

# 基于 Mask R-CNN 和多特征融合的实例分割

姜世浩, 齐苏敏, 王来花, 贾惠

(曲阜师范大学 软件学院, 山东 曲阜 273165)

**摘要:**为了能够充分地利用图像特征信息,提升实例分割的效果,提出了一种基于 Mask R-CNN 网络结构和多特征融合的实例分割模型。首先,在 Mask R-CNN 模型的基础上引入两条分支:一条基于整体嵌套边缘检测(HED)模型的边缘检测分支生成偏重于边缘信息的边缘特征图,一条基于全卷积网络(FCN)的语义分割分支生成偏重于空间位置信息的语义特征图。然后,在进行感兴趣区域对齐(ROIAlign)时,为了充分利用特征金字塔的各层信息,将感兴趣区域(ROI)同时映射到相应的金字塔层及其相邻层。最后,融合以上得到的多个特征图,生成信息更加丰富的新特征用于后续的检测和分割任务。实验结果表明,该方法有效提高了检测和分割的准确性。在使用 Resnet50-FPN 作为骨干网络且没有附加条件的情况下,与 Mask R-CNN 相比,该模型的检测和分割平均精度(mAP)分别提升了 1.2% 和 1.0%。

**关键词:**实例分割;深度学习;Mask R-CNN;全卷积网络;特征融合

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2020)09-0065-06

doi:10.3969/j.issn.1673-629X.2020.09.012

## Instance Segmentation Modal Based on Mask R-CNN and Multi-feature Fusion

JIANG Shi-hao, QI Su-min, WANG Lai-hua, JIA Hui

(School of Software Engineering, Qufu Normal University, Qufu 273165, China)

**Abstract:**To fully utilize image features to improve the effect of instance segmentation, an instance segmentation model based on Mask R-CNN network structure and multi-feature fusion scheme is proposed. Firstly, two branches are introduced on the basis of Mask R-CNN. One is an edge detection branch based on holistically-nested edge detection (HED) model to generate edge feature graph with emphasis on edge information, the other is a semantic segmentation branch based on fully convolution network (FCN) to generate semantic feature graph with emphasis on rich spatial location information. Secondly, when performing ROIAlign, regions of interest (ROI) are mapped to the corresponding pyramid layer and its adjacent layers to make full use of the information of each layer of the feature pyramid. Finally, the above multiple feature graphs are fused, and the new features with richer information can be generated for subsequent detection and segmentation tasks. Experiment shows that the proposed method effectively improves the accuracy of detection and segmentation. With Resnet50-FPN as the backbone network and no bells and whistles, the box AP is increased by 1.2% and the mask AP is increased by 1.0% compared to Mask R-CNN.

**Key words:**instance segmentation; deep learning; Mask R-CNN; fully convolution networks; feature fusion

## 0 引言

实例分割<sup>[1-2]</sup>是指给定一个图像,在正确检测出图像中所有目标的同时对每一个目标进行分类以及像素级别的分割,是一项具有挑战性的计算机视觉任务,与计算机视觉中的两个经典任务—目标检测<sup>[3-5]</sup>与语义分割<sup>[6-9]</sup>密切相关。实例分割的结果中包含丰富的信息,在自动驾驶、智能监控、生物医疗、人机交互等领域有着极大的利用价值。

2014 年 Hariharan 等人提出了 SDS<sup>[10]</sup>模型,该模型使用多尺度可结合组(MCG)提取建议区域,对于每个区域,使用卷积神经网络(CNN)来提取前景特征,再对每个区域使用支持向量机(SVM)在 CNN 顶层特征上进行分类。2015 年 Pinheiro 等人提出了基于单个卷积网络的 DeepMask<sup>[11]</sup>模型。给定一个图像块作为输入,输出一个与类别无关的掩模和该图像块完全包含一个物体的概率。同年, Dai 等人提出了多任务网

收稿日期:2019-10-17

修回日期:2020-02-25

基金项目:国家自然科学基金(61601261)

作者简介:姜世浩(1994-),男,硕士研究生,研究方向为计算机视觉;齐苏敏,博士,副教授,研究方向为计算机视觉;王来花,博士,副教授,研究方向为信息处理。

络层级模型(MNC)<sup>[12]</sup>。该模型分为实例区分、掩模估计、目标分类三个子任务,在共享特征的基础上,形成层级的多任务结构。2016年Pinheiro等人又提出了SharpMask<sup>[13]</sup>模型,该模型利用底层信息优化了DeepMask的输出,产生具有更高保真度的掩模。上述方法中,分割先于识别,分割结果与目标类别无关,导致结果精度较低。2017年,Y Li等人基于全卷积网络(FCN)<sup>[14]</sup>提出了一种可用于实例分割的端到端模型FCIS<sup>[15]</sup>,该模型是首个全卷积、端到端的实例分割解决方案。2018年,He等人提出了一种简单通用且性能强大的实例分割模型Mask R-CNN<sup>[16]</sup>。该模型在Faster R-CNN<sup>[17]</sup>的基础上加入了一个基于全卷积网络(FCN)的掩模预测分支,并应用了先进的骨干网络—深度残差网络(Resnet)<sup>[18]</sup>与特征金字塔(FPN)<sup>[19]</sup>。此外,提出ROIAlign代替了ROIPooling操作,解决了ROIPooling产生的对齐问题,使该模型能够更好地适应像素级别的分割任务。

Mask R-CNN虽然采用特征金字塔(FPN)结构提取了多层次的丰富特征,但对各层信息尤其是边缘细节信息与空间位置信息的利用仍不充分。文中在Mask R-CNN的基础上提出了多特征融合的实例分割方法。首先,在Mask R-CNN结构中加入一条基于HED<sup>[20]</sup>的边缘检测分支以及一条基于FCN的语义分割分支,分别提供边缘细节信息与空间位置信息;其次,原始特征图在进行ROIAlign时将ROI映射到原先分配的特征层及其相邻层,以充分利用特征金字塔的各层信息;最后,将各分支得到的特征图进行融合,融合得到的新特征中既包含了丰富的边缘信息,能够提升分割结果的边缘精度以及使检测器更好的区分邻接或交错的物体,还包含了目标的空间位置信息,能够将

目标前景与自然界复杂的背景进行有效的区分。实验结果表明,该模型与Mask R-CNN相比检测和分割精度都得到了提升。

## 1 基于Mask R-CNN和多特征融合的实例分割模型

文中提出的网络结构分为三个部分:特征提取网络、区域建议网络以及检测分割网络,整体结构如图1所示。其中,特征提取网络在原有骨干网络(Resnet-FPN)的基础上增加了边缘检测分支与语义分割分支。新增的分支在骨干网络的结果之上进行构建,与原网络联合训练,以重用骨干特征减少额外的参数,如图1左侧虚线框所示。区域建议网络和检测分割网络与Mask R-CNN相同。此外在ROIAlign操作之后加入了特征融合操作,如图1右侧虚线框中所示。

### 1.1 Mask R-CNN 模型

Mask R-CNN模型在Faster R-CNN模型的基础上添加了一个基于FCN的掩模预测分支用于实例分割。如图1所示,输入图像首先通过骨干网络(Resnet-FPN)进行特征提取得到特征图,再通过区域建议网络(RPN)在特征图上生成感兴趣区域(ROI),并将感兴趣区域对应位置的特征池化为固定尺寸的特征,最后由检测分支进行目标框的分类和回归,由掩模预测分支对目标进行像素级别的分割。

在Faster R-CNN中,感兴趣区域进行池化(ROIPooling)时对区域划分坐标进行了取整操作,在结果中引入了量化误差,对像素级别的分割任务影响较大。针对该问题,Mask R-CNN作者提出了ROIAlign,保留了浮点数坐标并通过双线性插值得得各个坐标点的值。

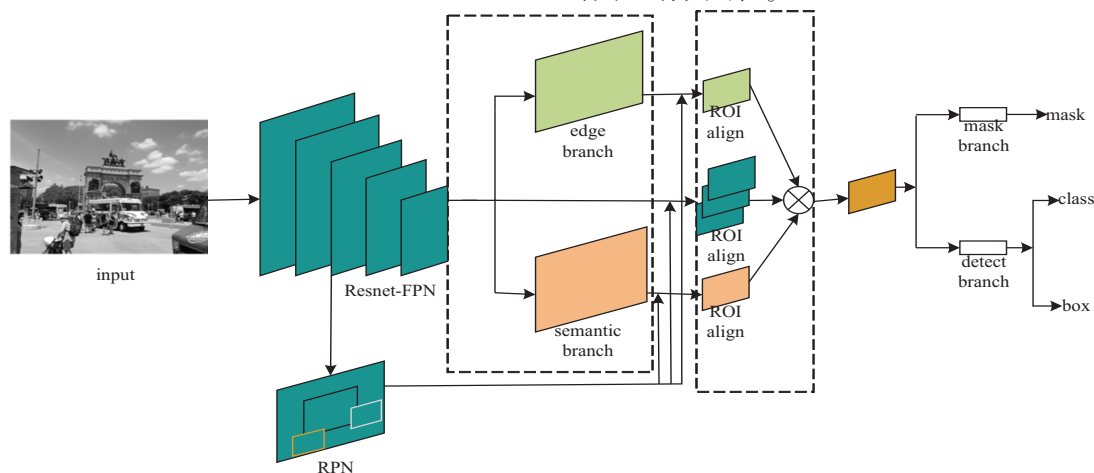


图1 网络结构

Mask R-CNN对于每一个ROI的损失函数定义为:

$$L_{\text{MaskR-CNN}} = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}} \quad (1)$$

式(1)中的 $L_{\text{cls}}$ 为分类损失函数:

$$L_{\text{cls}} = -\log p_u \quad (2)$$

其中, $p_u$ 为目标正确类别 $u$ 的预测概率。

式(1)中的  $L_{\text{box}}$  为边框回归损失函数:

$$L_{\text{box}} = \sum_{i \in \{x, y, w, h\}} \text{smooth}_{\text{L}}(t_i^u - v_i) \quad (3)$$

$$\text{smooth}_{\text{L}}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

其中,  $(t_x^u, t_y^u, t_w^u, t_h^u)$  表示对应真实分类的预测回归参数,  $(v_x, v_y, v_w, v_h)$  表示真实的平移缩放参数。

式(1)中的  $L_{\text{mask}}$  为分割损失:

$$L_{\text{mask}} = \text{BCE}(\hat{y}, y) \quad (5)$$

其中,  $\hat{y}$  为目标区域内像素对应正确类的掩码预测值,  $y$  为目标区域内像素的真实掩码标签, BCE 为式(6)所示的二值交叉熵。Mask 分支分别为每个类别预测掩模,仅在正确类别的掩模上计算误差,目的是避免类间竞争。

$$\text{BCE}(x, y) = -\frac{1}{n} \sum_{i=1}^n y_i \cdot \log x_i + (1 - y_i) \cdot \log(1 - x_i) \quad (6)$$

## 1.2 基于 HED 的边缘检测分支

整体嵌套边缘检测 (HED) 是一种基于深度学习的边缘检测算法<sup>[21-22]</sup>。该模型是一种端到端的边缘

检测模型,应用了多层次,多尺度预测的思想,将网络的不同层级的结果侧向输出并分别应用损失函数进行监督,最后用反卷积将高层输出上采样到原图大小并通过一个可训练的权重将各层的结果融合得到最终输出。

文中模型的边缘检测分支基于 HED 多层次预测的思想构建,以特征金字塔的输出结果作为输入,其结构如图 2 所示。由于特征金字塔的最高层(第五层)分辨率过低,通过反卷积操作上采样得到的结果过于模糊,对检测准确性的提升没有帮助,故舍弃该层,以前四层的输出结果作为输入。首先,每一层通过两个  $3 \times 3$  的卷积层提取各层的边缘信息,并经过一个  $1 \times 1$  的卷积层生成各自的边缘预测结果。然后将各层的预测结果上采样到最底层的大小,并通过一组可训练的权重将各层结果对应元素相加得到最终的边缘预测结果用于该分支的训练。最后,通过一个  $1 \times 1$  的卷积层将该结果映射到原始特征的特征空间得到边缘特征图。

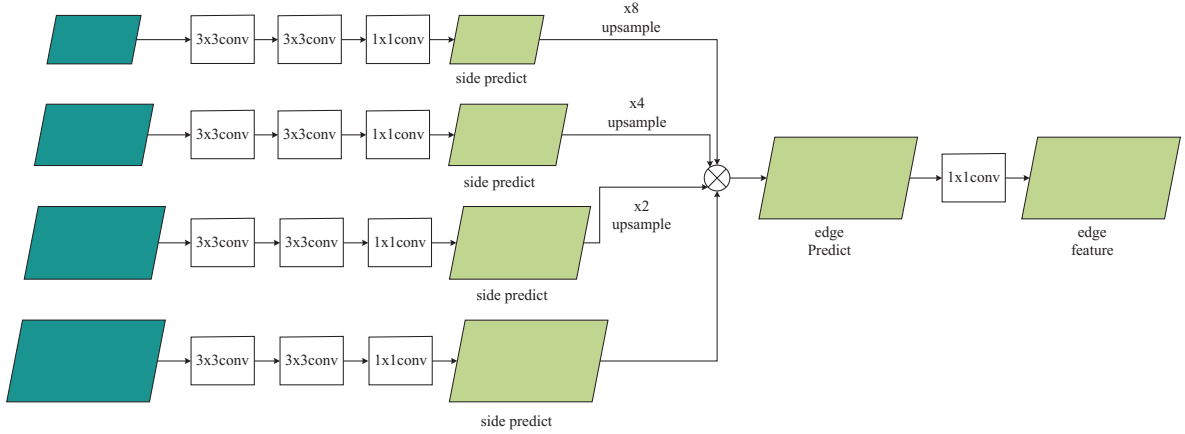


图2 边缘检测分支

在 HED 模型中,对每个侧边输出都应用单独的损失函数进行监督,使得边缘检测网络更容易响应物体内部的边缘纹理。而文中的边缘检测分支主要任务是识别物体的轮廓边缘,物体内部的边缘纹理反而会对目标检测任务造成干扰。因此,文中仅对边缘检测分支的最终融合结果进行监督,其损失函数表示为:

$$L_{\text{edge}} = l(\hat{y}_{\text{fuse}}, y) \quad (7)$$

其中,  $l_{\text{fuse}}(\hat{y}_{\text{fuse}}, y)$  为融合输出概率图  $\hat{y}_{\text{fuse}}$  与图像真实标签  $y$  的误差,表示为如式(8)所示的类别平衡交叉熵。 $\beta = \frac{Y_-}{Y}$ ,  $1 - \beta = \frac{Y_+}{Y}$ ,  $Y_-$  和  $Y_+$  分别表示标签中标记为边缘和非边缘的像素,  $j$  表示图像中的任一像素点。

$$l(\hat{y}, y) = -\beta \sum_{j \in Y_+} \log \hat{y}_j - (1 - \beta) \sum_{j \in Y_-} \log(1 - \hat{y}_j) \quad (8)$$

## 1.3 基于 FCN 的语义分割分支

语义分割分支同样以特征金字塔的各层输出结果为输入,其结构如图 3 所示。首先将每一层特征图分别通过一个  $1 \times 1$  的卷积层使得各层映射到相同的表示空间中,然后以第二层为基准,将高层特征图上采样,低层特征图下采样到第二层大小,并将各层特征图通过对应元素相加进行融合以结合低层特征的定位信息与高层特征的语义信息用于像素级别的语义分割。权衡分割精度与额外参数开销,以第二层为基准是最合适的。最后,将融合得到的特征通过四个  $3 \times 3$  的卷积层进一步获取语义信息,并分别通过两个  $1 \times 1$  的卷积层输出,其中一个输出语义分割的预测结果用于该分支的训练,另一个输出最终的语义特征图用于后续的特征融合。

文中模型的语义分割分支损失函数表示为逐像素

的交叉熵损失:

$$L_{\text{seg}} = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n y_{ji} \log \hat{y}_{ji} \quad (9)$$

其中,  $\hat{y}_{ji}$  与  $y_{ji}$  分别表示图像中像素  $j$  类别为  $i$  的预测概率和真实概率,  $m$ ,  $n$  分别表示图像的像素个数和类别数。

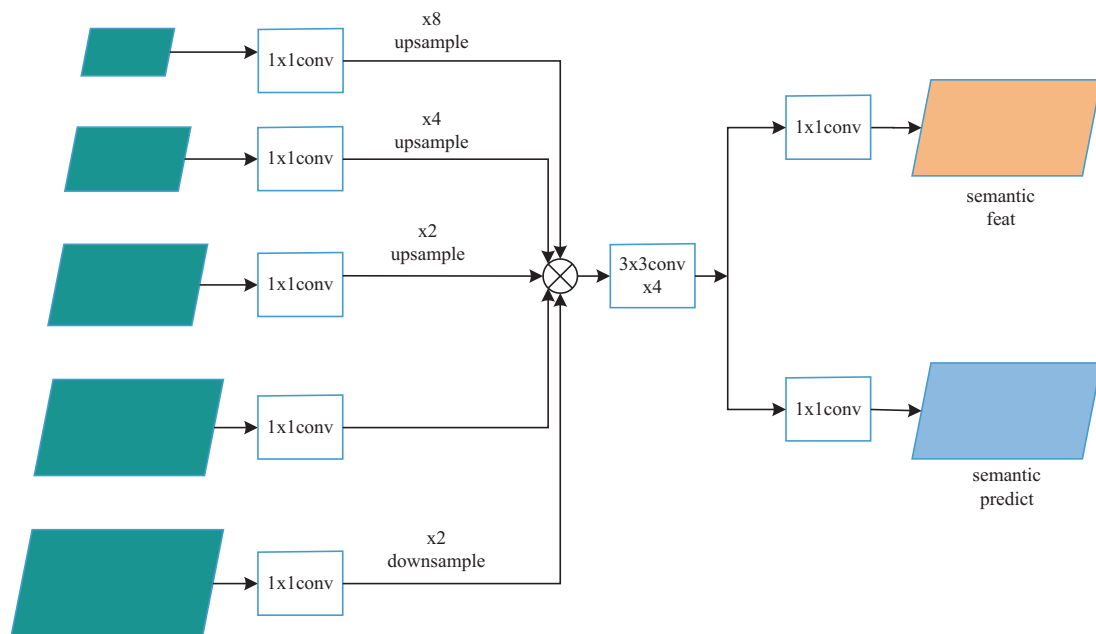


图 3 语义分割分支

#### 1.4 多特征融合

在 Mask R-CNN 中,特征金字塔在进行 ROIAlign 操作时,根据 ROI 的大小对 ROI 进行分配,较大的 ROI 分配到低层,较小的 ROI 分配到高层。在此基础上,文中将 ROI 同时分配给相邻层进行 ROIAlign 操作,得到多个特征图,以充分利用特征金字塔的特征信息。最后,将前述分支得到的边缘和语义特征图经过 ROIAlign 操作,并将所有特征图通过元素相加进行融合,生成信息更加丰富的边框特征和掩模特征,用于后续的检测和分割任务。此外,将所有 ROI 特征图的分辨率提升到  $28 \times 28$  以适应边缘和语义分割的细粒度特征。

#### 1.5 损失函数

整个模型以端到端的方式进行训练,损失函数在原 Mask R-CNN 的损失函数的基础上增加了边缘损失与分割损失用来监督边缘检测分支与语义分割分支的输出结果。整体的损失函数如下:

$$L = L_{\text{MaskR-CNN}} + \alpha L_{\text{edge}} + \beta L_{\text{seg}} \quad (10)$$

其中,  $L_{\text{MaskR-CNN}}$  为 Mask R-CNN 的损失函数,如式(1)所示;  $L_{\text{edge}}$  为边缘检测分支的损失函数,如式(7)所示;  $L_{\text{seg}}$  为语义分割分支的损失函数,如式(9)所示。参数  $\alpha$  和  $\beta$  分别表示边缘误差和语义分割误差在整体误差中的权重系数,通过实验得出,预测精确度对  $\alpha$  和  $\beta$  的变动并不敏感,故这里将  $\alpha$  和  $\beta$  默认设置为 0.5。

## 2 实验结果与分析

文中在 COCO 数据集上对模型进行训练与测试。

首先,利用 COCO2017train 对模型进行训练,然后使用 COCO2017test 和 COCO2017val 对提出的模型进行测试和验证。对于语义分割分支,文中使用 COCO stuff 数据集中的训练标签进行训练。由于 COCO 数据集中并不包含边缘检测的标签信息,所以需要通过 COCO 数据集中的分割标签生成边缘标签。对于每一张图片,首先遍历该图片中所有目标的掩模标签,并将每个掩模赋予不同的非零值后合并成与原图像大小相同的掩模图,未被掩模标记的部分值为 0。然后,基于掩模图判断图中的每个像素是否为边缘,若一个像素相邻的四个像素(上,下,左,右)不为同一个值,则将该像素记为边缘。对数据的其他处理与 Mask R-CNN 模型相同。文中以不同的 IOU 阈值(0.50 ~ 0.95,步长为 0.05)及不同大小目标的平均精度(mAP)作为评价标准。

使用 2 个 Tesla P100 GPU 进行实验,实验模型使用 Pytorch 进行搭建。训练时使用随机梯度下降(SGD)对模型进行优化,初始学习率设置为 0.005,动量设置为 0.9,权重衰减系数设置为 0.000 1,共迭代 12 次,学习率在第 8 和第 11 次迭代时降低为原来的 0.1 倍。用于对照的基线模型使用官方开源代码在相同的实验环境下运行,训练参数与官方代码相同。文中模型与 Mask R-CNN 在实验中均使用 Resnet-50-FPN 作为骨干网络。

为了进一步验证文中提出模型的有效性,将其与 Mask R-CNN 模型进行了比较,图 4 展示了文中方法

与 Mask R-CNN 分割效果的对比。可以看出,文中方法的分割结果边缘上与目标更加贴合,缺失和冗余更少,例如在对照组(a)中,与文中模型的长颈鹿分割结果相比,Mask R-CNN 的长颈鹿分割结果在足部边缘处有多处明显的缺失,且脖颈处有较明显的冗余。此

外,文中模型的分割结果对相邻目标的边界区分更加清晰,例如在对照组(d)中,Mask R-CNN 的分割结果与文中模型相比,在左边人物的肩膀和右边人物手臂的交界处,两个目标的像素出现了严重的重叠,无法辨别出明显的边界。

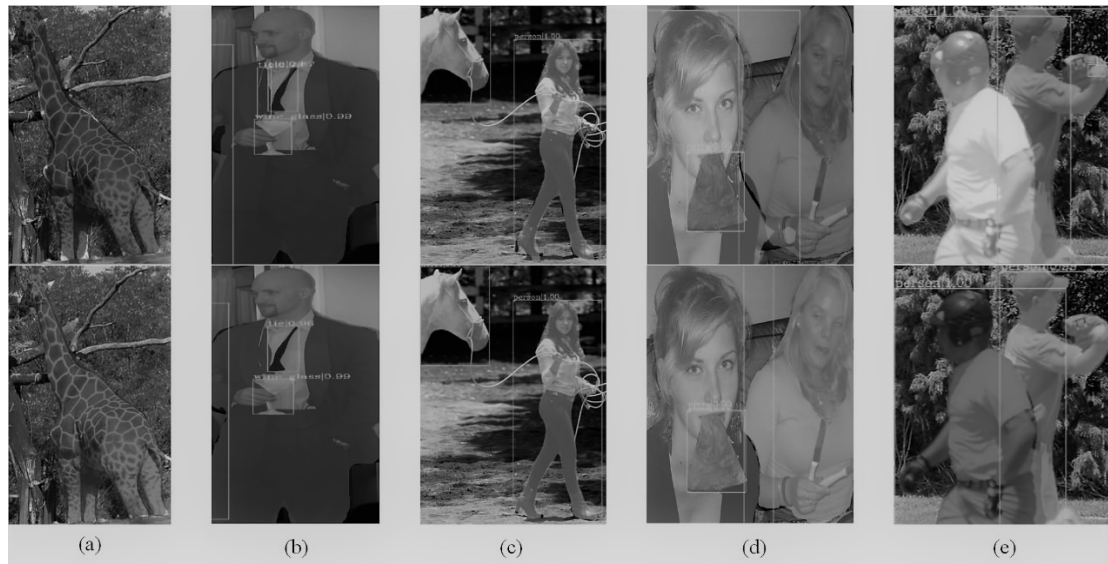


图4 Mask R-CNN(上)与文中模型(下)的分割效果对比

Mask R-CNN 与文中模型检测与分割的评价结果分别如表1和表2所示。可以看出,文中模型相比 Mask R-CNN 在检测和分割的精度上都得到了提升, bbox 与 mask 的 mAP 分别提升了 1.2% 与 1.0%。其中对于大物体的分割精度提升最为显著, mAP<sub>L</sub> 与 Mask R-CNN 相比提升了 1.6%, 但对于小物体的检测与分割精度提升较低。对该现象进行了分析,认为大物体所包含的边缘轮廓特征较为丰富,且 COCO 数据

集中大物体的分割标签在边缘细节上刻画得比较细致,因此边缘检测分支更容易在大物体中提取到丰富的边缘信息,从而能够在大物体的分割任务上取得较大的提升。而小物体的边缘轮廓较为模糊,且 COCO 数据集中小物体的分割标签与大物体相比较为粗糙,因此边缘检测分支和语义分割分支在小物体上无法提取到更多的特征,故模型在小物体上的表现与原 Mask R-CNN 相比提升较小。

表1 COCO2017 检测结果比较

Method	mAP <sub>bbox</sub>	mAP <sub>50</sub>	mAP <sub>75</sub>	mAP <sub>S</sub>	mAP <sub>M</sub>	mAP <sub>L</sub>
Mask R-CNN	37.3	58.9	40.5	22.3	40.6	48.1
文中方法	38.5	60.9	41.6	22.4	42.3	49.2

表2 COCO2017 分割结果比较

Method	mAP <sub>mask</sub>	mAP <sub>50</sub>	mAP <sub>75</sub>	mAP <sub>S</sub>	mAP <sub>M</sub>	mAP <sub>L</sub>
Mask R-CNN	34.3	55.7	36.3	18.5	37.4	46.6
文中方法	35.3	57.4	37.2	18.5	38.5	48.2

### 3 结束语

基于 Mask R-CNN 提出了一种多特征融合的实例分割方案。该方法在 Mask R-CNN 的基础上加入了边缘检测和语义分割分支,分别用于提取带有更丰富边缘信息和语义信息的特征图,并融合特征金字塔的多级特征得到包含更多信息的新特征用于检测和分割任务,提高了检测和分割的精度。实验结果表明,在 COCO 数据集上,与 Mask R-CNN 相比文中模型的 box

AP 提升了 1.2%, mask AP 提升了 1.0%。该模型对小目标的检测和分割精度提升较小,在今后的工作中将继续探究并加以改进。

### 参考文献:

- [1] ROMERA-PAREDES B, TORR P H S. Recurrent instance segmentation[C]//European conference on computer vision. Amsterdam: Springer International Publishing, 2016: 312 - 329.

- [2] 邓疏元,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.
- [3] 吕培建,陈佳鹏,袁飞,等. 基于上下文以及多尺度信息融合的目标检测算法[J]. 计算机科学,2019,46(6A):279-283.
- [4] 施泽浩,赵启军. 基于全卷积网络的目标检测算法[J]. 计算机技术与发展,2018,28(5):55-58.
- [5] GIRSHICK R. Fast R-CNN[C]//International conference on computer vision. Santiago; IEEE,2015:1440-1448.
- [6] 田莹,王亮,丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报,2019,30(2):440-468.
- [7] 代具亨,汤心溢,刘鹏. 基于深度学习的语义分割网络[J]. 红外,2018,39(4):33-38.
- [8] 刘丹,刘学军,王美珍. 一种多尺度 CNN 的图像语义分割算法[J]. 遥感信息,2017,32(1):57-64.
- [9] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Washington, DC; IEEE,2014:580-587.
- [10] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Simultaneous detection and segmentation[C]//European conference on computer vision. Zurich, Switzerland; Springer,2014:297-312.
- [11] PINHEIRO P O, COLLOBERT R, DOLLAR P. Learning to segment object candidates[C]//Advances in neural information processing systems. Montreal, Canada; [s. l.], 2015:1990-1998.
- [12] DAIJ F, HE K M, SUN J. Instance-aware semantic segmentation via multi-task network cascades[C]//Computer vision and pattern recognition. [s. l.]; IEEE,2016:3150.
- [13] PINHEIRO P O, LIN T, COLLOBERT R, et al. Learning to refine object segments[C]//European conference on computer vision. [s. l.]; Springer International Publishing,2016:75.
- [14] LONG J, SHELHAMER E, DARR ELL T. Fully convolutional networks for semantic segmentation[C]//IEEE conference on computer vision and pattern recognition. [s. l.]; IEEE,2015:3431-3440.
- [15] LI Y, QI H Z, DAI J F, et al. Fully convolutional instance-aware semantic segmentation[C]//IEEE conference on computer vision and pattern recognition. Honolulu; IEEE,2017:2359-2367.
- [16] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[C]//Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). [s. l.]; IEEE,2017:2961-2969.
- [17] 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.
- [18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//IEEE conference on computer vision and pattern recognition. [s. l.]; IEEE,2015:770-778.
- [19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE conference on computer vision and pattern recognition. Honolulu; IEEE,2017:2117-2125.
- [20] XIE S N, TU Z W. Holistically-nested edge detection[J]. International Journal of Computer Vision,2015,125(1-3):3-18.
- [21] SHEN W, WANG X G, WANG Y, et al. Deep contour: a deep convolutional feature learned by positive-sharing loss for contour detection[C]//Computer vision and pattern recognition. [s. l.]; IEEE,2015:3982-3991.
- [22] BERTASIUS G, SHI J B, TORRESANI L. High-for-low and low-for-high: efficient boundary detection from deep object features and its applications to high-level vision[C]//IEEE international conference on computer vision. Santiago, Chile; IEEE,2015:504-512.