基于轮廓点掩模细化的单阶段实例分割网络

张绪义*，曹家乐
天津大学电气自动化与信息工程学院，天津 300072

摘要 针对现有的实例分割方法 PolarMask 中分割结果边缘信息模糊的问题，通过对轮廓点角度偏置和距离的预测，基于轮廓点细化的单阶段实例分割网络准确提取出实例轮廓。同时，为了进一步提升实例分割的性能，利用语义分割子网络对实例边缘进行了进一步细化。实验结果表明，所提方法在大规模实例分割数据集 MS COCO 的测试集上的分割精度为 32.5%，比现有的实例分割方法(PolarMask)提高了 2.1 个百分点，证明了所提方法的有效性。

关键词 机器视觉；实例分割；语义分割；深度学习；卷积神经网络；角度预测
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/AOS202040.2115001

Contour-Point Refined Mask Prediction for Single-Stage Instance Segmentation

Zhang Xuyi*, Cao Jiale
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract To solve the fuzzy problem of edge information in mask results by single-stage PolarMask, a contour-point refined network is proposed herein. By predicting the angel offset and distance for each contour point, a more accurate contour can be generated. Moreover, an extra semantic segmentation is added to further refine the edge information. Experiments show that the proposed method achieves a segmentation accuracy of 32.5% on the MS COCO test dataset, 2.1 percentages higher than the fundamental PolarMask, demonstrating the effectiveness of the proposed method.

Key words machine vision; instance segmentation; semantic segmentation; deep learning; convolutional neural network; angle prediction
OCIS codes 150.0155; 040.1880; 100.4996

1 引 言

深度卷积神经网络被广泛应用于计算机视觉的各种应用中，例如物体检测[1]、空中侦察[2]、无人驾驶[3]、遥感监测[4]等领域。实例分割是重要且具有挑战性的计算机视觉任务之一，不仅需要对不同类别物体进行像素级分割，而且需要对不同物体进行区分。近年来，基于深度卷积神经网络的实例分割取得了巨大的成功。

基于深度卷积神经网络的实例分割方法主要分为两类，包括两阶段的实例分割和单阶段的实例分割。两阶段的实例分割可分为自上而下[5-8]和自下而上[9-11]的方法。自上而下的方法首先利用目标检测的方法检测出物体的区域，然后对该区域进行像素级分割。FCIS[5]利用全卷积神经网络生成中间特征图和共享特征图，并使用位置敏感的特征融合方法进行特征提取，从而进行实例分割。Mask R-CNN[6]是基于 Faster R-CNN[1]通过添加一个分割分支和一种区域特征聚集方法 RoI-Align 来获得精确的感兴趣区域并产生实例分割结果。MS R-CNN[7]提出了一个 Mask IoU 分支来学习预测出来的实例分割图的质量。PANet[8]以 Mask R-CNN 为基础，通过自底向上的路径增强、动态特征池化和全连接层融合来提高检测性能。自下而上的方法则是先进行像素级别的语义分割，再通过聚类、度量学习等手段来生成不同的实例。文献[9]中先进行语义分割操

收稿日期：2020-06-08；**修回日期**：2020-07-05；**录用日期**：2020-07-15
基金项目：国家自然科学基金青年科学基金(61906131)、天津市新一代人工智能科技重大专项(18ZXZNGX00320)
* **E-mail**：zxy1996@tju.edu.cn

作，再通过判别损失函数来训练网络，最后使用mean-shift的方法输出不同的实例。SGN[10]则是把实例分割问题分为一系列的子分组问题。SSAP[11]则是选择一个像素对关联金字塔，即判断两个像素属于同一实例的概率，然后通过级联分区提取实例分割图。自下而上的方法的效果通常要比自上而下的方法差。

单阶段实例分割是直接对每个物体进行分类、定位和像素级分割。由于不需要先检测后分割，该类方法一般具有更快的检测速度。受到单阶段的目标检测研究的影响，单阶段的实例分割可以分为基于锚点和无需锚点的方法。基于锚点的方法以YOLACT[12]和SOLO[13]为代表。YOLACT基于RetinaNet[14]通过全卷积网络(FCN)生成K个掩码，并预测K个线性组合系数，通过线性组合生成实例掩模。SOLO根据实例的位置和大小为实例中的每个像素分配类别，从而将实例分割转化为一个可分类的问题。TensorMask[15]采用密集滑动窗口的方式，为每个像素在局部窗口分割实例。无需锚点的方法以PolarMask[16]为代表，提出一种无需预设框的实例分割框架，利用极坐标系建模方式代替直角坐标系，用于预测实例的轮廓点与实例中心点之间的距离，从而生成实例分割图。选定极坐标的原点为实例中心，实例轮廓点由预测距离和固定角度确定，通过对多个轮廓点依次连线生成该实例分割图，从而将实例分割任务简化，使其和目标检测任务具有相同的复杂度。

PolarMask这种基于固定角度回归轮廓的方法对实例形状的鲁棒性不高。若实例形状不规则时，PolarMask在进行轮廓点依次连接时会引入背景信息或将部分实例切割掉，进而导致最终生成的实例分割结果十分不准确。针对上述问题，本文在实例分割网络PolarMask中引入两个子网络：掩模预测子网络，用于预测轮廓点的距离和角度偏置，以提取实例更准确的包络；语义分割子网络，用于预测每一类物体的语义分割图，将生成的实例分割结果与语义分割结果相结合，从而得到轮廓更加精确的实例分割结果。实验证明，该方法在MS COCO数据集评价指标下取得了有效的提升。

2　网络结构

针对PolarMask边缘信息过于粗糙的问题，在PolarMask的基础上，提出一种基于轮廓点掩模细化的单阶段实例分割网络，该方法主要包括四个子网络：骨干网络、特征金字塔子网络、语义分割子网络和掩模预测子网络。网络首先读取输入图像，通过骨干网络和特征金字塔子网络提取图像特征。语义分割子网络用于融合特征金字塔子网络输出的浅层细节特征和深层语义特征，生成该图各个类别的语义分割图。掩模预测子网络在多个尺寸的特征平面上预测每个实例轮廓点的距离和角度偏置生成实例轮廓，保证实例的完整性，最后与语义分割图相结合生成精细的实例掩模结果。

2.1　PolarMask

首先对基础方法PolarMask进行简单的介绍，该网络结构是基于无需锚点的目标检测网络FCOS[17]建立的，PolarMask提出极坐标建模的实例分割，通过寻找图像中实例的轮廓来进行建模，把实例分割问题转化为实例中心点的分类问题和密集距离的回归问题。该网络结构由骨干网络、特征金字塔网络(FPN)和检测子网络构成。由于完成的任务不同，将FCOS的检测框预测分支改为掩模距离预测分支，并将预测的通道数由4替换为36，表示实例中36个轮廓点，相对于实例原点的长度。同时对极坐标中心点进行采样，选择出高质量的正样本，降低了低质量样本的损失权重。此外，考虑到Smooth-L1损失函数在进行掩模距离回归时没有考虑轮廓点之间的信息，设计了Polar IOU损失函数，无须调整损失函数的权重就能使掩模分支快速且稳定收敛。

PolarMask证明了无需锚点的方法可以成为实例分割的一个新的方向，但从PolarMask的可视化结果可以看出，在图1标记的区域，其将背景误分为实例或将实例的部分区域进行了切割，其原因是PolarMask采用固定角度来预测轮廓点且轮廓点并不密集，导致轮廓点依次相连时包含背景或切割实

图1　PolarMask结果可视化

Fig. 1　PolarMask result visualization

例，使分割精度下降。为了解决这一问题，本文提出了语义分割子网络和掩模预测子网络，通过掩模预测子网络分别预测轮廓点的距离和角度偏置，再与语义分割子网络进行融合，从而生成精细的实例轮廓。

整体的网络结构如图 2。网络主要分为四个部分：骨干网络，特征金字塔子网络，语义分割子网络和掩模预测子网络。骨干网络主要是用于提取出检测和分割所需特征，浅层的特征图包含更多的细节信息，深层特征图包含更多的语义信息。常用的骨干网络有 VGG16[18]、Resnet50[19]、Resnet101[19]等。特征金字塔子网络主要是针对不同尺度物体的感受野的不同来提取对应的特征图，并对多尺度策略特征金字塔输出的特征图分别分配相应的预测子网络，以提高目标重叠情况下的预测性能。语义分割子网络用于生成图像中每一类别的语义分割图。掩模预测子网络用于生成图像中实例的最大轮廓和所属类别，最终将掩模预测子网络输出的实例轮廓和其对应类别的语义分割图相融合生成最终精细的实例分割结果。

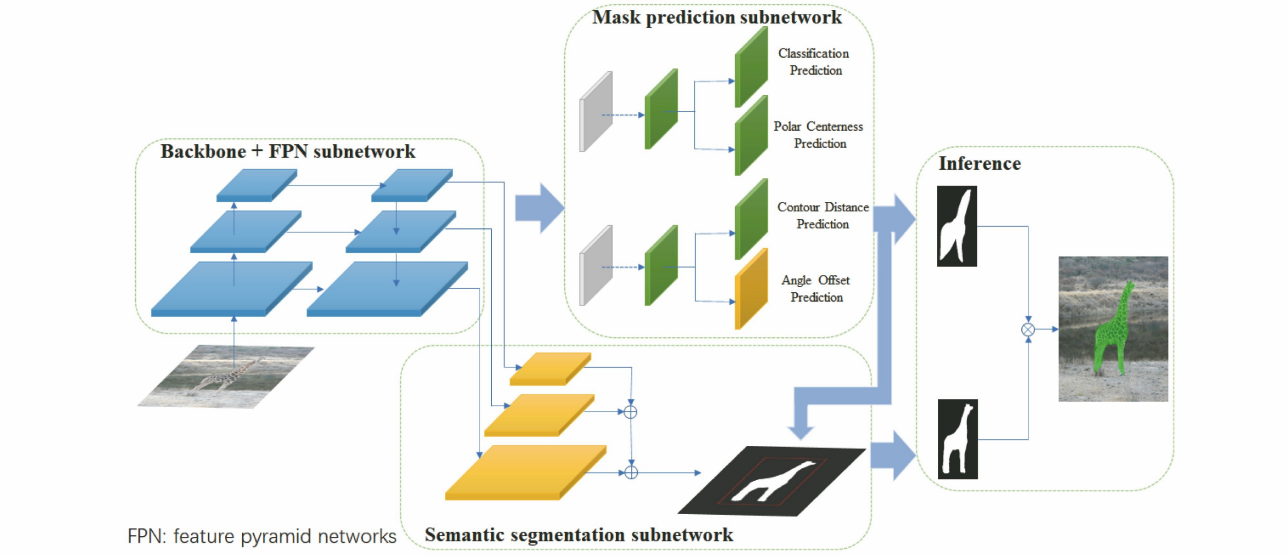

图 2　整体网络结构

Fig. 2　Network architecture of our overall framework

2.2　掩模预测子网络

掩模预测子网络包含四个分支，分别是分类预测、中心度预测、轮廓点距离预测和轮廓点角度偏置预测。分类预测用于预测以该点为中心的实例所属类别，中心度预测用于判断该点对于损失函数的权重大小，中心度越大，表示该点距离真实实例的中心越近，对损失函数的贡献越大，以此关注实例中心点的轮廓回归。通过距离和角度偏置预测来生成实例的整体轮廓。轮廓点距离表示实例原点(x_{center}，y_{center})到轮廓点(x_i，y_i)的距离 r_i，轮廓点的角度偏置表示每个轮廓点在初始角度 θ_i 上的偏置 $\Delta\theta_i$，如图 3 所示。

使用极坐标系进行建模时，需要制作极坐标系下轮廓点的标签，使用 36 个轮廓点组成一个实例轮廓。每个轮廓点的初始角度为 θ_i(θ_i=5°，15°，25°，…，355°)，由于卷积神经网络的固有特性，对于区间较大的数值的拟合效果较差，因此将角度偏置范围确定在(−5°，5°)之间。若角度偏置范围存在多个真实标签的轮廓点，本文选择距离实例原点最远的轮廓点作为回归的轮廓点，该轮廓点距离实例原点的距离即为标签在该初始角度 θ_i 的偏置角度为 $\Delta\theta_i$ 的回归距离 r_i。若在偏置范围内，真实标签不存在轮廓，则将 r_i 设为 10^{-6}，$\Delta\theta_i$ 设为 0。轮廓点对应的横纵坐标为

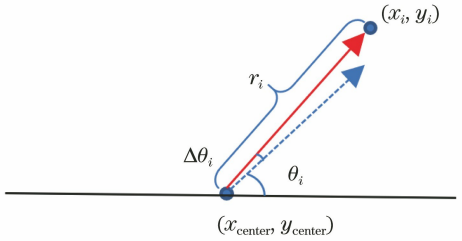

图 3　轮廓点的极坐标示意图

Fig. 3　Polar coordinates of the contour points

$$x_i = x_{\mathrm{center}} + r_i \times \sin(\theta_i + \Delta\theta_i), \tag{1}$$

$$y_i = y_{\mathrm{center}} + r_i \times \cos(\theta_i + \Delta\theta_i), \tag{2}$$

其中(x_i，y_i)为轮廓点在直角坐标系下的坐标。

在训练时由于需要回归 36 个轮廓点的距离和

角度偏置，相比目标检测更为复杂，所以选择 Polar IoU loss[16] 损失函数从整体上回归目标，目的是自动保持密集距离预测的分类损失与回归损失之间的平衡，其表达式为

$$l_{\text{mask}}=\log\frac{\sum_{i=1}^{n}d_{\max}}{\sum_{i=1}^{n}d_{\min}},\tag{3}$$

其中 $d_{\max}$ 表示预测的轮廓点到中心点的距离和对应的真实距离之间的最大值，$d_{\min}$ 表示预测的轮廓点到中心点的距离和对应的真实距离之间的最小值，n 表示每个实例的轮廓点数。角度偏置预测的损失函数选择 Smooth-L1 损失函数，可以避免梯度值过大、回归时出现梯度爆炸等问题。

在测试过程中，将特征图中每个点网络输出的分类得分和对应的中心度相乘得到最终的置信度得分，设置置信度阈值为 0.05，从特征金字塔输出置信度得分最高的 1000 个掩模结果，并用阈值为 0.5 的非极大值抑制（NMS）来生成掩模预测子网络输出的实例结果，用于之后和语义分割子网络的结果相结合。

2.3　语义分割子网络

充分利用浅层细节信息与深层语义信息对于图像语义分割任务非常重要。语义分割子网络如图 4 所示，通过将特征金字塔子网络输出的三个不同层级的特征层进行相加融合，再使用一个 3×3 的卷积层和一个 1×1 的卷积层将输出的特征维度进行压缩，输出类别数目相同的维度。

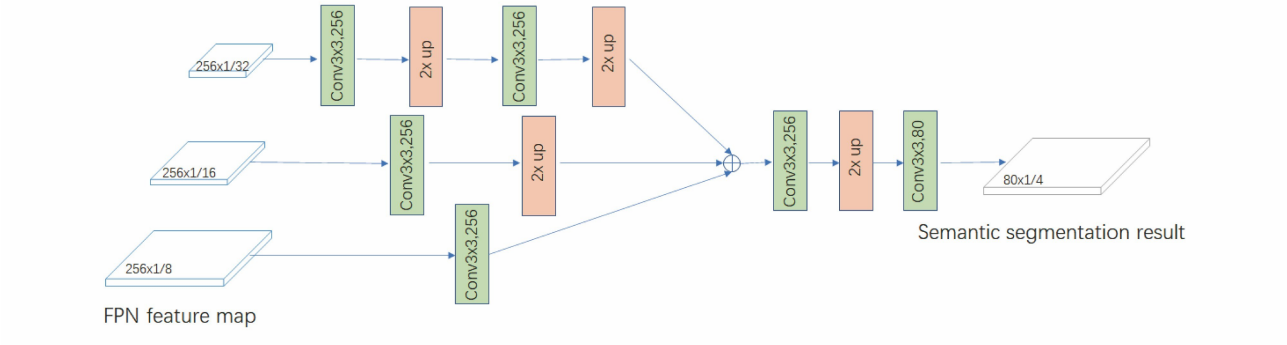

图 4　语义分割子网络

Fig. 4　Semantic segmentation subnetwork

上下层级的特征融合可以有效地提取出实例特征。由于图像中不同实例的尺寸都不相同，为了能让网络可以准确地分割出不同尺寸的实例，充分利用不同尺寸特征图中不同的感受野的特性，即相对大的感受野的特征图更关注实例的语义信息，相对小的感受野的特征图更关注实例的细节信息，通过将不同层的特征图进行 3×3 的卷积操作再进行上采样，将不同大小的特征图转换为相同尺度进行相加，使融合的特征图包含充分的语义信息和细节信息。该操作更有利于进行之后的语义分割。

该任务不同于全景分割，不需要对前景和背景同时进行语义分割，只需对实例类别进行语义分割即可，但前景的实例在大多数图像中的比例小于背景的比例，这便导致语义分割损失函数难以收敛，致使所有的像素点都被预测为背景。为了解决这一问题，提出了一种裁剪关注区域的方案，即只对关注区域进行语义分割损失函数的计算。根据掩模预测子网络输出的实例框出包含整个实例中的最大掩模矩形框。设掩模预测子网络输出的轮廓点距离为 r_i，输出的角度偏置为 $\Delta\theta_i$，由(1)式、(2)式得到矩形框左上坐标和右下坐标。如图 5 所示，图片中仅有两个实例，分别对每个实例取对应的矩形框作为其语义分割关注区域，即图 5(b)中矩形框的区域。仅对这个区域产生的损失函数进行计算，即可保证前景和背景比例均衡，有利于损失函数的收敛。矩形框由 36 个轮廓点确定，表达式为

$$x_{\max}=\max x_i+l_{\text{crop}},\tag{4}$$

$$x_{\min}=\min x_i-l_{\text{crop}},\tag{5}$$

$$y_{\min}=\min y_i-l_{\text{crop}},\tag{6}$$

$$y_{\max}=\max y_i+l_{\text{crop}},\tag{7}$$

式中 (x_i,y_i) 表示第 i 个轮廓点坐标（$i=1,2,\cdots,36$），l_{crop} 表示扩大的矩形框的距离，$(x_{\min},y_{\max})$ 为矩形框的左上角坐标，$(x_{\max},y_{\min})$ 表示矩形框的右下角坐标。

2.4　网络损失函数

损失函数的选择是卷积神经网络训练过程中重要的一部分，通过梯度下降算法将损失函数减小，使网络逐渐逼近最优参数，从而使神经网络达到较好的性能。本文定义整个网络的训练损失函数 l 由五部分组成，即

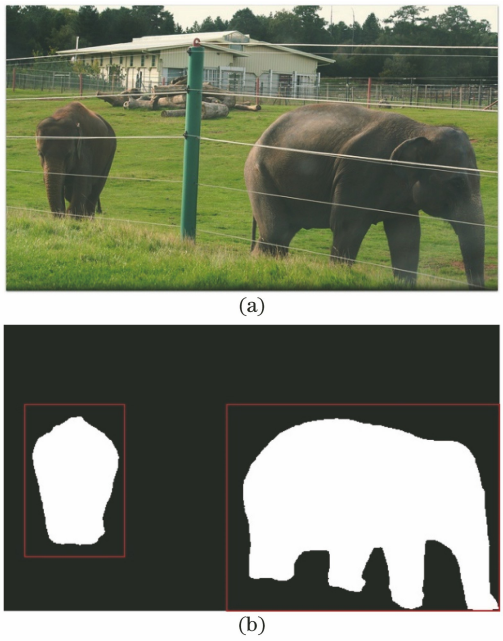

图 5　语义分割图像。(a)原始图像;(b)语义分割关注区域

Fig. 5　Semantic segmentation of an image. (a) Original image; (b) semantic segmentation of area of concern

$$l = l_{\text{cls}} + \lambda_{\text{centerness}} l_{\text{centerness}} + \lambda_{\text{mask}} l_{\text{mask}} + \lambda_{\text{angle}} l_{\text{angle}} + \lambda_{\text{segm}} l_{\text{segm}}, \tag{8}$$

式中:$\lambda_{\text{centerness}}$ 为中心度损失函数的权重;λ_{mask} 为掩模距离预测损失函数的权重;λ_{angle} 为掩模角度偏置预测损失函数的权重,λ_{segm} 为语义分割子网络损失函数的权重。其中 $\lambda_{\text{centerness}}$ 和 λ_{mask} 参考 Xie 等[16]的超参数设计,选取 $\lambda_{\text{centerness}}$ 值为 1,λ_{mask} 值为 1;λ_{segm} 和 λ_{angle} 一般可通过经验方式进行确定,人为设定几组不同的超参数值对比网络性能,从中选择最优的一组超参数,最终选取 λ_{mask} 值为 1,λ_{angle} 值为 0.2。分类损失函数 l_{cls} 选用 Focal loss 损失函数,表达式为

$$l_{\text{cls}} = -\alpha(1-p)^{\gamma}\log p, \tag{9}$$

其中 $\gamma=0.2$,$\alpha=0.25$,p 表示所分类别的置信度。相较于交叉熵损失函数可以有效解决多分类任务样本不平衡的现象。中心度(centerness)是为了提取高质量的正样本,使中心点的损失函数相对较大,使网络更关注于中心点的轮廓点回归,中心度的损失函数 $l_{\text{centerness}}$ 选用交叉熵函数。l_{mask} 采用 Polar IoU loss 可以在无需调整权重的情况下使掩模分支更快地收敛。角度偏置的损失函数 l_{angle} 选用 Smooth-L1 损失函数,表达式为

$$L_{\text{angle}}(\Delta\theta, \Delta\theta^{\text{gt}}) = \sum_{i=1}^{36} \text{smooth}_{\text{L1}}(\Delta\theta_i - \Delta\theta_i^{\text{gt}}), \tag{10}$$

式中 θ_i 表示第 i 个轮廓点的角度偏置,θ_i^{gt} 表示第 i 个轮廓点的角度偏置标签。语义分割领域中最常用的损失函数是交叉熵函数,本研究使用交叉熵损失函数来惩罚各像素点的预测值和真实值的差。

3　实验结果与分析

3.1　实验数据集

为了证明所提方法的有效性,在 MS COCO 数据集[20]上进行实例分割实验。MS COCO 数据集是在实例分割、目标检测和全景分割等领域使用十分广泛的数据集,该数据集共有 80 个类别,使用 11.5 万张图像作为训练集,5000 张图像作为验证集,2 万张图像作为测试集。MS COCO 数据集使用不同交并比(IOU)阈值下的平均精准度(Mean Average Precision)进行评估。其中测试集数据没有提供真实标签,需要提交至其官方服务器进行测试。均使用 1 倍训练策略,即单尺度训练和单尺度测试。

3.2　训练细节

为研究本文所提各个部分的作用,在切片实验中,均采用 ResNet-50 作为骨干网络,采用与 PolarMask 相同的超参数。具体地来讲,实验采用随机梯度下降(SGD)进行 9 万次迭代训练,初始学习率为 0.002,batch size 为 4。当迭代次数为 6 万次和 8 万次时,学习率分别降低 90%。权重衰减(Weight decay)设为 0.0001,动量(Momentum)设为 0.9。使用在大规模图像分类数据集 ImageNet[21]上预训练的权重初始化骨干网络。输入图像通常会被调整为同一尺度,短边等于 768,长边小于等于 1280。

3.3　语义分割子网络设计

为验证 2.3 小节提出的语义分割子网络的有效性,表 1 比较了特征金字塔不同特征层的融合方法对于实例分割结果的影响,并给出在各种情况下不同 IoU 阈值时的平均精度均值(AP)。实验结果表明,使用不同特征层进行相加的方法(sum)要优于采用通道级联再使用 1×1 卷积层进行降维的方法(Concat+1×1 conv)。

表 1　语义分割特征融合方式的比较

Table 1　Comparison of semantic segmentation feature fusion methods　　unit: %

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Concat+1×1 conv	30.5	51.0	31.9	12.5	32.7	46.3
Sum	30.8	51.5	32.3	12.4	32.0	45.2

3.4　掩模预测子网络设计

为验证 2.2 节提出的一种掩模预测子网络设计的有效性，比较了将角度偏置预测和距离预测分别使用不同预测子网络的方法[图 6(a)]以及共用同一预测子网络的方法[图 6(b)]时的实验结果。从表 2 中可以看出使用同一预测子网络的方法比分别使用不同的预测子网络的方法的 AP 性能要高 1.4 个百分点。原因是角度偏置预测和距离预测对象都是轮廓点，共用一个预测子网络可以有效减小计算量，提高运行速度。

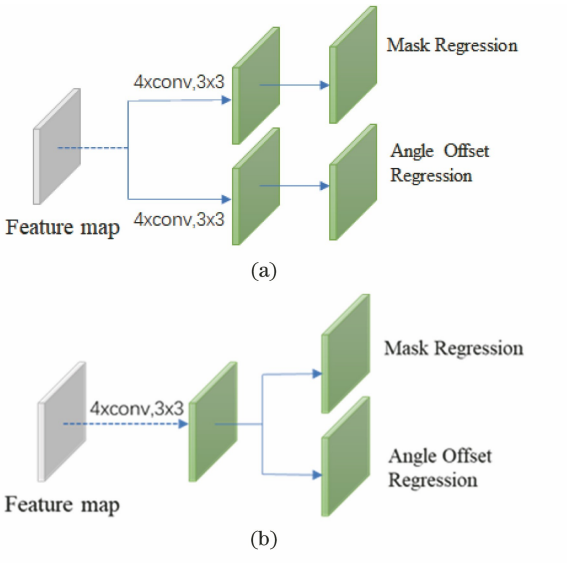

图 6　不同掩模预测子网络网络结构。(a)使用不同预测子网络；(b)使用同一预测子网络

Fig. 6　Network structures of different mask prediction subnetworks. (a) Using different prediction subnetworks (DPS); (b) using the same prediction subnetwork (SPS)

表 2　不同的掩模预测子网络的实验结果比较

Table 2　Comparison of experimental results of different mask prediction subnetworks　　unit: %

Method	Figure	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DPS	Fig. 6(a)	29.4	49.8	29.9	12.3	31.6	43.2
SPS	Fig. 6(b)	30.8	51.5	32.3	12.4	32.0	45.2

3.5　损失函数网络超参数对实验结果的影响

针对损失函数的权重超参数 λ_{segm} 与 λ_{angle} 的取值进行实验，损失函数的权重设计需要在一个数量级下进行操作，以保证不同的损失函数对于整体网络的贡献程度尽量相同。从实验中观察发现，语义分割子网络的损失 λ_{segm} 与其他部分的损失数量级相同，因此首先固定 λ_{angle} 的取值，确定 λ_{segm} 的大小，由表 3 的实验结果表明 λ_{segm} 值为 1.0 时效果更优。确定 λ_{segm} 的大小后，由于角度偏置预测的损失 λ_{angle} 比其他的损失稍大，因此分别通过实验验证 λ_{angle} 取值为 1.0、0.5、0.3、0.2、0.1 时对实验结果的影响。实验结果表 3 表明，λ_{segm} 值为 1，λ_{angle} 值为 0.2 时，能达到最高的实例分割精度。

表 3　采用不同的损失函数权重得到的实验结果

Table 3　Experimental results obtained by different loss function weights　　unit: %

λ_{segm}	λ_{angle}	AP	AP_{50}	AP_{75}
1.0	1.0	30.2	50.7	31.8
0.5	1.0	30.0	50.5	31.8
1.0	0.5	30.5	51.2	32.0
1.0	0.3	30.6	51.3	32.2
1.0	0.2	30.8	51.5	32.3
1.0	0.1	30.7	51.3	32.3

3.6　所提子网络对实验结果的影响

表 4 中的实验结果表明，在只使用语义分割子网络的情况下，平均准确度并没有提高，小尺寸和中尺寸实例的平均准确度小幅下降，大尺寸实例物体的平均准确度显著提高，其主要原因是语义分割子网络对大物体的分割能力比强，有效改进了实例的细节信息。但中小实例具有少量像素点，语义分割子网络经过多层的上采样难以提取其细节特征，导致性能结果下降。在只使用掩模预测子网络时，实验结果表明大中小尺度的实例平均准确度都有所提高，引入预测每个角度范围距离中心点最远的轮廓点，能最大限度地提取实例的轮廓，平均精度相对于基础网络提高了 0.3 个百分点。在同时使用语义分割子网络和掩模预测子网络的情况下，保证掩模预测的实例最大轮廓与语义分割子网络预测的分割图融合生成最终精细分割结果。实验结果表明，较基础网络 PolarMask，平均精度提升了 1.7 个百分点，大物体的实例分割性能提升最为显著，提升了 2.9 个百分点。

为了进一步验证基于轮廓点掩模细化的单阶段实例分割网络的有效性，对网络的各个部分的输出结果进行可视化，如图 7 所示。图 7(b)表示语义分割子网络输出的语义分割结果，包含实例丰富的细节信息；图 7(c)表示掩模预测子网络输出的实例分割结果，包含实例的最大轮廓；最后将掩模预测子网络的分割结果和语义分割子网络的分割结果融合，生成精细的实例分割结果，如图 7(d)所示。最终的实例分割结果具有语义分割结果的细节信息和掩模预测的实例信息，从而有效提升了实例分割结果。

使用提出的掩模预测子网络，通过预测最远轮廓的角度偏置和距离让掩模包含整个实例，对比

表 4　MS COCO 验证集下各个子网络的效果对比

Table 4　Comparison of each module under MS COCO-validation dataset　　unit：%

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	29.1	49.5	29.7	12.6	31.8	42.3
Baseline＋Semantic segmentation subnetwork	29.1	50.6	30.1	11.7	30.6	44.7
Baseline＋Mask prediction subnetwork	29.4	50.5	29.9	12.8	31.8	43.3
Ours	30.8	51.5	32.3	12.4	32.0	45.2

(a)　(b)　(c)　(d)

图 7　各个阶段分割结果。(a)原始图像；(b)语义分割子网络的分割结果；(c)掩模预测子网络的分割结果；(d)最终的实例分割结果

Fig. 7　Segmentation results of each stage. (a) Original image; (b) segmentation results of semantic segmentation subnetwork; (c) segmentation results of mask prediction subnetwork; (d) final instance segmentation results

图 8 中大象腿部等(圆圈区域)，发现本文方法相对于基础方法有较大的改善。再与语义分割结果相结合，生成最终精细的分割结果。相较于 PolarMask，本文方法通过去除实例中的背景、优化实例边缘信息，提高了平均准确度。针对形状较不规则的实例，效果提升更为明显，例如对于伸展状态的人，基础方法 PolarMask 很难包含人手和人脚等实例边缘(如图 8)，本文提出的办法可以有效改善这一问题。

(a)　(b)　(c)

图 8　不同方法的实例分割结果。(a)原始图像；(b) PolarMask；(c)本文方法

Fig. 8　Instance segmentation of different methods. (a) Original image; (b) PolarMask; (c) our method

3.7　实验结果对比

为了验证本文方法在实例分割任务上的优越性，在 MS COCO 数据集的测试集上评估本文方法，并将本文方法和其他先进方法进行比较。从表 5 中可以看出，在相同的实验设置情况下，本文方法相对于基础方法 PolarMask，平均准确度提高了 2.1 个百分点，并且本文方法在英伟达 GTX1070 显卡上的检测速度为 5 frame/s，基础方法 PolarMask 的检测速度为 6 frame/s，由此证明本文提出的方法对于改善实例分割结果有明显作用。

表 5　在 MS COCO 测试集下不同方法的性能比较

Table 5　Performance comparison of different methods under the MS COCO test dataset　　unit：%

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC[22]	Resnet101	24.6	44.3	24.8	4.7	25.9	43.6
FCIS[5]	Resnet101	29.2	49.5	-	7.1	31.3	50.0
YOLACT[12]	Resnet101	31.2	50.6	32.8	12.1	33.3	47.1
PolarMask[16]	Resnet101	30.4	51.9	31.0	13.4	32.4	42.8
Ours	Resnet101	32.5	53.6	34.3	13.1	34.3	48.0

图 9 展示了本文提出的网络在 MS COCO 测试集中的可视化结果，能够从图中看出对于小尺寸和大尺寸的实例，该网络都能够准确地将其轮廓检测出来，并且边缘轮廓比较清晰。

(a)　(b)　(c)

图 9　所提算法在 MS COCO 测试集中的结果

Fig. 9　Results of the proposed algorithm under the MS COCO test dataset

4　结　　论

提出了一种单阶段实例分割方法，包括四个子网络：骨干网络、特征金字塔子网络、掩模预测子网络和语义分割子网络。骨干网络和特征金字塔子网络用于生成不同尺度的特征层。基于生成的特征层，采用掩模预测子网络和语义分割子网络对不同尺度的实例进行分割。具体地来讲，首先采用掩模预测子网络分别预测实例轮廓点的距离和角度偏置，生成实例轮廓的最大包络；之后采用语义分割子网络通过融合多层级特征生成语义分割图；最终，基于实例最大包络和语义分割结果生成精细的实例分割结果。基于 MS COCO 数据集的相关实验表明，相对于基础方法 PolarMask，所提方法在不显著增加计算量的情况下将实例分割准确度提升 2.1 个百分点，证明本文所提算法的有效性。

参　考　文　献

[1]　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.

[2] 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.

[3] Hua X, Wang X Q, Wang D, et al. Multi-objective detection of traffic scenes based on improved SSD[J]. Acta Optica Sinica, 2018, 38(12): 1215003.
华夏, 王新晴, 王东, 等. 基于改进 SSD 的交通大场景多目标检测[J]. 光学学报, 2018, 38(12): 1215003.

[4] Zhu T Y, Dong F, Gong H X. Remote sensing building detection based on binarized semantic segmentation[J]. Acta Optica Sinica, 2019, 39(12): 1228002.
朱天佑, 董峰, 龚惠兴. 基于二值语义分割网络的遥感建筑物检测[J]. 光学学报, 2019, 39(12): 1228002.

[5] Li Y, Qi H Z, Dai J F, et al. Fully convolutional instance-aware semantic segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4438-4446.

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

[7] Huang Z J, Huang L C, Gong Y C, et al. Mask scoring R-CNN[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 6402-6411.

[8] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[9] de Brabandere B, Neven D, van Gool L. Semantic instance segmentation with a discriminative loss function[EB/OL]. (2017-08-08) [2020-06-04]. https://arxiv.org/abs/1708.02551.

[10] Liu S, Jia J Y, Fidler S, et al. SGN: sequential grouping networks for instance segmentation[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 3516-3524.

[11] Gao N Y, Shan Y H, Wang Y P, et al. SSAP: single-shot instance segmentation with affinity pyramid[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 642-651.

[12] Bolya D, Zhou C, Xiao F Y, et al. YOLACT: real-time instance segmentation [C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9156-9165.

[13] Wang X, Kong T, Shen C, et al. SOLO: segmenting objects by locations[EB/OL]. (2019-12-10)[2020-06-04]. https://arxiv.org/abs/1912.04488.

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

[15] Chen X L, Girshick R, He K M, et al. TensorMask: a foundation for dense object segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 2061-2069.

[16] Xie E, Sun P, Song X, et al. PolarMask: single shot instance segmentation with polar representation[EB/OL]. (2019-09-29)[2020-06-04]. https://arxiv.org/abs/1909.13226.

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2020-06-04]. https://arxiv.org/abs/1409.1556.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[20] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[21] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

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