软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(7):1959–1968 [doi: 10.13328/j.cnki.jos.005933]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

联合姿态先验的人体精确解析双分支网络模型*

高明达[1,2], 孙玉宝[1,2], 刘青山[1,2], 邵晓雯[1,2]

[1](江苏省大数据分析技术重点实验室 (南京信息工程大学 自动化学院),江苏 南京 210044)

[2](江苏省大气环境与装备技术协同创新中心(南京信息工程大学 自动化学院),江苏 南京 210044)

通讯作者: 孙玉宝, E-mail: sunyb@nuist.edu.cn

摘 要: 人体解析旨在将人体图像分割成多个具有细粒度语义的部件区域,进行形成对人体图像的语义理解.然而,由于人体姿态的复杂性,现有的人体解析算法容易对人体四肢部件形成误判,且对于小目标区域的分割不够精确.针对上述问题,联合人体姿态估计信息,提出了一种人体精确解析的双分支网络模型.该模型首先使用基干网络表征人体图像,将人体姿态估计模型预测到的姿态先验作为基干网络的注意力信息,进而形成人体结构先验驱动的多尺度特征表达,并将提取的特征分别输入至全卷积网络解析分支与检测解析分支.全卷积网络解析分支获得全局分割结果,检测解析分支更关注小尺度目标的检测与分割,融合两个分支的预测信息可以获得更为精确的分割结果.实验结果验证了该算法的有效性,在当前主流的人体解析数据集 LIP 和 ATR 上,所提方法的 mIoU 评测指标分别为 52.19%和 68.29%,有效提升了解析精度,在人体四肢部件以及小目标部件区域获得了更为准确的分割结果.

关键词: 人体解析;语义分割;人体姿态估计;部件检测;卷积神经网络

中图法分类号: TP391

中文引用格式: 高明达,孙玉宝,刘青山,邵晓雯.联合姿态先验的人体精确解析双分支网络模型.软件学报,2020,31(7):1959–1968. http://www.jos.org.cn/1000-9825/5933.htm

英文引用格式: Gao MD, Sun YB, Liu QS, Shao XW. Posture prior driven double-branch network model for accurate human parsing. Ruan Jian Xue Bao/Journal of Software, 2020,31(7):1959–1968 (in Chinese). http://www.jos.org.cn/1000-9825/5933.htm

Posture Prior Driven Double-branch Network Model for Accurate Human Parsing

GAO Ming-Da[1,2], SUN Yu-Bao[1,2], LIU Qing-Shan[1,2], SHAO Xiao-Wen[1,2]

[1](Jiangsu Key Laboratory of Big Data Analysis Technology (School of Automation, Nanjing University of Information Science and Technology), Nanjing 210044, China)

[2](Jiangsu Province Atmospheric Environment and Equipment Technology Collaborative Innovation Center (School of Automation, Nanjing University of Information Science and Technology), Nanjing 210044, China)

Abstract: Human parsing aims to segment a human image into multiple parts with fine-grained semantics and provides more detailed understanding of image contents. When the human body posture is complicated, the existing human parsing methods are easy to cause misjudgment to the human limb components, and the segmentation of the small target is not accurate enough. In order to solve the above problems, a double-branch networkjointingposture prior is proposed for accurate human parsing. The model first uses the backbone network to acquire the characteristics of the human body image, and then uses the pose prior information predicted by the human pose estimation model as the attention information to form a multi-scale feature expression driven by the human body structure prior. The multi-scale features are fed into the fully convolution network parsing branch and detection parsing branch separately. The fully

* 基金项目: 国家自然科学基金(61825601, 61532009, 61672292); 江苏省级项目(BRA2019077, DZXX-037)

Foundation item: National Natural Science Foundation of China (61825601, 61532009, 61672292); Jiangsu Provincial Project (BRA2019077, DZXX-037)

本文由“多媒体内容的多维度相似性计算与搜索”专题特约编辑蒋树强研究员、刘青山教授、孙立峰教授、李波教授推荐.

收稿时间: 2019-04-30; 修改时间: 2019-07-11; 采用时间: 2019-09-17; jos 在线出版时间: 2020-01-13

CNKI 网络优先出版: 2020-01-14 11:25:39, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.1125.016.html

convolutional network obtains global segmentation results, and the detection parsing branch pays more attention to the detection and segmentation of small-scale targets. The segmentation results of the two branches are fused to obtain the final parsing result, which can be more accurate. The experiment results verify the effectiveness of the proposed algorithm. Our Thisapproach has achieved 52.19% mIoU on LIP dataset, 68.29% mIoU on ATR dataset, which improves the human parsing accuracy effectively and achieves more accurate segmentation results in the human limb components and small target componentsn parsing accuracy effectively and achieves more accurate segmentation results in the human limb components and small target components.

Key words: human parsing; semantic segmentation; human pose estimation; object detection; convolution neural network

人体解析旨在将图像中的人体部件(例如头发,帽子等)进行精细化分割,是一项细粒度级别的语义分割任务.人体解析任务具有重要的研究意义,能够使机器更好地理解以人为中心的图像内容,也可为行人重识别[1,2]、行为识别[3]、视频事件检测、智能安防等相关任务的发展提供技术支撑.然而,人体解析仍是一个具有挑战性的工作,一方面主要体现在现实场景中以人为中心的图像具有复杂的背景,将人体的部件从复杂背景精确分割具有一定的难度;另一方面,人体图像中存在的遮挡以及人体姿态变化幅度大等问题都会增加人体解析任务的挑战性.

人体解析任务与语义分割[4]具有内在的关联性,语义分割方法的发展对于人体解析任务的研究也具有积极的借鉴作用.近年来,深度卷积网络在图像分类、目标检测等许多视觉问题中获得了广泛的应用,有效推动了这些问题的研究进展.许多研究者也将卷积神经网络推广应用于语义分割任务.Long 等人[5]将全卷积神经网络应用于语义分割任务,并提出跳跃结构,将不同尺寸的特征相互融合,分割效果得到显著提升.Chen 等人[6]提出空洞卷积并将其应用到语义分割,空洞卷积实现了在保存语义特征分辨率的前提下提高特征的语义信息.Chen 等人[7]在以往工作的基础上,基于空洞卷积提出了一种空间金字塔状的卷积操作,有效地提取多尺度特征,并且结合 Conditional Random Fields(CRF)后端处理,进一步推动语义分割领域的发展.

语义分割网络的研究积累也促进了人体分割算法的发展.Liang 等人[8]提出了一种上下文相关的卷积神经网络框架,该模型可以同时获取图像多层的上下文信息,有效提升了模型对衣服特征的辨别能力.Liang 等人[9]提出一种深度局部-全局长短记忆模型,该方法可以在特征提取时有效地利用像素位置的空间依赖关系.Chen 等人[10]利用了多尺度信息,并提出了一种注意力机制以更好地将多尺度特征进行融合.

人体解析任务主要是以人体为研究对象,为了达到更好的分割精度,需要充分利用人体结构信息.Gong 等人[11]提出了一种名为 SSL 的自监督损失函数,该方法使用的人体结构信息是通过直接提取对应的语义分割区域中心而得到,巧妙地将人体结构信息应用到人体解析任务,但是该方法提取到的人体姿态信息较为粗糙,当人体姿态复杂时分割效果不理想.Liang 等人[12]提出了 JPPNet,将人体姿态估计与人体解析两个任务有效地结合到一起,该网络结构在预测人体姿态信息的同时进行人体解析任务,并将两个任务的预测结果进行多次融合迭代,进而达到相互辅助与相互促进的作用.虽然该方法有利于改进四肢部件的分割效果,但在增强语义特征的过程中特征图分辨率会有所降低,造成像素位置信息的损失[13],小目标人体部件分割性能不尽人意.总体而言,尽管现有方法在人体解析任务中取得了一定成果,但是,由于人体姿态的复杂性,当前方法[10,11,14,15]往往不能准确理解人体的结构信息,对于四肢等人体部件容易造成误判,同时对于小尺度目标区域的分割精度不够理想,如何建立更为准确的人体解析模型仍是需要深入研究的难点问题.

针对上述问题,本文提出了一种联合姿态先验的双分支人体精确解析方法.该方法通过估计人体姿态先验信息,更好地理解人体各个部件之间的联系,为后续解析任务提供指导信息.并且,设计了双分支融合的解析模型,其中全卷积网络解析分支得到全局分割结果,检测解析分支则引入目标检测策略,可对检测到的小目标部件进行精细分割.本文的人体解析方法能够有效提升四肢部件和小目标部件的分割性能,在 LIP[11]数据集上 Mean Intersection over union(mIoU)的评测指标为 52.19%,相对于目前最好方法 JPPNet 取得的 51.37%结果提升 0.82%,尤其是左鞋和右鞋分别提升 9.87%和 9.58%,在 ATR[14]数据集上的分割结果也验证了本文方法的有效性.

1 人体精确解析网络模型

本文的主要工作是提出了联合姿态先验的人体解析双分支网络模型,具体网络结构如图 1 所示.首先主干网络联合人体结构先验提取多尺度语义特征,然后通过两个分支得到人体解析结果.其中一个是全卷积解析分支,用来直接预测每个像素点的类别;另外一个为检测解析分支,首先对图像中的人体部件进行检测,然后对检测框所对应的区域中的人体部件进行分割;网络的最终输出是两个分支融合的结果.

1.1 人体结构先验驱动的主干网络

为了实现像素粒度的人体解析,特征提取主干网络在增强特征的语义信息的同时应尽量提高特征图的分辨率.为此,本文使用基于深度残差网络 ResNet[13]的特征金字塔网络[16]提取人体图像的多尺度特征,由顶向下逐步提取语义信息,然后通过横向结构将不同分辨率的特征由底向上进行融合,最终融合得到的多尺度特征分别记作$\{P_2,P_3,P_4,P_5\}$,使其兼具高层语义与底层位置信息.

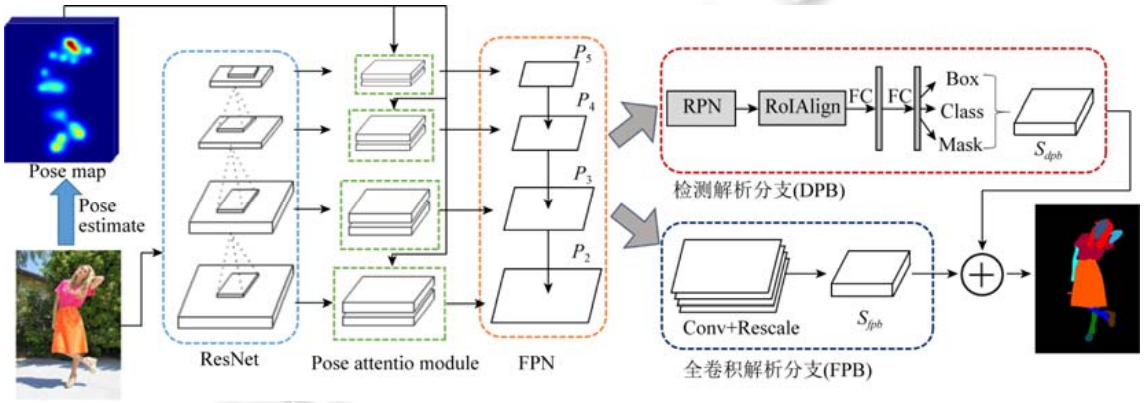

Fig.1 Posture prior driven double-branch human parsing network architecture

图 1 联合姿态先验的人体解析双分支网络模型

人体结构信息有利于增强人体特征的表达能力,可以用于指导人体解析任务.为此,本文的主干网络进一步将人体姿态先验作为注意力信息,融合至多尺度特征的横向结构中,记作 Pose Attention 模块,网络结构如图 2 所示.本文基于 Openpose[17]模型估计人体关键点置信图 $P=[p_1,p_2,...,p_M], p_i\in\mathbb{R}^{h'\times w'}$,其中,$M$ 为人体关键点个数,本文方法将 M 设为 26,h'和 w'分别为人体姿态置信图的长和宽.在 Pose Attention 模块中,对于来自 ResNet 的特征通过 1×1 的卷积操作进一步提高语义特征同时将特征固定为 256 维,记作 $X=\{x_1,x_2,...,x_{256}\}, x_i\in\mathbb{R}^{h\times w}$,其中,$h$ 和 w 分别为特征图的长和宽;然后将 X 与 P 进行拼接得到 $\hat{X}=[X,P]$,并使用 1 个 3×3 的卷积和一个 Relu 层使其更好地进行融合,同时将拼接后的特征还原为 256 维.该模块在引入人体姿态指导的同时,保证了输出特征维度与输入特征维度的一致性.

1.2 全卷积解析分支

受金字塔模型的层级语义解析思想[18]启发,融合不同尺寸的特征能够更好地提取全局特征.为此,本文分别将$\{P_3,P_4,P_5\}$上采样使其与 P_2 维度一致,并将 4 个特征拼接在一起,然后通过一个 1×1 的卷积对该特征进行降维的同时使多尺度特征更好地进行融合.最后使用一个 1×1 的卷积经过 softmax 并将其进行 4 倍上采样得到分类的置信图 $S_{fpb}=\{s_1,s_2,...,s_n\}, s_i\in\mathbb{R}^{h\times w}$,其中,$h$ 和 w 分别为输入图像的高度与宽度,n 为类别数.上述过程记作全卷积解析分支(FPB),结构如图 1 所示,该分支得到全局分割结果.该分支更注重于大目标部件,对于大目标部件的分割较为理想.

1.3 检测解析分支

在人体解析任务中有许多小目标,例如帽子、鞋子、太阳眼镜和袜子.基于全卷积网络的分割模型为增强特征的表达能力特征维度会降低,这会导致小目标丢失许多信息,无法对其进行准确分割.为解决该问题,本文引入目标检测来辅助人体解析,该分支记作检测解析分支(DPB).

借鉴 Mask R-CNN[19]的思想,首先通过 RPN[20]网络提取出目标的候选框,然后根据候选框的大小将其分配到$\{P_2,P_3,P_4,P_5\}$上,通过 ROIAlign[19]将候选框所对应的特征采样为固定大小进行分类与分割.由于感兴趣区域(RoI)之间可能有重叠,这就导致同一个像素点可能同时预测为多个类别,因此,本文不是直接使用 RoI 对应的分割结果.首先,对于检测到的 RoI 使用非极大值抑制得到最终的检测框集合 *Box*,以及检测框集合对应的类别集合 *Cls* 及其预测为该类别的得分集合 *score* 和 *Box* 包含的检测框内图像预测为前景的像素集合 *mask*;最终该分支得到的预测结果记作 $S_{dpb}=\{s_1,s_2,...,s_n\}$, 其中, $s_i\in\mathbb{R}^{h\times w}, i\in\{1,...,n\}$, h 和 w 分别为图像的高度与宽度,n 为类别数.首先将 S_{dpb} 初始化成值为 0 的矩阵,然后对检测框逐个进行处理.假设第 i 个检测框 Box_i 预测的类别 $Cls_i=k,k\in\{1,...,n\}$, 则 $s_k[j]=score_i, j\in mask_i$, 对于重叠部分本文取最大分值的得分.最后将 FPB 与 DPB 两个分支得到的结果进行融合,得到本文最终的人体部件分割得分 $S=S_{fpb}+S_{dpb}$.

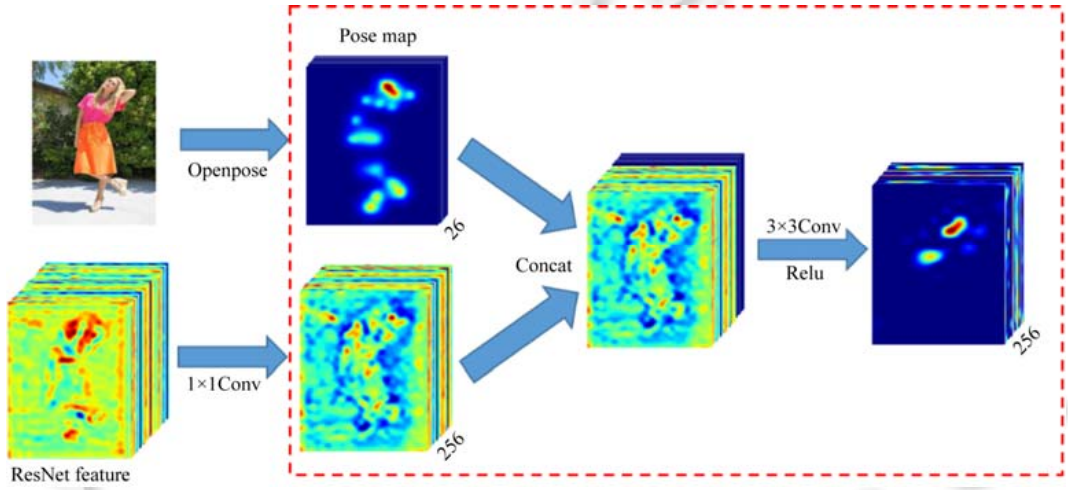

Fig.2 Pose Attention module

图 2 Pose Attention 模块结构

2 网络学习

本文模型具有双分支结构,各分支具有自身的学习任务.为了更好地对模型进行优化,本文采用端到端的训练方法,整体损失函数为两个分支损失函数之和,具体可以表示为

$$L=L_{FPB}+L_{DPB} \tag{1}$$

其中,L_{FPB}、L_{DPB} 分别对应于 FPB 分支与 DPB 分支的损失函数.

关于 FPB 分支,L_{FPB} 具体定义为逐像素点 softmax 交叉熵损失:

$$L_{parsing}=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}1\{y^{(i)}=j\}\log(p_j) \tag{2}$$

m 为像素点个数,n 为类别数,$y^{(i)}$为第 i 个像素点类别,$\log(p_j)$为第 i 个像素点预测为第 j 类的概率值.

对于 DPB 分支,其损失函数具体包括 4 项: $L_{DPB}=L_{class}+L_{bbox}+L_{mask}+L_{rpn}$,其中,$L_{class}$ 为目标检测的分类损失函数,人体解析数据集样本之间存在类别不均衡问题,针对该问题,采用 softmax focal loss[21]:

$$L_{class}=-\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{n}1\{y^{(i)}=t\}\alpha(1-p_t)^{\gamma}\log(p_t);\gamma=2,\alpha=0.25 \tag{3}$$

上式中 m 为样本个数,n 为类别数, $y^{(i)}$ 为第 i 个样本类别,p_t 为第 i 个样本预测为第 t 类的概率值.L_{bbox} 为边框回归损失,采用 smooth $L1$ 回归损失函数[22]:

$$L_{bbox}=smooth_{L1}(x)=\begin{cases}0.5x^2, & \text{if } |x|<1\\ |x|-0.5, & \text{otherwise}\end{cases} \tag{4}$$

L_{mask} 应用于检测解析分支的分割分支,与 Mask R-CNN 设置相同,采用 sigmoid 二值交叉熵损失.

在训练候选框提取器(RPN)时的损失函数为 $L_{rpn}=L_{rpn\_cls}+L_{rpn\_box}$,其中, L_{rpn_cls} 为二值交叉熵损失, L_{rpn_box} 为 smooth $L1$ 边框回归损失函数.以在 LIP 数据集上的网络学习为示例,本文采用动量(momentum)随机梯度下降算法作为优化方法,图 3 给出了各个损失函数在训练阶段随迭代次数的变化曲线,横坐标表示训练的迭代次数,纵坐标表示在对应迭代次数下损失函数的值.可以看出,各项损失函数值随着迭代次数都能够有效衰减,并趋于收敛.

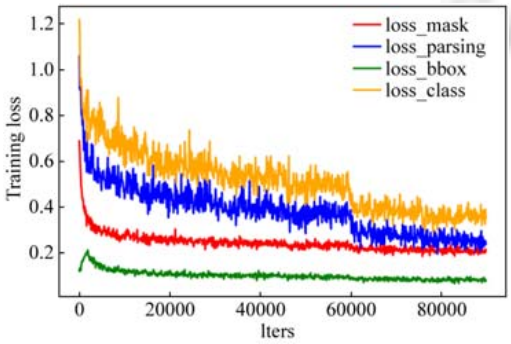

Fig.3 The plot of loss function versus number of iterations

图 3 损失函数衰减曲线

3 实验结果与分析

本文主要在两个比较流行的单人人体解析数据集上进行了研究,分别为 LIP 数据集和 ATR 数据集,并进一步与当前比较先进的方法进行比较.本文所有实验均使用基于 ResNet-50 的 FPN 网络作为基准网络.本文选取标准评测指标 Intersection over union(IoU)、Mean accuracy 和 Overall accuracy 作为实验的评估方法.本文网络模型是使用 caffe2 实现,以基于 ResNet-50 的 FPN 网络作为基准网络.在训练和测试时,输入图像的长宽中较短的一条边缩放为 400,另一条边缩放相同的比例,但是设定最大长度为 666.每个 GPU 的 mini-batch 设定为 4.本文在一个显存为 12G 的 GTX Titan X GPU 上训练,初始学习率设为 0.002 5,基于 ResNet-50 的模型迭代 90K 次,在迭代 60K 和 80K 时学习率相继缩小 10 倍.本文采用动量(momentum)随机梯度下降算法作为优化方法,将 momentum 设定为 0.9,weight decay 设定为 0.000 1.在训练时使用在 coco 数据集预训练的参数对网络初始化,只使用左右翻转进行数据增广.

3.1 数据集

LIP 数据集是一个用于人体解析任务的大规模数据集.该数据集包含背景一共有 20 类,共有 50 462 个样本,其中训练集 30 462 张图像,测试集与验证集分别为 10 000 张图像.该数据集背景较为复杂,人体姿态幅度大并且存在遮挡问题,目前该数据集依旧具有很大挑战性.

ATR 数据集包含 17 706 张图像,共有 17 类目标和 1 个背景类.ATR 数据集没有划分训练集和测试集,本文在其中随机选取 3 000 张图像作为测试集,剩余 14 706 张图像作为训练集.

3.2 本文方法分析

为探究 Pose Attention 模块、DPB 分支的有效性,本文主要在难度较大的 LIP 数据集上进行对比实验,首先将各模块去除,仅保留全卷积网络作为本文的基准网络 B,并通过递进组合的方式,验证不同模块的有效性.由于

P2 特征尺寸为原输入的 1/4,需将基准网络 B 最后预测的结果上采样 4 倍还原为输入尺寸.如表 1 所示,基准网络 B 在 mIoU 的评测指标下达到 44.23%的准确率.使用基准网络 B 对输入图像进行分割,一些失败的样本如图 4 所示.经过分析与观察预测图 4 中结果,主要有以下问题:(1) 当姿态复杂时,四肢等部件容易造成左右混淆,尤其是当图像中的行人为背影或侧身时;(2) 对于小目标分割不精确.图 4 中,B 表示基准网络,P、D 分别表示 Pose Attention 模块和 DPB 模块,*表示分类损失为 focal loss.

本文首先在基准网络 B 中引入 Pose Attention 模块,分别在生成$\{P_2,P_3,P_4,P_5\}$的横向结构中插入 Pose Attention 模块.如表 1 所示,此模块使预测结果在 mIoU 的测量指标下得到 4.83%的提升,主要表现在鞋子、腿部、胳膊等四肢人体部件的提升,l-shoes 和 r-shoes 的性能分别得到 18.46%和 18.13%的显著提升.实验结果表明,引入人体姿态信息能够有效提升模型对具有左右属性的四肢部件的判别能力.表 1 中,B 表示基准网络,P、D 分别表示 Pose Attention 模块和 DPB 模块,*表示分类损失为 focal loss.

Table 1 The results of each module on the validation set of LIP and the result of JPPNet

表 1 本文模型的各个模块在 LIP 测试集的对比实验结果及当前最新方法 JPPNet 结果

方法	B	B+P	B+P+D	B+P+D*	JPPNet[12]
bkg	84.82	85.12	82.26	85.09	86.26
hat	61.74	62.12	64.88	65.09	63.55
hair	67.38	67.39	68.63	68.53	70.20
gloves	32.47	34.39	38.56	39.17	36.16
sunglasses	29.78	27.05	32.22	34.32	23.48
u-clothes	62.51	63.97	65.27	65.11	68.15
dress	21.23	25.66	28.76	29.64	31.42
coat	47.88	50.81	52.87	53.34	55.65
socks	39.93	40.02	44.88	45.42	44.56
pants	67.07	68.75	69.52	69.34	72.19
jumpsuits	18.99	22.01	23.52	25.15	28.39
scarf	15.27	15.70	15.36	17.89	18.76
skirt	19.31	20.83	24.40	24.98	25.14
face	70.67	70.65	71.69	71.77	73.36
l-arm	52.19	59.74	61.59	62.00	61.97
r-arm	54.63	61.07	63.38	63.35	63.88
l-leg	37.90	52.59	56.63	58.56	58.21
r-leg	37.38	53.12	56.36	58.43	57.99
l-shoe	31.61	50.07	53.91	53.89	44.02
r-shoe	31.79	49.92	53.73	53.67	44.09
avg	44.23	49.05	51.55	52.19	51.37

3.3

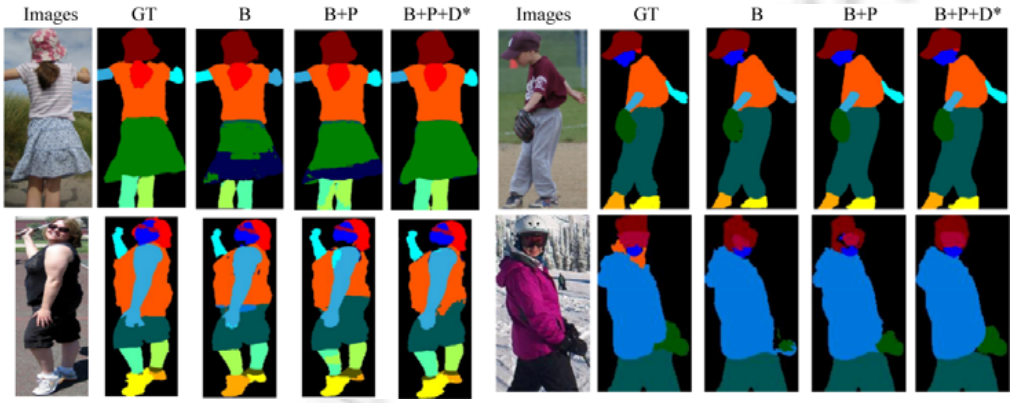

Fig.4 The visualization of each Blob's prediction results on the validation set of LIP

图 4 各个模块在 LIP 测试集的预测结果可视化

Pose Attention 模块作用在生成多尺度特征$\{P_2,P_3,P_4,P_5\}$的横向结构中,为了进一步验证其有效性,本文做了如下对比实验.只在生成特征 P_2 的横向结果中加入 Pose Attention 模块,该方法记作 PL2;与上述描述类

似,PL3、PL4、PL5 分别表示只在生成特征 P_3、P_4、P_5 的横向结构中加入 Pose Attention 模块.如表 2 所示,人体姿态信息引入各个生成多尺度特征的横向结构中都能使结果有相应提升.

人体姿态关键点能够帮助我们更好地理解人体结构信息,能够有效地提升与人体四肢相关的部件的分割性能.本文为进一步分析不同部位的关键点对模型性能的影响,在测试阶段分别将对应部位的关键点响应置为 0.为验证与胳膊相关的关键点对模型的作用,本文将手肘、手腕、肩膀关键点对应的通道响应设为 0,结果见表 3,与使用所有关键点的结果对比,l-arm 与 r-arm 的结果下降 10.37%与 9.80%,其他部件的分割性能基本不变.当将腿部相关的关键点响应置为 0 时,在 IoU 的评测指标下,l-leg 与 r-leg 的结果分别为 49.25%与 49.21%;当将脚部相关的关键点响应置为 0 时,l-shoe 与 r-shoe 的结果分别为 28.17%与 23.01%.同时,为了研究人体姿态估计误差对模型性能的影响,本文将人体姿态估计的标签作为人体结构信息.由于只有 LIP 数据集有人体关键点标注,因此,本文只在该数据集进行此对比实验,在 mIoU 的评测指标下结果为 51.93%,与使用 Openpose 模型的预测结果接近,实验结果表明,本文方法对人体姿态估计误差具有一定的鲁棒性.

Table 2 The mIoU pose attention module at different locations

表 2 Pose Attention 模块在不同位置处的 mIoU 指标比较

方法	mIoU
PL2	47.47
PL2+PL3	47.68
PL2+PL3+PL4	48.29
PL2+PL3+PL4+PL5	49.05

Table 3 The effect of different human poses

表 3 不同人体姿态关键点对模型性能的影响

方法	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe
使用所有关键点	62.00	63.35	58.56	58.43	53.89	53.67
去掉胳膊关键点	51.63	53.55	58.32	58.22	53.77	53.63
去掉腿部关键点	61.84	63.18	49.25	49.21	53.89	53.72
去掉足部关键点	61.72	63.17	54.84	54.78	28.17	23.01

为了验证 DPB 分支的作用,本文在引入 Pose Attention 模块的基础上进一步引入 DPB 模块.如表 1 所示,引入 DPB 能够使鞋子、太阳眼镜、袜子、围巾等小目标的分割性能得到显著提升,其中,l-shoe 提升 3.82%,r-shoe 提升 3.75%,整体性能提升 3.14%.

根据文献[11]对 LIP 数据集样本进行统计,本文观察到样本类别存在较大程度的不均衡问题,包含脸部、头发等目标的样本比较多,围巾、连体衣等在数据集中出现的次数较少.针对上述问题,本文在检测分支预测类别时选用 focal loss.如表 1 所示,当分类损失为 focal loss 时 mIoU 为 52.19%,相对于仅使用交叉熵损失时的结果 51.55%提升 0.64%,这主要表现在数量较少的样本类别中.

本文采用双分支的网络结构,最终人体各部件分割得分 $S=S_{fpb}+S_{dpb}$.关于 FPB 分支与 DPB 分支结果融合策略的探讨,本文做如下对比实验.首先,将 S_{fpb} 与 S_{dpb} 拼接,然后通过一个 3×3 的卷积与一个 1×1 的卷积得到最终分类的置信图 S,在 mIoU 的评测指标下结果为 51.66%.实验结果表明,本文的融合策略更加简洁、高效.

3.4 对比实验

对于 LIP 数据集,本文与当前几种比较好的方法比较,结果见表 4.在 Mean accuracy 与 mIoU 的评测指标下,本文的方法比当前最好的方法 JPPNet 的结果分别提升 3.13%和 0.82%.将人体姿态估计信息作为人体结构先验,模型能够更好地理解人体各个部件之间的关系,有利于人体部件的分割.在卷积神经网络提取特征的过程中小尺度目标的信息损失较为严重,检测解析分支能够将小目标部件所对应的区域放大,对其进行精确分割.JPPNet 在 mIoU 的评测指标下各类的得分见表 1,本文方法对于四肢部件和小目标部件的分割性能均优于 JPPNet,其中左鞋和右鞋分别得到 9.87%和 9.58%的显著提升,太阳眼镜提升 10.84%.值得注意的是,本文方法的基准网络使用的是 ResNet-50 提取的特征,而 JPPNet 的基准网络使用的是 ResNet-101,而且本文在测试时只使用原始图像,没有对输入进行翻转以及多尺度测试.

对于 ATR 数据集主要与当前主流的 3 种方法进行对比,见表 5,由于 JPPNet 方法在训练阶段需要人体姿态关键点的标注信息,但是 ATR 数据集并未提供该标注信息,JPPNet 在该数据集上并不适用,因此表 5 中没有与此方法进行对比.本文的方法比 Attention+SSL[11]的结果在 Mean accuracy 的评测指标下高 4.94%,在 Mean IoU 的评测指标下要高 5.38%.

本文选取 ATR 数据集部分预测结果可视化,如图 5 所示,第 1 列为输入图像,第 2 列为 Attention+SSL 方法的预测结果,第 3 列为本文方法的可视化结果,同时,为方便对比分析,本文将标签可视化,如第 4 列所示.Attention+SSL 方法由于缺少准确的人体结构先验信息,当人体姿态复杂时,容易对左右鞋、左右腿等人体四肢部件造成误判,本文方法在添加可靠的人体结构信息的情况下可对人体四肢相关部件进行准确分割;如图 5 第 3 行可视化结果所示,本文方法对于小目标部件中的太阳眼镜分割效果明显优于 Attention+SSL 方法,可对其边缘部位准确分割;对于鞋子这类小目标,本文方法在对其准确分类的同时还能对其边缘进行精确分割.

Table 4 Parsing performance of multiple methods on validation set of LIP

表 4 本文及其对比方法在 LIP 测试集上的解析结果

方法	Overall accuracy	Mean accuracy	Mean IoU
DeeplabV2[7]	82.66	51.64	41.64
Attention[10]	83.43	54.39	42.92
Attention+SSL[11]	84.36	54.94	44.73
JPPNet[12]	86.48	62.25	51.37
Ours	85.33	65.38	52.19

Table 5 Parsing performance of multiple methods on ATR dataset

表 5 本文及其对比方法在 ATR 数据集上的解析结果

方法	Overall accuracy	Mean accuracy	Mean IoU
DeeplabV2[7]	94.28	72.66	58.97
Attention[10]	94.88	73.68	61.55
Attention+SSL[11]	95.08	74.97	63.06
Ours	95.45	79.91	68.44

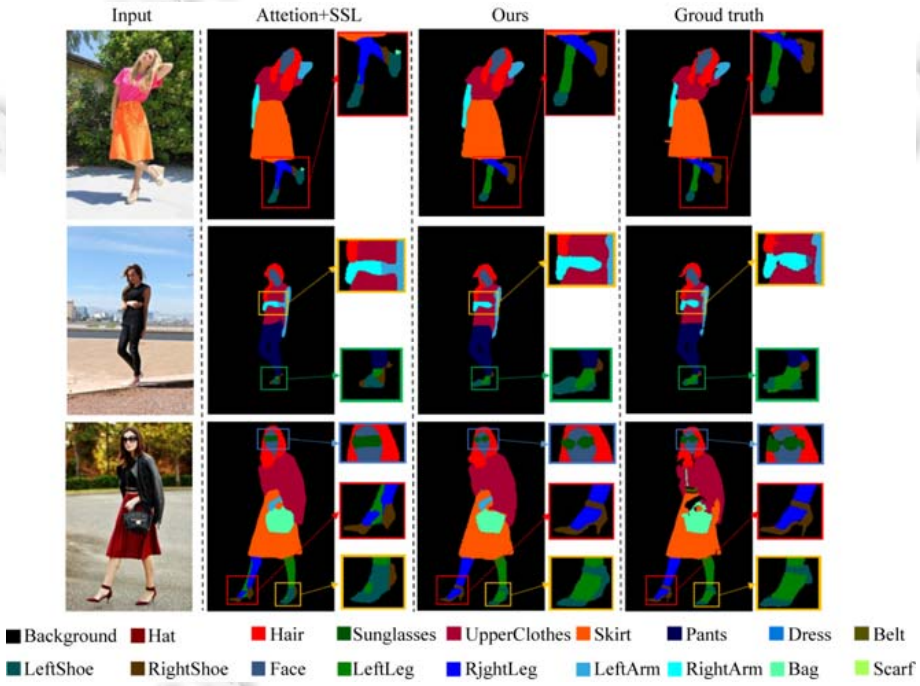

Fig.5 The visualization of segmentation results upon the ATR dataset

图 5 ATR 数据集分割结果可视化

4 结 论

针对人体四肢等部件和小目标分割不精确的问题,本文提出了一种联合姿态先验的人体解析双分支网络模型.实验结果表明,该方法能够实现人体部件的精确分割,对太阳眼镜等小目标和与四肢相关的部件有较好的分割效果.下一步研究工作的重点将是如何提升四肢部件和小目标部件除外的人体部件的分割性能,以及如何优化算法提升模型的检测速度.

References:

[1] Zhao R, Ouyang W, Wang X. Unsupervised salience learning for person re-identification. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2013. 3586−3593.

[2] Cai H, Wang Z, Cheng J. Multi-scale body-part mask guided attention for person re-identification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2019.

[3] Gan C, Lin M, Yang Y, *et al*. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016.

[4] Tian X, Wang L, Ding Q. Review of image semantic segmentation based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(2):440−468 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5659.htm [doi: 10.13328/j.cnki.jos.005659]

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Annals of the History of Computing, 2017,(4):640−651.

[6] Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFS. arXiv Preprint arXiv:1412.7062, 2014.

[7] Chen LC, Papandreou G, Kokkinos I, *et al*. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,40(4):834–848.

[8] Liang X, Xu C, Shen X, Yang J, Liu S, Tang J, Lin L, Yan S. Human parsing with contextualized convolutional neural network. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1386–1394.

[9] LiangX, ShenX, Xiang D, Feng J, Lin L, Yan S. Semantic object parsing with local-global long short-term memory. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3185–3193.

[10] Chen LC, Yang Y, Wang J, *et al*. Attention to scale: Scale-aware semantic image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 3640–3649.

[11] Gong K, Liang X, Zhang D, *et al*. Look into Person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. 2017. [doi: 10.1109/CVPR.2017.715]

[12] Liang X, Ke G, Shen X, *et al*. Look into Person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2018,(99):1.

[13] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770–778.

[14] Liang X, Yang J, Yang J, *et al*. Deep human parsing with active template regression. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2015,37(12):2402.

[15] Yang L, Song Q, Wang Z, *et al*. Parsing R-CNN for instance-level human analysis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 364–373.

[16] Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2117–2125.

[17] Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. In: Proc. of the IEEE Conf. on Computer Vision & Pattern Recognition. 2017.

[18] Zhao H, Shi J, Qi X, *et al*. Pyramid scene parsing network. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2881–2890.

[19] He K, Gkioxari G, Dollár P, *et al*. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). 2017.

[20] Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. 2015. 91–99.

[21] Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2017,(99):2999–3007.

[22] Girshick R. Fast R-CNN. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 1440–1448.

附中文参考文献:

[4] 田萱,王亮,丁琪.基于深度学习的图像语义分割方法综述.软件学报,2019,30(2):440−468. http://www.jos.org.cn/1000-9825/5659.htm

高明达(1994－),女,硕士,主要研究领域为图像分割.

刘青山(1975－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为图像与视频理解,模式识别.

孙玉宝(1983－),男,博士,副教授,CCF 专业会员,主要研究领域为主要从事深度学习理论,压缩感知重建,人体解析.

邵晓雯(1996－),女,硕士,主要研究领域为行人重识别.