A Top-Down Unified Framework for Instance-level Human Parsing

Haifang Qin[15]
qhfpku@pku.edu.cn
Weixiang Hong[25]
weixiang.hong@u.nus.edu
Wei-Chih Hung[3]
whung8@ucmerced.edu
Yi-Hsuan Tsai[4]
ytsai@nec-labs.com
Ming-Hsuan Yang[35]
mhyang@ucmerced.edu

[1] Peking University
[2] National University of Singapore
[3] University of California, Merced
[4] NEC Laboratories America
[5] Google Cloud

1 Introduction

In this supplementary material, we present 1) more visual comparisons of the proposed approach and the state-of-the-art PGN [2] method in crowded scenes from Figure 1 to Figure 3, 2) video results by applying our method for instance-level human parsing in Figure 4 and in the supplementary video, and 3) an ablation study on the impact of the batch size during training in Table 1.

Table 1: Ablation study for batch size on the CIHP [2] dataset.

Method	Batch Size	IoU threshold			AP^r_{vol}	Mean IoU
		0.5	0.6	0.7		
PGN [2]	4	35.8	28.6	20.5	33.6	55.8
Ours	4	43.1	36.0	26.6	37.9	54.5
Ours	8	44.0	36.8	27.2	38.6	55.2

2 Results and Analysis

We show more results on the CIHP [2] and PASCAL-Person-Part [1] dataset in crowded scenes. As shown in Figure 2 and 3, our approach can distinguish different instances and parts under the situation that instances are heavily occluded by each other. In Figure 1, failure cases in our method are caused by two main reasons: 1) confusion between classes, *i.e.*, Upper Clothes and Coat, Coat and Dress, Skirt and Dress, and 2) over detected instances,

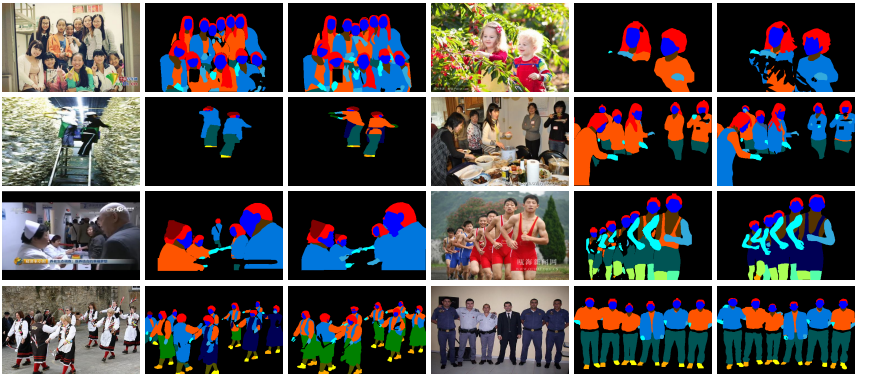

Figure 1: Failure cases in our approach on the CIHP [2] dataset. We show input images, Our$_P$ as our parsing results, and GT$_P$ as ground truths. Best viewed in color.

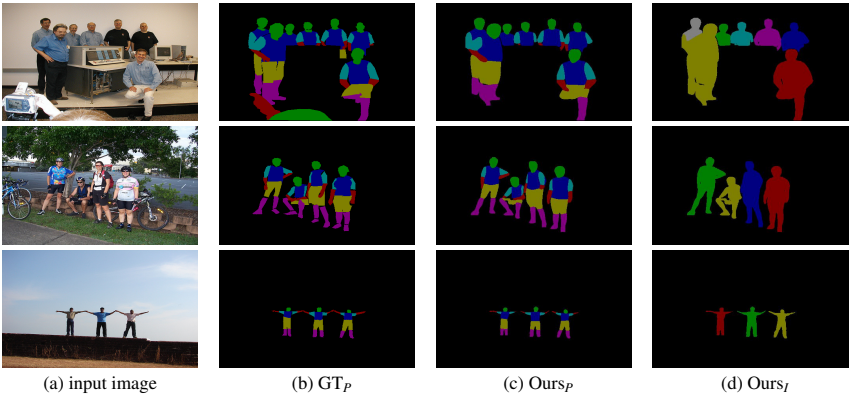

Figure 2: Instance-level human parsing comparisons on the PASCAL-Person-Part [1] dataset. From left to right, we show (a) input image, (b) GT$_P$ as ground truth parsing, (c) Ours$_P$ as our parsing results, (d) Ours$_I$ as instance masks, in which different colors mean different instances. Best viewed in color.

i.e., there are instances which are not annotated as ground truths while our method still detects them, which will be treated as false positives for evaluation.

We further apply the proposed method on video sequences in a frame-by-frame manner as in the attached supplementary videos from the DAVIS [3] dataset. The video sequences show that the proposed top-down unified framework is sufficiently robust to produce visually pleasing results and is able to handle challenging cases such as fast movement, occlusions, complex backgrounds for multiple instances even though there is no temporary information or post-processing to smooth the videos. Examples of still frames are shown in Figure 4.

3 Ablation Study for Batch Size

In Table 1, we show the impact of batch size on the CIHP [2] dataset. With the same batch size as PGN [2] (i.e., 4), our approach still performs favorably and only influences the performance marginally compared to the one using batch size as 8.

(a) input images (b) PGN$_I$ c) Ours$_I$ (d) PGN$_P$ (e) Ours$_P$

Figure 3: More results compared with PGN [7]. We show that our framework can handle challenging cases in crowded scenes. From left to right, we show (a) input images, (b) PGN$_I$ as instance results of PGN, (c) Ours$_I$ as instance results of ours, (d) PGN$_P$ as parsing results of PGN, (e) Ours$_P$ as parsing results of our approach. Different colors indicate different instances. Best viewed in color.

Figure 4: Instance-level human parsing results on the DAVIS [3] dataset. We show instance results and parsing results of still frames. Best viewed in color.

References

[1] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. *arXiv preprint arXiv:1808.00157*, 2018.

[3] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.