Renovating Parsing R-CNN for Accurate Multiple Human Parsing

Lu Yang[1 [0000−0003−3857−3982]], Qing Song[1 ⋆[0000−0003−4616−2200]], Zhihui Wang[1 [0000−0002−7547−8864]], Mengjie Hu[1 [0000−0001−7712−3322]], Chun Liu[1 [0000−0002−2834−9461]], Xueshi Xin[1 [0000−0002−9326−4499]], Wenhe Jia[1 [0000−0002−2516−957X]], and Songcen Xu[2 [0000−0002−0022−0906]]

[1] Beijing University of Posts and Telecommunications Beijing 100876, China
[2] Noah's Ark Lab, Huawei Technologies
{soeaver, priv, wangzh, mengjie.hu, chun.liu, xinxueshi, srxhemailbox}@bupt.edu.cn
xusongcen@huawei.com

Abstract. Multiple human parsing aims to segment various human parts and associate each part with the corresponding instance simultaneously. This is a very challenging task due to the diverse human appearance, semantic ambiguity of different body parts, and complex background. Through analysis of multiple human parsing task, we observe that human-centric global perception and accurate instance-level parsing scoring are crucial for obtaining high-quality results. But the most state-of-the-art methods have not paid enough attention to these issues. To reverse this phenomenon, we present Renovating Parsing R-CNN (RP R-CNN), which introduces a global semantic enhanced feature pyramid network and a parsing re-scoring network into the existing high-performance pipeline. The proposed RP R-CNN adopts global semantic representation to enhance multi-scale features for generating human parsing maps, and regresses a confidence score to represent its quality. Extensive experiments show that RP R-CNN performs favorably against state-of-the-art methods on CIHP and MHP-v2 datasets. Code and models are available at https://github.com/soeaver/RP-R-CNN.

Keywords: Multiple Human Parsing, Region-based Approach, Global Semantic Enhanced FPN, Parsing Re-Scoring Network

1 Introduction

Multiple human parsing [8] [21] [40] is a fundamental task in multimedia and computer vision, which aims to segment various human parts and associate each part with the corresponding instance. It plays a crucial role in applications in human-centric analysis and potential down-stream applications, such as person re-identification [22] [27], action recognition [5], human-object interaction [1] [29] [7], and virtual reality [14].

⋆ The corresponding author is Qing Song.

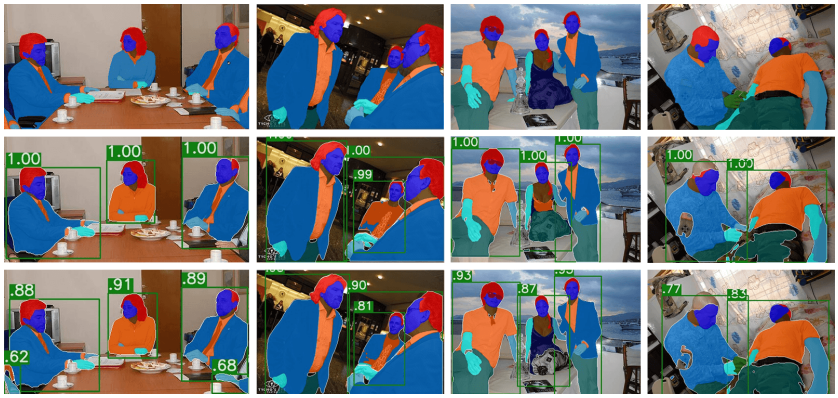

Fig. 1: Comparison of results between Parsing R-CNN and RP R-CNN on CIHP dataset. The first row is the ground-truth, the second row is the results of Parsing R-CNN, and the third is the predictions of RP R-CNN.

Due to the successful development of convolutional neural networks [33] [13], great progress has been made in multiple human parsing. Current state-of-the-art methods can be categorized into bottom-up, one-stage top-down, and two-stage top-down methods. The bottom-up methods [8] [9] [11] regard multiple human parsing as a fine-grained semantic segmentation task, which predicts the category of each pixel and grouping them into corresponding human instance. This series of methods will have better performance in semantic segmentation metrics, but poor in instance parsing metrics, especially easy to confuse adjacent human instances. Unlike bottom-up methods, the one-stage top-down [40] [30] and two-stage top-down methods [32] [24] [16] locate each instance in the image plane, and then segment each human parts independently. The difference between one-stage and two-stage is whether the detector is trained together with the sub-network used to segment the human part in an end-to-end manner. Compared with the bottom-up, the top-down methods are very flexible, which can easily introduce enhancement modules or train with other human analysis tasks (such as pose estimation [37], dense pose estimation [10] [40] or clothing parsing [35]) jointly. So it has become the mainstream research direction of multiple human parsing. But the human parts segmentation of each instance is independent and cannot make full use of context information, so the segmentation of some small scale human parts and human contours still needs to be improved. In addition, it is worth noting that neither bottom-up or top-down methods have a good way to evaluate the quality of predicted instance parsing maps. Resulting in many low-quality results that cannot be filtered.

In this paper, we are devoted to solving the problem of missing global semantic information in top-down methods, and evaluating the quality of predicted instance parsing maps accurately. Therefore, we propose Renovating Parsing R-CNN (RP R-CNN), which introduces a global semantic enhanced feature pyra-

mid network and a parsing re-scoring network to renovate the pipeline of top-down multiple human parsing. The global semantic enhanced feature pyramid network (GSE-FPN) is built on the widely used FPN [23]. We up-sample the multi-scale features generated by FPN to the same scale and fuse them. Using the global human parts segmentation to supervise and generate the global semantic feature, then fusing the global semantic feature with FPN features on the corresponding scales. GSE-FPN encourages the semantic supervision signal to directly propagate to the feature pyramid, so as to strengthen global information of learned multi-scale features. Global semantic enhanced features are passed to the Parsing branch through the RoIAlign [12] operation, ensuring that each independent human instance can still perceive the global semantic information, thereby improving the parsing performance of small targets, human contours, and easily confused categories. On this basis, the parsing re-scoring network (PRSN) is used to sense the quality of instance parsing maps and gives accurate scores. The score of instance parsing map is related to filtering low-quality results and sorting of instances, which is very important in the measurement of method and practical application. However, almost all the top-down methods use the score of detected bounding-box to represents the quality of instance parsing map [32] [40]. This will inevitably bring great deviation, because the score of bounding-box can only indicate whether the instance is human or not, while the score of the instance parsing map needs to express the segmentation accuracy of each human part, and there is no direct correlation between them. The proposed PRSN is a very lightweight network, taking the feature map and heat map of each human instance as input, using MSE loss to regress the mean intersection over union (mIoU) between the prediction and ground-truth. During inference, we use the arithmetic square root of predicted mIoU score multiply box classification score as human parsing final score.

Extensive experiments are conducted on two challenging benchmarks, CIHP [8] and MHP-v2 [45], demonstrating that our proposal RP R-CNN significantly outperforms the state-of-the-art for both bottom-up and top-down methods. As shown is Figure 1, RP R-CNN is more accurate in segmenting small parts and human edges, and the predicted parsing scores can better reflect the quality of instance parsing maps. The main contributions of this work are summarized as follows:

- A novel RP R-CNN is proposed to solve the issue of missing global semantic information and inaccurate scoring of instance parsing maps in top-down multiple human parsing.
- We introduce an effective method to improve the multiple human parsing results by fusing global and instance-level human parts segmentation.
- The proposed RP R-CNN achieves state-of-the-art on two challenging benchmarks. On CIHP `val` set, RP R-CNN yields 2.0 points mIoU and 7.0 points AP^p_{50} improvements compared with Parsing R-CNN [40]. On MHP-v2 `val` set, RP R-CNN outperforms Parsing R-CNN by 13.9 points AP^p_{50} and outperforms CE2P [32] by 6.0 points AP^p_{50}, respectively.

Our code and models of RP R-CNN are publicly available.

2 Related Work

Multi-Scale Feature Representations. Multi-scale feature is widely used in computer vision tasks [23] [19] [25] [38]. Long *et al.* [25] combine coarse, high layer information with fine, low layer information to generate fine features with high resolution, which greatly promotes the development of semantic segmentation. Lin *et al.* [25] present the feature pyramid network (FPN), and adopt it in object detection, greatly improve the performance of the small object. FPN is a feature pyramid with high-level semantics throughout, through top-down pathway and lateral connections. With the success of FPN, some researches introduce it into other tasks. Panoptic feature pyramid network (PFPN) [19] is proposed by Kirillov *et al.* and applied to panoptic segmentation. PFPN up-samples the feature pyramids and fuse them to the same spatial resolution, then a semantic segmentation branch is attached to generate high-resolution semantic features. However, the computation cost of PFPN is too large, and it only has a single scale semantic feature. Our GSE-FPN solves the above problems well, which adopts a lightweight up-sampling method, and use the global semantic feature to enhance the multi-scale feature.

Instance Scoring. Scoring the predicted instance is a challenging question. The R-CNN series [6] [31] of object detection approaches use the object classification score as the confidence of detection results. Recent studies [17] [34] [46] believe that it cannot accurately reflect the consistency between the predicted bounding-box and the ground-truth. Jiang *et al.* [17] present the IoU-Net, which adopts a IoU-prediction branch to predict the IoU between the predicted bounding box and the corresponding ground truth. Tan *et al.* [34] propose the Learning-to-Rank (LTR) model to produce a ranking score, which is based on IoU to indicate the ranks of candidates during the NMS step. Huang *et al.* [15] consider the difference between classification score and mask quality is greater in instance segmentation. They proposed Mask Scoring R-CNN, which uses a MaskIoU head to predict the quality of mask result. Different from these studies, this work analyzes and solves the inaccurate scoring of instance parsing maps for the first time. The proposed PRSN is concise yet effective, and reducing the gap between score and instance parsing quality.

Multiple Human Parsing. Before the popularity of convolutional neural network, some methods [39] [26] [41] using hand-crafted visual features and low-level image decompositions have achieved considerable results on single human parsing. However, limited by the representation ability of features, these traditional methods can not be well extended to multiple human parsing. With the successful development of convolutional neural networks [33] [20] [13] and open source of large-scale multiple human parsing datasets [8] [45], some recent researches [8] [32] [40] have achieved remarkable results in instance-level multiple human parsing. Gong *et al.* [8] present the part grouping network (PGN), which is a typical bottom-up method for instance-level multiple human parsing. PGN reformulates multiple human parsing as semantic part segmentation task and instance-aware edge detection task, the former is used to assign each pixel as human part and the latter is used to group semantic part into different human

Backbones	GT-box	GT-parsing	GT-score	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
R50-FPN				56.2	64.6	54.3	60.9
	✓			58.4$_{(+2.2)}$	58.0$_{(-6.6)}$	51.5$_{(-2.8)}$	62.3$_{(+1.4)}$
		✓		87.8$_{(+31.6)}$	91.4$_{(+26.8)}$	83.6$_{(+29.3)}$	90.6$_{(+29.7)}$
			✓	57.4$_{(+1.2)}$	73.7$_{(+9.1)}$	60.8$_{(+6.5)}$	60.9$_{(+0.0)}$

Table 1: Upper bound analysis of instance-level multiple human parsing via using ground-truth. All models are trained on CIHP `train` set and evaluated on CIHP `val` set. We replace the Bbox branch output with ground-truth box, replace the Parsing branch output with ground-truth segmentation or replace the instance score with ground-truth IoU, respectively. The results suggest that there is still room for improvement in human parsing and scoring.

instances. Ruan *et al.* [32] rethink and analyze the problems of feature resolution, global context information and edge details in human parsing task, and propose Context Embedding with Edge Perceiving (CE2P) framework for single human parsing. CE2P is a very successful two-stage top-down method, and wins the 1st places on three tracks in the 2018 2nd LIP Challenge. Parsing R-CNN [40] is proposed by Yang *et al.*, which is a one-stage top-down method for multiple human parsing. Based on the in-depth analysis of human appearance characteristics, Parsing R-CNN has made an effective extension on region-based approaches [6] [31] [23] [12] and significantly improved the performance of human parsing. Our work is based on the Parsing R-CNN framework, firstly introducing the global semantic information and reducing the gap between score and instance parsing quality for the top-down methods.

3 Renovating Parsing R-CNN

Our goal is to solve the issue of missing global semantic information and inaccurate scoring of instance parsing map in top-down multiple human parsing pipeline. In this section, we will introduce the motivation, architecture, and components of RP R-CNN in detail.

3.1 Motivation

In the top-down multiple human parsing pipeline, the network outputs three results: bounding-box, instance parsing map and parsing score. The importance of three outputs to the network performance is different. We take Parsing R-CNN [40] as baseline, and make an upper bound analysis of the three outputs. As shown in Table 1, we replace the predicted bounding-box with ground-truth, the multiple human parsing increases by 2.2 points mIoU [25], but AP^p_{50} and AP^p_{vol} [45] decrease. But when we replace the corresponding network output with the ground-truth of parsing map and mIoU, all the evaluation metrics have significant improvements. In particular, after the adoption of ground-truth parsing map, each evaluation metric has increased by about 30 points. These experimental results show that the accuracy of bounding-box has no significant impact on

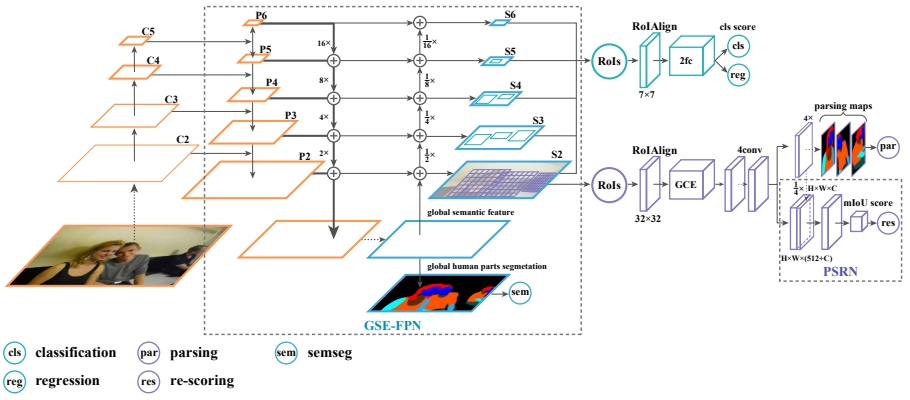

Fig. 2: RP R-CNN architecture. The input image is fed into a backbone with Global Semantic Enhanced FPN [23] to generate RoIs via RPN [31] (not shown in the figure) and RoI features via RoIAlign [12]. The global human parts segmentation is used to supervise and generate the global semantic feature. The BBox branch is standard component of Faster R-CNN which is used to detect human instance. The Parsing branch is mainly composed of GCE module [40] and Parsing Re-Scoring Network for predicting parsing maps and mIoU scores.

the multiple human parsing performance.However, the predicted parsing map and score still have a lot of room for improvement and not been paid enough attention by current studies. This is the motivation for our work.

3.2 Architecture

As illustrated in Figure 2, the proposed RP R-CNN involves four components: Backbone, GSE-FPN, Detector (RPN and BBox branch), and Parsing branch with PRSN. The settings of Backbone and Detector are the same as Parsing R-CNN [40]. The GSE-FPN is attached to the Backbone to generate multi-scale features with global semantic information. The Parsing branch consists of GCE module, parsing map output and Parsing Re-Scoring Network.

3.3 Global Semantic Enhanced Feature Pyramid Network

RoIAlign [12] aims to obtain the features of a specific region on the feature map, so that each instance can be processed separately. However, this makes the instance unable to directly perceive the global (context) information in branch. Global representation is crucial for human parsing, because we not only need to distinguish human body and background, but also give each pixel corresponding category through understanding the pose and recognizing the clothes the person wears [9]. Therefore, the information about the environment and objects around the human body is helpful for network learning. Some methods [43] perceive

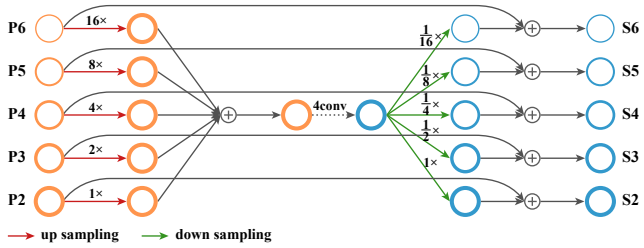

Fig. 3: Global semantic enhanced feature pyramid network (GSE-FPN). Circle is used to represent the feature map, and the circle thickness of circle is used to represent the spatial scale. The semantic segmentation loss is omitted here.

more valuable information by changing the area selected by RoIPool/RoIAlign. Different from these, we hope that by explicitly enhancing the global semantic representation of multi-scale features before RoIAlign. As a concrete example, the proposed GSE-FPN are illustrated in Figure 3, we adopt group normalization [36] and ReLU activation [28] after each convolutional layer.

High-resolution Feature. For semantic segmentation, high-resolution feature is necessary for generating high-quality results. Dilated convolution is an effective operation, which is adopted by many state-of-the-art semantic segmentation methods [42] [44] [3]. But dilated convolution substantially increases computation cost, and limits the use of multi-scale features. To keep the efficiency of network, and generate high-resolution features, we extend the multi-scale outputs of FPN [23]. Specifically, we up-sample the FPN generated multi-scale features to the scale of '$P2$' level by bilinear interpolation, which is 1/4 resolution of the original image. Each feature map is followed by a 1×1 256-d convolutional layer for aligning to the same semantic space, then these feature maps are fused together to generate high-resolution features.

Global Semantic Feature. As shown in Figure 3, we stack four 3×3 256-d convolutional layers after the high-resolution features to generate global semantic feature. Such a design is simple enough, but also can improve the representation ability of the network. In fact, we have tried some popular enhancement modules for semantic segmentation tasks, such as PPM [44] and ASPP [3] [4], but experiments show that these modules are not helpful to improve human parsing performance. A 1×1 C-dimension (C is the category number) convolutional layer is attached to the global semantic feature to predict human part segmentation.

Multi-scale Features Fusion. Through the above structure, we can get high-resolution global semantic feature. It is well known that semantic representation can bring performance gains to bounding-box classification and regression [12] [2]. Therefore, we down-sample the global semantic feature to the scales of $P3{\sim}P6$, and use element-wise sum to fuse them with the same scale FPN features. The generated new features are called global semantic enhanced multi-scale features, and denoted as $S2{\sim}S6$. We follow the Proposals Separation

Sampling [40] strategy that the $S2{\sim}S6$ level features are adopted for extracting region features for BBox branch and only $S2$ level is used for Parsing branch.

3.4 Parsing Re-Scoring Network

Parsing Re-Scoring Network (PRSN) aims to predict accurate mIoU score for each instance parsing map, and can be flexibly integrated into the Parsing branch.

Concise and Lightweight Design. PRSN follows the concise and lightweight design, which will not bring too much computation cost to model training and inference. PRSN receives two inputs, one is the $N{\times}512{\times}32{\times}32$ dimension parsing feature map, the other is the $N{\times}C{\times}128{\times}128$ dimension segmentation probability map (N is the number of RoIs, C is the category number). A max pooling layer with stride = 4 and kernel = 4 is adopted to make the probability map has the same spatial scale with parsing feature map. The down-sampled probability map and parsing feature map are concatenated together, then followed by two $3{\times}3$ 128-d convolutional layers. A final global average pooling layer, two 256-d fully connected layers, and MSE loss to regress the mIoU between the predicted instance parsing map and ground-truth.

IoU-aware Ground-truth. We define the mIoU between the predicted instance parsing map and matched ground-truth as regression target for PRSN. The common Parsing branch can output the segmentation probability map of each human instance, and calculate the loss with the segmentation ground-truth through a cross entropy function. Therefore, the mIoU between them can be calculated directly in the existing framework. It is worth noting that since the instance parsing ground-truth depends on the predicted region of Bbox branch, there is some deviation from the true location of human instance. However, we find that this deviation does not affect the effect of the predicting parsing score, so we do not make corrections to this deviation.

3.5 Training and Inference

As we introduce new supervision into RP R-CNN, there are some changes in the training and inference phases compared with the common methods [40] [30].

Training. There are three losses for global human parts segmentation and Parsing branch: $\mathcal{L}_{\text{sem}}$ (segmentation loss), $\mathcal{L}_{\text{par}}$ (parsing loss), $\mathcal{L}_{\text{res}}$ (re-scoring loss). The segmentation loss and parsing loss are computed as a per-pixel cross entropy loss between the predicted segmentation and the ground-truth labels. We use the MSE loss as re-scoring loss. We have observed that the losses from three tasks have different scales and normalization policies. Simply adding them degrades the overall performance. This can be corrected by a simple loss re-weighting strategy. Considering the losses of the detection sub-network, the whole network loss $\mathcal{L}$ can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{rpn}} + \mathcal{L}_{\text{bbox}} + \lambda_{\text{p}}\mathcal{L}_{\text{par}} + \lambda_{\text{s}}\mathcal{L}_{\text{sem}} + \lambda_{\text{r}}\mathcal{L}_{\text{res}}. \tag{1}$$

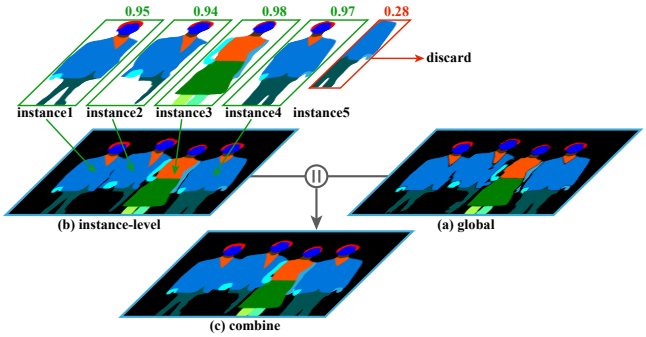

Fig. 4: Combination strategy for generating semantic segmentation results. ' ‖ ' symbol represents element-wise OR operation.

Dataset	Method	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
CIHP	Parsing R-CNN [40]	56.3	63.7	53.9	60.1
	Parsing R-CNN (our impl.)	56.2	64.6	54.2	60.9
	Δ	*-0.1*	*+0.9*	*+0.4*	*+0.8*
MHP-v2	Parsing R-CNN [40]	36.2	24.5	39.5	37.2
	Parsing R-CNN (our impl.)	35.5	26.6	40.3	37.9
	Δ	*-0.7*	*+2.1*	*+0.8*	*+0.7*

Table 2: Results of Parsing R-CNN [40] on the CIHP and MHP-v2 datasets. 'our impl.' denotes our implementation of Parsing R-CNN, which uses GN [36] in Parsing branch to stabilize the training.

The $\mathcal{L}_{\text{rpn}}$ and $\mathcal{L}_{\text{bbox}}$ are losses of RPN and BBox branch, each of which is composed of classification loss and box regression loss. By tuning λ_{p}, λ_{s} and λ_{r}, it is possible to make the network converge to optimal performance.

Inference. For network inference, we select top 100 candidate bounding-boxes per image from the human detection results. These candidates are fed into Parsing branch to predict instance parsing map and mIoU score. However, the mIoU score is only trained by positive samples, which leads to the lack of the ability to suppress negative samples. So we fuse mIoU score $\mathcal{S}_{\text{iou}}$ and classification score $\mathcal{S}_{\text{cls}}$ to generate the final parsing score $\mathcal{S}_{\text{parsing}} = \sqrt{\mathcal{S}_{\text{cls}} * \mathcal{S}_{\text{iou}}}$. In addition, we also find that the global human parts segmentation results are complementary to the Parsing branch result, the former has higher recall for each foreground, and the latter has better details. Thus, when generating semantic segmentation results, we adopt a new combination strategy, as shown in Figure 4. We filter out the low quality results based on the $\mathcal{S}_{\text{parsing}}$, and then generate instance-level human parts segmentation (b) *instance-level*. The instance-level (b) was and global human parts segmentation (a) *global* do element-wise OR operation to get the final result (c) *combine*. It is worth noting that if the results of (a) and (b) are different at the same pixel, and both of them are predicted as non-background

λ_p	AP^{bbox}	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
0.0	69.1	–	–	–	–
0.5	67.7	55.9	64.4	53.7	59.9
1.0	68.5	55.9	63.9	53.8	60.6
2.0	68.3	**56.2**	**64.6**	**54.3**	**60.9**
3.0	67.8	55.9	**64.6**	54.2	60.7

Table 3: Weight (λ_p) of parsing loss. All models are trained on CIHP `train` set and evaluated on CIHP `val` set (with $\lambda_s = 0.0$ and $\lambda_r = 0.0$).

λ_s	AP^{bbox}	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
0.0	69.1	56.2	64.6	54.3	60.9
0.5	67.9	57.0	65.1	54.6	61.1
1.0	67.7	57.8	66.5	55.0	61.7
2.0	67.4	**58.2**	**67.4**	**55.5**	**62.1**
3.0	67.1	58.0	67.4	55.3	61.8
Δ		*+2.0*	*+2.8*	*+1.2*	*+1.2*

Table 4: Weight (λ_s) of semantic segmentation loss (with $\lambda_r = 0.0$).

λ_s	AP^{bbox}	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
0.0	69.1	56.2	64.6	54.3	60.9
0.5	68.2	56.3	70.1	57.4	61.1
1.0	68.3	**56.4**	**70.3**	**57.6**	**61.3**
2.0	68.2	56.3	70.2	57.5	**61.3**
3.0	68.1	56.2	**70.3**	57.5	61.1
Δ		*+0.2*	*+5.7*	*+3.3*	*+0.4*

Table 5: Weight (λ_r) of re-scoring loss (with $\lambda_s = 0.0$).

category, we directly adopt the results of (b). This is because the (b) has a more accurate perception of the human parts inside each instance.

4 Experiments

In this section, we describe experiments on multiple human parsing of RP R-CNN. All experiments are conducted on the CIHP [8] and MHP-v2 [45] datasets. We follow the Parsing R-CNN evaluation protocols. Using mean intersection over union (mIoU) [25] to evaluate the human part segmentation. And using average precision based on part (AP^p) [45] as instance evaluation metric(s).

4.1 Implementation Details

Training Setup. All experiments are based on Pytorch on a server with 8 NVIDIA Titan RTX GPUs. We use 16 batch-size (2 images per GPU) and adopt ResNet50 [13] as backbone. The short side of input image is resized randomly sampled from [512, 864] pixels, and the longer side is limited to 1,400 pixels; inference is on a single scale of 800 pixels. Each image has 512 sampled RoIs for Bbox branch and 16 sampled RoIs for Parsing branch. For CIHP dataset, there are 135,000 iterations (about 75 epochs) of the training process, with a learning rate of 0.02 which is decreased by 10 at the 105,000 and 125,000 iteration. For MHP-v2 dataset, the max iteration is half as long as the CIHP dataset with the learning rate change points scaled proportionally.

Parsing R-CNN Re-Implementation. In order to better illustrate the advantages of RP R-CNN, we have re-implemented Parsing R-CNN according to

Inference methods	mIoU	Pixel acc.	Mean acc.
baseline	56.2	89.3	67.0
(a) semseg	50.2	88.0	61.3
(b) parsing	57.4	89.8	67.1
(c) combine	**58.2**	**90.2**	**69.0**
Δ	*+2.0*	*+0.9*	*+2.0*

Table 6: Semantic segmentation results of different inference methods on CIHP dataset. All models are trained on `train` set and evaluated on `val` set.

Methods	GSE-FPN	PRSN	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
RP R-CNN			56.2	64.6	54.3	60.9
	✓		58.2$_{(+2.0)}$	67.4$_{(+2.8)}$	55.5$_{(+1.2)}$	62.1$_{(+1.2)}$
		✓	56.4$_{(+0.2)}$	70.3$_{(+5.7)}$	57.6$_{(+3.3)}$	61.3$_{(+0.4)}$
	✓	✓	58.2$_{(+2.0)}$	71.6$_{(+7.0)}$	58.3$_{(+4.0)}$	62.2$_{(+1.3)}$

Table 7: Ablations of RP R-CNN on CIHP dataset. All models are trained on `train` set and evaluated on `val` set.

the original paper [40]. We find that the training of Parsing R-CNN is not very stable. We solve this issue by adding group normalization [36] after each convolutional layer of Parsing branch. As shown in Table 2, our re-implemented Parsing R-CNN achieves comparable performance with original version both on CIHP and MHP-v2 datasets. Therefore, this work takes our re-implemented Parsing R-CNN as the baseline.

4.2 Ablation Studies

In this sub-section, we assess the effects of different settings and components on RP R-CNN by details ablation studies.

Loss Weights. To combine our GSE-FPN with PRSN in Parsing R-CNN, we need to determine how to train a single, unified network. Previous studies demonstrate that multi-task training is often challenging and can lead to degraded results [18]. We also observe that adding the losses of all tasks directly will not give the best results. But grid searching three hyper-parameters (λ_p, λ_s and λ_r) is very inefficient, so we first determine λ_p, and then determine λ_s and λ_r separately to improve efficiency. As shown in Table 3, we find that the network performance is the best when $\lambda_p = 2.0$. We consider the group normalization layer makes the network convergence more stable, so that a proper large loss weight will bring higher accuracy. Table 4 shows that $\lambda_s = 2.0$ is the proper loss weight for semantic segmentation. Although increasing the weight of global human parts segmentation loss will slightly reduce the accuracy of human detection, the overall performance is improved. Table 5 shows that the re-scoring task is not sensitive to the loss weight, and its loss scale is smaller than other losses, so it has no significant impact on the optimization of other tasks in multi-task training. To sum up, we choose $\lambda_p = 2.0$, $\lambda_s = 2.0$ and $\lambda_r = 1.0$ as the loss weights of Eqn.(1).

Methods	GSE-FPN	PRSN	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
RP R-CNN			35.5	26.6	40.3	37.9
	✓		37.3 (+1.8)	28.9 (+2.3)	41.1 (+0.8)	38.9 (+1.0)
		✓	35.7 (+0.2)	39.6 (+13.0)	44.9 (+4.6)	38.2 (+0.3)
	✓	✓	37.3 (+1.8)	40.5 (+13.9)	45.2 (+4.9)	39.2 (+1.3)

Table 8: Ablations of RP R-CNN on MHP-v2 dataset. All models are trained on `train` set and evaluated on `val` set.

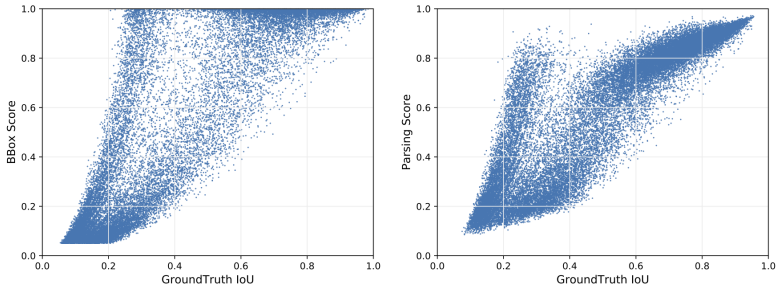

Fig. 5: Comparisons of ground-truth IoU vs. bbox score (**Left**) and ground-truth IoU vs. parsing score (**Right**) on CIHP dataset. All models are trained on `train` set and evaluated on `val` set.

Inference Methods. The combination inference method proposed in Figure 4 utilizes the complementarity of global human parts segmentation and Parsing branch results, and we give the detailed results in Table 6. The performance of using (a) *semseg* or (b) *parsing* alone is poor. The combination method (c) *combine* can significantly improve the metrics of semantic segmentation, and outperforms the baseline by 2 points mIoU.

Ablations on RP R-CNN. In Table 7, we perform the additional ablations of RP R-CNN on CIHP dataset. We observe that GSE-FPN is very helpful to the global human parts segmentation, which yields 2.0 points mIoU improvement. In addition, the global semantic feature also improves the instance metrics, AP^p_{50}, AP^p_{vol} and PCP_{50} increase 2.8, 1.2 and 1.2 points respectively. With PRSN, the improvements of instance metrics are very significant, AP^p_{50} improves 5.7 points, and AP^p_{vol} improves 3.3 points. With GSE-FPN and PRSN, our proposed RP R-CNN achieves 58.2 mIoU and 71.6 AP^p_{50} on CIHP. The additional ablations on MHP-v2 dataset is shown in Table 8. Through GSE-FPN and PRSN, the performance of human parsing is also significant improved. Particularly, AP^p_{50} and AP^p_{vol} are raised considerably by PRSN, 13.0 points and 4.6 points, respectively.

4.3 Comparison with State-of-the-Arts

We evaluate RP R-CNN on the CIHP and MHP-v2 datasets and compare the results to state-of-the-art including bottom-up and one-stage/two-stage top-down

Dataset	Methods	Backbones	Epochs	mIoU	AP^p_{50}	AP^p_{vol}	PCP_{50}
CIHP [8]	Bottom-Up						
	PGN† [8]	ResNet101	∼80	55.8	34.0	39.0	61.0
	DeepLab v3+ [4]	Xception	100	58.9	–	–	–
	Graphonomy [9]	Xception	100	58.6	–	–	–
	GPM [11]	Xception	100	60.3	–	–	–
	Grapy-ML [11]	Xception	200	60.6	–	–	–
	Two-Stage Top-Down						
	M-CE2P [32]	ResNet101	150	59.5	–	–	–
	BraidNet [24]	ResNet101	150	60.6	–	–	–
	SemaTree [16]	ResNet101	200	60.9	–	–	–
	One-Stage Top-Down						
	Parsing R-CNN [40]	ResNet50	75	56.3	63.7	53.9	60.1
	Parsing R-CNN† [40]	ResNeXt101	75	61.1	71.2	56.5	67.7
	Parsing R-CNN (our impl.)	ResNet50	75	56.2	64.6	54.3	60.9
	Unified [30]	ResNet101	∼37	55.2	51.0	48.0	–
	RP R-CNN (ours)	ResNet50	75	58.2	71.6	58.3	62.2
	RP R-CNN (ours)*	ResNet50	150	60.2	74.1	59.5	64.9
	RP R-CNN (ours)*†	**ResNet50**	150	**61.8**	**77.2**	**61.2**	**70.5**
MHP-v2 [45]	Bottom-Up						
	MH-Parser [21]	ResNet101	–	–	17.9	36.0	26.9
	NAN [45]	–	∼80	–	25.1	41.7	32.2
	Two-Stage Top-Down						
	M-CE2P [32]	ResNet101	150	**41.1**	34.5	42.7	**43.8**
	SemaTree [16]	ResNet101	200	–	34.4	42.5	43.5
	One-Stage Top-Down						
	Mask R-CNN [12]	ResNet50	–	–	14.9	33.8	25.1
	Parsing R-CNN [40]	ResNet50	75	36.2	24.5	39.5	37.2
	Parsing R-CNN (our impl.)	ResNet50	75	35.5	26.6	40.3	37.9
	RP R-CNN (ours)	ResNet50	75	37.3	40.5	45.2	39.2
	RP R-CNN (ours)*	**ResNet50**	150	38.6	**45.3**	**46.8**	**43.8**

Table 9: Multiple human parsing on the CIHP and MHP-v2 datasets. * denotes longer learning schedule. † denotes using test-time augmentation.

methods, shown in Table 9. On CIHP dataset, our proposed RP R-CNN achieves 58.2 mIoU and 71.6 AP^p_{50}, which surpasses Parsing R-CNN [40] in all respects. Compared with PGN [8], the performance advantage of RP R-CNN is huge, and AP^p_{50} is even 37.6 points ahead. With longer learning schedule (150 epochs), RP R-CNN achieves compared performance with one-stage top-down methods, *e.g.* M-CE2P and BraidNet. Even though we have adopted a lighter backbone (ResNet50 vs. ResNet101). Finally, using test-time augmentation, RP R-CNN with RestNet50 achieves state-of-the-art performance on CIHP.

On MHP-v2 dataset, RP R-CNN achieves excellent performance. We can observe that our RP R-CNN outperforms Parsing R-CNN consistently for all the evaluation metrics. And compared with M-CE2P, RP R-CNN yields about 10.8 point AP^p_{50} and 4.1 points AP^p_{vol} improvements. With RestNet50 backbone, it gives new state-of-the-art of 45.3 AP^p_{50}, 46.8 AP^p_{vol} and 43.8 PCP_{50}.

4.4 Analysis and Discussion

Effect of Parsing Re-Scoring Network. Figure 5 shows that the correlation between mIoU of the predicted parsing map with the matched ground-truth and the bbox/parsing score. As shown, the parsing score has better correlation with the ground-truth, especially for high-quality parsing map. However, it is

Fig. 6: Qualitative results of RP R-CNN on the CIHP and MHP-v2 datasets.

difficult to score low-quality parsing map, which is the source of some false positive detections. Therefore, the evaluation of low-quality prediction is still a problem to be solved.

Qualitative results. We visualize multiple human parsing results of RP R-CNN in Figure 6. We can observe that RP R-CNN has a good applicability to dense crowds and occlusions. In addition, the parsing score predicted by RP R-CNN reflects the quality of the parsing map.

5 Conclusions

In this paper, we proposed a novel Renovating Parsing R-CNN (RP R-CNN) model for solving the issue of missing global semantic information and inaccurate scoring of parsing result in top-down multiple human parsing. By explicitly introducing global semantic enhanced multi-scale features and learning the mIoU between instance parsing map and matched ground-truth, our RP R-CNN outperforms previous state-of-the-art methods consistently for all the evaluation metrics. In addition, we also adopt a new combination strategy, which improves the results of multiple human parsing by global semantic segmentation and instance-level semantic segmentation. We hope our effective approach will serve as a cornerstone and help the future research in multiple human parsing.

References

1. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV (2015) 1
2. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: Hybrid task cascade for instance segmentation. In: CVPR (2019) 7
3. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017) 7
4. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 7, 13
5. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: NIPS (2017) 1
6. Girshick, R.: Fast r-cnn. In: ICCV (2015) 4, 5
7. Gkioxari, G., Girshick, R., Dollar, P., He, K.: Detecting and recognizing human-object interactions. In: CVPR (2018) 1
8. Gong, K., Liang, X., Li, Y., Chen, Y., Lin, L.: Instance-level human parsing via part grouping network. In: ECCV (2018) 1, 2, 3, 4, 10, 13
9. Gong, K., Gao, Y., Liang, X., Shen, X., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: CVPR (2019) 2, 6, 13
10. Guler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018) 2
11. He, H., Zhang, J., Zhang, Q., Tao, D.: Grapy-ml: Graph pyramid mutual learning for cross-dataset human parsing. In: AAAI (2020) 2, 13
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 3, 5, 6, 7, 13
13. He, K., Zhang, X., Ren, S., , Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 2, 4, 10
14. Hsieh, C.W., Chen, C.Y., Chou, C.L., Shuai, H.H., Liu, J., Cheng, W.H.: Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In: ACM MM (2019) 1
15. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: CVPR (2019) 4
16. Ji, R., Du, D., Zhang, L., Wen, L., Wu, Y., Zhao, C., Huang, F., Lyu, S.: Learning semantic neural tree for human parsing. arXiv:1912.09622 (2019) 2, 13
17. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: ECCV (2018) 4
18. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018) 11
19. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: CVPR (2019) 4
20. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012) 4
21. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J.: Multi-human parsing in the wild. arXiv:1705.07206 (2017) 1, 13
22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018) 1
23. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 3, 4, 5, 6, 7

24. Liu, X., Zhang, M., Liu, W., Song, J., Mei, T.: Braidnet: Braiding semantics and details for accurate human parsing. In: ACM MM (2019) 2, 13
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 4, 5, 10
26. Luo, P., Wang, X., Tang, X.: Pedestrian parsing via deep decompositional network. In: ICCV (2013) 4
27. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV (2019) 1
28. Nair, V., Hinton, G.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010) 7
29. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018) 1
30. Qin, H., Hong, W., Hung, W.C., Tsai, Y.H., Yang, M.H.: A top-down unified framework for instance-level human parsing. In: BMVC (2019) 2, 8, 13
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015) 4, 5, 6
32. Ruan, T., Liu, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y., Huang, T.: Devil in the details: Towards accurate single and multiple human parsing. In: AAAI (2019) 2, 3, 4, 5, 13
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV (2015) 2, 4
34. Tan, Z., Nie, X., Qian, Q., Li, N., Li, H.: Learning to rank proposals for object detection. In: ICCV (2019) 4
35. Tangseng, P., Wu, Z., Yamaguchi, K.: Retrieving similar styles to parse clothing. TPAMI (2014) 2
36. Wu, Y., He, K.: Group normalization. In: ECCV (2018) 7, 9, 11
37. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018) 2
38. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV (2018) 4
39. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: CVPR (2012) 4
40. Yang, L., Song, Q., Wang, Z., Jiang, M.: Parsing r-cnn for instance-level human analysis. In: CVPR (2019) 1, 2, 3, 4, 5, 6, 8, 9, 11, 13
41. Yang, W., Luo, P., Lin, L.: Clothing co-parsing by joint image segmentation and labeling. In: CVPR (2014) 4
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016) 7
43. Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z., Zhou, H., Wang, X.: Crafting gbd-net for object detection. TPAMI (2017) 6
44. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 7
45. Zhao, J., Li, J., Cheng, Y., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: ACM MM (2018) 3, 4, 5, 10, 13
46. Zhu, B., Song, Q., Yang, L., Wang, Z., Liu, C., Hu, M.: Cpm r-cnn: Calibrating point-guided misalignment in object detection. arXiv:2003.03570 (2020) 4