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

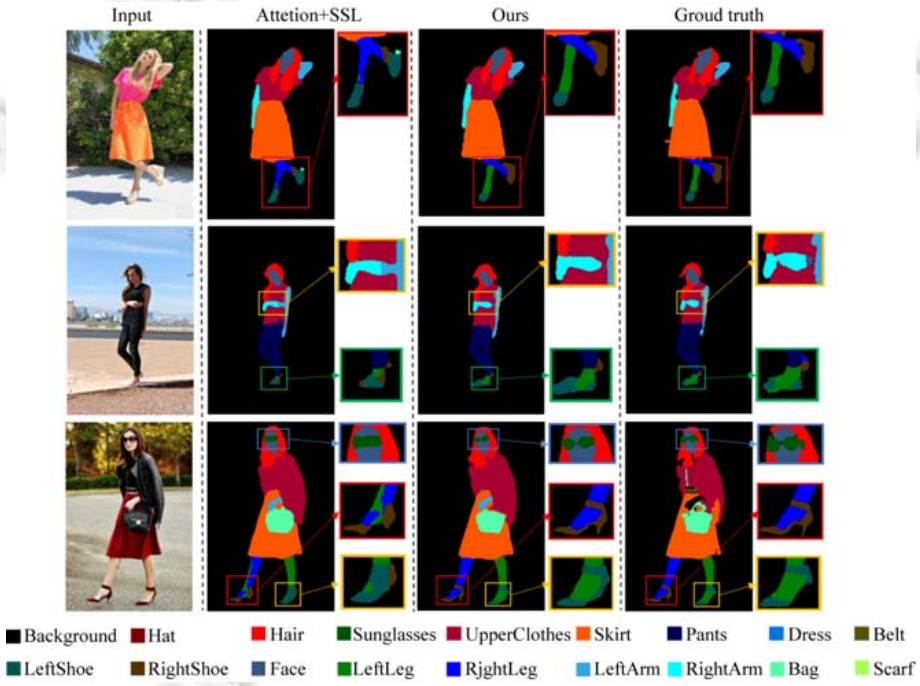

Fig.5 The visualization of segmentation results upon the ATR dataset

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.