基于改进特征金字塔的 Mask R-CNN 目标检测方法

任之俊，蔺素珍*，李大威，王丽芳，左健宏
中北大学大数据学院，山西 太原 030051

摘要 提出了一种基于改进特征金字塔的 Mask R-CNN 目标检测方法。实验结果表明，在目标边缘和包围盒两项检测中，相比于 Mask R-CNN 检测框架，所提方法在不同的交并比阈值下的平均准确率分别提高了约 2.4% 和 3.8%。尤其对于中等尺寸目标的检测准确率有较大的提高，分别为 7.7% 和 8.5%，具有较强的稳健性。

关键词 机器视觉；模式识别；目标检测；卷积神经网络；特征金字塔
中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.041502

Mask R-CNN Object Detection Method Based on Improved Feature Pyramid

Ren Zhijun, Lin Suzhen*, Li Dawei, Wang Lifang, Zuo Jianhong
College of Big Data, North University of China, Taiyuan, Shanxi 030051, *China*

Abstract The Mask R-CNN (mask region-based convolutional neural network) object detection method is proposed based on the improved feature pyramid. The experimental results show that compared with the Mask R-CNN detection structure, the mean average precision (mAP) under different Intersection-over-Union (IoU) thresholds increases by 2.4% and 3.8% in the detection of object edge and bounding box, respectively. In particular, the detection accuracy of medium size objects is greatly improved by 7.7% and 8.5%, respectively, which indicates strong robustness.

Key words machine vision; pattern recognition; object detection; convolutional neural network; feature pyramid
OCIS codes 150.0155; 100.5010; 100.4996

1 引 言

目标检测广泛应用于智能监控、自动驾驶、人机交互等领域，其任务是从复杂场景中标识出目标的分类信息与位置信息，用于后续的跟踪[1]、识别[2-3]以及更为复杂的任务。并且，目标检测需要同时解决分类和定位的问题，还要注意目标数量和目标尺寸。因此，目标检测一直是计算机视觉领域研究的热点和难点。

传统的目标检测方法，如尺度不变特征变换(SIFT)[4]、梯度方向直方图(HOG)[5]及可变形组件模型(DPM)[6]等，根据先验知识设计特征，虽在特定的场景达到了较高的检测速度与精度，但由于该类方法依赖先验知识，导致自适应性及泛化性较差。近年来，基于深度学习机制的目标检测方法能自适应提取目标不同层级的特征，并将训练好的模型应用在不同的场景中，有效提高了检测精度及泛化性。基于深度学习的目标检测模型根据分类回归与区域提取是否分开分为以下两类。1)基于回归的目标检测模型，根据特征映射图预先划定默认框，进而对目标分类。典型方法有：YOLO[7]、SSD[8]及 YOLOv3[9]，上述算法采用回归的思想提取边界回归框，极大地提高了检测速度，但检测精度较差[10]；2)基于区域候选的目标检测模型，先对特征映射图进行边界框提取，再将其输出与特征映射图一同输入至兴趣区域(RoI)池化层，以实现对目标的分类与定位，该类方法是近几年的研究热点。从 R-CNN[11]首次将深度学习机制引入目标检测领域

收稿日期：2018-08-27；**修回日期**：2018-08-30；**录用日期**：2018-09-04
基金项目：山西省应用基础研究项目(201701D121062)
* **E-mail**：lsz@nuc.edu.cn

实现了目标自适应检测起，诸多研究者对其进行改进，如：SPP-Net[12]在 R-CNN 中引入空间金字塔池化层，在减少输入图像尺寸对网络影响的同时提升了检测精度；Fast R-CNN[13]进一步在 SPP-Net 的空间金字塔池化层的基础上，采用单尺度池化，极大提高了检测速度；Faster R-CNN[14]又在 Fast R-CNN 提取候选区域过程中引入区域建议网络(RPN)，实现了端到端的训练，提高了区域提取的精度及网络训练速度，是当前应用较广的检测框架[15]；Mask R-CNN[16]将 Faster R-CNN 的 RoI 池化层改进为 RoIAlign，并采用双线性插值法降低了边界回归框的位置误差，同时加入了掩模生成任务，一定程度上提高了检测的精度。Mask R-CNN 尽管嵌入了特征金字塔网络(FPN)[17]，可以学习到丰富的特征，但由于边界回归框尺寸的限制，只能利用少数几层特征映射图的信息，这就会不可避免地导致其他层次可利用信息的丢失，进而影响后续的分类回归任务。

为此，本文提出了基于改进特征金字塔的 Mask R-CNN 目标检测方法。采用填零扩充对原图像进行预处理，在 FPN 中增加自下向上的反向侧边连接路径，再对所有层次的特征映射图进行自上而下的上采样侧边连接，将连接后的特征映射图分别输入 RPN 和 RoIAlign，RPN 对其提取边界框再输入 RoIAlign，对 RoIAlign 的输出进行分类和回归，以有效提取出目标的空间位置信息，最终达到提高检测准确率的目的。

2 基本原理

2.1 方法框架

1) 数据预处理：对输入图像四周进行填零补足，因所使用的数据集[18]中的图像宽高最大为 640 pixel，所以为方便之后的网络处理，将图像扩充为 1024 pixel×1024 pixel。

2) 特征提取：采用改进的 FPN 进行特征提取，对 FPN 增加自下而上的反向侧边连接并融合特征映射图，输入至 RPN 和 RoIAlign。

3) RPN 和 RoIAlign 分类回归：利用 RPN 对特征映射图进行边界框提取并映射到特征映射图上，输入 RoIAlign，根据损失函数对输出结果进行分类回归。

4) 修改与完善：网络完成目标检测后，根据评价指标分析各因素对检测效果的影响，进一步对网络进行修改与完善。

总体框架如图 1 所示。

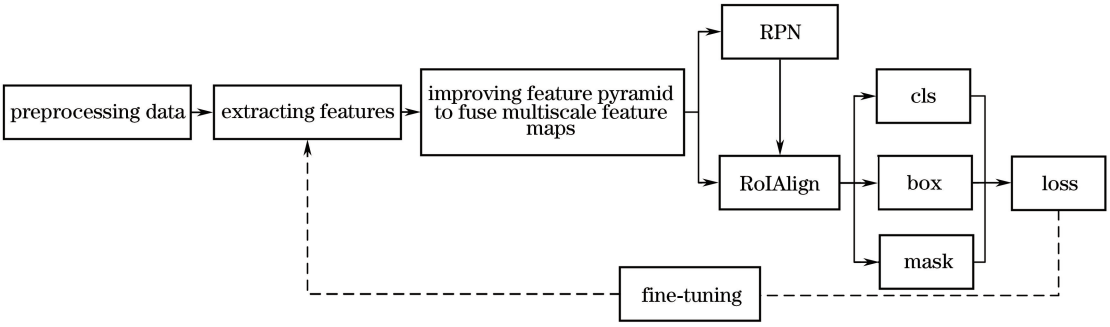

图 1 检测方法框架

Fig. 1 Flow chart of detection method

2.2 基于改进特征金字塔的 Mask R-CNN 目标检测方法

2.2.1 Mask R-CNN 基本原理及流程

Mask R-CNN 属于 R-CNN 系列检测框架，在 Faster R-CNN 的基础上添加一个掩模预测分支，并将 FPN 结合到 ResNet 中，改进 RoIPooling 层为 RoIAlign 层，在预测框提取过程中使用双线性插值法代替了原方法中简单的四舍五入取整，具体步骤如图 2 所示。图中，C_i($2\leqslant i\leqslant 5$)为共享卷积层第 i 阶段的特征映射图；P_j($2\leqslant j\leqslant 4$)为 FPN 由 C_i($2\leqslant i\leqslant 4$)及 P_{j+1}($j=i$)经过侧边连接生成的第 j 阶段特征映射图；由于尺寸问题，P_5 为直接由 C_5 经过卷积操作生成的特征映射图，且并未改变尺寸。先将图像输入至共享卷积层提取特征，生成多尺度特征映射图，再进行侧边连接，将每一阶段的特征映射图二倍上采样后与相邻低层进行张量相加，利用 RPN 对不同尺寸的特征映射图生成候选区域，并将其与特征映射图输入 RoIAlign 得到预测框，最后，对预测框进行分类和回归。但是，由于 FPN 只使用了自上而下路径，对于多层特征映射图的利用并不充分，尤其是对于高层特征映射图处理时会造成部分信息丢失，无法达到更好的检测效果。

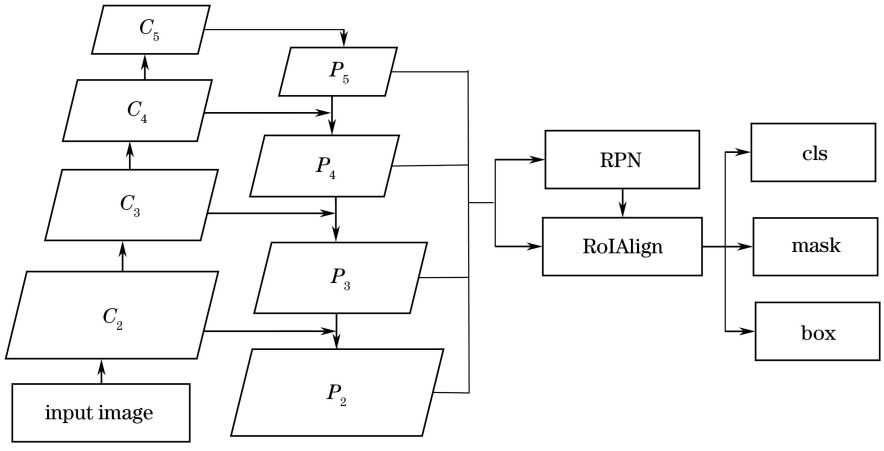

图 2 Mask R-CNN 检测流程图

Fig. 2 Flow chart of Mask R-CNN detection

2.2.2 改进的 Mask R-CNN 检测流程

Mask R-CNN 框架结合的 FPN 对于多尺度的特征映射图采用了侧边连接的方法，将高层语义信息融合进低层精确的定位信息中，在实验结果中有良好表现[16]。但是，从 FPN 结构分析来看，虽然它利用了多尺度的信息，但是文中的侧边连接方法只有自上而下的路径，而且对于 RPN 的输入是在这一组特征映射图中选取单一尺寸进行处理。这样会导致两个问题：其一，最高层的特征映射图与原特征提取网络[17]结构的最终输出是一样的，而大尺寸[18]目标的信息主要由此层特征映射图提供，所以对于大目标检测的准确率与原网络相近甚至略低；其二，对自上而下路径结构进行分析，可以知道对于 FPN 输出的一组特征映射图中，每一层包含本层和更高层的信息而不包含更低层的信息，而对 RPN 又是从中选取最优尺寸特征映射图进行输入，这样就会导致无法充分利用所有尺寸特征映射图的信息，造成检测准确率并非更优值。

为解决现有方法所存在的问题，提出了改进的 Mask R-CNN 目标检测方法，其整体目标检测框架流程如图 3 所示。其中，C_i（$2 \leqslant i \leqslant 5$）、$P_j$（$2 \leqslant j \leqslant 5$）、$N_l$（$2 \leqslant l \leqslant 5$）和 M_k（$2 \leqslant k \leqslant 5$）分别为共享卷积层、FPN 及所提方法生成的特征映射图。检测方法的步骤如下。

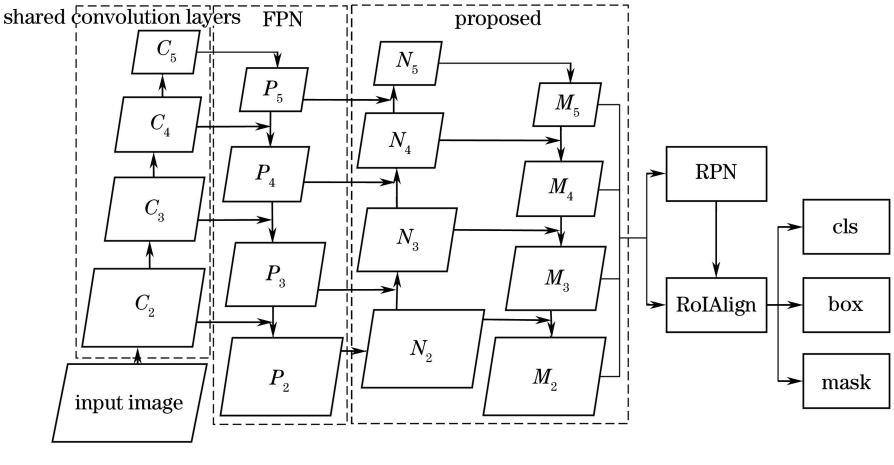

图 3 改进后的检测框架流程图

Fig. 3 Flow chart of improved detection framework

1）自下而上的反向侧边连接。对 FPN 生成不同尺寸的特征映射图，增加一条自下而上的路径，如图 3 中 $N_2 \sim N_5$ 所示。具体步骤如图 4 所示，N_2 与 P_2 尺寸相同，N_l（$2 \leqslant l \leqslant 4$）经步长为 2 的 3×3 卷积，得到与 P_{j+1}（$2 \leqslant j \leqslant 4$）相同尺寸的特征映射图并与 P_{j+1}（$2 \leqslant j \leqslant 4$）进行相加，再进行卷积操作得到 N_{l+1}（$2 \leqslant l \leqslant 4$）。上述卷积操作的卷积核数量均为 256。

2）融合多尺度特征映射图。根据图 5 所示，具体步骤为：N_5 经 1×1 卷积得到 M_5，将 M_k（$3 \leqslant k \leqslant$

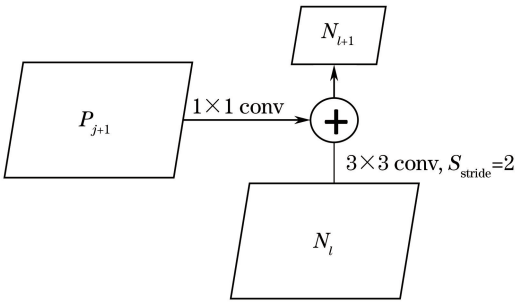

图 4 反向侧边连接具体步骤

Fig. 4 Concrete steps of reverse lateral connection

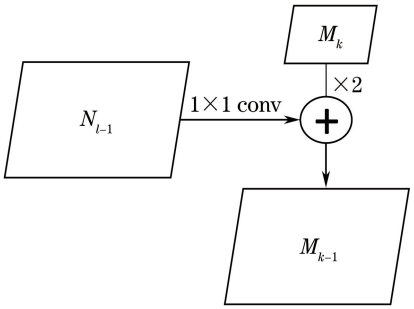

图 5 融合特征映射图

Fig. 5 Fused feature map

5)进行二倍上采样得到与 N_{l-1}($l=k$)尺寸相同的特征映射图并与 N_{l-1} 相加,再经 3×3 卷积得 M_{k-1}($3\leqslant k\leqslant 5$),再采用 sigmoid 激活函数得到 RPN 的输入[16]。

3 实验结果与分析

实验采用公开的 coco 数据集[18]进行训练与测试,将经典的 ResNet-101[19]作为特征提取的骨架网络,并利用 coco 数据集及 imagenet 数据集预训练该特征提取网络。采用的评价指标为不同 T_{IoU} 阈值($0.50\leqslant T_{IoU}\leqslant 0.95$)及不同尺寸目标下所有类别的平均准确率均值(mAP)。实验中,共享卷积层经过了 imagenet 数据集的预训练,根据实验测试对网络学习率从 0.02 调整为 0.001,并对所改进部分单独训练,然后将学习率调整为 0.0001,从而对网络进行整体的微调。

如表 1 所示,P_{AP} 是以 0.05 为步长从 0.50 至 0.95共 10 个 T_{IoU} 阈值下的平均准确率均值。同理,P_{AP50} 和 P_{AP75} 分别是 T_{IoU} 阈值为 0.50 和 0.75 时的平均准确率均值。P_{APs}、P_{APm}、P_{APl} 分别为小、中、大不同尺寸目标的平均准确率均值。MNC 和 FCIS 为 instance segmentation 任务中的经典模型。分析表 1 可知,由于采用反向侧边连接的方法生成特征映射图,包含了低层准确的定位信息和高层语义信息,避免了由于 Mask R-CNN 中 FPN 只有一条自上而下的侧边连接路径导致的高层特征映射图无法有效包含低层定位信息的问题,将中、小尺寸目标的识别准确率提高了 4%。并且,所提方法对改进 FPN 的中间层融合了更多的信息,在 P_{APm} 上提升了 7.7%。

如表 2 所示,P_{APbb} 为目标包围盒检测的平均准确率均值,根据不同阈值的 T_{IoU} 进行区分。对结果进行分析可知,相对于 Mask R-CNN 的结果,在不同的 T_{IoU} 阈值中,所提方法在 P_{AP75bb} 中的提高更为显著,为 3.9%,说明在所检测到的目标包围盒中,本方法的结果更精确。P_{APmbb} 的结果提高了 8.5%,显著高于另外两项的提升。根据分析结果和网络结构可知,FPN 和所添加的自下而上路径对于中间两层的信息既融合了高层语义信息,又融合了低层准确的定位信息。所以相对于最高层和最低层中主要判断大目标和小目标的信息,中间层对于中等目标的检测准确率更高。

表 1 Instance segmentation 的 mAP 结果比较

Table 1 Comparison of mAP results in instance segmentation %

Method	Backbone	P_{AP}	P_{AP50}	P_{AP75}	P_{APs}	P_{APm}	P_{APl}
MNC[20]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS[21]+OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Proposed	ResNet-101-improved FPN	37.2	58.2	39.4	18.4	45.8	52.7

表 2 Bounding box 的 mAP 结果比较

Table 2 Comparison of mAP results in bounding box %

Method	Backbone	P_{APbb}	P_{AP50bb}	P_{AP75bb}	P_{APsbb}	P_{APmbb}	P_{APlbb}
Faster R-CNN	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN with FPN	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Proposed	ResNet-101-improved FPN	42.3	61.4	45.6	24.2	49.6	51.3

4 结　论

提出了基于改进特征金字塔的 Mask R-CNN 目标检测方法，针对 Mask R-CNN 在 FPN 提取特征阶段无法充分利用所有尺度特征映射图信息的问题，结合 FPN 和反向侧边连接，融合连接后的特征映射图。通过在 coco 公开数据集上的训练和测试，实验结果表明相比较于原 Mask R-CNN 检测框架，所提方法对于不同 T_{IoU} 阈值下的 mAP 在目标边界和包围盒两项检测中分别提高了 2.4% 和 3.8%，尤其对于中等尺寸目标的检测准确率提高较多，分别为 7.7% 和 8.5%。

所提方法对于大目标的检测准确率和 P_{AP50} 指标提高幅度较小，原因是在使用多层特征映射图时将无用的冗余信息进行处理，对结果造成了一定的影响。下一步工作是在改进的 FPN 基础上对冗余信息进行排除，以达到更好的检测效果。

参 考 文 献

[1] Lin S Z, Zheng Y, Lu X F, *et al*. Adaptive tracking algorithm for aerial small targets based on multi-domain convolutional neural networks and autoregression model[J]. Acta Optica Sinica, 2017, 37(12): 1215006.
蔺素珍, 郑瑶, 禄晓飞, 等. 基于多域卷积神经网络与自回归模型的空中小目标自适应跟踪方法[J]. 光学学报, 2017, 37(12): 1215006.

[2] Liu F, Shen T S, Ma X X. Convolutional neural network based multi-band ship target recognition with feature fusion[J]. Acta Optica Sinica, 2017, 37(10): 1015002.
刘峰, 沈同圣, 马新星. 特征融合的卷积神经网络多波段舰船目标识别[J]. 光学学报, 2017, 37(10): 1015002.

[3] He Z C, Zhao L Z, Chen C. Convolution neural network with multi-resolution feature fusion for facial expression recognition[J]. Laser & Optoelectronics Progress, 2018, 55(7): 071503.
何志超, 赵龙章, 陈闯. 用于人脸表情识别的多分辨率特征融合卷积神经网络[J]. 激光与光电子学进展, 2018, 55(7): 071503.

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

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C] // IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2005, San Diego, CA, USA. IEEE: New York, 2005: 886-893.

[6] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C] // IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE, 2008: 1-8.

[7] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection [C] // IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[8] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C] // Leibe B, Matas J, Sebe N, *et al*. European Conference on Computer Vision, Cham: Springer, 2016, 9905: 21-37.

[9] Redmon J, Farhadi A. Yolov3: an incremental improvement[EB/OL]. (2018-04-08)[2018-07-31]. https://arxiv.org/abs/1804.02767.

[10] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection [C] // IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2999-3007.

[11] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[12] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[13] Girshick R. Fast R-CNN [C] // IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[14] 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.

[15] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved faster R-CNN[J]. Acta Optica

Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[16] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN [C] // 2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 2980-2988.

[17] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection [C] // IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[18] Lin T Y, Maire M, Belongie S, *et al*. Microsoft coco: common objects in context [C] // Fleet D, Pajdla T, Schiele B, *et al*. European Conference on Computer Vision, Cham: Springer, 2014, 8693: 740-755.

[19] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[20] Dai J F, He K M, Sun J. Instance-aware semantic segmentation via multi-task network cascades[C] // IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 3150-3158.

[21] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 4438-4446.