DOI: 10.11992/tis.202109043
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20211220.1100.004.html.

基于深度学习的实例分割研究综述

苏丽[1,2], 孙雨鑫[1], 苑守正[1]
（1.哈尔滨工程大学 智能科学与工程学院, 黑龙江 哈尔滨 150001; 2.哈尔滨工程大学 船舶装备智能化技术与应用教育部重点实验室, 黑龙江 哈尔滨 150001）

摘　要: 深度学习在计算机视觉领域已经取得很大发展, 虽然基于深度学习的实例分割研究近年来才成为研究热点, 但其技术可广泛应用在自动驾驶, 辅助医疗和遥感影像等领域。实例分割作为计算机视觉的基础问题之一, 不仅需要对不同类别目标进行像素级别分割, 还要对不同目标进行区分。此外, 目标形状的灵活性, 不同目标间的遮挡和繁琐的数据标注问题都使实例分割任务面临极大的挑战。本文对实例分割中一些具有价值的研究成果按照两阶段和单阶段两部分进行了系统性的总结, 分析了不同算法的优缺点并对比了模型在 COCO 数据集上的测试性能, 归纳了实例分割在特殊条件下的应用, 简要介绍了常用数据集和评价指标。最后, 对实例分割未来可能的发展方向及其面临的挑战进行了展望。

关键词: 计算机视觉; 实例分割; 图像分割; 卷积神经网络; 深度学习; 目标检测; 两阶段实例分割; 单阶段实例分割
中图分类号: TP183　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)01−0016−16

中文引用格式: 苏丽, 孙雨鑫, 苑守正. 基于深度学习的实例分割研究综述 [J]. 智能系统学报, 2022, 17(1): 16–31.
英文引用格式: SU Li, SUN Yuxin, YUAN Shouzheng. A survey of instance segmentation research based on deep learning[J]. CAAI transactions on intelligent systems, 2022, 17(1): 16–31.

A survey of instance segmentation research based on deep learning

SU Li[1,2], SUN Yuxin[1], YUAN Shouzheng[1]
(1. College of Intelligent Science and Engineering, Harbin Engineering University, Harbin 150001, China; 2. Key Laboratory of Ministry of Education on Intelligent Technology and Application of Marine Equipment, Harbin Engineering University, Harbin 150001, China)

Abstract: Deep learning has made great progress in the field of computer vision. Although instance segmentation research based on deep learning has only become a research hotpot in recent years, relevant techniques can be widely used in the fields of autonomous driving, complementary medicine and remote sensing imaging. Instance segmentation, as one of the fundamental problems of computer vision, requires not only pixel-level segmentation of different classes of targets, but also differentiation of different targets. In addition, the flexibility of target shapes, the occlusion between different targets and the tedious data annotation problems all make the instance segmentation task extremely challenging. In this paper, firstly, some valuable research results in instance segmentation are systematically reviewed according to two-stage instance segmentation and one-stage instance segmentation. Secondly, the advantages and disadvantages of different algorithms are analyzed and the testing performance of different models on the COCO dataset is compared. In addition, the applications of instance segmentation under special conditions are summarized, and common datasets and evaluation metrics are briefly introduced. Finally, the possible future directions of instance segmentation and the challenges it faces are prospected.

Keywords: computer vision; instance segmentation; image segmentation; convolutional neural network; deep learning; object detection; two-stage instance segmentation; one-stage instance segmentation

近年来, 深度学习和统一计算设备构架 (compute unified device architecture, CUDA) 等并行计算技术迅速发展直接推动了计算机视觉和图像处理领域进入了新的技术时代, 实例分割作为计算机视觉基础研究问题之一, 其技术可广泛应用于汽

收稿日期: 2021−09−30.　**网络出版日期**: 2021−12−21.
基金项目: 国家重点研发计划项目（2018YFB1601502）; 国际合作项目（MC-201920-X01）.
通信作者: 孙雨鑫. E-mail: heu_syx@hrbeu.edu.cn.

车自动驾驶, 机器人控制, 辅助医疗和遥感影像等领域。

在计算机视觉的基本任务中目标检测是预测图像中目标位置和类别。语义分割则是在像素级别上对目标分类。而实例分割可看作目标检测和语义分割的结合体, 旨在检测图像中所有目标实例, 并针对每个实例标记属于该类别的像素。即不仅需要对不同类别目标进行像素级别分割, 还要对不同目标进行区分。与其他计算机视觉研究问题相比, 实例分割的挑战性在于:

1) 需要预测并区分图像中每个目标的位置和语义掩码, 并且由于实例的不可知形状使得预测实例分割的掩码比目标检测任务预测矩形边界框更灵活;

2) 密集目标的相互遮挡与重叠使网络很难有效区分不同实例, 并且小目标的实例分割由于普遍缺少细节导致分割精度仍然很低;

3) 繁琐精细的数据标注耗费大量人力与时间, 如何减少成本, 有效利用现有未标注或粗糙标注的数据提升实例分割精度仍是一个亟待解决的问题。

1980 年日本学者福岛邦彦[1] 提出的神经认知机模型可以称为卷积神经网络的前身, Lecun[2] 提出反向传播算法使网络训练成为可能, 之后 2012 年 AlexNet[3] 在 ImageNet 图像识别大赛上获得冠军。从此深度卷积神经网络引起人们关注, 研究者用它解决计算机视觉任务。近年来, 实例分割的研究基本是建立在基于卷积神经网络的目标检测和语义分割基础之上。因此, 从研究发展来看实例分割任务是卷积神经网络成功运用在计算机视觉领域的产物[4]。实例分割方法主要归纳为两阶段与单阶段两类, 其中两阶段实例分割有两种解决思路, 分别是自上而下基于检测的方法和自下而上基于分割的方法。而单阶段实例分割可细化为感知实例分割, 建模掩码, Transformer 嵌入及一些其他方法。

本文从实例分割的研究现状, 算法优缺点和主流方法性能对比, 特殊条件下的应用, 常用数据集与权威评价指标等角度出发对一些具有启发性的研究成果进行整理, 归纳和分析, 为相关研究提供有价值的参考。

1　实例分割研究现状

从研究时间线来看, 实例分割技术根据处理过程目前主要归纳为两类: 两阶段和单阶段, 如图 1 所示, 本文将分别进行介绍。

1.1　两阶段的实例分割

两阶段实例分割是以处理阶段划分, 其中自上而下的基于检测方法是先检测出图像中实例所在区域, 再对候选区域进行像素级别分割。而自下而上的基于分割思想则将实例分割看作一个聚类任务, 通过将像素分组为图像中呈现的任意数量的目标实例, 最后判断每组的类别来生成实例掩码, 这种不需要束缚于目标框的影响。

1.1.1　自上而下的实例分割

自上而下的实例分割研究受益于目标检测的丰硕成果。下面介绍一下代表性的方法。

2014 年 Hariharan 等[5] 在 SDS 中首次实现检测和分割同时进行, 也是最早的实例分割算法, 奠定了后续研究基础[6]。如图 2 所示, 具体分为 4 步: 1) 建议框生成, 使用非极大值抑制 (non-maximum suppession, NMS)[7] 为每张图片产生 2000 个候选区域; 2) 特征提取, 联合训练两个不同的卷积神经网络 (convolutional neural network, CNN) 网络同时提取候选区域和区域前景特征; 3) 区域分类, 利用 CNN 中提取到的特征训练 SVM 分类器对上述区域进行分类; 4) 区域细化, 采用 NMS 来剔除多余区域, 最后使用 CNN 中的特征来生成特定类别的粗略掩码预测, 以细化候选区域将该掩码与原始候选区域结合起来可以进一步高分割效果。

虽然 SDS 效果逊色后续方法, 但 SDS 先用检测生成候选区域再对其语义分割的思想为后续实例分割提供了重要的研究启发。2015 年该团队又对 SDS 重新分析认为, 只使用 CNN 最高层的特征来解决实例分割问题存在着掩码细节粗糙的缺陷, 即高层特征的语义信息丰富有利于目标分类, 但缺少精确的位置信息。例如: 在底层特征图中可以定位目标部件, 但是没有丰富语义信息判别区分这个目标部件具体属于哪个物体。所以, 引入 Hypercolumns[8](所有 CNN 层对应该像素位置的激活输出值所组成的向量) 作为特征描述符, 将底层特征与高层特征融合从而提升分类的精确性并改善目标分割细节。

之后, CFM[9] 算法首次将掩码(mask)这一概念引入到实例分割中。CFM 通过矩形框生成特征图的掩码, 并将任意区域生成固定大小的特征以方便处理。这里是从卷积特征中提取掩码而非原始图像中提取。

DeepMask[10] 是首个直接从原始图像数据学习产生分割候选的工作。简单讲, 给定一个图片块作为输入, 输出一个与类别无关的 mask 和相应

的分数。它最大的特点是不依赖于边缘、超像素或者其他任何辅助形式的分割，是用分割的方法来生成高召回率的候选区域。但缺点是只能捕捉目标大致外形，不能准确描绘目标边界。为了优化 DeepMask 的掩码，SharpMask[11] 先在前向反馈通道中生成粗略的掩码，并在自上而下的通道中引入较低层次富有位置的特征逐步加以细化，最后产生具有更高保真度的能精确框定物体边界的掩码。

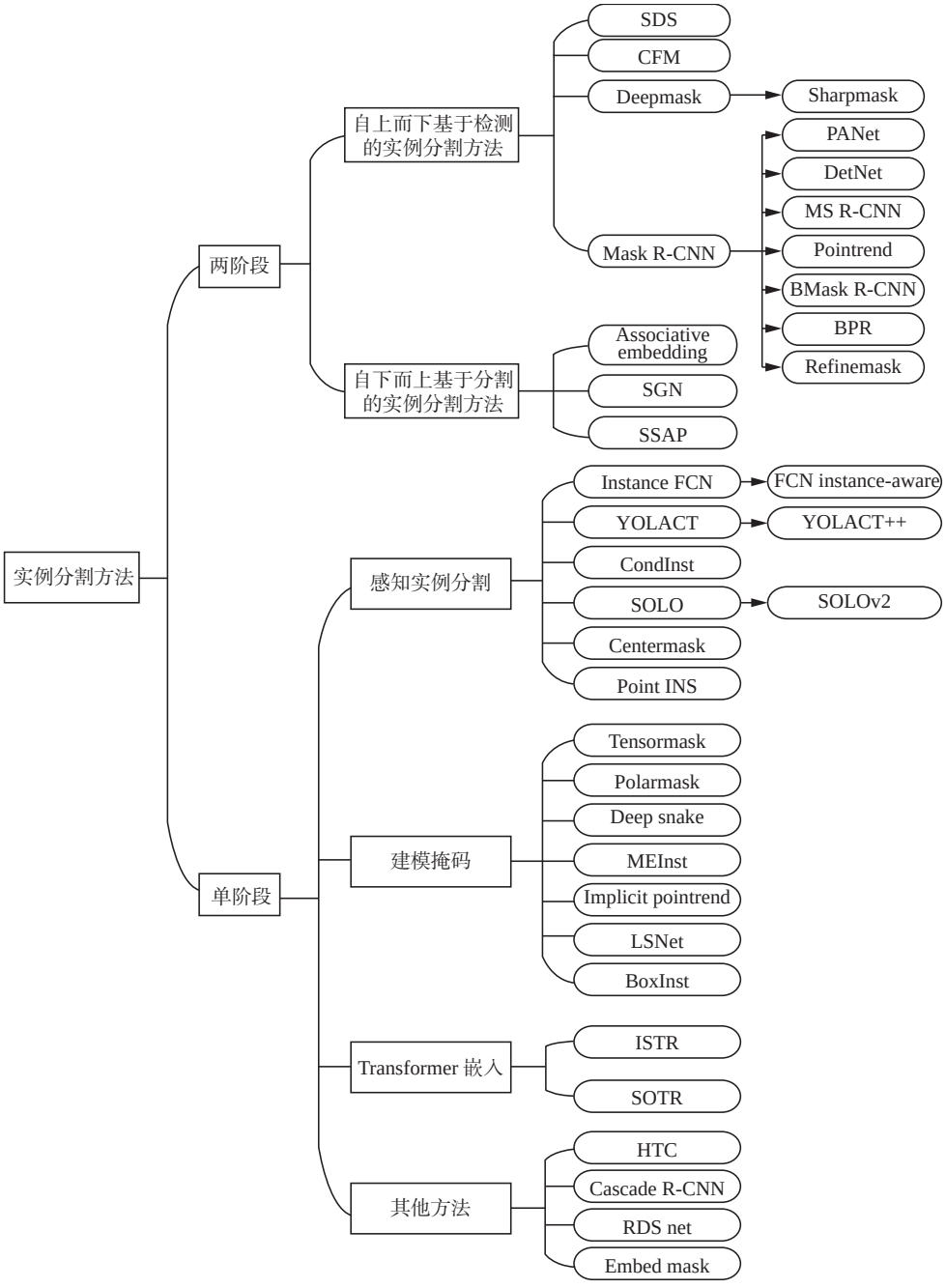

图 1　本文涉及的实例分割方法

Fig. 1　Paper focuses on the instance segmentation methods

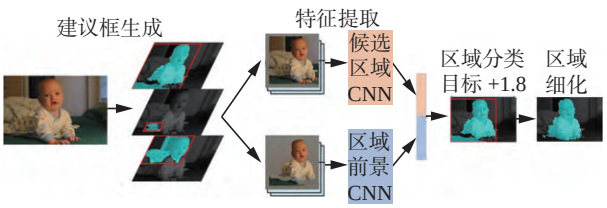

图 2　SDS 网络

Fig. 2　SDS network

但是上面提到的方法都需要先在原图生成掩膜候选区域，没有充分利用深度学习特征及大规模训练数据的优势并且推断时间缓慢，这些都是影响实例分割准确性的瓶颈。2016 年，何凯明团队在多任务网络级联 (MNC)[12] 中提出了一种级联结构，如图 3 中将实例分割任务分解为目标定位、掩码生成以及目标分类 3 个子任务，共用一

个主干网络, 将 3 个不同功能的网络分支级联起来。每个阶段都以前一阶段的结果作为输入。整个网络是端到端的。这样主干网络的训练可以共享 3 个子任务的监督, 有利于训练出更好的特征。这种设计另一个优点是可以快速地进行推断。

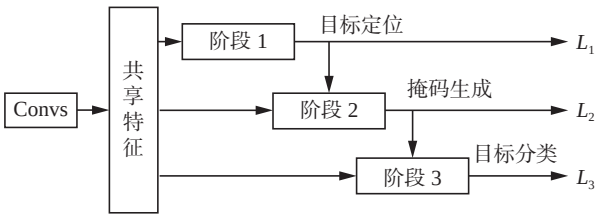

图 3　MNC 网络
Fig. 3　MNC network

随着计算机并行处理数据能力的提升和目标检测网络性能的快速更新, 实例分割研究趋势打开了一个新的局面。前沿的设计思想和领域的认识革新碰撞出新的学术火花。

2017 年何凯明团队提出简单通用且性能强大的两阶段 Mask R-CNN[13], 是 Faster R-CNN[14] 思想应用在实例分割的经典之作, 用于许多衍生应用的基线算法, 也是现今使用最多, 效率最高的实例分割算法。它的成功又激起实例分割领域新的技术浪潮。Mask R-CNN[13] 在目标分类和回归分支上增加了用于预测每个感兴趣区域 (region of interest, ROI) 的语义分割分支。网络结构如图 4 所示, 基础网络中采用了当时较为优秀的 ResNet-FPN[15-16] 结构, 多层特征图有利于多尺度物体及小物体的检测。首先, 将输入图片送入到特征提取网络得到特征图, 然后对特征图的每一个像素位置设定固定个数的 ROI(也可以称为锚框), 然后将 ROI 区域送入区域推荐网络 (region proposal network, RPN) 进行二分类 (前景和背景) 以及坐标回归, 以获得修正后的 ROI 区域。为了保证特征分辨率, 对 ROI 执行提出的 ROI Align[13] 操作替换原始的 ROI Pooling[14], 取消了取整操作, 而是通过双线性插值的方法保留所有的浮点数。最后增加了一个 mask 掩码分支来预测每一个像素的类别。采用了全卷积神经网络 (fully convolutional network, FCN)[17] 的网络结构, 利用卷积与反卷积构建端到端的网络, 对每一个像素分类, 实现了较好的分割效果。同时, 2018 年 Masklab[18] 也改进了 Faster R-CNN[14], 并产生两个额外的输出, 即语义分割和实例中心方向。由于 Mask R-CNN 对实例分割研究具有重要的启发意义, 后续涌现了一系列相关的工作, 具体方法如下。

2018 年 PANet[19] 在 Mask R-CNN 基础上引入自下而上的路径改进并扩展了金字塔特征提取网络, 使用自适应融合的 ROI 区域特征池化, 很好地融合了不同层次的特征信息。DetNet[20] 将空洞卷积加到骨干结构中既保证了特征分辨率同时又增大感受野, 并提出重新对检测、分割任务训练骨干网络以提高特征表达能力。

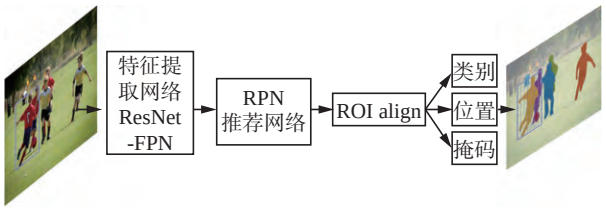

图 4　Mask R-CNN 网络
Fig. 4　Mask R-CNN network

2019 年 MS R-CNN[21] 提出现有的 mask 打分策略是使用分类的指标, 缺乏针对性的评价机制。故在 Mask R-CNN 基础上修改了 mask 评价标准, 通过添加 Mask IOU 分支来预测 mask 并且给其打分来提升模型实例分割性能。同年, 何凯明团队提出 PointRend[22] 将实例分割看作图像处理中渲染问题, 细化 Mask R-CNN 产生的粗糙掩码边缘, 先在边缘上选几个点再提取点的特征进行迭代计分计算达到细化掩码的目的。

2020 年 BMask R-CNN[23] 则将目标边缘信息加入 Mask R-CNN 中用于监督网络以增强掩码预测。

2021 年 BPR[24] 提出一个后处理细化模块以提高 Mask R-CNN 的边界质量。RefineMask[25] 利用边缘信息和语义分割信息细化 Mask R-CNN 生成的粗糙掩码边缘。姜世浩等[26] 在 Mask R-CNN 基础上引入两条分支, 基于整体嵌套边缘检测 2021 年 BPR[24] 提出一个后处理细化模块以提高 Mask R-CNN 的边界质量。BPR 利用一种裁剪细化的策略, 先通过实例分割网络 (如 Mask R-CNN) 得到粗糙的掩码。随后在掩码的边界上提取出一系列的方块, 这些方块被送入一个细化网络作二分类的前景与背景分割, 进而实现对边界处的方块进行优化。该网络可以解决 Mask R-CNN 预测的掩码存在边界粗糙的问题。RefineMask[25] 则利用边缘信息和语义分割信息细化 Mask R-CNN 生成的粗糙掩码边缘。通过多阶段的方式在实例分割过程中逐级融合更多的细粒度信息, 因此逐步精细化了实例掩模。最后, RefineMask 成功地克服了以往分割中所遇到的困难案例 (如物体的弯曲部分被过度平滑), 并输出了准确的边界。模型生成边缘特征图, 一条基于 FCN 生成偏重于空间位置信息的语义特征图。融合以上得到的多个特征图, 生成信息更加丰富的新特征。

但是上述自上而下的实例分割方法缺点在于:

1) 在一定程度上严重依赖精确的目标检测且得到的实例掩码分辨率相对较低；

2) 对于多实例的复杂场景，由于两阶段方法在前期需要单独设计网络生成大量建议区域，其推理时间与建议框的数量成正比，因此在推断速度上缓慢；

3) 仍然无法很好地区分同一类别重叠的不同实例个体且掩码分割细节不够平滑。

1.1.2　自下而上的实例分割

为了摆脱目标检测边界框对后续分割的限制，研究者们从另一个角度思考实例分割问题，将实例分割看作是一个图像聚类任务。也就是需要将图像中属于一个物体的所有像素聚成一个集合，并判断这个物体的类别。这种基于分割的方法通常会学习经过特殊设计的转换形式或实例边界，并以类似嵌入的方式将点聚类到实例掩码中。下面介绍几种代表方法。

BAI M 等[27] 利用 FCN 网络来学习分水岭变换的能量，然后利用能量分割，将图像分成若干个区域，每个区域就代表了一个实例。Associative embedding[28] 利用学习到的关系嵌入成组来分配像素。Brabandere 等[29] 引入判别损失函数通过推开属于不同实例的像素并拉近同一实例中的像素来有效地学习像素级别的实例嵌入。SGN[30] 使用序列组合网络将实例分割问题分解成一系列子类分组问题。每个网络都解决了语义复杂度不断提高的子分组问题，以便逐步从像素中组成对象。Gao 等[31] 学习像素对亲和力金字塔，即两个像素属于同一实例的概率，并通过级联图分区顺序生成实例。Fathi 等[32] 和 Brabandere 等[33] 把问题分解成为逐像素语义分割，逐像素对应实例的坐标进行预测和区分类别的实例个数。同时，尝试利用特征嵌入的方式，为每一个像素学习一个特征，并根据特征的距离对像素进行聚类。

这类方法通过将像素分组为图像中呈现的任意数量的对象实例来生成实例掩码，与自上而下的方法相比，自下而上方法的缺点是：

1) 严重依赖于密集的预测质量，导致性能不够标准或产生碎片掩码；

2) 由于聚类过程使得很难将其应用于复杂的情况，通常在准确性上落后。尤其是在具有不同复杂场景和语义类别较多的数据集上泛化能力有限。

3) 预测之后的处理技术很复杂。

综上分析两类方法，自上而下严重依赖目标检测效果，自下而上虽然天然克服了基于建议框的缺陷，转为对每个像素的嵌入学习和分组进行处理，但一般无法端到端训练，且受限于聚类算法，性能一般有限。那么，是否存在一种方法，可以绕过这些条条框框来直接作实例分割呢？

1.2　单阶段的实例分割

受单阶段的目标检测启发，现有方法将实例分割统一到 FCN[17] 框架下，如以单阶段全卷积一阶段目标检测 (fully convolutional one-stage object detection, FCOS)[34] 为目标检测框架衍生出一系列单阶段的实例分割算法。还有一些研究重新思考了掩码的合理表征方式，从而提升实例分割精度。近两年研究人员也将自然语言处理中的 Transformer 模型成功应用到图像实例分割领域且有了较好的效果。此外，其他方法则结合了实例分割和目标检测的优势加以实现。单阶段的实例分割任务难点在于不添用建议框的辅助下如何直接区分不同物体，特别是同类别的不同实例和如何完整的保存逐像素点含有的位置信息和语义信息。

1.2.1　感知实例分割

本质上，实例分割可看作实例位置感知的语义分割，需要在区域级别上进行操作，并且同一像素在不同区域中可能具有不同的语义，如图 5 所示。

图 5(a) 中的 FCN 网络简洁、高效已经广泛应用在语义分割中。最早利用 FCN 网络实现实例分割的是图 (b) 中的 Instance FCN[35]，它以位置敏感图的形式将实例信息引入语义分割中实现了平移可变性。将原有 FCN 单一输出通道变为多个对实例位置敏感的通道，通过聚合位置敏感图得到每个实例掩码。

图 5(c) 中的全卷积感知实例分割 (FCN instance-aware)[36] 改进 Instance FCN 不能输出对象类别信息的问题，提出内和外的位置敏感评分图来同时进行检测和分割实例。两个子任务不仅共享卷积特征而且共享位置敏感评分图。随后，王子愉等[37] 在全卷积感知实例分割的检测分支中使用了具有大型可分离卷积来获得更精确的边界框。同时设计了一个包含边界细化操作的分割模块以获得更精确的掩模。

2019 年加利福尼亚大学提出一个新的实例分割算法[38]。这个简单的全卷积实例分割模型是一个实时实例分割算法，这比以前的任何算法都要快得多，尽管它的精度不是很高。YOLACT 将整个任务拆分为了两个子部分，一个部分是得到类似于 FCN 的语义分割原型图，另一个部分得到检测框，通过融合原型图和检测框，得到掩码。该团队接下来对此 YOLACT 算法进行了加强，提出 YOLACT 的改进版 YOLACT++[39]，速度高达 33.5 fps，将实时实例分割又推上了新的高潮。

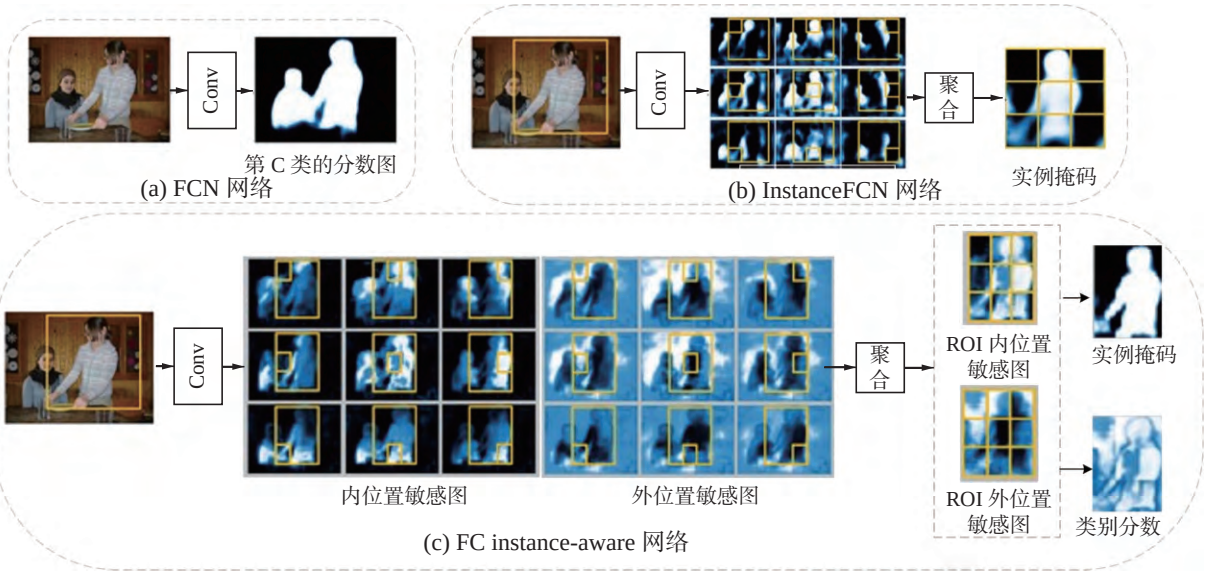

图 5　FCN 系列网络

Fig. 5　FCN series network

2020 年, CondInst[40] 可以做到真正的高速, 同时保证高精度, 彻底去掉了检测器的辅助。它用动态卷积的思想生成实例敏感的滤波器来编码实例信息, 不依赖目标框及 ROI Pooling 等特征对齐手段。用 FCOS 检测实例类别, 然后用动态网络生成的掩码头参数结合提取到含相对坐标信息的掩码特征图执行卷积操作生成最终的实例 mask。同年, SOLO[41] 将图像作为输入, 在全卷积特征图上输出相应类别概率直接输出实例蒙版, 无目标框监督, 既不需要 ROI Pooling 也不需要进行检测后处理过程[42]。SOLO 先划分正样本的栅格, 并且把每一个栅格对应一个物体的掩码,相当于一个正样本对应一张图, 这张图中只有这个正样本的掩码。SOLOv2[43] 在文献 [40] 基础上又提出动态学习分割目标掩码的思想, 将其分解为学习掩码核和生成掩码两个支路。此外, 引入专门针对掩码的后处理方法 Matrix NMS 可以一次完成具有并行矩阵运算, 比传统用于目标检测的 NMS[44] 能产生更好的结果。

Point INS[45] 从基于点特征实例分割的两个难点入手: 如何用点特征进行更强健的掩码表达和解决一阶段潜在存在特征错误分配建议框而带来后续分割错乱的问题, 提出实例感知卷积 (实例无关的特征和感知权重相配合)。而 CenterMask[46] 从另一方面解决基于点特征的一阶段实例分割即不同目标实例的区分和逐像素特征对齐。将其分解为两个子任务: 1) 局部掩码 (使用目标中心点) 来表示分离实例, 特别在多目标重叠环境下效果显著全局显著; 2) 在整张图片中生成全局的分割掩码。最后, 融合粗略但实例感知的局部掩码和精确但实例未知的全局掩码以形成最终实例的掩码。BlendMask[47] 是通过更合理的 blender 模块融合高层和底层的语义信息辅助来提取更准确的实例分割特征。AdaptIS[48] 的思想与其他主流方法不同, 它输入不仅是一张图像, 还需要人为指定一个点。即只需要一个目标身上的点, 就可以分割出这个目标实例的分割掩码。

1.2.2　建模掩码

传统的掩码 mask 表征方式是二值化, 也就是用一个矩阵表示, 矩阵中元素只有 0 和 1, 1 表示该位置是物体, 0 表示背景。目前, 大多数掩码局限于二维矩形框, 而现实世界中的物体大多都是不规则的多边形, 所以一些研究人员从如何合理建模掩码的角度出发研究实例分割问题。

2019 年 Tensormask[49] 通过 4D 的结构化张量在空间域中构建掩码, 是一种基于局部掩码的编码方式, 也是首个密集滑动窗口实例分割系统。虽然思想新颖但它的推理速度慢于两阶段 Mask R-CNN 且训练时间是 Mask R-CNN 的 6 倍。2020 年, Polarmask[50] 则提出了一种新的掩码编码形式, 使用极坐标建模来表示多边形目标, 将每个像素的掩码预测转变成在极坐标系下中心点分类和距离回归问题。而分析 Polarmask 的分割结果发现存在边缘信息模糊的问题, 因此提出了轮廓点细化的方法, 通过对轮廓点角度偏置和距离的预测, 使网络能够提取出更准确的实例轮廓。同年, Deep Snake[51] 用边缘建模的方式表征物体。并结合传统 snake 算法, 先给定一个初始边缘, 在

提取好的特征图上给边缘的每个节点提取一个特征，这样得到一个定义在边缘上的特征。然后用循环卷积 (circular convolution)[51] 构成的网络进行边缘上的特征学习，最后映射为指向物体轮廓的偏移，用于变形边缘。

虽然上面基于轮廓建模的方法具有易于优化和快速推断的优点。但是也有着天生的缺点，没有有效的表征目标中出现的空洞。因此，MEInst[52] 脱离目标检测的影响，考虑对掩码的宽度×高度进行压缩，从信息论的角度来说传统掩码表示中一定存在着信息冗余，因此可以更低的比特数对其进行表征，通过使用主成分分析法将掩码编码成一个统一的矩阵。

2021 年实例分割建模掩码的核心则是在没有实例像素标注时如何完成实例分割任务。LSNet[53] 类比 Polarmask 提出一种通用建模方式可用于检测，实例分割和姿态估计领域。Implicit PointRend[54] 提出基于点的实例级别标注，是实例分割中的一种新的弱监督形式。它可以将标准的边界框标注与标签点结合起来。BoxInst[55] 提出仅利用边界框监督完成实例分割，核心思想是重新设计实例分割中掩码损失，而无需修改分割网络本身。新的损失函数可以监督掩码训练，而无需依赖掩码注释。

1.2.3　Transformer 嵌入

最近，Transformer 模型在自然语言处理中的突破引起了计算机视觉社区的极大兴趣。Transformer 的关键部件是多头注意力，这可以显着提高模型的能力。目前，已有研究人员将 Transformer 应用到图像实例分割领域且有了较好的效果。ISTR[56] 是首个基于 Transformer 的端到端实例分割框架。ISTR 通过预测低维掩码嵌入和循环细化策略同时检测和分割实例，与自下而上和自上而下的框架相比，为实现实例分割提供了新的视角。SOTR[57] 利用 Transformer 简化了分割流程，使用两个并行子任务：1) 通过 Transformer 预测每个实例类别；2) 利用多级上采样模块动态生成分割掩码。此外提出的双 Transformer 在一定程度上提高了分割精度和训练收敛性。可见，编码器–解码器 Transformer 模型可以通过一系列可学习的掩码嵌入将实例分割任务统一。与 CNN 相比，视觉 Transformer 在实例分割领域具有很强的竞争力。

1.2.4　其他方法

经过上面介绍可知实例分割在一定程度上依附于目标检测任务，近年来出现了非常多优秀的算法解决这两个问题，且都取得了优异的效果。实际上，目标检测属于目标级别的任务，这类任务更关注物体级别的特征，对分辨率的需求不高，但需要更多的高级语义信息。而实例分割任务属于像素级别的任务，这类任务需要给出逐像素的输出，对分辨率的需求较高，需要更多的细节信息。但是，却鲜有文章深入分析两者之间的关联。这里介绍目前的几种工作。

HTC[58] 一项具有代表性的工作，它采用级联体系结构逐步完善了两个任务，并取得了可喜的成就。但是，这种多阶段设计带来了成本相对高的计算量。Cascade R-CNN[59] 为每个级联阶段添加了一个分割分支，将级联架构扩展到实例分段任务。RDS Net[60] 设计了双流网络在很大程度上克服了实例掩码的低分辨率，对目标框的严重依赖性以及边界框的定位错误。它引入 3 个模块即目标框辅助实例掩码关系模块，掩码修剪模块和掩码细化目标定位模块。Embed Mask[61] 通过引入建议框嵌入和像素嵌入的概念将基于建议框的方法和基于细分的方法结合在一起，以便根据实例建议框的嵌入相似性将像素分配给实例建议框。

综上所述，单阶段的实例分割算法种类繁杂，解决思路比较开阔，目前从精度和速度上看是最有效的算法，同时也摆脱了检测框的限制，是未来研究的趋势。

1.3　算法优缺点对比和实验结果比较

本小节对文中涉及到的部分实例分割算法进行优缺点比较和性能分析。表 1 是不同实例分割算法的优缺点对比。表 2 是不同实例分割算法在 COCO (microsoft common objects in context) 数据集上的性能对比。本文在最大程度上选择相同的基础网络且没有引入任何训练技巧，以保证算法性能比较的公平性。比较的结果均在 COCO 公开测试数据集上测试，因为 COCO 数据集是实例分割最常用的数据集，图片背景复杂，目标种类和数量多，目标尺寸相对较小，有很大难度。算法性能主要比较的参数是精度（COCO 评价标准，详细介绍见 3.5）及模型参数量（#Params）和推断速度（fps）。其中，fps 指每秒帧数，值越大算法速度越快，“–”表示未知，学习率规则“1×”表示模型训练 12 个 epoch（180K iterations），“3×”为 36 个 epoch，以此类推。表 1 和表 2 主要从两阶段和单阶段这两类对通用场景下的实例分割算法进行分类总结。从精度上看，相同基础网络时两阶段普遍优于单阶段，且模型所需训练迭代次数少。从速度上看，单阶段则快于两阶段，且精度也是处于平均水平，但以大量训练迭代次数为代价。因

此, 应用时需要根据具体需求选择合适的算法。综上分析, 没有一种算法可以兼顾精度高和速度快, 未来还可以从同时提升实例分割的速度与精度入手。

表 1　不同实例分割算法的优缺点对比

Table 1　Comparison of the advantages and disadvantages of different instance segmentation algorithms

	算法	年份	技术	优点	缺点
两阶段	SDS[5]	2014	检测生成掩码候选区域再语义分割	最早实例分割算法	掩码粗糙
	Hypercolumns[8]	2015	改进SDS将底层特征与高层特征融合	提升分类的精确性并改善目标分割细节	
	CFM[9]	2015	从卷积特征中提取掩码而非原始图像中提取	首次将掩码（Mask）引入	
	DeepMask[10]	2015	生成高召回率的掩码候选区域	不依赖于边缘、超像素等其他形式	只能捕捉目标大致外形, 不能准确描绘目标边界
	MNC[12]	2016	级联结构,3个不同功能的网络级联	快速地进行推断	—
	SharpMask[11]	2016	优化DeepMask,引入较低层位置特征加以细化	更精确框定物体边界的掩码	推断时间缓慢
	Mask R-CNN[13]	2017	Faster RCNN, ResNet-FPN, ROI Align, FCN的mask分支	可并行完成目标检测和分割两项任务	依赖目标检测结果
	PANet[19]	2018	自下而上特征路径, 自适应融合的ROI池化	增强不同尺度间的信息融合和特征利用	—
	DetNet[20]	2018	空洞卷积加入骨干网络	提高特征表达能力	—
	MS R-CNN[21]	2019	修改mask评价标准	合理评价掩码结果	—
	PointRend[22]	2019	将实例分割看作图像处理中渲染问题	细化Mask R-CNN掩码	结果处理比较复杂
	BMask R-CNN[23]	2020	目标边缘信息加入掩码分支	细化掩码边缘	—
	Associative embedding[28], SCI[33] ,SGN[30]	2017	将学习关系嵌入成组来分配像, 引入判别损失函数, 序列组合网络	不依赖检测的候选框	精度很低
单阶段	Instance FCN[35]	2016	位置敏感图	实例位置感知的FCN	精度很低
	FCIS[36]	2017	内/外的位置敏感图	改进[35]不能输出目标类别	
	YOLACT[38]/ YOLACT++[39]	2019	融合原型图和检测框	实时实例分割	精度较两阶段低
	HTC[58]	2019	级联结构	不同任务互惠互利	参数计算量高
	SOLO[41]/ SOLO V2[43]	2019	提取目标的点特征, 划分栅格, Matrix NMS	速度快, 精度高	训练时间长
	TensorMask[49]	2019	4Dtensor构建mask	密集滑动窗口分割	推理分割速度慢, 计算复杂度高
	Polarmask[50]	2020	极坐标建模mask	方法新颖	边缘信息模糊
	Deep Snake[51]	2020	边缘建模mask	易于优化和快速推断	不能很好表征目标中出现的空洞
	MEInst[52]	2020	矩阵编码mask	去除信息冗余	精度差些
	CenterMask[46]	2020	分解为局部和全局掩码	兼顾速度和精度	—
	BlendMask[47]	2020	blender融合高层和底层的特征	更准确分割mask	—
	CondInst[40]	2020	动态网络直接输出掩码	高速高精度	大目标缺少分割细节
	LSNet[53]	2021	可用于检测, 实例分割和姿态估计领域的通用建模方式	无需精细的掩码标注	精度差些

表 2　不同实例分割算法的性能对比

Table 2　Performance comparison of different instance segmentation algorithms

算法	基础网络	学习规则	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#Params	fps	GPU
Mask R-CNN[13]	ResNet-101 FPN	1×	35.7	58.0	37.8	15.5	38.1	52.4	135.0M	5.1	V100
Mask R-CNN[13]	ResNeXt-101-FPN	1×	37.1	60.0	39.4	16.9	39.9	53.5	137.1M	7.9	V100
Masklab[18]	ResNet-101	1×	35.4	57.4	37.4	16.9	38.3	49.2	—	—	—
PANet[19]	ResNeXt-101-FPN	1×	40.0	62.8	43.1	18.8	42.3	57.2	—	23.8	V100
MS R-CNN[21]	ResNet-101 FPN	1×	38.3	58.8	41.5	17.8	40.4	54.4	208.6M	5.9	V100
Point Rend[22]	ResNet-50-FPN	1×	36.3	—	—	—	—	—	147.2M	—	—
BMask [23]	ResNet-101-FPN	1×	37.7	59.3	40.6	16.8	39.9	54.6	195.4M	—	—
FCIS++[36]	ResNet-101-C5	1×	33.6	54.5	—	—	—	—	—	—	—
YOLACT550[39]	ResNet-101-FPN	4×	29.8	48.5	31.2	9.9	31.3	47.7	—	33.3	V100
Cascade Mask R-CNN[59]	ResNet-101-FPN	1×	38.4	60.2	41.4	20.2	41.0	50.6	252M	8.1	V100
HTC[58]	ResNet-101-FPN	1×	39.7	61.8	43.1	21.0	42.2	53.5	326.4M	2.4	V100
SOLO[41]	ResNet-101-FPN	6×	37.8	59.5	40.4	16.4	40.6	54.2	67.4M	22.8	V100
SOLOv2[43]	ResNet-101-FPN	6×	39.7	60.7	42.9	17.3	42.9	57.4	65.5M	31.4	V100
Tensormask[49]	ResNet-101-FPN	6×	37.1	59.3	39.4	17.4	39.1	51.6	—	2.7	V100
Polarmask[50]	ResNet-101-FPN	1×	30.4	51.9	31.0	13.4	32.4	42.8	—	12.3	V100
MEInst[52]	ResNet-101-FPN	1×	33.0	56.4	34.0	15.2	35.3	46.3	36. 9M	16.2	V100
CenterMask[46]	Hourglass-104	—	34.5	56.1	36.3	16.3	37.4	48.4	—	12.3	V100
BlendMask[47]	ResNet-101-FPN	3×	38.4	60.7	41.3	18.2	41.5	53.3	54.7M	9.8	1080Ti
CondInst[40]	ResNet-101-FPN	3×	39.1	60.9	42.0	21.5	41.7	50.9	54.3M	12.0	1080Ti
ISTR[56]	ResNet-101-FPN	3×	39.9	—	—	22.8	41.9	52.3	—	11.0	1080Ti
SOTR[57]	ResNet-101-FPN	3×	40.2	61.2	43.4	10.3	59.0	73.0	—	7.14	V100

2　实例分割的特殊应用

实例分割作为像素级别的目标识别任务，目前已广泛应用在遥感影像[62-67]，文字检测[68-70]，人脸检测[71-72]，辅助驾驶系统[73-76]，医疗图像处理[77-78]等各个场景下。

遥感图像中需要对标的物体进行识别，进而分析与测绘[79]。李澜[80]将 Mask R-CNN 应用于高分辨率光学遥感影像的目标检测和实例分割任务中，目的是在地图上找到遗漏的地理实体并提高矢量地图的质量。瑚敏君等[65]在 Mask R-CNN 原有的特征提取中每个层级的特征图后再增加一层卷积操作。然后，在原有掩码预测结构的基础上增加一个分支实现了高效、准确的高分辨率遥感影像建筑物提取算法。王昌安[79]则用于光遥感影像中近岸舰船的检测任务。

辅助驾驶系统不仅需要在行驶过程中识别不同的车道线，进行驾驶模式的决策，而且也需要对周围的车辆、行人等进行分析，判断周围的驾驶环境等这些都用到了实例分割[81-82]。邓琉元等[83]针对无人驾驶中用到的环视相机所呈环形图像中存在目标几何畸变难以建模问题，在 Mask R-CNN 中引入可变形卷积和可变形 ROI Pooling 来提升网络对几何形变的建模能力以实现环视鱼眼图像中准确的交通目标实例分割。蔡英凤等[73]和田锦等[74]将实例分割模型用于车道线检测解决了传统的车道线检测算法易受光照变化、阴影遮挡等环境干扰的缺陷。最后，所提算法可以完成复杂交通场景下的多车道线实时检测。除此之外，陈健雄[84]提出实例分割模型也可以有效识别中低速磁浮列车上接触轨固件的松动状态，保证了城市轨道交通的安全运行。

医疗图像处理需要对血管、骨骼、细胞等区域进行分割与检测，帮助医生进行诊断和研究[81]。同时降低误诊率和漏诊率，所以实例分割也是重要的关键技术之一。赵旭[77]研究基于实例分割的乳腺超声肿瘤识别，分割出乳腺超声图像的肿

瘤区。郑杨等[78] 在 Mask R-CNN 中加入空洞卷积完成宫颈细胞图像分割。吴宇[85] 则提出一个级联的 3D 椎骨分割网络。

可见, 实例分割应用已经非常广泛, 都是建立在两阶段 Mask R-CNN[13] 基础之上并有很好的算法效果。未来, 实例分割技术一定会有更大的发展应用前景。

3 数据集与评价指标

深度学习领域关注的是通过使用计算机算法自动发现数据中的规律性, 并通过使用这些规律性来采取一些行动。可见, 数据规模驱动深度学习领域的发展, 收集一个大规模的数据集也是实例分割研究中重要的工作。目前, 公开的大型数据集大多是由公司、科研团队或特别举办的专业比赛等收集创建的, 需要大量人工进行手动标注, 时间成本高[86]。本节简要归纳几种常用的实例分割数据集及评价指标。

3.1 COCO 数据集

COCO[87] 起源于 2014 年由微软出资标注的 Microsoft COCO 数据集, 与 ImageNet 竞赛一样, 被视为是计算机视觉领域最受关注和最权威的比赛之一。COCO 数据集是一个大型的、丰富的目标检测, 实例分割和字幕数据集。这个数据集以场景理解为目标, 主要从复杂的日常场景中截取, 图像中的目标通过精确的分割进行位置的标定。图像包括 91 类目标, 328 000 个影像和 2 500 000 个标签。目前为止有实例分割的最大且使用最广泛的数据集, 提供的类别有 80 类, 有超过 33 万张图片, 其中 20 万张有标注, 整个数据集中个体的数目超过 150 万个。使用时划分为训练集、验证集和测试集 3 个部分, 已成为比较实例分割算法性能最重要的公开数据集。

3.2 Cityscapes 数据集

Cityscapes[88] 是一个大规模城市场景数据集, 主要用于语义分割任务, 拥有 5 000 张在城市环境中驾驶场景的图像(2 975 张训练集, 500 张验证集, 1 525 张测试集)记录了 50 个不同城市的街道场景。它具有 19 个类别的密集像素标注(97% coverage), 其中 8 种类别具有实例级别分割标注。

3.3 Mapillary Vistas 数据集

Mapillary Vistas[89] 数据集是一个新建立的, 大场景的街景数据集, 用于图像语义分割以及图像实例分割, 旨在进一步开发用于视觉道路场景理解的先进算法。它包括 25 000 张高分辨率的彩色图像, 分成 66 个类, 其中有 37 个类别是特定的附加于实例的标签。对物体的标签注释可以使用多边形进行稠密, 精细的描绘。与 Cityscapes 相比, Mapillary Vistas 的精细注释总量大了 5 倍, 并包含来自世界各地在各种条件下捕获的图像, 包括不同天气, 季节和时间的图像。

3.4 LVIS 数据集

LVIS[90](large vocabulary instance segmentation) 是由 Facebook AI Research 于 2019 年建立的大型词汇实例分割数据集。目前公布的实例分割数据集的目标类别还是较少, 与实际应用场景下存在大量(未知)类别相违背。故 LVIS 收集了 164 000 张图像, 对 1 000 多个对象类别标注, 共有 220 万个高质量的实例分割掩码标签。相比于 COCO 数据集, LVIS 人工标注掩码具有更大的重叠面积和更好的边界连续性, 更精确的掩码。并且在数据成长尾分布 (类别种类多而单类的实例个数少) 时仍有很好的训练效果。

3.5 评价指标

这里以常用 COCO 数据集的评价指标为例。COCO 数据集官方评价标准如表 3 所示。AP 代表所有类别的平均精度, 作为最终 COCO 评价整体标准。AP 的定义使用并交比 (intersection-over-union, IOU) 的标准, 即两个实例掩码的重叠度。表 3 中 area 是指分割掩码 mask中像素的数量。同时 AP 也计算不同尺度目标如大目标, 中目标以及小目标的实例分割精度。

表 3　COCO 数据集的评价指标

Table 3　Evaluation index of COCO dataset

评价指标	含义
AP	IOU=0.50:0.05:0.95
AP_{50}	IOU=0.50
AP_{75}	IOU=0.75
AP_S	area < 32^2
AP_M	32^2 < area < 96^2
AP_L	area > 96^2

4 未来展望

综合来看, 实例分割技术正趋向兼并算法实时性和性能高精度的方向发展。单阶段的实例分割在性能上不弱于两阶段的实例分割, 但相较于两阶段法的网络架构更为简洁, 高效且易于训练。由现存算法的性能比较来看还有提升空间。所以, 总体期望发展的方向应该是在追求精度提升的基础上实现快速实时实例分割, 更好地适用

于实际应用。此外，

1) 笔者认为实例分割与目标检测，语义分割等其他高级计算机视觉任务可以互惠互利，可重点研究在不同图像感知任务之间的相互关系。此外，自然语言处理和计算机视觉两大任务可以彼此互鉴。最近，自然语言处理中常用的 Transformer[91, 92] 在计算机视觉 (computer vision, CV) 领域已经做了一些初步探索，未来针对 CV 的特点设计更适配视觉特性的 Transformer 将会带来更好的性能提升[93]。

2) 目标间遮挡和交叠情况仍然是实例分割最具挑战性的问题，可借鉴图卷积神经网络，胶囊网络和目标检测中的推理关系网络来有效解决遮挡情况下的实例分割问题。

3) 目前实例分割只针对单独的目标，没有考虑目标间的关系。从目标检测的经验来看，图像中不同目标是具有空间和语义的上下文联系，这种信息的流动和融合有助于目标检测精度的提升。实例分割可以借鉴注意力机制，图神经网络的方法来建立目标在空间布局以及几何形状之间的联系。

4) 从现有算法的精度来看，小目标的实例分割问题仍然是一个挑战。COCO 数据集中定义像素总数小于 32^2 为小目标。可见其在图像中像素面积占比很小，经过多次采样和池化等定会缺少很多细节。而实例分割是一个需要精确和完整的像素信息才能完成的任务，两者产生矛盾。未来的研究可以小目标检测为切入点，结合超分辨率图像任务、生成对抗网络、尺度自适应和注意力机制等策略来提高小目标的实例分割精度。

5) 实例分割大多是有监督学习，其数据采用人工手动进行像素标注的方式，繁琐的数据标注耗费大量的人力和时间。为了减少成本，使用自监督学习、弱监督学习方式从已有未标注或少量标注数据中自动生成标签实现实例分割。也可利用现有的已标注边界框作为先验信息辅助锁定目标范围。

6) 从实际应用的角度，现有网络设计的复杂度高，占用内存大，速度和准确度之间还不能达到平衡。轻量化的网络架构，满足速度快和精度高的需求将是实例分割未来探究的重要内容。

参考文献：

[1] FUKUSHIMA K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980, 36(4): 193–202.

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

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the acm, 2017, 60(6): 84–90.

[4] 董俊杰, 刘华平, 谢珺, 等. 基于反馈注意力机制和上下文融合的非模式实例分割 [J]. 智能系统学报, 2021, 16(4): 801–810.
DONG Junjie, LIU Huaping, XIE Jun, et al. Feedback attention mechanism and context fusion based amodal instance segmentationChinese Full Text[J]. CAAI transactions on intelligent systems, 2021, 16(4): 801–810.

[5] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]//Computer vision—ECCV 2014. Berlin, German: Springer, 2014: 297−312.

[6] 梁新宇, 林洗坤, 权冀川, 等. 基于深度学习的图像实例分割技术研究进展 [J]. 电子学报, 2020, 48(12): 2476–2486.
LIANG Xinyu, LIN Xikun, QUAN Jichuan, et al. Research on the progress of image instance segmentation based on deep Learning[J]. Acta electronica sinica, 2020, 48(12): 2476–2486.

[7] ARBELÁEZ P, PONT-TUSET J, BARRON J, et al. Multiscale combinatorial grouping[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014: 328−335.

[8] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Hypercolumns for object segmentation and fine-grained localization[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 447−456.

[9] DAI Jifeng, HE Kaiming, SUN Jian. Convolutional feature masking for joint object and stuff segmentation [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 3992−4000.

[10] PINHEIRO P O, COLLOBERT R, DOLLAR P. Learning to segment object candidates [M]//Advances in neural information processing systems. Morgan Kaufmann Publishers, 2015. 1990−1998.

[11] PINHEIRO P O, LIN T Y, COLLOBERT R, et al. Learning to refine object segments[M]//Computer Vision – ECCV 2016. Berlin, German: Springer, 2016:

75−91.

[12] DAI Jifeng, HE Kaiming, SUN Jian. Instance-aware semantic segmentation via multi-task network cascades [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 3150−3158.

[13] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 2980−2988.

[14] REN Shaoqing, HE Kaiming, 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.

[15] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 770−778.

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017: 936−944.

[17] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 3431−3440.

[18] CHEN Liangchieh, HERMANS A, PAPANDREOU G, et al. MaskLab: instance segmentation by refining object detection with semantic and direction features[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2018: 4013−4022.

[19] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2018: 8759−8768.

[20] LI Zeming, PENG Chao, YU Gang, et al. DetNet: design backbone for object detection[M]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 339−354.

[21] HUANG Zhaojin, HUANG Lichao, GONG Yongchao, et al. Mask scoring R-CNN[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2019: 6402−6411.

[22] KIRILLOV A, WU Yuxin, HE Kaiming, et al. PointRend: image segmentation as rendering[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 9796−9805.

[23] CHENG Tianheng, WANG Xinggang, HUANG Lichao, et al. Boundary-preserving mask R-CNN[C]//Computer Vision – ECCV 2020. Berlin, German: Springer, 2020: 660−676.

[24] TANG CHUFENG, CHEN HANG, LI XIAO, et al. Look closer to segment better: boundary patch refinement for instance segmentation[EB/OL]. (2021-04-12) [2021-09-30].https://arxiv.org/abs/2104.05239.

[25] ZHANG GANG, LU XIN, TAN JINGRU, et al. RefineMask: towards high-quality instance segmentation with fine-grained features[EB/OL]. (2021-04-12) [2021-09-30].https://arxiv.org/abs/2104.08569.

[26] 姜世浩, 齐苏敏, 王来花, 等. 基于 Mask R-CNN 和多特征融合的实例分割 [J]. 计算机技术与发展, 2020, 30(9): 65–70.
JIANG Shihao, QI Sumin, WANG Laihua, et al. Instance segmentation modal based on mask R-CNN and multi-feature Fusion[J]. Computer technology and development, 2020, 30(9): 65–70.

[27] BAI Min, URTASUN R. Deep watershed transform for instance segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017: 2858−2866.

[28] NEWELL A, HUANG ZHIAO, DENG JIA. Associative embedding: end-to-end learning for joint detection and grouping[EB/OL]. (2016-11-16) [2021-09-30].https://arxiv.org/abs/1611.05424.

[29] DE BRABANDERE B, NEVEN D, VAN GOOL L. Semantic instance segmentation for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. New York, USA: IEEE, 2017: 478−480.

[30] LIU Shu, JIA Jiaya, FIDLER S, et al. SGN: sequential grouping networks for instance segmentation[C]//2017 IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 3516−3524.

[31] GAO Naiyu, SHAN Yanhu, WANG Yupei, et al. SSAP: single-shot instance segmentation with affinity pyramid[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 642−651.

[32] FATHI A, WOJNA Z, RATHOD V, et al. Semantic in-

stance segmentation via deep metric learning[EB/OL]. (2017-03-30) [2021-09-30].https://arXiv:1703.10277.

[33] BRABANDERE B D, NEVEN D, GOOL L V. Semantic instance segmentation with a discriminative loss function[EB/OL]. (2017-08-08) [2021-09-30].https://arxiv.org/abs/1708.02551v1.

[34] TIAN Zhi, SHEN Chunhua, CHEN Hao, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 9626−9635.

[35] DAI Jifeng, HE Kaiming, LI Yi, et al. Instance-sensitive fully convolutional networks[M]//Computer Vision – ECCV 2016. Berlin, German: Springer, 2016: 534−549.

[36] LI Yi, QI Haozhi, DAI Jifeng, et al. Fully convolutional instance-aware semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017: 4438−4446.

[37] 王子愉, 袁春, 黎健成. 利用可分离卷积和多级特征的实例分割 [J]. 软件学报, 2019, 30(4): 954–961.
WANG Ziyu, YUAN Chun, LI Jiancheng. Instance segmentation with separable convolutions and multi-level features[J]. Journal of software, 2019, 30(4): 954–961.

[38] BOLYA D, ZHOU Chong, XIAO Fanyi, et al. YOLACT: real-time instance segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 9156−9165.

[39] Bolya D, Zhou C, Xiao F, et al. Yolact++: Better real-time instance segmentation[EB/OL]. (2020-08-05) [2021-09-30].https://pubmed.ncbi.nlm.nih.gov/32755851/

[40] TIAN Zhi, SHEN Chunhua, CHEN Hao. Conditional convolutions for instance segmentation[M]//Computer Vision – ECCV 2020. Berlin, German: Springer, 2020: 282−298.

[41] WANG Xinlong, KONG Tao, SHEN Chunhua, et al. SOLO: segmenting objects by locations[M]//Computer Vision – ECCV 2020. Berlin, German: Springer, 2020: 649−665.

[42] 李晓筱, 胡晓光, 王梓强, 等. 基于深度学习的实例分割研究进展 [J]. 计算机工程与应用, 2021, 57(9): 60–67.
LI Xiaoxiao, HU Xiaoguang, WANG Ziqiang, et al. Survey of instance segmentation based on deep learning[J]. Computer engineering and applications, 2021, 57(9): 60–67.

[43] WANG X, ZHANG R, KONG T, et al. SOLOv2: Dynamic and fast instance segmentation[EB/OL]. (2020-03-23) [2021-09-30]. https://arxiv.org/abs/2003.10152.

[44] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition. New York, USA: IEEE, 2006: 850−855.

[45] QI LU, ZHANG XIANGYU, CHEN YINGCONG, et al. PointINS: point-based instance segmentation[EB/OL]. (2020-03-13) [2021-09-30].https://arxiv.org/abs/2003.06148v1.

[46] WANG Yuqing, XU Zhaoliang, SHEN Hao, et al. CenterMask: single shot instance segmentation with point representation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 9310−9318.

[47] CHEN Hao, SUN Kunyang, TIAN Zhi, et al. BlendMask: top-down meets bottom-up for instance segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 8570−8578.

[48] SOFIIUK K, SOFIYUK K, BARINOVA O, et al. AdaptIS: adaptive instance selection network[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 7354−7362.

[49] CHEN Xinlei, GIRSHICK R, HE Kaiming, et al. TensorMask: a foundation for dense object segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 2061−2069.

[50] XIE Enze, SUN Peize, SONG Xiaoge, et al. PolarMask: single shot instance segmentation with polar representation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 12190−12199.

[51] PENG Sida, JIANG Wen, PI Huaijin, et al. Deep snake for real-time instance segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 8530−8539.

[52] ZHANG Rufeng, TIAN Zhi, SHEN Chunhua, et al. Mask encoding for single shot instance segmentation [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020: 10223−10232.

[53] DUAN KAIWEN, XIE LINGXI, QI HONGGANG, et al. Location-sensitive visual recognition with cross-IOU

loss[EB/OL]. (2021-04-11) [2021-09-30].https://arxiv.org/abs/2104.04899.

[54] CHENG BOWEN, PARKHI O, KIRILLOV A. Pointly-supervised instance segmentation[EB/OL]. (2021-04-13) [2021-09-30].https://arxiv.org/abs/2104.06404

[55] LEE Jungbeom, YI Jihun, SHIN Chaehun, et al. BBAM: bounding box attribution map for weakly supervised semantic and instance segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2021: 2643−2651.

[56] HU J, CAO L, LU Y, et al. ISTR: End-to-End Instance Segmentation with Transformers [EB/OL]. (2020-11-30) [2021-09-30].https://arxiv.org/abs/2011.14503v4.

[57] GUO R, NIU D, QU L, et al. SOTR: Segmenting Objects with Transformers [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2021: 7157−7166.

[58] CHEN Kai, PANG Jiangmiao, WANG Jiaqi, et al. Hybrid task cascade for instance segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2019: 4969−4978.

[59] CAI Zhaowei, VASCONCELOS N. Cascade R-CNN: high quality object detection and instance segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(5): 1483–1498.

[60] WANG Shaoru, GONG Yongchao, XING Junliang, et al. RDSNet: a new deep architecture forReciprocal object detection and instance segmentation[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(7): 12208–12215.

[61] Ying H, Huang Z, Liu S, et al. Embedmask: Embedding coupling for one-stage instance segmentation [EB/OL]. (2019-12-04) [2021-09-30].https://arxiv.org/abs/1912.01954.

[62] 惠健, 秦其明, 许伟, 等. 基于多任务学习的高分辨率遥感影像建筑实例分割 [J]. 北京大学学报(自然科学版), 2019, 55(6): 1067–1077.
HUI Jian, QIN Qiming, XU Wei, et al. Instance segmentation of buildings from high-resolution remote sensing images with multitask learning[J]. Acta scientiarum naturalium universitatis pekinensis, 2019, 55(6): 1067–1077.

[63] 何代毅, 施文灶, 林志斌, 等. 基于改进 Mask-RCNN 的遥感影像建筑物提取 [J]. 计算机系统应用, 2020, 29(9): 156–163.
HE Daiyi, SHI Wenzao, LIN Zhibin, et al. Building extraction from remote sensing image based on improved mask-RCNN[J]. Computer systems & applications, 2020, 29(9): 156–163.

[64] 宋师然. 高分遥感城市建筑物对象化识别方法研究 [D]. 北京: 北京建筑大学, 2020.
SONG Shiran. Research on object recognition method of urban buildings in high spatial resolution remote sensing imagery[D]. Beijing: Beijing University of Civil Engineering and Architecture, 2020.

[65] 瑚敏君, 冯德俊, 李强. 基于实例分割模型的建筑物自动提取 [J]. 测绘通报, 2020(4): 16–20,62.
HU Minjun, FENG Dejun, LI Qiang. Automatic extraction of buildings based on instance segmentation model[J]. Bulletin of surveying and mapping, 2020(4): 16–20,62.

[66] 于志文. 基于语义分割和实例分割的高分辨率遥感影像建筑物提取方法研究 [D]. 桂林: 桂林理工大学, 2020.
YU Zhiwen. Research on building extraction method of high resolution remote sensing images based on semantic segmentation and instance segmentation[D]. Guilin: Guilin University of Technology, 2020.

[67] 朱意星. 基于深度学习的文本与遥感图像目标检测研究 [D]. 合肥: 中国科学技术大学, 2020.
ZHU Yixing. Research on deep learning based object detection for text and aerial image[D]. Hefei: University of Science and Technology of China, 2020.

[68] 刘春, 田倬韬, 刘绍辉, 等. 一种改进的图像中的文本检测模型 [J]. 微电子学与计算机, 2020, 37(6): 83–88.
LIU Chun, TIAN Zhuotao, LIU Shaohui, et al. An improved text detection model in image[J]. Microelectronics & computer, 2020, 37(6): 83–88.

[69] 张小爽. 基于实例分割的场景图像文字检测 [D]. 杭州: 浙江大学, 2018.
ZHANG Xiaoshuang. Scene text detection based on instance segmentation[D]. Hangzhou: Zhejiang University, 2018.

[70] 李煌, 王晓莉, 项欣光. 基于文本三区域分割的场景文本检测方法 [J]. 计算机科学, 2020, 47(11): 142–147.
LI Huang, WANG Xiaoli, XIANG Xinguang. Scene text detection based on triple segmentation[J]. Computer science, 2020, 47(11): 142–147.

[71] 邓宏杰. 基于 Mask R-CNN 的人脸检测分割的改进研究 [J]. 现代计算机, 2020(27): 57–62,67.

DENG Hongjie. Research on the improvement of face detection and segmentation algorithm based on mask R-CNN[J]. Modern computer, 2020(27): 57–62,67.

[72] 王耀东. 基于 Mask RCNN 神经网络的行人重识别研究 [D]. 西安: 西安科技大学, 2020.
WANG Yaodong. Person Re-identification research based on mask RCNN neural network[D]. Xi'an: Xi'an University of Science and Technology, 2020.

[73] 蔡英凤, 张田田, 王海, 等. 基于实例分割和自适应透视变换算法的多车道线检测 [J]. 东南大学学报 (自然科学版), 2020, 50(4): 775–781.
CAI Yingfeng, ZHANG Tiantian, WANG Hai, et al. Multi-lane detection based on instance segmentation and adaptive perspective transformation[J]. Journal of southeast university (natural science edition), 2020, 50(4): 775–781.

[74] 田锦, 袁家政, 刘宏哲. 基于实例分割的车道线检测及自适应拟合算法 [J]. 计算机应用, 2020, 40(7): 1932–1937.
TIAN Jin, YUAN Jiazheng, LIU Hongzhe. Instance segmentation based lane line detection and adaptive fitting algorithm[J]. Journal of computer applications, 2020, 40(7): 1932–1937.

[75] 栗杰. 基于深度学习的现实交通场景下目标检测算法研究 [D]. 大连: 大连海事大学, 2020.
LI Jie. Research on object detection algorithm in real traffic scene based on deep learning[D]. Dalian: Dalian Maritime University, 2020.

[76] 王中宇, 倪显扬, 尚振东. 利用卷积神经网络的自动驾驶场景语义分割 [J]. 光学 精密工程, 2019, 27(11): 2429–2438.
WANG Zhongyu, NI Xianyang, SHANG Zhendong. Autonomous driving semantic segmentation with convolution neural networks[J]. Optics and precision engineering, 2019, 27(11): 2429–2438.

[77] 赵旭. 基于医学先验的多尺度乳腺超声肿瘤实例分割方法 [D]. 哈尔滨: 哈尔滨工业大学, 2019.
ZHAO Xu. Medical knowledge constrained instance segmentation method for multi-scale breast ultrasound tumor[D]. Harbin: Harbin Institute of Technology, 2019.

[78] 郑杨, 梁光明, 刘任任. 基于 Mask R-CNN 的宫颈细胞图像分割 [J]. 计算机时代, 2020(10): 68–72.
ZHENG Yang, LIANG Guangming, LIU Renren. Cervical cell image segmentation based on Mask R-CNN[J]. Computer era, 2020(10): 68–72.

[79] 王昌安. 遥感影像中的近岸舰船目标检测和细粒度识别方法研究 [D]. 武汉: 华中科技大学, 2019.
WANG Chang'an. Detection and fine-grained recognition of inshore ships on optical remote sensing images[D]. Wuhan: Huazhong University of Science and Technology, 2019.

[80] 李澜. 基于 Mask R-CNN 的高分辨率光学遥感影像的目标检测与实例分割 [D]. 武汉: 武汉大学, 2018.
LI Lan. Object detection and instance segmentation in optical high-resolution remote sensing imagery based on mask R-CNN[D]. Wuhan: Wuhan University, 2018.

[81] 刘一丁. 基于像素亲和性和语义信息的图像实例分割研究 [D]. 合肥: 中国科学技术大学, 2020.
LIU Yiding. Pixel-affinity-aware and semantic-aware image instance segmentation[D]. Hefei: University of Science and Technology of China, 2020.

[82] 张继凯, 赵君, 张然, 等. 深度学习的图像实例分割方法综述 [J]. 小型微型计算机系统, 2021, 42(1): 161–171.
ZHANG Jikai, ZHAO Jun, ZHANG Ran, et al. Survey of image instance segmentation methods using deep learning[J]. Journal of Chinese computer systems, 2021, 42(1): 161–171.

[83] 邓琉元, 杨明, 王春香, 等. 基于环视相机的无人驾驶汽车实例分割方法 [J]. 华中科技大学学报(自然科学版), 2018, 46(12): 24–29.
DENG Liuyuan, YANG Ming, WANG Chunxiang, et al. Surround view cameras based instance segmentation method for autonomous vehicles[J]. Journal of Huazhong university of science and technology (natural science edition), 2018, 46(12): 24–29.

[84] 陈健雄. 基于实例分割的中低速磁浮接触轨紧固件松动识别 [D]. 成都: 西南交通大学, 2019.
CHEN Jianxiong. Fastener looseness recognition of medium-low speed maglev contact rail based on instance segmentation[D]. Chengdu: Southwest Jiaotong University, 2019.

[85] 吴宇. 基于深度学习的椎骨实例分割算法研究 [D]. 成都: 电子科技大学, 2020.
WU Yu. Research of vertebra instance segmentation based on deep learning[D]. Chengdu: University of Electronic Science and Technology of China, 2020.

[86] 范丽丽, 赵宏伟, 赵浩宇, 等. 基于深度卷积神经网络的目标检测研究综述 [J]. 光学 精密工程, 2020, 28(5): 1152–1164.

FAN Lili, ZHAO Hongwei, ZHAO Haoyu, et al. Survey of target detection based on deep convolutional neural networks[J]. Optics and precision engineering, 2020, 28(5): 1152–1164.

[87] LIN Tsungyi, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Computer vision—ECCV 2014. Berlin, German: Springer, 2014: 740–755.

[88] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 3213–3223.

[89] NEUHOLD G, OLLMANN T, BULÒ S R, et al. The mapillary vistas dataset for semantic understanding of street scenes[C]//2017 IEEE International Conference on Computer Vision. New York, USA: IEEE, 2017: 5000–5009.

[90] GUPTA A, DOLLÁR P, GIRSHICK R. LVIS: a dataset for large vocabulary instance segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2019: 5351–5359.

[91] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017(29): 5998–6008.

[92] LIU ZE, LIN YUTONG, CAO YUE, et al. Swin transformer: hierarchical vision transformer using shifted windows[EB/OL]. (2021-03-25) [2021-09-30].https://arxiv.org/abs/2103.14030.

[93] HAN KAI, WANG YUNHE, CHEN HANTING, et al. A survey on visual transformer[EB/OL]. (2020-12-23) [2021-09-30]. https://arxiv.org/abs/2012.12556v1.

作者简介：

苏丽，副教授，主要研究方向为环境感知与智能控制、智能船舶、机器视觉检测技术、多传感器信息融合、先进控制理论、智能监控。

孙雨鑫，博士研究生，主要研究方向为计算机视觉。

苑守正，博士研究生，主要研究方向为智能船舶系统、水面无人船控制、水面无人船自动靠泊和非线性系统预测控制。