计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191200063

卷积神经网络低层特征辅助的图像实例分割方法

樊 玮[1] 刘 挺[1] 黄 睿[1] 郭 青[2] 张 宝[2]
1 中国民航大学计算机科学与技术学院 天津 300300
2 天津大学智能与计算学部 天津 300350
(wfancauc@163.com)

摘 要 流行的实例分割网络 Mask R-CNN 在进行实例分割时,存在目标分割边界和分割轮廓粗糙的问题,导致分割精度低。针对此问题,提出在 Mask R-CNN 分割分支中引入网络的低层卷积特征进行高精度的实例分割方法。首先从特征提取网络中选择特征,通过插值算法将其缩放至固定尺度(输入图像的 1/8)作为低层特征;然后通过 RoI 对齐操作提取当前待分割目标的特征后与原始的 Mask R-CNN 的分割分支对应目标的特征进行拼接,并将其作为精细化目标分割的特征。低层网络特征引入了更多的低级纹理和轮廓信息,可以有效地提高物体的分割精度。在 COCO2017 数据集上,所提方法使用 ResNet-101-FPN 作为特征提取网络得到的分割结果的平均准确度(AP)相对于 Mask R-CNN 提高了 1.2%。实验结果表明,所提方法在使用不同特征提取网络时具有较好的鲁棒性和有效性。

关键词:深度学习;深度神经网络;实例分割;特征融合;低层特征
中图法分类号 TP391.4

Low-level CNN Feature Aided Image Instance Segmentation

FAN Wei[1], LIU Ting[1], HUANG Rui[1], GUO Qing[2] and ZHANG Bao[2]
1 College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China
2 College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

Abstract The popular instance segmentation network, Mask R-CNN, has rough target segmentation boundaries and segmentation contours when performing instance segmentation, which leads to low segmentation accuracy. To solve this problem, a high-precision instance segmentation method is proposed by introducing the low-level features of the network into the segmentation branch of Mask R-CNN. Specifically, it selects the convolutional features from lower layers of feature extraction network at first. And then, it resizes the features to a fixed scale (1/8 of the input image) by interpolation algorithm to form the low-level features. It concatenates the features of original segmentation branch of Mask R-CNN with the features extracted by RoI Align operation from low-level features for current target. Since low-level features introduce more low-level texture and contour information, it can effectively improve the accuracy of instance segmentation. Compared with Mask R-CNN, the proposed method obtains 1.2% relative average precision (AP) improvement on the COCO2017 dataset by using ResNet-101-FPN as the feature extraction network. Experimental results show that the proposed method is robust and effective when using different feature extraction networks.

Keywords Deep learning, Deep neural network, Instance segmentation, Feature fusion, Low-level feature

1 引言

实例分割结合了目标检测和语义分割,是一个基础且富有挑战的计算机视觉问题[1]。实例分割在对图像中所有关注实例进行检测的同时需要对实例进行准确的分割。实例检测结果使用边界框(Bounding Box)标记,实例分割结果使用掩膜(Mask)标记。实例分割由于可以同时获得检测和分割结果,因此被用于图像语义理解[2-3]、自动驾驶[4-5]、视频监控[6]、跟踪[7-8]及图像搜索[9-10]等领域。

实例分割与目标检测、语义分割有着密切的联系[11-12]。根据解决思路的不同,目前基于深度神经网络的实例分割方法可以分为两种类型:基于检测的方法和基于分割的方法。

基于检测的方法的主要思想是先检测到图像中的所有实例,然后对每个实例进行分割。早期的研究使用滑动窗口的

到稿日期:2019-12-07 返修日期:2020-05-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:天津市教委科研计划项目(2019KJ126);中国民航大学中央高校基金项目(3122018C021,3122018C020)
This work was supported by the Scientific Research Project of Tianjin Education Commission (2019KJ126) and Foundation Project for Central University of CAUC(3122018C021,3122018C020).
通信作者:黄睿(rhuang@cauc.edu.cn)

方法得到实例的可能区域，然后计算每个区域是物体的可能性，并对其进行分割。具有代表性的工作有 DeepMask[13]，SharpMask[14]，InstanceFCN[15]，MNC[16]，BAIS[17] 等。DeepMask[13] 首先采用滑动窗口方式得到目标区域，然后设计两个分支分别用于实例分割和目标类别的预测。不同于 DeepMask，SharpMask[14] 改变了实例分割分支的网络结构，通过自上而下的结构来提升 Mask 预测的效果。InstanceFCN[15] 针对每个目标区域输出多个位置敏感特征图，然后组合不同的特征图生成最终的 Mask 预测结果。MNC[16] 将实例分割分解为实例定位、类别预测和 Mask 估计 3 个子任务，并将这 3 个子任务串联起来形成多级任务级联结构的网络，取得了较好的效果。BAIS[17] 针对预测的边界框进行多尺度的反卷积操作，以解决由于边界框不准确造成的 Mask 不完整的问题。但是这种方法在实例被遮挡而分为多部分时无法取得较好的效果。

为了降低使用滑动窗口引入的巨大的计算量，近几年，基于检测的方法选择先进的目标检测器先预测出目标的定位边界框，然后依据边界框预测相应目标的 Mask。具有代表性的检测器有 R-FCN[18]，Fast R-CNN[19] 和 Faster R-CNN[20] 等。FCIS[21] 在 InstanceFCN[15] 的基础上提出了内外位置敏感得分图（Position-sensitive Inside/outside Score Maps），内位置得分图和外位置得分图分别对应感兴趣区域（Region of Interest，RoI）中前景和背景的语义信息。位置敏感得分图被不同的感兴趣区域所共享，虽然能够区分每一个感兴趣区域中的前景和背景，但增加了输出得分图的通道数量。MaskLab[22] 基于 Faster R-CNN，引入了语义分割分支，能够同时得到边界框检测、语义分割、方向分割 3 个输出。其中，方向分割分支也采用了类似 FCIS 中得分图的方式，组合不同的方向通道的信息得到方向预测，能够区分边界框中同一语义类别（类不可知）的不同实例。Mask R-CNN[23] 扩展了 Faster R-CN，是一个灵活、简单的网络框架，近期提出的很方多法都基于 Mask R-CNN。如 MS R-CNN[24] 从直接预测 Mask 质量分数的角度解决分类分数和 Mask 质量不配准的问题。HTC[25] 将 Mask R-CNN 与 Cascade R-CNN[26] 结合形成级联结构，并在此基础上提出增强结构，逐步改善边界框预测和 Mask 分割的效果。PANet[1] 和 Sun 等[27] 提出的方法与本文的方法类似，二者均在 FPN[28] 的基础上增加了自下而上的特征融合结构来提升实例分割精度。PANet 通过改进的特征金字塔结构、Adaptive Feature Pooling 和全连接融合结构加强了低层特征和 Mask 分支特征。与 PANet 不同，本文方法将特征提取网络中的低层特征与 Mask 分支末端 RoI 特征相融合，提高了 RoI 特征的表达能力，改善了目标分割的质量。

与基于检测的方法不同，基于分割的方法首先得到全图的像素级的分割，然后通过聚类把相同类别的像素聚集在一起产生目标实例[22,24-25]。Liang 等[29] 提出的方法使用了现成的聚类算法，利用语义分割结果、图中实例的数目和图中每一个实例位置坐标聚类得到最终的实例分割结果。Bai 等[30] 将经典的分水岭变换方法[31] 与深度神经网络相结合产生图像的能量图，每个实例在能量图中以一个能量凹陷区域表示，然后用单一水平级切割能量图得到每个连续的实例。Kirillov 等[32] 首先使用两个分支分别输出实例不可知的语义分割结果和实例的边缘检测结果，然后使用超像素对边缘检测结果进行分割，并将其分割结果和语义分割结果合并得到实例分割结果。Jin 等[33] 提出的方法使用语义分割分支和标签转换分支，通过融合两个分支的输出结果得到最终的实例分割。Liu 等[34] 提出以多个级联的网络逐步解决实例分割子聚类问题，从预测断点到生成线段，线段聚集成面，最后聚集为每个实例。另外，Ren 等[35] 和 Romera-Paredes 等[36] 引入循环神经网络（RNN）[37]，顺序寻找目标并一次仅分割一个目标。然而，前者在结构中采用了多个 ConvLSTM[38] 单元来得到多个目标分割结果，后者采用 end-to-end 的方式，除了使用 LSTM[37] 模块还将整个网络搭建为一个循环式架构，在每一次循环中得到一个分割结果。

深度神经网络高层学到的是能够区分类别、属性等语义的特征，而网络低层学到的是颜色、边缘、轮廓、纹理等低层特征[39]。低层特征包含的边界、轮廓信息有益于提高分割质量，使分割预测结果更加细致、精准。本文方法属于基于检测的方法，旨在充分利用低层特征所包含的低级语义信息缩短网络低层信息流向高层的路径长度。本文方法改进了 Mask R-CNN 的实例分割分支，改进的分割分支在原分割分支的基础上增加了网络低层特征，将其作为辅助信息来增强分割分支特征的表达能力，改善分割质量。考虑到计算效率，将从特征提取网络中选择的特征缩放至原图的 1/8 作为向分割分支传递的低层特征，使用 RoI Align 层将从 RPN（Region Proposal Network）得到的 RoIs 映射到低层特征并从中提取相应的目标特征。对提取到的 RoI 特征采用单个卷积层以减少通道数量，进而减少运算量。将压缩后的特征与原分割分支的特征拼接，并使用卷积自适应融合低级和高级语义信息，最后用于实例分割。本文工作的主要贡献如下：

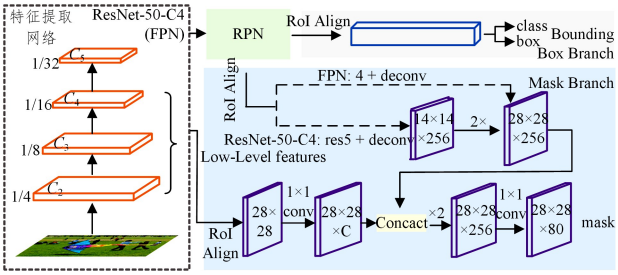
（1）提出了网络低层特征辅助的实例分割方法，改进了 Mask R-CNN 中的分割分支，改善了分割分支的特征表示，增强了分割分支特征对于目标物体的边界和轮廓的表达能力，从而有效提高了分割质量。

（2）在 COCO2017 数据集上的实验结果表明，本文方法对不同特征提取网络具有较好的鲁棒性和有效性。

2　低层特征辅助分割方法

本文方法旨在改善 Mask R-CNN 的分割质量，改进的 Mask R-CNN 网络结构如图 1 所示，其包括 4 个部分：特征提取网络（灰色虚线矩形框部分）、区域建议网络（Region Proposal Network，RPN）（浅绿色背景部分）、Bounding box 预测分支（灰色背景部分）和改进后的 Mask 预测分支（浅蓝色背景部分）。其中，特征提取网络计算得到输入图像的多尺度特征（见图 1 中橙色块），并将其作为 RPN 和改进后的 Mask 分支的输入。RPN 模块预测并输出目标粗略位置的定位边界框，将其作为 Bounding box 分支和 Mask 分支的输入。Bounding box 分支提取目标区域的特征并通过计算（图 1 中浅蓝色长方体）进一步准确定位目标位置。改进的 Mask 分支提取目标区域特征后与特征提取网络中的低层特征（Low-Level

Features)相融合，最后预测得到目标的 mask。



注：虚线表示原 Mask 分支中针对不同特征提取网络采取的不同操作；数字代表特征空间分辨率和通道；2×表示缩放 2 倍；×2 表示使用两个卷积层；×4 代表使用连续的 4 个卷积层；除了标有 1×1conv 的卷积层，其他所有卷积核尺寸为 3×3，反卷积核尺寸为 2×2，步长为 2，激活函数采用 ReLU；res5 表示 ResNet 中的第 5 个 stage

图 1 低层特征辅助分割网络(电子版为彩色)

Fig. 1 Low-level feature aided segmentation network

特征提取网络可以采用 ResNet-50-C4 和 ResNet-50/101-FPN 结构。根据命名习惯，当使用 ResNet-50-C4 时，特征提取网络的特征从下到上依次命名为 C_1，C_2，C_3，C_4，C_5；当使用 ResNet-50-FPN 时，特征提取网络的特征从下到上依次命名为 P_1，P_2，P_3，P_4，P_5。下面首先介绍 Mask R-CNN 的主要结构，然后详细叙述改进的 Mask 分支的细节。

2.1 Mask R-CNN 的主要结构

Mask R-CNN 主要包括 4 个部分：特征提取网络、区域建议网络(RPN)、Bounding box 预测分支和 Mask 分支。

特征提取网络可以使用不同的结构从图像 I 中提取不同尺度的卷积特征 $\boldsymbol{X}_k$，其中 $k=1,\cdots,6$。若特征提取网络采用 ResNet-50-C4，则仅使用 C_4 层的卷积特征 $\boldsymbol{X}_4$；若特征提取网络采用 ResNet-50/101-FPN，则使用 C_2 至 C_6 层的卷积特征 $\boldsymbol{X}_2-\boldsymbol{X}_6$，其中 $\boldsymbol{X}_6$ 由 $\boldsymbol{X}_5$ 下采样 1/2 得到。区域建议网络(RPN)根据输入图像 $\boldsymbol{I}$ 和相应的特征 $\boldsymbol{X}_k$ 得到目标区域建议边界框(Region Proposals)，然后 RPN 对目标区域建议边界框进行前景背景分类和边界框回归，并使用非极大值抑制(Non-maximum Suppression，NMS)筛选产生感兴趣目标区域(RoI)$b_i(i=1,2,\cdots,n,n$ 为 RoI 的数量)。

在原始的 Mask R-CNN 网络中，Bounding box 预测分支和 Mask 分支都将 RoI Align 层的输出作为输入。RoI Align 用于将 RoI 映射到特征图，RoI 对应区域的特征称为 RoI 特征。Bounding box 分支根据 RoI 特征对 RPN 给出的 RoI b_i 进行多类别分类和边界框回归，并使用 NMS 去除重复边界框。Mask 分支使用小型的 FCN 对 b_i 进行分割得到对应的二值掩码。

目前，在 Mask R-CNN 中较为常用的特征提取网络是 FPN (Feature Pyramid Networks)。FPN 利用不同尺度的卷积特征构造特征金字塔，是一种自上而下的具有内在多尺度特征的网络。FPN 可用于不同的检测和分割模型，其主要作用是通过不同分辨率的特征图来捕获目标多尺度信息以提升检测和分割的性能。

此外，Mask R-CNN 通过卷积、ReLU、池化等操作提取输入图像的不同尺度的特征。从低层到高层，特征的抽象能力和语义表达能力逐渐增强，有助于从复杂的背景中分割出前景目标。然而，特征金字塔高层特征缺乏目标轮廓和边界等信息，使 Mask 分支难以预测出轮廓准确、细节丰富的分割结果[1]。尽管采用 ResNet-50/101-FPN 特征提取网络能够利用特征金字塔在不同尺度上分割对应尺度的目标，但仍不足以分割出轮廓准确、边缘细致的结果。Mask R-CNN 在使用 ResNet-50-C4 作为特征提取网络时得到的物体检测和实例分割结果如图 2 所示。其中第一行第一幅图中，马的后腿不连续，脊背部分近似直线，因此可以看出 Mask R-CNN 的分割结果在物体的边界处无法准确地贴合实际的物体轮廓。鉴于低层网络特征可以提供低级的颜色、边界和轮廓等信息，本文方法引入网络的低层特征作为 Mask 分支的辅助信息来提升 Mask 的分割质量和分割精度。本文所提出的网络结构如图 1 所示，类似于 Mask R-CNN，改进后的网络仍然包括 4 个部分，分别是特征提取网络、区域建议网络(RPN)、Bounding box 预测分支和 Mask 分支。本文主要针对 Mask 分支进行改进。

图 2 本文方法和 Mask R-CNN 的实例检测和分割结果对比

Fig. 2 Comparison of instance detection and segmentation of proposed method and Mask R-CNN

2.2 低层特征辅助的 Mask R-CNN

考虑到计算效率，本文选择特征提取网络的第 2—4 层特征，即 $\boldsymbol{X}_2$，$\boldsymbol{X}_3$，$\boldsymbol{X}_4$。将不同尺度的特征缩放至输入图像尺寸的 1/8，并将其作为传递到分割分支的低层特征。$\boldsymbol{X}_2$，$\boldsymbol{X}_3$，$\boldsymbol{X}_4$ 可单独使用，也可以将特征缩放后拼接使用。如选择 C_2 和 C_4 层网络特征，则传递到分割分支的低层特征 $\boldsymbol{X}_l$ 为：

$$\boldsymbol{X}_l=Concact\left[\mathcal{R}\left(\boldsymbol{X}_2,\frac{1}{2}\right);\mathcal{R}(\boldsymbol{X}_4,2)\right] \tag{1}$$

其中，$R\left(\cdot,\frac{1}{2}\right)$ 表示缩放函数，1/2 表示缩放倍数；$Concact[;]$ 表示沿通道方向对特征进行拼接。

用于分割分支的特征由高层的语义特征 $\boldsymbol{X}_h^B$ 和低层 RoI 特征 $\boldsymbol{X}_l^B$ 组成。Mask R-CNN 中原 Mask 分支中的 RoI 特征经过 4 个卷积层或 res5 模块，并通过反卷积处理得到高层的语义特征 $\boldsymbol{X}_h^B$。反卷积操作用于对不同特征提取网络的特征进行缩放，若特征提取网络为 ResNet-50-C4，反卷积后特征尺寸为 14×14，需要将其上采样 2 倍；若特征提取网络为 FPN，反卷积后输出特征尺寸为 28×28，不需要使用上采样进行缩放。对于低层特征 $\boldsymbol{X}_l$，首先使用 RoI Align 操作从 $\boldsymbol{X}_l$ 中提取当前感兴趣区域对应的特征，其输出尺寸为 28×28，记为 $\boldsymbol{X}_l^B$。若 $\boldsymbol{X}_l^B$ 的通道数量高于 $\boldsymbol{X}_h^B$ 的通道数量，会使低层特征信息比重高于高层特征信息，因此特征融合后会引入过多次要信息。因此，使用 1×1 卷积操作对 $\boldsymbol{X}^B$ 进行降维，降维后特征的通道

数目为 C。本文在消融实验部分分析了特征通道数目 C 对分割精度的影响。

融合高层语义特征和低层特征可以使用按元素相加或拼接的方法，但文献[27]已经证明使用特征拼接的方法比按元素相加的方法灵活，有益于提高网络的性能。因此，本文使用拼接操作以及 2 个卷积层处理特征，能够自适应地融合高层语义信息和低层纹理信息。第一个卷积层的输入通道数为 $256+C$，输出通道数为 256；第二个卷积层的输入和输出通道数均为 256。经过卷积处理后得到最终的预测 mask 的特征 $\boldsymbol{X}^S$ 的尺寸为 $28\times28\times256$。$\boldsymbol{X}^S$ 的计算过程可形式化为：

$$\boldsymbol{X}^S=Conv(Conv(Concact[\boldsymbol{X}_h^\beta;Conv(\boldsymbol{X}_l^\beta)])) \tag{2}$$

其中，$Conv(\cdot)$ 表示卷积操作。

$\boldsymbol{X}^S$ 用于预测每个实例的分割结果。如图 2 所示，在第二行第一幅图中，改进后的 Mask R-CNN 可以完整地检测到马的后腿，分割的轮廓线更贴合马的脊背曲线。从图 2 可以看出，改进后的 Mask R-CNN 的分割结果在物体的边界处的效果明显优于原始的 Mask R-CNN。

3　实验及结果分析

3.1　数据集与评价指标

COCO 数据集[40]是目前较常用且具有挑战性的实例分割数据集之一。COCO2017 数据集有 80 类物体的实例标注信息，可以分为 3 个部分，分别为 train2017，val2017 和 test2017。train2017 包含约 115 000 张图像，val2017 包含 5 000 张图像，test2017 包含 40 000 张图像。其中 test2017 又被分为 test-dev 和 test-challenge，分别有 20 000 张图像。本文采用 train2017 进行训练，采用 test-dev2017 进行模型评估，采用 val2017 进行验证和消融实验。此外，在 test-dev2017 和 val2017 上分别进行了预测阶段的运行时间对比实验。

实验结果使用标准 COCO 评价指标。AP 表示 IoU 阈值从 0.5 至 0.95 每隔 0.05 取值下得到的多个 AP 的平均值；AP_{50} 和 AP_{75} 分别表示 IoU 阈值取值为 0.5 和 0.75 时的 AP 值；AP_S，AP_M 和 AP_L 表示不同尺度目标的 AP 值。

3.2　实施细节

实验 GPU 使用 1 个 RTX2080Ti，深度学习框架采用 Pytorch。实验模型的特征提取网络包括 ResNet-18-FPN，ResNet-50-C4 和 ResNet-50/101-FPN，均在训练之前加载公开的预训练权重。训练阶段每一批次输入 2 张图片，验证和测试阶段每一批次输入 1 张图片。每个实验均训练 720 000 个迭代共约 12 个 epoch，初始学习率为 0.002 5，并分别于第 480 000 个迭代和第 640 000 个迭代时将学习率乘以 0.1。实验的其他设置均与 maskrcnn-benchmark[41]一致。

3.3　定量结果

本文在与 Mask R-CNN[23]不同的实验条件下重新训练了 Mask R-CNN 作为本文基线(baseline)。COCO2017val 数据集上，Mask R-CNN 与本文方法在采用不同特征提取网络时的分割结果对比如表 1 所列。表 1 表明相比基线，本文提出的低层特征辅助的实例分割网络能够在采用不同特征提取网络时提高分割精度。当特征提取网络为 ResNet-50-C4 时，AP 提高 1.6%。当特征提取网络为 ResNet-50/101-FPN 时，分割精度提升幅度较小。不同实例分割方法在 COCO2017test-dev 数据集上的分割结果对比如表 2 所列。表 2 表明当特征提取网络为 ResNet-101-FPN 时，本文基线的 AP 值比 Mask R-CNN 高出 0.7%，而本文提出的方法相比基线提高了 0.5%。

表 1　COCO2017val 数据集上，Mask R-CNN 与本文方法在采用不同特征提取网络时的分割结果对比

Table 1　Segmentation results of Mask R-CNN and proposed method using different feature extraction networks on COCO2017val dataset

特征提取网络	低层特征	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
ResNet-50-C4	—	31.6	52.9	33.2	12.9	35.0	49.5	9.7
	√	**33.2**	53.7	35.4	13.1	36.4	51.6	8.3
ResNet-50-FPN	—	34.3	56.1	36.2	15.6	36.9	51.1	15.3
	√	**35.1**	56.4	37.4	15.6	37.5	52.5	14.6
ResNet-101-FPN	—	36.1	58.3	38.3	17.3	39.1	53.0	11.7
	√	**36.7**	58.6	39.2	16.8	39.4	54.0	11.0

注：— 表示重新训练的 Mask R-CNN，√ 表示引入低层特征辅助 Mask 分支进行实例分割

表 2　不同实例分割方法在 COCO2017test-dev 数据集上的分割结果对比

Table 2　Comparison of segmentation results of different instance segmentation methods on COCO2017test-dev dataset

方法	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC	ResNet-101	24.6	44.3	24.8	4.7	25.9	43.6
FCIS++	ResNet-101	33.6	54.5	—	—	—	—
Mask R-CNN[23]	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN[23]	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
MaskLab	ResNet-101	35.4	57.4	37.4	16.9	38.3	42.9
Mask R-CNN(retrain)	ResNet-50-C4	31.9	53.5	33.4	12.5	34.3	48.3
Ours		**33.6**	54.2	35.7	13.4	36.0	50.5
Mask R-CNN(retrain)	ResNet-50-FPN	34.6	56.7	36.7	15.4	36.3	49.6
Ours		**35.2**	56.8	37.3	15.7	36.9	50.5
Mask R-CNN(retrain)	ResNet-101-FPN	36.4	58.9	38.7	16.7	38.4	51.9
Ours		**36.9**	59.1	39.2	16.6	39.1	53.4

综合分析表 1 和表 2，本文方法相比本文基线在不同的特征提取网络下都有相应的精度提升。特征提取网络为 ResNet-50-C4 结构的模型在不同数据集上精度的提升幅度均比具有 FPN 结构的模型提升幅度大。由于在 ResNet-50-C_4 结构的模型中，不同尺度目标的 RoI 特征仅来自 C_4 特征，C_4 特征具有丰富的语义信息但缺乏准确的边缘、轮廓和定位信息，因此在分割分支融合了特征提取网络中的低层信息之后对分割精度的提升较明显。ResNet-50/101-FPN 模型的特征提取网络采用自顶向下的多尺度特征结构，并且从上至下特征所包含的语义信息逐渐减少，轮廓和定位信息逐层增加。

在 RoI Align 过程中，较大尺度目标的 RoI 特征提取自较小分辨率的特征图，小尺度目标的 RoI 特征提取自较大分辨率的特征图。因此较大尺度目标和小尺度目标的 RoI 特征中所包含的信息不同，导致低层特征对不同尺度目标的 RoI 特征的信息补充作用不同。表 1 和表 2 中，特征提取网络为 ResNet-50/101-FPN 的模型均在 AP_L 指标上提升幅度较大，而在 AP_S 指标上的提升不明显，因此当特征提取网络为 ResNet-50/101-FPN 时 AP 提升幅度较小。

本文方法由于引入了低层特征使网络增加了一定的计算开销。由表 1 和表 3 中预测阶段的运行时间对比可知，引入低层特征辅助 Mask 分支进行实例分割对分割速度造成的影响较小。

表 3 不同实例分割方法在 COCO2017test-dev 数据集上预测时的运行速度对比

Table 3 Comparison of running speed of different instance segmentation methods in prediction on COCO2017test-dev dataset

方法	Backbone	FPS
FCIS+	ResNet-101	8.4
Mask R-CNN[23]	ResNet-101-FPN	5.0
Mask R-CNN(retrain)	ResNet-50-C4	9.2
Ours		8.3
Mask R-CNN(retrain)	ResNet-50-FPN	14.9
Ours		13.9
Mask R-CNN(retrain)	ResNet-101-FPN	11.5
Ours		11.2

注：运行速度使用单 GPU RTX2080Ti 测量得到

3.4 消融实验

在 COCO2017val 数据集上进行所有消融实验，实验模型的特征提取网络均为 ResNet-18-FPN。设置初始学习率为 0.000 5，分别于迭代 480 000 次和 640 000 次时将学习率乘以 0.1。消融实验主要考虑特征提取网络中特征选择策略和特征融合前低层特征通道数目对分割精度的影响。

3.4.1 特征提取网络中的特征选择策略

特征提取网络中不同层的特征包含不同的语义信息。本文分别选择特征提取网络中 P_2，P_3 和 P_2+P_4 卷积特征进行实验对比。实验结果如表 4 所列，选择 P_3 卷积特征作为低层特征时的表现优于其他两种选择策略。对于 ResNet-18-FPN 特征提取网络，P_3 特征对 Mask 分支的辅助作用更大，性能提升更多。选择多层特征拼接后作为低层特征的方法的提升效果不明显。P_4 特征缺乏边缘和轮廓等信息，与高分辨率的 P_2 特征拼接后没有达到信息互补的效果，从而造成了不利的影响。

表 4 选择特征提取网络中不同层的特征时的分割结果对比

Table 4 Comparison of segmentation results of selecting different layers of feature extraction network

低层特征	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
P_2	21.9	38.6	22.1	9.1	22.6	33.7
P_3	**22.3**	39.2	23.1	9.8	22.9	34.9
P_2+P_4	21.8	38.6	22.1	9.1	22.4	33.8

3.4.2 低层特征通道数目的影响

低层特征通道数目的减少能够减小计算量，以及避免特征融合后其比重超过 Mask 分支的高层特征而引入过多噪声。实验对比了低层特征经过 1×1 卷积后得到的特征通道数目分别为 24，48 和 72 时的分割结果。

表 5 低层特征通道数目对分割结果的影响

Table 5 Effect of number of channels of low-level feature on segmentation results

通道数	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
24	22.2	39.3	22.6	9.3	23.0	33.9
48	22.3	39.2	23.1	9.8	22.9	34.9
72	22.4	39.4	22.8	10.2	23.0	34.2

表 5 表明不同通道数目的低层特征对 AP 的影响并不显著，而在 AP_S 和 AP_L 指标上具有明显的差异。低层特征通道数为 24 时的 AP_S 和 AP_L 均明显低于其他两种情形，这表明通道数目较少的特征对分割性能的辅助作用较小。低层特征通道数目为 48 时的 AP_L 高于通道数目为 72 时的 AP_L，而其 AP_S 却低于通道数目为 72 时的 AP_S。低层特征通道数目的增加有利于特征表达更丰富的信息并且提高分割精度。因此，相比具有挑战性的小尺度目标，大尺度目标更易于分割。通道数目过多的低层特征易造成信息冗余而降低分割精度。

结束语 本文提出了引入网络低层信息辅助 Mask R-CNN 分割分支提升实例分割精度的方法，以解决分割边界和分割轮廓粗糙的问题。本文从特征提取网络中选择网络低层卷积特征并将其缩放至固定尺度作为低层特征，在 Mask R-CNN 分割分支中引入低层特征从而引入更多的纹理和轮廓信息用于实例分割。通过 COCO 数据集上的不同实验可知，使用 ResNet-50-C4 或 ResNet-50/101-FPN 作为特征提取网络时，本文方法的实例分割精度都有所提高，从而验证了本文方法的有效性和鲁棒性。但由于引入更多的低层特征会带来计算开销，因此在未来的工作中将研究如何兼顾速度和精度，以提高模型的整体性能。

参 考 文 献

[1] LIU S, QI L, QIN H, et al. Path Aggregation Network for Instance Segmentation[J]. arXiv:1803.01534.

[2] LUO J, SAVAKIS A E, SINGHAL A. A Bayesian network-based framework for semantic image understanding[J]. Pattern Recognition, 2005, 38(6): 919-934.

[3] LI L, JIANG S Q, HUANG Q M. Learning Hierarchical Semantic Description Via Mixed-Norm Regularization for Image Understanding[J]. IEEE Transactions on Multimedia, 2012, 14(5): 1401-1413.

[4] LOZANO S, MÖLLER K, BRENDLE A, et al. AUTOPILOT-BT: A system for knowledge and model based mechanical ventilation[J]. Technology and Health Care, 2008, 16(1): 1-11.

[5] THEIS J, OSSMANN D, THIELECKE F, et al. Robust autopilot design for landing a large civil aircraft in crosswind[J]. Control Engineering Practice, 2018, 76: 54-64.

[6] ZHU J, LAO Y W, ZHENG Y F. Object Tracking in Structured Environments for Video Surveillance Applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(2): 223-235.

[7] GILBERT A L, GILES M K, FLACHS G M, et al. A Real-Time Video Tracking System[J]. IEEE Transactions on Pattern Ana-

lysis and Machine Intelligence，1980(1)：10.

[8] SALTI S，CAVALLARO A，DI STEFANO L. Adaptive Appearance Modeling for Video Tracking：Survey and Evaluation [J]. IEEE Transactions on Image Processing，2012，21(10)：4334-4348.

[9] YEE K P，SWEARINGEN K，LI K，et al. Faceted metadata for image search and browsing[C]// Proceedings of the 2003 Conference on Human Factors in Computing Systems(CHI 2003). Ft. Lauderdale，Florida，USA，2003.

[10] WANG M，LI H，TAO D C，et al. Multimodal Graph-Based Reranking for Web Image Search[J]. IEEE Transactions on Image Processing，2012，21(11)：4649-4661.

[11] LI X，LIU Z，LUO P，et al. Not All Pixels Are Equal：Difficulty-aware Semantic Segmentation via Deep Layer Cascade [J]. arXiv：1704.01344.

[12] LIU Z，LI X，LUO P，et al. Semantic Image Segmentation via Deep Parsing Network[J]. arXiv：1509.02634.

[13] PINHEIRO P O，COLLOBERT R，DOLLAR P. Learning to Segment Object Candidates[J]. arXiv：1506.06204.

[14] PINHEIRO P O，LIN T Y，COLLOBERT R，et al. Learning to Refine Object Segments[J]. arXiv：1603.08695.

[15] DAI J，HE K，LI Y，et al. Instance-sensitive Fully Convolutional Networks[J]. arXiv：1603.08678.

[16] DAI J，HE K，SUN J. Instance-aware Semantic Segmentation via Multi-task Network Cascades[J]. arXiv：1512.04412.

[17] HAYDER Z，HE X，SALZMANN M. Boundary-Aware Instance Segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2017.

[18] DAI J，LI Y，HE K，et al. R-FCN：Object Detection via Region-based Fully Convolutional Networks[J]. arXiv：1605.06409.

[19] GIRSHICK R. Fast r-cnn[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE，2016.

[20] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[21] LI Y，QI H，DAI J，et al. Fully Convolutional Instance-Aware Semantic Segmentation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI：IEEE，2017：4438-4446.

[22] CHEN L C，HERMANS A，PAPANDREOU G，et al. MaskLab：Instance Segmentation by Refining Object Detection with Semantic and Direction Features[J]. arXiv：1712.04837.

[23] HE K，GKIOXARI G，PIOTR DOLLÁ R，et al. Mask R-CNN [C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE，2017.

[24] HUANG Z，HUANG L，GONG Y，et al. Mask Scoring R-CNN [J]. arXiv：1903.00241.

[25] CHEN K，PANG J，WANG J，et al. Hybrid Task Cascade for Instance Segmentation[J]. arXiv：1901.07518.

[26] CAI Z，VASCONCELOS N. Cascade R-CNN：Delving into High Quality Object Detection[J]. arXiv：1712.00726.

[27] SUN Y，P P S K，SHIMAMURA J，et al. Concatenated Feature Pyramid Network for Instance Segmentation[J]. arXiv：1904.00768.

[28] LIN T Y，DOLLÁR P，GIRSHICK R，et al. Feature Pyramid Networks for Object Detection[J]. arXiv：1612.03144.

[29] LIANG X，LIN L，WEI Y，et al. Proposal-free Network for Instance-level Object Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2018，40(12)：2978-2991.

[30] BAI M，URTASUN R. Deep Watershed Transform for Instance Segmentation[J]. arXiv：1611.08303.

[31] BEUCHER S，C ANTUÉL. Use of Watersheds in Contour Detection[C]// International workshop on image processing，real-time edge and motion detection. CCETT，1979.

[32] KIRILLOV A，LEVINKOV E，ANDRES B，et al. InstanceCut：from Edges to Instances with MultiCut[J]. arXiv：1611.08272.

[33] JIN L，CHEN Z，TU Z. Object Detection Free Instance Segmentation With Labeling Transformations[J]. arXiv：1611.08991.

[34] LIU S，JIA J，FIDLER S，et al. SGN：Sequential Grouping Networks for Instance Segmentation[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice：IEEE，2017：3516-3524.

[35] REN M，ZEMEL R S. End-to-End Instance Segmentation with Recurrent Attention[C]// Computer Vision & Pattern Recognition. IEEE，2017.

[36] ROMERA-PAREDES B，TORR P H S. Recurrent Instance Segmentation[J]. Computer Science，2016，9910(10)：312-329.

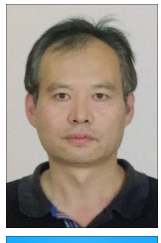
[37] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.

[38] SHI X，CHEN Z，WANG H，et al. Convolutional LSTM Network：A Machine Learning Approach for Precipitation Nowcasting[J]. arXiv：1506.04214.

[39] ZEILER M D，FERGUS R. Visualizing and Understanding Convolutional Networks[J]. arXiv：1311.2901.

[40] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft COCO：Common Objects in Context [C] // European Conference on Computer Vision. Springer International Publishing，2014.

[41] MASSA F，GIRSHICK R. maskrcnn-benchmark：Fast，modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch[OL]. https：//github.com/facebookresearch/maskrcnn-benchmark.



FAN Wei，born in 1968，Ph.D，professor，is a member of China Computer Federation. His main research interests include machine learning and revenue management.

HUANG Rui，born in 1987，Ph.D，lecturer，is a member of China Computer Federation. His main research interests include computer vision and machine learning.