Journal of Image and Signal Processing 图像与信号处理, 2023, 12(1), 61-75
Published Online January 2023 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2023.121007

Hans汉斯

图像语义分割方法研究进展

汤婧婧，章　盛，徐立中*

河海大学计算机与信息学院，江苏　南京

收稿日期：2022年12月19日；录用日期：2023年1月9日；发布日期：2023年1月29日

摘　要

图像语义分割任务是将图像的每个像素分类到每一个实例中，每个实例对应一个类。该任务是场景理解概念的一部分，由于深度学习的图像语义分割方法能更好地解释图像的全局上下文，越来越受到计算机视觉和机器学习研究者的关注，并广泛应用于室内导航、自动驾驶，甚至虚拟或增强现实系统等领域。本文介绍图像语义分割的术语的概念，回顾传统和现有的深度学习方法，强调了它们在该领域的贡献和意义，以及语义分割算法的评价指标与常用数据集，最后，我们对当前语义图像分割任务中存在的一些问题进行讨论，并提出相关解决方法和研究展望。

关键词

机器学习，计算机视觉，深度学习，语义分割

Recent Progress in Semantic Image Segmentation

Jingjing Tang, Sheng Zhang, Lizhong Xu*

College of Computer and Information, Hohai University, Nanjing Jiangsu

Received: Dec. 19th, 2022; accepted: Jan. 9th, 2023; published: Jan. 29th, 2023

Abstract

The semantic image segmentation task consists of classifying each pixel of an image into an instance, where each instance corresponds to a class. This task is a part of the concept of scene understanding or better explaining the global context of an image. Image semantic segmentation is more and more being of interest for computer vision and machine learning researchers. Many applications on the rise need accurate and efficient segmentation mechanisms: indoor navigation,

*通讯作者。

文章引用: 汤婧婧, 章盛, 徐立中. 图像语义分割方法研究进展[J]. 图像与信号处理, 2023, 12(1): 61-75.
DOI: 10.12677/jisp.2023.121007

autonomous driving, and even virtual or augmented reality systems to name a few. This paper provides the concept of image semantic segmentation terms, and reviews traditional and existing deep learning methods with emphasizing their contributions and significance in this field, as well as the evaluation indicators of semantic segmentation algorithms and commonly used datasets. At last, we discuss some problems in current semantic image segmentation tasks, and propose relevant solutions and research prospects.

Keywords

Machine Learning, Computer Vision, Deep Learning, Semantic Segmentation

1. 引言

图像语义分割是基于标签数据的像素到像素的分类，将像素(Pixel)按照图像中表达语义含义的不同进行分组(Grouping)/分割(Segmentation)，也称图像语义标注(Image semantic labeling)、像素语义标注(Semantic pixel labeling)或像素语义分组(Semantic pixel grouping)，具体实例如图 1 所示。图像语义分割是获取一些有意义的信息以提取所需信息的基本技术[1]，是当今最具挑战性的领域之一。它在教育、医学、天气预报、气候变化预测等各个领域都有着更多的应用[2]。特别是近年来，基于深度学习分割方法的出现，图像语义分割研究发展趋势强劲上升[3]。研究表明，完美的分类可以导致更好的语义分割结果，所以不断有最先进技术用于语义分割，基于区域的分割、基于图的分割、图像分割、实例分割，他们都具有相同语义分割基础。全面认识语义分割领域的进步，有利于图像分割技术的发展和壮大，也有力的促进计算机视觉的研究。

(a) 原图　　(b) 语义分割

Figure 1. Semantic image segmentation
图 1. 图像语义分割

2. 图像语义分割方法发展历程

图像语义分割方法始于上世纪七十年代，直到深度学习算法的应用，才出现各种各样的图像语义分

割方法，应用场景也越来越丰富，极大的促进其发展。经过几十年的研究，相关技术不断被完善。以深度学习应用于图像语义分割时间节点，将图像语义分割发展史分 2 个阶段：传统图像语义分割方法阶段和基于深度学习图像语义分割方法阶段：

1) 传统方法的图像语义分割方法时期

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

2) 基于深度学习方法的图像语义分割方法时期

当卷积神经网络(CNN) [4]出现后，学者们开始利用神经网络模型训练像素的特征分类器实现语义分割，这种方法受到传统语义分割方法诸多不足的限制，准确性普遍较低。由 Long [5]等人提出了全卷积神经网络(FCN)，至此图像语义分割方法进入到了全卷积神经网络时期。全卷积神经网络在深度学习中表现出了强大的潜力，计算机在图片通过深度学习网络进行深度学习后能够清楚地归纳出输入图片中的具有相同语义含义的像素点。深度学习方法成为了现今解决语义分割问题的主流。对比传统方法，基于全卷积神经网络深度学习的语义分割技术能够获得更高的精度以及更好的运算效率，因此这一时期的语义分割技术新方法多，进展快。

3. 传统的图像语义分割方法

传统的图像语义分割算法始于上世纪七十年代，由于当时计算机硬件设备不足的限制，研究者只能根据图像的颜色、纹理信息和空间结构等特征将图像分割成不同的区域，同一区域内具有一致的语义信息，不同区域之间属性不同，以手工为主对图像中的目标物进行分割，开发出许多各种不同方法用以图像语义分割，从最简单的阈值分割、区域生长、边缘检测到图划分的分割方法。图划分是经典的传统图像语义分割方法，其中最常用的就 Normalized cut 和 Grab cut 方法，N-cut 是一种考虑全局信息的方法来进行图割，用以改变经典的 min-cut 算法操作中的不足，创新点在于将两个分割部分与全图节点的连接权重也考虑进算法之中，根据图像中的像素给出的阈值将图像一分为二。缺点在于这种分割方式比较简单直接，只能利用图像的像素进行分割，对于整体物体的影响考虑不周。为了改进这一缺点，Grab cut 的创新在于预先将图片中需要进行分割处理的部分进行人工标定，在计算机处理的时候也需要人工进行干预，对图像进行标注，指导辅助计算机进行判断分割。总之，传统的图像分割算法由于没有数据训练阶段，计算复杂度不高，在较困难的分割任务上，分割性能的提升空间有限。几类传统的图像分割方法详细阐述如下：

1) 基于阀值的图像分割方法：原理是需要人为找出相关阀值，并将图像中各像素值与阈值进行比较，根据对比结果将各像素划分到不同区域，典型方法如最大熵法、模糊阀值法、自适应阀值法等。这类方法在初期凭借着运算量小且易于实现等优点逐步在各领域进行了广泛的应用，但其一般适用于单通道的灰度图像数据，无法适用后期大量出现的深通道彩色图像，是其存在的较大局限性。

2) 基于交互式的图像分割方法：原理主要基于早期的二分类理论，具体表现为以人工方式进行 seed (种子点)标注，seed 一般位于图像目标物的边界处，接着算法将以此作为约束条件，演算出最终的分割结果。很明显这类方法无法快速高效的批量处理各种复杂场景下的图像，且需要耗费较大的人力财力。

3) 基于边缘检测的图像分割方法：则是借助了图像不同区域块的特征差异性较大的特点，这种差异性具体表现为某像素值相邻两侧的灰度值跳跃过大，因此可以利用这一特性找出各区域块的边界从而实现图像目标物的边缘检测。若待分割的图像背景噪声较少，则该类方法一般能得到比较理想的结果。然

而遇到图像各目标边缘复杂、有重叠或有大量噪声干扰的情况时，该类方法难以得到较好的结果。

4) 基于概率图模型(PGM)的分割方法：以 PGM 作为基石对像素进行分类分割。PGM 一般分为生成模型和判别模型，生成模型原理是通过在各概率图节点与各像素点之间建立起对应关系，从而达到分割图像的目的。譬如在 1997 年，Friedman N 等人[6]提出的 Bayesian network (贝叶斯网络)就是创建出一个具备科学严谨的推导且操作简便的有向图模型，这个有向图模型由变量节点之间的条件概率分布情况构成。然而，图像各像素点之间通常具有上下文语义信息关联，并不完全独立，而概率图的构造是根据各变量节点的独立性， 因此研究人员难以为各像素建立完全准确的数学相关性。综合来看，基于 PGM 的分割算法比较适用于单一场景下的图像分割。

5) 基于聚类的图像分割方法：类比机器学习领域当中聚类的思想，根据图像中像素的灰度值相近与否对像素点进行分类，最终将图像划分为多个不同区域，该类方法的特点是不需任何先验信息且不需要特征提取这一操作，关键是要找准各大聚类中心初始化点的位置。总体来讲，基于聚类的分割方法其精确度一般不够理想。

4. 基于深度学习的图像语义分割方法

由于基于深度学习的语义分割模型数量繁多，以下按照模型架构进行介绍。

4.1. 全卷积网络(FCN)

Long 等人[5]提出了第一批用于语义图像分割的深度学习方法之一，使用了全卷积网络(FCN)。一个 FCN (图 2)仅包括卷积层，这使得它能够拍摄任意大小的图像并产生相同大小的分割图。该网络不同于现有的 CNN 架构，如 VGG16 和 GoogLeNet，用全卷积层替换所有全连接层来管理非固定大小的输入和输出。因此，模型输出的是空间分割图，而不是分类分数。

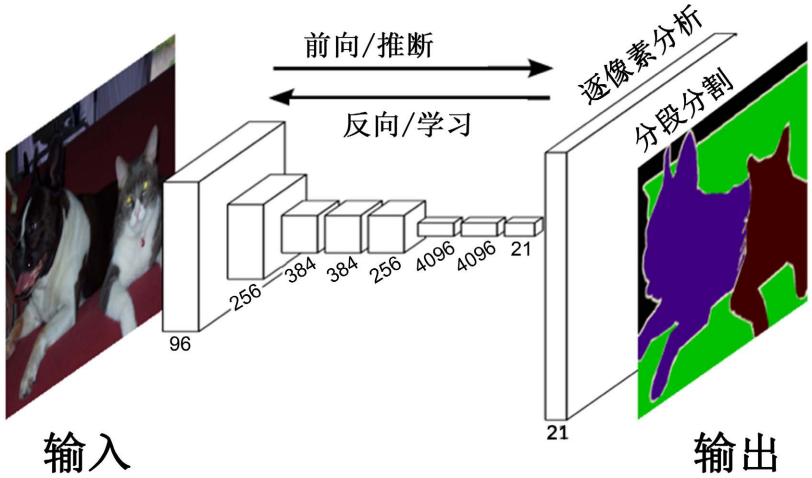

Figure 2. The FCN for making pixel-accurate predictions [5]
图 2. 用于像素精确预测的 FCN 网络[5]

通过使用跳过连接，其模型最终层的特征图被上采样并与较早层的特征图融合，该模型结合了语义信息(深的、粗糙的层)和外观信息(浅的、精细的层)，以便产生精确和详细的分割。该模型在 PASCAL VOC、NYUDv2 和 SIFT Flow 上进行了测试，取得了最先进的分割性能。

这项工作被认为是图像分割的里程碑，证明了深度网络可以在可变大小的图像上以端到端的方式进行语义分割。然而，尽管传统的 FCN 模型流行且有效，但它也有一些局限性——它对于实时推理来说不

够快，没有以有效的方式考虑全局上下文信息，并且它不容易转换成 3D 图像。一些学者试图克服 FCN 的部分局限性。例如，Liu 等人[7]提出了一个名为 ParseNet 的模型，来解决 FCN 的一个问题——忽略了全局上下文信息。ParseNet 通过使用图层的平均要素来扩充每个位置的要素，从而将全局上下文添加到 FCN 中。图层的要素地图被合并，整个图像产生一个上下文向量。该上下文向量被归一化和去池化，以产生与初始特征图相同大小的新特征图。然后将这些要素地图连接起来。简而言之，ParseNet 是一个 FCN，所描述的模块取代了卷积层(图 3)。

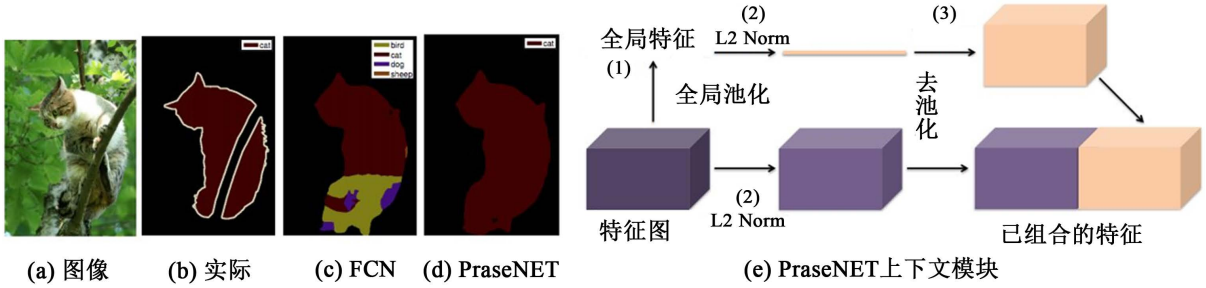

Figure 3. ParseNet produces a segmentation [7]
图 3. ParseNet 生成分割[7]

4.2. 带有图形模型的卷积模型

如前所述，FCN 忽略了潜在有用的场景级语义上下文，为了集成更多上下文，一些方法将概率图形模型，例如条件随机场(CRF)和马尔可夫随机场(MRF)结合到 DL 架构中。Chen 等[8]提出了一种基于细胞神经网络和全连接条件随机场相结合的语义分割算法(图 4)。深层 CNN 的最后一层的响应对于精确的对象分割来说不够局部化(由于使 CNN 适用于诸如分类之类的高级任务的不变性)。为了克服深层 CNN 较差的定位特性，他们将最终 CNN 层的响应与完全连接的 CRF 相结合。与以前的方法相比，他们的模型能够以更高的准确率定位线段边界。

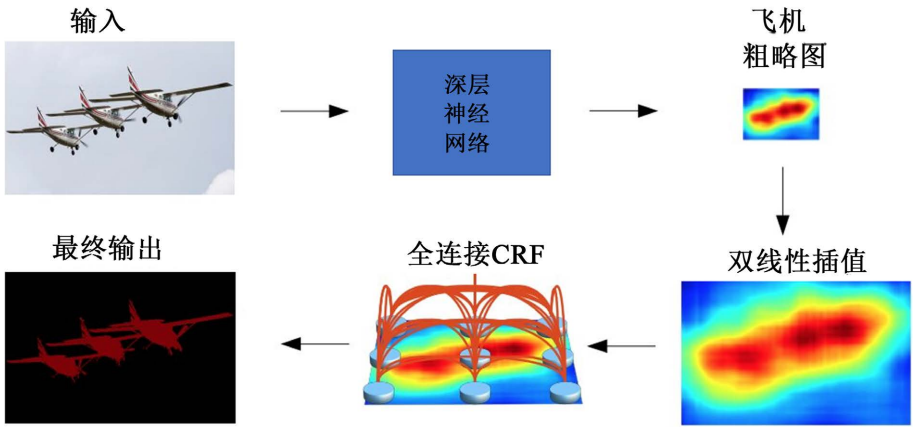

Figure 4. CNN + CRF model [8]
图 4. CNN + CRF 模型[8]

Schwing 和 Urtasun [9]提出了用于图像分割的全连通深度结构化网络，该网络是一种联合训练 CNN 和全连接 CRF 进行语义图像分割的方法，并在具有挑战性的 PASCAL VOC 2012 数据集上取得了令人鼓舞的结果。Lin 等人[10]提出了一种基于上下文深度条件随机场的高效语义切分算法。他们探索了“补丁–补丁”上下文(图像区域之间)和“补丁–背景”上下文，以通过使用上下文信息来改进语义分割。

4.3. 基于编码器-解码器的模型

另外用于图像语义分割的模型有很多是基于卷积编码器－解码器架构的，其中大多数基于深度学习的分割工作使用编码器－解码器模型。将这些工作分为两类，用于一般图像分割的编码器－解码器模型和用于医学图像分割的编码器－解码器模型。

4.3.1. 用于一般图像分割的编码器－解码器模型

Noh 等人[11]提出了一篇基于反卷积(又名转置卷积)语义分割的方法，他们的模型(图 5)由两部分组成，一个编码器使用 VGG 16 层网络的卷积层，一个去卷积网络将特征向量作为输入并生成逐像素类别概率图。去卷积网络由去卷积图层和去卷积图层组成，用于识别按像素分类的标注并预测分割掩膜。

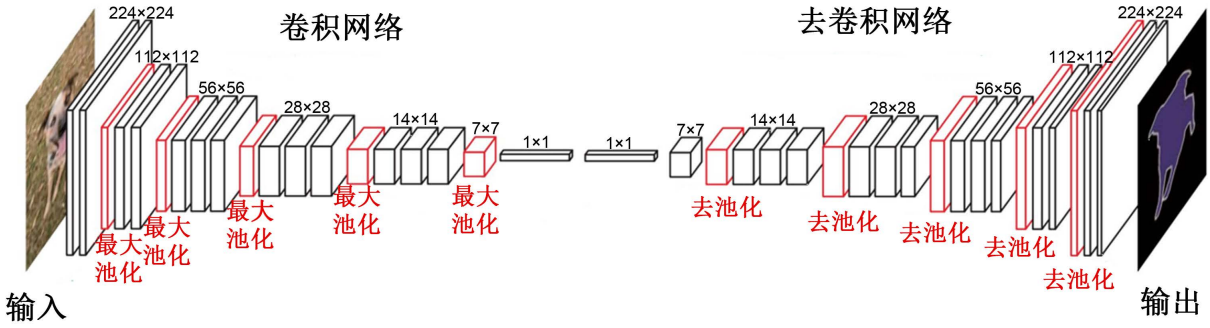

Figure 5. Deconvolutional semantic segmentation [11]
图 5. 反卷积语义分割[11]

Badrinarayanan 等人[12]提出了一种用于图像分割的卷积编解码器架构，类似于去卷积网络，SegNet 的核心可训练分段引擎包括编码器网络和相应的解码器网络，编码器网络在拓扑上与 VGG16 网络中的 13 个卷积层相同，解码器网络之后是逐像素分类层。与其他架构相比，SegNet 的可训练参数数量也少得多。同时作者还提出了 SegNet 的贝叶斯版本，以对用于场景分割的卷积编码器－解码器网络固有的不确定性进行建模[12]。

其他部分方法采用转置卷积或编码器－解码器进行图像语义分割，Fu 等人[13]提出了改善网络中的信息流和梯度传播并增强区分特征表示，利于网络优化的堆叠解卷积网络(SDN)，其他如高分辨率网络(HRNet)，Linknet，W-Net，以及用于 RGB-D 分割的位置敏感去卷积网络。

4.3.2. 用于医学和生物医学图像分割的编码器－解码器模型

受 FCNs 和编码器－解码器模型的启发，最初用于医学/生物医学图像语义分割的模型包括 U-Net [14]和 V-Net [15]，现在也用于医疗领域之外。

Ronneberger 等人[14]提出了用于分割生物显微图像的 U-Net。他们的网络和训练策略依赖于使用数据增强来有效地从很少的带注释的图像中学习。U-Net 架构，包括两个部分，捕获上下文的收缩路径和实现精确定位的对称扩展路径。下采样或收缩部分有一个类似 FCN 的结构，用 3 × 3 卷积提取特征。上采样或扩展部分使用上卷积(或去卷积)，减少了增加特征地图的数量，同时增加它们的尺寸网络的下采样部分的特征映射被复制到上采样部分，以避免丢失模式信息。最后，1 × 1 卷积处理特征图以生成对每个特征图进行分类的分割图，输入图像的像素。

由 Milletari 等人提出[15]的 V-Net 是另一个著名的基于 FCN 的模型，用于三维医学图像分割。对于模型训练，他们引入了一种基于 Dice 系数的新目标函数，使模型能够处理前景和背景中体素数量之间存

在强烈不平衡的情况。关于医学图像分割的一些其他相关工作包括用于从胸部 CT 图像中快速和自动分割肺叶的渐进密集 V-net (PDV-Net)等，以及用于病变分割的 3D-CNN 编码器。

4.4. 基于多尺度和金字塔网络的模型

多尺度分析是图像处理中的一个相当古老的概念，已经被部署在各种神经网络架构中。Lin 等人提出的特征金字塔网络(FPN)就是这类模型中最经典的一个[16]，它主要是为对象检测而开发的，但后来也应用于图像分割。Zhao 等[17]提出了金字塔场景解析网络(PSPN)，这是一个多尺度网络，用于更好地学习场景的全局上下文表示。使用残差网络(ResNet)作为特征提取器，利用扩展网络，从输入图像中提取不同的模式。这些特征地图随后被输入金字塔汇集模块，以区分不同尺度的模式。它们以四种不同的规模汇集在一起，每一种对应于金字塔等级并降低它们的维数。输出的金字塔等级被上采样并与初始特征图连接，以捕捉局部和全局上下文信息。最后，卷积层用于生成逐像素预测。

还有其他使用多尺度分析进行分割的模型，如 DM-Net (动态多尺度过滤网络)，背景对比网络和门控多尺度聚集(CCN)，自适应金字塔上下文网络(APC-Net)，多尺度上下文交织(MSCI)，以及显著对象分割。

4.5. 基于 R-CNN 的模型

区域卷积网络(R-CNN)及其扩展(Fast R-CNN、Faster R-CNN、Mask-RCNN)已被成功运用于目标检测应用中，其中，Faster R-CNN [18]架构(图 6)使用区域提议网络(RPN)来提议边界框候选。RPN 提取感兴趣区域(RoI)，RoI pool 层从这些建议中计算特征，以便推断边界框坐标和对象的类别。R-CNN 的一些扩展已经被大量用于解决实例分割问题，即同时执行对象检测和语义分割的任务。

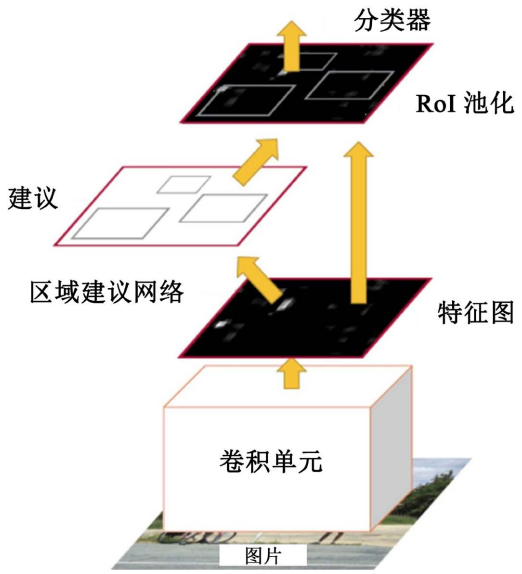

Figure 6. Faster R-CNN architecture [18]
图 6. Faster R-CNN 架构[18]

在这个模型的一个扩展中，He 等人[19]提出了一个用于对象实例分割的掩模 R-CNN，该 R-CNN 在许多 COCO 挑战中击败了所有以前的基准。该模型有效地检测图像中的对象，同时为每个实例生成高质量的分割掩模。Mask R-CNN 本质上是一个更快的 R-CNN，有 3 个输出分支(图 7)，第一个计算边界框坐标，第二个计算关联的类，第三个计算二进制遮罩以分割对象。掩模 R-CNN 损失函数组合了边界框坐标、预测类和分割掩模的损失，并联合训练它们。

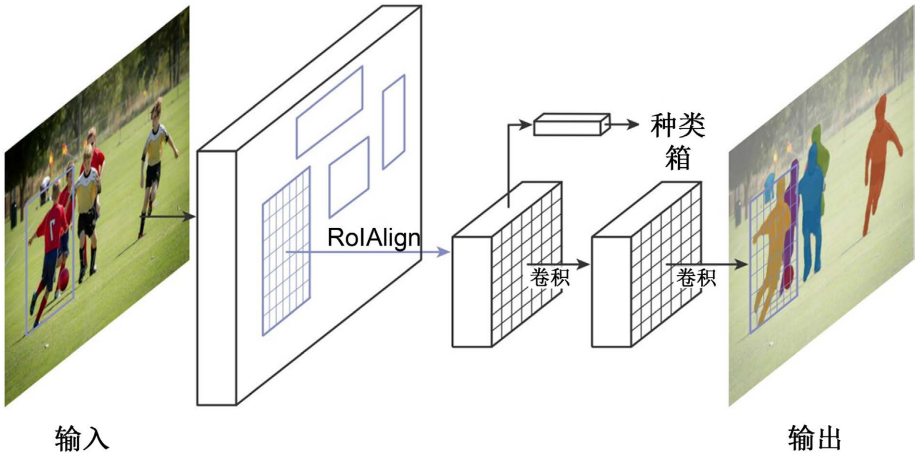

Figure 7. Mask R-CNN architecture [19]
图 7. Mask R-CNN 架构[19]

Liu 等提出的路径聚合网络[20]是基于掩模 R-CNN 和 FPN 模型。网络的特征提取器使用具有新的增强的自底向上路径的 FPN 架构，改进了低层特征的传播。第三条途径的每一个阶段都以大脑皮层的特征图作为输入，并且用 3 × 3 卷积层处理它们。输出被添加到相同阶段的特征图中并且这些特征图提供给下一阶段。如同在掩模 R-CNN 中一样，自适应特征池层的输出给三个分支。前两个使用完全连接的层来生成边界框坐标和相关对象类的预测，第三个用 FCN 处理 RoI 以预测对象遮罩。

Chen 等[21]提出了一个实例细分模型 MaskLab，通过基于更快的 R-CNN 使用语义和方向特征来细化对象检测。该模型产生三个输出，盒子检测、语义分割和方向预测。基于更快的 RCNN 对象检测器，预测框提供对象实例的精确定位。在每个感兴趣的区域内，MaskLab 通过结合语义和方向预测来执行前景/背景分割。

相关学者基于 R-CNN 也提出了许多其他模型，Lee 和 Park [22]提出了将新空间注意力引导 Mask (SAG-Mask)分支，在 Mask R-CNN 中添加到无锚点单级对象检测器(FCOS)中从而关注信息像素并抑制噪声的 CenterMask，还有 TensorMask，R-FCN，DeepMask，PolarMask 等。

4.6. DeepLab 系列

Chen 等人开发的 DeepLabv1 [8]和 DeepLabv2 [23]是目前最流行的图像语义分割模型之一，后者有三个关键特征，首先是使用扩张卷积来解决网络中分辨率下降的问题(由最大汇集和跨越引起)。第二个是阿特鲁空间金字塔池(ASPP)，它以多种采样率使用过滤器探测传入的卷积要素图层，从而在多种尺度下捕捉对象和图像上下文，以在多种尺度下稳健地分割对象。第三是通过深度学习方法来改进对象边界的定位，例如 VGG-16 或 ResNet-101 之类的 CNN 模型以全卷积方式使用，使用扩展卷积，双线性插值阶段将特征图放大到原始图像分辨率。最后，完全连接的 CRF 细化分割结果以更好地捕捉对象边界[23]。图 4 展示了 DeepLab 流程图，主要区别是使用了扩张卷积和 ASPP。

随后，Chen 等人[24]提出了 DeepLabv3，它结合了扩张卷积的级联和并行模块。并行卷积模块被分组在 ASPP。在 ASPP 中增加了 1 × 1 卷积和批量归一化。所有的输出被连接并通过另一个 1 × 1 卷积处理，以创建每个像素的具有 logits 的最终输出。进一步 Chen 等人[25]提出了 Deeplabv3+，它采用了一种编解码架构，包括由深度方向卷积(输入的每个通道的空间卷积)和逐点卷积(以深度方向卷积作为输入的 1 × 1 卷积)。他们使用了 DeepLabv3 框架作为编码器。最相关的模型具有修改的例外主干，具有更多层、扩展的深度方向可分离卷积，而不是最大汇集和批量标准化。

4.7. 基于递归神经网络的模型

虽然 CNN 是计算机视觉问题的天然解决方案，但它们并不是唯一的可能性，RNN 在对像素之间的短期/长期依赖性进行建模以(潜在地)改进分割图的估计方面是有用的。使用 RNNs，可以将像素链接在一起并顺序处理，以模拟全局上下文并改进语义分割。然而，一个挑战是图像的自然 2D 结构。

Visin 等人[26]提出了一个基于 RNN 的语义分割模型，称为 ReSeg。这种模式主要是根据 ReNet，它是为图像分类而开发的。每个 ReNet 层由四个 RNN 组成，它们在两个方向上水平和垂直扫描图像，对补丁/激活进行编码，并提供相关的全局信息。使用 ReSeg 模型执行图像分割，ReNet 层堆叠在提取通用局部特征的预训练 VGG-16 卷积层之上。ReNet 层之后是上采样层，以在最终预测中恢复原始图像分辨率。使用门控循环单元(gru)是因为它们在内存使用和计算能力之间提供了良好的平衡。

在另一项工作中，Byeon 等人[27]开发了使用长－短时记忆(LSTM)网络的场景图像的像素级分割和分类。他们研究了自然场景图像的二维(2D) LSTM 网络，考虑了标签的复杂空间依赖性。在这项工作中，分类，分割和上下文集成都是由 2D LSTM 网络进行的，允许在单个模型中学习纹理和空间模型参数。

Xiang 与 Fox [28]提出了数据关联递归神经网络(DA-RNNs)，用于联合 3D 场景映射和语义标记。DA-RNNs 使用一种新的递归神经网络架构对 RGB-D 视频进行语义标记。网络的输出与 Kinect-Fusion 等映射方法相结合，以便将语义信息注入到重建的 3D 场景中。

4.8. 基于注意力的模型

近年来，注意力机制也被应用于图像语义分割。Chen 等[29]提出了一种注意机制，学习在每个像素位置对多尺度特征进行软加权。他们采用了强大的语义分割模型，并用多尺度图像和注意力模型对其进行联合训练(图 8)。其中，注意力模型学习给不同尺度的物体分配不同的权重，例如，该模型对小人(绿色虚线圆)上分配大的权重比例为 1.0，而对大孩子(粉色虚线圆)上分配权重比例为 0.5。注意机制优于平均池和最大池，它使模型能够评估不同位置和尺度的特征的重要性。

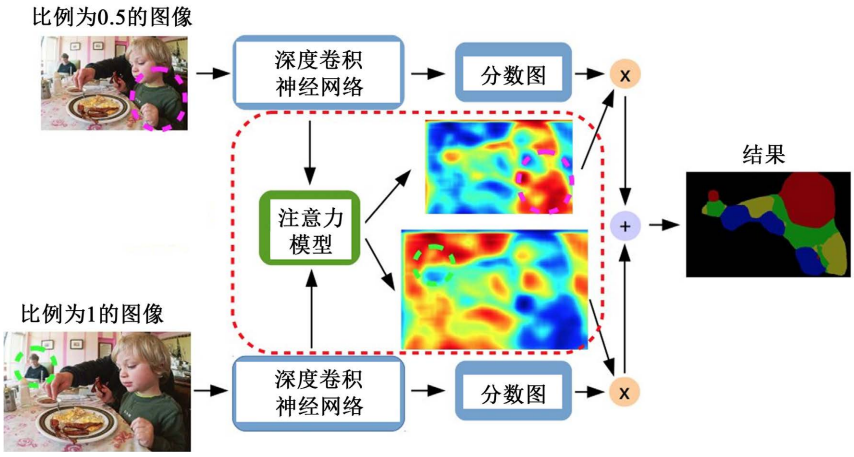

Figure 8. Attention-based semantic segmentation model [29]
图 8. 基于注意力的语义分割模型[29]

Fu 等人[30]提出了一种用于场景分割的双注意网络，该网络能够基于自我注意机制捕获丰富的上下文依赖。具体来说，他们在扩展 FCN 的基础上附加了两种类型的注意力模块，这两种模块分别在空间和通道维度上对语义相互依赖进行建模。位置注意模块通过所有位置的特征的加权和来选择性地聚集每个位置的特征。

相关学者将注意力机制应用于语义分割的研究，Choi 等人[31]提出了根据像素的垂直位置选择性地强调信息性特征或类，利用属性来有效处理城市场景数据集中像素级分布不同的用于改进城市场景图像语义分割高度驱动注意力网络(HANet)。Zhang 等人[32]提出了以 split-attention blocks 构造的 ResNeSt，可以作为其它任务的骨架并达到先进的性能。其他如端对端实例分割，采用了受自我注意力机制启发的对象上下文池的 OCNet，用于场景解析的 PSANet 和用于语义分割的判别特征网络，CCNet，期望最大化注意力(EMANet)等。

4.9. 生成模型与对抗性训练

GANs 已经被广泛应用于计算机视觉中的任务，并且已经被用于图像语义分割。Luc 等人[33]提出了一种用于语义分割的对抗性训练方法。他们训练了一个卷积语义分割网络，以及将地面实况分割图与由分割网络生成的分割图区分开的对抗网络，该方法未增加测试时使用的模型复杂性，并提高了在 Stanford Background 和 PASCAL VOC 2012 数据集上的标记精度。

Hung 等人[34]提出了一个使用对抗网络的半监督语义分割框架。考虑到空间分辨率，他们设计了一个 FCN 鉴别器来区分预测的概率图和地面真实分割分布。该模型考虑的损失函数包含三项：基于分割事实的交叉熵损失、鉴别器网络的对抗性损失和基于置信图的半监督损失：鉴别器的输出。Xue 等[35]提出了一种用于医学图像分割的具有多尺度 L1 损失的对抗网络，使用 FCN 作为分割器来生成分割标签图，并提出了一种新的具有多尺度 L1 损失函数的对立评论家网络，以迫使评论家和分割器学习全局和局部特征，这些特征捕捉像素之间的长程和短程空间关系。

还有其他基于对抗训练的分割模型，例如 Xu 等人[36]提出了利用线性可分离性进行聚类 GAN 实现无监督语义分割，使用 GANs 的细胞图像分割等。

5. 语义分割算法的评价指标与数据集

5.1. 图像语义分割评价指标

图像语义分割技术经过几十年的发展，出现 1000 多种的算法、网络结构和模型，还有新方法、新方案在不断的涌现。同对分割算法的研究一样，针对分割算法的性能评价也一直是研究的热点问题，为此，Garcia-Garcia [37]等人在 2017 年的 CVPR 会议上，专门发表了一篇对诸多数据集和网络模型进行评估的方法综述，提出了被认为是现在统一的标准和公认的算法评估指标。为了保证算法评价的公正性，衡量语义分割算法的性能，需要使用通用的客观评测指标。目前，运行时间、显存占用和准确率是 3 种常用的算法评测指标。本节介绍一些图像语义分割评价指标。

1) 运行时间。神经网络运行的时间包括网络模型的训练时间和测试时间。大多数算法需要实时预测分割结果。在某些情况下，提供算法确切的运行时间可能比较困难，因为运行时间非常依赖硬件设备及后台实现。然而，提供算法运行硬件的信息及运行时间有利于评估方法的有效性，以及在保证相同环境的条件下测试最快的执行方法。

2) 显存占用。数据的规模对神经网络模型的训练至关重要，因此训练神经网络模型需要高性能的硬件设施和软件实现。图形处理单元(GPU)具有高度并行特性以及高内存带宽，但是相比于传统的中央处理器(CPU)，时钟速度更慢以及处理分支运算的能力较弱。在某些情况下，对于操作系统及机器人平台，其显存资源相比高性能服务器并不具优势，即使是加速深度网络的 GPU，显存资源也相对有限。因此，在运行时间相同的情况下，记录算法运行状态下显存占用的极值和均值都很有意义。

3) 准确率。像素准确率(pixel accuracy, PA)是指分类正确的像素占总像素的比例，定义为公式(1)：

$$\mathrm{PA}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

其中，TP (True Positive)：真正例，预测为正例，实际是正例；FP (False Positive)：假正例，预测为正例，实际是反例；FN (False Negative)：假反例，预测为反例，实际是正例；TN (True Negative)：真反例，预测为反例，实际是反例。

4) 交并比 (intersection over union, IoU)是像素的真实值与预测值的交集除以像素的真实值和预测值的并集，定义公式(2)：

$$\mathrm{IoU}=J\left(A,B\right)=\frac{\left|A\cap B\right|}{\left|A\cup B\right|} \tag{2}$$

其中，A：像素的真实值，B：像素的预测值。

而当测试集出现类别不均衡(不同类别：样本数量差别很大)情况时，像素准确率并不能客观反映模型性能。因此定义平均像素准确率(mean pixel accuracy, MPA) (公式 3)和平均交并比(mean intersection-over-union, MIoU) (公式 4)两种评测指标，其中平均交并比是最重要的性能评测指标，更能反映模型的准确程度。

$$\mathrm{MPA}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ij}}{\sum_{j=0}^{k}p_{ij}} \tag{3}$$

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

其中，假设共有 $k+1$ 个类(其中包含一个空类或背景)，p_{ii} 表示预测与实际相符的像素，p_{ij} 表示实际属于类 i 但被预测为类 j 的像素，p_{ji} 表示实际属于类 j 但被预测为类 i 的像素。i 表示真实值，j 表示预测值，p_{ij} 表示将 i 预测为 j 的像素个数。

5.2. 图像语义分割数据集

图像语义分割数据库是图像目标提取乃至图像内容理解的算法研发、模型建立与算法测试的重要依赖环境。图像库的规模大小、所包含的目标的成像条件(光照、遮挡、姿态、尺等)变化、目标标准分割的结果(Ground-truth)等都对分割算法的鲁棒性以及算法评测的合理性与可信性有着很大的影响，在测试两种不同的网络结构或者模型的时候需要存在着一个统一的标准，同样的在研发出新的基于深度学习图像语义分割技术的网络模型时候也需要以一个共同的标准来进行判断，一些算法可以在给出的特定数据集上执行；它不会在其他数据集上提供相同的结果[38] [39]。本节介绍一些常用公共数据集，见表 1。

Table 1. Common datasets for semantic image segmentation
表 1. 图像语义分割常用数据集

年代	数据集名称	内容特点	应用场景
2009	CamVid	包括了 146,617 个二维多边形和 58,657 个具有精确对象方向的 3D 边界框，数据集中包含了 NYU Depth V2、Berkeley B3DO，适合于场景理解任务	道路场景
2009	SBD	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	室外场景

Continued

2011	SiftFlow	收集了 731 个包含着 102,206 帧的视频作为实验数据库是 LabelMe 的数据集子集，图像主要包含着 8 种不同的户外场景	自然景观
2012	PASCAL VOC	含有 20 种类别，道路场景数据有着 11,520 张图片，包含着 27,450 个注释对象	道路行人车辆
2012	NYU Depth V2	主要提供了 1449 个 RGBD 图像的新数据集，其中捕获了 464 个不同的室内场景，并附有详细的标注，能够验证 3D 场景的提示和推断，实现更好的对象分割	室内物体
2013	KITTI	有 389 对立体图像和光流图、39.2 km 视觉测距序列和超过 200,000 幅带有 3D 标注目标的图像组成，11 个类别，包含了市区、乡村和高速公路等场景的真实图像数据，每幅图像中最多有 15 辆车和 30 个行人以及各种程度的遮挡	道路行人车辆 3D 模型
2014	PASCAL-CONTEXT	包含了 10,103 张训练图像的像素级别的标注，共 540 类	道路行人车辆
2014	PASCAL-Part	数据集中训练集和验证集共 10,103 幅，测试集 9637 幅。该数据集还为目标提供轮廓标注	道路行人车辆场景
2014	MS COCO	包括 200 000 个图像和 8 个图像实例，已经公开了 5,000,000 个对象实例，数据集中主要包括了室内场景和室外场景	室内室外的常用场景
2015	Cityscape	城市道路场景数据集，来自 50 个不同的城市街景记录的立体视频序列，包括 20,000 张弱注释图片和 5000 张的高质量的强注释的图片，涵盖了各种时间及天气变化下的街道动态物体	道路车辆、行人、街景
2015	SUN-RGBD	由 4 个 RGB-D 传感器获取而得，其中包含了 10,000 个 RGB-D 图像，比例类似于 PASCAL VOC，整个数据集包括了 146,617 个二维多边形和 58,657 个具有精确对象方向的 3D 边界框，数据集中包含了 NYU Depth V2、Berkeley B3DO，适合于场景理解任务	室内物体 3D 模型
2015	ILSVRC	1400 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	室内室外景观
2018	ADE20K	包含 SUN 和 Places 数据集的场景范畴，可视化目标，目前已有超过 250 个带有注解示例的目标，以及带有超过 10 个注解示例的部件	室内室外景观
2019	CityFlow	从 10 个路口提取的 40 个摄像头收集到的视频，是目前都市环境中最大规模的数据集，包含超过 20 万个目标框	道路场景

6. 图像语义分割研究中存在的主要问题及解决方法

6.1. 存在的主要问题

1) 深度学习的图像语义分割算法都需要 CNN 作为最基础的框架，在进行语义分割时，除了语义信息还需要细节信息、上下文信息，对于物体边缘的分割效果不理想；还有在图像初始阶段输入到网络之时，由于 CNN 的卷积核不会太大，模型只能利用局部信息理解输入图像，影响编码器最后提取的特征的可区分性。

2) 模型的通用性能低，一些算法可以在特定数据集上执行；它不会在其他数据集上提供相同的结果[38]这是因为不同的数据集在进入训练阶段和测试阶段之前没有执行相同的操作；其次是在机器学习过程中，认为整个数据集没有任何歧义，这样就产生了有效和准确的最佳结果[40]；有时数据集的样本很少或

很多时，会出现模型过度拟合和拟合不足的问题[41]。这不仅是语义分割任务上存在的问题，只要基于深度学习的任务时都会面临的难题。

3) 耗费显存问题。

4) 图像遮挡区域语义分割问题。

6.2. 解决方法

对于深度学习的图像语义分割算法研究中存在上面的主要问题，研究人员一直根据各个算法的缺陷特点找出相应解决办法，有些办法需要进一步的改进和完善。具体的解决方案如下：

1) 语义分割对于物体边缘的分割效果不理想的边缘问题，可采取对网络输出的分割的边界增加额外的损失、让网络对边界的特征和区域内部的特征分开建模学习、简单有效的方法是提高输入图像的输入分辨率和中间层特征图的分辨率。到目前为止，还没有具体的针对边缘的好坏的评价指标，只能靠目测来确定，易产生不公平现象，如何客观公正也值得进一步探讨。

2) 增强模型的通用性，采用数据增强、正则化等很多基础的方法能增强模型的通用性，效果不理想。对不可获取的测试数据集，要从模型本身出发，迫使模型能够学习到更为鲁棒的特征；对可获取，但没有标签的测试数据集，要用无监督域语义分割方法：如 FCAN 与 ADVENT 的基于对抗学习法，使目标域与源域在同一 Encoder 后编码的特征能够尽量相似；CycleGAN 风格迁移法，能转换源域图片的风格使得其与目标域相似；自监督学习法，在目标域上形成伪标签来训练模型。

3) 耗费显存问题：内存占用是评估基于深度学习的图像语义分割算法性能的主要指标之一，是语义分割算法研究中的一个重要影响因素，在多数场景下，内存是可以扩充的，由于显存是 GPU 计算中的稀缺资源，在语义分割网络训练常常遇到显存不足，为了解决这个问题，常用的方法就是调参：网络参数调整，比如减小训练图像大小，降低 FC output 个数，使用小的 conv kernel size 等。还有从卷积的实现(采用 FFT、二值量化、不采用 bias 等)、卷积的方式(深度可分离卷积、各项异性卷积、空洞卷积等)、特征提取块(bottleneck 结构、add 替代 concat 融合特征、简化解码器 module 等)、网络结构(降低网络深度、宽度、输入分辨率，多分支网络等)等方面进行改进，在网络中构建不同图像之间损失或者特征交互模块[42]。

4) 图像遮挡区域语义分割问题：对于带有遮挡的图像区域，当前的语义分割方法效果不理想，实现对遮挡的图像区域进行正确的语义分割将会使基于语义分割的图像理解技术更加接近人类图像理解的水平，有利于拓宽语义分割技术在现实中的应用场景，利用合理的上下文建模机制，能帮助网络猜测遮挡部分的语义信息[43] [44]，香港科技大学研究人员将图像建模为两个重叠图层，为网络引入物体间的遮挡与被遮挡关系，提出了一个轻量级的能有效处理遮挡的实例分割算法，大幅提升遮挡处理性能[45]。

7. 总结与展望

图像语义分割是计算机视觉领域的重要研究方向之一，深度学习的出现明显提升了语义分割技术发展机遇。本文介绍了图像语义分割的研究发展历程，以及传统的与深度学习的图像语义分割方法、评价指标、常用数据库，提出基于深度学习图像语义分割研究中存在的部分问题及解决方法，在图像语义分割研究中仍然存在着很多的未知内容，需要今后深入探究。随着图像语义分割技术的不断发展，视频语义分割、三维数据集语义分割、实时语义分割等将是未来的研究方向，其挑战性更大，前景也更广阔。

基金项目

本文得到国家自然科学基金(No. 51979085)的资助。

参考文献

[1] Lu, X., Wang, W., Shen, J., *et al*. (2022) Zero-Shot Video Object Segmentation with Co-Attention Siamese Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 2228-2242.

[2] Batra, A., Singh, S., Pang, G., *et al*. (2019) Improved Road Connectivity by Joint Learning of Orientation and Segmentation. 2019 *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-20 June 2019, 10377-10385. https://doi.org/10.1109/CVPR.2019.01063

[3] Sehar, U. and Naseem, M.L. (2022) How Deep Learning Is Empowering Semantic Segmentation: Traditional and Deep Learning Techniques for Semantic Segmentation: A Comparison. *Multimedia Tools and Applications*, **81**, 30519-30544. https://doi.org/10.1007/s11042-022-12821-3

[4] Sharif Razavian, A., Azizpour, H., Sullivan, J., *et al*. (2014) CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 806-813. https://doi.org/10.1109/CVPRW.2014.131

[5] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Net-Works for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[6] Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian Network Classifiers. *Machine Learning*, **29**, 131-163. https://doi.org/10.1023/A:1007465528199

[7] Liu, W., Rabinovich, A. and Berg, A.C. (2015) ParseNet: Looking Wider to See Better.

[8] Chen, L.-C., Papandreou, G., *et al*. (2014) Semantic Image Segmentation with Deep Convolutionalnets and Fully Connected CRFs.

[9] Schwing, A.G. and Urtasun, R. (2015) Fully Connected Deep Structured Networks.

[10] Lin, G., Shen, C., *et al*. (2016) Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3194-3203. https://doi.org/10.1109/CVPR.2016.348

[11] Noh, H., Hong, S. and Han, B. (2015) Learning Deconvolution Network for Semantic Segmentation. 2015 *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1520-1528. https://doi.org/10.1109/ICCV.2015.178

[12] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[13] Fu, J., Liu, J., *et al*. (2019) Stacked Deconvolutional Network for Semantic Segmentation. *IEEE Transactions on Image Processing*. https://doi.org/10.1109/TIP.2019.2895460

[14] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Berlin, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[15] Milletari, F., Navab, N. and Ahmadi, S.-A. (2016) V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *International Conference on 3D Vision IEEE*, Stanford, 25-28 October 2016, 565-571. https://doi.org/10.1109/3DV.2016.79

[16] Lin, T.-Y., Dollar, P., *et al*. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[17] Zhao, H., Shi, J., *et al*. (2017) Pyramid Scene Parsing Network. 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[18] Ren, S., He, K., *et al*. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*'15: *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Volume 1, 91-99.

[19] He, K., Gkioxari, Dollar, G.P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2961-2969. https://doi.org/10.1109/ICCV.2017.322

[20] Liu, S., Qi, L., *et al*. (2018) Path Aggregation Network for Instance Segmentation. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[21] Chen, L.-C., Hermans, A., *et al*. (2018) Masklab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 4013-4022. https://doi.org/10.1109/CVPR.2018.00422

[22] Lee, Y. and Park, J. (2020) CenterMask: Real-Time Anchor-Free Instance Segmentation. 2020 *IEEE Conference on*

Computer Vision and Pattern Recognition, Seattle, 13-19 June 2020, 13906-13915. https://doi.org/10.1109/CVPR42600.2020.01392

[23] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[24] Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H. (2018) Rethinking Atrous Convolution for Semantic Image Segmentation.

[25] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. 2018 *European Conference on Computer Vision*, Munich, 8-14 September 2018, 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

[26] Visin, F., Kastner, K., *et al*. (2015) ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks.

[27] Byeon, W., Breuel, T.M., *et al*. (2015) Scene Labeling with LSTM Recurrent Neural Networks. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3547-3555. https://doi.org/10.1109/CVPR.2015.7298977

[28] Xiang, Y. and Fox, D. (2017) DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. https://doi.org/10.15607/RSS.2017.XIII.013

[29] Chen, L.-C., Yang, Y., *et al*. (2016) Attention to Scale: Scale-Aware Semantic Image Segmentation. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3640-3649. https://doi.org/10.1109/CVPR.2016.396

[30] Fu, J., Liu, J., *et al*. (2019) Dual Attention Network for Scene Segmentation. 2019 *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3146-3154. https://doi.org/10.1109/CVPR.2019.00326

[31] Choi, S., Kim, J.T. and Choo, J. (2020) Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 9373-9383. https://doi.org/10.1109/CVPR42600.2020.00939

[32] Zhang, H., Wu, C., *et al*. (2020) Resnest: Split-Attention Networks.

[33] Luc, P., Couprie, C., *et al*. (2016) Semantic Segmentation Using Adversarial Networks.

[34] Hung, W.-C., Tsai, Y.-H., *et al*. (2018) Adversarial Learning for Semi-Supervised Semantic Segmentation.

[35] Xue, Y., Xu, T., *et al*. (2018) SegAN: Adversarial Network with Multi-Scale L1 Loss for Medical Image Segmentation. *Neuroinformatics*, **16**, 383-392. https://doi.org/10.1007/s12021-018-9377-x

[36] Xu, J.J., Zhang, Z.X. and Hu, X.L. (2022) Extracting Semantic Knowledge from GANs with Unsupervised Learning.

[37] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., *et al*. (2017) A Review on Deep Learning Techniques Applied to Semantic Segmentation.

[38] Xu, K., Wen, L., *et al*. (2019) Spatiotemporal CNN for Video Object Segmentation. 2019 *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 1379-1388. https://doi.org/10.1109/CVPR.2019.00147

[39] Cao, Z., Hidalgo, G., Simon, T., *et al*. (2018) OpenPose: Real-Time Multi-Person 2D Pose Estimation Using Part Affinity Fields.

[40] Zhu, Y., Zhou, Y., Xu, H., *et al*. (2019) Learning Instance Activation Maps for Weakly Supervised Instance Segmentation. 2019 *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3111-3120. https://doi.org/10.1109/CVPR.2019.00323

[41] Chen, K., *et al*. (2019) Hybrid Task Cascade for Instance Segmentation. 2019 *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 4969-4978. https://doi.org/10.1109/CVPR.2019.00511

[42] Zhuang, P., Wang, Y. and Qiao, Y. (2020) Learning Attentive Pairwise Interaction for Fine-Grained Classification. 2020 *AAAI Conference on Artificial Intelligence*, New York, 7-12 February 2020, 13130-13137. https://doi.org/10.1609/aaai.v34i07.7016

[43] He, J., Deng, Z., Zhou, L., *et al*. (2019) Adaptive Pyramid Context Network for Semantic Segmentation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 7511-7520. https://doi.org/10.1109/CVPR.2019.00770

[44] Liu, J., He, J., *et al*. (2020) Learning to Predict Context-Adaptive Convolution for Semantic Segmentation. 2020 *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 769-786. https://doi.org/10.1007/978-3-030-58595-2_46

[45] Ke, L., Tai, Y.W. and Tang, C.K. (2021) Deep Occlusion-Aware Instance Segmentation with Overlapping Bi-Layers. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 4018-4027. https://doi.org/10.1109/CVPR46437.2021.00401