计算机科学
COMPUTER SCIENCE

深度卷积神经网络图像实例分割方法研究进展

胡伏原，万新军，沈鸣飞，徐江浪，姚睿，陶重犇

引用本文

胡伏原，万新军，沈鸣飞，徐江浪，姚睿，陶重犇. 深度卷积神经网络图像实例分割方法研究进展[J]. 计算机科学，2022, 49(5): 10-24.

HU Fu-yuan, WAN Xin-jun, SHEN Ming-fei, XU Jiang-lang, YAO Rui, TAO Zhong-ben. Survey Progress on Image Instance Segmentation Methods of Deep Convolutional Neural Network[J]. Computer Science, 2022, 49(5): 10-24.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200038

深度卷积神经网络图像实例分割方法研究进展

胡伏原[1,2]　万新军[1,3]　沈鸣飞[1]　徐江浪[1,3]　姚　睿[4]　陶重犇[1,2]
1 苏州科技大学电子与信息工程学院　江苏 苏州 215009
2 苏州科技大学苏州市虚拟现实智能交互及应用技术重点实验室　江苏 苏州 215009
3 苏州科技大学苏州市大数据与信息服务重点实验室　江苏 苏州　215009
4 中国矿业大学计算机科学与技术学院　江苏 徐州 221116
(fuyuanhu@usts.edu.cn)

摘　要　图像实例分割是图像处理和计算机视觉技术中关于图像理解的重要环节,随着深度学习和深层卷积神经网络日趋成熟,基于深度卷积神经网络的图像实例分割方法取得了跨越性进展。实例分割任务实际上是目标检测和语义分割两项任务的结合,可以在像素层面完成识别图像中目标轮廓的任务。实例分割不仅可以定位图像中目标的位置,从像素层面上分割所有目标,还可以标注出图像中同一类别的不同个体,既是对图像的像素级分割,又是实例级理解。首先,阐述了图像实例分割产生的原因和深度卷积神经网络的作用。然后,根据图像实例分割方法的过程和特征,分别从两阶段和单阶段的角度介绍了图像实例分割的研究进展,详细阐述了两类方法的优势和不足,进而总结了各类实例分割方法对区域、特征提取和掩膜的设计思路。此外,归纳了图像实例分割方法的性能评价标准和常用的公开数据集,并在此基础上对比和评估了主流的图像实例分割模型的分割精度。最后,指出了当前图像实例分割存在的问题及解决思路,并对其未来发展进行了总结和展望。

关键词:实例分割;深度卷积神经网络;目标检测;语义分割;两阶段;单阶段
中图法分类号　TP391.4

Survey Progress on Image Instance Segmentation Methods of Deep Convolutional Neural Network

HU Fu-yuan[1,2],WAN Xin-jun[1,3],SHEN Ming-fei[1],XU Jiang-lang[1,3],YAO Rui[4] and TAO Zhong-ben[1,2]
1 School of Electronic & Information Engineering,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China
2 Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China
3 Suzhou Key Laboratory for Big Data and Information Service,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China
4 School of Computer Science and Technology,China University of Mining and Technology,Xuzhou,Jiangsu 221116,China

Abstract　Image instance segmentation is an important part of image processing and computer vision technology about image understanding. With the development of deep learning and deep convolutional neural network,image instance segmentation method based on deep convolutional neural network has made great progress. Instance segmentation task is actually the combination of target detection and semantic segmentation,which can complete the task of recognizing the target contour in the image at the pixel level. Instance segmentation can not only locate the position of the object in the image,segment all the objects from the pixel level,but also mark different individuals of the same category in the image,which is not only the pixel level segmentation of the image,but also the instance level understanding. Firstly,the reason of image segmentation and the function of deep convolution neural network are described. Then,according to the process and characteristics of image instance segmentation methods,the research progress of image instance segmentation is introduced from two-stage and single-stage perspectives,and the advantages and disadvantages of the two methods are described in detail. Then,the design ideas of region,feature extraction and mask are summarized. In addition,the performance evaluation criteria and common public data sets of image instance segmentation methods

到稿日期:2021-02-03　返修日期:2021-07-09
基金项目:国家自然科学基金(61876121,61801323);江苏省重点研发计划项目(BE2017663);江苏省高等教育自然科学研发项目(19KJB520054,19KJB110021,20KJB520018)
This work was supported by the National Natural Science Foundation of China(61876121,61801323),Primary Research & Development Plan of Jiangsu Province(BE2017663) and Foundation of Natural Science Research Program in Jiangsu Province Higher Education(19KJB520054,19KJB110021,20KJB520018).
通信作者:万新军(wanxinjun1030@126.com)

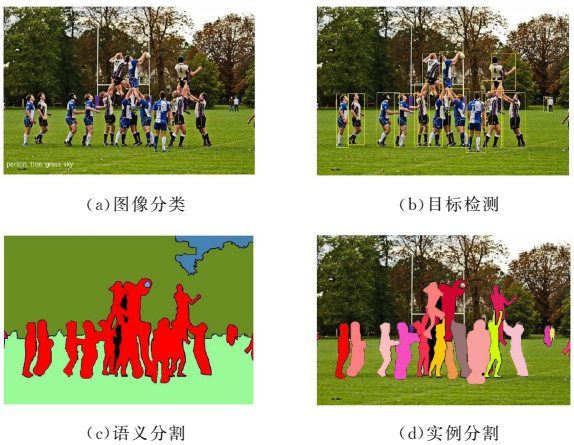
are summarized，and on this basis，the segmentation accuracy of mainstream image instance segmentation models is compared and evaluated. Finally，it points out the problems and solutions of the current image instance segmentation，summarizes the development of image instance segmentation and prospects for the future.

Keywords Instance segmentation，Deep convolutional neural network，Object detection，Semantic segmentation，Two stage，Single stage

1 引言

图像实例分割[1-2]是图像处理和计算机视觉领域的一个重要课题，它结合了目标检测[3]和语义分割[4]两项任务，可以在实例级检测到目标位置，并在像素级对不同类别的目标进行分割。图像实例分割在自动驾驶[5]、机器人导航[6]、视频监控[7]和医学图像分割[8]等众多领域具有十分重要的应用。

目标检测[9-11]不仅提供图像目标的类别，还以边界框的形式标记出图像目标的位置，如图 1(b)所示。语义分割[12-13]是对图像中每个像素点的标签进行预测，然后对目标和背景区域进行分类标记，如图 1(c)所示。但是，但语义分割网络中，一个像素点只能对应一种固定的语义，无法区分同一类别目标的不同个体，在进一步理解复杂场景图像内容的层面上，无法准确解析图像中不同目标的语义信息。而实例分割的出现很好地解决了这个问题，实例分割既是对图像的像素级分割，又是实例级理解，为属于同一类别的不同对象实例提供了不同的标签，如图 1(d)所示。相较于目标检测的边界框，实例分割可精确到目标的边缘；相较于语义分割，实例分割需要标注出图像中同一类别目标的不同个体。同为图像分割问题，语义分割主要研究图像各部分区域的像素级类别，实例分割更关注图像前景中的目标个体。语义分割图是一个像素级图层，不同区域位置的像素可用不同的物体类别数字表示，而实例分割图等效于多个图层叠加，每个前景目标占据一个图层。因此，利用检测技术识别图像内不同的目标个体是实例分割的特性问题，且实例数量的概念也是语义分割问题没有的。



(a)图像分类 (b)目标检测

(c)语义分割 (d)实例分割

图 1 计算机视觉中关于图像理解的 4 个基本任务

Fig. 1 Four basic tasks of image understanding in computer vision

在研究思路上，语义分割只需将图像分割成具有一定语义信息的区域块，并识别出每个区域块的语义类别，实现从底层到高层的语义推理过程，最终得到具有逐像素语义标注的分割图像。自全卷积网络被提出以来，现有的语义分割框架大都基于编码器-解码器(Encoder-Decoder)范式，其中编码器用于压缩原始输入图像的空间分辨率，并逐步地提取更加高级、抽象的语义特征；解码器则用于将编码器所提取到的高级特征上采样到原始输入分辨率，以进行像素级的预测。因此，语义分割的研究主要关注充分利用上下文信息，即扩大感受野来提高模型的表示能力。而实例分割扩展了语义分割的研究内容，除了像素级的语义类别，更关注不同的目标个体，一方面继续采用全卷积网络进行语义分割；另一方面，引入检测技术识别图像中的实例个体。

实例分割早期采用的方法有条件随机场(CRF)[14]、递归神经网络(RNN)[15-16]和模板匹配[17]等，但这些方法无法正确理解图像的语义信息，在分割性能和分割效率方面还有待改进。本文主要研究基于深度卷积神经网络(Deep Convolutional Neural Networks，DCNN)[18]的图像实例分割方法。随着深度学习[19-21]和深度卷积神经网络[22-24]的应用，多层卷积神经网络可以自动学习图像中的特征，许多实例分割框架被提出，比较经典的有 InstanceFCN[25]，FCIS[26]，Mask R-CNN[27]，PANet[28]和 MS R-CNN[29]等。

随着 GPU 和大数据带来的历史机遇，DCNN 可学习到更加丰富的特征，浅层提供的低层次细节特征有利于定位，深层包含的高级语义信息有利于分类，DCNN 表达能力更强，在许多视觉识别任务中均达到了最先进的性能。为了在 DCNN 卷积层中自适应地捕获空间相关性，更多复杂而有效的卷积网络架构被提出。一定条件下，随着网络层级的增加，模型精度得到提升，网络的表示能力和网络性能也会更好。DCNN 在图像中自适应地捕获局部上下文信息，在各种密集预测任务上取得了极大的性能提升，基于 DCNN 的图像实例分割方法也取得了跨越式进步。

本文根据图像实例分割方法的过程和特征，将实例分割方法分为两阶段和单阶段，阐述了基于深度卷积神经网络的图像实例分割方法的研究进展。图 2 是各时间节点的代表性实例分割算法。两阶段方法根据两个阶段的图像分割过程和掩膜生成方式分为基于检测的方法和基于分割的方法。单阶段方法按照有无锚框分为基于锚框的方法和无锚框的方法。本文首先详细介绍了两阶段和单阶段两类实例分割方法的优势和不足，然后总结了各类实例分割方法对区域、特征提取和掩膜的设计思路。此外，本文还整理归纳了图像实例分割方法的性能评价标准和常用的公开数据集，并在此基础上对比和评估了主流的图像实例分割模型在不同条件下的分割精度。最后，指出了当前实例分割存在的问题及解决思路，并对未来的发展进行了总结和展望。

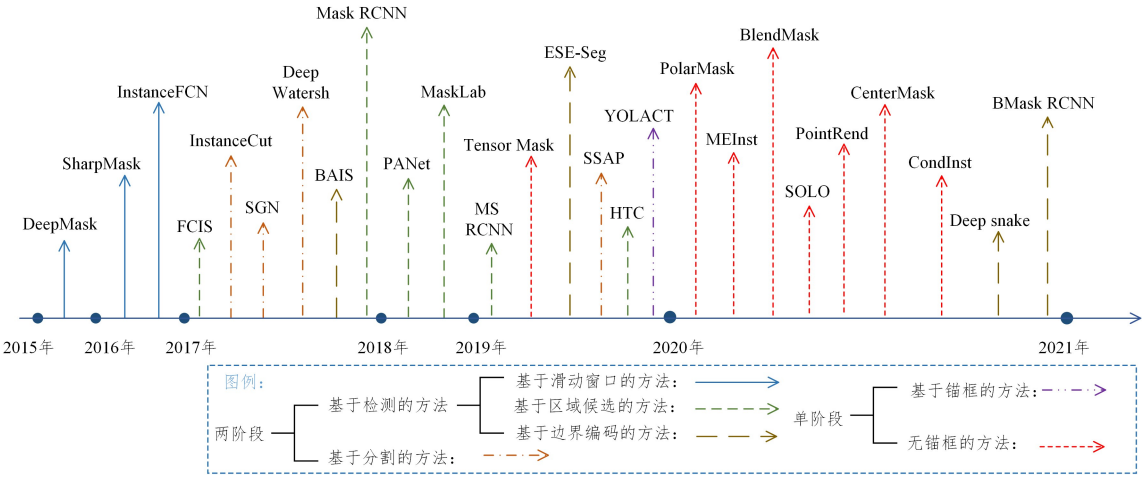


图 2 各时间节点的代表性实例分割算法

Fig. 2 Representative instance segmentation algorithm of each time node

2 两阶段图像实例分割方法

图像实例分割的研究长期以来都依赖于较为复杂的两阶段方法[25-29]，本文根据两阶段的图像分割过程和掩膜生成方式将两阶段方法分为两类，即基于检测的实例分割方法和基于分割的实例分割方法。

2.1 基于检测的实例分割方法

基于检测的方法首先通过检测器识别出目标的类别和位置，并用矩形框标记，然后在特定的区域内进行语义分割，同一类别的不同目标实例输出不同的分割掩膜。本节根据实例分割方法中采用的检测技术、获得候选子区域的方法以及是否关注实例边界，将基于检测的实例分割方法分为基于滑动窗口的方法、基于区域候选的方法和基于边界编码的方法。基于检测的实例分割方法可以使用前沿的目标检测框架，并且在检测框内进行实例分割，可以有效避免背景的干扰，实现精准分割。

2.1.1 基于滑动窗口的方法

两阶段实例分割方法在目标检测阶段首先生成候选子区域，然后在特定区域内进行目标识别和分割。生成候选子区域最直接的方法就是滑窗法，它使用不同大小的窗口在图像上进行滑动来寻找目标，并使用分类器判别滑动框中存在目标的概率。图 3 给出了基于滑动窗口方法进行实例分割的通用框架，即通过多尺度、多位置的目标掩膜和目标得分两个子分支进行实例分割。

Pinheiro 等[30]提出 DeepMask，该方法采用滑动窗口技术，将模型密集地用于图像的多个位置和多个尺度，在保证图像中的每个目标都至少有一个图像块能够被完全覆盖的前提下，每次以固定的像素步进行滑动。整个网络前半部分共享参数，得到底层特征，后半部分网络有两个分支，一个分支给出候选目标分割掩膜，另一个分支给出目标的分数。DeepMask 使用图像块作为卷积神经网络的输入，它会产生一个类别未知的分割掩膜和当前图像块中包含一个对象的可能性。它不依赖于超像素、边缘或者其他任何形式的低层级分割，直接从原始图像数据学习产生分割候选，最终能够达到更高的召回率以及更好的识别准确率。

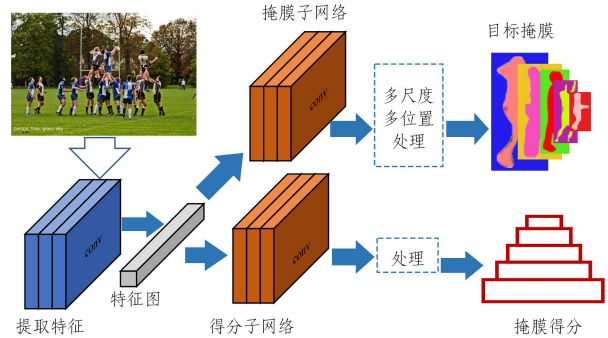


图 3 基于滑动窗口方法进行实例分割的通用框架

Fig. 3 General framework for instance segmentation based on sliding window method

Pinheiro 等[31]为了提炼 DeepMask 的输出，提出 SharpMask 算法。DeepMask 使用了一个非常简单的前馈网络来产生粗略的目标掩膜，而不是像素级别的准确分割。目标实例分割需要的是目标级别和像素级别的信息，但是对于前馈网络来说，卷积网络中下层获取了大量的空间信息，而顶层主要由目标水平的信息组成，在姿势和外形变化时不能达到很好的效果。SharpMask 从低层级图像中得到目标精确的边缘信息，与网络高层级的目标信息相结合，能够得到比较精确的边缘。该方法与 DeepMask 网络结合，采用自上而下的细化方法，将低维特征和高维语义信息相结合，提高了实例分割性能。

语义分割对同一类别的不同目标不作区分，但是实例分割需要像素针对不同的目标有不同的响应，因此 Dai 等[25]提出了 InstanceFCN。传统的全卷积网络(FCN)[32]中，每个图像只生成一个得分图(Score Map)，每个像素的值表示该像素是否属于目标的概率。在 InstanceFCN 中，会生成 $k\times k$ 个得分图，每个像素的值表示该像素是否属于某一类的某个相对位置的概率。例如，对于单个目标，将它分成 9 份，每个小窗口标记这个目标的不同位置，产生 9 个特征图，每个特征图识别该目标是否属于对应的位置，最终 9 个窗口对应的位置拼接成为一个窗口，窗口中每个像素有前景和背景两种标签。InstanceFCN 采用的是实例敏感得分图(Instance-Sensitive Score Map)，取得了更好的效果。

2.1.2 基于区域候选的方法

滑窗法类似穷举，其对图像子区域进行搜索，但是，一般图像中大部分子区域并没有目标。因此，为了提高计算效率，可以采用区域候选的方法生成候选子区域，只对图像中最有可能包含目标的区域进行检测。传统的提取候选框的方法是选择性搜索（Selective Search）[33]，但是比较耗时。2017 年 Faster RCNN 算法[34]提出区域候选网络（Region Proposal Network，RPN），通过 RPN 提取候选框生成感兴趣区域（Region of Interest，RoI），再进行分类和回归。实例分割则通过 RPN 区分前景和背景，再进行前景的目标分割。RPN 的使用极大地促进了两阶段实例分割的发展，基于区域候选网络的实例分割方法成为主流。

Li 等[26]提出了一个全卷积、端到端的网络框架 FCIS，该网络建立在 InstanceFCN[25]的基础上，基于 Faster RCNN[34]框架进行改进，使用位置敏感的特征融合方法进行特征提取，不仅共享卷积特征，也共享得分图。传统的全卷积网络[32]由于平移不变性、像素对位置不敏感，在任何位置都具有相同的响应。然而，实例分割要求同一像素在不同候选区域应该具有不同的语义信息，像素需要一定程度的位置敏感性。因此，FCIS 引入了区分同一像素在目标实例所属位置关系中的内部/外部得分图，与 RPN 网络共同作用并进行特征融合。FCIS 网络无额外参数，通过卷积特征表示和分数图完全共享给对象分割和检测子任务，网络结构高度集成且高效，最终在 COCO 2016 图像分割竞赛中获得了第一名。

FCIS 优化了网络结构，解决了同一像素在图像的不同区域具有不同响应的问题，但是分割精度仍然没有较大突破。2017 年 He 等[27]提出了一个两阶段实例分割框架 Mask RCNN，如图 4 所示。该方法在 Faster RCNN[34]的基础网络上加入了 mask 分支，用于生成目标掩模，同时把感兴趣池化（RoI pooling）修改为 RoI Align，用于处理掩膜与原图目标不对齐的问题，双线性差值法使候选区域和卷积特征的对齐不因量化而损失信息。Mask RCNN 框架的第一阶段扫描图像并生成候选区域，第二阶段对候选区域分类并生成边界框和掩膜，掩膜分支只进行语义分割，类别分支负责类型预测。Mask RCNN 算法让研究者看到了目标检测网络的高效使用对实例分割任务的促进作用，因此基于检测的两阶段实例分割算法也相继被提出，如 PANet[28]，MS RCNN[29]，HTC[35]等。

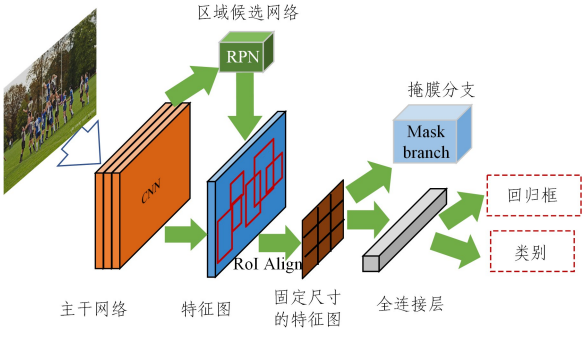


图 4 Mask R-CNN 算法的基本框架图

Fig. 4 Basic framework of Mask R-CNN algorithm

为提取图像多尺度特征，Mask RCNN[27]采用特征金字塔网络（FPN）[36]通过自上而下的多层特征融合线路，来增加每层特征的丰富性，图 5(a)给出了 FPN 主干网络。但是，整个网络中低层特征流向高层特征的路线过长，导致高层特征中包含的定位信息较少。针对此问题，Liu 等[28]提出 PANet 网络，加入了自底向上（Bottom-up）的短线路，如图 5(b)所示的自底向上的路径扩充，通过缩短高低层特征融合的路径，来增强信息在网络中的传播，提高了生成预测掩膜的质量。图 5(c)给出了自适应的特征池化，与 FPN 相比更为灵活，FPN 的 RoI 池化只从高层特征取值，而 PANet 在各个尺度特征中进行操作。图 5(d)给出了框分支，图 5(e)给出了全连接特征融合。该网络在 FPN[36]和 Mask RCNN[27]模型的基础上进行了改进，显著地提升了模型在目标检测和实例分割网络上的性能。

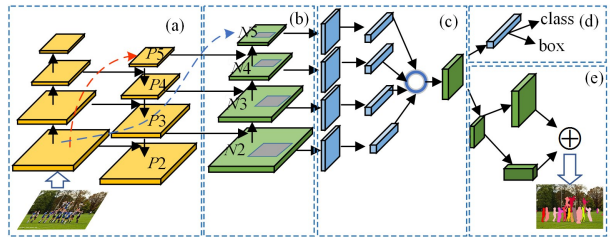


图 5 PANet 算法框架图

Fig. 5 PANet algorithm framework

通过 Mask RCNN[27]解决实例分割任务，掩膜分割质量是由检测分支的分类置信度决定。然而，掩膜的质量通常与分类置信度没有太大的关联。虽然输出结果是一个掩膜，但得分却是与目标检测的边界框共享的，都是针对目标区域分类置信度计算出的分数。它和图像分割掩膜的质量未必一致，用来评价预测掩膜的质量，可能存在偏差。于是 Huang 等[29]提出了 MS RCNN 框架，它的计分方式很简单，即同时考虑分类得分和模型针对掩膜的得分规则——MaskIoU head。实验证明，使用 MS RCNN 的掩膜得分评估时，在不同主干网络上，*AP* 始终提升近 1.5%。

Chen 等[37]根据实例分割的概念提出了 MaskLab 算法，该模型产生了 3 个输出：box 检测、语义分割和方向预测。MaskLab 采用 Faster RCNN[34]作为基础检测器，预测框提供了对象实例的准确定位。在每个感兴趣区域内，通过组合语义和方向预测来执行前景和背景分割。其中，语义分割有助于网络区分包括背景在内的不同语义类的对象，而方向预测估计每个像素朝向其相应中心的方向，实现分离同一语义类的实例。此外，MaskLab 还采用空洞卷积来提取更密集的特征图，使用超列特征进行细化掩模分割，使用多网格来捕获不同尺度的背景，以及利用可变形裁剪和调整大小操作获得上下文信息，MaskLab 算法最终取得了 35.4%的平均精度。

MaskLab 算法[37]采用多种技巧取得了一定的性能提升，但是 box 检测、语义分割和方向预测 3 个输出特征未能充分融合。因此，有研究人员提出了级联（Cascade）结构，它通过多阶段优化提升了各种任务的性能，将级联的思想整合到实例分割中将十分有效地解决多任务的复杂问题。较早的实例分割模型 MNC[38]采用多任务方法，在共享底层特征的基础上，形成级联的多任务结构。Wen 等[39]提出一种联合多任务

级联结构，在全卷积网络(FCN)[32]分支中引入了特征融合过程，有效地融合了高层和低层特征，增强了图像语义特征的上下文信息。Chen 等[35]提出混合任务级联(HTC)，它通过在每个阶段合并级联和多任务来改善信息流，并利用空间上下文来进一步提高分割准确性。HTC 将检测和分割功能交织在一起，从而有效地将级联集成到实例分割中，以进行联合多阶段处理。它在 COCO 基准测试中取得了显著的性能提升。

2.1.3 基于边界编码的方法

基于区域候选的实例分割方法依靠全卷积网络(FCN)[32]进行像素级分类，FCN 平等地对待所有的像素，忽略了目标的形状和边界信息，导致掩模预测结果粗糙，定位模糊。然而，靠近边界的像素点很难分类，像素级分类器很难保证精确的掩膜分割。因此，采用基于边界编码的实例分割方法可以提供更好的定位性能，使目标掩膜更加清晰。近年来这类方法得到了更多的关注。

Hayder 等[40]提出了边界感知实例分割网络(BAIS)，它使用基于边界的距离变换来预测超出边界框的掩模像素。其首先设计了一个具有残差反卷积结构的目标掩模网络(OMN)，提取特征并将其解码成最终的二值目标掩模。然后设计了一个深度网络，输入图片，输出普适性的实例掩模，它可以超过最初候选框的范围，并且以端到端的方式学习。这种方法能摆脱传统区域候选中候选框的范围局限，并且对不够准确的区域候选具有鲁棒性。

但是，当图像中的目标数量较多时，BAIS[40]等方法通常按顺序预测实例掩膜，速度会受限。Xu 等[41]提出了一种基于显式形状编码的 ESE-Seg 实例分割框架，如图 6 所示。它使用边界编码的方法，推理时间不受图像中目标数量的影响，利用张量运算显式地解码多个目标形状，极大地减少了实例分割的计算消耗。显式形状表示通常基于轮廓，明确的形状表示不需要额外的解码器网络训练，可以很容易地实现对图像中所有对象的并行解码，通过一次遍历得到所有的形状。ESE-Seg 框架[41]基于形状特征内中心半径和切比雪夫多项式拟合设计，内中心半径首先定位目标线段内的一个内中心，然后根据这个内中心将轮廓点转换为极坐标。为了使形状向量更切合，应用切比雪夫多项式进行函数逼近。这样可以通过快速张量运算来实现实例形状解码，从而使实例分割达到目标检测的速度。

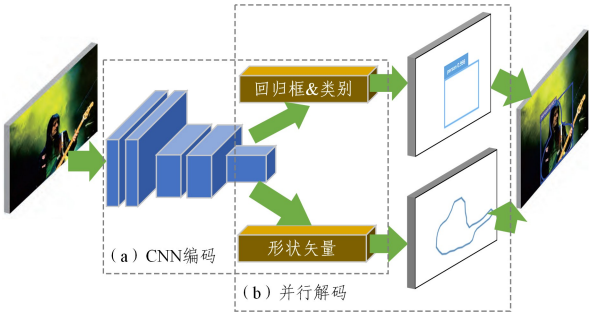


图 6 ESE-Seg 算法框架图

Fig. 6 ESE-Seg algorithm framework

ESE-Seg 框架[41]显著提高了实例分割的速度，但是精度提升有限。因此，Cheng 等[42]提出了 BMask RCNN 框架，它明确地预测了实例级边界，可以获得更准确的实例形状信息，从而更好地进行掩膜定位。与语义分割相比，实例分割中的边界与掩膜具有双重关系。因此，通过构建融合块，相互学习边界和掩模特征，改进掩模定位表示，可以使掩模预测更加关注边界，从而使目标掩膜更加清晰，达到更好的分割性能。

在给定的回归框中进行逐像素分割存在诸多限制，比如依赖回归框的准确性、计算量大等。因此，Peng 等[43]借鉴传统的 snake 算法[44]提出了 Deep snake 算法，其选择用轮廓结构化特征学习来表示目标的形状，其中轮廓是一组首尾相连的有序点。这种方法的参数量远小于稠密像素，速度上限更高，而且轮廓更适用于细胞、文字这些目标的分割。该算法通过目标检测的定位来初始化建议轮廓，然后对建议轮廓进行变形，通过迭代式轮廓调整得到目标形状。Deep snake 算法采用轮廓调整和循环卷积，不仅提升了性能还减少了计算量，保持了实时性。

2.2 基于分割的实例分割方法

基于分割的实例分割方法首先利用语义分割子网络对图像进行像素级别的分割，再通过聚类、度量学习等方法将每个对象的像素组合，不同的目标实例输出不同的分割掩膜。如图 7 给出了基于语义分割方法进行实例分割的基本框架。

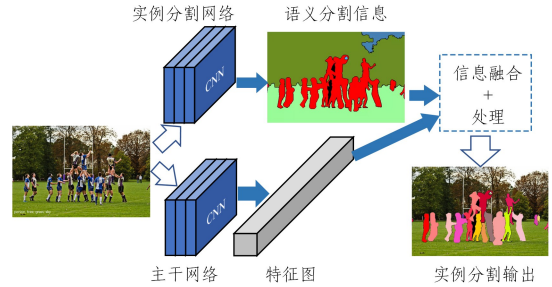


图 7 基于语义分割方法的实例分割框架

Fig. 7 Instance segmentation framework based on semantic segmentation method

Uhrig 等[17]利用预测的实例中心和像素方向将实例与全卷积网络进行分组。Kirillov 等[45]提出 InstanceCut，采用边界将语义分割划分为实例级分割。Liu 等[46]提出序列分组网络 SGN，将实例分割的复杂任务用一系列神经网络来解决，每个神经网络解决一个增加语义复杂度的子类问题，从而使用简单的结构逐渐构造目标实例。

2017 年有几种方法利用深度度量学习来学习实例的嵌入，对像素进行分组，形成实例级分割。De Brabandere 等[47]采用自定义判别损失函数训练网络学习到一种度量，即从像素空间到高纬度空间的映射，使得同一实例中的像素映射到高维空间后，得到的嵌入向量(Embedding Vector)之间的距离相近，最后使用聚类的方法输出不同实例，完成分割任务。Fathi 等[48]采用深度度量学习的方法，引入种子模型，为每个像素学习种子得分，分数决定像素是否是扩展掩膜的良好候选，帮助网络分类并拾取最佳种子。Kong 等[49]提出用于实例分组的递归像素嵌入，提出在超球面上进行嵌入，并利用余弦距离来度量像素的接近程度。

Bai 等[50]结合传统的分水岭算法和深度学习算法，提出用于实例分割的深分水岭变换。该方法生成能量图，能量图中的每一个实例对应一个能量盆地，然后单个能级执行切割，直接生成与实例对应的组件。其中定义了一个中间任务，即学习能量下降的方向，然后将学得的方向信息传给下一部分网络层，从而得到最终的能量图。对于每个实例，学习其中每个像素与该实例中距离该像素最近的边界之间的距离。该方法可以进行端到端的训练，并生成快速、准确的估算，但是依然存在过分割的问题。

Gao 等[51]提出了 SSAP 框架，该方法基于亲和金字塔来区分实例，可以与像素级语义类标签在一个单一的骨干网络中共同学习，以分层的方式计算两个像素属于同一实例的概率，实现从粗到细生成目标实例。

基于分割的实例分割方法一直以来都不是主流，与基于检测的方法相比，这类方法的结果并没有竞争力，一般无法进行端到端训练。基于分割的方法虽然保持了更好的低层特征，如细节信息和位置信息，但会导致非最优的分割，并且泛化能力差，无法应对类别多的复杂场景。但是，它们比基于检测的方法更简单，也能避免检测框存在的缺陷。如果能够将检测和分割的思想结合起来使用，将会产生更好的效果。

3 单阶段图像实例分割方法

两阶段实例分割方法虽然精度不错，但是很难达到实时分割的速度。自从单阶段目标检测框架被提出，其简单、灵活、速度快、精度高的优点极大地影响了实例分割的研究。相较于单阶段目标检测，单阶段的实例分割更困难，一个阶段意味着需要同时定位、分类和分割对象。不同于目标检测用两个角的坐标即可表示目标的预测回归框，实例分割掩膜的形状和大小都更为灵活，很难用固定大小的向量来表示。如何区分不同的目标实例，尤其是同一类别的不同目标也较为困难。

当前主流的单阶段实例分割方法根据有无锚框可分为基于锚框的实例分割方法和无锚框的实例分割方法。

3.1 基于锚框的实例分割方法

锚框(anchor)是预定义的框集合，其宽度和高度与数据集内目标的宽度和高度相匹配。在两阶段实例分割方法中，候选区域类似于锚框。实例分割算法借助锚框更容易回归不同尺度、不同长宽比的目标，促进实例掩膜的预测。

Bolya 等[52]提出了实时的实例分割框架 YOLACT，如图 8 所示，YOLACT 算法将整个任务划分为两个并行的分支：第一个分支使用 FCN 网络[32]为每个锚框生成一组图像大小的原型掩膜(Prototype Masks)；第二个分支给原版的目标检测网络添加了额外的输出，用来预测原型掩膜中每个锚框里实例的掩模系数(Mask Coefficients)。然后利用非极大值抑制(NMS)来消除重复的锚框，并将两分支的输出结果进行线性组合。最后根据回归框进行裁剪、阈值化，得到每个实例对应的掩膜。其中，消除多余的框采用的是改版后的 Fast NMS，其比标准 NMS 快 12 ms，并且性能损失很小。

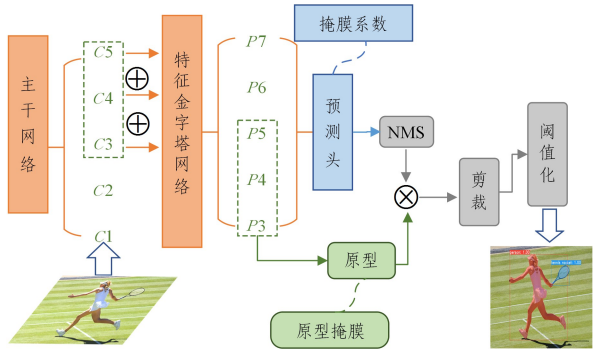


图 8 YOLACT 算法网络结构图

Fig. 8 YOLACT algorithm network diagram

预测一组原型掩膜和掩膜系数需要更丰富、更高级的特征，所以在网络设计上，YOLACT 使用 ResNet-101 结合 FPN 作为主干网络，与原版 RetinaNet[53]检测器相比，YOLACT 检测头的设计更轻量，速度更快。由于采用了并行结构与轻量化的组合方式，其在单阶段检测网络的基础上只添加了少许计算量，在主干网络为 ResNet-101＋FPN 的情况下可达到 30FPS。两阶段实例分割方法通过 ROI Align 等特征定位步骤保留了空间信息，同时使用卷积层输出掩膜，但是这些操作都必须等待 RPN 来完成，极大地影响了效率。在 YOLACT 中，全卷积层负责预测语义标签，卷积层负责预测原型掩膜和掩膜系数。两分支并行，最后通过矩阵乘法组合，既保留了空间的相关性，获得无损失的特征信息，又保持了单阶段的模型结构，速度更快。但是，YOLACT 网络在注重速度的同时，分割精度并不突出。由于掩膜是在合成以后裁剪得到的，没有抑制框外部噪声的功能，如果锚框定位不准，或多个大小差异较大的同类实例相隔较远，可能导致掩膜泄露。

2019 年 Bolya 等[54]在 YOLACT 算法[52]的基础上提出 YOLACT＋＋算法，在保证实时性的前提下，YOLACT＋＋算法对原版算法作出了 3 点改进，极大地提高了分割精度。首先，YOLACT＋＋受 MS RCNN[29]启发，添加了 mask re-scoring 分支，使得模型可以结合框和 mask 的质量对预测结果进行综合排序，与仅依赖分类置信度进行排序的算法相比，收效显著；接着，参考可变形卷积的思路，将主干网络 ResNet 中 C3—C5 层的各个标准 3×3 卷积替换成 3×3 可变形卷积，掩膜的分割精度提高了 1.8%；最后，优化了预测头的结构，以更好的锚定比例和长宽比改进了锚框的设计。YOLACT ＋＋模型以 33.5 FPS 的速度在 COCO 基准上实现了 34.1%的分割精度。

此外，Liu 等[55]在 YOLACT 算法[52]的基础上提出了 YolactEdge 算法，这是第一个在小型边缘设备上以实时速度运行的实例分割算法。在 550×550 分辨率的图像上，以 ResNet-101 为主干网络的 YolactEdge 在 Jetson AGX Xavier 上的运行速度高达 30.8 FPS，在 RTX 2080 Ti 上的运行速度为 172.7 FPS。

3.2 无锚框的实例分割方法

YOLACT[52]采用的单阶段目标检测网络 RetinaNet[53]严重依赖于预定义的锚框，锚框对超参数(如输入大小、高宽

比、比例等)和不同的数据集非常敏感。此外,由于密集地放置锚框以提高召回率,过多的锚框会造成正负样本不平衡和较高的计算成本。

2019 年,CenterNet[56] 和 FCOS[57] 等单阶段无锚框目标检测网络被提出,此类方法不依赖预定的锚框,直接预测回归框所需的全部信息,如位置、框的大小和类别等。此外,在每个位置只预测一个目标的情况下,FPN[36] 结构弥补了多尺度信息,Focal Loss[53] 促进了对中心区域的预测。无锚框架构简单、灵活、速度快,因此成为单阶段实例分割的首选。本节根据无锚框实例分割方法使用的编码方法、设计思想、融合技巧和卷积核形式将无锚框的方法大致分为 7 类,即基于轮廓编码的方法、基于掩膜编码的方法、基于渲染的方法、融合检测和分割的方法、融合局部和全局信息的方法、基于动态卷积核的方法和基于位置信息的方法。

3.2.1 基于轮廓编码的方法

Xie 等[58] 参考目标检测模型 FCOS[57] 提出了 PolarMask 算法,相比 RetinaNet[53],FCOS 将基于锚框的回归变成了中心点估计与上下左右 4 个边界距离的回归,而 PolarMask 则进一步细化了边界的描述,相比 FCOS 把 4 根射线发散到 36 根射线,使其能够适应掩膜的表示。PolarMask 算法基于极坐标系建模轮廓,把实例分割问题转化为实例中心点分类问题和密集距离回归问题。同时,PolarMask 算法还提出 Polar Center Ness 和 Polar IoU Loss,分别用来优化高质量正样本采样和密集距离回归的损失函数。

PolarMask 算法通过寻找目标的轮廓建模,虽然在精度和速度上并不突出,但它提供了一种新的实例分割建模方式和研究思路。

3.2.2 基于掩膜编码的方法

基于轮廓编码的方法使用一组轮廓系数来对目标形状进行编码,但是这类方法预测的掩膜不可避免地出现空心衰减,即无法很好地表示含有孔洞的目标,因此它们只能描绘具有单个轮廓的实例。Zhang 等[59] 提出了一种基于掩膜编码的 MEInst 算法,与直接预测二维掩膜不同,MEInst 算法将其提取为紧凑而固定维度的表示向量,通过在现有的单阶段目标检测器上附加并行回归分支来解决实例分割任务,从而得到了简单高效的实例分割框架。MEInst 网络以 FCOS[57] 为基线,包含骨干网络和特征金字塔部分,以及两个用于分类、框回归和中心度计算的专用头(它们共享同一分支),然后采用并行分支来预测编码掩膜系数。

图像中区分像素通常沿着对象边界分布,而其目标主体中的大多数像素都具有类别连续和类别一致的属性,因此给定一个结构化实例掩膜,MEInst 算法可以找出其表示形式中的冗余,并将二维掩膜编码成更加紧凑的表示向量,从而提升实例分割的速度和精度。

3.2.3 基于渲染的方法

Kirillov 等[60] 借鉴经典的计算机图形学思想将图像分割视为一个渲染问题,提出基于迭代细分的 PointRend 神经网络模块,在自适应选择的位置执行基于点的分割预测。图片中低频区域的点大概率属于同一个物体,因此不需要使用太多采样点,且采样点较多相当于过采样。而图片中高频区域的点大概率靠近物体边界,采样点太稀疏会导致分割出的物体边界过于平滑、不真实,相当于欠采样;相反,采样点越多,分割出来的物体边界会更精细、真实。借鉴渲染的思路,渲染器将 3D 网格映射到像素的规则网格,先通过一种合理的采样方式进行非均匀采样,计算出每个采样点的分割结果,然后再将结果映射到规则的方格里。

PointRend 模块主要由 3 部分组成。1)点选择策略。灵活、自适应地选择合适的采样点,高频区域多采点,低频区域少采点。通过计算与其近邻的值显著不同的位置,来高效渲染高分辨率图像。2)逐点表示。通过组合低层特征和高层特征,在选定的点上构造逐点特征。实值点的特征通过在特征图上使用双线性插值得到。3)点头。小型神经网络用于基于逐点特征表示,预测标签。PointRend 使用简单的多层感知器进行逐点分割预测,多层感知器在所有点上共享权重,可以输出清晰的对象边界。

3.2.4 融合检测和分割的方法

两阶段实例分割方法分为基于检测和基于分割两种思路,但是单独采用一种思路都存在一些问题:基于检测的模型存在冗余的特征提取,特征和掩膜之间的局部一致性会丢失,并且由于使用了缩小特征图的卷积,会损失位置信息;基于分割的模型严重依赖逐像素预测的质量,容易导致非最优分割,由于掩膜在低维提取,模型对复杂场景的分割能力有限。因此,如果能结合两种思路的优势,充分融合基于检测的方法提供的高级全局信息和基于分割的方法包含的细节和位置信息,则实例分割的性能必然会得到极大的提升。

Chen 等[61] 基于单阶段目标检测器 FCOS[57] 提出 BlendMask 算法,将基于检测的方法生成的实例级高维信息和基于分割的方法生成的逐像素预测信息进行融合。BlendMask 借鉴了 FCIS[26] 的裁剪思想和 YOLACT[52] 的加权思想,提出了 Blender 模块,其能够更好地融合包含实例级的全局性信息和提供细节和位置信息的低层特征。BlendMask 算法的网络架构如图 9 所示,其整体架构包括 FCOS[57] 检测器网络和掩膜分支。

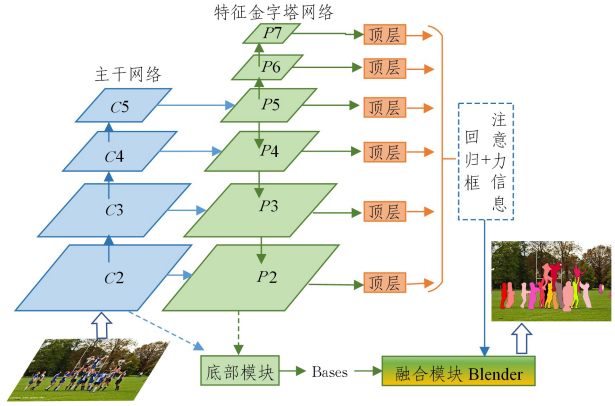


图 9 BlendMask 算法网络结构图

Fig. 9 BlendMask algorithm network diagram

掩膜分支包括 3 个部分:1)用于预测分数图的底部模块;2)用于预测实例注意力的顶层;3)用于将分数图与注意力

合并的 Blender 模块。底部模块与基于候选的全卷积方法类似，BlendMask 框架添加了一个称为 Bases 的底部模块来预测分数图，它的输入可以选择主干特征图或者 FPN 的特征图。顶层在每个检测塔的顶端单独添加了一个卷积层来预测顶层的注意力信息，比 YOLACT[52] 网络多了一个注意力图。Blen-der 模块是 BlendMask 的核心部分，它结合了与位置相关的注意力信息来生成最终的预测结果。

BlendMask 使用注意力引导的融合模块来计算全局特征，计算量更小，并且它属于密集像素预测，输出的分辨率不会受到限制，推理速度也更快，最终取得了 38.4% 的分割精度。

Ying 等[62] 同样融合了自上而下和自下而上两种思路，基于 FCOS 网络提出了 EmbedMask 算法。与基于检测的方法一样，EmbedMask 建立在检测模型之上，具有强大的检测能力。同时，EmbedMask 应用额外的嵌入模块来生成像素和候选的嵌入，如果像素嵌入属于同一实例，则由候选嵌入来指导像素嵌入。通过这个嵌入耦合过程，如果像素嵌入相似，则将像素分配为掩膜。像素级聚类使 EmbedMask 能够在不丢失定位细节的情况下生成高分辨率的掩膜，而候选嵌入的存在简化和加强了聚类过程，以获得比基于分割的方法更快的速度和更好的性能。EmbedMask 算法的关键是特别设计了额外的模块来学习像素嵌入、候选嵌入和候选边距，以提取实例掩膜。作为一种单阶段的实例分割方法，EmbedMask 算法可以在 COCO 基准测试中获得与 Mask RCNN[26] 相当的分数，同时提供比 Mask RCNN 更高的质量，运行速度更快。

3.2.5 融合局部和全局信息的方法

单阶段实例分割主要面临两个挑战：对象实例区分和像素级特征对齐。因此，Wang 等[63] 基于单阶段检测器 CenterNet[56] 提出了 CenterMask 算法，将实例分割分解为两个并行的子任务：局部形状预测和全局显著性生成。第一个分支从目标的中心点表示中获取粗糙的形状信息，用于约束不同目标的位置区域以自然地将不同的实例进行区分；第二个分支以像素到像素的方式分割整个图像，对整张图像预测全局的显著图，用于保留准确的位置信息，实现精准的分割。最后，通过将两个分支的输出相乘来构造每个实例的掩膜。

可视化结果表明，仅具有局部形状分支的中心掩膜可以很好地分离对象，而仅具有全局选择性分支的中心掩膜在对象不重叠的情况下表现良好。在复杂和对象重叠的情况下，这两个分支相结合可以区分实例，同时实现像素级分割。此外，该方法还可以方便地嵌入到 FCOS[57] 等单阶段目标检测器中，具有良好的性能，体现了它的通用性。

3.2.6 基于动态卷积核的方法

目前效果较好的实例分割方法通常依靠 ROI 操作来获取最终的实例掩膜，但是 ROI 通常使用矩形框，对于形状不规则的对象，框内可能包含过多的不相关内容，如背景和其他目标，使用旋转的 ROI 可以缓解此问题，但是计算量更大。为了区分前景实例和背景，掩膜分支需要较大的感受野来对上下文信息进行编码，大量 3×3 卷积的使用增加了计算复杂度。此外，为了使不同大小的 ROI 能进行批量计算，需要将其调整为统一尺寸，这就限制了实例分割的输出分辨率，而大型实例需要更高的分辨率才能保留边缘细节。考虑到基于 ROI 的方法存在很多不足，Tian 等[64] 提出了 CondInst 算法，这是一个简单而有效的实例分割框架。

CondInst 算法的网络结构如图 10 所示，其中上半部分沿用了目标检测网络 FCOS[57] 的基本结构，$C3$，$C4$ 和 $C5$ 是主干网络的特征图，$P3$ 到 $P7$ 是特征金字塔网络的特征图。head 层的输出分为类别头和控制头，类别头用于预测位置 (x,y) 处目标实例的类别概率 $p(x,y)$，与 FCOS 中相同。控制头采用以实例为条件的动态实例感知网络动态生成卷积核参数，应用于掩膜分支的动态卷积头，应用的次数等于图像中目标实例的数量。

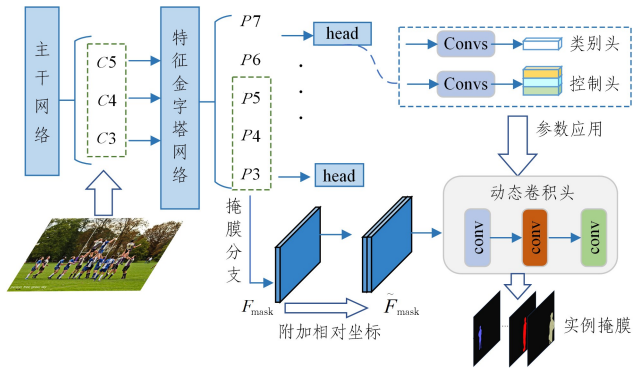


图 10 CondInst 算法的网络结构图

Fig. 10 CondInst algorithm network diagram

下半部分的动态卷积头是 CondInst 算法的核心，掩膜分支中会以实例为条件动态生成卷积，并且仅要求卷积预测一个实例的掩膜，从而减轻了卷积的负载。掩膜分支的结构就是一般的 FCN[32] 网络，但是它的特点在于 FCN 的参数是动态的，不同的实例有不同的参数，功能上类似于用不同的框区分不同的实例。该框架通过全卷积网络解决实例分割任务，无须进行 ROI 裁剪和特征对齐即可获得具有更高精确度的高分辨率实例掩膜。

3.2.7 基于位置信息的方法

语义分割是逐像素的语义类别分类，即预测每个像素所在目标的语义类别。同理，实例分割也可以定义为逐像素的实例类别分类，预测每个像素所在目标的实例类别。Wang 等[65] 提出了 SOLO 算法，将实例类别定义为目标的位置和尺寸，并将实例分割任务分为类别预测和生成实例掩膜两个子任务，思路非常巧妙，模型精度也不错，在 COCO 评测中超越了 Mask R-CNN[27] 和其他单阶段实例分割模型。SOLO 算法的思想是：将图片分成 $S\times S$ 的网格，若目标质心的位置处于某个网格中，该网格负责预测两项任务：1）目标的语义类别；2）目标的实例掩膜。SOLO 参照语义分割，将定义的目标中心位置的类别放在通道维度上，保留了几何结构上的信息。对于尺寸的处理，SOLO 使用特征金字塔网络将不同尺寸的目标分配到不同层级的特征图上，依次作为目标的尺寸类别，以区分所有的实例，再使用实例类别来分类目标。

SOLO 在预测实例掩膜时，有 $S\times S$ 个网格，但其中有很多是冗余的，并不是每个网格都有目标。为节省资源，SOLO

将分割图分解为 X 方向的分割图和 Y 方向的分割图，称作解耦头(Decoupled Head)。原来的实例掩膜分支有 $S\times S$ 个通道，但是有实例的网格远远少于没有实例的网格，所以解耦头将通道减少为 $S+S$，减少显存的同时精度并没有降低。最终的实例分割掩膜是 Y 分支与 X 分支掩膜对应位置相乘后的结果。SOLO 算法分类问题监督明确，使用 focal loss[53] 和 FPN[36]，把一张图的结果分散到不同的特征图中，降低了每张图的复杂程度。但是当场景非常复杂时，分类问题的难度将呈指数级递增。此外，SOLO 仅仅通过绝对位置坐标来关联融合，高维信息不够丰富。

SOLOv2[66] 是原作者对 SOLO[65] 的改进，主要基于位置信息生成实例掩膜，在掩膜分支的设计中借鉴了动态卷积核的思想。SOLOv2 提出动态头(Dynamic Head)，将 SOLO 中的掩膜分支改为内核分支(Kernel Branch)和特征分支(Feature Branch)两个分支，内核分支由卷积核生成，每个网格对应一个卷积核。每个网格对应的实例分割图为该网格预测的卷积核与特征图的卷积结果。与 SOLO 相比，除了将 SOLO 中的掩膜分支(Mask Branch)替换为动态头外，其他地方未做明显改动，但是与 SOLO 相比，其速度更快、精度更好。SOLOv2 将原始的掩膜预测直接解耦到卷积核学习和特征学习，不需要锚框，无须标准化，不需要边界框检测，直接将输入图像映射到所需的对象类和对象掩膜，训练和推论都更简单。

4 实例分割设计思路

实例分割针对图像目标进行像素级预测，无论是两阶段还是单阶段的分割方法，都在图像范围内基于区域进行操作，提取目标特征并生成目标掩膜。因此，实例分割的设计思路可分为 3 步：1)设计区域的范围；2)提取区域内目标特征；3)预测和表征区域内的目标掩膜。

4.1 区域设计

实例分割中关于区域的设计，无论是滑窗法中的窗口，还是 RPN 产生的感兴趣区域(ROI)，又或者是锚框，都是区域的不同表示形式。目标检测是实例分割的子任务，目标检测框架中关于区域的设计为：Cornernet[67] 定义为角点，ExtremeNet[68] 定义为极值点和中心点，FSAF[69] 和 FoveaBox[70] 定义为矩形框的中间区域，FCOS[57] 设计的区域虽然是矩形框，排除低质量的框后也是将回归区域定义为矩形框中心区。因此，实例分割模型中的区域设计可以借鉴目标检测算法中关于真实回归区域的设定。

设计区域范围时首先要保证区域足以覆盖图像中的所有目标，然后尽可能减少区域的数量，避免更高的计算成本，找到范围和数量的折中。

4.2 特征提取设计

传统的特征提取方法有 SIFT[71](尺度不变特征变换)、HOG[72](方向梯度直方图)等，实例分割采用的是深度学习的方法，它是一种自学习的特征表达方式，相比 SIFT，HOG 等依靠先验知识设计的特征，其表达效果更好。实例分割任务中，如何设计高效的特征提取方式，使其可以适配多尺度、多分辨率的区域，并可以充分融合区域特征与局部特征，是实例分割过程中十分重要的一步。

目前特征提取设计有很多选择，如 ROI Align、ROIConv、可变形卷积[73]、动态卷积和深度可分离卷积[74]等，甚至普通的 3×3 卷积[75]，也可以完成规则区域的特征提取。

4.3 掩膜设计

在不同区域内对目标掩膜进行预测和表征是实例分割的核心问题，也是当前主流方法的创新点。同时，为了解决目标掩膜的信息冗余，还可以将目标掩膜参数进行压缩以减少运算量。本节将从局部图像掩膜、全局图像掩膜和掩膜压缩的角度分析当前各类实例分割方法的设计思路。图 11 为局部掩膜和全局掩膜示意图。

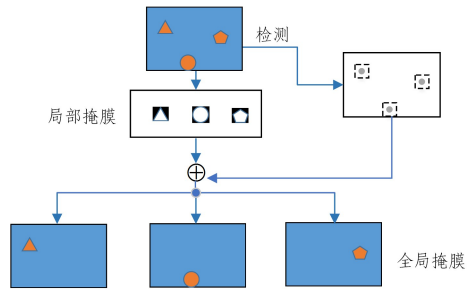


图 11 局部掩膜和全局掩膜示意图

Fig. 11 Schematic diagram of local mask and global mask

4.3.1 局部图像掩膜

局部掩膜的方法基于图像局部的信息输出实例的分割结果。以 Mask RCNN[27] 为代表的两阶段实例分割方法，结合目标检测的定位功能，在 RPN 生成的局部区域内预测和表征目标掩膜。基于轮廓、边界编码的实例分割方法也属于局部图像掩膜。如 PolarMask[58] 用目标中心点发出的射线组成的多边形来描述目标的轮廓；ESE-Seg[41] 通过显式形状编码，利用张量运算快速地解码多个目标形状；TensorMask[76] 利用结构化的 4D 张量来表示空间域上的掩膜。

局部图像掩膜通常更为紧凑，并且可以和目标检测算法结合起来，同时速度相对较快，占用的显存更小，但是它必须与掩膜位置一起使用才能恢复到全局掩膜。

4.3.2 全局图像掩膜

全局掩膜的方法首先基于整张图像生成中间共享特征图，然后组合提取的特征以形成每个实例的最终掩膜。YOLACT[52] 首先生成全局图像的多张原型掩膜，然后利用针对每个实例生成的掩膜系数对原型掩膜进行组合。BlendMask[61] 提出的 Blender 模块用于融合包含实例级的全局性信息和低层特征。CenterMask[63] 将实例分割分解为两个并行的子任务：局部形状预测和全局显著性生成。第二个分支利用全局掩膜保留准确的位置信息，实现精准分割。CondInst[64] 的掩膜分支直接通过基于全卷积的动态实例感知网络生成全局掩膜。

基于全局掩膜的方法直接从原图分割出掩膜，能够较好地保留目标的位置信息，实现像素级的特征对齐，小目标不会丢失全局信息，大型目标的边界也会更加清晰。但是全局掩膜的表征比较昂贵，显存和速度都可能成为瓶颈。

4.3.3 掩膜压缩

预测目标掩膜时，并非所有参数都在模型中发挥作用，存在一定的信息冗余，因此需要压缩掩膜的参数。假设掩膜的分辨率相同，区域的个数设为 N，每个区域内掩膜的分辨率设为 $w\times h$，最终需要预测 $N\times w\times h$ 的张量。但是，直接计算会导致计算量大和运算浪费。如果使用矩阵分解的方法压缩掩膜张量，可减少模型的冗余和计算量，最终提高掩膜的质量和运算效率。

(1)压缩区域个数 N：并非所有区域都需要预测掩膜，如果可以提前确定某个区域中并不包含目标，那么这个区域就可以直接过滤掉。

(2)压缩掩膜分辨率 $w\times h$：将二维掩膜提取为紧凑而固定维度的表示向量，基于掩膜进行编码；或者通过解耦头将 $w\times h$ 的矩阵拆分为 $w\times 1$ 和 $1\times h$ 的两个向量的乘积，减少通道的数量，参数量变少，计算成本降低。

(3)压缩需要预测的张量 $N\times w\times h$：把 $N\times w\times h$ 分解为两个矩阵的乘积，通过 $N\times k$ 和 $k\times w\times h$ 的两个矩阵相乘得到最终的 $N\times w\times h$ 矩阵 。

当前绝大多数算法都是上述 3 种设计思路的排列组合，虽然取得了一定的性能提升，但仍存在局限性。首先，当前实例分割方法难以摆脱区域的限制，区域的范围和数量影响了分割的效率和精度；其次，实例分割网络架构不够简洁，参数量多，训练起来耗时费力，而轻量网络的实时分割精度欠佳；最后，实例分割基本上按照固定的模式设计网络，类似 PointRend[60] 和 PolarMask[58] 这样的新思路太少，且性能并不突出。

5 公开数据集与评估标准

本节主要介绍图像实例分割的常用数据集，总结了实例分割的性能评估标准，并根据该标准在 COCO 数据集[77]、Cityscapes 数据集[78]、PASCAL VOC 数据集[79] 和 SBD 数据集[80] 上对比了主流图像实例分割方法的分割精度。

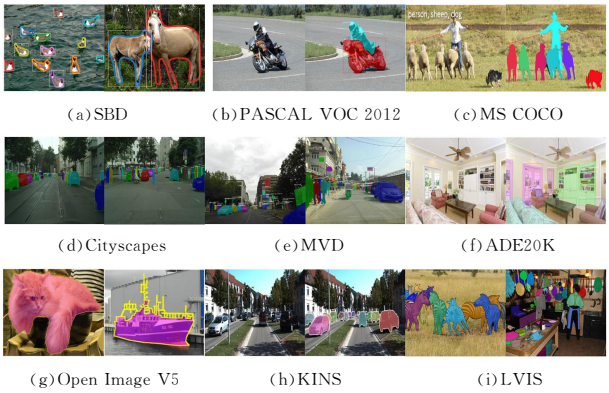
5.1 常用数据集

至今，基于深度卷积神经网络的图像实例分割方法已取得跨越性发展，除了卷积神经网络强大的特征表达能力和不断创新的实例分割模型，计算机视觉领域的相关数据集也起到了推动作用，表 1 列出了数据集汇总信息，图 12 为不同数据集的标注图像示意图。在大量标注训练集的监督训练、测试集的模型调参以及验证集的模型泛化能力评估中，实例分割模型不断得到改进。

表 1 常用实例分割数据集汇总

Table 1 Summary of commonly used instance segmentation datasets

Dataset name	Year	Description	Main application scenarios
SBD[80]	2011	使用实例级边界重新注释 Pascal VOC 2011 数据集中的 11355 幅图像，共 20 个对象类别，分为 5623 幅训练图像和 5732 幅测试图像	多场景
PASCAL VOC[79]	2012	含有 20 种类别，道路场景数据有 11520 张图片，包含 27450 个注释对象	道路行人车辆
Cityscapes[78]	2016	来自 50 个不同的城市街景记录的立体视频序列，将 30 个对象类划分为 8 个与城市场景相关的类别，数据集包含 5000 幅精细标注图像和 20000 幅粗略标注图像	道路、车辆、行人和街景
MS COCO[77]	2017	包含 200000 幅图像和 80 幅图像实例，共有 118000 幅训练图像、5000 幅验证图像和 41000 幅测试图像，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
MVD[81]	2017	来自世界各地在各种条件下捕获的图像，包含 25000 幅高分辨率的带注释的图像，分成 66 个类，其中有 37 个类别是特定的附加于实例的标签	道路行人车辆
ADE20K[82]	2017	拥有超过 25000 张图像，包括室内和室外场景中的 150 个类别，训练集和验证集分别包含 20210 幅和 2000 幅图像，测试集有 3000 幅图像	室内和室外场景
Open Images V5[83]	2018	数据集包含 280 万个物体实例的分割掩码，覆盖 350 个类别，在验证集和测试集上包含 99000 万个非常注重质量的手工标注掩码	多场景
KINS[84]	2019	来自 KITTI 数据集总共 14991 幅被注释的图像，其中 7474 幅图像用于训练，另外 7517 幅用于测试	道路行人车辆
LVIS[85]	2019	针对超过 1000 类物体进行了约 200 万个高质量的实例分割标注，包含 164000 幅图像	多场景



(a)SBD (b)PASCAL VOC 2012 (c)MS COCO

(d)Cityscapes (e)MVD (f)ADE20K

(g)Open Image V5 (h)KINS (i)LVIS

图 12 不同数据集的标注图像示意图

Fig. 12 Schematic diagram of annotated images of different data sets

5.2 性能评价标准

测试不同的实例分割模型，需要统一的性能评价标准，目前常用平均精度(Average Precision，AP)来衡量模型的性能。例如，针对某一分类的多个样本，假设它有 m 个正例，每一个正例对应一个召回率 R 值$(1/m,2/m,\cdots,1)$，对每一个召回率计算最大准确率 P，然后对这 m 个 P 值求均值，如式(1)所示：

$$AP=\frac{1}{m}\sum_{i=1}^{m}P_i=\int P(R)\mathrm{d}R \tag{1}$$

AP 针对某一个类，而一个数据集往往包含相当多的分类。假设有 C 类样本，则对数据集中所有类别的 AP 求均值就得到 mAP，如式(2)所示。一个数据集通常包含多类样本，用一类样本的平均准确率来衡量显然不合适，因此目前平均

精度 AP 默认为数据集的平均准确率。

$$mAP=\frac{1}{C}\sum_{j=1}^{C}AP_j \tag{2}$$

5.3 算法评估

本节在常用数据集上对当前主流实例分割模型的分割精度进行了对比和评估。其中 AP_{50} 是 IoU 阈值为 0.5 时 AP 的值，AP_{75} 是 IoU 阈值为 0.75 时 AP 的值。AP_S，AP_M 和 AP_L 分别是小、中、大 3 种不同尺度目标的 AP 值，短线表示数据不可用。

表 2 列出了当前主流的两阶段和单阶段图像实例分割模型的实验结果。由于两阶段基于分割的实例分割方法在速度和精度上缺乏竞争力，表 2 中未列出。早期的两阶段方法，如 MNC[38] 和 FCIS[26]，平均精度相对较低，自 2017 年 Mask RCNN 算法[27] 被提出，它就成为了实例分割领域的标杆，平均精度达到了 36.2%。目前每提出一种新的实例分割算法都要和 Mask RCNN[27] 进行对比，甚至最新的实例分割算法在平均精度上还低于它。当然，最近两年单阶段实例分割发展迅猛，许多新方法新思路相继被提出，打破了 Mask RCNN 框架的垄断地位。比如 YOLACT 算法[52] 虽然精度稍差，只有 31.2%，但是它强调的是实时性，速度快。PolarMask 算法[58] 也只有 32.1% 的精度，但它采用的是新型的轮廓建模方法，胜在思路新奇。TensorMask 算法[76] 采用 4D 张量，平均精度有所提高，但是模型比较复杂，没有压缩的设计，速度很慢。PointRend 算法[60] 也另辟蹊径，将渲染的思想融入实例分割，达到了 36.3% 的精度，并且生成的边界掩膜很精细，达到了抗锯齿化的效果。BlendMask 算法[61] 借鉴了 YOLACT 算法的思想，提出了高效的融合模块，模型精度达到了38.4%。SOLO 算法[65] 和 SOLOv2 算法[66] 利用位置信息，完全不依赖锚框，提出了实例类别的概念，并分别使用解耦头和动态卷积核压缩了张量，提高了分割效率和精度。CondInst 算法[64] 采用了以实例为条件的动态实例感知网络，显著地提高了分割精度，达到了 39.1%，是目前效果较好的实例分割算法。总体来看，单阶段实例分割方法效果更好。

表 2 MS COCO 数据集上实例分割模型的平均精度对比

Table 2 Comparison of average accuracy of instance segmentation models on MS COCO dataset

	Method	Year	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two stage	MNC[38]	2016	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
	FCIS[26]	2017	ResNet-101-C5	29.2	49.5	—	7.1	31.3	50.0
	Mask RCNN[27]	2017	ResNet-101-FPN	36.2	58.6	38.4	16.4	38.4	52.1
	PANet[28]	2018	ResNet-50-FPN	36.6	58.0	39.3	16.3	38.1	53.1
	MaskLab[37]	2018	ResNet-101-FPN	35.4	57.4	37.4	16.9	38.3	49.2
	HTC[35]	2019	ResNet-50-FPN	38.4	60.0	41.5	20.4	40.7	51.2
	MS RCNN[29]	2019	ResNet-101-FPN	37.5	58.7	40.2	17.2	39.5	53.0
Single stage	YOLACT[52]	2019	ResNet-101-FPN	31.2	50.6	32.8	12.1	33.3	47.1
	TensorMask[76]	2019	ResNet-101-FPN	37.1	59.3	39.4	17.4	39.1	51.6
	MEInst[59]	2020	ResNet-101-FPN	33.9	56.2	35.4	19.8	36.1	42.3
	PolarMask[58]	2020	ResNet-101-FPN	32.1	53.7	33.1	14.7	33.8	45.3
	PointRend[60]	2020	ResNet-50-FPN	36.3	—	—	—	—	—
	SOLO[65]	2020	ResNet-101-FPN	37.8	59.5	40.4	16.4	40.6	54.2
	SOLOv2[66]	2020	ResNet-101-FPN	39.7	60.7	42.9	17.3	42.9	57.4
	BlendMask[61]	2020	ResNet-101-FPN	38.4	60.7	41.3	18.2	41.5	53.3
	CondInst[64]	2020	ResNet-101-FPN	39.1	60.9	42.0	21.5	41.7	50.9
	CenterMask[63]	2020	ResNet-101-FPN	36.1	58.7	38.0	16.5	38.4	51.2

表 3 列出了 Cityscapes 数据集[78] 上部分实例分割模型的平均精度对比。其中 Cityscapes 验证子集的结果表示为 AP[val]，Cityscapes 测试子集的结果表示为 AP。早期的两阶段方法，如 BAIS[40] 和 SGN[46] 的分割精度较低，而 Mask RCNN[27] 的出现刷新了记录，PANet[28] 由于缩短高低层特征融合的路径，实现了较高的分割精度。Deep snake[43] 和 BMask RCNN[42] 基于边界编码进行实例分割，关注目标边缘的分割精度，性能也有较大提升。

表 3 Cityscapes 数据集上实例分割模型的平均精度对比

Table 3 Comparison of average accuracy of instance segmentation models on Cityscapes dataset

Method	Year	AP[val]	AP	AP_{50}	person	rider	car	truck	bus	train	mcycle	bicycle
BAIS[40]	2017	—	17.4	36.7	—	—	—	—	—	—	—	—
SGN[46]	2017	29.2	25.0	44.9	21.8	20.1	39.4	24.8	33.2	30.8	17.7	12.4
InstanceCut[45]	2017	15.8	13.0	27.9	10.0	8.0	23.7	14.0	19.5	15.2	9.3	4.7
Mask RCNN[27]	2017	36.4	32.0	58.1	34.8	27.0	49.1	30.1	40.9	30.9	24.1	18.7
PANet[28]	2018	41.4	36.4	63.1	41.5	33.6	58.2	31.8	45.3	28.7	28.2	24.1
Deep snake[43]	2020	37.4	31.7	58.4	37.2	27.0	56.0	29.5	40.5	28.2	19.0	16.4
PointRend[60]	2020	—	35.8	—	—	—	—	—	—	—	—	—
BMask RCNN[42]	2020	35.0	29.4	54.7	34.3	25.6	52.6	24.2	35.1	24.5	21.4	17.1

表 4 列出了部分实例分割模型在 PASCAL VOC 2012[79] 数据集和 SBD[80] 数据集上的平均精度对比。其中 mAP^r_{50} 和 mAP^r_{70} 是 IoU 阈值分别为 0.5 和 0.7 时数据集的平均精度。

表4 PASCAL VOC 2012和SBD数据集上实例分割模型的平均精度对比

Table 4 Comparison of average accuracy of instance segmentation models on PASCAL VOC 2012 and SBD datasets

Method	Year	Datasets	mAP^r_{50}	mAP^r_{70}
SDS[1]	2014	PASCAL VOC 2012	49.7	25.3
MNC[38]	2016	PASCAL VOC 2012	59.1	36.0
		SBD	63.5	41.5
InstanceFCN[25]	2016	PASCAL VOC 2012	61.5	43.0
FCIS[26]	2017	PASCAL VOC 2012	65.7	52.1
		SBD	65.7	52.1
Mask RCNN[27]	2017	PASCAL VOC 2012	68.5	40.2
ESE-Seg[41]	2019	PASCAL VOC 2012	69.3	36.7
		SBD	40.7	12.1
YOLACT-500[52]	2019	SBD	72.3	56.2
Deep snake[43]	2020	SBD	62.1	48.3

6 问题与解决方案

随着深度卷积神经网络的应用，图像实例分割已经取得显著性成果，但是当前仍存在诸多挑战，如目标尺度变化多样、目标边缘分割精度不高、显存消耗过大和数据集标注代价昂贵等，本节将详细阐述此类问题并提出可能的解决思路。

(1)目标尺度变化多样：图像中通常含有不同类别的视觉要素，它们之间往往存在尺度差异。此外，CNN在特征提取阶段会存在若干池化操作，小目标在池化过程中会造成信息丢失，即使通过上采样恢复空间分辨率也无法找回底层信息。因此，解决好目标尺度问题才能更好地提高分割精度。

首先，可参考SNIP[86]、SNIPER[87]、SSD[11]、空洞卷积和FPN[36]等多尺度图像处理方法；其次，可利用CNN不同阶段的特征图，通过上下采样等方法将特征图插值到相同分辨率再进行特征融合，或者使用空间金字塔池化，利用不同系数的池化层生成对应多个尺度的分支，也可使用多尺度注意力机制融合多尺度特征；最后，训练阶段使用过采样和复制粘贴的策略增强小目标的数量，也可有效提升小目标的分割性能。

(2)目标边缘分割精度不高：相邻像素对应感受野内的图像信息过于相似，如果临近的像素都属于所需分割区域的内部，这种“相似”是有利的；但是，如果相邻像素刚好处在所需分割区域的边界上，那么这种“相似”就会影响目标边缘的分割精度。此外，目标间的相互遮挡也影响了目标边缘的分割精度。

目前有几种可能的解决思路：1)对网络输出的分割边界增加额外的损失[89]；2)让网络对边界的特征和区域内部的特征分开建模学习，其本质思想还是让网络同时做两个任务——分割和边缘检测；3)提高输入图像的输入分辨率和中间层特征图的分辨率也可能提高边缘分割精度；4)参考合理的上下文建模机制，帮助网络预测遮挡部分的语义信息，解决目标边缘遮挡问题。

(3)显存消耗过大：使用一个12GB的GPU训练ResNet-101分类模型，batch_size(一次训练所选取的样本数)能设置到32左右。但是要训练一个以ResNet-101为基础网络的分割模型，batch_size只能设置为4左右，甚至更低，这样会导致模型收敛速度非常慢。因为分割不仅需要高层的语义信息，同时也需要高分辨率的特征图来恢复图像的细节特征。如果用一张卡训练类似COCO的大型数据集，可能需要两周。现实情况是，不是每个应用场景的硬件设备都有这么大的显存，也不是每个研究组都有服务器集群可供训练模型。

因此，轻量化网络架构可以有效解决显存消耗问题，它的核心是在尽量保持精度的前提下，从体积和速度两方面对网络进行轻量化改造。随着移动端、嵌入式设备的普及和使用，架构简单、计算量少、速度快且能满足准确率要求的轻量化网络将是未来实例分割的发展方向。

(4)数据集标注代价昂贵：图像分割数据集对于实例分割网络的模型训练非常重要，但是当前数据集的标注是耗时费力的。想要训练一个成熟的分割模型，至少需要上万张的训练图像，而且如果实际应用场景不一样，为了达到更好的效果，每个场景都需要标注。如果外包给数据集标注公司，时间和成本也是一个问题。此外，如果标注规则定得不太合理，标注的数据集没有训练出来效果好的模型，还需要返工。

目前，数据集问题可以通过减少训练样本来解决，利用小样本、半监督或者弱监督学习的方法来减少昂贵的数据标注，这些方法也是实例分割未来的发展方向。

结束语 图像实例分割是计算机视觉领域具有挑战性的研究方向，当前按照图像分割的过程和特征，可将图像实例分割方法分为两阶段和单阶段，其中两阶段方法效果不错但是速度较慢，因此近年来单阶段实例分割方法成为主流，其速度快、精度高，有些模型已经满足实时实例分割的要求。但是目前仍存在目标尺度变化多样、目标边缘分割精度不高、显存消耗过大和数据集标注代价昂贵等问题。接下来对实例分割的未来发展方向进行几点展望：1)当前实例分割网络结构复杂，运行时间长，未来应该考虑轻量化网络架构与实际工程应用相结合，注重技术落地和场景推广；2)单阶段无锚框的实例分割方法是未来发展的主流，特别是利用位置信息进行实例分割的方法彻底摆脱了锚框和回归框的限制，未来还有很大的发展空间；3)使用多尺度特征提取和融合，以及全局特征和局部特征的信息融合让网络学习到更好的特征值得进一步研究；4)最近，同时利用CNN的特征提取能力、transformer的内容和位置自注意力机制的混合方式在实例分割任务上取得了显著的性能提升，同时减少了参数量，未来应该多关注transformer在实例分割中的应用；5)为了丰富图像分割任务，实现高维度场景理解，全景分割、视频实例分割跟踪和3D点云实例分割将成为未来的研究热点。

参 考 文 献

[1] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]// Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2014: 297-312.

[2] FAN W, LIU T, HUANG R, et al. Low-level CNN Feature Aided Image Instance Segmentation[J]. Computer Science, 2020, 47(11): 186-191.

[3] TANG Q K, HU Y. Single stage target detection algorithm based on positive and negative anchor frame equalization and feature alignment [J]. Chinese Journal of Computer Aided Design and Graphics, 2020, 32(11): 1773-1783.

[4] ZHOU P C, GONG S R, ZHONG S, et al. Image Semantic Segmentation Based on Deep Feature Fusion[J]. Computer Science, 2020, 47(2): 126-134.

[5] MOU L, ZHU X X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(11): 6699-6711.

[6] XU Z, LIU S, SHI J, et al. Outdoor RGBD instance segmentation with residual regretting learning[J]. IEEE Transactions on Image Processing, 2020, 29: 5301-5309.

[7] MANINIS K K, CAELLES S, CHEN Y, et al. Video object segmentation without temporal information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1515-1530.

[8] XU Y, LI Y, WANG Y, et al. Gland instance segmentation using deep multichannel neural networks[J]. IEEE Transactions on Biomedical Engineering, 2017, 64(12): 2901-2912.

[9] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788.

[11] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]// Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 21-37.

[12] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]// Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241.

[13] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(12): 2481-2495.

[14] ZHANG Z, SCHWING A G, FIDLER S, et al. Monocular object instance segmentation and depth ordering with cnns[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2614-2622.

[15] ROMERAPAREDES B, TORR P H S. Recurrent instance segmentation[C]// Proceedings of the 2016 European Conference on Computer Vision. Heidelberg: Springer, 2016: 312-329.

[16] REN M, ZEMEL R S. End-to-end instance segmentation with recurrent attention[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 293-301.

[17] UHRIG J, CORDTS M, FRANKE U, et al. Pixel-level encoding and depth layering for instance-level semantic labeling[C]// Proceedings of the German Conference on Pattern Recognition. Switzerland: Springer, 2016: 14-25.

[18] ZHANG S, GONG Y H, WANG J J. Development of deep convolutional neural network and its application in computer vision [J]. Chinese Journal of Computers, 2019, 42(3): 453-482.

[19] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[20] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61: 85-117.

[21] GOODFELLOW I, BENGIO Y, COURVILLE A, et al. Deep learning[M]. Cambridge: MIT press, 2016.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.

[23] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778.

[24] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4700-4708.

[25] DAI J, HE K, LI Y, et al. Instance-sensitive fully convolutional networks[C]// Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 534-549.

[26] LI Y, QI H, DAI J, et al. Fully convolutional instance-aware semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2359-2367.

[27] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2980-2988.

[28] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8759-8768.

[29] HUANG Z, HUANG L, GONG Y, et al. Mask scoring r-cnn [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 6409-6418.

[30] PINHEIRO P O, COLLOBERT R, DOLLÁR P. Learning to segment object candidates[J]. Advances in Neural Information Processing Systems, 2015, 28: 1990-1998.

[31] PINHEIRO P O, LIN T Y, COLLOBERT R, et al. Learning to refine object segments[C]// European Conference on Computer Vision. Heidelberg: Springer, 2016: 75-91.

[32] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440.

[33] VAN DE SANDE K E A, UIJLINGS J R R, GEVERS T, et al. Segmentation as selective search for object recognition[C]// Proceedings of the 2011 International Conference on Computer Vision. Spain: IEEE, 2011: 1879-1886.

[34] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks

[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[35] CHEN K,PANG J,WANG J,et al. Hybrid task cascade for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press,2019:4974-4983.

[36] LIN T Y,DOLLÁR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2017:936-944.

[37] CHEN L C,HERMANS A,PAPANDREOU G,et al. Masklab: Instance segmentation by refining object detection with semantic and direction features[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press,2018:4013-4022.

[38] DAI J,HE K,SUN J. Instance-aware semantic segmentation via multi-task network cascades[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2016:3150-3158.

[39] WEN Y, HU F, REN J, et al. Joint multi-task cascade for instance segmentation[J]. Journal of Real-Time Image Processing,2020,17(6):1983-1989.

[40] HAYDER Z,HE X,SALZMANN M. Boundary-aware instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press,2017:5696-5704.

[41] XU W,WANG H,QI F,et al. Explicit shape encoding for real-time instance segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press,2019:5168-5177.

[42] CHENG T,WANG X,HUANG L,et al. Boundary-preserving mask R-CNN[C]// Proceedings of the European Conference on Computer Vision. Heidelberg:Springer,2020:660-676.

[43] PENG S,JIANG W,PI H,et al. Deep snake for real-Time instance segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2020:8533-8542.

[44] KASS M,WITKIN A,TERZOPOULOS D. Snakes: Active contour models[J]. International Journal of Computer Vision, 1988,1(4):321-331.

[45] KIRILLOV A,LEVINKOV E,ANDRES B,et al. InstanceCut: from edges to instances with multicut[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2017:5008-5017.

[46] LIU S,JIA J,FIDLER S,et al. Sgn: Sequential grouping networks for instance segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press,2017:3496-3504.

[47] DE BRABANDERE B,NEVEN D,VAN GOOL L. Semantic instance segmentation with a discriminative loss function[EB/OL]. [2020-09-10]. https://arxiv.org/abs/1708.02551.

[48] FATHI A,WOJNA Z,RATHOD V,et al. Semantic instance segmentation via deep metric learning[EB/OL]. [2020-09-10]. https://arxiv.org/abs/1703.10277.

[49] KONG S,FOWLKES C C. Recurrent pixel embedding for instance grouping[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press,2018:9018-9028.

[50] BAI M,URTASUN R. Deep watershed transform for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press,2017:5221-5229.

[51] GAO N,SHAN Y,WANG Y,et al. Ssap: Single-shot instance segmentation with affinity pyramid[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos:IEEE Computer Society Press,2019:642-651.

[52] BOLYA D,ZHOU C,XIAO F,et al. Yolact: Real-time instance segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos:IEEE Computer Society Press,2019:9157-9166.

[53] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos:IEEE Computer Society Press,2017:2980-2988.

[54] BOLYA D,ZHOU C,XIAO F,et al. Yolact++: Better real-time instance segmentation[EB/OL]. [2020-09-10]. https://arxiv.org/abs/1912.06218.

[55] LIU H,SOTO R A R,XIAO F,et al. YolactEdge: Real-time Instance Segmentation on the Edge(Jetson AGX Xavier: 30 FPS, RTX 2080 Ti: 170 FPS)[EB/OL]. [2020-09-10]. https://arxiv.org/abs/2012.12259.

[56] DUAN K W,BAI S,XIE L X,et al. CenterNet: object detection with keypoint triplets[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press,2019:6569-6578.

[57] TIAN Z,SHEN C,CHEN H,et al. Fcos: Fully convolutional one-stage object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press,2019:9627-9636.

[58] XIE E,SUN P,SONG X,et al. Polarmask: Single shot instance segmentation with polar representation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 12193-12202.

[59] ZHANG R,TIAN Z,SHEN C,et al. Mask encoding for single shot instance segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2020:10226-10235.

[60] KIRILLOV A,WU Y,HE K,et al. Pointrend: Image segmentation as rendering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2020:9799-9808.

[61] CHEN H,SUN K,TIAN Z,et al. BlendMask: Top-down meets bottom-up for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos:IEEE Computer Society Press,2020:8573-8581.

[62] YING H,HUANG Z,LIU S,et al. Embedmask: Embedding coupling for one-stage instance segmentation[EB/OL]. [2020-09-10]. https://arxiv.org/abs/1912.01954.

[63] WANG Y, XU Z, SHEN H, et al. CenterMask: single shot instance segmentation with point representation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9313-9321.

[64] TIAN Z, SHEN C, CHEN H. Conditional Convolutions for Instance Segmentation[EB/OL]. [2020-09-10]. https://arxiv.org/abs/2003.05664.

[65] WANG X, KONG T, SHEN C, et al. Solo: Segmenting objects by locations[C]// Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 649-665.

[66] WANG X, ZHANG R, KONG T, et al. SOLOv2: Dynamic and fast instance segmentation[J]. Advances in Neural Information Processing Systems, 2020, 33: 17721-17732.

[67] LAW H, DENG J. CornerNet: Detecting Objects as Paired Keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[68] ZHOU X, ZHUO J, KRAHENBUHL P. Bottom-up object detection by grouping extreme and center points[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 850-859.

[69] ZHU C, HE Y, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 840-849.

[70] KONG T, SUN F, LIU H, et al. Foveabox: beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[71] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[72] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE, 2005: 886-893.

[73] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 764-773.

[74] WANG Z Y, YUAN C, LI J C. Instance segmentation using separable convolution and multilevel features [J]. Journal of Software, 2019, 30(4): 954-961.

[75] CHEN C, QI F. Review on Development of Convolutional Neural Network and Its Application in Computer Vision[J]. Computer Science, 2019, 46(3): 63-73.

[76] CHEN X, GIRSHICK R, HE K, et al. Tensormask: A foundation for dense object segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 2061-2069.

[77] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]// Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755.

[78] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 3213-3223.

[79] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The PASCAL visual object classes(VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[80] HARIHARAN B, ARBELÁEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]// Proceedings of the 2011 IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 991-998.

[81] NEUHOLD G, OLLMANN T, ROTA BULO S, et al. The mapillary vistas dataset for semantic understanding of street scenes [C]// Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4990-4999.

[82] ZHOU B, ZHAO H, FERNANDEZ F X P, et al. Scene parsing through ade20k dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 633-641.

[83] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[EB/OL]. [2020-09-10]. https://arxiv.org/abs/1811.00982.

[84] QI L, JIANG L, LIU S, et al. Amodal instance segmentation with kins dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3014-3023.

[85] GUPTA A, DOLLAR P, GIRSHICK R. Lvis: A dataset for large vocabulary instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 5356-5364.

[86] SINGH B, DAVIS L S. An analysis of scale invariance in object detection snip[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3578-3587.

[87] SINGH B, NAJIBI M, DAVIS L S. SNIPER: efficient multi-scale training[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 9333-9343.

HU Fu-yuan, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning and computer vision.

WAN Xin-jun, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include deep learning and computer vision.

(责任编辑:喻藜)