面向复杂环境的图像语义分割方法综述

王嫣然[1] 陈清亮[1] 吴俊君[2]
(暨南大学信息科学技术学院 广州 510632)[1] (佛山科学技术学院机电工程学院 广东 佛山 528225)[2]

摘 要 图像语义分割是视觉智能方向最重要的基础性技术之一,语义分割效果关系着智能系统对其应用场景的理解能力,因此在诸如无人驾驶、机器人认知与导航、安防监控与无人机着陆系统等重要领域均具有较大的应用价值。由于复杂环境下的目标存在非结构化、目标多样化、形状不规则化以及光照变化、视角变化、尺度变化与物体遮挡等各种干扰因素,给图像的语义分割带来了较大挑战。近年来,受益于深度学习理论的快速发展,图像语义分割方向涌现了一大批具有典型意义的研究成果。为启发图像语义分割领域的学术研究及其相关智能系统的工程化开发,文中首先全面阐述了图像语义分割方法的研究发展历程,并将其划分为:传统的图像语义分割方法、传统方法与深度学习相结合的图像语义分割方法、基于深度学习的图像语义分割方法;其次从复杂环境下图像语义分割面临的问题出发,重点对近年来涌现的各种面向复杂环境的语义分割方法的模型、算法、性能及存在的问题进行了详细地分析与对比,并按照强监督、弱监督、无监督图像语义分割方法分类进行阐述;然后归纳了当前主流的PASCAL VOC,Cityscape,SUN RGB-D等9类包含各种复杂环境的数据集,以及3项评估指标PA,mPA和mIoU;最后对面向复杂环境的图像语义分割研究工作进行了总结,并对其在实时视频分割、三维场景重构及无监督语义分割等方向的发展进行了展望。

关键词 语义分割,视觉智能,深度学习,图像分割,卷积神经网络

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.005

Research on Image Semantic Segmentation for Complex Environments

WANG Yan-ran[1] CHEN Qing-liang[1] WU Jun-jun[2]
(College of Information Science and Technology, Jinan University, Guangzhou 510632, China)[1]
(School of Mechatronics Engineering, Foshan University, Foshan, Guangdong 528225, China)[2]

Abstract Image semantic segmentation is one of the most important fundamental technologies for visual intelligence. Semantic segmentation can greatly enable intelligent systems to understand their surrounding scenarios, so it has enormous value in application domains such as unmanned vehicles, robot cognition and navigation, video surveillance and drone landing systems. Great challenges also exist in the semantic segmentation of images, due to various interfering factors of targets in complex environments, such as unstructured targets, diversity of objectives, irregular shapes, illumination changes, different viewing angles, scale variation, object occlusion, etc. In recent years, benefiting from the great advancements in deep learning techniques, a large number of research approaches with practical significance emerge in image semantic segmentation. For having a comprehensive survey and inspiring the academic research, this paper extensively discussed the existing state-of-the-art image semantic segmentation methods, and further classified them into the traditional image semantic segmentation ones, the ones combining traditional and deep learning techniques, and those based purely on deep learning. In order to address these problems in complex environments, various semantic segmentation methods for complex environment emerged in recent years were analyzed and compared in detail, including the models, algorithms and performance with the category of strong supervised, weak supervised and unsupervised semantic segmentation methods. Furthermore, the current main datasets such as PASCAL VOC, Cityscape, SUN RGB-D, which contains various complex environments and 3 evaluation indicators of PA, mPA, mIoU were summarized. Finally, the existing research of image semantic segmentation for complex environment was summarized, and its future trends were prospected such as optimization in real-time video, 3d scene reconstruction and unsupervised semantic segmentation techniques.

Keywords Semantic segmentation, Visual intelligence, Deep learning, Image segmentation, Convolutional neural network

---

到稿日期:2019-02-17 返修日期:2019-04-19 本文受国家自然科学基金(61603103,61673125),广东省自然科学基金(2016A030310293),广州市科技计划科学研究专项(201707010013)资助。

王嫣然(1995—),女,硕士,CCF学生会员,主要研究方向为计算机视觉、图像语义分割;**陈清亮**(1980—),男,博士后,教授,主要研究方向为人工智能;**吴俊君**(1981—),男,博士,副教授,CCF会员,主要研究方向为移动机器人和视觉智能研究,E-mail:jjunwu@fosu.edu.cn(通信作者)。

1　引言

图像语义分割(Image Semantic Segmentation)是指按照图像中每个像素点所表达的语义内容对其进行分类的图像分割技术。它是场景理解的基础性技术，对智能驾驶、机器人认知层面的自主导航、无人机着陆系统以及智慧安防监控等无人系统具有至关重要的作用。

由于传统语义分割方法在非结构化复杂环境下的场景理解能力及工作效率欠佳，近年来面向复杂环境的语义分割问题成为了研究热点，并取得了一系列显著成果。本文从图像语义分割的研究发展历程、原理、模型、方法性能以及存在的问题等角度，对近年来涌现的一批典型成果进行详细的综合性分析与评测总结。

已有文献对图像语义分割成果进行了综述，如文献[1-5]主要对传统图像分割方法进行综述，为利用传统方法进行语义分割提供了思路；文献[6]评述了基于内容的图像语义分割方法；文献[7]对语义分割发展历程及现状进行了简要阐述。但是，针对近年来出现的一批基于深度学习理论的面向复杂环境的语义分割方法尚未有全面的综述性文献，因此本文进行了相关工作，本文对促进图像语义分割研究及相关应用开发均具有积极意义。

图像语义分割效果直接关系到无人系统对场景理解的准确度。目前深度学习技术在图像语义分割领域取得了可喜的进展，但是复杂环境的非结构化、目标多样化、形状不规则化以及光照变化、物体遮挡等各种因素，都给语义分割精度带来了极大的挑战。下面列举了几类常见的语义分割问题，具体如图 1 所示。

(1)较难分割小目标物体，以及目标物体的较小条状区域。例如，室内场景下桌椅的脚，以及道路场景下电线杆、路灯等较细的条状部分。

(2)较难区分具有相似外观的不同目标，以及具有不同外观的同一目标。例如，与树具有相似纹理及外观的地面被误分为树木一类。

(3)对复杂环境下光照、季节变化的适应能力不强，鲁棒性欠佳。例如，将道路旁的光照阴影误分为草地，或不能正确分割不同季节的同一场景。

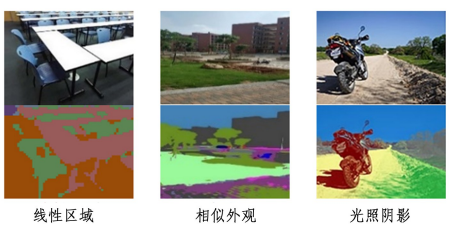

图 1　语义分割问题的示例

Fig. 1　Examples of semantic segmentation challenges

本文第 1 节对图像语义分割技术进行了概述；第 2 节详述了面向复杂环境的图像语义分割技术的研究与发展历程，并将其依次划分为传统语义分割方法、结合了深度学习技术的传统语义分割方法、基于深度学习的语义分割方法 3 个阶段，每个阶段均选取了几种典型的方法进行评述；第 3 节重点对当前的研究热点、基于深度学习的语义分割方法进行了分析、测评与对比总结，将它们分为有监督、弱监督与无监督 3 种类别，并针对每种类别选取了几种代表性方法，对其网络结构及模型原理、特点和不足等方面进行评述；第 4 节对当前方法的实验方案、测试数据集以及评估指标进行了归纳，并对一系列主流方法的技术特性与工作性能进行了测评与分类总结；最后，对本文研究方向的发展趋势进行了展望。

2　图像语义分割发展历程

图像语义分割技术的发展先后经历了 3 个时期。

(1)传统语义分割方法时期。早期受到计算能力的制约，图像分割方法主要依赖于图像颜色、纹理和形状等简单的浅层特征[8]，因此分割精度不高且分割结果没有语义标注。

(2)传统方法与深度学习相结合的语义分割方法时期。该类分割方法与目标检测相似，即先利用传统的图像分割方法进行处理，得到 patch 级的分割结果，然后利用卷积神经网络训练一个特征分类器，对分割后的 patch 进行分类，从而得到语义分割结果。此类方法的精度仍然受到传统分割方法的制约。

(3)基于深度学习的语义分割方法时期。由于深度学习技术在鲁棒性特征的自主学习与分类等方面表现出了强大的能力，因此直接采用卷积神经网络(Convolution Neural Network，CNN)[9]训练出语义分割结果逐渐成为当前的主流方法，并且取得了比前两类方法更好的效果。该类方法也是本文重点讨论的内容。

2.1　传统图像语义分割

传统图像语义分割根据灰度、色彩、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得目标与背景分离。这一阶段的方法包括基于阈值的分割[10]、基于边缘的分割[11]、基于聚类的分割[12]、基于图论的分割[4]以及基于区域的分割[13]。其中最常用的是基于图论的分割。

传统语义分割算法一般采用马尔科夫随机场(Markov Random Fields，MRFs)[14]和条件随机场(Conditional Random Fields，CRFs)[15]来构建概率图模型，并使用图论的方法来求解。MRFs 是传统计算机视觉中普遍应用的模型，属于无向图的概率图模型，其主要思想是为每个特征和像素分配一个随机向量，通过计算每个像素属于每个类的概率来确定该像素的分类。CRFs 在 MRFs 的基础上，对随机向量加入观察值，从本质上讲，CRFs 是给定了观察值集合的 MRFs。

除此之外，传统语义分割的经典之作还包括“Normalized Cut”[16]和“Grab Cut”[17]等基于图论的分割方法。N-Cut 的提出是为了解决最小分割算法[18]在分割时只考虑局部信息的缺点。N-Cut 考虑全局信息进行图划分，可完成多语义的图像分割。Grab Cut 利用了图像中的纹理信息和边界信息，以尽可能少的用户交互得到较好的前景与背景分割结果。基于此，徐海霞等[19]先使用顶点数不断减少的凝聚图序列逼近原图，再利用 N-Cut 方法生成粗粒度分割，最后依据图像统计特征，使用混合模型的最大后验概率优化分割结果，得到细粒度分割结果。刘磊等[20]在 Grab Cut 的基础上引入高阶势能项，来描述像素本身的细节信息以及像素之间的关联信息，并

修改了代价函数,提高了模型的表达精度。

2011 年,Arbeláez 等提出了一种轮廓检测算法 GPB-UCM[21],该算法首先使用 GPB (Globalized Probability of Boundary)方法计算每一个像素作为边缘的概率,接着使用分水岭算法[22]将概率结果转化为多个闭合区域,最后通过 UCM (Ultrametric Contour Map)方法将闭合区域集转化为层次树,以生成轮廓图。2016 年,Zhang 等提出了随机决策森林 (Random Decision Forests)[23]分割方法,该方法中的分类器由多种决策树组合而成。根据以上两种方法,Pont-Tuset 等于 2017 年提出了 MCG (Multiscale Combinatorial Grouping)[24]算法,该方法利用 GPB-UCM 算法得到图像的多个轮廓分割块,再利用随机森林组合分类器得到最终的预测对象,是传统方法的巅峰之作。

2.2 传统方法与深度学习相结合的图像语义分割方法

传统图像语义分割方法利用浅层视觉特征进行图像目标分割,然后使用人工标注语义信息,来完成图像理解任务。随着深度学习技术的兴起,研究者们开始将深度学习模型引入到传统的语义分割方法中,即在利用传统方法分割出目标区域的基础上,进一步采用卷积神经网络等方法学习目标特征并训练分类器,对目标区域进行分类,从而实现目标区域的自动语义标注。

2013 年,Farabet 等首次尝试了结合卷积神经网络的语义分割,提出了第一个将深度学习应用于语义分割的方法[25]。该方法使用拉普拉斯金字塔 (Laplacian Pyramid)[26]得出的图像不同尺度的副本来训练一个多尺度卷积网络,同时通过超像素、分割树获取原始图像分割轮廓,对多尺度卷积网络进行监督学习,最后经过超像素、CRFs、无参数多级解析[27]3 种后处理方法得出最终标记。

同年,Couprie 等使用图像和深度图对室内场景进行语义分割[28]。其提出的算法结构更为简单,即首先将深度图和 RGB 图像进行滤波及卷积特征提取,将不同尺度上的特征图相融合,构成分类器,然后利用该分类器对超像素分割后的 RGB 图像进行分类。然而超像素分割不稳定,当图片中的小块物体或者单个物体中存在杂色时,存在许多分类错误的情况。另外,超像素分割难以处理弱边界区域,因此该方法具有很大的局限性。

2.3 基于深度学习的图像语义分割

1998 年,Lecun 最早提出了 LeNet 网络[29],并设计了卷积神经网络的 3 层结构:卷积层、池化层、非线性层。该结构为深度学习技术在图像领域的成功应用奠定了坚实的理论基础。

当前基于深度学习的图像语义分割方法的主流思想是将图像分类的经典网络作为基网络,根据具体的应用场景,对基网络进行改进并提升语义分割性能,以适应场景理解的需要。

例如:2012 年,Hinton 研究组提出了 AlexNet[30],首创了深度卷积神经网络模型。该网络在 LeNet 的基础上调整了网络架构并加深了网络深度。AlexNet 在当年的 ImageNet[31]竞赛中表现优异并获得了冠军。

2014 年,牛津视觉几何研究团队的 Simonyan 等提出卷积神经网络 VGG[32],赢得了 2014 年 ImageNet 竞赛的冠军。VGG 采用与 AlexNet 相似的 5 层结构,将网络分为 5 组,使用 3×3 过滤器,并将其组合作为一个卷积序列进行处理。VGG 网络与之前模型的主要不同在于:VGG 网络在第 1 层使用了一批小感受野 (Receptive field)[33]的卷积层,使得模型的参数更少,非线性更强,也因此使得决策函数更具区分度,模型更好训练。

为减少神经网络的计算开销,2014 年 Szegedy 等设计了第一个 Inception 架构的网络 GoogLeNet[34]。Inception 的思路是减少每一层的特征过滤器的数目,从而减少运算量。这种新的方法证实了 CNN 层可以有更多的堆叠方式,而不仅仅是标准的序列方式。

此外,2016 年 He 等提出的 ResNet 网络[35]以其高达 152 层的深度以及引入的残差模块而闻名。残差模块使得网络下一层可以同时掌握前一层的输出以及原始的输入,从而调整学习;该模块的连接方式也协助解决了梯度消失问题。

4 种基本网络结构的对比总结如表 1 所列。

表 1 AlexNet、VGG-16、GoogLeNet、ResNet 的对比结果

Table 1 Comparison of AlexNet, VGG-16, GoogLeNet and ResNet

网络	提出时间	层数	卷积层数	卷积核大小	全连接层数	Top-5 准确率/%
AlexNet	2012	8	5	11,5,3	3	84.60
VGG-16	2014	19	16	3	3	92.70
GoogLeNet	2014	22	21	7,1,3,5	1	93.30
ResNet	2016	152	151	7,1,3,5	1	96.40

鉴于神经网络的优良表现,研究学者纷纷将深度学习应用到语义分割领域,随后又提出了 FCN(Fully Convolution Network, FCN)[36], SegNet[37], DeepLab[38], RefineNet[39], PSPNet[40],BiSeNet[41]等模型。其中 FCN 不仅开启了像素级语义分割,更开拓了之后语义分割算法使用全卷积网络的新思路,使语义分割打破了传统方法的限制,从而提高了分割精度。

3 基于深度学习的语义分割方法

如何在复杂环境中精确地实现目标的语义分割是当前研究的重点和难点。FCN 模型初步实现了像素级的语义分割,将图像语义分割的精度推向了新高度,这是该方向的标志性成果之一,为复杂环境的精确语义分割提供了可能性。至此,大量基于 CNN 的语义分割方法相继出现,具体包括基于强监督的语义分割方法、基于弱监督的语义分割方法以及基于无监督的语义分割方法,其主要优缺点如表 2 所列。

表 2 强监督、弱监督及无监督语义分割方法的优缺点对比结果

Table 2 Advantages and disadvantages of strong supervision, weak supervision and unsupervised semantic segmentation methods

类别	优点	缺点
强监督语义分割	基于密集标记的数据集,分割精度较高	过度依赖密集标记的数据集,无法进行迁移,对未知场景的分割精度差
弱监督语义分割	只需要图像级标注数据集即可完成训练	需要进行大量数据集的训练,耗时较长,精度低于强监督分割
无监督语义分割	不依赖于人工密集标注数据集,对未知环境具有较强的适应能力	域适应难度较大,目前分割精度不高

本节将从解决复杂环境语义分割问题的角度，分类阐述有关图像语义分割技术的新进展。

3.1　强监督语义分割方法

基于强监督的语义分割方法，依赖于人工密集标注的数据集进行训练，对训练场景的分割精度较高，但对未知环境的适应能力较差。

在复杂环境中，由于目标物体的大小不一致、边界模糊、光照和季节变化等因素的影响，导致神经网络提取到的特征较为粗糙，通过分类网络简单扩展的 FCN 无法较好地提取物体边界，分割效果不佳。近年来，国内外研究学者提出多种方法来提升复杂环境的图像语义分割性能，包括编解码方法、多尺度方法等。其中多尺度方法因具有整合全局信息的能力而成为近年来较为常用的方法，如 DeepLab v3+[42] 使用多尺度方法结合编解码结构获得了目前最好的分割效果。

3.1.1　基于编解码的方法

在基于深度神经网络的图像语义分割方法中，比较经典的是编解码（Encoder-Decoder）结构的方法。该类方法的网络结构如图 2 所示，其包含编码器和解码器两大部件，由一系列卷积层和上采样层组成。编码器为卷积层和上采样组成的分类网络，用于产生低分辨率的图像表示或特征映射。解码器与编码器相对称，用于将前一阶段获得的低分辨率图像映射到像素级预测上。最后一层为 softmax 分类器，用于对每个像素标签进行分类预测。编解码方法可产生与原图相同的高分辨率预测图像。编码器和解码器之间可直接进行信息沟通，用于在解码阶段更好地恢复目标细节。

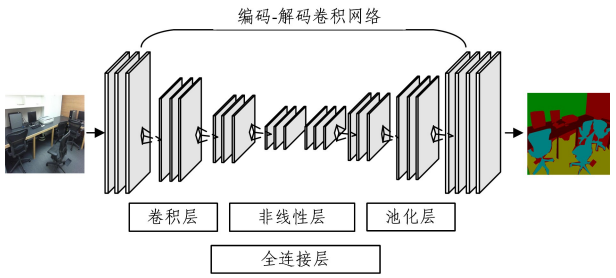

图 2　编解码网络结构图

Fig. 2　Encoder-Decoder network structure

2014 年，Long 等[36] 提出的全卷积网络，是第一个使用类似编解码方法的网络。该奠基性成果基于 AlexNet，VGG，ResNet，GoogleNet 等基本网络，对复杂环境下的图像语义分割具有较强的应对能力。但由于其直接对原图进行填充，引入了噪声，且未考虑有用的上下文信息，导致分割精度不高，参数规模巨大，计算效率欠佳。

为解决 FCN 现存的缺点，Chen 等[43] 将 CRF 作为 FCN 分割网络的后处理算法，利用 CRF 对网络分割结果进行细化，得到了较好的效果。但 FCN 与 CRF 未能很好地融合，有悖于神经网络端到端的训练目标。2015 年，Zheng 等提出的 CRF as RNN 模型[44] 使 FCN 的分割准确率有了显著提高。该工作的主要贡献在于将密集 CRF 重写为带有成对势能的形式，并将其视为 RNN 结构，成功地将 CRF 与 RNN 整合在一起，使其成为一个完整的端到端网络。

2015 年，基于 FCN 框架，Badrinarayanan 等提出了 SegNet 网络[37]，它是典型的编码-解码网络，用于道路、车辆的分割。该网络的优点在于池化层记录像素点的空间位置，在后续恢复图像分辨率时，能够有效地将其映射回对应位置，保留像素空间信息；然而，SegNet 不能很好地识别物体轮廓，物体边缘的分割精度较差。

此外，Noh 等提出的 DeconvNet[45] 对卷积层进行镜像处理，构成 Encoder-Decoder 结构，以编解码结构改善了 FCN 效果。随后，Hong 等[46] 模仿 DeconvNet，在 FCN 的基础上将卷积层与全连接层全部进行镜像处理，导致网络结构的参数规模很大，结果不如 DeconvNet。Paszke 等提出的 ENet[47] 在卷积之间添加 BN 层和 ReLU，依然采用 Encoder-Decoder 结构，分割效果较好。Yang 等[48] 提出的 CEDN，在使用 Encoder-Decoder 结构的同时，采用 contour 概率图结合 MCG 方法进行分割，效果良好，但速度较慢。

3.1.2　基于多尺度信息的方法

若要解决图像语义理解带来的挑战，模型除了需要具有优秀的网络结构，还需要具备对各种信息进行整合的能力，包括对尺度空间信息的整合，以及对局部与全局信息的平衡。研究者提出了许多方法使模型具有获取全局信息的能力，例如用 CRFs 作为后处理调优结果、多尺度聚合以及将对上下文的建模延伸到另一种深度模型中。常用于获取多尺度信息的方法架构如图 3 所示[49]。

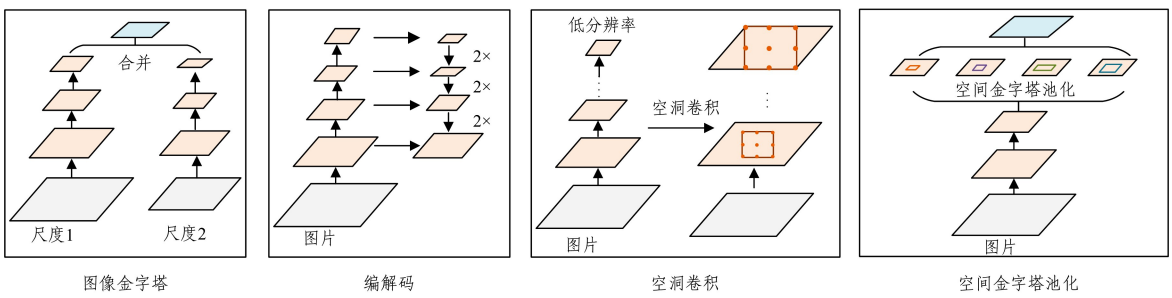

图 3　用于捕获多尺度信息的可选架构

Fig. 3　Alternative architecture for capturing multi-scale information

基于多尺度的方法可以将粗糙的高层语义特征和细粒度底层特征进行融合，以有效获取详细的空间信息，并改善多尺

度目标引起的物体分割精度降低的问题。图 3 中的图像金字塔结构将输入图片放缩成不同比例，分别应用于卷积神经网络，并将预测结果融合得到富含上下文信息的输出；编解码结构将编码阶段的多尺度特征运用到解码阶段，在解码阶段融合多尺度特征，同时恢复空间分辨率；空洞卷积结构通过使不同大小的空洞卷积核对原图进行卷积操作，以获取不同尺度的特征图，这种方法能够在不引入额外参数和不降低特征分辨率的情况下扩大感受野，从而获得更多的上下文信息；空间金字塔池化具有不同采样率和多种视野的卷积核，能够以多尺度捕捉对象。

为了扩大感受野，有效地利用上下文信息，Yu 等[50]提出了空洞卷积（Dilated Convolutions）模型。空洞卷积核是 Kronecker 卷积核[51]的一种泛化，可以在不丢失分辨率的情况下指数级增大感受野。此模型中卷积核孔洞倍数从小到大，先使用小尺度卷积核获取局部特征，再用大尺度卷积核将特征分到更多区域中。该结构能够在避免空间层级化信息丢失的同时，保留图像的内部数据结构，解决了语义分割任务的瓶颈问题。Chen 等[49]和 Wang 等[52]对空洞卷积进行了更详细的讨论。

在利用多尺度信息的一类方法中，最为经典的是 Google 提出的 DeepLab 系列方法。2014 年公布的 DeepLab v1 模型[43]使用空洞卷积，按照目标物体大小修改 hole，进而调整感受野；最后采用 CRFs 对分割结果进行细节增强。此后，研究者们发现多尺度可以使模型表现得更好。2017 年，Chen 等又提出了 DeepLab v2[38]版本，在 v1 版本上增加了一个多视野域，打造多尺度以提高模型的表现力，该结构被称为基于洞的空间金字塔（Atrous Spatial Pyramid Pooling，ASPP）[38]。同年，Chen 等提出的 DeepLab v3[49]在 v2 版本的基础上改进了 ASPP 模块，加入了 1×1 的卷积层和全局平均池化层。DeepLab v3 使用空洞卷积与空间金字塔池化结构，在该结构之后，DeepLab v3 还对特征图进行了 8 倍或 16 倍的上采样。但该方法较为粗糙，导致场景中的细节信息被忽略。

为改进 DeepLab v3 的缺陷，2018 年 Chen 等又提出了 DeepLab 的 v3＋版本[42]，v3＋版本使用改进版的 Xception 作为基础网络，实验证明该网络比 ResNet 101 的表现更好，且该版本设计了基于 v3 的解码模块，进一步保护了物体的边缘细节信息。

针对当前语义分割丢失信息的问题，2017 年，Lin 等提出了 RefineNet 网络[53]，其使用链式残差连接，能够有效地将下采样中缺失的信息进行融合，从而产生高分辨率的预测图像。这种方法可以将粗糙的高层语义特征和细粒度底层特征进行融合，使用残差连接和恒等映射思想，实现端到端训练。该模型在各种复杂场景中都有较好的分割效果。

2017 年，Zhao 等[54]提出的 PSPNet 也使用空洞卷积改善 ResNet 结构，并添加了一个金字塔池化模块。金字塔池化模块使用大内核池化层来捕获全局信息，其中内核分别覆盖了图像的整个区域、半个区域和小块区域。在图像进入金字塔池化模块阶段后，除了主分支的损失之外又新增了附加损失。

2018 年，Yu 等提出了双向分割网络（Bilateral Segmentation Network，BiSeNet）[41]，其主要用于实时语义分割。BiSeNet 网络是在现有实时语义分割模型加速方法的基础之上，针对现有方法牺牲精度以求速度的缺点所提出的。该网络包含两个部分：空间路径和上下文路径，分别用于解决空间信息缺失和感受野缩小的问题。

同年，Yu 等提出了判别特征网络（Discriminative Feature Network，DFN）[55]。该网络利用注意力模块（Convolutional Block Attention Module，CBAM）[56]选择更具判别力的特征，有效解决了语义分割的两个基本问题：类内不一致与类间无差别问题。Zhang 等[57]提出上下文语义编码模块与类别预测模块，在某种程度上减轻了分割问题中类间样本不均衡的问题。

2019 年，何恺明团队[58]提出“全景 FPN”，聚焦于图像全景分割任务，将分别用于语义分割和实例分割的 FCN 和 Mask R-CNN 相结合，设计了 Panoptic FPN。该方法使用丰富的多尺度特征，同时对语义分割和实例分割有效，兼具稳健性和准确性。

3.2 弱监督语义分割方法

强监督的语义分割方法是基于深度神经网络语义分割发展初期的研究热点。由于像素级的语义标注不仅难以获取，且依靠像素级标注训练得到的网络具有局限性，因此研究者们考虑采用一些相对于像素级标签更容易获取的标注作为监督信息，如物体框、线条、点，以及图像标签等，构建训练图像中图像标签和像素之间的关联，实现弱监督语义分割。

2016 年，Wei 等[59]提出了基于目标区域的方法来构建图像标签与语义之间的关联。该方法通过训练一个多标签的分类网络，对图片产生的目标区域进行分类，最后将类别上置信度较高的目标区域映射回原图片，从而获得语义标签和位置之间的关联，并将定位图作为监督信息训练语义分割网络。该方法虽然获得了较高的分割性能，但有两个明显的缺点：1）该方法需要对所有目标区域做一次分类，损耗较高；2）直接将目标区域内的像素点作为物体区域会引入错误像素。

为弥补基于目标区域的方法的缺点，2017 年 Wei 等[60]提出了一种由简单到复杂的方法。该方法首先需要从网络下载大量的简单图片，然后通过显著性检测技术获取其对应的显著图，进而利用显著图和语义标签作为监督信息训练一个初始分割网络，使之对简单图片具备一定的分割能力。随后，利用该网络预测出所有简单图片的标签，并利用这些标签训练一个增强版语义分割网络。最后通过强化语义分割网络来预测更多复杂图片的标签，并训练出一个更好的语义分割网络，即 Powerful DCNN。该方法的缺点是，必须收集大量简单图片对初始网络进行训练，且训练样本多、训练时间长。

鉴于上述不足，研究者们希望发现一种不依赖于简单图片的方法。2016 年，Zhou 等[61]提出了 CAM（Class Activation Mapping）方法，其中分类网络可以通过 Top-down 的方

式定位出图片上对物体分类贡献较大的区域。然而，CAM 方法的主要问题在于它只能发现最具判别力的物体区域，这些区域通常分布稀疏且只属于目标物体的一部分，这与语义分割需要定位完整物体的目标并不一致。2017 年，Wei 等提出了对抗性擦除方法[62]。该方法主要通过不断擦除物体上最具判别力的一些区域，使得分类网络能够发现物体更多的区域。虽然该方法用更简单的方式获得了更高的性能，但也存在两个明显的问题：1）需要多次训练分类网络，时耗较长；2）对于每张训练图片很难确定何时停止擦除操作。

针对对抗性擦除的缺点，2018 年 Zhang 等进一步提出了一种对抗补充学习的方法[63]。该方法首先利用一个分类网络发现一些物体的判别区域；然后将这些区域从中间的特征图中擦除，并将擦除后的特征图输入到另外一个分支中进行训练，进而获得同第一个分支互补的物体定位图；最后，将从两个分支获得的物体定位图合并获得最终的结果。

3.3　无监督语义分割方法

许多研究成果表明：基于大规模数据集训练出来的深度神经网络具有良好的工作性能。然而，给定一个新的数据集，训练好的神经网络往往表现不佳，典型的解决方案是仍然执行密集的手动标记，重新训练网络。另一种方案是利用可自动生成语义标注的计算机合成数据[64]进行训练。然而，由于“Domain Shift”现象[65]，重复使用合成数据训练的模型可能会损害其在实际数据中的性能。因此，在解决该问题时需要使用无监督域适应方法，构建源域的标记示例和目标域中未标记示例之间的映射关系，来减少目标数据上的预测误差。

无监督域自适应的一般做法是通过最小化 Domain Shift 度量来建立跨域的不变性，例如相关距离[66]或最大均值差异[67]。2015 年，Tzeng 等[68]提出 DC 方法，利用二元域分类器和域混淆损失来激励预测的域标签均匀分布。2017 年，Tzeng 等[69]提出对抗性判别域适应（Adversarial Discriminative Domain Adaptation，ADDA），通过对抗性训练优化适应模型。Hoffman 等提出的 FCNWild 方法[70]，仅利用对域适应的全卷积对抗训练来解决跨域分割问题。Zhang 等[71]提出了全卷积自适应网络（Fully Convolutional Adaptation Networks，FCAN），这是一种用于语义分割的新型深层体系结构，它结合了图像域适应网络和特征自适应网络，探索了如何使用合成图像提升真实图像的语义分割性能。该方法对从 GTA5（游戏视频）到城市街景的语义分割进行了泛化实验，与最先进的无监督自适应技术相比，该方案取得了优异的成果。

4　经典语义分割算法的对比

本节首先对当前图像语义分割算法的性能评估标准进行归纳，然后梳理了测试价值较高的公开数据集，最后在此基础上对经典的语义分割算法的性能进行了综合性对比与评述。

4.1　评估指标

经过近几年的快速发展，图像语义分割算法性能的主要评估指标可归纳为：准确率、时间复杂度及内存消耗 3 项，其中准确率是最关键的指标。

虽然现有的文献对图像语义分割成果采用了许多不同的精度衡量方法，但本质上它们都是像素精度及图像交并比的变种。例如：像素精度（Pixel Accuracy，PA）[36]、均像素精度（mean Pixel Accuracy，mPA）[36]、平均交并比（mean Intersection over Union，mIoU）[36]等。其中，mIoU 由于其简洁和较强的代表性成为了最常用的度量标准，大多数研究者都使用该标准报告其结果。

假设共有 $k+1$ 个类（从 L_0 到 L_k，其中包含 1 个空类或背景），p_{ij} 表示本属于类 i 但被预测为类 j 的像素数量。即，p_{ii} 表示真正的数量，而 p_{ij} 和 p_{ji} 则分别被解释为假正和假负。那么，像素精度（PA）为标记正确的像素占总像素的比例。

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

均像素精度（mPA）为 PA 的提升，其计算每个类内被正确分类的像素数的比例，之后求所有类的平均。

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

平均交并比（mIoU）即计算两个集合的交集和并集之比，在语义分割问题中计算的是真实分割与预测分割之间的交并比，即用真实正例的数量除以真实正例、错误负例、错误正例的总数量。

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

4.2　图像语义分割数据集

为了科学、一致地评价各类图像分割算法的性能，需要使用标准的图像数据集进行测试和对比。语义分割使用的数据集大多包含像素级标签，根据包含的信息可将数据集分为：2 维平面 RGB 图片数据、2.5 维带有深度信息的 RGB-D 数据，以及 3 维数据。本节主要介绍复杂环境领域的 2 维及 2.5 维数据集，其简要说明及使用频率如表 3 所列。

2009 年，第一个道路、驾驶场景理解数据集 CamVid[72-73]被发布，该数据集源于 5 个视频序列，来自一个安装在汽车仪表盘上的 960×720 分辨率的摄像机，包含 32 类物体，701 张图片。

Sift flow[74]数据集包含 2688 张完全标注的图像，是 LabelMe[75]数据集的子集。多数图像基于 8 种不同的户外场景，包括街道、高山、田地、沙滩、建筑等。

PASCAL VOC[76]数据集含有 21 类物体以及 9993 张图片，该数据集是目前最受欢迎的语义分割数据集。2014 年，在 PASCAL VOC 的基础上，PASCAL CONTEXT[77]将物体类别以及图片数量都大幅度地进行了扩充。

Cityscape[78]数据集是一个大规模城市街道场景数据集，提供了 8 种 30 个类别的标注。该数据集包括约 5000 张精细标注的图片，20000 张粗略标注的图片。其涵盖不同的时间

以及不同天气情况下的城市街景，包含大量动态物体、变化的场景布局以及变化的背景。

SYNTHIA[79-80]数据集是一个大规模虚拟城市合成数据集，带有语义信息，为自动驾驶或城市场景规划等研究领域而提出。

NYU Depth v2 数据集[81]是常用的 2.5 维数据，带有深度信息，包含 40 类室内物体，1449 张由微软 Kinect 设备捕获的 RGB-D 图像。该数据集重点刻画室内场景，可用于家庭机器人训练任务。但是由于数据集规模较小，限制了其在深度神经网络的应用。

SUN3D[82]数据集与 NYU Depth v2 相似，该数据集包含了一个大规模的 RGB-D 视频数据，每一帧包含了场景中的物体语义分割信息以及摄像机位态信息，部分图片涵盖一天中不同时段拍摄的情况。

SUN RGB-D[83]数据集由 4 个 RGB-D 传感器获取，包含 10000 张 RGB-D 图像，尺寸与 PASCAL VOC 一致。该数据集包含了 NYU Depth v2，Berkeley B3DO[84]，以及 SUN3D 数据集中的图像，整个数据集均为密集标注，包括多边形、带方向的边界框以及三维空间，适合于场景理解任务。

表 3 常用语义分割数据集汇总表

Table 3 Commonly used semantic segmentation datasets

时间	数据集名称	说明	使用频率
2009	CamVid [72-73]	包含 32 类物体，701 张图片。Sturgess 等[85]将数据集按照 367∶100∶233 的比例分为训练集、验证集、测试集	★★
2011	Sift flow [74]	类别数为 33 类，含有 2688 张图片。Liu 等[74]将该数据集分成 2488 个训练集，200 个测试集	★★
2012	PASCAL VOC [76]	包含 1 个已标注的图像数据集和 5 个不同的竞赛。竞赛分为检测、分类、分割、动作分类、任务布局。分割比赛的目标主要是预测测试集中每幅图像每个像素所属的物体类别。该数据集通常被划分为 2 个子集，包含 1464 张训练图像和 1449 张验证图像，测试集在竞赛中不公开	★★★★
	NYU Depth v2 [81]	专注于室内场景数据集，数据规模较小，限制了其在深度网络中的应用	★★
2013	SUN3D[82]	包含 8 个已标注视频序列，将扩展为 415 个序列，该数据集从 41 座建筑中的 254 个空间获取，某些地方在一天中的多个时段被重复拍摄	★
2014	PASCAL CONTEXT[77]	包含了 10103 张训练图像的像素级别的标注，共 540 类	★★★
2015	Cityscape[78]	大规模关注于城市街道场景理解的数据集，提供了 8 种 30 个类别的语义级别、实例级别以及密集像素标注。数据是从 50 个城市中持续数月采集而来，涵盖不同的时间以及好的天气情况	★★★★
	SUN RGB-D[83]	包含 37 个类别，5285 张训练图片，5050 张测试图片	★★★
2016	SYNTHIA [79-80]	提供了 13407 张训练图像的 11 个类别物体细粒度的像素级标注	★★

4.3 语义分割算法的性能对比

深度学习技术在各种高级视觉任务上取得的巨大成功离不开本文综述的各种优秀算法。本节将对面向复杂环境的图像语义分割方法进行分析比较，结果如表 4 所列，其中主要的对比因素有算法分类、方法名称、提出时间、算法特点、使用的数据集以及在该数据集上的 mIoU 等。

在传统图像语义分割方法中，N-Cut，Grab Cut 等经典算法得到了广泛的应用，且具有标准开源代码，但这些算法的效率较低。在此基础上改进的算法，如 Random Decision Forests，GPB-UCM，MCG 等，吸取了经典算法的优点，同时打破了常规，在生成的图像分割块的质量以及算法时间复杂度上都有更好的表现。其中，MCG 更是传统语义分割的巅峰之作。

为打破传统分割方法在完成语义分割任务上的限制，现阶段图像语义分割方法大多放弃使用人工设计特征算子，转而使用以卷积神经网络为代表的自动特征提取技术。基于神经网络的图像语义分割使用端到端的方法，通过改造经典网络，如 AlexNet，VGG 等，将图像分类网络转换为逐像素预测的网络。网络训练一般耗时较长，但训练完成后，预测图片的时间仅为毫秒级。为应对像素级标注样本不足的挑战，一些基于弱标注样本和计算机生成图像样本的弱监督和无监督方法相继出现。如文献[60,63]提出的基于显著性检测技术的弱监督方法，通过显著性检测技术检测目标物体的判别区域，以寻求图像标签和像素语义之间的关联，从而完成分类。文献[71]提出的无监督方法，将计算机合成图片域适应到目标图片，以完成图像的语义分割任务。

表 4 图像语义分割方法的分析归纳

Table 4 Analysis and summary of image semantic segmentation methods

分类	方法	提出时间	算法特点	数据集	分割精度 mIOU/%
传统语义分割	Grab Cut[17]	2004	利用图像的纹理信息和边界信息，依靠少量人工干预得到较好的前景与背景分割	—	—
	GPB-UCM[21]	2011	计算每个像素作为边缘的概率，检测目标轮廓，生成轮廓图，完成分割。步骤繁杂，复杂度高	BSDS[86]	—
	Random Decision Forests[23]	2016	使用多决策树组合成分类器	—	—
	MCG[24]	2017	在 GPB-UCM 的基础上，利用生成的多个轮廓分割块，结合随机森林分类器，得到预测对象	BSDS	—

（续表）

分类		方法	提出时间	算法特点	数据集	分割精度 mIOU/%
基于深度学习的语义分割	强监督	FCN-8s[36]	2014	以 CNN 为基础架构，引入跳层链接。未考虑上下文信息，分割精度不高	PASCAL VOC/ NYU Depth v2/ Sift flow	62.7/34.0/ 39.5
		DeepLab v1[43]	2014	使用全连接 CRF 辅助 CNN 定位图中物体的位置，优化图像分割效果	PASCAL VOC	71.6
		CRF as RNN[44]	2015	将 CRF 重写为带有成对势能的形式，并与 RNN 整合在一起，形成一个完整的端到端网络，改进了 FCN 的分割性能	PASCAL VOC	74.7
		SegNet[37]	2015	典型的编解码结构，采用上采样方法恢复图像尺寸，能够将相关值精准地映射到对应位置，提高图像恢复精度	CamVid/ SUN RGB-D	60.1/31.84
		Dilated Convolution[50]	2015	使用扩张卷积，在不丢失分辨率的情况下增大感受野	PASCAL VOC	67.6
		DeepLab v2[38]	2017	将扩张卷积与金字塔模块融合，提出空洞金字塔，打造多尺度，提高模型的表现力	PASCAL VOC/ Cityscapes	79.7/70.4
		DeepLab v3[49]	2017	改进空洞金字塔，引入 1×1 卷积层及全局平均池化层，提出使用空间金字塔池化提取多尺度信息	PASCAL VOC/ Cityscapes	85.7/81.3
		PSPNet[54]	2017	使用金字塔池化模块，并引入扩张卷积，融合多尺度信息，以提高分割精度	PASCAL VOC/ Cityscapes	85.4/80.2
		RefineNet[53]	2017	使用 long-range 残差连接，将粗糙的高层语义特征和细粒度底层特征进行融合，从而产生高分辨率的预测图像	NYU Depth v2/ Cityscapes/ PASCAL VOC/ SUN-RGB-D	46.5/73.6/ 83.4/45.9
		DeepLab v3+[42]	2018	使用改进版的 Xception 作为基础网络，并设计了基于 v3 版本的解码模块，进一步保护边缘细节信息	PASCAL VOC/ Cityscapes	89.0/82.1
		BiseNet[41]	2018	包含 Spatial Path 和 Context Path，分别用来解决空间信息缺失和感受野缩小的问题	Cityscapes	68.4
		DFN[55]	2018	利用注意力模块选择更具判别力的特征，有效解决类内不一致与类间无差别的问题	PASCAL VOC/ Cityscapes	86.2/80.3
		FPN[58]	2019	将 FCN 和 Mask R-CNN 相结合，使用丰富的多尺度特征，可同时解决语义分割与实例分割任务	Cityscapes	79.1
	弱监督	Proposal based[59]	2016	训练多标签分类网络，获取置信度较高的 proposal 构建图像标签与语义之间的关联。该方法获得了较高的分割性能，但损耗较大，且容易引入错误像素	PASCAL VOC	43.2
		Simple to Complex[60]	2017	通过下载简单图片，利用显著图和语义标签作为监督信息，训练分割网络。但需要大量简单图片进行训练，训练时间较长	PASCAL VOC	51.2
		Adversarial Erasing[62]	2017	不断擦除最具判别力区域，使分类网络发现该物体的更多其他区域，完整定位目标物体	PASCAL VOC	55.7
		Adversarial Complementary Learning[63]	2018	擦除特征图中已发现的物体判别区域，并将擦除后的特征图输入到另一分支进行训练，得到物体定位图，并将两支结果进行融合	ILSVRC[31]	—
	无监督	FCNWild[70]	2016	利用对域适应的全卷积对抗训练来解决跨域分割问题	GTA5/SYNTHIA→ Cityscapes	27.1/17.0
		ADDA[69]	2017	通过对抗性训练优化适应模型	NYU Depth v2	—
		FCAN[71]	2018	结合图像域适应网络以及特征自适应网络，探索使用合成图像来提升真实图像语义分割的性能	Cityscapes	47.75

注：—表示论文中无此项数据

结束语　本文主要阐述了面向复杂环境的图像语义分割方法，介绍了复杂环境下语义分割任务所面临的挑战，以及传统图像语义分割方法和基于深度学习的语义分割方法，重点致力于深度学习这一正在崛起的研究领域，涵盖了最前沿的相关工作。本文既为读者提供了必要的语义分割任务背景知识，也使读者了解到现阶段语义分割研究的进展。本文主要从模型结构以及算法特点方面综述了为完成复杂环境下语义分割任务做出卓越贡献的方法，以及复杂环境领域常用的数据集。语义分割问题正被许多优秀的方法推进，深度学习技术也被证明对于解决语义分割问题具有有效性，但该领域仍然存在着开放的问题。未来图像语义分割领域有如下一系列研究方向。

（1）实时视频分割。现阶段语义分割方法优劣的评估标准聚焦在准确率上，随着现有方法准确率的提高，以及许多现实场景的应用，分割系统被要求具有越来越短的响应时间。因此，未来工作需要在准确率与运行时间之间寻求平衡。

（2）三维数据的应用。三维数据对真实场景至关重要，目前对三维数据进行语义分割的方法陆续出现，但大量语义分割的工作仍围绕二维图像数据，因此将语义分割迁移到三维图像数据是未来的一个研究方向。

（3）弱监督以及无监督图像语义分割。在初期发展迅速的强监督语义分割由于极高的样本标注成本而面临精度提升瓶颈。为解决训练数据标注困难的问题，设计出更灵活、扩展性更强的弱监督以及无监督语义分割方法将得到越来越多的重视。弱监督乃至无监督语义分割是未来发展的趋势。

参 考 文 献

[1] GÓMEZ D，YÁÑEZ J，GUADA C，et al. Fuzzy image segmentation based upon hierarchical clustering[J]. Knowledge-Based Systems，2015，87(7)：26-37.

[2] NAZ S,MAJEED H,IRSHAD H. Image segmentation using fuzzy clustering: A survey[C]// International Conference on Emerging Technologies. Islamabad:IEEE,2010:181-186.

[3] PENG B,ZHANG L,ZHANG D. A survey of graph theoretical approaches to image segmentation[J]. Pattern Recognition, 2013,46(3):1020-1038.

[4] LIU S T,YIN F L. The Basic Principle and Its New Advances of Image Segmentation Methods Based on Graph Cuts[J]. Acta Automatica Sinica,2012,38(6):911-922. (in Chinese)
刘松涛,殷福亮. 基于图割的图像分割方法及其新进展[J]. 自动化学报,2012,38(6):911-922.

[5] YI F,MOON I. Image segmentation:A survey of graph-cut methods[C]// International Conference on Systems and Informatics. Yantai:IEEE,2012:1936-1941.

[6] JIANG F,GU Q,HAO H Z,et al. Survey on Content-Based Image Segmentation Methods[J]. Journal of Software, 2017, 28(1):160-183. (in Chinese)
姜枫,顾庆,郝慧珍,等. 基于内容的图像分割方法综述[J]. 软件学报,2017,28(1):160-183.

[7] GARCIA-GARCIA A,ORTS-ESCOLANO S,OPREA S,et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation[J]. arXiv:1704.06857,2017.

[8] SMEULDERS A W M,WORRING M,SANTINI S,et al. Content-Based Image Retrieval at the End of the Early Years[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2000,22(12):1349-1380.

[9] DESAI A D,GOLD G E,HARGREAVES B A,et al. Technical Considerations for Semantic Segmentation in MRI using Convolutional Neural Networks[J]. arXiv preprint arXiv: 1902.01977,2019.

[10] MARDIA K V,HAINSWORTH T J. A Spatial Thresholding Method for Image Segmentation[J]. IEEE transactions on pattern analysis and machine intelligence,1988,10(6):919-927.

[11] LAKSHMI S,SANKARANARAYANAN D V. A study of edge detection techniques for segmentation computing approaches [J]. International Journal of Computer Applications, 2010, CASCT(1):35-41.

[12] GIANNAKEAS N,KARVELIS P S,EXARCHOS T P,et al. Segmentation of microarray images using pixel classification-Comparison with clustering-based methods[J]. Computers in biology and medicine,2013,43(6):705-716.

[13] ADAMS R,BISCHOF L. Seeded region growing [J]. IEEE Transactions on pattern analysis and machine intelligence,1994, 16(6):641-647.

[14] LI S Z. Markov random field models in computer vision[C]// European conference on computer vision. Heidelberg:Springer, 1994:361-370.

[15] LAFFERTY J D,MCCALLUM A,PEREIRA F C N. Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data[C]// International Conference on Machine Learning. Williamstown:Morgan Kaufmann, 2001:282-289.

[16] SHI J,MALIK J. Normalized Cuts and Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.

[17] ROTHER C,KOLMOGOROV V,BLAKE A. "GrabCut":interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics,2004,23(3):309-314.

[18] HENZINGER M,NOE A,SCHULZ C,et al. Practical Minimum Cut Algorithms[J]. ACM Journal of Experimental Algorithmics,2018,23(1):1-8.

[19] XU H X,TIAN Z,DING M T. Multiscale Segmentation for SAR Image Based on Spectral Clustering and Mixture Model [J]. Journal of Image and Graphics,2010,15(3):450-454. (in Chinese)
徐海霞,田铮,丁明涛. 基于谱聚类与混合模型的 SAR 图像多尺度分割[J]. 中国图象图形学报,2010,15(3):450-454.

[20] LIU L,SHI Z G,SU H R,et al. Image Segmentation Based on Higher Order Markov Random Field[J]. Journal of Computer Research and Development,2013,50(9):1933-1942. (in Chinese)
刘磊,石志国,宿浩茹. 基于高阶马尔可夫随机场的图像分割[J]. 计算机研究与发展,2013,50(9):1933-1942.

[21] ARBELAEZ P,MAIRE M,FOWLKES C C,et al. Contour Detection and Hierarchical Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5):898-916.

[22] VINCENT L,SOILLE P. Watersheds in Digital Spaces:An Efficient Algorithm Based on Immersion Simulations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991,13(6):583-598.

[23] ZHANG C,XUE Z,ZHU X,et al. Boosted random contextual semantic space based representation for visual recognition[J]. Information Sciences,2016,369(6):160-170.

[24] PONT-TUSET J,ARBELAEZ P,BARRON J T,et al. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(1):128-140.

[25] FARABET C,COUPRIE C,NAJMAN L,et al. Learning Hierarchical Features for Scene Labeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(8):1915-1929.

[26] GHIASI G,FOWLKES C C. Laplacian pyramid reconstruction and refinement for semantic segmentation[C]// European Conference on Computer Vision. Amsterdam:Springer,2016:519-534.

[27] FAVREAU J D,LAFARGE F,BOUSSEAU A,et al. Extracting Geometric Structures in Images with Delaunay Point Processes[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE,2019:1-1.

[28] COUPRIE C,FARABET C,NAJMAN L,et al. Indoor Semantic Segmentation using depth information[J]. arXiv preprint arXiv: 1301.3572,2013.

[29] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[30] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Ad-

vances in neural information processing systems. Nevada：ACM，2012：1097-1105.

[31] 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.

[32] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv preprint arXiv：1409.1556，2014.

[33] LIU Y，YU J，HAN Y. Understanding the effective receptive field in semantic image segmentation[J]. Multimedia Tools and Applications，2018，77(17)：22159-22171.

[34] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE，2015：1-9.

[35] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：770-778.

[36] LONG J，SHELHAMER E，DARRELL T. Fully Convolutional Networks for Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Massachusetts：IEEE，2015：3431-3440.

[37] BADRINARAYANAN V，KENDALL A，Cipolla R. Segnet：A deep convolutional encoder-decoder architecture for image segmentation[J]. arXiv preprint arXiv：1511.00561，2015.

[38] CHEN L-C，PAPANDREOU G，KOKKINOS I，et al. DeepLab：Semantic Image Segmentation with Deep Convolutional Nets，Atrous Convolution，and Fully Connected CRFs[J]. IEEE transactions on pattern analysis and machine intelligence，2017，40(4)：834-848.

[39] LIN G，MILAN A，SHEN C，et al. RefineNet：Multi-Path Refinement Networks for High-Resolution Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii：IEEE，2017：5168-5177.

[40] ZHAO H，SHI J，QI X，et al. Pyramid Scene Parsing Network[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii：IEEE，2017：6230-6239.

[41] YU C，WANG J，PENG C，et al. BiSeNet：Bilateral Segmentation Network for Real-Time Semantic Segmentation[C]// European Conference on Computer Vision. Cham：Springer，2018：334-349.

[42] CHEN L C，ZHU Y，PAPANDREOU G，et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[J]. arXiv preprint arXiv：1802.02611，2018.

[43] CHEN L C，PAPANDREOU G，KOKKINOS I，et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. arXiv preprint arXiv：1412.7062，2014.

[44] ZHENG S，JAYASUMANA S，ROMERA-PAREDES B，et al. Conditional Random Fields as Recurrent Neural Networks[C]// IEEE International Conference on Computer Vision. Santiago：IEEE，2015：1529-1537.

[45] NOH H，HONG S，HAN B. Learning deconvolution network for semantic segmentation[C]// IEEE International Conference on Computer Vision. Santiago，Chile：IEEE，2015：1520-1528.

[46] HONG S，NOH H，HAN B. Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation[C]// Neural Information Processing Systems. Montreal：IEEE，2015：1495-1503.

[47] PASZKE A，CHAURASIA A，KIM S，et al. Enet：A deep neural network architecture for real-time semantic segmentation[J]. arXiv preprint arXiv：1606.02147，2016.

[48] YANG J，PRICE B，COHEN S，et al. Object contour detection with a fully convolutional encoder-decoder network[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：193-202.

[49] CHEN L C，PAPANDREOU G，SCHROFF F，et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. arXiv preprint arXiv：1706.05587，2017.

[50] YU F，KOLTUN V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. arXiv：1511.07122，2015.

[51] ZHOU S，WU J N，WU Y，et al. Exploiting Local Structures with the Kronecker Layer in Convolutional Networks[J]. arXiv preprint arXiv：1512.09194，2015.

[52] WANG P，CHEN P，YUAN Y，et al. Understanding convolution for semantic segmentation[C]// IEEE Winter Conference on Applications of Computer Vision. Nevada：IEEE，2018：1451-1460.

[53] LIN G，MILAN A，SHEN C，et al. Refinenet：Multi-path refinement networks for high-resolution semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii：IEEE，2017：5168-5177.

[54] ZHAO H，SHI J，QI X，et al. Pyramid scene parsing network[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii：IEEE，2017：2881-2890.

[55] YU C，WANG J，PENG C，et al. Learning a Discriminative Feature Network for Semantic Segmentation[J]. arXiv preprint arXiv：1804.09337，2018.

[56] WOO S，PARK J，LEE J Y，et al. Cbam：Convolutional block attention module[C]// European Conference on Computer Vision. Cham：Springer，2018：3-19.

[57] ZHANG H，DANA K，SHI J，et al. Context encoding for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：7151-7160.

[58] KIRILLOV A，GIRSHICK R，HE K，et al. Panoptic Feature Pyramid Networks[J]. arXiv preprint arXiv：1901.02446，2019.

[59] WEI Y，LIANG X，CHEN Y，et al. Learning to segment with image-level annotations[J]. Pattern Recognition，2016，59(1)：234-244.

[60] 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 and machine intelligence，2017，39(11)：2314-2320.

[61] ZHOU B，KHOSLA A，LAPEDRIZA A，et al. Learning deep features for discriminative localization[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：2921-2929.

[62] WEI Y,FENG J,LIANG X,et al. Object region mining with adversarial erasing:A simple classification to semantic segmentation approach[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii:IEEE,2017:6488-6496.

[63] ZHANG X,WEI Y,FENG J,et al. Adversarial complementary learning for weakly supervised object localization[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:1325-1334.

[64] RICHTER S R,VINEET V,ROTH S,et al. Playing for data:Ground truth from computer games[C]// European Conference on Computer Vision. Amsterdam:Springer,2016:102-118.

[65] YAO T,PAN Y,NGO C W,et al. Semi-supervised domain adaptation with subspace learning for visual recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston:IEEE,2015:2142-2150.

[66] SUN B,FENG J,SAENKO K. Return of frustratingly easy domain adaptation[C]// The Thirty-Second AAAI Conference on Artificial Intelligence. Arizona:ACM,2016:2058-2065.

[67] TZENG E,HOFFMAN J,ZHANG N,et al. Deep domain confusion:Maximizing for domain invariance[J]. arXiv preprint arXiv:1412.3474,2014.

[68] TZENG E,HOFFMAN J ,DARRELL T,et al. Simultaneous deep transfer across domains and tasks[C]// IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4068-4076.

[69] TZENG E,HOFFMAN J,SAENKO K,et al. Adversarial discriminative domain adaptation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Hawaii:IEEE,2017:4.

[70] HOFFMAN J,WANG D,YU F,et al. Fcns in the wild:Pixel-level adversarial and constraint-based adaptation[J]. arXiv preprint arXiv:1612.02649,2016.

[71] ZHANG Y,QIU Z,YAO T,et al. Fully Convolutional Adaptation Networks for Semantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:6810-6818.

[72] BROSTOW G J,SHOTTON J,FAUQUEUR J,et al. Segmentation and Recognition Using Structure from Motion Point Clouds [C]// European Conference on Computer Vision. Marseille: Springer,2008:44-57.

[73] 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.

[74] LIU C, YUEN J, TORRALBA A. Sift flow:Dense correspondence across scenes and its applications[J]. IEEE transactions on pattern analysis and machine intelligence,2011,33(5):978-994.

[75] RUSSELL B C,TORRALBA A,MURPHY K P,et al. LabelMe:A Database and Web-Based Tool for Image Annotation [J]. International Journal of Computer Vision,2008,77(1/2/3):157-173.

[76] EVERINGHAM M,ESLAMI S M A,GOOL L J V,et al. The Pascal Visual Object Classes Challenge:A Retrospective[J]. International Journal of Computer Vision,2015,111(1):98-136.

[77] MOTTAGHI R,CHEN X,LIU X,et al. The Role of Context for Object Detection and Semantic Segmentation in the Wild [C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus:IEEE,2014:891-898.

[78] CORDTS M,OMRAN M,RAMOS S,et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE,2016:3213-3223.

[79] ROS G,SELLART L,MATERZYNSKA J,et al. The SYNTHIA Dataset:A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE,2016:3234-3243.

[80] HERNANDEZ-JUAREZ D,SCHNEIDER L,ESPINOSA A,et al. Slanted Stixels: Representing San Francisco's Steepest Streets[J]. arXiv:1707.05397,2017.

[81] SILBERMAN N,HOIEM D,KOHLI P,et al. Indoor segmentation and support inference from rgbd images[C]// European Conference on Computer Vision. Florence:Springer,2012:746-760.

[82] XIAO J,OWENS A,TORRALBA A. Sun3d:A database of big spaces reconstructed using sfm and object labels[C]// IEEE International Conference on Computer Vision. Sydney,Australia: IEEE,2013:1625-1632.

[83] SONG S,LICHTENBERG S P,XIAO J. SUN RGB-D:A RGB-D scene understanding benchmark suite[C]// IEEE Conference on Computer Vision and Pattern Recognition. Massachusetts: IEEE,2015:567-576.

[84] JANOCH A,KARAYEV S,JIA Y,et al. A category-level 3-D object dataset:Putting the Kinect to work[C]// IEEE International Conference on Computer Vision. Barcelona:IEEE,2011: 1168-1174.

[85] STURGESS P,ALAHARI K,LADICKY L,et al. Combining appearance and structure from motion features for road scene understanding[C]// British Machine Vision Conference. London:British Machine Vision Association,2009:7-10.

[86] MARTIN D,FOWLKES C,TAL D,et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]// IEEE International Conference on Computer Vision. Vancouver: IEEE,2001:416-425.