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

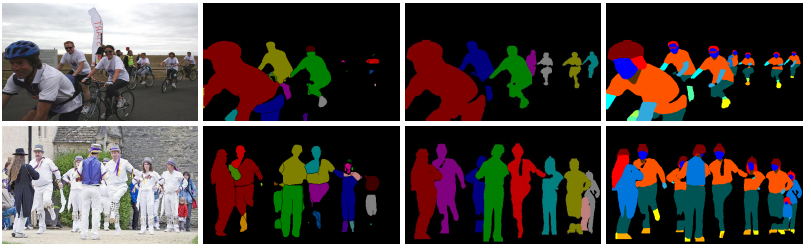
Abstract

Instance-level human parsing is one of the essential tasks for human-centric analysis which aims to segment various body parts and associate each part with the corresponding human instance simultaneously. Most state-of-the-art methods group instances upon multi-human parsing results, but they tend to miss instances and fail in grouping under the crowded scene. To address this problem, we propose a top-down unified framework to simultaneously detect human instance and parse every part within that instance. To better parse the single human, we also design an attention module, which is aggregated to our parsing network. As a result, our approach is capable of obtaining fine-grained parsing results and the corresponding human mask in a single forward pass. Experiments show that the proposed algorithm performs favorably against state-of-the-art methods on the CIHP and PASCAL-Person-Part datasets.

1 Introduction

Instance-level human parsing is one of the challenging tasks in computer vision. It aims to segment various body parts of human and associate them with corresponding instances. This task widely benefits the human-centric analysis in the wild and plays an essential role in high-level application domains, such as video surveillance and human behaviour analysis. Existing methods for instance-level human parsing can be grouped into two categories: segmentation-based methods [15], and proposal-based approaches [24, 31, 53].

The segmentation-based methods, such as PGN [15], first perform multi-human parsing to predict per-pixel classification, and group pixels of the same category into different instances. Although these methods have achieved state-of-the-art performance on several public benchmarks, they tend to wrongly group the disjoint parts together within an instance



(a) Input Image (b) PGN$_I$ (c) Ours$_I$ (d) Ours$_p$

Figure 1: Examples of instance grouping of the PGN [15] and our method. From left to right, we show (a) input image, (b) instance grouping results from PGN. (c) instance masks of ours, (d) multi-human parsing results of ours. Different colors mean different instances. Best viewed in color.

or combine parts from different instances. Figure 1(b) shows one example, where some human instances in the crowded scenario are missed. On the other hand, the proposal-based methods benefit from the success of human detection. Existing algorithms [24, 31, 53] perform instance detection and multi-human parsing independently, and then aggregate the results from these two branches to produce the final instance-level human parsing. Despite the promising results, these methods are not end-to-end trainable for instance-level human parsing and require heavy post-processing.

Observing the drawbacks of both categories, our goal is to develop a method that can address the aforementioned issues while maintaining their benefits. First, to avoid the grouping strategy that may easily fail even in trivial situations (see Figure 1(b)), we focus on the proposal-based approach instead of relying on bottom-up cues such as edges of instances. Second, to address the misalignment between detection and parsing results, we propose a top-down unified framework that can simultaneously perform instance detection and single-human parsing to produce the final instance-level human parsing result in a single forward pass. As shown in Figure 1(c)(d), the proposed method recognizes instances and parses human parts with less confusions in the crowded scenes. To this end, we decompose the objective of instance-level human parsing into two subtasks: instance detection and single-human parsing. With the support of the instance detection branch, the multi-human parsing task could be reduced to multiple single-human parsing. Thus, we design a segmentation branch that is able to handle the multi-category segmentation task, i.e., single-human parsing, given the detected instance bounding boxes from the detection branch. The noise from other instances still exists even though the bounding box is given. To further decrease the noise from other instances within the given bounding box, we design an attention module to enhance the signal from the foreground region of the target instance. To retain the end-to-end trainable merit, we additionally utilize a binary segmentation branch. Since the attention maps are generated upon the given proposals instead of the entire image, they can effectively suppress the noise from pixels of background or other instances. We then train the proposed framework in an end-to-end fashion and generate instance-level human parsing without heavy post-processing compared to existing approaches.

To evaluate the proposed framework, we conduct experiments on two benchmark datasets including CIHP [15] and PASCAL-Person-Part [7]. Experimental results show that our method performs favorably against state-of-the-art algorithms, obtaining improvement over 15%, 23% and 29% in terms of AP^r_{vol}, AP^p and PCP on the CIHP dataset, respectively.

Furthermore, we demonstrate that the proposed attention module helps the human parsing performance, especially under challenging situations such as crowded scenes. The main contributions of this work are: 1) we propose a unified proposal-based framework to obtain the instance-level human parsing results in a single forward pass without any post-processing, while achieving state-of-the-art performance, 2) we decompose a more difficult problem, instance-level human parsing problem, into instance detection and several single-human parsing tasks, and 3) we introduce an ROI-level attention module to provide additional human segmentation signals to improve the final human parsing result.

2 Related Work

Instance Segmentation. Instance segmentation requires predicting class label and pixel-wise instance masks to localize a varying number of instances in each image. There are mainly two groups of methods in instance segmentation: proposal based [19, 29] and segmentation based [3, 21, 25, 28, 34, 37, 40, 51, 52] methods. Proposal-based methods have a strong connection to object detection. In R-CNN [13], object proposals are fed into the network to extract features for classification. Fast R-CNN [12], faster R-CNN [35], and SPPNet [17] speed up the process by pooling features from global feature maps. Based on the detection methods, Mask-RCNN [19] and PANet [29] propose to add a mask head to predict the instance segmentation.

The other group methods are mainly segmentation-based. They first learn a transformation for converting segmentation maps to instance maps [3, 25, 28, 40] or instance boundaries [21], and then decode instance masks from predicted transformation. DIN [1] fuses predictions from object detection and semantic segmentation systems. In addition, graphical models are used in [51, 52] to infer the order of instances, while RNNs [34, 37] are utilized to propose one instance at each time step. For these approaches, however, the transformation converting the segmentation map to the instance map/boundary tends to fail due to complicated instance appearances. Thus, obtaining instance results upon the semantic segmentation results remains a challenge.

Human Parsing. Recently, numerous research efforts have been devoted to single human parsing [5, 14, 26, 30, 39, 44, 47, 48] for advancing human-centric analysis research. For example, Liang *et al*. [26] propose a Co-CNN architecture that integrates multiple levels of image contexts into a unified network. Gong *et al*. [14] design structure-sensitive learning to enforce the produced parsing results semantically consistent with the human joint structures. Though rapid progress has been made in single human parsing domain, muti-human parsing in crowded scenes remains a challenging problem due to the confusion across different instances.

Instance-level Human Parsing. With the recent proposed CIHP [15] and PASCAL-Person-Part [7] datasets, the community has achieved significant advances in human analysis [15, 24, 53]. Instead of solving the single human parsing problem, the crowded multi-human parsing has attracted attention. Multi-human parsing requires the correct parsing of all humans in an image, while on the instance-level, it requires the association of parts of each human. Beyond multi-human parsing, instance-level human parsing which provides fine-grained parsing results and corresponding instance mask is more critical for human central analysis in the real-world scenario.

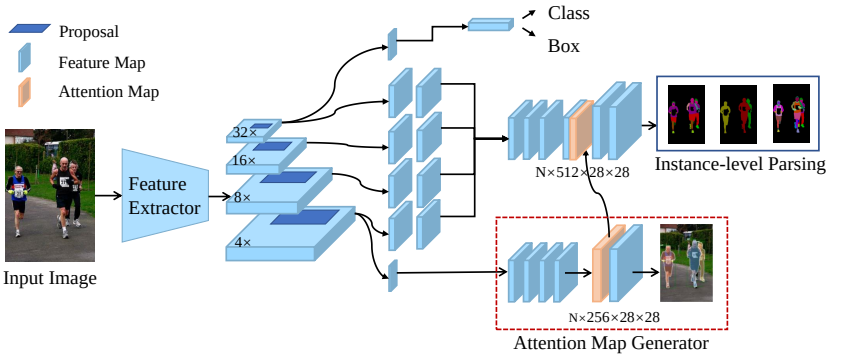


Figure 2: Illustration of our proposed method. Each input image first goes through an FPN [17] backbone to extract pyramid features. The pyramid features include four levels, having downsample ratio in range from 4 to 32. After RoI Align, the pyramid features are fed into the head for prediction. The heads in our approach consist of three parts: 1) detection branch, 2) parsing branch, 3) attention map generator.

Recently, PGN [15] predicts multi-human body parts using the edges of instances and groups the body parts into instances through a part grouping post-processing. However, it tends to predict inaccurate instance grouping. Li *et al*. [24] obtains the category-level segmented parts and the human bounding box through a detector independently, and then associate them via a differentiable conditional random field. Zhou *et al*. [53] extend Mask R-CNN [19] with ASPP module [6] to perform multi-human parsing and aggregate instance mask results from Mask R-CNN to obtain the final instance-level part results. All of these recent works [24, 31, 53] treat this problem as a combination of multi-human parsing problem and human detection, thereby viewing this task as the segmentation-based method with different grouping strategies, which requires heavy post-processing procedures. In contrast, the proposed method decomposes multi-human parsing into a joint task of instance detection and single human parsing, producing the instance parsing result without any post-processing.

Attention Mechanism. It is observed that attention plays an essential role in human perception [8, 20, 36]. One important property of a human visual system is that one does not process am entire scene at once. Instead, humans exploit a sequence of partial glimpses and selectively focus on salient parts to capture visual structure [22]. Recently, numerous works aim to incorporate the attention mechanism into deep learning frameworks [10, 22, 32]. Such attempts has been proved effective in many vision tasks including classification [45], detection [2], image captioning [38, 46, 50], and image-question-answering [49]. Mnih *et al*. [32] learn an attention model that adaptively selects a sequence of regions or locations for processing. Chen *et al*. [5] obtain several attention masks to fuse feature maps or predictions from different branches. Vaswani *et al*. [41] present a self-attention model for machine translation. Wang *et al*. [42] obtain attention masks by calculating the correlation matrix between each spatial point in the feature map. In our work, we add an auxiliary loss to supervise feature maps that focus more on the foreground region of the target instance.

3 Proposed Method

Figure 2 illustrates the framework of the proposed method. We first present an overview of our network and then introduce our proposed attention map generator in Section 3.1. Section 3.2 illustrates the implementation details.

3.1 Unified Framework for Instance-level Human Parsing

As shown in Figure 1, segmentation based methods such as PGN [15] may lead to inaccurate grouping and miss instances. To address these issues, the proposed approach first detects instances and then parses human parts within each person instance. By adopting this methodology, we divide the original multi-human parsing task as a set of easier problems, *i.e.*, multiple single-human parsing. Thus, we can parse human parts easier within an instance proposal.

Overall Framework. Built upon the Faster-RCNN [35] detection framework, we incorporate a parsing branch that enables the feature sharing between detection and parsing tasks. As shown in Figure 2, given an input image, we first take advantages of the Feature Pyramid Network (FPN) [27] to extract multi-scale features. Second, the features and the proposals generated from the region proposal network (RPN) are fed into the parsing/detection branch individually to produce final results. The intuition of our design is based on 1) multi-scale features account for various size of instances while preserving both fine details and global context information, and 2) parsing and detection branches share the same feature extractor, in which network parameters are jointly learned during optimization.

In the proposed framework, the detection branch is responsible to localize multiple human in an image, whiling sending the successfully detected human proposals to the instance-level parsing branch. As such, the parsing branch only needs to handle human parsing within a proposal, i.e., ideally within one instance. More importantly, the collaboration between these two branches jointly optimizes the overall objective for both human detection and parsing in a unified framework, achieving the multi-human instance-level parsing in a single forward pass.

Attention Module. We have introduced an end-to-end trainable network for instance-level human parsing. However, there could be challenges that harm the final parsing results, mainly from two factors. First, some complicated scenes may contain multiple human instances. Such cases can be relaxed by detecting every single human and parsing them individually in our framework. Second, even if the bounding box of each instance is accurate, each bounding box may contain parts from other instances, which leads to confusions. To address this problem, we further propose an attention map generator to emphasize the foreground of the target instance within a proposal. Consequently, the noisy part caused by other instances or the background within a proposal region can be mitigated effectively.

To this end, we incorporate an attention branch using an auxiliary loss to predict the foreground map and aggregate them into the parsing branch (as shown in Figure 2). Note that the generated attention maps are on the ROI-level instead of on the image-level. With the generated attention map that focuses on the foreground region, the parsing branch receives additional features of the human mask, further enhancing the parsing ability and reducing the noise.

Optimization Objective. We have described each component in the proposed unified framework, including the detection, instance-level parsing, and attention branches. As such, our final objective with multi-task loss functions are defined as:

$$L = L_d + L_a + \alpha L_p, \tag{1}$$

where L_d is the detection loss including bounding box regression and category classification, L_a is the binary cross-entropy loss to generate the human mask, and L_p is the cross-entropy loss that parses human into parts. α is the weight to balance the parsing loss. Note that, the parsing loss is defined only on positive proposals received from the detection branch.

3.2 Implementation Details

Training Details. We jointly train the three branches in the proposed model. During training, the input images are resized such that the shorter edge is 800 pixels, and the max size of the longer side is set to $1,333$ pixels. Each image has 512 sampled proposals, i.e., RoIs, with a ratio of $1:3$ for positives and negatives. Each RoI is considered positive when it has IoU larger than 0.5 with respect to the ground truth box. Each mini-batch has 1 image per GPU and we train on 8 GPUs with the effective mini-batch size as 8 for 130K iterations. The learning rate is set to 0.002, and it is decreased by 10 at the 100K and 115K iteration. We use a weight decay of 0.0001 and momentum of 0.9. The loss weight α is set to 5 to balance the values from different losses. During inference, we set the proposal number as $1,000$ and run the box prediction on these proposals followed by the non-maximum suppression [33]. We implement our framework with Pytorch and Tesla $V100$ GPUs.

Model Details. Our detection baseline model is built upon the Faster-RCNN [35] framework with the ResNet-101 [18] architecture. To obtain features from all levels of pyramid features, we apply RoI Align [19] to the features and fuse them with max operation in the parsing branch like [29]. We set the size in RoI Align as 14×14 for the detection/attention branch and as 28×28 for the parsing branch. We set the number of convolution layers after RoI Align to 4 in all branches. Each convolution layer is with kernel size 3×3 and 1 padding size. We add group normalization [43] to each convolution layer in the parsing branch. To calculate the loss, we upsample the network output by 2, while the ground truth of each instance is resized to 28×28 for the attention map and 56×56 for the parsing result via the nearest interpolated operation.

4 Experimental Results

We compare our method with state-of-the-art methods on the CIHP [15] and the PASCAL-Person-Part [7] datasets. Comprehensive ablation studies of our approach are conducted on the CIHP [15] validation dataset. More results and images are available in the supplementary material. All the source code and trained models will be made available to the public.

4.1 Dataset

The CIHP [15] dataset is the most challenging dataset for instance-level human parsing. It contains 28,280 training images, 5,000 validation images, and 5,000 test images with 19 semantic human part annotation. We evaluate our method in terms of AP^r [23] following

Table 1: Comparisons in terms of AP^r and Mean IoU on the CIHP [15] test set. *ms* represents multi-scale testing and *flip* represents flip testing.

Method	IoU threshold			AP^r_{vol}	Mean IoU
	0.5	0.6	0.7		
PGN [15] (+ ms + flip)	35.8	28.6	20.5	33.6	**55.8**
Ours w/o attention	41.0	33.4	23.8	36.2	53.4
Ours	41.8	34.0	24.2	37.0	53.5
Ours + ms + flip	**44.0**	**36.8**	**27.2**	**38.6**	55.2

Table 2: Comparisons in term of AP^p results on the CIHP [15] validation dataset.

Method	AP^p	IoU thresholds		
		0.5	0.6	0.7
PGN [15]	0.39	0.34	0.17	0.06
Our method	**0.48**	**0.51**	**0.26**	**0.09**

Table 3: Comparisons in term of *PCP* results on the CIHP [15] validation dataset.

Method	*PCP*	IoU thresholds		
		0.5	0.6	0.7
PGN [15]	0.34	0.61	0.25	0.13
Our method	**0.44**	**0.77**	**0.36**	**0.19**

PGN [15]. We also report the results in terms of AP^p [23] and *PCP* [23], which indicate how well the parsing results within the corresponding human instance. We will discuss more about the metrics in the next section.

The PASCAL-Person-Part [7] dataset contains 1,716 images for training and 1,817 for testing. Following Chen *et al*. [6], the annotations are merged to include six person parts: Head, Torso, Upper arms, Lower arms, Upper legs, and Lower legs. Following the state-of-the-art method [15], we evaluate the performance in terms of AP^r with different IOU thresholds.

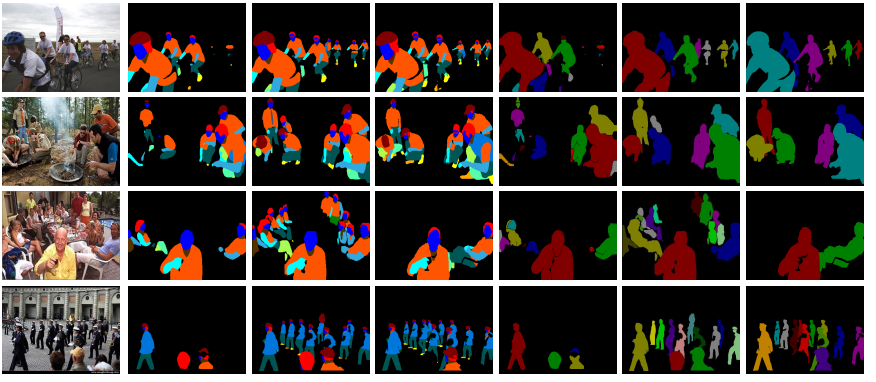
4.2 Evaluate Metric

AP^r **(Mean Average Precision)** is first proposed for evaluating instance segmentation results by Hariharan *et al*. [16]. Recent works [15, 23, 24] adopt it to evaluate the instance-level human parsing results. After producing the parsing results and instance masks, the part-level instances can be generated. AP^r only takes part-level instance into considerations, which means that AP^r cannot accurately measure the quality of instance-level human parsing.

AP^p **(Average Precision based on Part)** was first proposed by Li *et al*. [23]. Different from AP^r, AP^p uses part-level Intersection over Union (IoU) of different semantic part categories within a person to determine if one instance is a true positive. Specifically, when comparing one predicted semantic part parsing map with one ground truth parsing map, the average IoU of all the semantic part categories is used as the measure of overlap. In other words, AP^p emphasizes how well a specific human instance has been parsed.

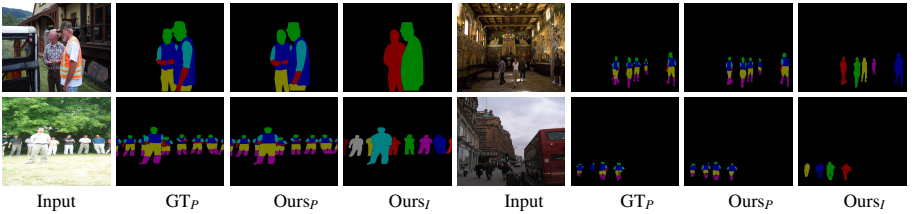
PCP **(Percentage of Correctly Parses Body Parts)** was first proposed by Li *et al*. [23] to evaluate the parsing quality on the semantic parts within a person instance.For each true-positive person instance, *PCP* considers all the categories (excluding background) with pixel-level IoU larger than a threshold as correctly parsed. *PCP* of one person is the ratio between the correctly parsed categories and the total number of categories of that person. The overall *PCP* is the average *PCP* for all human instances.

To summarize, AP^p together with *PCP* can better measure the instance-level human pars-



(a) Input Images (b) PGN$_P$ (c) Ours$_P$ (d) GT$_P$ (e) PGN$_I$ (f) Ours$_I$ (g) GT$_I$

Figure 3: Comparisons of instance-level human parsing of PGN [15] and our method. From left to right, we show (a) input images, (b) parsing results of PGN, (c) parsing results of our approach, (d) parsing ground truth, (e) instance results of PGN, (f) instance results of ours, (g) instance ground truth. Different colors indicate different instances. Better viewed in color.



Input GT$_P$ Ours$_P$ Ours$_I$ Input GT$_P$ Ours$_P$ Ours$_I$

Figure 4: Instance-level human parsing qualitative comparisons on the PASCAL-Person-Part [7] dataset. From left to right, we show input images, GT$_P$ is ground truth parsing, Ours$_P$ is our parsing results, Ours$_I$ is instance masks in which different colors mean different instances. Better viewed in color.

ing quality than AP^r. Therefore, we also evaluate on the CIHP [15] validation dataset in terms of AP^p and PCP. As shown in Table 2 and 3, the proposed approach performs favorably against the current state-of-the-art method PGN [15] by a significant margin.

4.3 Performance Evaluation

Comparisons on the CIHP Dataset. The CIHP [15] dataset is one of the most extensive datasets for instance-level human parsing. We first conduct a baseline study to demonstrate the usefulness of jointly learning the detection/parsing networks. The proposed method achieves 37.0% in terms of AP^r_{vol}, which is better than 32.3% that uses a separate parsing network. Here, we separately utilize the Mask-RCNN [19] detection network and use detected human instance for training the DeepLab-v2 [4] parsing network. The experiment suggests that using a unified framework for instance-level human parsing is beneficial in terms of both performance and computational complexity.

We report performance of our method on the CIHP [15] testing set as well as validation set. As shown in Table 1, our method without strategies like multi-scale/flipping testing outperforms PGN [15], which applies these techniques in both training and testing. To measure the parsing quality concerning a particular human instance, we also report the results in

Table 4: Results on the PASCAL-Person-Part [7] dataset.

Method	AP^r_{vol}	IoU threshold		
		0.5	0.6	0.7
MNC [8]	36.7	38.8	28.1	19.3
Holistic [23]	38.4	40.6	30.4	19.1
PGN [15]	39.2	39.6	29.9	20.0
Ours	**43.1**	**48.1**	**38.3**	**25.7**

Table 5: Results on the CIHP [15] validation set.

Attention structure	AP^r_{vol}	IoU threshold		
		0.5	0.6	0.7
dot_after_conv3_sum	32.8	36.5	29.7	21.0
dot_after_conv4_sum	33.1	36.1	28.4	19.3
concat_after_conv4_trans	34.1	38.1	30.9	21.6
concat_after_conv4	**36.2**	**40.7**	**33.9**	**25.1**

terms of AP^p and PCP to better characterize the effects of instance-level human parsing. As shown in Table 2, our method performs favorably against PGN [15] by 23% in terms of AP^p at every threshold. Similarly in Table 3, our method outperforms PGN [15] by 29% in terms of PCP. Experimental results demonstrate that our method is able to effectively generate human parsing results and associate the parts to the corresponding instance.

We also present visual comparisons of instance-level human parsing results in Figure 3. Compared to our results, PGN [15] usually misses human instances in the crowded scene when there are multiple instances overlapped or occluded to each other. In contrast, our approach parses all the instances even when they are not annotated as the ground truth, which shows a strong generalization ability. Furthermore, PGN [15] tends to wrongly group the disjoint parts together within an instance or combine parts from different instances. In contrast, our approach handles such disjoint situations well by focusing on the correct parts within a proposal.

Comparisons on the PASCAL-Person-Part Dataset. The PASCAL-Person-Part [7] dataset is a subset of PASCAL-VOC-2010 [11] with additional annotations. For a fair comparison, we report the AP^r on the testing set for multi-human parsing.

We report the performance comparisons of the proposed approach with three state-of-the-art methods in terms of AP^r at IoU threshold from 0.5 to 0.7 and AP^r_{vol} in Table 4. For a fair comparison, we evaluate our approach with multi-scale training and testing similar to PGN [15]. The proposed method improves the performance of the 2^{nd} best exisitng method by 28% for $AP^r_{0.7}$ and 10% for AP^r_{vol}. We also visualize the qualitative instance-level human parsing results in Figure 4. It shows that the proposed approach is able to handle challenging situations such as various scales, complex environments and occlusion issues.

Ablation Study on Attention Module. We conduct the ablation studies to investigate the effects with different operation types and layers in the proposed attention structure. Here, we use conv3 and conv4 to represent the third layer and four layer after RoI Align, respectively. We place it either between conv3 layer and conv4 layer, denoted as "after_conv3" or between conv4 and upsample layer, denoted as "after_conv4". For feature fusion types, we compare dot and concatenate. As shown in Table 5, the proposed attention module has better performance when placed closer to the classification layer, *i.e.*, both dot and concat between conv4 and upsample layer performs better than conv3 and conv4. We then use "concat_after_conv4" as our attention configuration in the final framework.

5 Conclusion

In this work, we propose a unified approach for instance-level human parsing. Instead of directly parsing the input images, we leverage a human detector to decompose the multi-human

parsing problem into several sub-problems. so that each of them tackles a single-human parsing task that is much easier. In addition, an attention map generator is developed to further suppress the noises from other overlapped instances or the background. We conduct extensive experiments on two benchmark datasets and show that our method performs favorably against other state-of-the-art instance-level human parsing approaches.

References

[1] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] 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.

[8] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201, 2002.

[9] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 2016.

[10] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 24(8):2151–2184, 2012.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Proceedings of the International Journal of Computer Vision*, 2010.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[14] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] 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.

[16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 297–312, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 346–361, 2014.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[20] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[21] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *Proceedings of the Advances in Neural Information Processing Systems*, 2010.

[23] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017.

[24] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017.

[25] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*, 2016.

[26] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[27] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[31] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Thomas Huang. Devil in the details: Towards accurate single and multiple human parsing. *arXiv preprint arXiv:1809.05996*, 2018.

[32] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Proceedings of the Advances in Neural Information Processing Systems*, 2014.

[33] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *Proceedings of the International Conference on Pattern Recognition*, 2006.

[34] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[36] Ronald A Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1-3): 17–42, 2000.

[37] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent instance segmentation. In *Proceedings of the European Conference on Computer Vision*, 2016.

[38] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[39] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. A high performance crf model for clothes parsing. In *Proceedings of the Asian Conference on Computer Vision*, 2014.

[40] Jonas Uhrig, Marius Cordts, Uwe Franke, and Thomas Brox. Pixel-level encoding and depth layering for instance-level semantic labeling. In *Proceedings of the German Conference on Pattern Recognition*, 2016.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, 2017.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 10, 2017.

[43] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, 2018.

[44] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *Proceedings of the European Conference on Computer Vision*, 2016.

[45] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, 2015.

[47] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[48] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[49] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[51] Ziyu Zhang, Alexander G Schwing, Sanja Fidler, and Raquel Urtasun. Monocular object instance segmentation and depth ordering with cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[52] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[53] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. *arXiv preprint arXiv:1808.00661*, 2018.