第 59 卷 第 16 期/2022 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于双流卷积神经网络的人体实例分割

马子彤，王国栋*

青岛大学计算机科学技术学院，山东 青岛 266071

摘要 人体实例分割是以人为中心的场景理解和识别的核心问题。然而人体实例体型的差异性、人们之间的互动等导致了空间关系的复杂性，给分割任务带来了极大的挑战。目前主流的实例分割方法大都严重依赖物体的边界框检测，因此通常无法很好地将两个高度重合的对象分开。利用已经具有完备数据标注的人体骨骼特征为人体实例分割任务提供先验知识，提出了一种双流的网络结构，用来分别提取骨骼特征和图片上下文特征。接着，特征融合模块(FFB)自适应地融合来自不同流的特征并将其送入分割模块，得到最终的分割结果。实验结果表明，所提算法在COCOPersons、OCHuman数据集上的平均精确度分别为59.5%、56.7%，相比其他算法均有一定的提升。

关键词 图像处理；卷积神经网络；双流卷积神经网络；注意力机制

中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP202259.1610004

Human Instance Segmentation Based on Two-Stream Convolutional Neural Network

Ma Zitong, Wang Guodong*

College of Computer Science & Technology, Qingdao University, Qingdao 266071, *Shandong, China*

Abstract Segmentation of human instances is a fundamental problem in human-centered scene understanding and recognition. However, due to the diversity of human body shapes and interactions, spatial relations become complex, posing significant challenges for segmentation tasks. At the moment, most of the mainstream instance segmentation methods rely heavily on the boundary box detection of objects, and thus, are usually unable to effectively separate two highly overlapping objects. In this paper, human skeleton features with complete data annotation are used to provide a priori knowledge for the human instance segmentation task, and a two-stream network structure is proposed to extract skeleton and context features, respectively. The feature fusion module (FFB) then adaptively combines the features from different streams and sends them into the segmentation module, where the final segmentation result is obtained. The proposed algorithm's average accuracy on the COCOPersons and OCHuman datasets is 59.5% and 56.7%, respectively, which is improved better than other algorithms.

Key words image processing; convolutional neural network; two-stream convolutional neural network; attention mechanism

1 引 言

实例分割[1-6]为图像中的每个像素点分配一个标签，同时区分相同类别的不同个体。近年来，随着自动驾驶[7]、人脸识别[8-10]、姿态识别[11-15]等技术的迅猛发展，计算机视觉中有关“人”的研究受到广泛关注。目前有关人体实例分割主要面临两个挑战：首先，人类的身高和体型相差较大，在姿态上也具有可变性；其次，在日常情境中，人类个体通常会与其他个体或者物品产生关联，从而导致遮挡现象。

当前的实例分割方法大多是基于Mask R-CNN[1]的，也就是基于检测框的分割。具体来说，首先用一个检测器检测出物体所在位置，然后在检测框内进行分割。然而这类方法存在如下弊端：首先，如果检测框本身不准确，比如没有完全覆盖物体，那么就算框内的分割做得再好也无法得到正确的实例掩

收稿日期：2021-05-18；**修回日期**：2021-06-15；**录用日期**：2021-06-27

基金项目：山东省自然科学基金(ZR2019MF050)、山东省高等学校优秀青年创新团队支持计划(2020KJN011)

通信作者：*doctorwgd@gmail.com

模；其次，这类方法通常会在第一阶段生成的众多检测框中使用非极大值抑制（NMS），从而只保留置信度最高的候选框。因此当多个目标发生重叠时，极有可能出现漏检的情况。为了解决上述问题，Zhang等[16]提出了一种利用人体姿态特征生成实例候选区域，然后在候选区域内进行实例分割的方法。但是如何有效地嵌入姿态表示并将其与图像特征相结合，实现人体实例的精确分割是一个有待解决的问题。不加区分地直接结合骨架特征仍然不能得到令人满意的结果。

本文提出了一种双流网络结构，在传统实例分割任务中有效引入骨骼信息。其中，姿态流更加关注人体骨骼信息，如位置信息和方向信息，图片流则主要关注图片上下文特征，两个流并行提取对应特征并使用特征融合模块（FFB）进行加权融合，最后将融合结果输入分割模块进行分割。所提方法能充分利用局部特征和全局特征，从而提高模型的表征能力。

2 基本原理

2.1 人体实例分割

实例分割既具备语义分割的特点，即在像素层面上的分类，也具备目标检测的特点，即需要定位不同实例。当前主流的实例分割方法分为自上而下[1]的基于检测的分割和自下而上[16]的基于像素的分割。自上而下的方法例如Mask R-CNN[1]，在原有Faster R-CNN[5]的分类和位置回归两个并行分支外再加入一个实例分割的并行分支，生成二值掩模。自下而上的分割方法先对每个像素进行分类标记，然后通过聚类等方法将像素分类到对应的实例中。这类方法通常需要更大的计算能力。

许多方法尝试将人体姿态估计与实例分割相结合，以更好地解决人体实例分割问题。Mask R-CNN[1]将目标检测、姿态估计及实例分割集合到一个模型中，但是并没有取得令人满意的效果。Pose2Instance[17]在自上而下模型的区域生成网络（RPN）与掩模预测头之间加一个姿态检测器，预测关键点，然后其直接与CNN特征堆叠输入掩模预测头中进行分割。PersonLab[18]首次使用自下而上的方法进行实例分割，即首先检测关键点，再将关键点聚类成人体实例。类似的，Pose2Seg[16]设计了一个级联模型，很好地解决了重合人像的实例分割问题，精度超越了Mask R-CNN，并且提出了一个新的具有挑战性的数据集OCHuman。

2.2 姿态表示

2D人体姿态估计定位并检测出人体关键点，将关键点按照关节顺序相连，得到人体的躯干，进而得到关于人体的单人姿态估计（SPPE）和多人姿态估计（MPPE）两个部分。

在深度学习之前，传统方法一般使用方向梯度直方图（HOG）[19]、形状匹配之形状上下文（shape context）[20]、尺度不变特征变换（SIFT）[21]等浅层手工设计特征来对特征建模，主要是希望解决单人人体姿态估计问题。随着AlexNet[22]的出现，深度学习开始快速发展。在2013年，Jain等[12]第一次引入CNN来解决单人姿态估计的问题，并且沿用了传统骨架的思路。Openpose[11]联合关键点置信图（PCM）和部分亲和场（PAF）进行姿态表示，解决了CPM[12]中多人肢体连接的问题。其中PCM预测人体关键点的热度图（joints heatmap），可以看到每个人体关键点上都有一个高斯的峰值，代表网络预测出这里是一个人体的关键点。PAF可以看作是记录关节位置和方向的2D向量，学习身体部位和对应个体的关联。在本文中，为了更好地定位人体实例，引入Openpose[11]中定义的PCM和PAF的真实值来表征姿态信息。

2.3 注意力机制

计算机视觉中的注意力机制就是想让机器能够关注重点信息而忽略无关信息，实质就是学习权重分布，主要分为两种，一种是软注意力，这是一种确定的注意力，而且是可微的，即可以通过神经网络的前向传播和后向反馈来学习权重分布。它对关注的不同通道域[23]、空间[24]等赋予不同的权重，抽取出更加关键的信息以使模型做出更加精准的判断。另一种是硬注意力，是一个随机的预测过程并且是不可微的，它更加关注点，即对特征图上每个点进行加权。

软注意力被广泛应用于计算机视觉的相关领域，如检测[25]、分割任务等。在人体实例分割任务中，人类通常处于背景复杂、遮挡严重的图像中，传统的实例分割算法通常无法很好地处理复杂人体姿态与背景的关系。所提算法通过构造FFB，自动地去学习姿态流、图片流的权重，进而精炼来自不同流的输出特征，进一步提升分割效果。

3 算法实现

3.1 整体结构

所提算法的整体结构如图1所示。

1）首先将图片输入图1（b）中图片流的主干网络，提取特征。图1（b）采用50层的ResNet[26]结合特征金字塔网络（FPN）[27]，融合多层特征。其中采用卷积核为7×7、步长为2的卷积层Conv1进行浅层特征提取；为防止梯度爆炸，在每一个卷积层后加入BN层进行批量归一化；最大池化层采用步长为2的3×3卷积操作；Conv2～5是基于瓶颈网络结构块（bottleneck block）的，目的是通过参数量的降低减少训练时间。每个瓶颈结构块涉及一个减少通道数的1×1卷积核、一个提取特征的3×3卷积核、最后恢复维度的1×1卷积核，Conv2～5层分别堆叠3、4、6、3次。FPN自下而上的路径以ResNet50的Conv2～5层产生的不同尺度

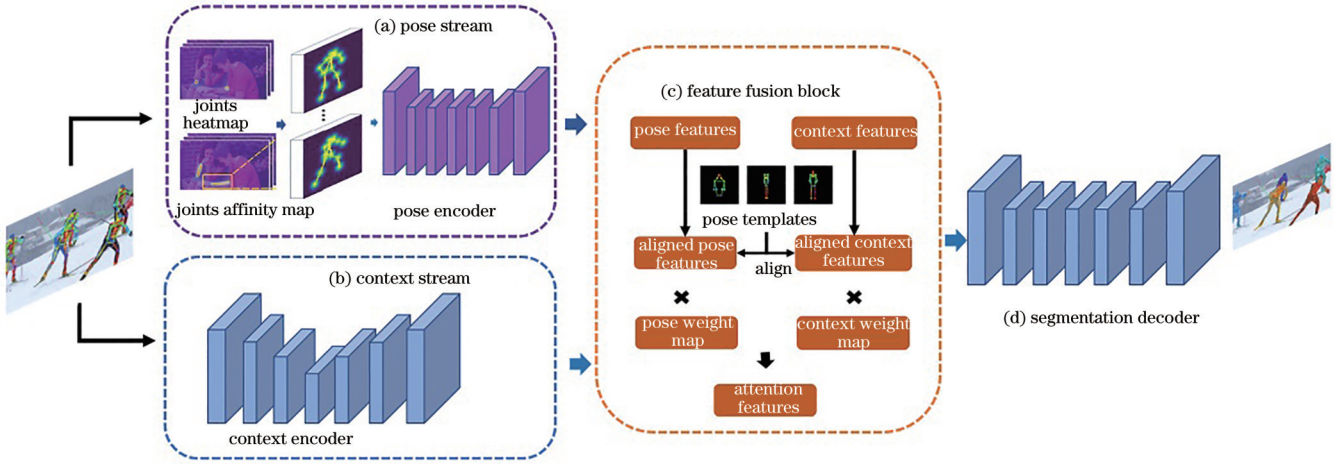

图 1　网络整体结构

Fig. 1　Overall network structure

低级语义特征图$\{\boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{C}_4, \boldsymbol{C}_5\}$作为输入，自上而下的路径对$\{\boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{C}_4, \boldsymbol{C}_5\}$进行二倍上采样后得到$\{\boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5\}$，然后将其与1×1大小的卷积核处理过的$\{\boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{C}_4, \boldsymbol{C}_5\}$对应层特征融合。

2）将PAF、PCM的真实值输入图1(a)中姿态流的网络中，以提取人体姿态骨骼特征。由于骨骼特征是手工设计的，本文希望将手工设计特征输入神经网络中，以提取到更高级和普遍的特征。因此这里仅使用较浅层的网络结构：在步长为2、卷积核大小为7×7的卷积操作后接5个瓶颈网络结构块。

3）在图1(c)，首先使用基于人体骨骼姿态的仿射对齐(Affine-Align)[16]将图片中的姿态与姿态模板对齐，并将每个实例区域对齐为统一的大小64 × 64。这样做的一个优点是可以调整图片中一些奇怪的姿势，从而降低神经网络的学习难度。然后将对齐后的图片特征及骨骼特征拼接起来送入FFB，通过学习对应的注意力向量后加权求和，得到融合后的特征表示。

4）接下来，将融合后的特征传递给图1(d)分割模块，以生成每个感兴趣区域(RoI)的实例分割。分割模块是一种简单的编码器-解码器体系结构，是基于校准的RoI的分辨率进行设计的。类似于图1(a)，不同的是这里需要堆叠10个瓶颈网络结构块以获得足够的感受野。然后通过上采样恢复分辨率，再通过一个瓶颈网络结构块后接1 × 1大小的卷积核得到分割结果，最后根据仿射变换[16]得到的矩阵$\boldsymbol{H}$反转对齐，将每个实例的分割掩模组合成一个最终分割掩模。

3.2　双流网络结构

由于每张图片中都可能包含未知数量、未知位置的人，同时人们之间的交互性、场景内容的多样性导致了空间关系的复杂性。为了获得高质量的分割结果，提出一种基于双流卷积网络的模型：图片流用于提取图片上下文特征，姿态流引入包含精准位置信息的人体骨骼姿态来定位图像中的人体实例，最终对两个分支的结果进行加权融合，更加准确地定位实例位置，获得最后的分割结果。

对于姿态流，通过PCM[11]、PAF[11]将人体骨骼姿态转换为一种骨骼特征图，从而有效解决严重遮挡情况下的实例分割问题。PCM是根据标记的2D关键点信息生成的一组以关键点为中心的高斯分布。每一张热力图都表示不同人的同一个关键点，图中有多少个关键点就对应多少个PCM，缺少的关键点用0补齐。对于检测到的关键点，为了正确地连接它们，组装成完整的人体骨骼姿态，引入Openpose[11]中定义的PAF真实值。PAF对应于每个肢体的每个像素点的一组既包含位置信息，又包含方向信息的二维向量。对于不同的肢体部位，都有对应的PAF与它相关的两个关键点的联系，从而得到完整的人体信息。使用的数据集是COCOPersons数据集，对每个人体实例定义了19个肢体、17个关键点，其中关键点包括眼睛、鼻子、耳朵、肩膀、肘部、手腕、臀部、膝盖和脚踝。除了鼻子外，其他均有左右之分。因此对于每个人体实例，PAFs是一个38通道的特征图，PCMs是一个17通道的特征图。

由于骨骼特征着眼于局部信息，缺少与全局特征的结合分析，对人体实例定位的精确度仍有不足。下一步引入FFB融合图片流和姿态流，以获得更为准确的分割结果。

3.3　特征融合模块

为了充分获取人体姿态信息，使用双流卷积神经网络建构特征描述后，通过一种自适应特征融合模块优化人体实例分割。FFB以自适应的方式融合图片特征和骨骼特征，即在训练过程中通过梯度反向传播动态地调整两种特征的权重，并按照所学到的权重去

提升有用的特征同时抑制无关特征。对于图片与人体骨架信息，分别在通道维度上学习不同特征。其中，0代表图片通道，1代表姿态通道，W、H、C分别代表特征图的宽、高、通道数。使用仿射对齐模块[16]，具体操作如下：首先根据COCOPersons数据集中标注的边框信息，从特征图中提取每个实例的RoI；从姿态模板中匹配与当前姿态$\boldsymbol{P}$适配度最高的模板$\boldsymbol{P}_{\mathrm{u}}$，即得分(score)最高的模板，通过计算最佳仿射变换矩阵$\boldsymbol{H}^{*}$，将RoIs变换到相同的旋转角度进行对齐。其中$\boldsymbol{H}$是一个包含旋转、缩放、x轴平移、y轴平移、左右翻转的2×3矩阵，计算方法为

$$\boldsymbol{H}^{*}=\arg\min_{H}\left(\boldsymbol{H}\cdot\boldsymbol{P}-\boldsymbol{P}_{\mathrm{u}}\right), \tag{1}$$

$$s_{\mathrm{score}}=\exp\left(-\left\|\boldsymbol{H}^{*}\cdot\boldsymbol{P}-\boldsymbol{P}_{\mathrm{u}}\right\|\right)。 \tag{2}$$

将$\boldsymbol{H}$应用于RoIs并使用双线性插值到固定的大小，得到对齐后的特征图$\boldsymbol{F}_{\mathrm{p}},\boldsymbol{F}_{\mathrm{c}}\in\mathbf{R}^{W\times H\times C}$，其中$\boldsymbol{F}_{\mathrm{p}},\boldsymbol{F}_{\mathrm{c}}$分别代表对齐后的姿态特征和图片特征，这相当于一个正则化操作，简化了图像分布。将对齐后的特征拼接起来(Concatenate)，经3×3大小的卷积核提取特征，利用Softmax输出每一个通道的权重$\boldsymbol{A}_{0},\boldsymbol{A}_{1}$，并与$\boldsymbol{F}_{\mathrm{p}},\boldsymbol{F}_{\mathrm{c}}$进行加权融合，得到融合后的特征表示，公式为

$$\boldsymbol{A}_{0},\boldsymbol{A}_{1}=\sigma\left[\mathrm{F}_{3\times3}\left(\boldsymbol{F}_{\mathrm{c}}\oplus\boldsymbol{F}_{\mathrm{p}}\right)\right], \tag{3}$$

$$\boldsymbol{X}=\mathrm{F}_{3\times3}\left(\boldsymbol{A}_{0}\otimes\boldsymbol{F}_{\mathrm{c}}+\boldsymbol{A}_{1}\otimes\boldsymbol{F}_{\mathrm{p}}\right), \tag{4}$$

式中：σ代表Softmax函数；$\oplus$表示拼接操作；+表示逐元素相加；$\mathrm{F}_{3\times3}$表示卷积核大小为3×3的卷积操作。

4 实验结果与分析

4.1 实验参数及数据集

使用的深度学习框架是Pytorch，在训练时使用NVIDIA GeForce GTX 1080 Ti显卡，每批次输入4张分辨率为512×512大小的图片。初始学习率设置为2×10^{-4}，使用Adam方法学习模型中的参数，用He等[28]提出的方法进行初始化。

所有实验训练和测试均是在具有丰富的人体标注信息的COCOPersons数据集上进行的。它既包含人体分割的掩模标注，又包含人体姿态关键点位置和可见性的标注，但是实例间的遮挡情况较轻。为了进一步验证所提算法在严重遮挡情况下的分割效果，同时在更加具有挑战意义的OCHuman数据集上进行了测试。OCHuman数据集由Zhang等[16]提出，通过MaxIoU衡量相同类别个体遮挡程度的严重性。该数据集包含8110个人体实例，人体实例的MaxIoU全部都在0.5以上，整个数据集的平均MaxIoU达0.67。同时，根据遮挡的严重程度，该数据集又分为OCHuman-Moderate（MaxIoU为0.5～0.75）和OCHuman-Hard(MaxIoU在0.75以上)两个子集。

4.2 评估指标

使用平均精确度(AP)来评估算法的性能。P(precision)为准确率，R(recall)为召回率。PR曲线下的面积，就是AP。mAP(mask AP)即掩模平均精度值，计算公式为

$$p_{\mathrm{mAP}}=\int_{0}^{1}p(r)\mathrm{d}r。 \tag{5}$$

通过计算不同交并比(IoU)的阈值平均精度来评估分割效果，其中$\mathrm{AP_M}$，$\mathrm{AP_L}$代表对图片中不同大小的人体实例计算所得的平均精度，$\mathrm{AP_H}$代表对OCHuman-Hard子集计算所得的平均精度。对于网络输出掩模A和掩模真实值B的交并比，计算公式为

$$p_{\mathrm{IoU}}(A,B)=\frac{|A\cap B|}{|A\cup B|}。 \tag{6}$$

4.3 损失函数

使用交叉熵函数作为损失函数来度量分割掩模真实值和实际生成的掩模之间的距离，计算公式为

$$L(\hat{y},y)=-\left[y\log\hat{y}+(1-y)\log(1-\hat{y})\right], \tag{7}$$

式中：$\hat{y}$代表预测值；y代表真实值。

4.4 消融实验

4.4.1 添加姿态流对分割结果的影响

为了验证引入的人体骨骼姿态信息对人体实例分割性能的影响，实验设置如下：One Stream代表不引入姿态信息，Two Streams代表所提双流网络，GT Kpt代表输入的是关键点的真实值。实验结果表明丰富的人体姿态标注信息有助于网络关注某个特定的个体，更适用于区分图像中高度重叠的人体实例。实验结果如表1所示，与未引入姿态信息相比，所提双流网络的AP值提高了25.3个百分点。图2为分割结果，表明所提算法在处理实例边界问题上具有一定的优势。

表1 在COCOPersons验证集上的消融实验结果

Table 1 Results of ablation study on COCOPersons validation set

Training method		AP	$\mathrm{AP_M}$	$\mathrm{AP_L}$
One Stream		0.342	0.333	0.385
Two Streams	+Concatenate	0.588	0.542	0.690
	+Max Pooling	0.591	0.545	0.689
	+Avg Pooling	0.584	0.542	0.684
	+SENet[23]	0.593	0.545	0.693
	+CA[29]	0.592	0.548	0.690
	+FFB(Ours)	0.595	0.548	0.697

4.4.2 不同特征融合方式对分割结果的影响

为了更好地分析特征融合方式对分割效果的影响，将FFB与常见的特征融合方式最大池化(Max Pooling)、平均池化(Average Pooling)以及拼接(Concatenate)进行了对比，同时也和注意力机制SENet[23]以及CA[29]进行了比较。所提算法的AP达

图 2　在 COCOPersons 数据集上的消融实验效果对比图

Fig. 2　Comparison of ablation study results on COCOPersons dataset

到了 0.595，相比其他方法，提高了约 1 个百分点，充分证明了所提算法可以更有效地强化目标区域的特征提取，实验结果如表 1 所示。

4.5　对比实验

为了评估所提算法在不同数据集上的泛化能力，分别在 OCHuman、COCOPersons 数据集上进行了测试，实验结果如表 2、3 所示，分割效果如图 3、4 所示。

5　结　　论

提出了一种双流神经网络模型，将 ResNet50-fpn 应用于图片流 CNN，将人体骨骼姿态特征引入姿态流 CNN，使得模型更适用于提取人体实例特征。然后对图片流和姿态流的输出进行加权融合，作为双流神经网络的输出结果，并输入分割模块中进行分割，最终实

表 2　不同算法在 OCHuman 数据集上的分割效果

Table 2　Segmentation results of different algorithms on OCHuman dataset

Method	Dataset	Backbone	AP	AP_H
Mask R-CNN[1]	OCH val	ResNet50-fpn	0.163	0.113
	OCH test	ResNet50-fpn	0.169	0.128
Pose2Seg(GT Kpt)[16]	OCH val	ResNet50-fpn	0.544	0.491
	OCH test	ResNet50-fpn	0.552	0.495
Pose2Seg	OCH val	ResNet50-fpn	0.222	0.150
	OCH test	ResNet50-fpn	0.238	0.175
Ours	OCH val	ResNet50-fpn	0.573	0.576
	OCH test	ResNet50-fpn	0.567	0.570

表 3　不同算法在 COCOPersons 数据集上的分割效果

Table 3　Segmentation results of different algorithms on COCOPersons dataset

Method	Backbone	AP	AP_M	AP_L
Mask R-CNN[1]	ResNet50-fpn	0.532	0.433	0.648
PersonLab[18]	ResNet101		0.476	0.592
PersonLab	ResNet101(ms scale)		0.492	0.621
PersonLab	ResNet152		0.483	0.595
PersonLab	ResNet152(ms scale)		0.497	0.621
Pose2Seg(GT Kpt)[16]	ResNet50-fpn	0.582	0.539	0.679
Pose2Seg	ResNet50-fpn	0.555	0.498	0.670
Ours	ResNet50-fpn	0.595	0.548	0.697

图 3 在 OCHuman 数据集上的分割效果对比

Fig. 3 Comparison of segmentation results on OCHuman dataset

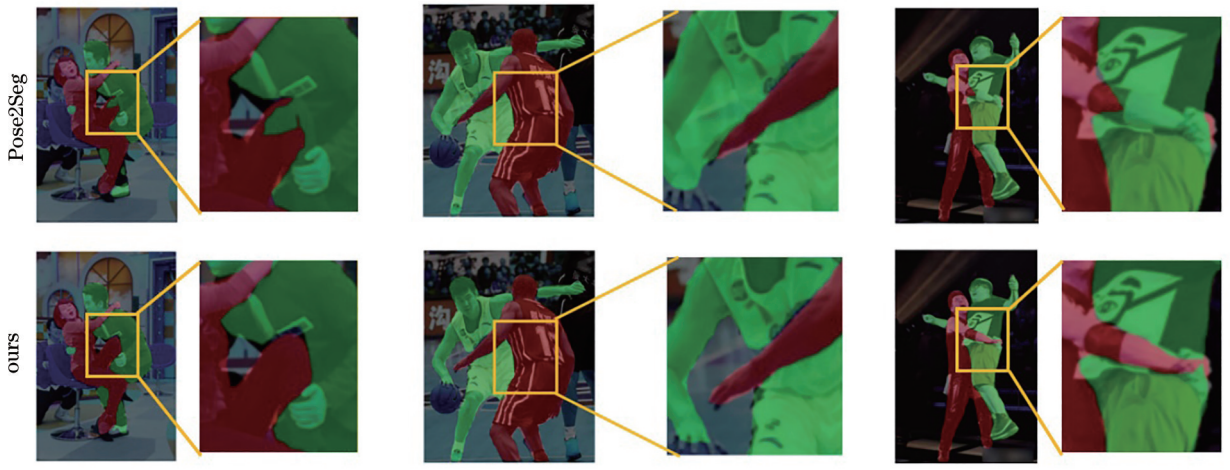

图 4 在 COCOPersons 数据集上的分割效果对比

Fig. 4 Comparison of segmentation results on COCOPersons dataset

现人体实例分割。为了证明所提算法的有效性，在 COCOPersons、OCHuman 数据集上进行了验证，平均精度均高于其他算法，充分证明了所提算法的鲁棒性。

参 考 文 献

[1] 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 Press, 2017: 2980-2988.

[2] 李琪琪, 花向红, 赵不钒, 等. 一种室内场景点云平面分割的新方法[J]. 中国激光, 2021, 48(16): 1604002.
Li Q Q, Hua X H, Zhao B F, et al. New method for plane segmentation of indoor scene point cloud[J]. Chinese Journal of Lasers, 2021, 48(16): 1604002.

[3] Chen H, Sun K Y, Tian Z, et al. BlendMask: top-down meets bottom-up for instance segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 8570-8578.

[4] 张绪义, 曹家乐. 基于轮廓点掩模细化的单阶段实例分割网络[J]. 光学学报, 2020, 40(21): 2115001.
Zhang X Y, Cao J L. Contour-point refined mask prediction for single-stage instance segmentation[J]. Acta Optica Sinica, 2020, 40(21): 2115001.

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

[6] Cai Z W, 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.

[7] Zhang Z Y, Fidler S, Urtasun R. Instance-level segmentation for autonomous driving with deep densely connected MRFs[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 669-677.

[8] George A, Marcel S. Cross modal focal loss for RGBD face anti-spoofing[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 7878-7887.

[9] Shi Y C, Yu X, Sohn K, et al. Towards universal representation learning for deep face recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020,

Seattle, WA, USA. New York: IEEE Press, 2020: 6816-6825.

[10] Ma X, Zhang F D, Li Y L, et al. Robust sparse representation based face recognition in an adaptive weighted spatial pyramid structure[J]. Science China Information Sciences, 2017, 61(1): 1-13.

[11] Cao Z, Simon T, Wei S H, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1302-1310.

[12] Jain A, Tompson J, Andriluka M, et al. Learning human pose estimation features with convolutional networks[EB/OL]. (2014-04-23) [2021-05-17]. https://arxiv.org/abs/1312.7302.

[13] Artacho B, Savakis A. UniPose: unified human pose estimation in single images and videos[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7033-7042.

[14] Lifkooee M Z, Liu C L, Liang Y Q, et al. Real-time avatar pose transfer and motion generation using locally encoded Laplacian offsets[J]. Journal of Computer Science and Technology, 2019, 34(2): 256-271.

[15] Newell A, Huang Z A, Deng J. Associative embedding: end-to-end learning for joint detection and grouping[EB/OL]. (2016-11-16) [2021-03-04]. https://arxiv. org/abs/1611.05424.

[16] Zhang S H, Li R L, Dong X, et al. Pose2Seg: detection free human instance segmentation[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: 889-898.

[17] Tripathi S, Collins M, Brown M, et al. Pose2Instance: harnessing keypoints for person instance segmentation [EB/OL]. (2017-04-04) [2021-05-04]. https://arxiv. org/abs/1704.01152.

[18] Papandreou G, Zhu T, Chen L C, et al. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 282-299.

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

[20] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509-522.

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

[22] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [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: 7132-7141.

[24] Jimmy B, Ryan K, Kyunghyun C, et al. Tell: neural image caption generation with visual attention[EB/OL]. (2015-02-10)[2021-03-04]. https://arxiv.org/abs/1502.03044.

[25] 崔海华, 漏华铖, 田威, 等. 轨道式爬行机器人制孔基准的视觉高精度定位[J]. 光学学报, 2021, 41(9): 0915002.
Cui H H, Lou H C, Tian W, et al. High-precision visual positioning of hole-making datum for orbital crawling robot[J]. Acta Optica Sinica, 2021, 41(9): 0915002.

[26] 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, June 26-July 1, 2016, Las Vegas, USA. New York: IEEE, 2016: 770-778.

[27] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2117-2125.

[28] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1026-1034.

[29] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13708-13717.