过度拟合并增强网络的学习能力;同时,还引入贝叶斯网络(Bayesian network)和高斯过程,基于后验概率计算像素类别,使网络在 ISS 过程中能更合理地模拟事件概率.

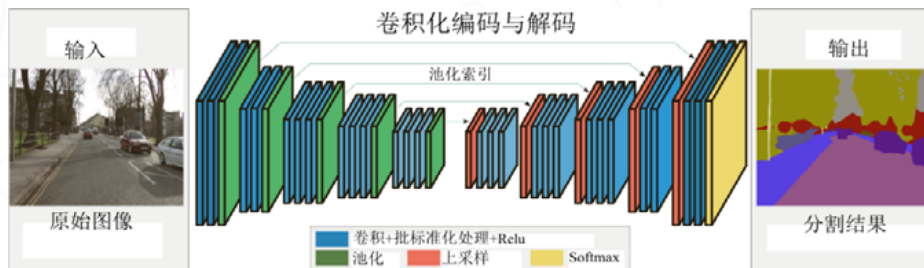


Fig. 7 Framework for SegNet-Basic^[60]

图 7 SegNet-Basic 网络框架^[60]

文献[33]在 FCN 的基础上提出一个完全对称的 DeconvNet 网络.该网络用反卷积替换 BI 算法,建立了一种完全对称机制;同时,将反卷积操作与上池化操作结合起来,在一定程度上避免了细节丢失和边界模糊等现象,更好地反映了物体的细节,提高了分割效果.

文献[32,33,60]虽然能够取得较好的分割结果,但在网络训练中也存在参数权重过多、计算量过大等问题,导致分割速度难以达到实时处理的要求.2016 年,文献[34]基于编码器-解码器结构提出一种高效神经网络 Enet (efficient neural network).ENet 在执行分割任务时采用分解滤波器策略,使用低阶近似(low-rank approximation)将卷积操作分解为更简单的操作,在保证分割精度的同时,显著降低了计算量,是一种可完成像素标注、场景解析等任务的实时分割网络.ENet 中编码部分的网络比解码部分复杂很多,是一种不对称的编码器-解码器结构.

文献[35]使用带有大型卷积核(large kernel)的卷积层代替 CNN 中的全连接层,提出了全局卷积网络(global convolution network,简称 GCN).该方法使用 ResNet 构建编码器,使用 GCN 和反卷积构建解码器,并且网络中加入了用于边界优化的简单残差块,能够在像素相对集中的小面积区域捕获图像信息,使得物体的分割边界更清晰、分割准确率更高.

3.1.4 基于概率图模型的方法

“未能充分考虑图像的上下文信息,以及局部特征和全局特征的利用率失衡”是 ISS 在发展过程中所面临的另一个问题,而将概率图模型(probabilistic graphical model,简称 PGM)^[61]用于 CNN 的后期处理,则能有效地捕获图像上下文信息,并且平衡局部特征与全局特征的利用率.PGM 建模时以像素点作为节点,像素点与像素点之间的概率相关关系作为边.PGM 可有效地获取各像素点之间的依赖关系,捕捉图像全局信息和像素级语义信息,进而为语义分割过程提供丰富的图像上下文信息.该类方法的处理流程如图 8 所示,先用 CNN 对原始图像进行特征提取,得到粗糙分割结果;再将其送入 PGM 中用于捕获语义信息和像素依赖关系,对粗糙分割结果进行边缘细化、精度加工等优化,从而得到精细化的分割结果.图 8 虚线框的内容是使用 PGM 对图像进行建模的过程.常用的 PGM 包括马尔可夫随机场(Markov random field,简称 MRF)、条件随机场、贝叶斯网络等.

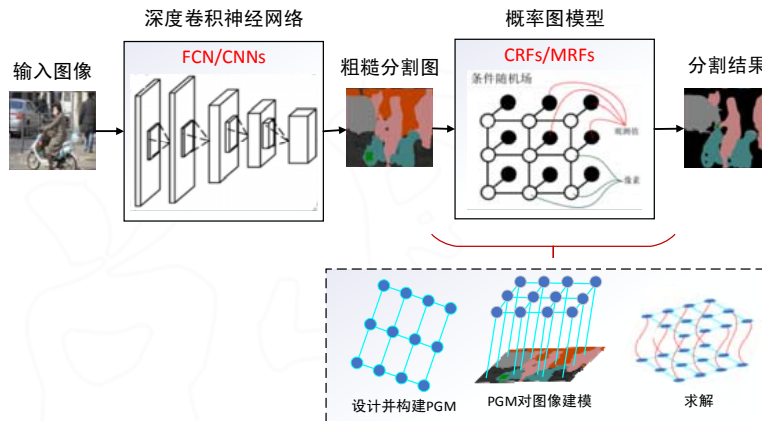


Fig.8 Workflow for ISS based on PGM

图 8 基于概率图模型的图像语义分割方法的处理流程

基于上述思想,文献[38]引入 MRF 来捕捉 ISS 中复杂的上下文信息,将 CNN 与 MRF 结合,提出一种端到端训练的深度解析网络(deep parsing network,简称 DPN).DPN 将高阶关系(highorder relation)、标签信息和语义信息结合在 MRF 中,通过 CNN 中的基础网络层来近似模拟 MRF 的一元项(unary term),通过 MRF 的额外层来近似估计 MRF 的成对项(pairwise term),其反向传播过程不需要额外的迭代计算便能够在 GPU(graphics processing unit)中进行并行加速处理,从而减少了计算量,提高了分割效率.

文献[62]以 CRF 为基础设计了一个目标团势函数(object clique potential),提出一个能够有效地检测并分割物体的新模型.文献[36]则综合利用 CNN 与 CRF 的优势来探索图像中复杂的上下文信息,提出一种深度结构化模型(deep structured model,简称 DSM).DSM 在建模 CRF 后,利用相邻图像块之间的语义关系来捕获“区域-区域上下文(path-path context)”;通过滑动金字塔池化技术连接由 CNN 输出的多尺度特征图,以此来捕获“区域-背景上下文(path-background context)”;最后,综合利用两种不同的图像上下文信息进行语义分割,提高了分割准确率.文献[63]基于上述思想^[36],将 CNN 与 CRF 组合在一起用于结构化预测,直接在消息的传递推理过程中预测消息,避免了大量计算梯度,是一种运行速度更快、运算量更小的 DSM.文献[37]对传统的 CRF 改进后将其嵌入到 CNN 中,提出一个从粗糙分割到精细分割逐步转变的网络模型.该网络^[37]在粗糙分割阶段,使用 FCN 提取图像的空间特征后进行特征组合,再用“语境 CRF(context CRF)”重构经过组合的特征;在精细分割阶段,用“指导 CRF(guidance CRF)”根据输入图像的边界来描绘物体轮廓,精细化分割对象的边界.

文献[36,37,63]在进行图像语义分割后续优化处理时,只将粗糙分割图像输入 CRF 的一元项和成对项进行结构化预测,忽略了对 CRF 中高阶势能项(higher order potential,简称 HOP)^[64]的有效利用,在捕获全局特征和图像上下文信息时存在一定的不确定性和误差.文献[42]将两种不同的 HOP 加入到 CRF 后,将其与 CNN 结合,提出一种新的 ISS 模型,避免了上述缺陷.文献[40]设计了一个基于目标检测的 HOP 和一个基于超像素^[65]的 HOP,

并将这两种 HOP 嵌入到 CNN 中进行端到端训练,提高了 ISS 的分割准确率。

文献[41]尝试用高斯条件随机场(Gaussian conditional random field,简称 GCRF)代替传统 CRF 执行对分割结果的后续优化任务,通过固定次数的迭代高斯平均场(Gaussian mean field,简称 GMF)提出高斯平均场网络(Gaussian mean field network,简称 GMF network),再将 GCRF、GMF network 与 CNN 结合在一起,共同处理 ISS 问题,得到一种端到端语义分割的高斯条件随机场网络(Gaussian conditional random fields network,简称 GCRF network)。文献[39]使用 CNN 分别学习 GCRF 的一元势函数和二元势函数,提出一种端到端训练参数的二次优化(quadratic optimization,简称 QO)模型,提高了 ISS 后续优化处理的效率。

3.1.5 基于特征融合的方法

“利用 CRF 等概率图模型作为 CNN 的后期处理”能够有效地捕获图像上下文信息,提高全局特征的利用率,但概率图模型方法在学习、推理过程中仍存在计算量过大、训练时间较长、严重消耗内存等缺点。特征融合是整合图像上下文信息并提高全局特征利用率的另一种策略,基于特征融合的方法主张兼顾图像的全局特征、局部特征以及高、中、底等各层次特征,通过融合不同层次特征、不同区域特征来捕获图像中隐含的上下文信息,可有效地避免使用概率图模型导致的问题。

文献[66]将提取的全局特征经过上池化处理后加入到局部特征中,两种特征融合后获得图像的上下文信息,再将上下文信息与融合后的特征一起用于下一层网络的处理。文献[43]通过拉普拉斯金字塔(Laplacian pyramid)算法^[67]将不同卷积层提取的低层特征进行重构,提出了 LRR(Laplacian pyramid reconstruction and refinement model)模型。LRR 模型把特征图表示为一组基函数的线性组合,使用跨层方法^[23]引入边界信息后,将低层特征与高层特征进行融合,可有效地捕获图像的上下文信息,并对粗糙分割结果求精。

文献[48]将原始图像进行尺度变换后并行输入 FCN,并引入注意力模型(attention model)^[68],对不同尺度的目标物体赋予不同的权重,再分别学习其对应的特征图,最后,融合多尺度特征进行像素分类。文献[44]提出一种能够进行多级并行处理的级联式 RefineNet 网络,其框架结构如图 9 所示。在该网络中,原始图像首先经过 CNN 处理得到 1/4、1/8、1/16 和 1/32 尺度的 4 种不同分辨率特征图,这些特征图再送入与之对应的精细模块(RefineNet block)融合。如此迭代数次,通过多路径优化处理,不同层次、不同分辨率的特征图融合得到优化的分割结果。RefineNet 中的精细模块由一系列经过残差连接的组件构成,每个组件对低分辨率特征图进行上采样处理后再融合高分辨率特征图,可有效整合不同尺度、不同层次的特征,对图像上下文信息的利用更充分、合理。

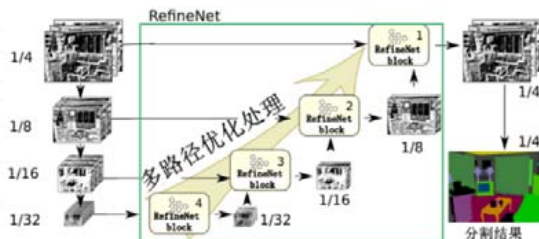


Fig.9 Framework for RefineNet^[44]

图 9 RefineNet 网络框架^[44]

文献[45]使用卷积核大小不同的 4 种 CNN 来捕捉局部特征和全局特征,对图像特征进行级联处理,将多个同一层次的图像特征进行融合,提出了 PSPNet 网络。PSPNet 的处理流程如图 10 所示:图像经过 CNN 处理后获得粗糙特征图,然后再通过空间金字塔池化(spatial pyramid pooling)^[20]模块对特征图进行级联处理,将 4 种不同尺度的特征图进行融合,得到完整的特征表达,能够分别获取不同区域的上下文信息,进一步提升了分割精度。文献[46]基于 PSPNet 网络,在兼顾分割精度的同时,为保证实时性,提出了能够实时分割的图像级联网络(image cascade network,简称 ICNet)。ICNet 对不同尺寸的输入图像进行下采样操作,将低分辨率图片通过整个 CNN 网络后得到粗糙分割图,然后利用级联特征融合单元(cascade feature fusion unit,简称 CFF)来融合高分辨率图片的特征,从而提高分割速度。

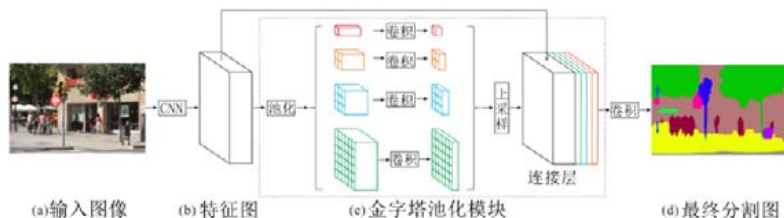


Fig.10 Workflow for PSPNet

图 10 PSPNet 网络的处理流程

文献[47]借鉴文献[69–71]的思想,使用区域卷积(region convolution)对每个阶段的感兴趣区域进行处理,而默认忽略其他不感兴趣的区域,提出了端到端训练的深层级联(deep layer cascade,简称 LC)方法.LC 方法具有一定的自适应能力和自主学习能力,能够将不同复杂度的图像区域分别放在不同深度的网络层进行处理,可以有针对性地处理不同难易程度的像素。

除上述图像特征融合方式外,一些学者主张将上一阶段卷积网络提取的图像特征融入到下一阶段卷积网络提取的特征中,将不同卷积网络提取的不同图像特征进行融合.基于这种思想,文献[72]在 VGG-16 网络中加入一系列不同尺度的卷积操作,从图像中提取出不同尺度的特征信息后,将上一阶段特征融入下一阶段的特征中进行卷积处理,设计出一种能够捕获高层和低层特征的多尺度深度卷积神经网络(multi-scale deep convolutional neural network).文献[49]在文献[50]的基础上增加了 3 个不同尺度的子网络,采用“先进行联合训练、后进行阶段训练”的策略,可独立完成深度估计、法向量估计和 ISS 这 3 个任务.文献[73]从结构上对上面的多尺度 CNN^[49,50]进行改进,将其扩展为 4 个子网,并利用其中一个子网进行粗糙分割,其余 3 个子网进行精细分割.文献[74]则在 FCN 的基础上增加多个不同尺度的卷积层,提出了多尺度全卷积网络(multi-scale fully convolutional network).

3.1.6 基于 RNN 的方法

针对“无法充分利用图像上下文信息、局部特征和全局特征的利用率失衡”等问题,另一种解决思路是:利用 RNN 可递归处理历史信息 and 建模历史记忆的特点,在分割图像过程中使用 RNN 来捕获图像上下文信息和全局特征.RNN 不仅可以学习当前时刻的信息,还可以依赖之前的序列信息,有利于建模全局内容和保存历史信息,促进了图像上下文信息的利用.基于 RNN 的方法进行图像语义分割时,将 RNN layer 嵌入到 CNN 中,在卷积层提取图像的局部空间特征,在 RNN layer 提取像素序列特征.其一般处理流程如图 11 所示,首先,输入图像经过 CNN 处理后得到特征图;然后,将特征图输入 RNN 中获取图像上下文信息,用 RNN layer 序列化像素、分析各像素的依赖关系后得到全局语义特征,再使用反卷积层进行上采样处理;最后,得到分割结果。

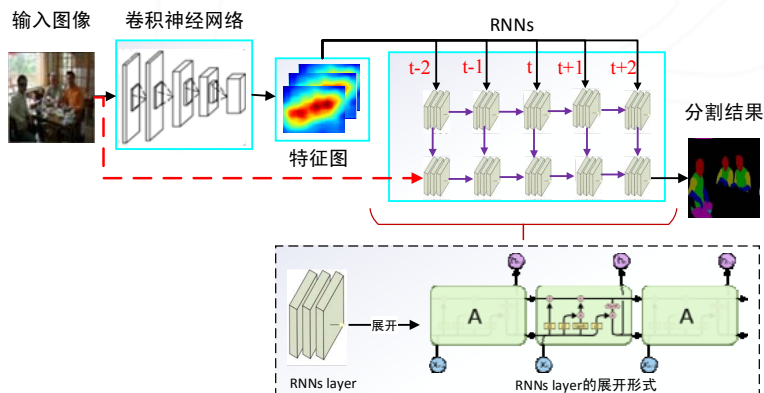


Fig.11 Workflow for ISS based on RNN

图 11 基于循环神经网络的图像语义分割方法处理流程

文献[51]借鉴 RNN 的循环思想,将泛化后的 RNN 应用于 ISS 领域.文献[53]利用 DAG(directed acyclic graph)的特性来弥补分割过程中“RNN 无法直接应用于 UCG(undirected cyclic graph)结构图像”的缺陷,使 RNN 能够直接对图像中的每个像素进行分类.文献[52]综合利用 CNN 与 RNN 的优点,使用 RNN 的衍生网络 ReNet 处理图片数据,提出了 ReSeg 网络.ReSeg 使用 4 个传统的 RNN 序列模型替代 CNN 中卷积层的卷积和池化操作,在水平与垂直两个方向分别切割图像并对其空间依赖关系建模,其框架结构如图 12 所示,输入图像经过 VGG-16 网络后得到图像的局部特征,然后将特征图送入 ReNet 网络逐步提取出图像的全局特征和上下文信息,最后使用由反卷积网络组成的上采样层逐步恢复特征图的分辨率,输出分割结果.同时,ReSeg 还使用 GRU 来平衡内存占用率和计算负载能力,具有很高的灵活性和适应性.文献[54]用 4 个不同方向的 RNN 子网来完成图像标注任务:输入图像被划分为多个非重叠窗口送入 4 个独立且不同方向的 LSTM 记忆块,在没有其他附加条件的情况下,捕获局部特征和全局特征.文献[75]则利用光度和深度 2 种不同类型的数据来建模全局特征,使用 LSTM 从多种途径获取上下文信息再将其整合到 CNN 中,增强了语义特征的表达效果.

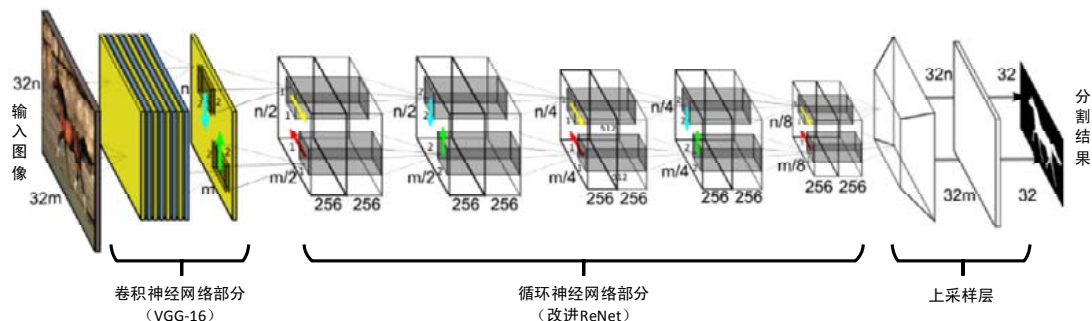


Fig.12 Framework for ReSeg^[52]

图 12 ReSeg 网络的框架结构^[52]

传统的 RNN 序列模型在训练过程中容易出现梯度爆炸或梯度消失等问题,使用其衍生网络 LSTM 或 GRU 配合梯度截断(gradient clipping)、批标准化(batch normalization)等策略则可避免该问题^[76,77].LSTM 和 GRU 利用某些特定的方法来规则遍历二维图像,图像上下文的关联性可转化为结构化的循环依赖关系,易于建模像素序列以及图像空间域的依赖关系^[52].在利用 LSTM 或 GRU 对图像进行建模并将其平滑为像素序列的过程中,需尽量保存图像的时间序列特征,并尽量避免丢失其高级语义信息和像素依赖关系,以提升分割效果^[78,79].

3.1.7 基于 GAN 的方法

“使用带有高阶势能项的 CRF 进行结构化预测”存在着模型复杂、计算量大、训练时间长、内存占用率高等问题,此外,HOP 也需要预先人工设置,不易实现.在 ISS 过程中,使用 GAN 代替 CRF 捕获图像上下文信息,不但能够避免上述问题,还可在不增加模型复杂度和训练时间的情况下增加长距离空间标签的连续性,具有较强的空间一致性.

基于 GAN 的方法进行图像语义分割时,基本框架结构如图 13 所示.该类方法一般使用 FCN,SegNet 或 PSPNet 等分割网络作为生成器网络,输入图像经过生成器网络处理后得到预测分割图像,将预测分割图像作为人造样本、真实标签图像(ground truth)作为真实样本输入判别器网络,判别器网络学习真实样本和人造样本的区别,并基于博弈思想进行对抗训练.待输出样本数据的真假后,其内部的反馈机制会对生成器网络与判别器网络进行调节,经过数次迭代训练后,生成器网络的分割准确率和判别器网络的鉴别能力不断提高.图 13 中,虚线表示 GAN 利用判断结果进行反馈微调;菱形标志表示选择“真实标签图像”或“预测分割图像”两者中的一种作为判别器网络的输入.当判别器网络的输入组合为“原始图像”与“预测分割图像”时,输出“假”代表正确结果;当输入组合为“原始图像”与“真实标签图像”时,输出“真”代表正确结果.

2016 年,文献[55]首次将 GAN 引入 ISS 领域,提出一种图像分割的新方法,原始图像在由 CNN 构成的分割网络中转变为分割结果,分割结果输入对抗网络后被判断出真假,两组网络进行对抗学习、彼此竞争,经过迭代

训练后,逐步提高分割网络的分割准确率.文献[56]基于 FCN,将 GAN 与领域适应性(domain adaptation)思想结合,将源域与目标域共享标记空间,并通过最优化目标损失函数来减少全局偏移和特定偏移的影响,提出用于 ISS 的领域适应性框架.文献[57]通过 GAN 来实现分割网络的参数规则化,使用未进行标注的图像训练分割网络(生成器网络).文献[58]则利用条件生成对抗网络(conditional generative adversarial network,简称 CGAN)^[80]产生人造样本进行对抗训练.

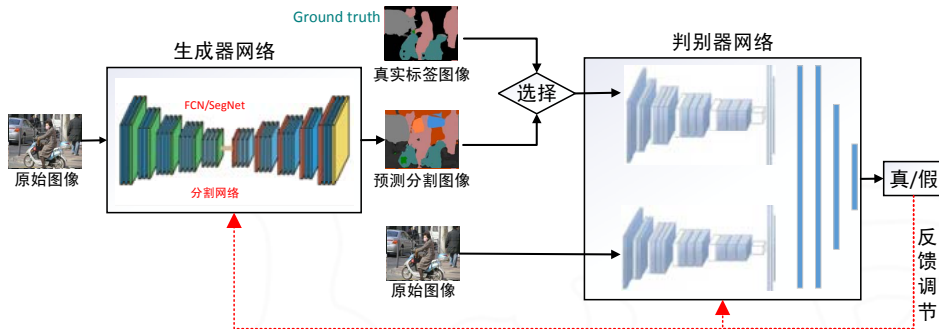


Fig.13 Framework for ISS based on GAN

图 13 基于 GAN 的图像语义分割方法的基本框架

GAN 模型拥有不断生成数据和辨别数据真假的能力,在一定程度上可减少 CNN、FCN 等网络在图像语义分割过程中带来的问题.同时,GAN 引入判别器来解决数据域分布不一致问题,通过对抗学习来近似不可解的损失函数,具有较好的分割效果^[81].但 GAN 模型的优化过程不稳定,训练时容易坍塌到一个鞍点上,在处理大规模图像数据时,其可解释性和延展性有待提高^[82].

3.2 弱监督学习图像语义分割方法

利用 CNN、FCN 等神经网络进行全监督学习的分割方法在 ISS 领域获得了较好的效果,但 ISSbFSL 方法在训练阶段所需要的像素级精确标注图像制作过程费时、费力,难以大批量获取.因此,一些学者开始研究以弱监督学习的方式进行图像语义分割,形成了一系列的 ISSbWSL 方法.ISSbWSL 方法使用经过粗略标记的弱标注图像进行训练,减少了标注时间和标注成本.在 ISS 领域,目前常见的弱标注数据大致有图像级标注、边框级标注和涂鸦级标注.如图 14 所示,边框级标注泛指一些人工标记的边界框(bounding box),涂鸦级标注泛指一些人工随机涂鸦的点或线条,图像级标注则把图像中的物体种类标签作为标注.

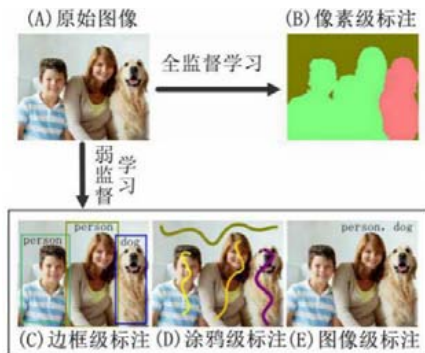


Fig.14 Examples of different image labeling

图 14 不同类型的图像标注示例

与像素级标注数据相比,弱标注数据无需过多人工操作,更容易获取.根据弱标注数据的不同类型,我们将 ISSbWSL 方法分为 4 类:基于边框级标注的方法、基于涂鸦级标注的方法、基于图像级标注的方法和多种弱

标注数据混合的方法.下面进行详细介绍.

3.2.1 基于边框级标注的方法

在 ISSbWSL 领域,基于边框级标注的方法使用边框级标注图像作为训练样本.与 ISSbFSL 方法相比,该类方法在标注边界框过程中时间成本较低,可节省大量人力物力;同时,其分割性能与同等条件下的全监督学习方法近似.

文献[83]以 FCN 为基础网络,用边框级标注的图像作为训练样本,通过循环迭代方式不断提高分割准确率,提出了 BoxSup 网络.BoxSup 的框架结构如图 15 所示,首先,用 MCG 算法^[19]得到初步的目标候选区域;然后,将该目标候选区域作为已知的监督信息输入到 FCN 中进行优化和更新;待 FCN 输出具有更高精度的候选区域后,再将输出的目标候选区域又重新输入 FCN 中进行训练.如此重复迭代,直到准确率收敛.文献[84]在 GrabCut 算法^[85]的基础上加入 CRF 和 CNN,提出了 DeepCut 方法.DeepCut 使用边框级标注的图像作为训练样本,通过在 CNN 中进行迭代训练,逐步提高图像的分割精度.

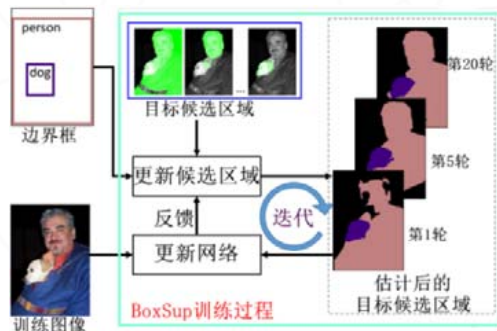


Fig.15 Framework for BoxSup

图 15 BoxSup 网络的框架结构

3.2.2 基于涂鸦级标注的方法

基于涂鸦级标注的方法使用涂鸦级标注的图像作为训练样本,分割过程较为简单,训练样本易于获取,降低了人工标注的工作量.

文献[86]用多个像素标注图像中的物体,提出了用随机涂鸦的点作为监督信息的点监督(point supervision)方法.该方法将监督信息与 CNN 模型中的损失函数相结合,取得了较好的分割效果.文献[87]使用涂鸦方式进行图像标注,将带有涂鸦线条的图像作为训练样本,提出了 ScribbleSup 方法.ScribbleSup 方法分为自动标记阶段和图像训练阶段:自动标记阶段首先根据涂鸦线条对图像生成像素块,然后将每个像素块作为图模型中的一个节点,再用 GraphCut 算法建模自动完成对所有训练图像的标注;图像训练阶段则是将上一阶段完成的标注图像送入 FCN 中训练,得到分割结果.

3.2.3 基于图像级标注的方法

与其他类型的标注相比,图像级标注使用类别标签作为训练标注,不需要进行像素标注,制作更简单、工作量更小,因而受到研究者更多的关注,成为 ISSbWSL 方法的主流.但是图像级标注只提供了物体种类信息,缺少位置、形状等信息,在分割过程中也面临诸多挑战.

文献[88]将多示例学习(multiple instance learning,简称 MIL)^[89]应用在 ISSbWSL,以构建图像标注与像素语义之间的关联;同时,还使用超像素、候选框和 MCG 算法等技术作为后续处理,取得了一定的分割效果.文献[90]使用约束卷积神经网络(constrained convolutional neural network,简称 CCNN)进行图像语义分割,CCNN 将图像级标注作为限制条件,通过内部的损失函数来预测像素类别,把训练过程看作是限制条件的最优化过程.文献[91]使用复合形式的损失函数进行图像语义分割,提出了 SEC(seed, expand and constrain)方法.SEC 方法中的复合损失函数由 3 个不同的目标损失函数组成,训练 CNN 时,3 个不同的目标损失函数分别完成不同的分割任务.

文献[92]提出一种可将分割结果由简单到复杂逐步转变的 STC(simple to complex)方法.该方法首先利用显

著性目标检测(salient object detection)算法检测出显著性区域,进行区域特征融合和构建像素间的语义关系后,由 CNN 产生一组显著性区域图.接着,再由一套迭代机制从简单到复杂地重复数次该过程,逐步提升分割精度.文献[93]在循环迭代的训练过程中引入增强反馈(augmented feedback)思想,先使用选择搜索算法和 MCG 算法进行目标定位,再根据反馈信息逐步提升分割能力,可在一定程度上避免 ISSbWSL 存在的误差累积问题.文献[94]基于 CNN 和期望值最大化(expectation maximization,简称 EM)算法,使用显著性目标检测算法和注意力图(attention maps)对图像进行分割,分割效果较好.文献[95]基于 FCN,采用弱监督学习方式识别出图像中多个不同的显著性区域(discriminative localization)后,捕获不同区域的局部特征,再利用局部特征对图像中的物体进行定位和分割.文献[96]采用图像级标注图像对分类网络进行训练,使用分类网络按照主次顺序逐步获取目标对象的显著性区域,并利用一种逐步擦除显著性区域、不断提高像素分类精度的方法来处理图像的显著性区域,使得分割精度迅速提升.

3.2.4 多种弱标注数据混合的方法

以上 3 种使用弱标注图像进行训练的分割方法极大地推动了 ISSbWSL 的发展,但由于弱标注图像的局限性,单纯使用某种图像级标注的训练效果较差,如果混合多种弱标注图像则可能取得较好的效果.多种弱标注数据混合的方法将多种弱标注图像与像素级标注图像相互混合,通过混合训练的方式进行半监督学习.

文献[97]引入半监督学习思想,将分类和分割相结合,提出了由分类网络和分割网络组合而成的 DecoupledNet 网络.训练时,DecoupledNet 先用大量的图像级标注训练分类网络,再用少量的像素级标注训练分割网络.这种方法没有循环迭代过程,不必考虑迭代收敛,因而具有较好的扩展性.文献[98]在 DeepLab 网络的基础上,将边框级标注与图像级标注一起用于训练,对于给定的边框级标注图像,先使用 CRF 对其做自动分割,再在分割结果上做全监督学习.同时,该方法^[98]还尝试将少量的像素级标注图像和大量的弱标注图像结合训练,并使用 EM 算法来预测未标记像素的类别,其分割结果与进行全监督学习的 DeepLab 网络十分接近.

总的来说,ISSbWSL 方法尝试在大量弱标注数据中找到图像的高级特征,允许计算机在无人指导的情况下进行学习,可使 ISS 摆脱对大量像素级标签数据的依赖.但目前大部分 ISSbWSL 方法没有充分考虑噪声干扰下弱标注图像分布的不确定性和复杂性,其分割性能与 ISSbFSL 方法相比仍有较大差距.如何灵活运用弱标注图像数据来提高分割准确率和抗干扰性,并减少大量弱标注图像所带来的计算复杂性,是该类方法目前亟需解决的问题.

4 图像语义分割实验分析与对比

在进行图像语义分割实验时,要对每种方法进行公平、客观的评价,就必须使用权威的数据集和统一的评价指标.本节将介绍一些在 ISS 实验中常用的公共数据集以及衡量算法性能的指标,并对前文中一些经典方法的实验结果进行系统的分析和对比.表 3 总结了图像语义分割的常用公共数据集.

Table 3 Common datasets for image semantic segmentation
表 3 常用的图像语义分割数据集

| 文献 | 数据集 | 设计目的/应用范围 | 种类数量 | 数据总量 | 分辨率 | 训练集 | 验证集 | 测试集 |
|-------|-----------------|-----------|---------|----------|-----------|--------|--------|--------|
| [99] | PASCAL VOC 2012 | 多种应用 | 21 | 9 993 | 不固定 | 1 464 | 1 449 | 1 452 |
| [100] | PASCAL-CONTEXT | 多种应用 | 59(540) | N/A | 不固定 | 10 103 | 10 103 | 9 637 |
| [101] | PASCAL-PART | 人体解析 | 21 | N/A | 不固定 | 10 103 | 10 103 | 9 637 |
| [102] | MS COCO | 多种应用 | 81 | 328 000 | 不固定 | 82 783 | 40 504 | 81 434 |
| [103] | ILSVRC | 多种应用 | 2 万多 | 1 400 多万 | 不固定 | N/A | N/A | N/A |
| [104] | KITTI | 城市街道场景解析 | 10 | N/A | 1226×370 | 140 | N/A | 112 |
| [105] | Cityscapes | 城市场景解析 | 8(30) | 2 万左右 | 2048×1024 | 22 973 | 500 | N/A |
| [106] | Sift Flow | 户外场景解析 | 33 | N/A | 256×256 | 2 688 | N/A | N/A |
| [107] | SBD | 户外场景解析 | 8 | N/A | 320×210 | 725 | N/A | N/A |
| [108] | NYUDv2 | 室内场景解析 | 40 | 40 多万 | 480×640 | 795 | 654 | N/A |
| [109] | SUN-RGBD | 室内场景解析 | 37 | 1 万左右 | 不固定 | 2 666 | 2 619 | 5 050 |

注:N/A 表示相关论文未提及、无法确定该信息

4.1 常用公共数据集

- (1) PASCAL Visual Object Classes(简称 PASCAL VOC)^[99].PASCAL VOC 是一个国际计算机视觉挑战赛,该组织提供了领域内知名度最高的图像测试数据集和计算机视觉领域的基准测试.2005 年~2012 年间,该组织每年都发布带标签的图像数据库并开展算法竞赛,由此产生了一系列数据集.目前,该系列最常用的数据集是 PASCAL VOC 2012.该数据集涉及物体共 21 种,包括人类、动物、交通工具和室内物品等,图片大小不固定,背景复杂多变.
- (2) PASCAL-CONTEXT^[100].PASCAL-CONTEXT 数据集由 PASCAL VOC 2010 数据集改进和扩展而来,里面增加了更多的物体标注和场景信息,总共包含 540 个语义类别的图像标注.在算法评估时,一般选取前 59 类作为分割评判标准.
- (3) PASCAL-PART^[101].PASCAL-PART 数据集中的图片大都出自 PASCAL VOC 2010,分为训练集、验证集和测试集这 3 个部分,每部分中的图像都含有像素级标注,能够提供丰富的细节信息.PASCAL-PART 每张图像中,目标物体的不同部位都有精确标注,可为物体解析和 ISS 任务提供详细标注的样本.
- (4) Microsoft Common Objects in Context(简称 MS COCO)^[102].MS COCO 数据集早先是微软公司进行图像测试的一个大型数据库,后来,微软公司将其开源和推广.MS COCO 数据集总共包含 81 种类别(包括背景)、328 000 张图像、2 500 000 个物体实例和 100 000 个人体关键部位标注,大部分图片从复杂的日常场景中获取,图中的物体具有精确的位置标注.
- (5) ImageNet Large Scale Visual Recognition Challenge(简称 ILSVRC)^[103].ILSVRC 也是一个著名的国际计算机视觉挑战比赛,提供的 ImageNet 数据集有 1 400 多万幅图片,涵盖 2 万多个类别,其中,超过百万的图片有明确的类别标注和物体位置标注.数据集文档详细,有团队维护,使用方便,在图像研究领域应用广泛,被称为图像算法性能检验的另一标准数据集.
- (6) Karlsruhe Institute of Technology and Toyota Technological Institute(KITTI)^[104].KITTI 是目前国际上用于自动驾驶场景检测的最大评测数据集,主要用于评测车载环境下路面分割、目标检测、目标跟踪等技术.KITTI 数据集包含市区、乡村和高速公路等真实场景图像,每张图像中含有不同程度的遮挡现象.
- (7) Cityscapes Dataset(简称 Cityscapes)^[105].Cityscapes 主要提供无人驾驶环境下的图像分割数据,用于评估算法在城区场景语义理解方面的性能.Cityscapes 包含 50 个城市不同环境、不同背景、不同季节的街道场景,提供 5 000 张精细标注的图片、20 000 张粗略标注的图片和 30 种标注物体.
- (8) Sift Flow^[106].该数据集以户外场景解析类的图片为主,包含街道、山脉、城市、海滩和建筑等 8 种户外类型场景,共有 2 688 张图片、33 种语义类目标和 3 种地理类目标.图片都具有像素级标注,分辨率为 256×256.
- (9) Stanford background dataset(简称 SBD)^[107].SBD 由斯坦福大学建立,用于衡量语义级场景解析算法的性能.该数据集包含 725 张图片,分别从 LabelMe、PASCAL VOC 等数据集中抽取而来.图片大多为户外场景类型,大小较为规整,每张图片至少包含 1 个前景对象.
- (10) NYU Depth Dataset V2(简称 NYUDv2)^[108].NYUDv2 中的图片大都出自微软 Kinect 数据库,分为 RGB 图像、深度图像和 RDB-D 图像这 3 个子数据集.该数据集由一系列表示各种室内场景的视频序列组成,总共包含 1 449 张 RDB-D 图像、26 种场景类型、464 种室内场景和 407 024 帧经过标记的图像数据.
- (11) SUN-RGBD^[109].SUN-RGBD 中的图像大多由 4 个不同的 RGB-D 图像传感器捕获而得,共包含 10 000 张 RGB-D 图像、146 617 个多边形标注、58 657 个边界框标注以及大量的空间布局信息和种类信息.

4.2 实验性能评价指标

在 ISS 领域,常用的性能评价指标主要包括平均召回率(average recall,简称 AR)^[110]、平均精度(average precision,简称 AP)^[110]、平均精度均值(mean average precision,简称 mAP)^[110]、像素准确率(pixel accuracy,简称 PA)^[23]、平均准确率(mean accuracy,简称 MA)^[23]、平均交并比(mean intersection over union,简称 mIoU)^[23]和带权交并比(frequency weighted intersection over union,简称 FWIoU)^[23].在结果评价时,一般选取 PA、MA 和 mIoU 这 3 种评价指标综合分析.其中,mIoU 表示分割结果与其真值的重合度,是目前 ISS 领域使用频率最高和最常见的评价指标.PA、MA 和 mIoU 的具体定义及计算公式如公式(1)~公式(3)所示.

(1) PA 用于计算正确分割的像素数量与图像像素总量的比例,其具体计算方法如公式(1)所示.

$$PA = \left(\sum_{i=1}^N X_{ii} \right) / \left(\sum_{i=1}^N T_i \right)$$

(1)

(2) MA 表示所有类别物体像素准确率的平均值,其具体计算方法如公式(2)所示.

$$MA = \left(\sum_{i=1}^N \frac{X_{ii}}{T_i} \right) / N$$

(2)

(3) mIoU 表示分割结果与原始图像真值的重合程度,其具体计算方法如公式(3)所示.

$$mIoU = \left(\sum_{i=1}^N \frac{X_{ii}}{T_i + \sum_{j=1}^N (X_{ji} - X_{ii})} \right) / N$$

(3)

其中, N 代表图像像素的类别数量; T_i 代表第 i 类的像素总数; X_{ii} 代表实际类型为 i 、预测类型为 i 的像素总数; X_{ji} 代表实际类型为 i 、预测类型为 j 的像素总数.

4.3 实验结果分析与对比

为便于说明算法效果,本节将按照图 1 中的分类对 ISSbRC、ISSbFSL 和 ISSbWSL 这 3 类方法的实验结果分别进行分析对比.事实上,随着计算机硬件技术的发展和数据处理能力的提高,当前 ISSbDL 领域中,大多数研究都以提升算法的“分割准确率”为研究重点,而对“计算性能”关注不多.特别是上述 3 类方法中的 ISSbRC 方法和 ISSbWSL 方法,由于其功能侧重点、应用场景和改进方式等原因,相关研究大都忽略了对算法“计算性能”的实验考察.因此,本节对这两类方法的效果仅从算法“分割准确率”方面进行分析对比.部分 ISSbFSL 方法因涉及到动态场景解析或实时图像语义分割等研究领域,比较重视算法的“计算性能”指标.因此,针对 ISSbFSL 方法,本节将从算法“分割准确率”与“计算性能”两个方面进行实验结果分析.

4.3.1 ISSbRC 方法的实验对比分析

ISSbRC 方法的实验对比见表 4,主要比较因素有关键技术、实验数据集和评价指标等.

Table 4 Experimental comparison of ISSbRC
表 4 基于区域分类的图像语义分割方法实验对比

| 分类 | 文献 | 方法名称 | 发表年份 | 关键技术 | 数据集 | 评价指标 | 数值(%) |
|-----------|------|-----------|------|-----------------|-----------------|------|-------|
| 基于候选区域的方法 | [12] | RCNN | 2014 | 选择搜索,候选区域 | PASCAL VOC 2011 | mIoU | 47.9 |
| | [13] | SDS | 2014 | MCG算法,候选区域 | PASCAL VOC 2011 | mIoU | 52.6 |
| | [14] | MPA | 2016 | 滑动窗口,尺度对齐 | PASCAL VOC 2012 | mIoU | 62.1 |
| | [15] | Mask-RCNN | 2017 | 候选区域,ROI Align层 | MS COCO | mIoU | 60 |
| 基于分割掩膜的方法 | [16] | DeepMask | 2015 | 检测框,分割掩膜 | MS COCO | AR | 33.1 |
| | [17] | SharpMask | 2016 | 精炼模型,跳跃结构 | MS COCO | AR | 66.4 |
| | [18] | MultiPath | 2016 | 跳跃连接,中心凹区域 | MS COCO | mAP | 45.4 |

从表 4 中可以看到,ISSbRC 方法大多选用 PASCAL VOC 和 MS COCO 数据集作测试数据集,因为这两种数据集更有权威性和说服力.在基于候选区域的方法中,RCNN^[12]的 mIoU 虽然不高,但因其提出时间最早、性能平稳、代码开源,已被实用化.其他算法,如 SDS^[13]、MPA^[14]、Mask-RCNN^[15]等,其框架结构都是在 RCNN 的基础

上改进而来,因此分割准确率逐步提升.基于分割掩膜的方法使用 RCNN 中的技术得到候选区域后,再对像素进行二分类处理生成分割掩膜,分割效果较好.在该类方法中,SharpMask^[17]和 MultiPath^[18]分别对 DeepMask^[16]进行改进,分割性能有较大的提升.其中,SharpMask 将 DeepMask 生成的粗略分割掩膜输入精细模块,进行逐步优化后最终生成精细的分割掩膜,其平均召回率比 DeepMask 提高了近 1 倍.

4.3.2 ISSbFSL 方法的实验对比分析

(1) 针对 ISSbFSL 方法分割准确率的实验结果对比见表 5,主要比较因素有基于的基础网络、关键技术、是否使用 PGM 方法、实验数据集和评价指标等.

Table 5 Experimental comparison of segmentation accuracy for ISSbFSL
表 5 基于像素分类的全监督学习图像语义分割方法的分割准确率实验对比

| 分类 | 文献 | 方法名称 | 年份 | 基础网络 | 关键技术 | PGM | 数据集 | mIoU (%) |
|-------------|------|-----------------|------|--------|-------------------------|-------|-----------------|----------|
| 基于 FCN 的方法 | [23] | FCN | 2014 | VGG-16 | 上采样,Skip Layer | × | PASCAL VOC 2012 | 62.2 |
| | [25] | DeepLab-V1 | 2014 | ResNet | 上采样,结构预测 | CRF | PASCAL VOC 2012 | 71.6 |
| | [26] | DeepLab-V2 | 2016 | ResNet | 带孔卷积,上采样,ASPP | CRF | PASCAL VOC 2012 | 79.7 |
| | [27] | DeepLab-V3 | 2017 | ResNet | 改进带孔卷积,改进 ASPP | CRF | PASCAL VOC 2012 | 86.9 |
| | [28] | CRFasRNN | 2015 | FCN-8s | 将 CRF 建模为 RNN 后嵌入 CNN 中 | CRF | PASCAL VOC 2012 | 74.7 |
| 基于优化卷积结构的方法 | [29] | Dilation10 | 2015 | VGG-16 | 扩张卷积&特征融合 | × | PASCAL VOC 2012 | 75.3 |
| | [30] | DUC+HDC | 2017 | ResNet | DUC+HDC | × | PASCAL VOC 2012 | 83.1 |
| | [31] | deformable | 2017 | ResNet | 可变形卷积 | N/A | PASCAL VOC 2012 | 75.3 |
| 基于编码解码的方法 | [32] | Bayesian SegNet | 2015 | FCN | 反卷积,上采样,DropOut 层 | × | CityScapes | 57.0 |
| | [33] | DeconvNet | 2015 | FCN | 反卷积,上池化 | × | PASCAL VOC 2012 | 69.6 |
| | [34] | ENet | 2016 | FCN | 分解滤波器,扩张卷积 | × | CityScapes | 58.3 |
| | [35] | GCN+ | 2017 | ResNet | 大核卷积,全局卷积网络 | × | PASCAL VOC 2012 | 82.2 |
| 基于概率图模型的方法 | [36] | DSM | 2016 | VGG-16 | 通过 CNN 建模 CRF | CRF | PASCAL VOC 2012 | 78.0 |
| | [37] | C&G | 2016 | FCN | 将 CRF 嵌入到 CNN | CRF | PASCAL VOC 2012 | 78.1 |
| | [38] | DPN | 2015 | VGG-16 | 将 CNN 与 MRF 融合 | MRF | PASCAL VOC 2012 | 77.5 |
| | [39] | QO | 2016 | ResNet | 二次优化 | G-CRF | PASCAL VOC 2012 | 80.2 |
| | [40] | HOCRF+ | 2016 | VGG-16 | 将 CRF 嵌入到 CNN | HOCRF | PASCAL VOC 2012 | 77.9 |
| 基于特征融合的方法 | [43] | LRR | 2016 | ResNet | 拉普拉斯金字塔 | × | PASCAL VOC 2012 | 76.8 |
| | [44] | RefineNet | 2016 | ResNet | 多路径优化,精炼模块 | × | PASCAL VOC 2012 | 83.4 |
| | [45] | PSPNet | 2016 | ResNet | 多尺度特征整合,空间金字塔池化 | × | PASCAL VOC 2012 | 85.4 |
| | [46] | ICNet | 2017 | ResNet | 级联模型,特征融合 | × | Cityscapes | 69.5 |
| | [47] | LC | 2017 | ResNet | 级联模型,区域卷积 | × | PASCAL VOC 2012 | 82.7 |
| 基于 RNN 的方法 | [51] | rCNN | 2014 | RNN | 多尺寸输入窗口 | × | SIFT Flow | N/A |
| | [52] | ReSeg | 2016 | ResNet | 将 ReNet 进行功能扩展 | × | CamVid | N/A |
| | [53] | DAG-GNN | 2016 | RNN | 使用 PGM 建模图像结构 | × | SIFT Flow | N/A |
| | [54] | 2D-LSTM | 2015 | RNN | 4 个不同方向的 RNN | × | SIFT Flow | N/A |
| 基于 GAN 的方法 | [55] | N/A | 2016 | CNN | GAN,对抗训练 | × | PASCAL VOC 2012 | 54.3 |
| | [58] | N/A | 2016 | FCN | GAN,域适应 | × | Cityscapes | 67.8 |

注:该表中的 N/A 表示相关论文未提及或无法复现该项,×表示此方法没有使用 PGM

从表 5 中可以看到,根据算法的应用场景和分割特色不同,选用的数据集也不同.当对常规静态图像进行图像语义分割时,大多选用 PASCAL VOC 2012 作为测试数据集;当进行动态场景解析或实时图像语义分割时,大多选用 CityScapes 作为测试数据集.

该类方法中,DeepLab-V3^[27]、PSPNet^[45]、RefineNet^[44]、DUC+HDC^[30]、LC^[47]、GCN+^[35]和 QO^[39]等算法在 PASCAL VOC 2012 数据集上的 mIoU 都超过了 80%,对图像中不同尺度的物体有较好的识别效果,分割结果的边界比较接近真实分割边界,是最具代表性的图像语义分割算法.其中,DeepLab-V3 算法因为集成了 FCN^[23]、PSPNet 和 DeepLab-V2 等众多网络的优点,其 mIoU 指标目前排名最高.而 PSPNet 与 RefineNet 通过多路径、多尺度方式对图像特征进行融合,可有效捕捉图像中丰富的上下文信息,在 mIoU 指标上分别排名第二和第三.

CRFasRNN^[28]、Dilation10^[29]、DeepLab-V1^[25]和 DeepLab-V2^[26]等算法则是基于 FCN 进行改进,在 PASCAL VOC 2012 数据集上的 mIoU 都超过 70%,在分割准确率方面与 FCN 相比有较大提升.其中,DeepLab-V2 由于具有代码开源时间早、性能稳定和分割准确率高等优点,在工业界备受青睐,被广泛用于分割静态图像,其 mIoU 达到了 79.7%.

其中,SegNet^[32]、ENet^[34]和 ICNet^[46]这 3 种算法由于主要用于无人驾驶、在线视频处理等领域,故而选择在满足实时图像语义分割性能测试的 CityScapes 数据集进行实验.实验结果表明,这 3 种算法的 mIoU 都超过了 50%,分割精度基本满足对街道场景图像进行语义分割的要求.其中,ICNet 将不同尺度的图像放在不同深度的神经网络中处理,并使用级联特征融合单元融合不同分辨率的特征图,在 CityScapes 上的 mIoU 为 69.5%,分割准确率相对于 SegNet 和 ENet 有明显提升,分割性能突出.

(2) 针对算法的计算性能,基于 Cityscapes 基准测试和 PASCAL VOC 基准测试中的有关内容^[99,105]以及相关参考文献^[23,46],我们从 ISSbFS 方法中选择了代表性较强、相关度较高的几种经典算法进行分析对比.各算法计算性能的实验测试均在 Cityscapes 数据集上进行,其测试结果见表 6,主要比较因素有算法名称、发表年份、运行时间和每秒帧数等,其中,“运行时间”代表分割一张图像所消耗的时间,“每秒帧数”代表每秒能够分割的图像总数量.

Table 6 Experimental comparison of computational performance for ISSbFSL
表 6 基于像素分类的全监督学习图像语义分割方法的计算性能实验对比

| 文献 | 方法名称 | 发表年份 | 运行时间(s) | 每秒帧数(fps) |
|------|------------|------|---------|-----------|
| [34] | ENet | 2016 | 0.013 | 76.9 |
| [46] | ICNet | 2017 | 0.033 | 30.3 |
| [32] | SegNet | 2015 | 0.06 | 16.7 |
| [23] | FCN-8s | 2014 | 0.5 | 2 |
| [28] | CRFasRNN | 2015 | 0.7 | 1.4 |
| [30] | DUC+HDC | 2017 | 0.9 | 1.1 |
| [44] | RefineNet | 2016 | 1.2 | 0.86 |
| [45] | PSPNet | 2016 | 2.2 | 0.45 |
| [38] | DPN | 2015 | 2.5 | 0.4 |
| [29] | Dilation10 | 2015 | 4.0 | 0.25 |
| [25] | DeepLab-V1 | 2014 | 4.0 | 0.25 |
| [26] | DeepLab-V2 | 2016 | 4.0 | 0.25 |

从表 6 中可以看到,各类具体算法的分割速度有较大差异.其中,ENet^[34]、ICNet^[46]和 SegNet^[32]这 3 种算法的运行时间分别为 0.013s、0.033s 和 0.06s,分割速度较快,实时性强,适用于实时图像分割;而 FCN^[23]由于在使用双线性插值算法进行上采样过程中耗时较长,导致分割速度不高,其运行时间为 0.5s,无法满足实时图像分割的需求;DeepLab-V1^[25]和 DeepLab-V2^[26]由于在利用 PGM 对图像进行结构化预测过程中计算较为复杂、耗时较长,导致其分割速度较低,也无法满足实时图像分割的需求;其他算法的分割速度都比 FCN 要低,也同样无法满足实时图像分割的需求,不适用于在线视频处理和动态场景解析等任务.

4.3.3 ISSbWSL 方法的实验对比分析

ISSbWSL 方法的实验结果对比见表 7,主要比较因素有关键技术、监督信息、是否使用 CRF 方法、实验数据集和评价指标等.在这些方法中,BoxSup^[83]、ScribbleSup^[87]和 WeaklySemi^[98]这 3 种方法在 PASCAL VOC

2012 数据集上的 mIoU 都超过了 70%,分割准确率较高,是以弱监督学习方式进行图像语义分割的典型算法.而基于图像级标注的方法因为仅仅使用带有种类标注的弱标注数据进行弱监督训练,分割效果不明显,分割边界粗糙且不连续,mIoU 都普遍较低.

Table 7 Experimental comparison of ISSbWSL
表 7 基于像素分类的弱监督学习图像语义分割方法实验对比

| 分类 | 文献 | 方法名称 | 年份 | 关键技术 | 监督信息 | CRF | 数据集 | mIoU (%) |
|--------------|------|-------------|------|------------|-------------|-----|-----------------|----------|
| 基于边框级标注的方法 | [83] | BoxSup | 2015 | MCG 算法 | 边框级 | × | PASCAL VOC 2012 | 75.2 |
| | [84] | DeepCut | 2016 | CRF | 边框级 | ✓ | N/A | N/A |
| 基于涂鸦级标注的方法 | [86] | WTP | 2015 | Objectness | 涂鸦级 | × | PASCAL VOC 2012 | 49.1 |
| | [87] | ScribbleSup | 2016 | 超像素 | 涂鸦级 | ✓ | PASCAL VOC 2012 | 71.3 |
| 基于图像级标注的方法 | [88] | MIL | 2015 | MCG 算法 | 图像级 | × | ImageNet | 42.0 |
| | [90] | CCNN | 2015 | Class Size | 图像级 | × | PASCAL VOC 2012 | 42.4 |
| | [91] | SEC | 2016 | 显著性检测算法 | 图像级 | ✓ | PASCAL VOC 2012 | 50.7 |
| | [92] | STC | 2015 | 显著性检测算法 | 图像级 | ✓ | PASCAL VOC 2012 | 49.8 |
| | [93] | AugFeed | 2016 | MCG 算法 | 图像级 | ✓ | PASCAL VOC 2012 | 54.34 |
| | [94] | EM | 2017 | 显著性检测 | 图像级 | ✓ | PASCAL VOC 2012 | 58.71 |
| | [97] | Decoupled | 2015 | N/A | 图像级、像素级 | ✓ | PASCAL VOC 2012 | 66.6 |
| 多种弱标注数据混合的方法 | [98] | WeaklySemi | 2015 | N/A | 图像级、边框级、像素级 | ✓ | PASCAL VOC 2012 | 73.9 |

注:该表中的 N/A 表示相关论文未提及或无法复现该项,✓表示此方法使用过 CRF,×表示没有使用 CRF

5 总结与展望

如今,深度学习技术已经广泛应用到图像语义分割领域.本文主要对基于深度学习的图像语义分割的经典方法与研究现状进行了较为细致的分类、梳理与总结.根据分割特点和处理粒度不同,将基于深度学习的图像语义分割方法分为基于区域分类的图像语义分割方法和基于像素分类的图像语义分割方法,把基于像素分类的图像语义分割方法进一步细分为全监督学习图像语义分割方法和弱监督学习图像语义分割方法.对每类方法的代表性算法进行了研究、分析和对比,并概括总结了每类方法的技术特点和优缺点.在现有研究成果的基础上,我们总结 ISS 研究领域的重点问题和发展趋势,认为该领域还存在如下一些具有挑战性的研究方向.

(1) 应用于场景解析任务的图像语义分割

场景解析任务处理的图像背景复杂、环境多变,现有 ISSbDL 方法无法有效地捕获图像的上下文信息和深度语义信息,在识别和分割图像中目标物体时仍存在较大的困难.文献[111]把迁移学习的思想引入场景解析任务,将图像像素特征与词汇概念相结合,提出一个开放式词汇解析网络(the open vocabulary parsing network,简称 OVPN).文献[112]提出一个针对该任务的语境循环残差网络(contextual recurrent residual network,简称 CRRN),通过继承序列模型和残差学习,建模远程语境依赖、学习视觉特征.这些方法目前都存在难以选择标注基元量化级别、未充分利用场景几何深度等问题,如何解决这些场景解析中的问题并实现有效分割是一个挑战.

(2) 实例级图像语义分割

实例级图像语义分割,有时也称为实例分割(instance segmentation,简称 IS),融合了分割与检测两个功能,可以分割出图像中同类物体的不同实例.文献[113]将多任务学习(multi-task learning)^[114]引入分割领域实现实例分割,其分割过程分为 3 个能够共享卷积特征的子任务,将上一任务的输出作为下一任务的输入,分割时,能够区分出不同的实例对象.文献[13,15]对 RCNN 进行改进后,既能用于 ISS,又能用于 IS.文献[115,116]对 FCN 进行改进,使用滑动窗口或物体框将不同的位置信息编码到特征图中,对每个实例进行语义分割.文献[117]在图像中使用聚类的方法构建分割树,并探索不同的实例.文献[118]使用多示例学习方法结合弱监督学习进行 IS.文献[119]使用一个可逆的 RNN 处理 IS 问题.这些方法在分割准确率和算法综合性能上都有很大的提升空间,如何平衡分割效果与时间复杂度,也是目前亟需解决的问题.

(3) 实时图像语义分割

实时图像语义分割以极高的分割速率处理图像或视频数据,并分析利用各图像(帧)之间的时空关系,是一种以高分割速率运行的 ISS 机制.文献[34]基于编码器-解码器结构,采用分解滤波器策略,使用低阶近似将卷积操作分解为更简单的操作,降低了计算量,初步实现了实时分割.文献[46]采用逐步提高分割精度的策略,逐渐减少图像经过的网络层数,利用级联特征融合单元来融合高分辨率与低分辨率图像的特征,提高了分割速度,基本达到了实时分割.实时图像语义分割常被应用于视频跟踪和多目标定位等任务,有巨大的商业价值,但目前的 ISS 方法大多无法满足实时分割的速度要求.探索如何进一步提高实时图像语义分割的速度与精度,是该领域的一个研究热点.

(4) 应用于三维数据的语义分割

目前,大多数 ISS 算法以处理静态图片数据为主,而针对点云、多边形网格等三维数据的分割方法却较少.文献[120–122]尝试使用三维卷积神经网络(3D convolutional neural network,简称 3D-CNN)对三维数据进行处理.文献[123]则对 3D-CNN 进行改进,设计了一个能够标注点云数据并进行语义分割的 3D-CNN.文献[124]直接以未排序的点云作为输入数据,提出一个能够直接对点云数据进行语义分割的 PointNet 网络.三维数据语义分割技术是近年来兴起的一项热门研究,由于三维数据的无序性和非结构化本质,如何合理离散化和结构化这些数据并有效地保留其空间位置信息,仍是一个有待解决的问题.而且,由于三维数据语义分割任务严重依赖大量数据集,如何大规模地获取三维数据并建立相关公共数据集,也是研究者们要努力的一个方向.

(5) 应用于视频数据的语义分割

视频语义分割是一种基于三维空间的 ISS 问题.目前,针对视频数据的语义分割方法较少.带有时间序列的视频数据在语义分割过程中能充分利用二维图像中的隐含信息,可更好地显示二维图像无法展现的时空特征.文献[125]基于 FCN 提出一种有效利用时空信息进行视频语义分割的循环全卷积网络(recurrent fully convolutional network,简称 RFCN).文献[126]将视频数据中的空间特征融入 FCN,提出一种融合时空特征的时空全卷积网络(spatio-temporal fully convolutional network,简称 STFCN).文献[127]则设计了一个定时全卷积网络(clockwork FCN),使用自适应时钟信号操纵定时卷积驱动进行视频分割.未来,研究如何充分利用视频丰富的时空序列特征具有重要意义;同时,如何从视频高效抽取高层语义信息也是一个难点.

致谢 在此,我们向对本文提出宝贵修改意见的各位同行及评审专家表示感谢.

References:

- [1] Csurka G, Perronnin F. An efficient approach to semantic segmentation. *Int'l Journal of Computer Vision*, 2011,95(2):198–212. [doi: 10.1007/s11263-010-0344-8]
- [2] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science*, 2006,313(5786):504–507. [doi: 10.1126/science.1127647]
- [3] Jiang F, Gu Q, Hao HZ, Li N, Guo YW, Chen DX. Survey on content-based image segmentation methods. *Ruan Jian Xue Bao/ Journal of Software*, 2017,28(1):160–183 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5136.htm> [doi: 10.13328/j.cnki.jos.005136]
- [4] Wei YC, Zhao Y. A review on image semantic segmentation based on DCNN. *Journal of Beijing Jiaotong University*, 2016,40(4): 82–91 (in Chinese with English abstract). [doi: 10.11860/j.issn.1673-0291.2016.04.013]
- [5] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 1998, 86(11):2278–2324. [doi: 10.1109/5.726791]
- [6] Pearlmutter BA. Gradient calculations for dynamic recurrent neural networks: A survey. *IEEE Trans. on Neural Networks*, 1995, 6(5):1212–1228. [doi: 10.1109/72.410363]
- [7] Goodfellow IJ, Pougetabadi J, Mirza M, Xu B, Wardefarley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. In: *Advances in Neural Information Processing Systems*. 2014. 2672–2680.

- [8] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997,9(8):1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
- [9] Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [10] Couprie C, Farabet C, Najman L, Lecun Y. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.
- [11] Farabet C, Couprie C, Najman L, Lecun Y. Learning hierarchical features for scene labeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013,35(8):1915–1929. [doi: 10.1109/TPAMI.2012.231]
- [12] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]
- [13] Hariharan B, Arbeláez P, Girshick R, Malik J. Simultaneous detection and segmentation. In: *Proc. of the European Conf. on Computer Vision*. Springer-Verlag, 2014. 297–312.
- [14] Liu S, Qi X, Shi J, Zhang H, Jia J. Multi-scale patch aggregation (MPA) for simultaneous detection and segmentation. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2016. 3141–3149. [doi: 10.1109/CVPR.2016.342]
- [15] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *Proc. of the IEEE Int'l Conf. on Computer Vision*. 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]
- [16] Pinheiro PO, Collobert R, Dollár P. Learning to segment object candidates. In: *Proc. of the Advances in Neural Information Processing Systems*. 2015. 1990–1998.
- [17] Pinheiro PO, Lin TY, Collobert R, Dollár P. Learning to refine object segments. In: *Proc. of the European Conf. on Computer Vision*. Springer-Verlag, 2016. 75–91.
- [18] Zagoruyko S, Lerer A, Lin TY, Pinheiro PO, Gross S, Chintala S, Dollár P. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016.
- [19] Arbeláez P, Pont-Tuset J, Barron JT, Marques F, Malik J. Multiscale combinatorial grouping. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*. 2014. 328–335. [doi: 10.1109/CVPR.2014.49]
- [20] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015,37(9):1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
- [21] Girshick R. Fast R-CNN. In: *Proc. of the IEEE Int'l Conf. on Computer Vision*. 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]
- [22] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Proc. of the Advances in Neural Information Processing Systems*. 2015. 91–99. [doi: 10.1109/TPAMI.2016.2577031]
- [23] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 2014,39(4):640–651. [doi: 10.1109/TPAMI.2016.2572683]
- [24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [25] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014.
- [26] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 2018,40(4):834–848. [doi: 10.1109/TPAMI.2017.2699184]
- [27] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [28] Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, Huang C, Torr PH. Conditional random fields as recurrent neural networks. In: *Proc. of the IEEE Int'l Conf. on Computer Vision*. 2015. 1529–1537. [doi: 10.1109/ICCV.2015.179]
- [29] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
- [30] Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G. Understanding convolution for semantic segmentation. In: *Proc. of the IEEE Winter Conf. on Applications of Computer Vision*. 2018. 1451–1460. [doi: 10.1109/WACV.2018.00163]

- [31] Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 764–773.
- [32] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015,39(12):2481–2495. [doi: 10.1109/TPAMI.2016.2644615]
- [33] Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1520–1528. [doi: 10.1109/ICCV.2015.178]
- [34] Paszke A, Chaurasia A, Kim S, Culurciello E. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147, 2016.
- [35] Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters—Improve semantic segmentation by global convolutional network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1743–1751. [doi: 10.1109/CVPR.2017.189]
- [36] Lin G, Shen C, Van Den Hengel A, Reid I. Efficient piecewise training of deep structured models for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3194–3203. [doi: 10.1109/CVPR.2016.348]
- [37] Shen F, Zeng G. Fast semantic image segmentation with high order context and guided filtering. arXiv preprint arXiv:1605.04068, 2016.
- [38] Liu Z, Li X, Luo P, Loy CC, Tang X. Semantic image segmentation via deep parsing network. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1377–1385. [doi: 10.1109/ICCV.2015.162]
- [39] Chandra S, Kokkinos I. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian CRFs. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 402–418.
- [40] Arnab A, Jayasumana S, Zheng S, Torr PH. Higher order conditional random fields in deep neural networks. In: Proc. of the European Conf. on Computer Vision: Springer-Verlag, 2016. 524–540. [doi: 10.1007/978-3-319-46475-6_33]
- [41] Vemulapalli R, Tuzel O, Liu MY, Chellapa R. Gaussian conditional random field network for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3224–3233. [doi: 10.1109/CVPR.2016.351]
- [42] Arnab A, Jayasumana S, Zheng S, Torr PH. Higher order potentials in end-to-end trainable conditional random fields. *Computer Science*, 2015.
- [43] Ghiasi G, Fowlkes CC. Laplacian reconstruction and refinement for semantic segmentation. arXiv preprint arXiv:1605.02264, 2016. [doi: 10.1007/978-3-319-46487-9_32]
- [44] Lin G, Milan A, Shen C, Reid I. RefineNet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. arXiv preprint arXiv:1611.06612, 2016.
- [45] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6230–6239. [doi: 10.1109/CVPR.2017.660]
- [46] Zhao H, Qi X, Shen X, Shi J, Jia J. ICNet for real-time semantic segmentation on high-resolution images. arXiv preprint arXiv:1704.08545, 2017.
- [47] Li X, Liu Z, Luo P, Loy CC, Tang X. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. arXiv preprint arXiv:1704.01344, 2017.
- [48] Chen LC, Yang Y, Wang J, Xu W, Yuille AL. Attention to scale: Scale-aware semantic image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3640–3649. [doi: 10.1109/CVPR.2016.396]
- [49] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 2380–2504. [doi: 10.1109/ICCV.2015.304]
- [50] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2014. 2366–2374.
- [51] Pinheiro P, Collobert R. Recurrent convolutional neural networks for scene labeling. In: Proc. of the Int'l Conf. on Machine Learning. 2014. 82–90.
- [52] Visin F, Ciccone M, Romero A, Kastner K, Cho K, Bengio Y, Matteucci M, Courville A. Reseg: A recurrent neural network-based model for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2016. 41–48. [doi: 10.1109/CVPRW.2016.60]

- [53] Shuai B, Zuo Z, Wang B, Wang G. DAG-recurrent neural networks for scene labeling. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3620–3629. [doi: 10.1109/CVPR.2016.394]
- [54] Byeon W, Breuel TM, Raue F, Liwicki M. Scene labeling with LSTM recurrent neural networks. In: Proc. of the Computer Vision and Pattern Recognition. 2015. 3547–3555. [doi: 10.1109/CVPR.2015.7298977]
- [55] Luc P, Couprie C, Chintala S, Verbeek J. Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408, 2016.
- [56] Hoffman J, Wang D, Yu F, Darrell T. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649, 2016.
- [57] Koziński M, Simon L, Jurie F. An adversarial regularisation for semi-supervised training of structured output neural networks. arXiv preprint arXiv:1702.02382, 2017.
- [58] Souly N, Spampinato C, Shah M. Semi and weakly supervised semantic segmentation using generative adversarial network. arXiv preprint arXiv:1703.09695, 2017.
- [59] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. 2015. 234–241.
- [60] Badrinarayanan V, Handa A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293, 2015.
- [61] Liu JW, Li HE, Luo XL. Learning technique of probabilistic graphical models: A review. Acta Automatica Sinica, 2014,40(6): 1025–1044 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2014.01025]
- [62] Qi X, Shi J, Liu S, Liao R. Semantic segmentation with object clique potential. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 2587–2595. [doi: 10.1109/ICCV.2015.297]
- [63] Lin G, Shen C, Reid I, Van Den Hengel A. Deeply learning the messages in message passing inference. Computer Science, 2015, 71(5):866–872.
- [64] Kohli P, Ladicky LU, Torr PH. Robust higher order potentials for enforcing label consistency. Int'l Journal of Computer Vision, 2009,82(3):302–324. [doi: 10.1007/s11263-008-0202-0]
- [65] Song XY, Zhou LL, Li ZG, Chen J, Zeng L, Yan B. Review on superpixel methods in image segmentation. Journal of Image and Graphics, 2015,20(5):599–608 (in Chinese with English abstract). [doi: 10.11834/jig.20150502]
- [66] Liu W, Rabinovich A, Berg AC. Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579, 2015.
- [67] Burt P, Adelson E. The Laplacian pyramid as a compact image code. IEEE Trans. on Communications, 1983,31(4):532–540. [doi: 10.1109/TCOM.1983.1095851]
- [68] Mnih V, Heess N, Graves A. Recurrent models of visual attention. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 2204–2212.
- [69] Li H, Lin Z, Shen X, Brandt J, Hua G. A convolutional neural network cascade for face detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 5325–5334. [doi: 10.1109/CVPR.2015.7299170]
- [70] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2001. 511–518. [doi: 10.1109/CVPR.2001.990517]
- [71] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261, 2016.
- [72] Raj A, Maturana D, Scherer S. Multi-scale convolutional architecture for semantic segmentation. Technical Report, CMU-RI-TR-15-21, Pittsburgh: Robotics Institute, Carnegie Mellon University, 2015.
- [73] Roy A, Todorovic S. A multi-scale CNN for affordance segmentation in RGB images. In: Proc. of the European Conf. on Computer Vision. Springer Int'l Publishing, 2016. 186–201. [doi: 10.1007/978-3-319-46493-0_12]
- [74] Bian X, Lim SN, Zhou N. Multiscale fully convolutional network with application to industrial inspection. In: Proc. of the IEEE Winter Conf. on Applications of Computer Vision. IEEE Computer Society, 2016. 1–8. [doi: 10.1109/WACV.2016.7477595]
- [75] Li Z, Gan Y, Liang X, Yu Y, Cheng H, Lin L. LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 541–557.

- [76] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 2342–2350.
- [77] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
- [78] McLaughlin N, Rincon JMD, Miller P. Recurrent convolutional network for video-based person re-identification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 1325–1334. [doi: 10.1109/CVPR.2016.148]
- [79] Lei T, Zhang Y. Training RNNs as fast as CNNs. arXiv preprint arXiv:1709.02755, 2017.
- [80] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.
- [81] Im DJ, Kim CD, Jiang H, Memisevic R. Generating images with recurrent adversarial networks. arXiv preprint arXiv:1602.05110, 2016.
- [82] Denton EL, Chintala S, Fergus R. Deep generative image models using a Laplacian pyramid of adversarial networks. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2015. 1486–1494.
- [83] Dai J, He K, Sun J. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1635–1643. [doi: 10.1109/ICCV.2015.191]
- [84] Rajchl M, Lee M, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, Rutherford M, Hajnal J, Kainz B, Rueckert D. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. on Medical Imaging, 2016, 36(2):674–683. [doi: 10.1109/TMI.2016.2621185]
- [85] Boykov YY, Jolly MP. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2001. 105–112. [doi: 10.1109/ICCV.2001.10011]
- [86] Bearman A, Russakovsky O, Ferrari V, Li FF. What's the point: Semantic segmentation with point supervision. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 549–565.
- [87] Lin D, Dai J, Jia J, He K, Sun J. ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2016. 3159–3167. [doi: 10.1109/CVPR.2016.344]
- [88] Pinheiro PO, Collobert R. From image-level to pixel-level labeling with convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1713–1721. [doi: 10.1109/CVPR.2015.7298780]
- [89] Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997,89(1-2):31–71. [doi: 10.1016/S0004-3702(96)00034-3]
- [90] Pathak D, Krahenbuhl P, Darrell T. Constrained convolutional neural networks for weakly supervised segmentation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1796–1804. [doi: 10.1109/ICCV.2015.209]
- [91] Kolesnikov A, Lampert CH. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 695–711. [doi: 10.1007/978-3-319-46493-0_42]
- [92] Wei Y, Liang X, Chen Y, Shen X, Cheng MM, Feng J, Zhao Y, Yan S. STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2017,39(11):2314–2320. [doi: 10.1109/TPAMI.2016.2636150]
- [93] Qi X, Liu Z, Shi J, Zhao H, Jia J. Augmented feedback in semantic segmentation under image level supervision. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 90–105.
- [94] Hou Q, Dokania PK, Massiceti D, Wei Y, Cheng M, Torr P. Bottom-up top-down cues for weakly-supervised semantic segmentation. In: Proc. of the Int'l Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition. Springer-Verlag, 2017. 263–277.
- [95] Durand T, Mordan T, Thome N, Cord M. WILDCAT: Weakly supervised learning of deep ConvNets for image classification, pointwise localization and segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 5957–5966. [doi: 10.1109/CVPR.2017.631]
- [96] Wei Y, Feng J, Liang X, Cheng MM, Zhao Y, Yan S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proc. of the Computer Vision and Pattern Recognition. IEEE, 2017. 6488–6496. [doi: 10.1109/CVPR.2017.687]

- [97] Hong S, Noh H, Han B. Decoupled deep neural network for semi-supervised semantic segmentation. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 1495–1503.
- [98] Papandreou G, Chen LC, Murphy K, Yuille AL. Weakly- and semi- supervised learning of a DCNN for semantic image segmentation. arXiv preprint arXiv:1502.02734, 2015.
- [99] Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes challenge: A retrospective. *Int'l Journal of Computer Vision*, 2015,111(1):98–136. [doi: 10.1007/s11263-014-0733-5]
- [100] Mottaghi R, Chen X, Liu X, Cho NG, Lee SW, Fidler S, Urtasun R, Yuille A. The role of context for object detection and semantic segmentation in the wild. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 891–898. [doi: 10.1109/CVPR.2014.119]
- [101] Chen X, Mottaghi R, Liu X, Fidler S, Urtasun R, Yuille A. Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 1979–1986. [doi: 10.1109/CVPR.2014.254]
- [102] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2014. 740–755.
- [103] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M. Imagenet large scale visual recognition challenge. *Int'l Journal of Computer Vision*, 2015,115(3):211–252. [doi: 10.1007/s11263-015-0816-y]
- [104] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. *The Int'l Journal of Robotics Research*, 2013, 32(11):1231–1237. [doi: 10.1177/0278364913491297]
- [105] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3213–3223.
- [106] Liu C, Yuen J, Torralba A. Nonparametric scene parsing: Label transfer via dense scene alignment. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2009). IEEE, 2009. 1972–1979. [doi: 10.1109/CVPRW.2009.5206536]
- [107] Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. In: Proc. of the 2009 IEEE 12th Int'l Conf. on Computer Vision. IEEE, 2009. 1–8. [doi: 10.1109/ICCV.2009.5459211]
- [108] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2012. 746–760.
- [109] Song S, Lichtenberg SP, Xiao J. Sun RGB-D: A RGB-D scene understanding benchmark suite. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 567–576. [doi: 10.1109/CVPR.2015.7298655]
- [110] Turpin A, Scholer F. User performance versus precision measures for simple search tasks. In: Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2006. 11–18. [doi: 10.1145/1148170.1148176]
- [111] Zhao H, Puig X, Zhou B, Fidler S, Torralba A. Open vocabulary scene parsing. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2018. 2021–2029. [doi: 10.1109/ICCV.2017.221]
- [112] Le T, Duong CN, Han L, Luu K, Savvides M, Pal D. Deep contextual recurrent residual networks for scene labeling. *Pattern Recognition*, 2018,80:32–41. [doi: 10.1016/j.patcog.2018.01.005]
- [113] Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3150–3158. [doi: 10.1109/CVPR.2016.343]
- [114] Caruana R. Multitask learning. *Machine Learning*, 1997,28(1):41–75. [doi: 10.1023/A:1007379606734]
- [115] Li Y, Qi H, Dai J, Ji X, Wei Y. Fully convolutional instance-aware semantic segmentation. In: Proc. of the Computer Vision and Pattern Recognition. 2017. 4438–4446. [doi: 10.1109/CVPR.2017.472]
- [116] Dai J, He K, Li Y, Ren S, Sun J. Instance-sensitive fully convolutional networks. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 534–549.
- [117] Silberman N, Sontag D, Fergus R. Instance segmentation of indoor scenes using a coverage loss. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2014. 616–631.
- [118] Chen YT, Liu X, Yang MH. Multi-instance object segmentation with occlusion handling. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3470–3478. [doi: 10.1109/CVPR.2015.7298969]

- [119] Liang X, Wei Y, Shen X, Jie Z, Feng J, Lin L, Yan S. Reversible recursive instance-level object segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 633–641. [doi: 10.1109/CVPR.2016.75]
- [120] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3D shapenets: A deep representation for volumetric shapes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1912–1920. [doi: 10.1109/CVPR.2015.7298801]
- [121] Prokhorov D. A convolutional learning system for object classification in 3-D LIDAR data. IEEE Trans. on Neural Networks, 2010,21(5):858–863. [doi: 10.1109/TNN.2010.2044802]
- [122] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3D shape recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 945–953. [doi: 10.1109/ICCV.2015.114]
- [123] Huang J, You S. Point cloud labeling using 3D convolutional neural network. In: Proc. of the Int'l Conf. on Pattern Recognition (ICPR). IEEE, 2016. [doi: 10.1109/ICPR.2016.7900038]
- [124] Qi CR, Su H, Mo K, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 77–85. [doi: 10.1109/CVPR.2017.16]
- [125] Valipour S, Siam M, Jagersand M, Ray N. Recurrent fully convolutional networks for video segmentation. In: Proc. of the Applications of Computer Vision. 2017. 29–36. [doi: 10.1109/WACV.2017.11]
- [126] Fayyaz M, Saffar MH, Sabokrou M, Fathy M, Klette R, Huang F. STFCN: Spatio-temporal FCN for semantic video segmentation. arXiv preprint arXiv:1608.05971, 2016.
- [127] Shelhamer E, Rakelly K, Hoffman J, Darrell T. Clockwork convnets for video semantic segmentation. In: Proc. of the European Conf. on Computer Vision. Springer-Verlag, 2016. 852–868. [doi: 10.1007/978-3-319-49409-8_69]

附中文参考文献:

- [3] 姜枫,顾庆,郝慧珍,李娜,郭延文,陈道蓄.基于内容的图像分割方法综述.软件学报,2017,28(1):160–183. <http://www.jos.org.cn/1000-9825/5136.htm> [doi: 10.13328/j.cnki.jos.005136]
- [4] 魏云超,赵耀.基于 DCNN 的图像语义分割综述.北京交通大学学报,2016,40(4):82–91. [doi: 10.11860/j.issn.1673-0291.2016.04.013]
- [61] 刘建伟,黎海恩,罗雄麟.概率图模型学习技术研究进展.自动化学报,2014,40(6):1025–1044. [doi: 10.3724/SP.J.1004.2014.01025]
- [65] 宋熙煜,周利莉,李中国,陈健,曾磊,闫镜.图像分割中的超像素方法研究综述.中国图像图形学报,2015,20(5):599–608. [doi: 10.11834/jig.20150502]



田莹(1976—),女,山东济宁人,博士,副教授,CCF 高级会员,主要研究领域为智能信息处理,数据挖掘,机器学习.



丁琪(1996—),女,硕士生,CCF 学生会员,主要研究领域为智能信息处理,数据挖掘,机器学习.



王亮(1992—),男,硕士生,CCF 学生会员,主要研究领域为数据挖掘,机器学习,智能信息处理.