全景分割研究综述

徐鹏斌 [1] 瞿安国 [1] 王坤峰 [1] 李大字 [1]

摘 要 在计算机视觉领域, 全景分割是一个新颖且重要的研究主题, 它是机器感知、自动驾驶等新兴前沿技术的基石, 具有十分重要的研究意义. 本文综述了基于深度学习的全景分割研究的最新进展, 首先总结了全景分割任务的基本处理流程, 然后对已发表的全景分割工作基于其网络结构特点进行分类, 并进行了全面的介绍与分析, 最后对全景分割任务目前面临的问题以及未来的发展趋势做出了分析, 并针对所面临的问题提出了一些切实可行的解决思路.

关键词 全景分割, 语义分割, 实例分割, 深度学习

引用格式 徐鹏斌, 瞿安国, 王坤峰, 李大字. 全景分割研究综述. 自动化学报, 2021, **47**(3): 549−568

DOI 10.16383/j.aas.c200657

A Survey of Panoptic Segmentation Methods

XU Peng-Bin[1] QU An-Guo[1] WANG Kun-Feng[1] LI Da-Zi[1]

Abstract In the field of computer vision, panoptic segmentation is a novel and important research topic. It is the cornerstone of emerging leading-edge technologies such as machine perception and autonomous driving, and has very important significance of research. This paper reviews the research progress of panoptic segmentation based on deep learning methods, summarizes the basic processing flow of panoptic segmentation, and then classifies the already published panoptic segmentation works based on the characteristics of its network structure, and makes a more comprehensive introduction and analysis. Finally, the current problems and future development trends of the panoptic segmentation task are analyzed, and some potential solutions are provided for solving the existing problems.

Key words Panoptic segmentation, semantic segmentation, instance segmentation, deep learning

Citation Xu Peng-Bin, Qu An-Guo, Wang Kun-Feng, Li Da-Zi. A survey of panoptic segmentation methods. *Acta Automatica Sinica*, 2021, **47**(3): 549−568

全景分割[1] 是将图像划分为语义区域 (Stuff) 和对象实例 (Things) 的任务, 是近年来新兴起的一个研究方向, 也是计算机视觉中一个重要的研究问题. 随着图像处理技术的发展, 数字图像已经成为日常生活中不可缺少的媒介, 每时每刻都在产生图像数据. 对图像中的物体进行快速准确的分割变得愈发重要.

全景分割包含语义分割和实例分割两大任务. 语义分割是将类别标签按图像中物体类别分配给每个像素, 即将输入图像中的像素分为不同的语义类别. 传统的语义分割方法大多基于模型驱动, 模型驱动方法可分为生成式和判别式[2]. 生成式模型首先学习图像的特征和标签概率, 然后计算输入图像特征时各个标签的后验概率, 依据此概率对图像进行标注. 马尔科夫随机场 (Markov random field, MRF) 是一种应用广泛的生成式模型[3], 它利用先验上下文信息和训练得到的结果, 提高分割性能. 但是当图像分辨率较大时, 分割速度和精度会大幅下降. 判别式模型假设图像特征与标签之间存在某种映射关系, 然后从历史数据中学习该映射关系的相关参数[2]. 典型的判别式模型包括支持向量机 (Support vector machine, SVM)、条件随机场 (Conditional random field, CRF) 等. SVM 因其可处理非线性且具有良好的泛化能力, 在语义分割研究中得到了广泛应用[3]. CRF 不仅可以利用图像局部上下文信息, 还可学习从局部到全局的上下文信息, 已经成功应用于图像标记[4]. 然而, 判别式模型存在收敛速度慢、无法解决存在隐变量的情况等问题.

近年来, 随着硬件计算能力的提高, 语义分割得到快速发展. 随着全卷积网络 (Fully convolutional network, FCN) 的出现[5], 深度学习推动语义分割任务快速发展, 并且在自动驾驶、人脸识别等领域得到应用.

收稿日期 2020-08-15 录用日期 2020-12-14
Manuscript received August 15, 2020; accepted December 14, 2020
国家自然科学基金 (62076020, 61873022), 北京市自然科学基金 (4182045) 资助
Supported by National Natural Science Foundation of China (62076020, 61873022) and Beijing Municipal Natural Science Foundation (4182045)
本文责任编委 胡清华
Recommended by Associate Editor HU Qing-Hua
1. 北京化工大学信息科学与技术学院 北京 100029
1. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029

实例分割实质上是目标检测和语义分割的结合, 目的是将输入图像中的目标检测出来, 并且对目标的每个像素分配类别标签. 实例分割能够对前景语义类别相同的不同实例进行区分, 这是它与语义分割的最大区别. 相比语义分割, 实例分割发展较晚, 因此实例分割模型主要基于深度学习技术, 但它也是图像分割一个重要的组成部分. 随着深度学习的发展, 实例分割相继出现了 SDS (Simultaneous detection and segmentation)[6]、DeepMask[7]、MultiPath network[8] 等方法, 分割精度和效率逐渐得到提升.

全景分割是语义分割和实例分割的综合. 全景分割任务不仅要求区分输入图像中的背景语义类别和前景语义类别, 还要将同一类别前景语义中的不同实例分割出来, 因此全景分割任务比语义分割、实例分割任务的难度更高. 全景分割由 Kirillov 等[1] 提出, 已经得到计算机视觉学界的高度重视, 涌现出 JSIS-Net (Joint semantic and instance segmentation network)[9]、TASCNet (Things and stuff consistency network)[10]、AUNet (Attention-guided unified network)[11] 等方法, 显著推动了全景分割的发展. 但是在真实环境下, 全景分割经常遇到以下挑战:

1) 分支融合冲突

全景分割任务是语义分割与实例分割两个任务的综合, 在网络结构方面, 现有大部分方法将输入图像的特征输送到语义分支与实例分支, 然后融合两个分支的输出, 得到全景输出. 但是在融合时会出现像素分配冲突, 影响全景预测质量.

2) 小物体检测分割

数据集中的图像会出现大小、距离不一的许多物体, 对于大物体, 诸多全景分割方法能够对其进行准确分割, 而当小物体出现时, 经常伴随被漏检或者分割不准确的问题, 这导致全景分割精度较低, 直接增加了全景分割的难度.

3) 分割对象交叠

在图像采集过程中, 会因为季节、天气、光照、距离等条件的变化, 出现不同的场景, 图像中物体会出现遮挡交叠等情况, 这使得分割方法无法准确判断像素的归属, 导致分割不精确.

为了克服上述挑战, 已经出现了一些全景分割方法, 它们在分支融合、小物体检测、遮挡处理方面提出了不同的应对策略, 在一定程度上解决了这些问题. 本文首先介绍全景分割的流程, 然后重点介绍深度学习在全景分割领域的研究进展.

本文内容安排如下: 第 1 节介绍全景分割的基本流程; 第 2 节对语义分割、实例分割等相关知识以及全景分割数据集进行介绍; 第 3 节介绍深度学习在全景分割领域的研究进展; 第 4 节讨论全景分割研究面临的挑战, 并对今后的发展趋势进行展望; 第 5 节对本文进行总结.

1 全景分割的基本流程

全景分割任务的重点在于为每个像素分配一个语义标签和实例 ID, 处理流程如图 1 所示, 主要分为三个步骤: 特征提取、语义分割和实例分割处理、子任务融合. 对于一幅输入图像, 首先提取特征. 然后将提取的特征输入语义分割与实例分割的子任务分支进行处理, 产生语义分割与实例分割输出. 最后是子任务融合, 将语义与实例分支产生的结果通过适当的策略进行融合, 产生最终的全景预测. 迄今已有许多工作采用上述基本流程, 如 JSIS-Net[9]、AUNet[11]、Single network[12]、OANet[13] 等. 本节将从特征提取、语义分割与实例分割处理、子任务融合三个方面总结应用于全景分割的关键技术.

1.1 特征提取

特征提取的主要任务是获得输入图像的特征, 为后续的两个任务提供必要信息, 如图 1 所示. 目前, 全景分割对特征的提取全部基于深度神经网络, 主要使用的骨干网络包括 VGGNet (Visual geometry group network)、ResNet (Residual network) 等.

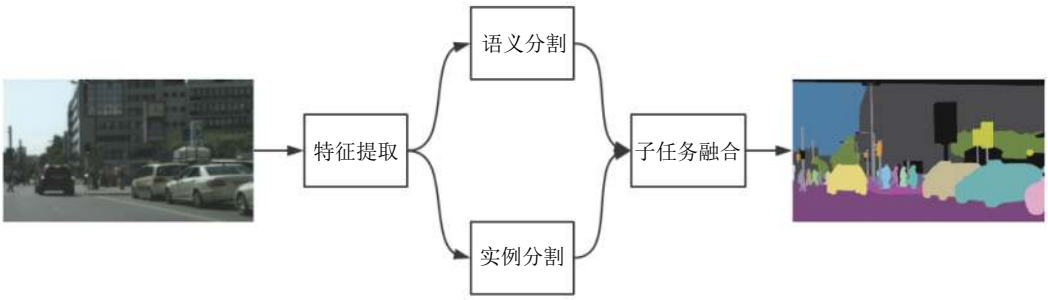


图 1 全景分割流程图

Fig. 1 The processing flow of panoptic segmentation

VGGNet[14] 具有灵活的结构和良好的性能, 在 ILSVRC (ImageNet large scale visual recognition challenge) 图像分类任务的 Top5 错误率为 7.5%, 受到广泛关注. VGGNet 性能的提升主要依靠使用大量小卷积核和池化核, 加深了网络结构, 因此提高了特征学习能力和非线性表达能力, 与之前的卷积神经网络相比, 对图像的适应性也更好.

ResNet 是 He 等[15] 提出的深度残差网络. 由于神经网络在反向传播过程中要不断地传播梯度, 当网络层数加深时, 梯度在传播过程中会逐渐消失, 导致网络难以训练. ResNet 的出现缓解了这个问题, 它的思想是让卷积层拟合残差映射, 而不是直接拟合所需的低层映射, 因此网络变得更易于优化, 同时精度也得到了提升. 由于具有以上良好特性, ResNet 在全景分割任务中得到了较多应用[12, 16−17].

许多特征提取网络提高模型性能的策略以网络加深或加宽为主, 而 Huang 等[18] 从特征的角度出发, 提出 DenseNet, 将当前层与之前的每一层相连, 通过建立不同层之间的连接关系, 充分利用特征以达到更好的效果. 相较于其他特征提取网络, DenseNet 减轻了梯度消失的问题, 减少了参数数量, 增强了特征信息流的传递.

为了满足实时性, 研究小而高效的模型至关重要, 为此, Google 提出了 MobileNet[19]. 该网络提出了深度可分离卷积, 将传统卷积的两个步骤分为 Depth-wise 和 Point-wise 卷积, 大幅减少了参数量, 在 Point-wise 改变通道数, 可大幅减小计算量. 但是 MobileNet 精度较低. 随后, Google 又提出了 MobileNet V2 和 MobileNet V3 网络[20−21], MobileNet V3 综合了前两个版本的优点, 改进了池化层和激活函数, 网络准确率和计算速度进一步得到了提升.

1.2 语义分割与实例分割处理

通过骨干网络提取特征后, 特征可以被语义分割和实例分割任务共享, 进行后续处理. 如文献 [13] 提出的 OANet, 在骨干网络之上添加语义分割分支与实例分割分支: 实例分割分支采用 Mask R-CNN 作为网络架构, 对骨干网络特征进行 RoIAlign 等操作产生实例分割预测; 语义分割分支通过对骨干网络特征进行上采样等操作产生语义分割预测. 文献 [10, 12, 22] 也对骨干网络提取的特征进行共享和处理, 以实现语义和实例预测.

1.3 子任务融合

子任务融合步骤是将上述两个分支的输出结果进行处理, 以产生最终的全景预测. 研究人员已经提出许多方法来解决这个问题, 这些方法大致分为两类:

1) 启发式方法

启发式方法是相对于最优化方法提出的, 它是依据有限的信息在短时间内找到问题解决方案的一种方式, 一般启发式方法有模拟退火法[23]、遗传算法[24]、神经网络[25] 等. 本文中以神经网络为主,如在文献 [11, 13, 26] 中, 融合步骤应用了启发式方法对语义、实例分支的预测结果进行融合, 形成全景分割预测.

2) 设置 Panoptic head 进行子任务融合

Panoptic head 由 Xiong 等[27] 提出, 将语义分割与实例分割的结果进行融合, 产生全景预测, 性能方面达到了一流水准. 后来一些工作受此启发, 也应用 Panoptic head 类似方法进行分支融合, 如文献 [22, 28−29] 等, 将来自两分支的预测、掩码或者逻辑输出输入 Panoptic head 进行处理, 进而得到全景预测. 第 3 节将对此进行详细介绍.

2 相关知识

全景分割预测主要依赖于语义分割与实例分割预测的融合, 本节从卷积神经网络、语义分割、实例分割等方面对全景分割的相关知识进行介绍.

2.1 卷积神经网络简介

卷积神经网络最早可以追溯至 1980 年日本学者福岛邦彦提出的神经认知机 (Neocognitron) 模型[30]. 随后 LeCun[31] 提出 LeNet-5, 将反向传播算法引入网络训练, 形成了卷积神经网络的雏形. 在 2012 年 ImageNet 图像识别大赛上, Krizhevsky 等[32] 提出 AlexNet, 该网络具有全新的深层结构并引入了 dropout 方法, 将图像识别错误率从 25% 降至 15%, 获得了当年的图像识别冠军. 从此, 深度卷积神经网络引起广泛关注, 许多学者采用深度卷积神经网络解决目标检测、图像分割等计算机视觉问题. 近年来, 卷积神经网络出现了诸多模型, 如 VGGNet、GoogLeNet、ResNet 等.

2.1.1 卷积神经网络的基本组成

卷积神经网络主要由输入层、卷积层、池化层和全连接层组成. 输入层主要对输入数据进行标准化处理, 即归一化. 输入数据可以是一维的, 也可以是多维的. 由于卷积神经网络在计算机视觉领域应用较多, 许多工作都假设输入为三维数据, 包括图像的宽度、高度和颜色通道数.

卷积层主要对输入图像进行特征提取, 其内部通常包含多个卷积核, 即滤波器. 卷积核对特定区域进行卷积运算, 产生不同的特征映射, 这个区域称为 “感受野”, 其大小取决于卷积核的大小.

卷积层提取特征后, 堆叠形成的输出会被传输至池化层进行处理. 池化层本质上是一种下采样操作, 可以降低特征映射的分辨率, 提取抽象语义. 典型的池化方法有最大池化和平均池化, 最大池化计算区域内的特征最大值, 平均池化计算区域内的特征平均值. 由于卷积神经网络经常用于图像识别, 最大池化能够更好地保留图像边缘信息, 因此最大池化比平均池化更常用.

全连接层的主要作用是将特征图转化为类别输出, 全连接层可以有多层, 通常在全连接层引入 dropout 以防止过拟合, 然后通过 softmax 函数层生成图像数据或样本的类别概率.

2.1.2 经典网络结构

卷积神经网络已经出现了许多网络结构, 本小节介绍几种典型结构.

1) LeNet-5

LeCun 等提出了 LeNet-5 模型[31], 在手写数字识别领域得到成功应用, 在 MNIST 数据集上识别正确率达到 99.2%. 该模型共有 7 层, 包括卷积层、池化层和全连接层, 如图 2 所示. 原始图像输入网络后, 首先归一化为 32 × 32 的矩阵, 然后输入卷积层、池化层进行处理. 卷积运算用于提取更高层的图像语义特征, 同时降低噪声干扰. 池化层用于降低特征图的维数, 一方面可以降低运算量, 另一方面可以防止过拟合. 最后经过全连接层的激活函数运算后到输出层, 输出原始图像所属的类别概率.

2) AlexNet

AlexNet 是 Krizhevsky 等[32] 于 2012 年提出的网络. AlexNet 包括 8 层, 有 5 个卷积层和 3 个全连接层. AlexNet 与 LeNet-5 的主要区别在于 AlexNet 使用了 ReLU 激活函数, 使得网络训练时的梯度计算更加简单, 并且在一定程度上缓解了梯度消失问题. 此外, 为了防止过拟合, 在网络训练时采用了数据增强和 dropout 方法, 卷积通道数比 LeNet 更多, 网络运行更快.

3) VGGNet

VGGNet 是由 Simonyan 等[14] 提出的卷积神经网络, 包括 VGG-16 (结构如图 3 所示) 和 VGG-19. VGGNet 采用 3×3 的卷积核和 2×2 的池化核, 该设置减少了网络参数, 有利于更好地提取特征, 同时节省计算资源. VGGNet 比 AlexNet 有更深的

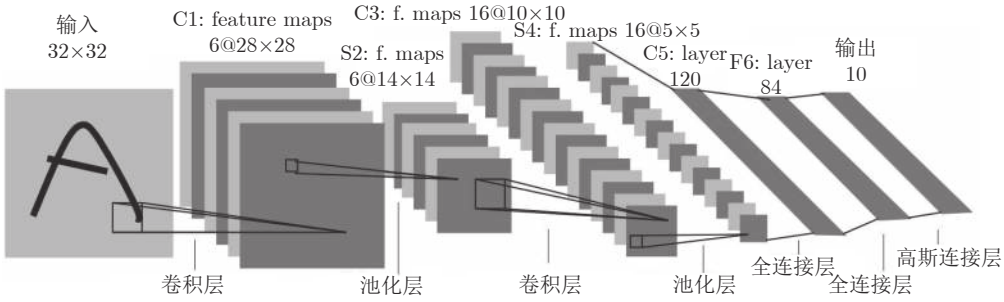


图 2 LeNet-5 的网络结构[31]

Fig. 2 The structure of LeNet-5[31]

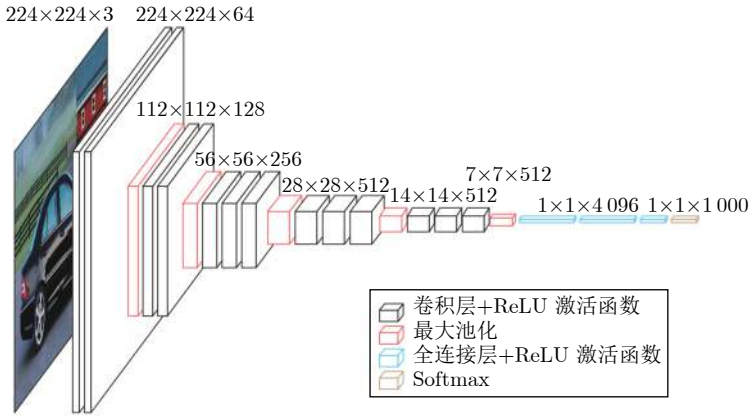


图 3 VGG-16 的网络结构

Fig. 3 The structure of VGG-16

网络结构, 进一步提升了网络性能.

4) GoogLeNet

GoogLeNet 是 Szegedy 等[33] 提出的一种卷积神经网络结构, 在 2014 年 ImageNet 图像识别挑战赛上获得了冠军. GoogLeNet 有 22 个卷积层, 通过提出并应用 Inception 模块, GoogLeNet 网络的参数更少, 精度更高. AlexNet、VGGNet等结构在增加网络深度 (层数) 时使得参数更多, 更容易出现过拟合, 同时耗费了大量计算资源. 为了解决这个问题, GoogLeNet 网络大量应用 1×1、3×3 和 5×5 的卷积核, 降低了运算量, 并且修正了网络的非线性问题, 输出结果能够得到显著的提升.

5) ResNet

ResNet 由 He 等[15] 提出, 在 2015 年 ImageNet 挑战赛的多个任务上获得冠军. 与之前的卷积网络相比, ResNet 能够更好地解决梯度消失和爆炸问题. 如图 4 所示, ResNet 包括两种映射: 一种是恒等映射, 即图中曲线部分; 另一种是残差映射, 即曲线以外的部分. 假设原始映射为 $H(x)$, 非线性拟合的另外一部分映射 $F(x) = H(x) - x$ 为残差部分, 那么原来的映射就可以转化为 $H(x) = F(x) + x$. 若将残差部分推至 0, 则原始映射直接为 x, 即曲线部分, 这时可以大大降低运算量, 但这只是一种极端情况, 实际很难达到. 在 ImageNet 数据集上, ResNet 比普通网络的表现更好, 将网络层数加深, 错误率逐渐下降, 直到 1 000 层以后, 出现过拟合.

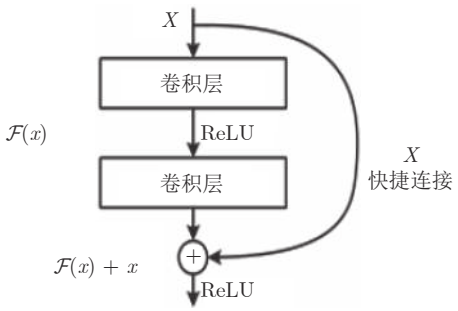


图 4　ResNet 网络的残差模块[15]

Fig. 4　The residual module of ResNet[15]

2.2　语义分割

随着深度学习技术的快速发展, 语义分割研究取得了显著突破, 出现了许多用于语义分割的卷积神经网络模型, 如 FCN、DeepLab、U-Net 等.

2.2.1　语义分割的基本组成

语义分割处理的基本流程包含三部分: 输入、分割处理和输出, 如图 5 所示.

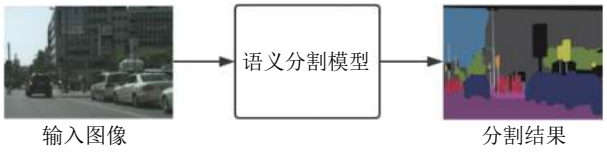


图 5　语义分割处理流程

Fig. 5　The processing flow of semantic segmentation

首先, 图像在输入层经过归一化等操作, 将像素设置到固定值. 其次, 将处理后的图像输入 CNN 语义分割模型, 在其中, 卷积层提取图像特征, 池化层降低特征维数, 许多模型还采用批规范化和 ReLU 激活函数, 在加快模型运行速度的同时, 缓解了过拟合问题. 最后, 网络通过 softmax 层计算所有类别的概率, 将类别标签分配给对应的像素, 输出图像像素所属的语义类别.

2.2.2　语义分割的典型方法

1) FCN

全卷积网络 (Fully convolutional network, FCN) 是 Shelhamer 等[5] 提出的一种重要的语义分割模型. 传统 CNN 在卷积层后使用全连接层和 softmax 进行分类和分配标签, 而 FCN 可以接受任意尺寸的输入图像, 通过将网络中的全连接层替换为卷积层, 在最后一个卷积层的输出进行上采样, FCN 得到像素语义类别的密集预测. 由于使用双线性插值进行上采样导致分割结果不够精细, 因此将抽象表达能力强的深层特征与分辨率高的低层特征进行跳跃连接 (Skip connection), 提高语义分割的精细度.

2) DeepLab

Chen 等[34] 提出了 DeepLab 语义分割网络. 该网络以 VGG-16 为基础, 将 VGG-16 的全连接层转换为卷积层, 在图像的原始分辨率上产生非常稀疏的计算分数, 为了得到更密集的计算分数, DeepLab 采用空洞卷积对特征图采样, 扩大感受野, 并缩小步长. 空洞卷积能够解决 CNN 反复池化和下采样带来的分辨率下降问题. 在此基础上又提出了 DeepLab V2[35], 它以 ResNet 和 VGG-16 为基础, 采用 ASPP (Atrous spatial pooling pyramid) 对网络进行修正, 以增强多尺度特征. DeepLab V3[36] 以级联的方式采用空洞卷积构建语义分割模块, ASPP 部分以并行的方式连接, 使得网络更深, 并且可以合并多个上下文.

3) U-Net

U-Net[37] 是 FCN 的一种变体, 最初用于医学图像分割, 由于分割效果良好而用于语义分割. U-Net 是一个 U 形的语义分割网络, 采用编码器—解码器结构, 简单高效. 编码器部分对输入图像进行特征

提取, 由卷积和下采样操作组成, 卷积核尺寸为 3×3, 步长为 1. 由于没有零填充 (Zero padding), 特征映射不断收缩, 尺寸减小. 解码器部分将特征尺寸恢复至原始图像大小, 主要由上采样和跳跃连接组成, 上采样增加特征维度, 最后使用 1×1 卷积将特征向量转换为类别标签. 扩张路径将特征与空间信息融合, 产生精准的分割. 但是, U-Net 运行速度较慢, 由于块的交叠, 部分运算多余.

4) PSPNet

PSPNet 由 Zhao 等[38] 提出, 网络利用全局特征信息来获得更好的分割结果. 在该网络中, ResNet 骨干网络提取特征形成特征图后, 金字塔池化模块对特征图进行不同尺度的池化, 池化后的尺寸为: 1×1、2×2、3×3、6×6. 最后 PSPNet 对池化得到的结果进行上采样, 并与另一分支的特征图结合, 通过卷积层输出分割结果. 该网络应用的金字塔池化模块汇聚了不同区域的上下文信息, 能够更好地理解图像, 提高获取全局信息的能力. PSPNet 在 2016 年 ImageNet 场景解析挑战赛以及 PASCAL VOC 2012 和 Cityscapes 数据集上均表现良好.

2.3 实例分割

随着深度学习技术的快速发展, 也出现了许多基于卷积神经网络的实例分割模型. 本小节介绍实例分割的典型模型.

2.3.1 实例分割的基本流程

实例分割模型一般由三部分组成: 图像输入、实例分割处理、分割结果输出. 图像输入后, 模型一般使用 VGGNet、ResNet 等骨干网络提取图像特征, 然后通过实例分割模型进行处理. 模型中可以先通过目标检测判定目标实例的位置和类别, 然后在所选定区域位置进行分割, 或者先执行语义分割任务, 再区分不同的实例, 最后输出实例分割结果.

2.3.2 实例分割的典型方法

1) DeepMask

DeepMask[7] 网络采用 VGGNet 对输入图像提取特征, 生成分割提议, 提取的特征为两个分支所共享, 第 1 个分支对选中的物体预测一个分割掩码, 第 2 个分支对输入的 Patch 预测一个目标得分. 该网络在 PASCAL VOC 2007 和 MS COCO 数据集上进行了验证, 分割精度良好.

2) Mask R-CNN

Mask R-CNN 由 He 等[39] 提出, 是在 Faster R-CNN[40] 基础上扩展而来的一种全新的实例分割模型. Mask R-CNN 属于两阶段方法, 第 1 阶段使用 RPN (Region proposal network) 来产生 ROI (Region of interest) 候选区域. 第 2 阶段模型对每个 ROI 的类别、边界框偏移和二值化掩码进行预测. 掩码由新增加的第 3 个分支进行预测, 这是 Mask R-CNN 与其他方法的不同点. 此外, Mask R-CNN 提出了 ROIAlign, 在下采样时对像素进行对准, 使得分割的实例位置更加准确.

3) PANet

PANet 是 Liu 等[41] 提出的一种两阶段实例分割模型. 为了缩短信息通路, 该模型利用低层精确的定位信息提升特征金字塔, 创建了自底向上的路径增强. 为了恢复候选区域和所有特征层之间被破坏的信息通路, Liu 等[41] 开发了自适应特征池化, 用来汇聚每个候选区域所有特征层的特征. 此外, 模型用全连接层来增强掩码预测, 由于具有全卷积网络的互补特性, 模型获得了每个候选区域的不同视图. 由于这些改进, PANet 在 MS COCO 2017 实例分割任务上排名第一.

4) Mask SSD

Mask SSD 是 Zhang 等[42] 提出的一种单阶段实例分割模型. Mask R-CNN 需要经过两阶段处理, 先通过 RPN 获得候选区域, 后进行目标检测和实例分割, 因此计算效率比较低. Mask SSD 在单阶段检测器 SSD 的基础上, 增加一个实例分割模块, 包括多个卷积层和一个反卷积层, 预测每个检测对象的前景掩码. 并且网络对特征表示和目标预测进行了优化, 使得 Mask SSD 具有与 Mask R-CNN 相近的检测分割精度, 同时将计算速度提高了约 50%.

2.4 常用数据集

为了促进全景分割的研究进展, 大规模的开源数据集必不可少. 目前, 全景分割常用数据集有 PASCAL VOC、MS COCO、Cityscapes、ADE20k 等. 本小节从图像数、类别数、样本数等方面介绍几种常用数据集.

1) PASCAL VOC 数据集[43]

PASCAL VOC 在 2005 ~ 2012 年每年发布关于图像分类、目标检测、图像分割等任务的子数据集, 并举行世界级的计算机视觉大赛. PASCAL VOC 数据集最初有 4 类, 最后稳定在 21 类, 对于分割任务, 这些类别有汽车、房屋、动物、飞机、自行车、船、公共汽车、小汽车、摩托车、火车等, 测试图像从早期的 1578 幅最后稳定在 11540 幅. PASCAL VOC 数据集包括训练集和测试集, 对于实际比赛有一个独立的测试集. 2012 年以后 PASCAL VOC 大赛停办, 但是数据集开源, 可以下载使用.

2) MS COCO 数据集[44]

MS COCO 是另一个图像识别分割数据集, 它

总共有 91 个物体类别, 32.8 万幅图像, 超过 8 万幅图像用于训练, 4 万多幅图像用于验证, 8 万多幅图像用于测试, 拥有 250 万个标注实例. MS COCO 数据集的每一类物体的图像数量多, 标注精细, 数据场景多样性高, 是目前比较流行的数据集.

3) Cityscapes 数据集[45]

Cityscapes 数据集是一个城市街道场景的数据集, 拥有精细标注的 5 000 幅城市驾驶场景图像, 其中 2 975 幅用于训练, 500 幅用于验证, 1 525 幅用于测试, 还有 20 000 幅粗标注的图像, 一般使用精细标注的那部分数据. 该数据集包含来自 50 个城市街道场景中记录的图像, 是一个流行的街道场景数据集.

4) ADE20K 数据集[46]

ADE20K 是一个新的场景理解数据集, 总共有 2 万多幅图像, 其中训练集有 20 210 幅图像, 验证集有 2 000 幅图像, 测试集有 3 352 幅图像, 以开放字典标签集密集注释. ADE20K 包含 151 个物体类别, 如汽车、天空、街道、窗户、草坪、海面、咖啡桌等, 每幅图像可能包含多个不同类型的物体, 物体尺度变化大, 因此检测难度高.

5) Mapillary Vistas 数据集[47]

Mapillary Vistas 是目前世界上最多样化的全景分割数据集之一, 它包含的图像像素分割精确, 并且实例被特别标注. Mapillary Vistas 包含 28 个 Stuff 类别和 37 个 Things 类别的图像, 共有图像 25 000 幅, 18 000 幅用于训练, 2 000 幅用于验证, 5 000 幅用于测试, 图像分辨率大, 地理位置范围广泛, 十分具有挑战性.

不同于之前的综述文章[2, 48−49], 本节从卷积神经网络开始对全景分割相关基础知识进行了简洁、全面的介绍, 综述了语义分割、实例分割经典模型并进行了讨论, 最后概括了全景分割任务现有数据集, 有助于读者对全景分割相关任务的理解.

3　深度学习在全景分割中的应用进展

自全景分割研究方向提出以来, 深度学习以其强大的性能, 一直是全景分割研究的关键. 目前很多全景分割模型, 如 UPSNet、AUNet、TASCNet, 都是基于卷积神经网络实现的.

根据实例分割分支采用的策略, 全景分割方法可以分为单阶段方法和两阶段方法, 它们各有优劣. 本节从单阶段和两阶段的角度, 分别介绍深度学习在全景分割中的应用进展.

3.1　单阶段方法

模型若没有区域提议步骤, 可以将这类模型称为单阶段模型. 单阶段模型计算速度快, 能够满足智能终端和边缘设备实时响应的需求, 但是精度比两阶段模型稍低, 许多研究人员致力于在满足实时响应的前提下提高模型精度. 本小节对单阶段方法进行综述.

BlitzNet[50] 是最先探索全景分割的方法之一, 它将目标检测和语义分割任务联合起来共同执行, 具有全景分割的雏形, 且没有区域提议这一步骤, 因此将其归为单阶段全景分割方法, 如图 6 所示. BlitzNet 基本流程如下: 首先用 ResNet-50 生成高层特征图; 依照 SSD[51] 目标检测方法, 迭代降低特征图的分辨率以执行边界框的多范围搜索, 防止有用信息丢失; 接下来用反卷积层对特征图进行放大, 以预测精确的分割图; 最后用单个卷积层实现预测. 整个网络以级联方式连接, 采用跳跃连接机制, 将缩小尺度和扩大尺度的特征图结合, 通过权值共享来进行多尺度检测和分割, 减少了计算量, 同时表明目标检测和语义分割在准确性上能够相互促进.

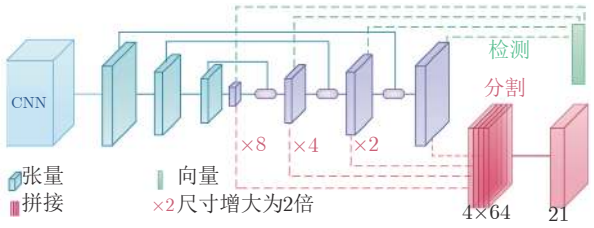


图 6　BlitzNet 网络的基本结构[50]

Fig. 6　The structure of BlitzNet[50]

DeeperLab 是另一种单阶段全景分割模型[52], 如图 7 所示. 语义图和实例图通过单向的全卷积网络产生预测掩码, 然后将预测掩码进行融合, 得到全景分割预测. 模型主要采用了深度可分离卷积, 使用具有两层预测头的共享解码器输出, 扩大内核大小, 同时采用硬数据挖掘的策略, 目的是减小计

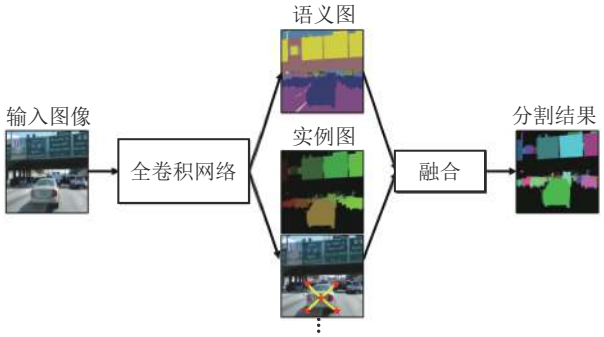


图 7　DeeperLab 全景分割结构[52]

Fig. 7　The structure of DeeperLab panoptic segmentation[52]

算量, 减少高分辨率输入的内存占用, 提升网络运行速度. 在融合策略方面, 对于语义分割预测的 Stuff 类像素分配一个实例标签, 其他像素的实例标签由实例分割决定, 语义标签由语义分割中投票数多的标签决定, 通过改进文献 [1] 的启发式方法, 进一步提升网络性能.

Eppel 等[53] 提出 Generator evaluator-selector 网络进行全景分割. 该网络包含两个独立的部分: 一部分是生成器网络, 主要用于提议与图像中的目标对象和不同区域相对应的各种分段 (Segment); 另一部分是评估器网络, 它选择要合并到分割图中的最佳分段. 网络使用了单指针 (Single pointer) 分割网络[54], 产生仅限于覆盖未分割图像区域的给定 ROI 的分段, 然后使用评估器网络对这些分段排序, 选择得分最高并且互相一致的分段, 减少像素分配冲突的问题. 然后使用精细化网络对选定的分段抛光, 使用分类网络对这些分段进行分类. 最后将处理好的分段填补到分割图中, 形成全景预测. 由于网络对于分段的处理较多, 耗费计算量大, 不易达到实时性要求.

De Geus 等[22] 提出了 FPSNet, 由于无需计算成本极高的实例掩码预测和融合启发算法, 而是通过将全景分割任务转换为自定义的像素级密集分类任务来实现, 极大提高了模型的计算速度, 同时能实现相似或者更好的分割性能. FPSNet 架构与一般全景分割网络相似, 为了减少计算损耗, 取消了实例分割预测以及后处理步骤, 通过引入全景头实现融合目的. 全景头将可以进行密集分割的特征图和表示 Things 类实例的掩码作为输入, 最终实现三个目的: 1) 对 Stuff 类实例进行语义分割; 2) 对于 Things 类实例, 将注意力掩码转换为完整的像素级实例掩码; 3) 在单个特征图中输出 Things 和 Stuff 类别预测, 在其上进行像素级分类. 当有来自普通目标检测器的边界框物体检测以及应用密集图像的单一特征映射时, 边界框产生注意力掩码, 表示图像中 Things 类实例的位置, 并确定实例的输出顺序, 缓解图像中出现的实例遮挡问题.

为了最佳利用物体位置信息, Chen 等[55] 提出了 SpatialFlow 模型, 这是第一个将物体空间信息纳入考虑范围的模型, 该模型在骨干网络上设计了 Things、Stuff、回归和分类子网络, 通过将回归子网络的信息输送给其他三个分支来增加网络对物体位置细节的提取, 提升网络对输入图像的理解能力, 达到空间位置敏感的目的. 这种方法的引入大幅提升了 SpatialFlow 在 MS COCO 数据集的表现, PQ (Panoptic quality) 值达到了 47.3%.

Weber 等[28] 提出了 Single-shot 全景分割模型, 该模型架构与文献 [52] 的架构相似, 不同之处在于 Single-shot 模型的实例中心预测只有一个分支, 输出边界框中心, 通过对像素添加预测偏移和计算 L2 距离来消除遮挡问题, 扩展与另外两个分支的融合. Weber 等[28] 提出了一个新的全景头来输出连贯的全景预测, 基本流程如下: 网络将语义逻辑输出分为 Things 和 Stuff 两类, Stuff 类复制到实例感知逻辑输出中, Things 类以检测分支结果代替. 然后利用检测分支预测中的类别作为索引, 选择 Things 中相应逻辑输出, 并基于边界框进行裁剪, 经滤除堆叠操作后形成新的逻辑输出. 最后, 为了使 argmax操作选择正确的实例, 提出最近中心策略, 利用边界框中心预测消除遮挡, 选择正确实例以形成全景预测. 由于网络融合步骤较复杂, 计算量大, 不易达到实时速度.

为了使模型达到实时速度并且保证分割精度, Hou 等[56] 提出了密集检测的实时全景分割模型, 该模型利用密集检测和一个全局自注意力机制, 确保达到目标. 在骨干网络上增加多个共享全景头部进行密集检测, 全景头部内含语义塔和位置塔, 位置塔输出边界框坐标和中心, 语义塔输出边界框分类信息, 防止图像信息丢失. 然后, 引入新的无参掩码构建方法, 通过 NMS (Non-maximum suppression) 选择具有最高置信分数的查询边界框, 构建全局掩码概率映射, 以密集框和查询框的自注意力近似这个概率映射, 用阈值限制形成实例掩码集, 大幅降低计算复杂度, 同时由于不需对预测进行后处理, 实现了显著的硬件加速.

Cheng 等[57] 在 DeeperLab 的基础上, 提出了一种简单、强大、快速的自底向上的全景分割方法. 如图 8 所示, Panoptic-DeepLab 语义分支采用 Dual-ASPP 结构, 设计与 DeepLab 相同, 实例分支采用 Dual-decoder 结构, 与类别无关, 增加网络的泛化能力. 在此基础上, 添加实例中心回归, 预测实例中心以及从每个像素到其对应中心的偏移量, 提升检测准确度. 然后, 该模型通过将像素分配到其最近的预测中心来实现简单的分组操作, 缓解像素的分配冲突问题.

相对于之前的全景分割模型, Bonde 等[58] 提出了无边界框全景分割模型 BBFNet, 网络架构有较大创新. 如图 9 所示, BBFNet 结构上取消了实例模块, 直接以语义模块提供 Things 和 Stuff 两类特征, 降低了计算成本. 语义模块使用可变形卷积块[59] 的中间特征, 经卷积处理形成语义预测. 然后使用传统的霍夫投票[60]、分水岭[61] 方法处理语义分割预测, 其中分水岭处理模块的特征与输入图像直接得到的特征进行连接, 形成了分水岭预测, 分水岭特

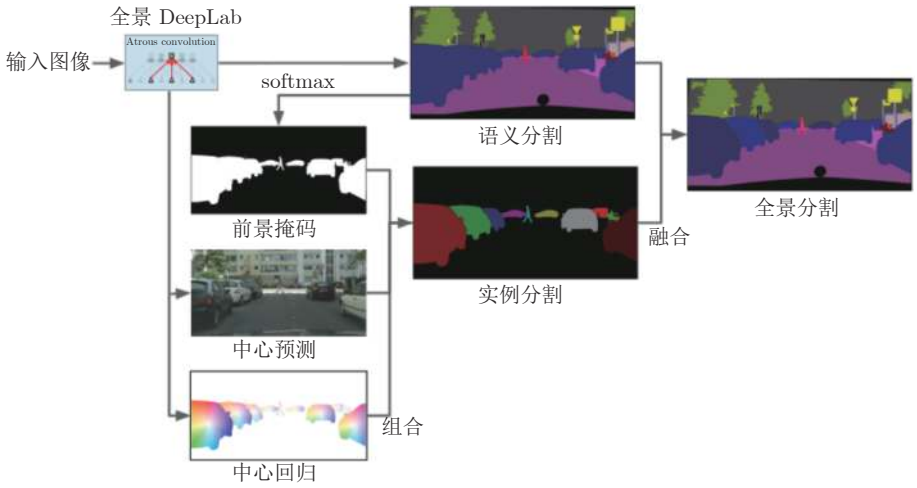


图 8　Panoptic-DeepLab 模型结构[57]

Fig. 8　The structure of Panoptic-DeepLab model[57]

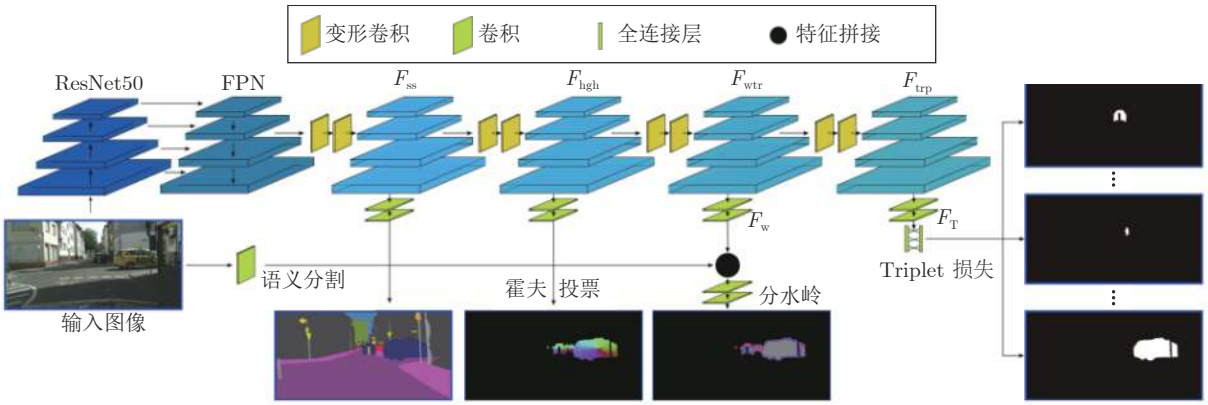


图 9　BBFNet 模型结构[58]

Fig. 9　The structure of BBFNet model[58]

征经变形卷积块处理, 中间特征进入三重损失网络, 三重损失网络能够细化、合并检测到的物体实例, 并检测新的物体, 因此使用三重损失函数可将属于同一实例的像素聚到一起, 将不属于同一实例的像素分开, 缓解像素分配冲突的问题, 使分割结果更准确.

为了解决实例分支耗费计算时间以及独立处理分支导致预测不一致等问题, Chang 等[62] 提出了 EPSNet 模型. 对于整幅图像, EPSNet 采用并行网络产生原型掩码, 并为语义和实例分割预测一组系数, 通过将原型与分支中的预测系数线性组合, 生成实例和语义段. 原型由实例和语义分支共享, 节省了产生大尺寸掩码的时间. 此外, Chang 等[62] 提出了一种新的融合模块, 称为跨层注意力融合模块, 该模块将特征金字塔网络的 F3、F4 和 F5 层作为源特征层, 在其上应用注意力模块以捕获特征图任意两个位置的空间相关性, 提高了共享特征的质量, 并生成预测输送至语义、实例两分支, 进一步增强了网络对低层丰富特征的应用. 在 MS COCO 数据集该网络以更快的推理速度获得了竞争性能.

Wang 等[63] 提出了 PCV (Pixel consensus voting) 模型, 如图 10 所示. PCV 的核心是基于广义霍夫变换的实例分割框架, 它能对包含实例质心的可能区域像素进行离散的概率投票, 投票分支针对每个像素预测该像素是否为实例掩码的一部分, 如果是, 则计算实例掩码质心的相对位置, 降低像素分配冲突. 语义分支采用 FCN 网络. PCV 的一项关键创新是将来自投票分支的预测聚集成投票热图的空洞卷积机制, 有助于实现有效的投票聚合和反投影, 热图局部最大值是检测候选对象, 在每个峰区域, 对查询过滤器进行卷积, 在所有其他峰上反投影支持该特定峰的像素, 以增加对图像细节的提取. 这些像素形成与类别无关的实例分割掩码. 最后, 使用简单的贪婪策略合并实例和语义分割掩码,

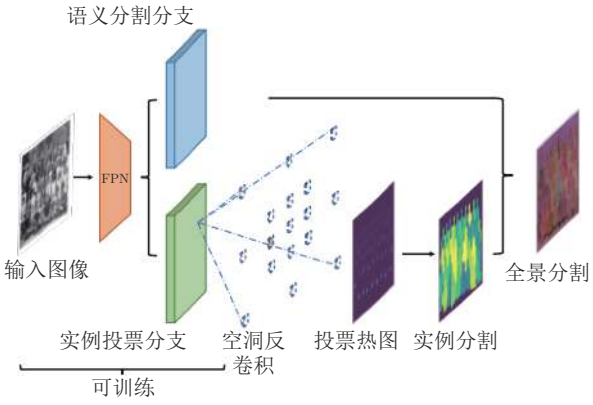


图 10 PCV 模型结构[63]

Fig. 10 The structure of PCV model[63]

产生完整的全景分割输出. 相比一些高度工程化的全景分割模型, 该模型计算简单, 在 MS COCO 数据集获得了良好表现.

Wang 等[64] 在 Panoptic-DeepLab 的基础上提出了应用于全景分割的独立轴对称机制. 如图 11 所示, 其核心是沿高度轴和宽度轴将二维注意力机制分解为两个一维注意力机制, 恢复独立注意力模型中的大感受野, 消除了注意力只在一个局部区域的限制, 降低计算复杂度, 并添加了相对位置编码, 形成位置敏感自注意力. 然后, 在图像的宽度轴上定义一个轴向注意层作为简单的一维位置敏感自注意, 并对高度轴使用类似的定义, 形成轴向注意, 目的是提升注意力在建模依赖于位置的交互时的能力. 网络提取输出步长为 16 的特征图, 采用双重卷积解码器进行全景分割, 采用与 Panoptic-DeepLab 完全相同的预测头部, 产生语义分割和类别无关的实例分割, 保证预测速度达到实时性的要求.

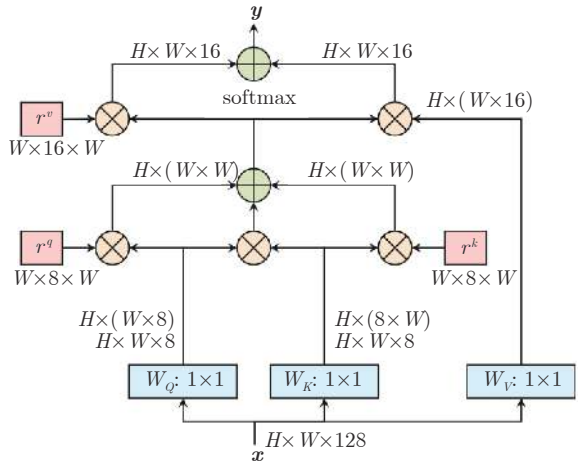


图 11 轴注意力模型结构[64]

Fig. 11 The structure of axial-attention model[64]

最后通过简单的多数表决[52] 合并形成最终的全景分割预测.

表 1 列举了一些典型的单阶段全景分割方法在开源数据集上实验结果的性能比较.

3.2 两阶段方法

两阶段方法有区域提议步骤, 相对于单阶段方法, 分割精度更高, 但是推理速度受到较大影响, 难以达到实时性要求. 两阶段方法由于高精度、高鲁棒性的特点而备受关注. 本小节总结两阶段的全景分割方法.

Li 等[16] 提出了一种弱监督全景分割方法, 结构采用动态实例化网络[65], 该网络包含一个语义子网络和一个实例子网络. 实例子网络将语义子网络和目标检测器的输出作为输入, 在目标检测器提示下, 语义分割输出转变为实例分割输出, 形成最终预测. 语义分割模型使用弱监督方法和图像先验知识来估计真值, 然后采用 GrabCut[66] 和 MCG (Multiscale combinatorial grouping)[67], 从边界框注释中获得前景掩码, 当两者表示一致时, 像素才能分配给由边界框表示的对象类, 解决分配冲突问题. 网络使用图像级注释, 利用训练的多类别分类器根据图像级标签提取弱定位线索, 形成的定位热度图中空间最小的优先级最高, 防止小目标遗漏. 但由于图像中实例数量难以预测, 该网络不适用于具有多个 Stuff 类的图像.

De Geus 等[9] 提出 Single JSIS-Net 网络产生全景预测. 该网络使用共享特征提取器为实例和语义分支提供特征, 语义分支使用金字塔池化模块产生特征映射, 实例分支基于 Mask R-CNN 设计, 产生相应输出, 最终通过启发式融合产生全景预测. 在融合步骤, 对于实例掩码交叠, 本文采用逐实例概率映射消除; 对于 Things 类别实例冲突, 首先将语义预测中的 Things 类移除, 代之以最相似的 Stuff 类, 之后将低于 4096 像素的 Stuff 类移除, 以得分最高的 Stuff 类填充, 解决融合冲突. 由于用最相似类别进行填充, 像素分配存在较大误差. 实验结果表明, 该网络的 PQ 值仅有 27.2%, 性能较差, 不能实现准确分割.

UPSNet[27] 网络架构采用与 JSIS-Net 网络相似的做法, 区别在融合步骤 UPSNet 提出全景分割头, 对两个分支的输出进行融合. 在全景分割头, 语义分割逻辑输出分为 Things 和 Stuff 两类, 通过双线性插值以及边框外语义分割掩码补零, 裁剪形成了新的实例分割掩码, 然后以 Stuff 类掩码和实例掩码之和表示当前掩码, 用 softmax 预测实例类别. 网络中另外设置了一个未知类别预测通道, 以防止

表 1　现有单阶段方法性能比较

Table 1　Performance comparison of existing single-stage methods

模型	数据集	PQ	mIoU	AP	mAP	Inference time
BlitzNet[50]	PASCAL VOC	—	—	—	83.8	24 帧/s
DeeperLab[52]	Mapillary Vistas validation set	31.95	—	—	—	—
Generator evaluator-selector net[53]	MS COCO	33.7	—	—	—	—
FPSNet[22]	CityScapes validation set	55.1	—	—	—	114 ms
SpatialFlow[55]	MS COCO 2017 test-dev split	47.3	—	—	36.7	—
Single-shot panoptic segmentation[28]	MS COCO val2017	32.4	27.9	33.1	—	21.8 帧/s
Panoptic-DeepLab[57]	MS COCO val set	39.7	—	—	—	132 ms
Real-time panoptic segmentation from dense detections[56]	MS COCO val set	37.1	—	—	—	63 ms
BBFNet[58]	MS COCO-2017 dataset	37.1	—	—	—	—
Axial-DeepLab[64]	Cityscapes test set	62.8	79.9	34.0	—	—
EPSNet[62]	MS COCO val set	38.6	—	—	—	53 ms
PCV[63]	Cityscapes val set	54.2	74.1	—	—	182.8 ms

出现类别误判的情况, 影响最终的分割质量, 最终将语义分割逻辑输出预测、当前实例掩码预测以及未知类别预测按序填入相应张量形成全景预测. 但是, 由于深层特征分辨率较低, 并且缺乏对浅层信息的有效利用, UPSNet 对图像中小目标分割指标不高.

为了解决语义和实例分支分割结果不一致的问题, Li 等[10] 提出了 TASCNet, 如图 12 所示. TASCNet 是一个端到端的全景分割模型, 基本架构与 UPSNet 类似, 不同之处是 TASCNet 两分支中间加入了 TASC (Things and stuff consistency) 模块, 该模块在训练过程中保持两个子任务 Things 类别实例输出分布对齐, 促使语义和实例分割头部的输出之间实现分离最小化, 减少有用信息的丢失, 同时实现掩码引导的融合, 降低全景预测前景目标像素对不齐导致的小目标分割指标低的问题. 在 MS COCO 数据集上, TASCNet 获得了良好的分割效果.

Li 等[11] 根据前景和背景之间的关系提出统一的 AUNet. 该网络包含前景和背景分支, 背景分支中加入了两个注意源 (RPN 和前景分段掩码), 分别提供对象级别和像素级别的注意, 对应的注意力模块分别为 PAM (Proposal attention module) 和 MAM (Mask attention module). 前景分支借助 RPN 输出的 ROI 产生前景掩码边界框和类别标签, 背景分支通过 PAM 将 RPN 中的信息提示与共享的特征结合, 产生的掩码作用于背景分支, 使得分割任务更多集中于局部物体, 分割更准确. MAM 将前景 Things 和背景 Stuff 之间的边界用前景分割掩码细化, 在注意力操作中集中更多注意力于

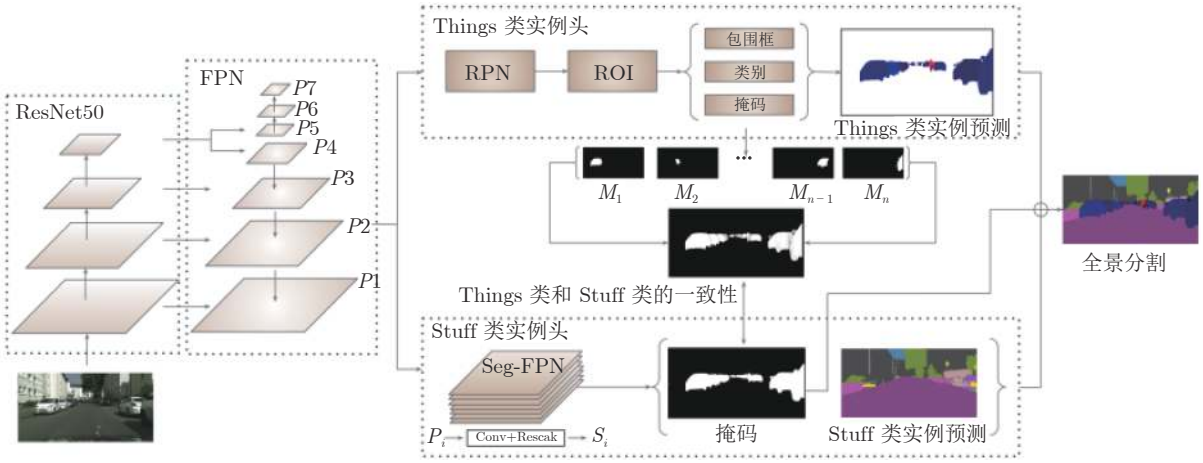


图 12　TASCNet 模型结构[10]

Fig. 12　The structure of TASCNet model[10]

Stuff 区域, 以获得更好的背景分割预测. 与其他将重点放在前景的方法相比, 该方法能够更好地处理背景内容, 在分割指标 PQ 和 AP (Average precision) 值上有了较大提升.

Kirillov 等[26] 在 Mask R-CNN 的基础上提出了一种全景金字塔网络, 该网络将特征金字塔网络 (Feature pyramid network, FPN) 作为全景分割的特征来源, 将语义分割和实例分割任务统一到单个网络中. FPN 能够产生分辨率高、表达能力强的多尺度特征, 将其引入全景分割解决了其他骨干网络提供特征少、分割不准确的问题. 在 MS COCO 数据集上, 全景金字塔网络取得了良好的精度, 成为全景分割任务的一个基准方法.

De Geus 等[12] 提出 Single network 来解决全景分割任务, 整个网络由骨干网络、语义分割分支和实例分割分支组成. 为了提升网络性能, 将语义分支与实例分支连接, 通过对 RPN 产生的区域提议增加边界框, 扩大实例分支产生的边界框, 将语义分支产生的 Things 类预测输入实例分支, 减少有用信息丢失, 实现对显式信息的利用. 此外, 语义分支 softmax 层前产生的预测与骨干网络提取的特征归一化后级联, 输入实例分支, 应用卷积处理后输入 RPN 网络, 实现对隐式信息的利用. 在融合策略方面, 优先考虑实例分类得分较高的类别填充争议部分, 解决交叠问题. 与 TASCNet相比, 该网络能更快地处理输入图像, 但是由于深度特征丢失了物体的位置细节信息, 造成定位性能差, PQ 指标性能较低.

为了解决启发式算法不能很好地融合实例的问题, Liu 等[13] 提出一种以端到端遮挡感知方式解决全景分割任务的模型 OANet, 其架构类似于 UPSNet 模型. 在融合策略上, 提出空间排序模块对两分支预测进行融合, 流程如下: 将实例分支预测映射至输入图像大小, 设置初始张量为 0, 映射值为 1, 然后应用扩张的大核卷积获得排序分数映射, 最后利用像素级交叉熵优化获得排序分数映射. 优化后的排序分数映射经过计算, 得到每个实例的分数, 进而确定实例先后次序进行融合. 相比之前的模型, 该模型融合策略更加可靠, 模型解决实例交叠的能力得到提升, 在 MS COCO 数据集上获得了良好的分割效果.

Lazarow 等[68] 为了解决具有检测置信度的实例与自然遮挡关系不相关的问题, 提出用于全景分割的学习实例遮挡网络, 即在分支部分模拟两个实例掩码如何以二进制关系交叠, 进而解决遮挡问题. 网络架构如图 13 所示, 该模型在 Things 分支掩码头部添加了遮挡头, 它根据两个掩码交叠部分与未交叠部分的比率和确定的阈值大小之比, 来决定是否消除遮挡, 若比率大于阈值, 则根据二进制遮挡关系决定交叠部分像素属于哪个实例, 这样融合过程就绕过了对置信度的依赖, 从而可以查询两个具有明显交集的实例中哪个实例获得的比率更高, 位于另一个实例的上方或下方, 以便融合过程与实例的初始顺序无关. 在 MS COCO 数据集上, 该方法的 PQ 值较其他方法有了显著提高, 但是不能在端到端训练中有效消除实例间遮挡.

类似于 OCFusion, SOGNet[29] 也致力于实例交叠问题的研究, 将交叠问题转化为场景交叠图, 提升模型分割性能, 其结构如图 14 所示. 受场景图解析任务中关系分类的启发, SOGNet 将交叠问题

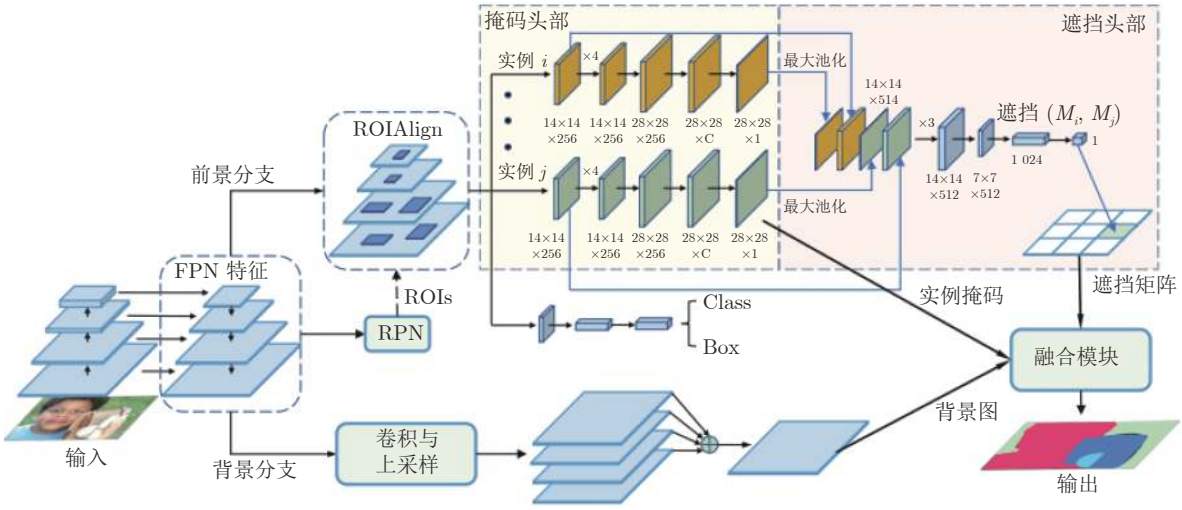


图 13 学习实例遮挡网络结构[68]

Fig. 13 The structure of learning instance occlusion for panoptic segmentation[68]

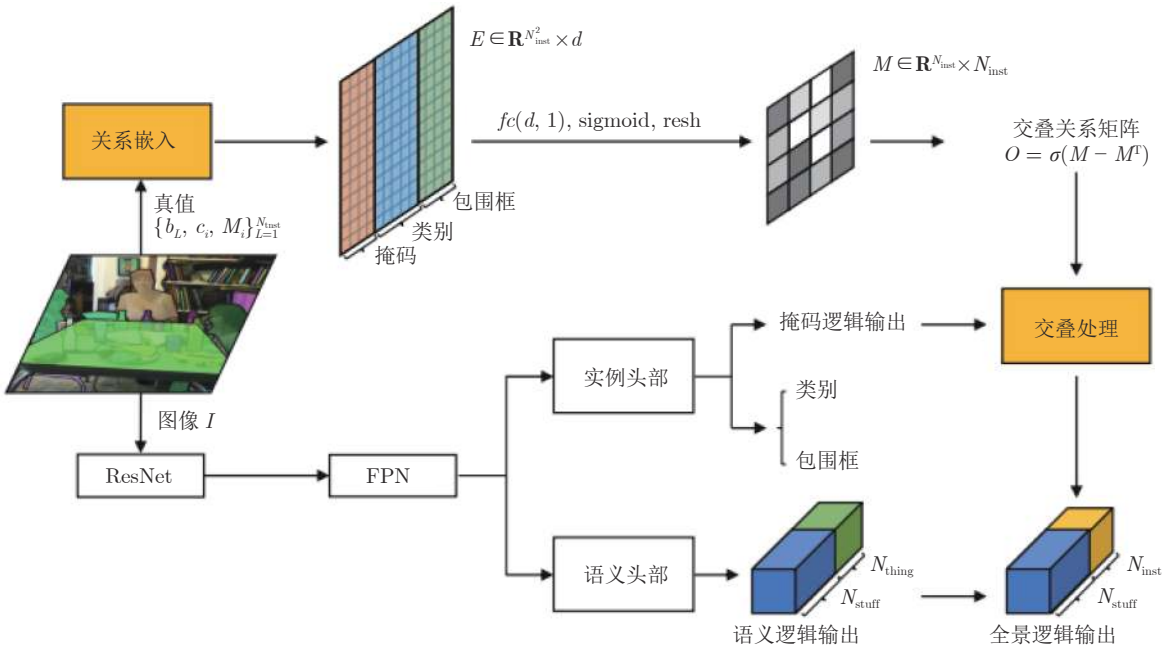


图 14　SOGNet 模型结构[29]

Fig. 14　The structure of SOGNet model[29]

表述为带有方向边的简化场景图, 其中实例与实例之间有三种关系类型: 无交叠、作为主体覆盖、作为对象覆盖. 作者利用对象的类别、几何形状和外观信息对场景交叠图执行边缘特征嵌入, 并提出交叠解决模块, 该模块以遮挡关系矩阵显式编码实例间的遮挡关系, 以可区分的方式解决任何一对实例之间的遮挡区域, 去除遮挡区域后的掩码逻辑输出随后用于全景头每个具有全景注释的实例, 以便进行分类, 形成全景预测. 该策略在解决目标重叠时像素分配冲突的问题上比以往的方法更好, 在 MS COCO 数据集上获得了较好的分割结果, 对于小物体分割也有更好的表现.

PanDA[69] 从数据集方面做出改进, 以提升全景分割模型的性能. PanDA 是一种用于全景分割数据集的像素空间数据扩充方法, 它首先将真值分为前景段和背景段, 其中包含未标记和背景类别实例的背景段以噪声图案填充, 生成和原图大小相同的新图, 然后将前景段覆盖于新的背景段上. 对于前景段, 使用 dropout、resize 等操作控制每个实例的不同方面, 对实例创建新的上下文信息, 同时根据片段面积大小进行排序, 解决目标遮挡问题. PanDA 可以利用原始数据集中图像合成新的数据集, 计算量小且不需训练, 在一流模型各种指标上获得比原始版本高 1.2% ~ 6.9% 的相对增益.

Jayasumana 等[17] 提出 Bipartite CRF (BCRF) 模型, 其架构如图 15 所示. 本文首先将 CRF 引入全景分割, 提出带二分随机变量的 CRF 公式, 捕获语义和实例标签之间的相互作用. 对于输入的语义和实例分支预测, BCRF 模块将语义分支预测变为与初始分类器一致, 同时提高语义标注的平滑性; 对于实例分支, 同样鼓励全景分割与实例分割概率一致. 此外, 通过惩罚将不同的实例标签分配给相似的像素, 让实例标签在整个图像上更加一致, 缓解像素分配冲突问题. 为了让模型性能更强, Jayasumana 等[17] 还引入了 CRF 中的交叉势能, 即任意像素的语义标签必须与该像素的实例标签兼容. 若实例分类的初始结果与语义分类的结果不一致, 则其中一个应根据 CRF 中其他项的相互作用进行纠正. 通过这些设置, BCRF 模块最终得到了较好的全景预测, 在 PASCAL VOC 和 MS COCO 数据集上取得了良好的表现.

Li 等[70] 提出了全景分割的统一训练和推理网络, 缩小了全景分割的训练和推理流程之间的差距. 该网络包含骨干网络、语义分割子模块、目标检测子模块和全景子模块四部分, 整个流程的核心是全景分割子模块动势头 (Dynamic potential head) 和密集实例亲和头 (Dense instance affinity head), 动势头是表示通道动态数量的全景实例的无参数步骤模块, 对分割和定位线索进行集合. 密集实例亲和头是参数化、高效且数据驱动的模块, 该模块预测并利用像素属于 Things 和 Stuff 的可能性, 清理和纠正粗略的特征线索, 以准确地描绘全景边界. 此外, 为了直接优化用于全景分割预测的全局目标, 本文提出了全景匹配损失函数, 该函数以及全景头

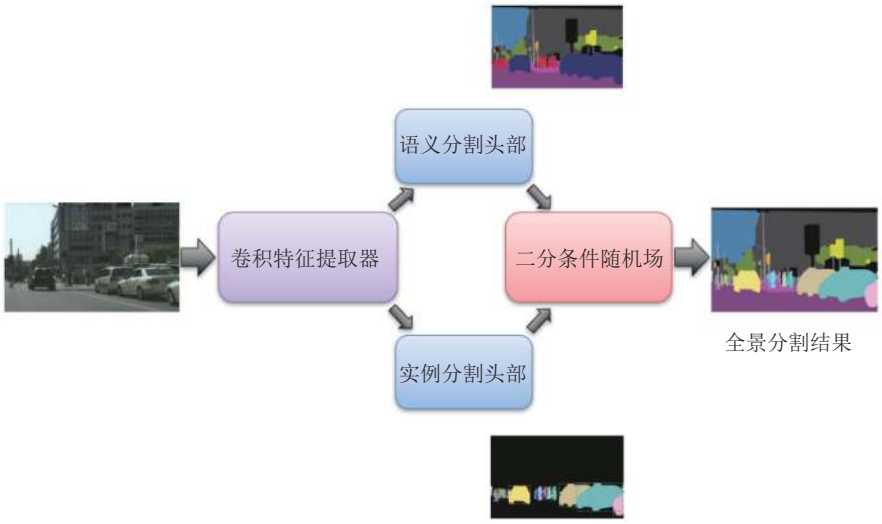


图 15 BCRF 网络结构[17]

Fig. 15 The structure of BCRF network[17]

的可区分性, 使网络能够以端到端的方式学习. 根据预测的亲和力强度连接和聚集整个图像上的全景图, 提出的参数化全景头能够修复来自前一阶段的不精确预测, 提升模型的分割性能, 且能够灵活工作, 在 MS COCO 数据集上, 该网络分割指标 PQ 值达到了 43.4%.

Behley 等[71] 提出一种基于 LiDAR (Light detection and ranging) 点云的全景分割基准, 将研究从二维空间扩展到三维空间, 为全景分割研究提供了新的方向. 该基准依托 KITTI数据集[72] 进行实验, KITTI 是自动驾驶场景的计算机视觉算法评测数据集, 常用于评测立体图像、光流、3D 目标检测等. 其基本流程如下: 首先, 对 KITTI数据集进行标注, 采用半自动化过程, 使用不同策略生成时间上一致的实例标注. 对于静态物体, 通过使用 SLAM 系统执行姿态校正, 然后通过考虑段的位置来简单地执行数据关联; 对于动态物体, 同时考虑物体和传感器的运动. 在网络设置方面, 语义分支采用了 KPConv[73] 和 RangeNet++[74] 网络, 目标检测器采用 Point-Pillars[75], 训练时分别用 KPConv 和 Point-Pillars、RangeNet++和 Point-Pillars 两种组合, 其中 Point-Pillars 使用定向边界框消除遮挡, 然后组合语义分割的预测, 并将每个边界框的实例 ID 分配给其中的每个像素点, 生成全景预测. 在 KITTI 数据集上, 第 1 种组合的分割性能明显超过第 2 种组合, 但由于深层特征图的分辨率较低, 导致两组网络分割指标较低.

EfficientPS 是 2020 年提出的全景分割方法[76], 其架构如图 16 所示. 该方法包括新的共享骨干网络、移动反向瓶颈单元、双向 FPN, 以及实例和语义分割头. 不同于以前直接利用 FPN 产生特征的方法, 双向 FPN 在原 FPN 的基础上添加了一个反向分支, 能够高效地利用多尺度特征. 另外, 还提出了一种具有可分离卷积的语义头, 在上下文特征和精细特征关联融合之前独立地聚合特征, 使得语义更精细, 在目标边界产生更好的细化边界. 对于实例头, 基于 Mask R-CNN 并通过可分离卷积和 iABN (inplace activated batch normalization) 同步层进行扩展. 为了彻底利用两个头部的逻辑输出, 还提出了一种无参数全景融合模块, 基于每个像素的头部预测适应性, 有选择地衰减或放大融合的逻辑输出分数来自适应融合实例. 在 Cityscapes 和 Mapillary Vistas 数据集上, 该网络达到当时最好的分割指标.

为了更好地利用语义分支和实例分支之间的互补性, Chen 等[77] 提出了 BANet, 该网络包括骨干网络、S2I (Semantic-to-instance) 模块、I2S (Instance-to-semantic) 模块和遮挡处理模块, 如图 17 所示. 当语义分割头通过 FPN 产生语义特征, RPN 网络产生实例候选后, S2I 模块利用 RoIAlign 裁剪的实例语义特征和选定的 FPN 特征, 其目的是用实例特征代替粗粒度的置信度分数映射来提高语义特征的区分能力, 然后聚合所裁剪的特征提供上下文信息, 提升实例分割性能. 为了使 I2S 模块有效处理实例头产生的特征, 提出了 RoIInlay 运算符, 该运算符能够恢复裁剪后实例特征的结构, 以便将其与语义特征聚合以进行语义分割, I2S 模块接收来自语义部分和实例头的特征, 在 SIM (Structure injection module) 和 OCM (Object context module) 中聚合处理, 以实例头特征提升语义分割预测效果.

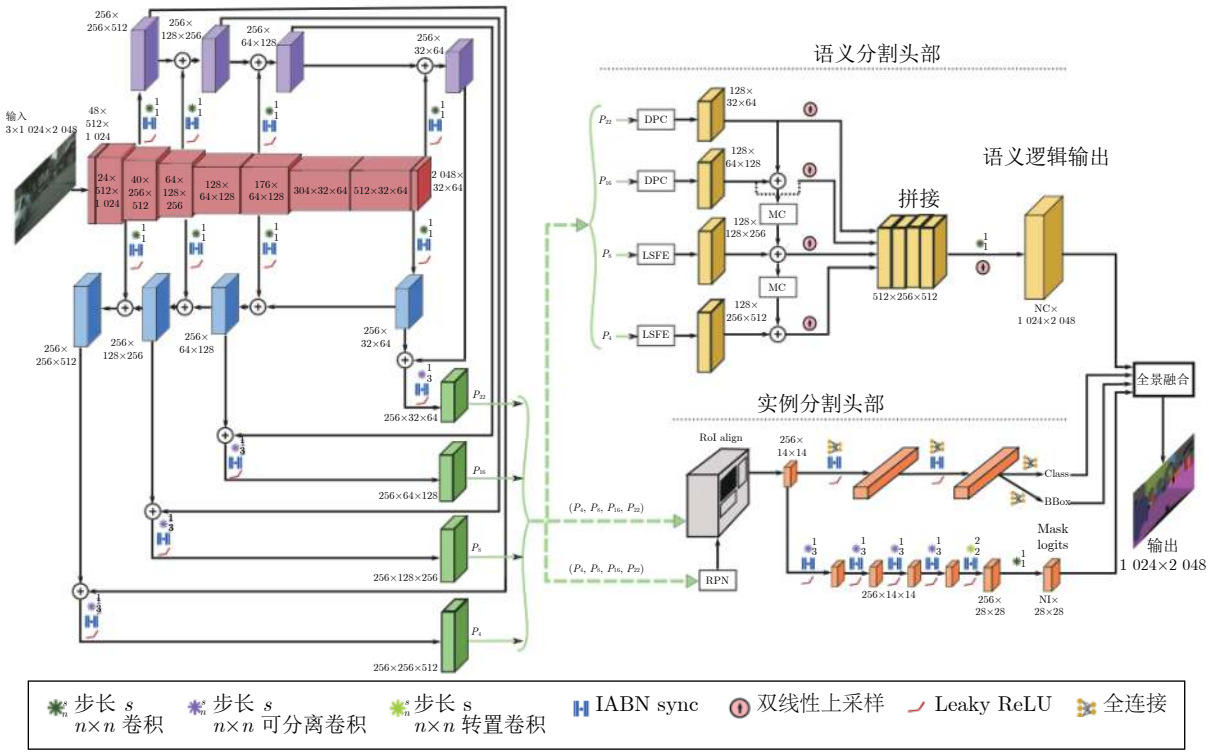


图 16　EfficientPS 模型结构[76]

Fig. 16　The structure of EfficientPS model[76]

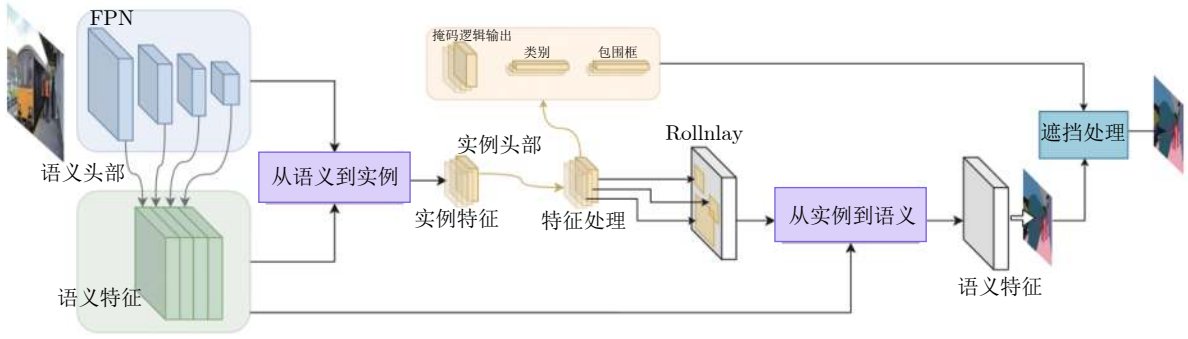


图 17　BANet 模型结构[77]

Fig. 17　The structure of BANet model[77]

遮挡处理模块以实例外观为参考, 通过计算像素与实例的相似度, 将像素分配给最相似的实例来解决遮挡问题. 该网络充分利用语义、实例分支的互补性, 获得良好的分割效果, 但是实例特征在传输过程中被 I2S 模块重复处理, 极大影响了网络计算效率.

全景跟踪是 Hurtado 等[78] 提出的基于全景分割的新研究方向, 用于帮助智能机器人提高对动态场景的理解能力. Hurtado 等提出了 Panoptic-TrackNet, 模型结构如图 18 所示, 在 Efficient-PS 的基础上增加了实例跟踪头和 MOPT(Multi-object panoptic tracking) 融合模块以解决跟踪任务. 实例跟踪头能够跨越连续帧跟踪对象实例, 根据任意帧预测的嵌入向量与先前对象实例的嵌入向量之间的欧氏距离来计算关联相似度, 然后采用 Hungarian 算法将多个实例关联起来, 同时在特定时间窗关联与其最接近的 IDs, 当分类置信分数高的实例与旧的 IDs 不关联时创建新的 IDs 与其关联. MOPT 融合模块跟踪 IDs 与相应的段匹配, 根据设定阈值选择置信分数高的段, 完成排序及恢复分辨率; 通过保留较高分数的段, 缓解遮挡问题; 将上一步产生的掩码与筛选得到的语义分支掩码融合, 注入预先准备的空 canvas 中, 形成全景跟踪输出.

表 2 列举了一些典型的两阶段全景分割方法在开源数据集上实验结果的性能比较.

4 面临问题与解决思路

4.1 全景分割面临的问题

目前全景分割任务的模型设计方法全部基于深度学习技术设计, 由于深度学习具有诸多优点, 未来若干年基于深度学习的全景分割研究仍将是主流. 虽然深度学习在全景分割领域获得了成功, 但是由于以下一些因素的影响, 全景分割的发展仍然十分受限.

1) 深度学习理论还不完善[79]. 深度学习技术学习能力强, 迁移性好, 因此得到了广泛关注与应用. 但是, 很多工作的创新点都是依靠以往的经验和直觉提出, 缺乏严格的理论指导. 为了大幅提升全景分割模型的性能, 使其能够应用于自动驾驶、机器人导航等领域, 需要从模型可解释性、泛化能力、数

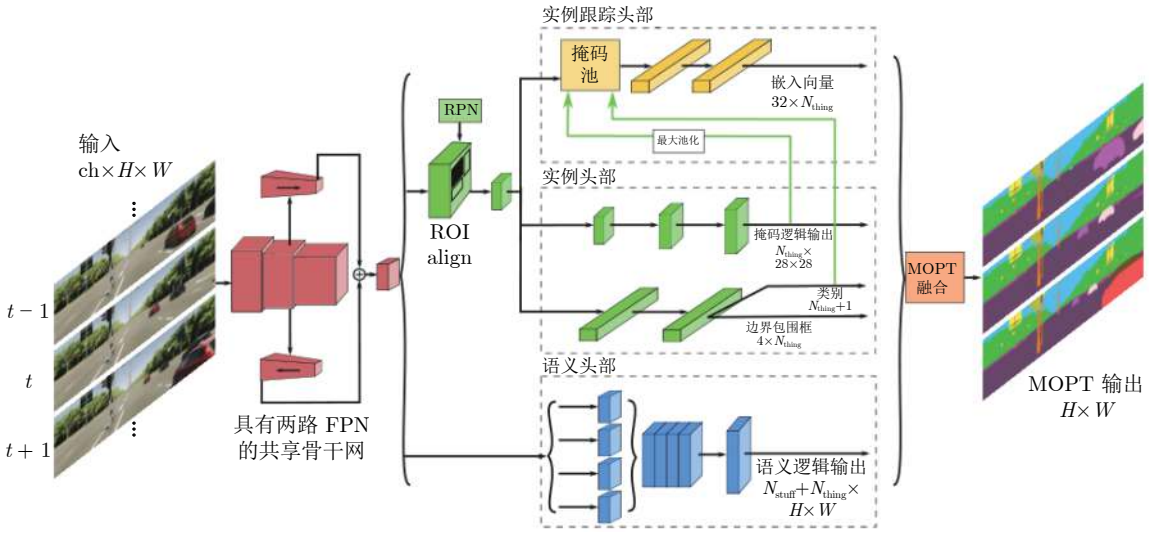


图 18 PanopticTrackNet 模型结构[78]

Fig. 18 The structure of PanopticTrackNet model[78]

表 2 现有两阶段方法性能对比

Table 2 Performance comparison of existing two-stage methods

模型	数据集	PQ	mIoU	AP	mAP	Inference time (ms)
Weakly- and semi-supervised panoptic segmentation[16]	VOC 2012 validation set	63.1	—	59.5	—	—
JSIS-Net[9]	MS COCO test-dev	27.2	—	—	—	—
TASCNet[10]	Cityscapes	60.4	78.7	39.09	—	—
AUNet[11]	Cityscapes val set	59.0	75.6	34.4	—	—
Panoptic feature pyramid networks[26]	MS COCO test-dev	40.9	—	—	—	—
UPSNet[27]	MS COCO	42.5	54.3	34.3	—	17
Single network panoptic segmentation for street scene understanding[12]	Mapillary Vistas	23.9	—	—	—	484
OANet[13]	MS COCO 2018 panoptic segmentation challenge test-dev	41.3	—	—	—	—
OCFusion[68]	MS COCO test-dev dataset	46.7	—	—	—	—
SOGNet[29]	MS COCO	43.7	—	—	—	—
PanDA[69]	MS COCO subsets	37.4	45.9	28.0	—	—
BCRF[17]	Pascal VOC dataset	71.76	—	—	—	—
Unifying training and inference for panoptic segmentation[70]	MS COCO test-dev set	47.2	—	—	—	—
BANet[77]	MS COCO val set	41.1	—	—	—	—
EfficientPS[76]	Cityscapes validation set	63.6	79.3	37.4	—	166

据和知识联合驱动等方面进一步完善深度学习基础理论, 为改进模型结构、提升模型性能提供科学的理论指导.

2) 面向全景分割的数据集规模较小, 训练样本数量较少. 目前应用于全景分割任务的数据集仅有PASCAL VOC、MS COCO、Cityscapes 等少数几种, 并且这些数据集只提供了少数标注类型, 如Cityscapes 总共有 34 种类型, PASCAL VOC 仅有 20 种. 数据集及标注类型的缺乏, 导致训练出的模型可靠性一般, 泛化能力差. 目前许多全景分割模型都基于 MS COCO 和 Cityscapes 数据集进行预训练, 根据具体任务再进行微调来满足要求. 如果有大规模数据集供模型训练使用, 全景分割模型的性能指标会得到进一步提升.

3) 图像中目标重叠时, 像素分配冲突问题很难有效解决. 由于全景分割任务要求图像中所有实例分割结果不能有重叠现象出现, 所以在子任务融合模块如何有效处理实例分割分支产生的重叠是首先要解决的问题, 即对于一个处于交叠部分的 Things 类像素, 会有来自实例分支的多个实例 ID标签, 这时该像素如何分配至正确的实例是一个关键问题. 目前提出的 UPSNet、JSISNet 等模型大多采用启发式方法进行融合, 一定程度上能够缓解这个问题, 但是无法有效解决, 因此影响全景分割结果的性能.

4) 全景分割模型的设计方法效率不高. 目前无论是单阶段方法还是两阶段方法, 几乎都遵循骨干网络提取特征, 然后通过子任务处理和融合来产生全景分割预测的范式. 虽然对其中的某些环节进行改进能够提升全景分割性能, 但是模型设计方法比较单一, 限制了进一步创新的思路, 无法有效提升分割性能.

4.2 解决思路

对于上述局限性, 有一些初步的解决思路. 一种思路是改进完善数据集. Panagiotis 等[80] 在 Cityscapes 和 PASCAL VOC 数据集的基础上, 提出了 Cityscapes-Panoptic-Parts 和 PASCAL-Panoptic-Parts 两个数据集, 采用人工或者人机协作的方式对图像创建语义、实例、部分三种级别的标注, 兼容全景分割, 标注丰富详细, 扩展了全景分割数据集的数量和规模. Liu 等[69] 提出像素空间数据扩充方法, 可以在原数据集的基础上, 将图像分成前景和背景, 通过对前景和背景进行处理, 排序合成新的全景数据集, 能够在一定程度上提升分割性能.

另一种思路是将传统算法创新应用于全景分割任务中, 如目前部分工作将霍夫算法[60] 应用于全景分割. 文献 [58] 提出的模型中, 当预测输出语义分割模块后, 借助语义分割包含的 Stuff 和 Things 类预测, 该模型利用霍夫投票、分水岭等传统图像分割算法分割融合产生全景预测. 在文献 [63] 中, 提出的 PCV 模型核心是基于广义霍夫变换的实例分割框架, 该框架通过投票过滤器确定像素的归属, 然后将产生的实例分割掩码和语义分割掩码通过贪心算法合并, 产生全景分割预测.

第三种思路是从弱监督学习的角度出发, 解决像素分配问题. 文献 [16] 以弱监督方式训练全景分割网络, 对于图像中所有像素, 在没有可靠标签的情况下, 利用弱监督与图像先验知识使图像中像素的一个子集逼近真值, 即对确定的像素分配标签, 不确定的标记为 “忽略”区域, 以此逼近. 然后使用子集中像素的估计标签训练网络. 该方法解决了图像中目标像素的分配冲突问题, 在输出的全景预测中以空白方式显示交叠区域, 为解决像素分配冲突提供了一个方向.

上述方法在一定程度上使全景分割性能得到提升, 但是在推理速度和分割精度上的表现还难以令人满意. 文献 [57] 提出的 Panoptic-DeepLab 模型虽然显著提升了推理速度, 但是分割精度不尽人意. 因此如何兼顾速度与精度, 提出有效的全景分割模型还有待研究. 此外, 深度学习理论还应继续完善, 如何改进模型结构, 创建新的优化方法以减少计算复杂度, 得到精度和效率并重的模型, 这些还需要深入研究.

5 结束语

全景分割任务在计算机视觉领域具有重要的研究意义和应用价值, 其研究进展可以直接推动自动驾驶、机器人等领域的发展. 深度学习作为目前的主流技术, 在全景分割任务中得到广泛应用. 本文综述了深度学习在全景分割中的研究进展, 介绍了全景分割数据集和相关背景知识, 重点介绍了基于深度学习技术的全景分割模型, 总结了深度学习在全景分割任务中的最新进展, 分析了全景分割现有方法存在的问题, 并提出了一些解决思路.

在今后的工作中, 首先需要从深度学习理论和方法入手, 进一步完善深度学习理论, 提升全景分割的性能指标. 此外, 在扩展数据集、结合传统算法与深度学习算法方面, 也应该进行重点研究. 这些工作有助于使得全景分割技术研究和应用更加成熟.

References

1 Kirillov A, He K M, Girshick R, Rother C, Dollar P. Panoptic segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 9396−9405

2 Hu Tao, Li Wei-Hua, Qin Xian-Xiang. A review on image semantic segmentation. *Measurement and Control Technology*, 2019, **38**(7): 8−12
(胡涛, 李卫华, 秦先祥. 图像语义分割方法综述. 测控技术, 2019, **38**(7): 8−12)

3 Yang Y, Hallman S, Ramanan D, Fowlkes C C. Layered object models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(9): 1731−1743

4 Ladicky L, Russell C, Kohli P, Torr P H S. Associative hierarchical CRFs for object class image segmentation. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 739−746

5 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 640−651

6 Hariharan B, Arbelaez P, Girshick R, Malik J. Simultaneous detection and segmentation. In: European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 297−312

7 Pinheiro P O, Collobert R, Dollar P. Learning to segment object candidates. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2015. 1990−1998

8 Zagoruyko S, Lerer A, Lin T Y, Pinheiro P O, Gross S, Chintala S, et al. A MultiPath network for object detection. [Online], available: https://arxiv.org/abs/1604.02135, April 4, 2016

9 De Geus D, Meletis P, Dubbelman G. Panoptic segmentation with a joint semantic and instance segmentation network. [Online], available: https://arxiv.org/abs/1809.02110, September 6, 2018

10 Li J, Raventos A, Bhargava A, Tagawa T, Gaidon A. Learning to fuse things and stuff. [Online], available: https://arxiv.org/abs/1812.01192, December 4, 2018

11 Li Y W, Chen X Z, Zhu Z, Xie L X, Huang G, Du D L, et al. Attention-guided unified network for panoptic segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 7019−7028

12 De Geus D, Meletis P, Dubbelman G. Single network panoptic segmentation for street scene understanding. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV). Paris, France: IEEE, 2019. 709−715

13 Liu H Y, Peng C, Yu C Q, Wang J B, Liu X, Yu G, et al. An end-to-end network for panoptic segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 6165−6174

14 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. [Online], available: https://arxiv.org/abs/1409.1556, September 4, 2014

15 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 770−778

16 Li Q Z, Arnab A, Torr P H S. Weakly- and semi-supervised panoptic segmentation. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 106−124

17 Jayasumana S, Ranasinghe K, Jayawardhana M, Liyanaarachchi S, Ranasinghe H. Bipartite conditional random fields for panoptic segmentation. [Online], available: https://arxiv.org/abs/1912.05307, December 11, 2019

18 Huang G, Liu Z, Van Der Maaten L, Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 2261−2269

19 Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. [Online], available: https://arxiv.org/abs/1704.04861, April 17, 2017

20 Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 4510−4520

21 Howard A, Sandler M, Chen B, Wang W J, Chen L C, Tang M X, et al. Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 1314−1324

22 De Geus D, Meletis P, Dubbelman G. Fast panoptic segmentation network. *IEEE Robotics and Automation Letters*, 2020, **5**(2): 1742−1749

23 Xiao Xiao-Wei, Xiao Di, Lin Jin-Guo, Xiao Yu-Feng. Overview on multi-objective optimization problem research. *Application Research of Computers*, 2011, **28**(3): 805−808
(肖晓伟, 肖迪, 林锦国, 肖玉峰. 多目标优化问题的研究概述. 计算机应用研究, 2011, **28**(3): 805−808)

24 Ge Ji-Ke, Qiu Yu-Hui, Wu Chun-Ming, Pu Guo-Lin. Summary of genetic algorithms research. *Application Research of Computers*, 2008, **25**(10): 2911−2916
(葛继科, 邱玉辉, 吴春明, 蒲国林. 遗传算法研究综述. 计算机应用研究, 2008, **25**(10): 2911−2916)

25 Wu Ying, Chen Ding-Fang, Tang Xiao-Bing, Zhu Shi-Jian, Huang Ying-Yun, Li Qing. Summarizing of neural network. *Science & Technology Progress and Policy*, 2002, **19**(6): 133−134
(巫影, 陈定方, 唐小兵, 朱石坚, 黄映云, 李庆. 神经网络综述. 科技进步与对策, 2002, **19**(6): 133−134)

26 Kirillov A, Girshick R, He K M, Dollar P. Panoptic feature pyramid networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 6392−6401

27 Xiong Y W, Liao R J, Zhao H S, Hu R, Bai M, Yumer E, et al. UPSNet: A unified panoptic segmentation network. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision andPattern Recognition. Long Beach, CA, USA: IEEE, 2019. 8810−8818

28 Weber M, Luiten J, Leibe B. Single-shot panoptic segmentation. [Online], available: https://arxiv.org/abs/1911.00764, November 2, 2019

29 Yang Y B, Li H Y, Li X, Zhao Q J, Wu J L, Lin Z C. SOGNet: Scene overlap graph network for panoptic segmentation, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(7): 12637−12644

30 Fukushima K. Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, **36**(4): 193−202

31 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

32 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, **60**(6): 84−90

33 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 1−9

34 Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. [Online], available: https://arxiv.org/abs/1412.7062, December 22, 2014

35 Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(4): 834−848

36 Chen L C, Hermans A, Papandreou G, Schroff F, Wang P, Adam H. MaskLab: Instance segmentation by refining object detection with semantic and direction features. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 4013−4022

37 Ronneberger O, Fischer P, Brox T, et al. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer, 2015. 234−241

38 Zhao H S, Shi J P, Qi X J, Wang X G, Jia J Y. Pyramid scene parsing network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 6230−6239

39 He K M, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2980−2988

40 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **39**(6): 1137−1149

41 Liu S, Qi L, Qin H F, Shi J P, Jia J Y. Path aggregation network for instance segmentation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 8759−8768

42 Zhang H, Tian Y L, Wang K F, Zhang W S, Wang F Y. Mask SSD: An effective single-stage approach to object instance segmentation. *IEEE Transactions on Image Processing*, 2019, **29**: 2078−2093

43 Everingham M, Eslami S M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 2015, **111**(1): 98−136

44 Lin T T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 740−755

45 Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 3213−3223

46 Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20K dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 5122−5130

47 Neuhold G, Ollmann T, Bulo S R, Kontschieder P. The Mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5000−5009

48 Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. [Online], available: https://arxiv.org/abs/1704.06857, April 22, 2017

49 Lateef F, Ruichek Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 2019, **338**: 321−348

50 Dvornik N, Shmelkov K, Mairal J, Schmid C. BlitzNet: A real-time deep network for scene understanding. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 4174−4182

51 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 21−37

52 Yang T J, Collins M D, Zhu Y K, Hwang J J, Liu T, Zhang X, et al. DeeperLab: Single-shot image parser. [Online], available: https://arxiv.org/abs/1902.05093, February 13, 2019

53 Eppel S, Aspuru-Guzik A. Generator evaluator-selector net: A modular approach for panoptic segmentation. [Online], available: https://arxiv.org/abs/1908.09108, August 24, 2019

54 Eppel S. Class-independent sequential full image segmentation, using a convolutional net that finds a segment within an attention region, given a pointer pixel within this segment. [Online], available: https://arxiv.org/abs/1902.07810, February 20, 2019

55 Chen Q, Cheng A D, He X Y, Wang P S, Cheng J. SpatialFlow: Bridging all tasks for panoptic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

56 Hou R, Li J, Bhargava A, Raventos A, Guizilini V, Fang C, et al. Real-time panoptic segmentation from dense detections. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 8520−8529

57 Cheng B W, Collins M D, Zhu Y K, Liu T, Huang T S, Adam H, et al. Panoptic-DeepLab. [Online], available: https://arxiv.org/abs/1910.04751, October 24, 2019

58 Bonde U, Alcantarilla P F, Leutenegger S. Towards bounding-box free panoptic segmentation. [Online], available: https://arxiv.org/abs/2002.07705, February 18, 2020

59 Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H, et al. Deformable convolutional networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 764−773

60 Ballard D H. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 1981, **13**(2): 111−122

61 Bai M, Urtasun R. Deep watershed transform for instance segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 2858−2866

62 Chang C Y, Chang S E, Hsiao P Y, Fu L C. EPSNet: Efficient panoptic segmentation network with cross-layer attention fusion. [Online], available: https://arxiv.org/abs/2003.10142, March 23, 2020

63 Wang H C, Luo R T, Maire M, Shakhnarovich G. Pixel consensus voting for panoptic segmentation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 9461−9470

64 Wang H Y, Zhu Y K, Green B, Adam H, Yuille A, Chen L C. Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. [Online], available: https://arxiv.org/abs/2003.07853, March 17, 2020

65 Arnab A, Torr P H S. Pixelwise instance segmentation with a dynamically instantiated network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 879−888

66 Rother C, Kolmogorov V, Blake A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions*

on Graphics, 2004, **23**(3): 309−314

67 Arbelaez P, Pont-Tuset J, Barron J, Marques F, Malik J. Multiscale combinatorial grouping. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 328−335

68 Lazarow J, Lee K, Shi K Y, Tu Z W. Learning instance occlusion for panoptic segmentation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 10717−10726

69 Liu Y, Perona P, Meister M. PanDA: Panoptic data augmentation. Liu Y, Perona P, Meister M. PanDA: Panoptic data augmentation. [Online], available: https://arxiv.org/abs/1911.12317, November 27, 2019

70 Li Q Z, Qi X J, Torr P H S. Unifying training and inference for panoptic segmentation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 13317−13325

71 Behley J, Milioto A, Stachniss C. A benchmark for LiDAR-based panoptic segmentation based on KITTI. [Online], available: https://arxiv.org/abs/2003.02371, March 4, 2020

72 Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, et al. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 9296−9306

73 Thomas H, Qi C R, Deschaud J E, Marcotegui B, Goulette F, Guibas L. KPConv: Flexible and deformable convolution for point clouds. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 6410−6419

74 Milioto A, Vizzo I, Behley J, Stachniss C. RangeNet++: Fast and accurate LiDAR semantic segmentation. In: Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE, 2019. 4213−4220

75 Lang A H, Vora S, Caesar H, Zhou L B, Yang J, Beijbom O. PointPillars: Fast encoders for object detection from point clouds. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 12689−12697

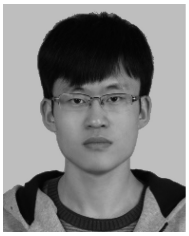
76 Mohan R, Valada A. EfficientPS: Efficient panoptic segmentation. [Online], available: https://arxiv.org/abs/2004.02307, April 5, 2020

77 Chen Y F, Lin G C, Li S Y, Bourahla O, Wu Y M, Wang F F, et al. BANet: Bidirectional aggregation network with occlusion handling for panoptic segmentation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020. 3792−3801

78 Hurtado J V, Mohan R, Burgard W, Valada A. MOPT: Multi-object panoptic tracking. [Online], available: https://arxiv.org/abs/2004.08189, April 17, 2020

79 Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. *Acta Automatica Sinica*, 2017, **43**(8): 1289−1305
(张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, **43**(8): 1289−1305)

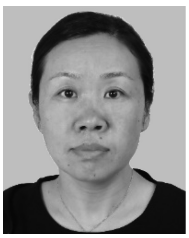
80 Meletis P, Wen X X, Lu C Y, de Geus D, Dubbelman G. Cityscapes-panoptic-parts and PASCAL-panoptic-parts datasets for scene understanding. [Online], available: https://arxiv.org/abs/2004.07944, April 16, 2020



徐鹏斌 北京化工大学信息科学与技术学院硕士研究生. 2019 年获得华北电力大学学士学位. 主要研究方向为深度学习, 计算机视觉, 图像全景分割.
E-mail: 2019210488@mail.buct.edu.cn
(**XU Peng-Bin** Master student at the College of Information Science and Technology, Beijing University of Chemical Technology. He received his bachelor degree from North China Electric Power University in 2019. His research interest covers deep learning, computer vision, and panoptic segmentation.)

瞿安国 北京化工大学信息科学与技术学院硕士研究生. 2018 年获得北京理工大学学士学位. 主要研究方向为深度学习, 计算机视觉, 图像全景分割.
E-mail: 2018210472@mail.buct.edu.cn
(**QU An-Guo** Master student at the College of Information Science and Technology, Beijing University of Chemical Technology. He received his bachelor degree from Beijing Institute of Technology in 2018. His research interest covers deep learning, computer vision, and panoptic segmentation.)

王坤峰 北京化工大学信息科学与技术学院教授. 主要研究方向为计算机视觉, 机器学习, 智能无人系统. 本文通信作者.
E-mail: wangkf@mail.buct.edu.cn
(**WANG Kun-Feng** Professor at the College of Information Science and Technology, Beijing University of Chemical Technology. His research interest covers computer vision, machine learning, and intelligent unmanned systems. Corresponding author of this paper.)



李大字 北京化工大学信息科学与技术学院教授. 主要研究方向为人工智能, 先进控制, 分数阶系统, 复杂系统建模与优化.
E-mail: lidz@mail.buct.edu.cn
(**LI Da-Zi** Professor at the College of Information Science and Technology, Beijing University of Chemical Technology. Her research interest covers artificial intelligence, advanced control, fractional order systems, and complex system modeling and optimization.)