Multi-Human Parsing with Pose and Boundary Guidance

Shunchen Du, Yigang Wang*, and Zizhao Wu

School of Media and Design, Hangzhou Dianzi University, China
yigang.wang@hdu.edu.cn

Abstract. In this work, we present a novel end-to-end semantic segmentation framework for multi-human parsing, which integrates both the high- and low-level features. Our framework includes three modules: segmentation module, pose estimation module, and boundary detection module. Ideally, the pose estimation model will guide high-level feature extraction to identify the part location; meanwhile, the boundary detection module will concentrate on the low-level feature extraction so as to distinguish the boundary. Both modules are united in the backbone segmentation module to generate the desired accurate high-resolution prediction for multi-human parsing. Experiment results on the PASCAL-Person-Part dataset demonstrate that our method achieves superior results over state-of-the-art methods. Code has been made available at https://github.com/scmales/MHP-with-Pose-and-Boundary-Guidance.

Keywords: human parsing · pose estimation · multi-task learning.

1 Introduction

Human parsing refers to assigning image pixels from the human body to a semantic category, e.g. hair, dress, leg. This task can also be considered as fine-grained semantic segmentation for human body. At present, human parsing is critical to the understanding of humans and supports lots of industrial applications, such as virtual try-on [23], dressing style recognition [33], human behavior analysis [20], and so on.

Recently, with the advance of deep convolutional neural networks, multi-human parsing has attracted a huge amount of interest [8, 22] and achieved great progress. Most of the previous works [21, 35] focus on single-human parsing, which simplifies the scene and ignores the human interaction, occlusion, and rich human body posture. However, in the real complex scenarios, there exist the diverse background composition, the physical contact between people, and the occlusion between human body parts, these kinds of approaches may fail to deal with these issues. As shown in Fig. 1, the inaccurate detection of edge and large areas of missing body parts are common problems.

To address these obstacles, some researchers [36, 42, 12, 18] have further studied on the multi-human parsing problem, and significant progress have been

* Corresponding author

made. For example, some researchers [36, 42] have tried to pre-process the scene by detecting each person, and then performing human parsing individually, thus reducing the complexity of the problem. Following these approaches, some researchers have also found that human parsing and human pose estimation are two popular applications of human body analysis [3, 10, 40], despite the differences in detail, these two tasks are highly correlated and complementary. On one hand, pose estimation focuses on the joint detection of human body, which can be regarded as the high-level abstraction of human body, while neglecting the other details. On the other hand, human parsing considers the categories of every pixel, while flaws in modeling high-level abstraction. Inspired from these points, they suggested to jointly analyze both tasks [12, 29, 36].

In this paper, rather than mutual learning human parsing and pose estimation, we propose an end-to-end unified framework for simultaneous human parsing, pose estimation, and boundary detection. Specifically, our framework includes three modules, they are semantic segmentation module, pose estimation module, and edge detection module. These modules are unified with our framework, i.e., in the forward propagation, we obtain the multi-resolution features of the pose estimation module and the edge detection module. Then, we integrate these features as the residual input to the semantic segmentation module based on the encoder-decoder structure, to achieve the accurate high-resolution multi-human parsing results. We evaluate our method on the PASCAL-Person-Part dataset, as well as the ablation studies, which demonstrate that our method is competitive among existing multi-human parsing models.

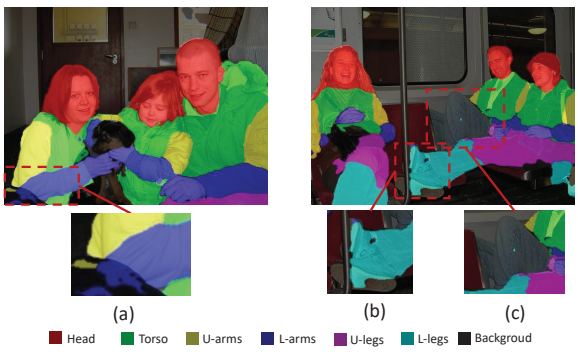

Fig. 1. The problem of multi-human parsing. (a) and (b) show the inaccurate edges for segmentation. (c) shows the lack of human body parts.

2 Related Work

2.1 Human Parsing

Human parsing aims to segment a human image into multiple semantic parts, which can also be considered as fine-grained semantic segmentation for human body. Benefiting from the success of deep learning and convolutional neural networks, significant progress have been achieved. For example, Fully Convolutional Network (FCN) [28] had made the first attempt to realize the end-to-end training of semantic segmentation. More recent works [31, 1] have been proposed based on a variety of encoder-decoder networks. Among the existing models, the representation learning of the network played a key role to the results. Ideally, the learned feature representation should capture both the global and local information of the image so as to make an accurate prediction. In order to achieve this goal, some specific modules have been devised to aggregate the context features on different regions, such as Atrous Spatial Pyramid Pooling (ASPP) module [7], Pyramid Scene Parsing (PSP) module [41]. More recently, some researchers focused their attention on the concrete human parsing problem by exploiting human parsing benchmark dataset [19, 26] and explored the potential connections between human body. For example, Liu et al. suggested a multi-path enhance network [22], which extracts multi-dimensional feature based on the different scales of input pictures to enhance the final results, but the complexity of the network is high, which leads the network hard to train.

2.2 Pose Estimation and Edge Detection

There are abundant of works [13, 17] have been investigated to the pose estimation problem. In the early stage of this task, CNN has been used to get the joint coordinates directly [32], but the results are not very satisfying. Later, Wei et al. [34] suggest a novel framework which combined the ideas of multi-stage and post-process refinement, to solve the problem of partial occlusion and invisibility. Following their work, Cao et al. [4] proposed a real-time multi-human pose estimation network based on the Part Affinity Fields (PAFs), which abstracts the connections between the joints. This network carried out human joint detection, the regression training, and then conducted the Hungarian algorithm to separate the different joints. Different from this method, some researchers [11, 14] suggested to firstly perform the human detection to separate the single human, and then performed the pose estimation individually, thus cast the multi-human pose estimation into the traditional single human pose estimation problem.

Edge detection has also been regarded as an important research field of image processing. The current mainstream work is based on CNN due to its high performance. For example, Xie et al. [37] proposed a holistically-nested edge detection method, which implements the end-to-end training of edge detection through a full convolution neural network and depth supervision, significant performance has been achieved of this method.

2.3 Multi-task Learning

Multi-task learning aims to learn multiple different tasks simultaneously while maximizing performance on one or all of the tasks. In the human parsing field, some researchers [24, 25, 38, 39] noticed that human pose estimation and human parsing are related tasks since both problems are strongly dependent on the human body representation and analysis, which inspired them to propose many multi-task learning approaches in the field. Particularly, Xia et al. [36] unified the pose estimation and semantic segmentation in their framework, Conditional Random Field (CRF) was further employed to fuse both tasks, and obtained significant performance. Gong et al. [12] explored a new self-supervising structure-sensitive learning method that does not require additional monitoring information but the joint coordinates, to improve the analytical results. Liang et al. [18] presented a novel joint human parsing and pose estimation network which incorporates the multi-scale feature connections and iterative location refinement in an end-to-end framework, to investigate efficient context modeling to facilitate the task of multi-task learning.

There exist some methods [2, 5] aim to proceed with edge detection and semantic segmentation simultaneously, while maximizing performance on one or all of the tasks. Among them, Ruan et al. [27] suggested a Context Embedding with Edge Perceiving (CE2P) framework which unifies several useful properties, including feature resolution, global context information, edge details, and leverage them to benefit the human parsing task.

In this paper, comparing with the traditional method [27, 19, 26], we fuse the multi-level features by using the multi-task learning structure. What's more, comparing with multi-task methods [36, 18], our advantage is that we construct a synchronous end-to-end network which is easy to train and fuse multiple related tasks for multi-human parsing.

3 Algorithm

Our architecture consists of three modules: backbone semantic segmentation modules, pose estimation module, edge detection module, and semantic segmentation module. Fig. 2 illustrates the details. In our architecture, we employ the Residual Network (ResNet-101 [15]) as the backbone architecture. The output of the backbone network is fed into the followed three modules to learn features globally and locally.

Given an input picture with a fixed size of $H \times W \times 3$, the final output of the entire model is the result of semantic segmentation defined as P, which contains c channels of confidence maps, and is consistent with the number of human semantic segmentation labels. Our network is multi-module based, which will generate additional outputs include m channels of the Part Confidence Maps (PCMs) K, n channels of the Part Affinity Fields (PAFS) F from the pose estimation module, and two channels of edge confidence map E from the edge detection module, where m denotes the number of labels for joints and background, and n

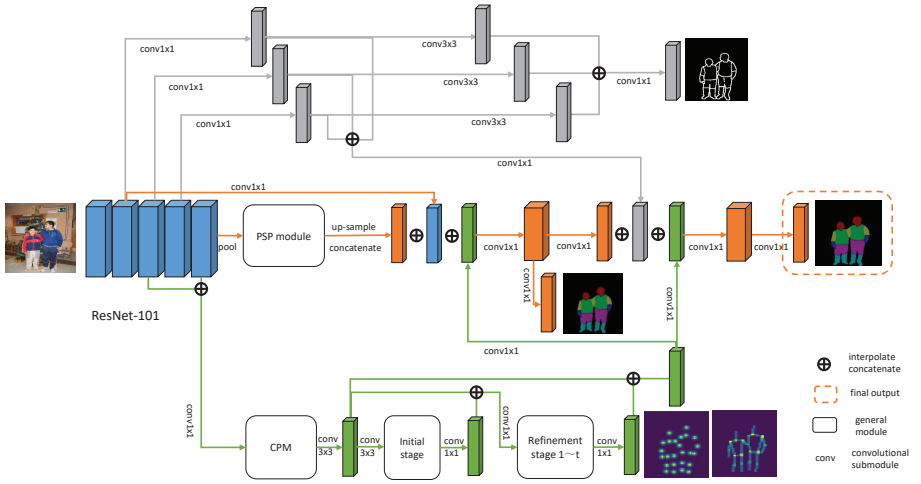

Fig. 2. The framework of our approach for multi-task learning. The architecture has three modules: semantic segmentation module, pose estimation module, and edge detection module.

denotes the double number of part vectors between the joints. For more details about PAFs, we refer the reader to OpenPose [4].

In our case, we set $c = 7, m = 15, n = 28$ to the PASCAL-Person-Part dataset, the values are defined as is shown in Fig. 2, given the output K and F of the pose estimation module, we stitch it with the output features of the pose estimation layer, and further concatenate them with the intermediate features of the edge detection network, finally, they are fed into the semantic segmentation network to guide the generation of human parsing results.

In the following, we describe the details of the modules in our network.

3.1 Pose Estimation Module

We use the bottom-up strategy to design the structure of the module. To reduce the cost of training, the depth-wise separable convolution layers have been introduced to construct a light-weight pose estimation network [30].

Fig. 2 illustrates the structure of our network, we simplify the description of refinement stages and only retain the last feature representation of PCMs and PAFs. Specifically, the module takes the features extracted from the third and last layer of ResNet-101 as the input, then we proceed the input through the conventional pose machine (CPM) layer, the initial stage and a total number of t refinement stages, these three stages are composed in the convolutional layers and the exponential linear units layers. Furthermore, the input of each refinement stage is the concatenated features between the output of the previous stage and the output of the CPM layer. PCMs and PAFs will be output at the initial stage

and each refinement stage. We get the final output of PCMs and PAFs of this module on the final layer.

The purpose of this structure is to construct an iterative optimization process and refine the result. Thus exploit the potential of the residual structure and ensures the stability and efficiency of the refining process. Given the $s(0 \leq s \leq t)$ stage, where 0 represents the initial stage, and $1 \sim t$ denotes the refine stage, the regression loss is defined as follows:

$$L_K^s = \frac{1}{N} \sum_{i=0}^{m} \sum_{p} W(p) \| K_i^s(p) - K_i'(p) \|^2 \tag{1}$$

$$L_F^s = \frac{1}{N} \sum_{i=0}^{n} \sum_{p} W(p) \| F_i^s(p) - F_i'(p) \|^2 \tag{2}$$

where N is the total number of effective pixels involved in the calculation process; L_K^s and L_F^s denote the loss function value of the PPM and PAF respectively in phase s; i represents the $i-th$ channel; p denotes the location of an image; W is a binary mask with $W(p) = 0$ when the pixel is added during data enhancement at location p; $K_i^s(p)$ represents the joint confidence maps prediction value in the i-th channel at the location p; $K_i'(p)$ denotes the label value as location p. The variables related to PAFs can also be described similarly.

In summary, the loss function of our pose estimation module is defined as:

$$L_{pose} = \sum_{s=0}^{t} (L_K^s + L_F^s) \tag{3}$$

3.2 Edge Detection Module

In the edge detection module, we aim to extract features in different spatial dimensions and enhance the performance of the edge detection, thus we emphasize more on the low-level features. Our module takes the output of the three modules in ResNet-101 as the input. As shown in Fig. 2, three modules of input are processed separately; then they have been fused in the last convolutional layer; finally, we obtain the output of the edge detection. In summary, the edge detection module consists of convolutional submodules, where each submodule contains the convolutional layer, the batch normalization layer, and the rectified linear unit layer. Overall, the edge detection can be regarded as the binary classification problem. We construct the loss function of edge L_{edge} based on cross-entropy loss of category balance.

3.3 Semantic Segmentation Module

The Semantic Segmentation module is based on the encoder-decoder framework. As shown in Fig. 2, in the encoder stage, PSP module [41] is employed as the method to obtain context features of the image through four different

scale pooling operations, to effectively increase the network receptive field, and then samples and splices the multi-scale features, and finally sends them to the decoder stage.

The content of the decoder stage is composed of several convolutional submodules, which contain the convolution layer, the normalization layer, and the rectified linear unit layer. Specifically, we use edge features and pose features for feature enhancement, and exploit the residual structure to ensure the learning performance. As shown in Fig. 2, we observe that the edge features are the concatenated features from all convolution outputs of the first layer of the edge detection module; the pose features are the concatenated features from the output of CPM layer, PCMs, and PAFs. The characteristics of feature enhancement are reflected as follows: firstly, the input to the first convolutional module is based on the output of the PSP module, the second layer of ResNet-101, and the residual features; secondly, the input to the intermediate convolutional submodule is based on the output of the last layer, the residual of edge module, and the residual of pose module. We argue that the features of pose estimation and semantic segmentation correspond to high-level semantics. To learn the transformation correlation between the high-level semantics better, we use the residual features twice.

Fig. 3 illustrates PCMs, PAFs, the intermediate features of edge detection, and their analyzable visualization after feeding input to the model. We can observe that the PCMs and PAFs are able to identify the abstract location of the individual body parts, and thus provide prior information for human part segmentation. Besides, the intermediate features of edge detection contain rich details of human body parts and can effectively supplement the details deficiency in semantic segmentation.

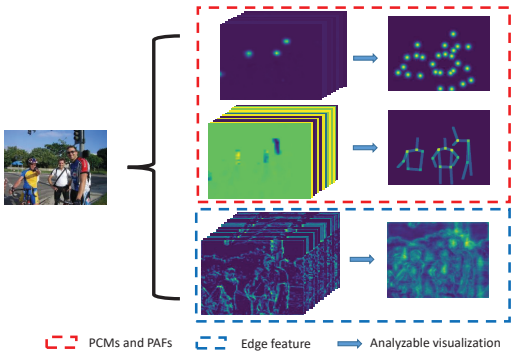

Fig. 3. A figure illustrating PCMs, PAFs and edge features.

The output of the module is the maps of human body with the model parameters. So the total loss of semantic segmentation module L_{seg} is a sum of two components, which is similarly defined as multi-class cross-entropy loss.

Finally, the loss defined on the whole model is defined as:

$$L = \alpha L_{pose} + \beta L_{edge} + \gamma L_{seg} \tag{4}$$

where α, β, γ are the weights defined on different losses. L_{edge} and L_{seg} belong to classification loss, and it has been proved in the paper [27] that adding the two in the same proportion is effective. Besides, inspired by the paper [16], we unify the regression loss L_{pose} and classification loss L_{seg} to the same order of magnitude in that avoiding the loss with a small gradient being carried away by the loss with a large gradient. We train our end-to-end model by minimizing the above loss function.

4 Experiments

Dataset. We evaluate our results on the PASCAL-Person-Part dataset [9], which is developed for human parsing evaluation and comes from PASCAL VOC dataset. The training set of PASCAL-Person-Part contains 1716 images, with a validate set contains 1817 pictures. Each image corresponds to a labeled image which is regarded as the benchmark segmentation results. Besides the segmentation results, the dataset also contains the pose results [36]. But we found that the open-source dataset lacked 40 pose labels for training and verification, so we removed them according to the method [36]. Overall, there are 7 semantic segmentation categories in the dataset, including background, head, trunk, upper arm, lower arm, upper leg, and lower leg; there also exist 14 types of joint point coordinates, including head, neck, left shoulder point, left elbow joint, left wrist, left hip, left knee, left ankle, right shoulder point, right elbow joint, right wrist, right hip point, right knee, and right ankle. The label of the PCMs is from the Gaussian distribution with the joint point coordinate as the center. The labels of the PAFs are composed of vectors between the joints, where the value can be defined as the L_2 distances between the joints. The edge detection tags mentioned are generated from semantic segmentation masks by tracing the 8-pixel neighborhood changes. We only experiment on this dataset because other datasets for multi-human parsing do not have both pose and part segment annotations.

Implementation Details. We test our method on a PC, which is equipped with an Ubuntu operating system, an Intel 4GHz i7 processor, 8GB RAM, and RTX 2080 Ti graphics card. The input size of images is set to 384. For our joint loss function, we set the weight of each term as $\alpha = 0.5; \beta = 1; \gamma = 1$. To augment the data, we also perform rotation and horizontal flipping to augment the original data. The initial learning rate is set to 0.001, the optimization strategy used in this work is the stochastic gradient descent (SGD) with a linear learning rate decay. We also employed the two norm regularization, the weight decay value is set to 0.0005.

Ablation studies. We further evaluate the effectiveness of coarse-to-fine schemes of our model, including three situations as follows: only semantic segmentation module, edge detection module with semantic segmentation module, the complete architecture with three modules. Fig. 4 illustrates the comparison

between the prediction with the ground-truth. For instance, as we can see from the second row, the result of Fig. 4(b) shows that the foot is missing or even detected by mistake, the result of Fig. 4(c) is better than that of Fig. 4(b), while the result of Fig. 4(d) correctly predicts the foot. Then from the third row of the figure, the accuracy of the three experimental results is also increasing compared with the part of the upper arm of the human body covered by the dog. In the same way, the results of the third image also show that the integrity of the prediction results of the network with pose estimation and edge detection is better than that of the other strategies after ablation study.

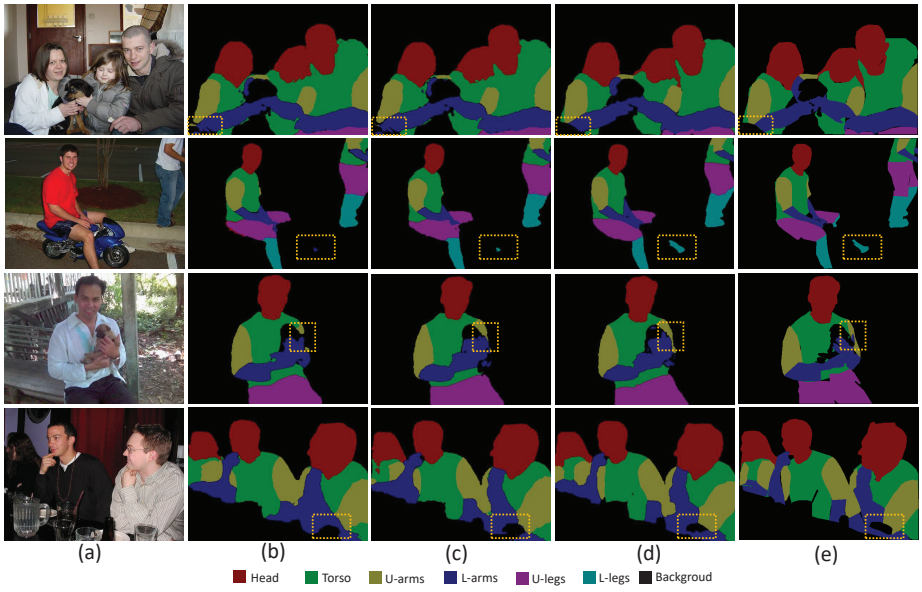

Fig. 4. Visual comparison of ablation study. (a) input images. (b) predictions based solely on the semantic segmentation module. (c) predictions based on semantic segmentation with edge detection module. (d) predictions based on complete architecture. (e) ground truth.

Results. The evaluation criterion used in this paper is mainly based on the Mean Intersection Over Union (MIoU) of each class. The higher the values are, the more accurate the results are. We conduct the prediction and the evaluation of the validation set of the PASCAL-Person-Part dataset. The experiment results are shown in Table 1. We note that our model achieves superior performance, when comparing with the existing mainstream algorithms [8, 35, 6, 12, 36], the MIoU of our model is 66.11%. We also note that our method gets the highest score on the cross-ratio of all body parts. The result demonstrates the efficiency of our method for multi-human parsing.

Table 1. Mean Pixel IOU (mIOU) of Human Parsing on PASCAL-Person-Part.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	background	Ave
Attention [8]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HANZ [35]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
LIP-SSL [12]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
DeepLabv3* [7]	84.06	66.96	54.26	52.80	48.08	43.59	94.79	63.50
Joint [36]	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
MuLA [29]	-	-	-	-	-	-	-	65.10
Our model(w/o pose/edge)	84.98	66.85	54.52	53.72	47.77	44.22	95.13	63.88
Our model(w/o pose)	85.45	68.28	56.78	56.63	49.31	46.56	**95.34**	65.48
Our model	**85.99**	**68.61**	**58.08**	**57.50**	**49.78**	**47.61**	95.23	**66.11**

5 Conclusion

In this work, we have proposed a multi-human parsing algorithm that integrates the tasks of pose estimation and edge detection, and realized synchronous end-to-end learning. Specifically, our method has suggested multi-module architecture, which composes of the pose estimation module and edge detection module, and provides the auxiliary features for the multi-human parsing tasks. The experiment results in the PASCAL-Person-Part dataset demonstrated the effectiveness of our method.

Acknowledgements:

This work was supported in part by the National Natural Science Foundation of China (61602139), and in part by the Zhejiang Province Science and Technology Planning Project (2018C01030).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Bertasius, G., Shi, J., Torresani, L.: Semantic segmentation with boundary neural fields. CoRR **abs/1511.02674** (2015)
3. Bourdev, L.D., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: IEEE 12th International Conference on Computer Vision. pp. 1365–1372. IEEE Computer Society (2009)
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. CoRR **abs/1812.08008** (2018)
5. Chen, L., Barron, J.T., Papandreou, G., Murphy, K., Yuille, A.L.: Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. CoRR **abs/1511.03328** (2015)

6. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
7. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017)
8. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. CoRR **abs/1511.03339** (2015)
9. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.L.: Detect what you can: Detecting and representing objects using holistic models and body parts. CoRR **abs/1406.2031** (2014)
10. Dong, J., Chen, Q., Xia, W., Huang, Z., Yan, S.: A deformable mixture parsing model with parselets. In: IEEE International Conference on Computer Vision. pp. 3408–3415. IEEE Computer Society (2013)
11. Fang, H., Xie, S., Lu, C.: RMPE: regional multi-person pose estimation. CoRR **abs/1612.00137** (2016)
12. Gong, K., Liang, X., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and A new benchmark for human parsing. CoRR **abs/1703.05446** (2017)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. CoRR **abs/1802.00434** (2018)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 386–397 (2020)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
16. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CoRR **abs/1705.07115** (2017)
17. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and A new benchmark. CoRR **abs/1812.00324** (2018)
18. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 871–885 (2019)
19. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2402–2414 (2015)
20. Liang, X., Wei, Y., Shen, X., Yang, J., Lin, L., Yan, S.: Proposal-free network for instance-level object segmentation. CoRR **abs/1509.02636** (2015)
21. Liang, X., Xu, C., Shen, X., Yang, J., Tang, J., Lin, L., Yan, S.: Human parsing with contextualized convolutional neural network. IEEE Trans. Pattern Anal. Mach. Intell. **39**(1), 115–127 (2017)
22. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. CoRR **abs/1611.06612** (2016)
23. Lin, J., Guo, X., Shao, J., Jiang, C., Zhu, Y., Zhu, S.: A virtual reality platform for dynamic human-scene interaction. In: SIGGRAPH ASIA, Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments. pp. 11:1–11:4. ACM (2016)
24. Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Cao, X., Yan, S.: Fashion parsing with video context. IEEE Trans. Multimedia **17**(8), 1347–1358 (2015)
25. Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Cao, X., Yan, S.: Fashion parsing with video context. IEEE Trans. Multimedia **17**(8), 1347–1358 (2015)

26. Liu, S., Liang, X., Liu, L., Shen, X., Yang, J., Xu, C., Lin, L., Cao, X., Yan, S.: Matching-cnn meets KNN: quasi-parametric human parsing. CoRR **abs/1504.01220** (2015)
27. Liu, T., Ruan, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y., Huang, T.S.: Devil in the details: Towards accurate single and multiple human parsing. CoRR **abs/1809.05996** (2018)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR **abs/1411.4038** (2014)
29. Nie, X., Feng, J., Yan, S.: Mutual learning to adapt for joint human parsing and pose estimation. In: Computer Vision - ECCV. vol. 11209, pp. 519–534. Springer (2018)
30. Osokin, D.: Real-time 2d multi-person pose estimation on CPU: lightweight openpose. CoRR **abs/1811.12004** (2018)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015)
32. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. CoRR **abs/1312.4659** (2013)
33. Wang, Y., Tran, D., Liao, Z., Forsyth, D.A.: Discriminative hierarchical part-based models for human parsing and action recognition. J. Mach. Learn. Res. **13**, 3075–3102 (2012)
34. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. CoRR **abs/1602.00134** (2016)
35. Xia, F., Wang, P., Chen, L., Yuille, A.L.: Zoom better to see clearer: Human part segmentation with auto zoom net. CoRR **abs/1511.06881** (2015)
36. Xia, F., Wang, P., Chen, X., Yuille, A.L.: Joint multi-person pose estimation and semantic part segmentation. CoRR **abs/1708.03383** (2017)
37. Xie, S., Tu, Z.: Holistically-nested edge detection. Int. J. Comput. Vis. **125**(1-3), 3–18 (2017)
38. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3570–3577. IEEE Computer Society (2012)
39. Yang, W., Luo, P., Lin, L.: Clothing co-parsing by joint image segmentation and labeling. CoRR **abs/1502.00739** (2015)
40. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition. pp. 1385–1392. IEEE Computer Society (2011)
41. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. CoRR **abs/1612.01105** (2016)
42. Zhao, J., Li, J., Cheng, Y., Zhou, L., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and A new benchmark for multi-human parsing. CoRR **abs/1804.03287** (2018)