Letter

Part Decomposition and Refinement Network for Human Parsing

Lu Yang, Zhiwei Liu, Tianfei Zhou, and Qing Song

Dear Editor,

This letter is concerned with human parsing based on part-wise semantic prediction. Human body can be regarded as a whole structure composed of different semantic parts, and the mainstream single human parser uses semantic segmentation pipeline to solve this problem. However, the differences between human parsing and semantic segmentation tasks bring some issues that are inevitable to avoid. In this paper, we propose a novel method called part decomposition and refinement network (PDRNet), which adopt part-wise mask prediction other than pixel-wise semantic prediction to tackle human parsing task. Specifically, we decompose the human body into different semantic parts and design a decomposition module to learn the central position of each part. The refinement module is proposed to obtain the mask of each human part by learning convolution kernel and convolved feature. In inference stage, the predicted human part masks are combined into a complete human parsing result. Through the decomposition, refinement and combination of human parts, PDRNet greatly reduces the confusion between the target human and the background human, and also significantly improves the semantic consistency of human part. Extensive experiments show that PDRNet performs favorably against state-of-the-art methods on several human parsing benchmarks, including LIP, CIHP and Pascal-Person-Part.

Introduction: The problem of assigning dense semantic labels to a human image, formally known as human parsing, is of great importance in computer vision as it finds many applications, including clothing retrieval, virtual reality, human-computer interaction [1], [2], etc. Generally speaking, the vast majority of the existing human parsing methods follow two paradigms: bottom-up and top-down. The bottom-up [3], [4] treats human parsing as a fine-grained semantic segmentation task, predicting the category of each pixel and grouping it into corresponding instances. The top-down [5]−[10] locates each instance in the image plane, and then segments each human part independently. Therefore, an accurate single human parser is particularly important for the top-down method. The mainstream single human parsers map the human body to the same size feature space [5], [7], [11], and use pixel-wise semantic segmentation pipeline to solve the problem. However, there are great differences between human parsing and semantic segmentation tasks. First of all, in the human parsing, all human bodies except the target human are regarded as the background, while semantic segmentation does not distinguish different human instances, but tends to treat the target human and the background human equally (background confusion errors). Secondly, each human part is an instance with boundary, and we need to assign the same semantic label to the whole part. However, semantic segmentation is a pixel-wise classification, which can not guarantee that all pixels in the one part can be predicted the same category (semantic inconsistency errors).

In this work, we are committed to solving the errors caused by the differences between method and objective in human parsing. We abandon the process of semantic segmentation, learn from the idea of instance segmentation [12]−[14], decompose the human body into different semantic parts, segment each part mask independently, and then combine them into a complete human structure. Specifically, we propose a decomposition module to predict the centers and categories of human parts on the feature map. The decomposition module encodes the position information of human parts into the spatial dimension, and encodes the category information into the channel dimension. Therefore, the prior geometric context of the human body is retained in the feature map, which effectively avoids the confusion between the target human and the background human. In order to obtain the mask of each part, we propose a refinement module. The refinement module consists of two branches, one is used to learn the convolution kernel at the center of each part, the other is used to learn the convolved feature. We use dynamic convolution to generate the mask for each part, which converts the traditional pixel-wise semantic segmentation problem into a more concise binary part-wise mask segmentation. In inference stage, we present a human parsing probability map combination method based on the predicted human part categories and masks. The predicted mask with the highest score of each category is sampled and weighted fusion is carried out according to the quality score [11], and finally combined into a complete human parsing result.

As shown in Figs. 1(a)−1(c) , we call the proposed method of decomposition, refinement and combination of human body as PDRNet. Experiments show that, PDRNet has achieved state-of-the-art performances on four benchmarks, including CIHP [3], Pascal-Person-Part [15] and LIP [16]. Meanwhile, we also verify that PDRNet can significantly reduce background confusion errors and semantic inconsistency errors through qualitative comparison.

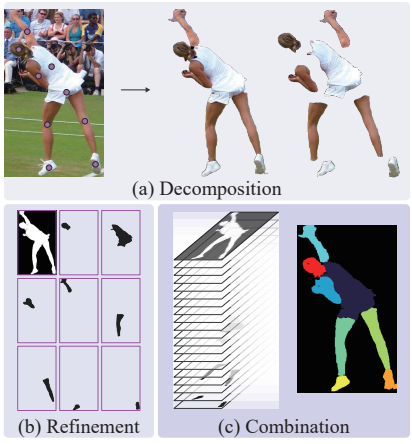

Fig. 1. Illustration of proposed PDRNet for human parsing. We decompose the human body into different semantic parts, segment each part mask independently, and then combine them into a complete human structure.

Related work:

• Human parsing: Human parsing has attracted a lot of research efforts in recent years [3], [15], [16]. Most of them regard it as a special case of semantic segmentation, and improve the performance by introducing attention mechanism [10], [17], auxiliary supervision [5], [15], [18], [19], human hierarchical structure [7], [8], [20] or quality estimation methods [9], [11]. Some earlier studies introduced human structure prior knowledge by designing hand-crafted features [21] or grammar model [22]. Attention mechanism [10] is then

Corresponding author: Qing Song.

Citation: L. Yang, Z. W. Liu, T. F. Zhou, and Q. Song, " Part decomposition and refinement network for human parsing," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 6, pp. 1111–1114, Jun. 2022.

L. Yang and Q. Song are with the Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: soeaver@bupt.edu.cn; priv@bupt.edu.cn).

Z. W. Liu is with the Institute of Automation, Chinese Academy of Sciences, National Laboratory of Pattern Recognition, Beijing 100190, China (e-mail: zhiwei.liu@nlpr.ia.ac.cn).

T. F. Zhou is with the ETH Zurich, Zurich 8092, Switzerland (e-mail: ztfei.debug@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105647

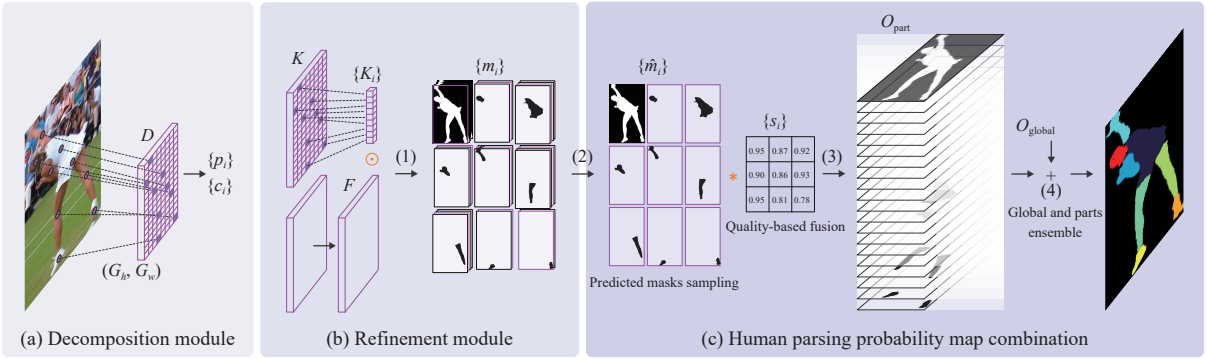

Fig. 2. Illustration of proposed decomposition module, refinement module and human parsing probability map combination.

adopted to construct the geometric context of the human body, which promotes the development of the community. In order to improve the ability of semantic segmentation network to understand human structure, some researchers uses keypoints [15] and edge [5] supervision to improve the model representation. Graph transfer learning [23], graph networks [7], [8] and semantic neural tree [20] are used to exploit the human representational capacity. However, it is difficult to eliminate the differences between the semantic segmentation method and the human parsing objective through these efforts. Urgently need a new perspective to solve the human parsing problem. Guided by this intuition, we try to decompose the human body into parts, segment each part mask independently, and combine them into a complete structure, which make a further step towards the consistency of method and objective in human parsing.

• Instance segmentation: Instance segmentation is a more challenging view of dense pixel prediction. It not only needs to predict the semantic categories at pixel level, but also distinguish different instances in the image simultaneously. According to whether the object proposal is explicitly adopted, the instance segmentation method can be divided into proposal-based [12] and proposal-free [13,] [14]. The first successful attempt is the proposal-based method, the milestone work is Mask R-CNN [12], which learns from the two-stage object detection framework by detecting the object box first and then segmenting the object mask in the box. The proposal-free methods attempt a more direct idea, using pixel grouping, object contour and other strategies [24] to obtain the object mask. Some pioneering efforts [13], [14] segment the mask directly without the box supervision, which has advantages in efficiency and performance. Our work is inspired by this idea, which can be viewed as a groundbreaking attempt to explore the method beyond semantic segmentation pipeline in the area of human parsing.

Methodology:

• Problem definition: Given an arbitrary human body image, the standard human parser [5], [6], [9] will adopt the semantic segmentation pipeline and regard it as a pixel-wise dense prediction problem: $\mathcal{O} = \mathcal{F} \odot \mathcal{K}$, where $\mathcal{F}$ denotes input feature map, $\mathcal{K}$ denotes convolution kernel, $\mathcal{O} \in \mathbb{R}^{N \times C \times H \times W}$ denotes output probability map (C is the number of parsing categories), $\odot$ is convolution operator. Generally speaking, cross entropy loss function is used to solve the problem. The human parsing is solved as a whole structure, and different parts are regarded as different categories. However, there are great differences between human parsing and semantic segmentation tasks, which lead to common background confusion error and semantic inconsistency error. In order to reduce these errors and improve the representation ability of human parser, our work adopts a more concise binary part-wise mask segmentation method. Inspired by some instance segmentation work [13], [14], we use decomposition module and refinement module to learn the position $\{\boldsymbol{p}_i\}$ (with category $\{\boldsymbol{c}_i\}$) and mask information $\{\boldsymbol{m}_i\}$ of each part respectively. Finally, using the post-processing method to combine learned masks of each part into a complete human parsing probability map $\mathcal{P}$ in inference stage.

• Decomposition module: The decomposition module is proposed to locate and classify each human part. This is the first step in whole process, and the subsequent steps depend on the output of decomposition module. Referring to [13], the decomposition module divides the input human body into a uniform grids, denoted as (G_h, G_w). As shown in Fig. 2(a), if the center of a part falls into a grid cell, that grid cell is responsible for predicting the semantic category. The non-existent part of the human body (due to the angle of view or special ornaments) is unnecessary to be predicted. This ensures that the decomposition module can predict the position and category of each part existing in the human body, that is, decompose the human body into different parts. The output of decomposition module is $\mathcal{D} \in \mathbb{R}^{N \times C \times G_h \times G_w}$, and most of the grids correspond to background category (negative samples). Therefore, we use focal loss [25] to classify each grid. Through the decomposition module, we get the position $\{\boldsymbol{p}_i\}$, $i \in \mathcal{D}_{pos}$ and category $\{\boldsymbol{c}_i\}$, $i \in \mathcal{D}_{pos}$ information of each part that existing in the human body.

• Refinement module: For an input human image, we get the position of each part through the decomposition module. We need to further obtain the masks on this basis. Therefore, we propose the refinement module to generate a unique refinement mask for each part. The different number of parts in each human body leads to the different number of positive samples predicted by the decomposition module. Thus, a dynamic scheme is required to generate an uncertain number of part masks. In the refinement module, we adopt the dynamic convolution, while two different branches are used to learn convolution kernel and convolved feature respectively. As shown in Fig. 2(b), the output feature $\mathcal{K}$ of kernel branch has the same dimension with the output of the decomposition module $\mathcal{D}$, that is $\mathcal{K} \in \mathbb{R}^{N \times C \times G_h \times G_h}$. According to the position information of human parts $\{\boldsymbol{p}_i\}$, we select the corresponding position features of $\mathcal{K}$ and concatenate them together as $\{\mathcal{K}_i\} \in \mathbb{R}^{N_{pos} \times C \times 1 \times 1}$. The feature branch outputs the convolved feature $\mathcal{F} \in \mathbb{R}^{N \times C \times H \times W}$, which has the same space dimension with the input human image. If we regard $\{\mathcal{K}_i\}$ as the 1×1 convolution kernel and $\mathcal{F}$ as the convolved feature, then we have

$$\{\boldsymbol{m}_i\} = \mathcal{F} \odot \{\mathcal{K}_i\} \tag{1}$$

where $\{\boldsymbol{m}_i\}$ is the learned masks of each human part. The area of each part in the human body is limited, so most of the pixels in $\{\boldsymbol{m}_i\}$ correspond to the background.We use dice loss to optimize each part mask in the refinement module. Thus, through the refinement module, we get the part masks $\{\boldsymbol{m}_i\}$ of the input human.

• Predicted masks sampling: In the decomposition module, in order to ensure that each human part is fully trained, we use the adaptive center range based on the part size when defining the centers. Therefore, a human part may be assigned to multiple grids, in other words, the refinement module will output redundant human part masks. One of the most intuitive ways to remove the redundant predictions is to use non-maximum suppression, e.g., greedy-NMS, soft-NMS [26], matrix-NMS [14]. But the human parsing task has its particularity, which every part of the human body is unique. As illustrated in Fig. 2(c) , we use a simpler way to sample the appropriate masks, only the mask with the highest quality score is sampled for each part

$$\{\hat{\boldsymbol{m}}_i\} = \max_s(\{\boldsymbol{m}_i^{(s)}\}), \quad s \in \{\boldsymbol{s}_i\} \tag{2}$$

where $\{\boldsymbol{s}_i\}$ denotes the quality scores of predicted part masks. According to the [11], the quality score of each mask is calculated by fusing the discrimination score of the decomposition module and the pixel score of the mask probability map.

• Quality-based human parts fusion: In order to obtain consistent results of human parsing, we need to combine $\{\hat{\boldsymbol{m}}_i\}$ into the whole human probability map $\mathcal{O}_{\text{part}} \in \mathbb{R}^{N \times C \times H \times W}$. We propose the quality-based human parts fusion method

$$\mathcal{O}_{\text{part}}^{(i)} = \begin{cases} (1-\hat{\boldsymbol{m}}_i) * \boldsymbol{s}_i, & i = 0 \\ \hat{\boldsymbol{m}}_i * \boldsymbol{s}_i, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $i \in C, C$ is the number of parsing categories. We use $(1-\hat{\boldsymbol{m}}_i)$, $i=0$ to represent the probability map of background category.

• Global and parts ensemble: The combined human probability map $\mathcal{O}_{\text{part}} \in \mathbb{R}^{N \times C \times H \times W}$ contains the information of human parts. We find the introduction of global human probability map can slightly improve the performance of human parsing. Therefore, we use the global and parts ensemble method to generate the final probability map $\mathcal{O}$ by adjusting the proportion of each other through hyper-parameter α

$$\mathcal{O} = \alpha * \mathcal{O}_{\text{part}} + (1-\alpha) * \mathcal{O}_{\text{global}}. \tag{4}$$

Here, $\mathcal{O}_{\text{global}}$ denotes the global human probability map.

Experiments:

• The decomposition and refinement modules: In order to verify the effectiveness of the two modules, we pursue concise design. As shown in Table 1, we investigate the conv. number in two modules and grid number in the decomposition module. Table 1(a) shows the conv. number in decomposition and refinement modules has slight impact on the performance, which proves that the performance improvement of PDRNet does not come from the larger network capacity. Therefore, in order to keep the computational efficiency, we choose two shared convolutions and one independent convolution. Table 1(b) shows the grid number (G_h, G_w) has a significant impact. Setting (G_h, G_w) to $(64, 48)$ can get the best performance, which can reach 57.97% mIoU.

Table 1. The Impact of Conv. Number (Top), grid Number (Bottom) of PDRNet on LIP val Set

Shared convs	Indep. convs	Pix Acc.	Mean Acc.	mIoU
1	1	88.38	68.66	57.52
2	**1**	88.53	**69.01**	**57.97**
4	1	88.44	68.89	57.78
4	2	**88.54**	68.94	57.95

(a) The Impact of Conv. Number in Two Modules

Grid num	Pix Acc.	Mean Acc.	mIoU
(32, 24)	87.61	66.43	55.01
(64, 48)	**88.53**	**69.01**	**57.97**
(96, 72)	87.94	68.17	57.21

(b) The Impact of Grid number in the Decomposition Module

• Human parsing probability map combination: Table 2 shows sampling the highest score mask (denoted as Top-1) can achieve higher performance than the NMS-based sampling [14]. We argue this is because each part of human body is unique, so there is no need for complex duplicate removal method. The core of probability map combination is to fuse the sampled part masks into a complete probability map. As shown in Table 3, the quality-base fusion is 0.75 points mIoU higher than direct fusion and 1.50 points mIoU higher than score-based fusion. By comparing direct fusion and quality-base fusion, we find that it is necessary to weight the mask quality, where some low quality predictions can be suppressed. Table 4 illustrates the influence of ensemble factor α on human parsing performance. Note that, it is not the best choice to use only the global probability map $\mathcal{O}_{\text{global}}$ (α = 0.0) or the part probability map $\mathcal{O}_{\text{part}}$ (α = 1.0). It can be observed that setting α = 0.75 achieves the best performance with 88.53% pix Acc., 69.01% mean Acc. and 57.97% mIoU.

• LIP [16]: We compare our proposed PDRNet with previous methods on LIP val set in Table 5. PDRNet achieves state-of-the-art with similar input size $(512 \times 384 \approx 473 \times 473)$, which yields 60.11% mIoU with HRNet-W48 backbone. In terms of pixAcc., mean Acc. and mIoU, PDRNet surpasses the best performing method QANet [11] by 0.22, 0.02 and 0.50 points, respectively.

Table 2. Performance Comparison of Different Kinds of Predicted Masks Sampling on LIP val Set

Methods	Pix Acc.	Mean Acc.	mIoU	Post-proc time (ms)
Hard-NMS	87.48	65.92	54.59	10
Matrix NMS	88.21	68.48	57.40	3
Top-1	**88.53**	**69.01**	**57.97**	1

Table 3. Performance Comparison of Different Kinds of Human Parts Fusion on LIP val Set

Methods	Pix Acc.	Mean Acc.	mIoU
Direct fusion	88.33	68.79	57.22
Score-based fusion	88.24	68.67	56.47
Quality-based fusion	**88.53**	**69.01**	**57.97**

Table 4. Ablation Study of Global and Parts Ensemble Factor α on LIP val Set

Methods	Pix Acc.	Mean Acc.	mIoU
Globle prob. only (α = 0.0)	88.10	68.63	56.89
Ensemble prob. (α = 0.25)	88.39	68.97	57.67
Ensemble prob. (α = 0.5)	88.46	68.95	57.88
Ensemble prob. (α = 0.75)	**88.53**	**69.01**	**57.97**
Part prob. only (α = 1.0)	88.31	68.42	57.37

• CIHP [3]: In the upper part of Table 6, we compare our method against 11 recent methods on CIHP val PDRNet achieves the best results in all metrics; specifically, 65.1% mIoU, 63.5% AP^p and 57.5% AP^r. Compare with previous state-of-the-art QANet [11], PDRNet yields 1.3 points mIoU and 1.7 points AP^p improvements.

• PASCAL-person-part [15]: PASCAL-person-part is a classic multiple human parsing benchmark with only 7 semantic categories. The lower part of Table 6 summarizes the quantitative comparison results with 9 competitors on PASCAL-person-part test set. Our PDRNet with HRNet-W48 backbone yields 73.3% mIoU, 63.9% AP^p and 59.1% AP^r, which again demonstrates our superior performance.

Conclusions: Using traditional semantic segmentation pipeline to process human parsing task will bring unavoidable semantic inconsistency and background confusion errors. This work draws on the idea of instance segmentation and proposes a new human parsing method to addresses these issues. Firstly, a decomposition module is designed to encode the human geometry prior and predict the center position of each part. Then, the refinement module is proposed to predict the part masks. In inference stage, combining the predicted human part masks into a complete human parsing probability map. We verify the superiority of our method on several benchmarks, and further prove that it can be flexibly combined with the existing human parsing frameworks.

Acknowledgments: This work was supported by the National Key Research and Development Program of China (2021YFF0500900).

Table 5. Comparison of Pixel Accuracy, Mean Accuracy and mIoU on LIP val Set

Methods	Input size	Pix Acc.	Mean Acc.	mIoU
Attention [29]	–	83.43	54.39	42.92
MMAN [30]	256×256	85.24	57.60	46.93
JPPNet [16]	384×384	–	–	51.37
CE2P [5]	473×473	87.37	63.20	53.10
BraidNet [28]	384×384	87.60	66.09	54.42
CorrPM [18]	384×384	–	–	55.33
OCR [31]	473×473	–	–	56.65
PCNet [10]	473×473	–	–	57.03
CNIF [7]	473×473	88.03	68.80	57.74
DTCF [19]	473×473	88.61	68.89	57.82
HHP [8]	473×473	89.05	70.58	59.25
SCHP [32]	473×473	–	–	59.36
QANet [11]	512×384	88.92	71.17	59.61
PDRNet (ours)	512×384	**89.15**	**71.19**	**60.11**

Table 6. Comparison With Previous Methods on Multiple Human Parsing on CIHP val and PASCAL-Person-Part Sets. Bold Numbers are State-of-the-Art on Each Dataset, † Denotes Using Multi-Scale Test Augmentation

Datasets	Methods	Backbones	Epochs	mIoU	AP^p	AP^p_{50}	PCP_{50}	AP^r	AP^r_{50}
CIHP [3]	Bottom-up								
	PGN† [3]	ResNet101	~80	55.8	39.0	34.0	61.0	33.6	35.8
	Graphonomy [23]	Xception	100	58.6	–	–	–	–	–
	CorrPM [18]	ResNet101	150	60.2	–	–	–	–	–
	One-stage Top-down								
	Parsing R-CNN [6]	ResNet50	75	56.3	53.9	63.7	60.1	36.5	40.9
	Unified [27]	ResNet101	~37	55.2	48.0	51.0	–	38.6	44.0
	RP R-CNN [9]	ResNet50	150	60.2	59.5	74.1	64.9	42.3	48.2
	Two-stage Top-down								
	CE2P [5]	ResNet101	150	59.5	–	–	–	42.8	48.7
	BraidNet [28]	ResNet101	150	60.6	–	–	–	43.6	49.9
	SemaTree [20]	ResNet101	200	60.9	–	–	–	44.0	49.3
	PCNet [10]	ResNet101	120	61.1	–	–	–	–	–
	QANet [11]	ResNet101	140	63.8	61.7	77.1	72.0	57.3	64.8
	PDRNet (ours)	ResNet101	140	**65.1**	**63.5**	**81.0**	**74.6**	**57.5**	**64.9**
PPP [15]	Bottom-up								
	PGN† [3]	ResNet101	~80	68.4	–	–	–	39.2	39.6
	Graphonomy [23]	Xception	100	71.1	–	–	–	–	–
	MGHP [4]	ResNet101	150	–	–	–	–	55.9	59.0
	One-stage Top-down								
	Parsing R-CNN [6]	ResNet50	75	62.7	49.8	58.2	48.7	40.4	43.7
	RP R-CNN [9]	ResNet50	75	63.3	50.1	58.9	49.1	40.9	44.1
	Two-stage Top-down								
	CNIF† [7]	ResNet101	150	70.8	–	–	–	–	–
	DTCF† [19]	HRNet-W48	200	70.8	–	–	–	–	–
	HHP† [8]	ResNet101	150	73.1	–	–	–	–	–
	QANet [11]	ResNet101	140	69.5	60.1	74.6	62.9	54.0	62.6
	PDRNet (ours)	ResNet101	140	70.3	60.4	75.1	63.5	54.8	64.0
	PDRNet (ours)†	HRNet-W48	140	**73.3**	**63.9**	**81.2**	**69.8**	**59.1**	**69.6**

References

[1] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. ECCV*, 2018.

[2] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," in *Proc. TPAMI*, 2021.

[3] K. Gong, X. Liang, Y. Li, Y. Chen, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. ECCV*, 2018.

[4] T. Zhou, W. Wang, S. Liu, Y. Yang, and L. V. Gool, "Differentiable multi-granularity human representation learning for instance-aware human semantic parsing," in *Proc. ICCV*, 2021.

[5] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang, "Devil in the details: Towards accurate single and multiple human parsing," in *Proc. AAAI*, 2019.

[6] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-CNN for instance-level human analysis," in *Proc. CVPR*, 2019.

[7] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning compositional neural information fusion for human parsing," in *Proc. ICCV*, 2019.

[8] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, and L. Shao, "Hierarchical human parsing with typed part-relation reasoning," in *Proc. CVPR*, 2020.

[9] L. Yang, Q. Song, Z. Wang, M. Hu, C. Liu, X. Xin, W. Jia, and S. Xu, "Renovating parsing r-CNN for accurate multiple human parsing," in *Proc. ECCV*, 2020.

[10] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang, "Part-aware context network for human parsing," in *Proc. CVPR*, 2020.

[11] L. Yang, Q. Song, Z. Wang, Z. Liu, S. Xu, and Z. Li, "Quality-aware network for human parsing," arXiv: 2103.05997, 2021.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," in *Proc. ICCV*, 2017.

[13] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *Proc. ECCV*, 2020.

[14] W. Xinlong, Z. Rufeng, K. Tao, L. Lei, and S. Chunhua, "Solov2: Dynamic, faster and stronger," in *Proc. NIPS*, 2020.

[15] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proc. CVPR*, 2017.

[16] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint human parsing and pose estimation network and a new benchmark," in *Proc. TPAMI*, 2018.

[17] L. Yang, Q. Song, Y. Wu, and M. Hu, "Attention inspiring receptive-fields network for learning invariant representations," in *Proc. TNNLS*, 2018.

[18] Z. Zhang, C. Su, L. Zheng, and X. Xie, "Correlating edge, pose with parsing," in *Proc. CVPR*, 2020.

[19] Y. Liu, L. Zhao, S. Zhang, and J. Yang, "Hybrid resolution network using edge guided region mutual information loss for human parsing," in *Proc. ACM MM*, 2020.

[20] R. Ji, D. Du, L. Zhang, L. Wen, Y. Wu, C. Zhao, F. Huang, and S. Lyu, "Learning semantic neural tree for human parsing," in *Proc. ECCV*, 2020.

[21] N. Wang and H. Ai, " Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. ICCV*, 2011.

[22] L. Zhu, Y. Chen, Y. Lu, C. Lin, and A. Yuille, "Max margin and/or graph learning for parsing the human body," in *Proc. CVPR*, 2008.

[23] K. Gong, Y. Gao, X. Liang, X. Shen, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proc. CVPR*, 2019.

[24] S. Wang, Y. Gong, J. Xing, L. Huang, C. Huang, and W. Hu, "RDSNet: A new deep architecture for reciprocal object detection and instance segmentation," in *Proc. AAAI*, 2020.

[25] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[26] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Improving object detection with one line of code," in *Proc. ICCV*, 2017.

[27] H. Qin, W. Hong, W.-C. Hung, Y.-H. Tsai, and M.-H. Yang, "A top-down unified framework for instance-level human parsing," in *Proc. BMVC*, 2019.

[28] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "BraidNet: Braiding semantics and details for accurate human parsing," in *Proc. ACM MM*, 2019.

[29] L. Chen, Y. Yang, J. Wang, W. Xu, and A. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. CVPR*, 2016.

[30] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. ECCV*, 2018.

[31] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. ECCV*, 2020.

[32] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," arXiv:1910.09777, 2019.