


**ORIGINAL RESEARCH**

# Enhancing human parsing with region-level learning

Yanghong Zhou<sup>1</sup> | P. Y. Mok<sup>1,2</sup> <sup>1</sup>The Hong Kong Polytechnic University, Hong Kong, Hong Kong<sup>2</sup>Laboratory for Artificial Intelligence in Design, Hong Kong Science Park, Hong Kong, Hong Kong**Correspondence**

P. Y. Mok.

Email: [tracy.mok@polyu.edu.hk](mailto:tracy.mok@polyu.edu.hk)**Funding information**

Research Grants Council, Hong Kong Special Administrative Region, Grant/Award Numbers: 152112/19E, 152161/17E; The Laboratory for Artificial Intelligence in Design under the InnoHK Research Clusters, Hong Kong Special Administrative Region, Grant/Award Number: RP1-1

**Abstract**

Human parsing is very important in a diverse range of industrial applications. Despite the considerable progress that has been achieved, the performance of existing methods is still less than satisfactory, since these methods learn the shared features of various parsing labels at the image level. This limits the representativeness of the learnt features, especially when the distribution of parsing labels is imbalanced or the scale of different labels is substantially different. To address this limitation, a Region-level Parsing Refiner (RPR) is proposed to enhance parsing performance by the introduction of region-level parsing learning. Region-level parsing focuses specifically on small regions of the body, for example, the head. The proposed RPR is an adaptive module that can be integrated with different existing human parsing models to improve their performance. Extensive experiments are conducted on two benchmark datasets, and the results demonstrated the effectiveness of our RPR model in terms of improving the overall parsing performance as well as parsing rare labels. This method was successfully applied to a commercial application for the extraction of human body measurements and has been used in various online shopping platforms for clothing size recommendations. The code and dataset are released at this link <https://github.com/applezhouyp/PRP>.

**KEYWORDS**

computer vision, image processing, image segmentation, pose estimation

## 1 | INTRODUCTION

Human parsing has attracted considerable attention in recent years, because it is the core technology that supports many research studies and applications in the fields of retailing [1–3], social science [4, 5], medicine [6, 7], and even security [8, 9]. The aim of human parsing is to segment the pixels of an input image into regions according to different labels of body parts and clothes. As a pixel-level classification method, human parsing can be regarded as a branch of semantic segmentation. However, the application of general semantic segmentation methods for human parsing tasks can rarely achieve optimal performance, since human parsing involves fine-grained segmentation targets [10]. To achieve better parsing accuracy, one of the key research strategies is to consider the contextual relationships between different semantic labels, and to exploit the physical structure of the human body in network designs [11–14]. For example, Ji et al. [12] designed a semantic neural tree to encode

the physiological structure of the human body and achieved better parsing results. Wang et al. [13] applied a graph structure to develop network model by leveraging the relationships between body parts.

Despite the effectiveness of this approach, we argue that there are still several challenges that have been overlooked or are insufficiently addressed in existing methods of human parsing. The two most significant ones are (i) negative transfer and (ii) imbalanced labels. The problem of negative transfer arises because existing models typically learn shared features for all the segmented parts, that is, all body parts and types of clothing. However, from the perspective of multi-task learning, this shared approach to feature learning can result in negative transfer when the tasks are less closely related [15, 16]. The negative transfer problem may deteriorate the overall performance of the model, and this idea was recently discussed in the task of pose estimation [17]. In fact, the negative transfer can also be a serious issue in human parsing, but has received little

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *IET Computer Vision* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

attention so far. The second problem of imbalanced labels arises because the number of training samples for parsing classes are different in human parsing tasks, meaning that human parsing data are naturally imbalanced [18]. For example, the face class is typically present in all images, while the bag class is not, meaning that the number of samples in the face class is much more numerous than those in the bag class. Previous models have treated these imbalanced classes equally, which has inevitably resulted in biased learning results. Zhao et al. [19] proposed to balance feature magnitude to address the problem of imbalanced problem in semantic segmentation. Nevertheless, they ignored the imbalance between segmentation regions.

To tackle these two problems and to achieve better parsing accuracy, we propose a Region-level Parsing Refiner (RPR) in this paper. The basic idea is to enhance the image-level parsing results by introducing region-level learning. More specifically, we first segment the whole human body into several regions, for example, a head region (with face, hair, glasses etc.), and a body region (with upper clothing, dress, left leg, right leg etc.), and then use a parsing branch for each region to learn a parser for all the semantic labels within the region. Each parser consists of a pyramid scene parsing module [20], a decoder and a convolutional layer. The region-level parsing results are then integrated with the image-level results to give enhanced parsing score maps.

The underlying idea of RPR is simple yet effective. Each region-level parser aims to parse a small region of the human body, for example, the head. This branch can learn the features of the labels used for parsing more effectively, because (i) the label imbalance problem is alleviated by focusing on a specific region rather than the whole image; (ii) the target label area for parsing is reduced; and (iii) the variation in scale over the region is smaller than that over the entire image. The results of image-level parsing, which can be obtained using any human parsing model [21, 22], can be further refined by region-level parsing learning.

The main contributions of this paper are as follows.

- We propose a RPR module that uses region-level learning to refine the results of image-level parsing.
- The proposed RPR module is highly portable, and can be easily integrated with existing human parsing models, such as CE2P [21] or DeepLab3 [22] to give performance improvements.
- We conduct extensive experiments on two benchmark datasets, ATR [23] and LIP [24], to evaluate our RPR module, and the results demonstrate its effectiveness and portability.
- We demonstrate the proposed method using a real-world application.

## 2 | RELATED WORK

### 2.1 | Semantic segmentation

Current mainstream semantic segmentation methods are based on Fully Convolutional Networks (FCNs) [25]. Although the

use of an FCN makes possible to achieve an end-to-end training for semantic segmentation, there are still many challenges, such as resolution recovery, contextual feature capturing and boundary preservation. Since the resolution of the segmentation prediction is reduced and many local details are lost due to a series of pooling layers and convolution strides, an FCN upsamples the prediction and fuses with the features extracted from lower layers. Some researchers [26–28] have used encoder-decoder structures that downsample the features in the encoder and then upsample in the decoder.

To capture contextual features, Chen et al. [22] designed an atrous spatial pyramid pooling (ASPP) module that used convolutions with different dilation rates, while Zhao et al. [20] proposed a pyramid pooling module (PPM) in which features were fused at different pyramid scales. Although both ASPP and PPM can effectively capture contextual features, they cannot capture the object features well. Recently, Yuan et al. [29] exploited initial segmentation results to generate object features and then used the relationship between the pixel features and object features to refine the segmentation results.

To preserve boundary information, some earlier works have used conditional random fields (CRFs) either as a post-processing step [26, 30] or by end-to-end training [31, 32]. Rather than using costly CRFs, some models have learnt boundary prediction using a separate branch, and the learnt boundary features have then been combined to refine the results [33–36]. For example, Gated-SCNN [34] used a shape stream module to learn the boundary features and a regular stream module to learn the segmentation features, these features were then fused to refine the segmentation results. Li et al. [36] learnt body features using smoothed images and then obtained the boundaries by deducing the body features from the segmentation features. Bai and Zhou [37] directly applied an edge decoder for edges. Wang et al. [38] proposed a position attention module to emphasise the detail edge information in low-level features, and a channel correlation coefficient attention module to learn between-channel correlation in high-level features.

### 2.2 | Human parsing

Recently, a great deal of research effort has been devoted to human parsing and part segmentation, such as for animals and cars. Human parsing can be divided into single-human parsing and multi-human parsing. Single-human parsing [39] assumes that there is only one human instance in the input images and tackle the category-level human segmentation while multi-human parsing [40] aims to segment the human parts of all human instances in the input images. It is noted that single-human parsing can be used with human detector [14, 21] or trained with pose estimation [41] to address the problem of multi-human parsing [39]. In this paper, we mainly focus on single-human parsing.

Similar to the case of semantic segmentation, these approaches have been based on FCNs. For example, Liang et al. [42] proposed a co-CNN architecture based on FCNs for human parsing, and designed a local-global-local structure to achieve

better feature learning. A long short term memory (LSTM) network was proposed in ref. [43] to improve feature learning by jointly capturing the local and global spatial dependencies at different distances for semantic object parsing. In an extension to this work, a graph LSTM method [44] was proposed in order to fully exploit the natural properties of the image (e.g. the local boundaries). The model took a superpixel of an arbitrary shape as a node of a graph, and connected it with other superpixel nodes based on their spatial neighbourhood connections.

The knowledge of the hierarchical structure of human body has been exploited to design networks for human parsing. Zhu et al. [11] developed a progressive cognitive network to recognise different human body parts gradually, based on a component-aware region convolution structure. Wang et al. [13] adopted bottom-up and top-down hierarchical views of a human body structure to reason about human body part segmentation. Li et al. [14] constructed a dual graph reasoning framework using a hierarchical approach, while Ji et al. [12] proposed a neural tree to encode the structure of human body.

Pose and boundary information have also been exploited to improve the performance of human parsing. Wang et al. [45] leveraged co-occurrence of pose skeleton and the clothing parts to improve the parsing performance based on a chain-CRFs model. Gong et al. [46] proposed a part grouping network that combined semantic part segmentation and instance-aware edge detection into a single network to tackle instance-level human parsing. Feature resolution, global context information and edge details were used to design a context embedding with

edge perceiving (CE2P) framework [21] for human parsing. Su et al. [47] used a pose estimation network module to provide pose heatmaps about the human pose information for human parsing. Zhang et al. [48] proposed a Correlation Parsing Machine to take advantage of both edge and pose features to improve human parsing. Zeng et al. [49] used neural network search technology to search a optimise network structure to joint human parsing and pose estimation. Zhou and Mok [39] used the global joint representation for human parsing and proposed a pose-aware global representation network model enhance feature learning in human parsing. Yang et al. [50] leveraged both hierarchical human body structure and pose estimation for human parsing, and proposed a pose-guided hierarchical semantic decomposition and composition framework.

All of these methods have treated the classes equally and used a shared network to learn the features of all classes. In contrast, we propose a range of region-level parsers that learn the specific features for the related classes within a particular region. This approach is beneficial in terms of solving the data imbalance problem and avoiding negative transfer.

### 3 | THE METHOD

Figure 1 shows the overall architecture of a human parsing model equipped with the proposed RPR modules. It consists of a backbone network and an image-level parsing branch, and

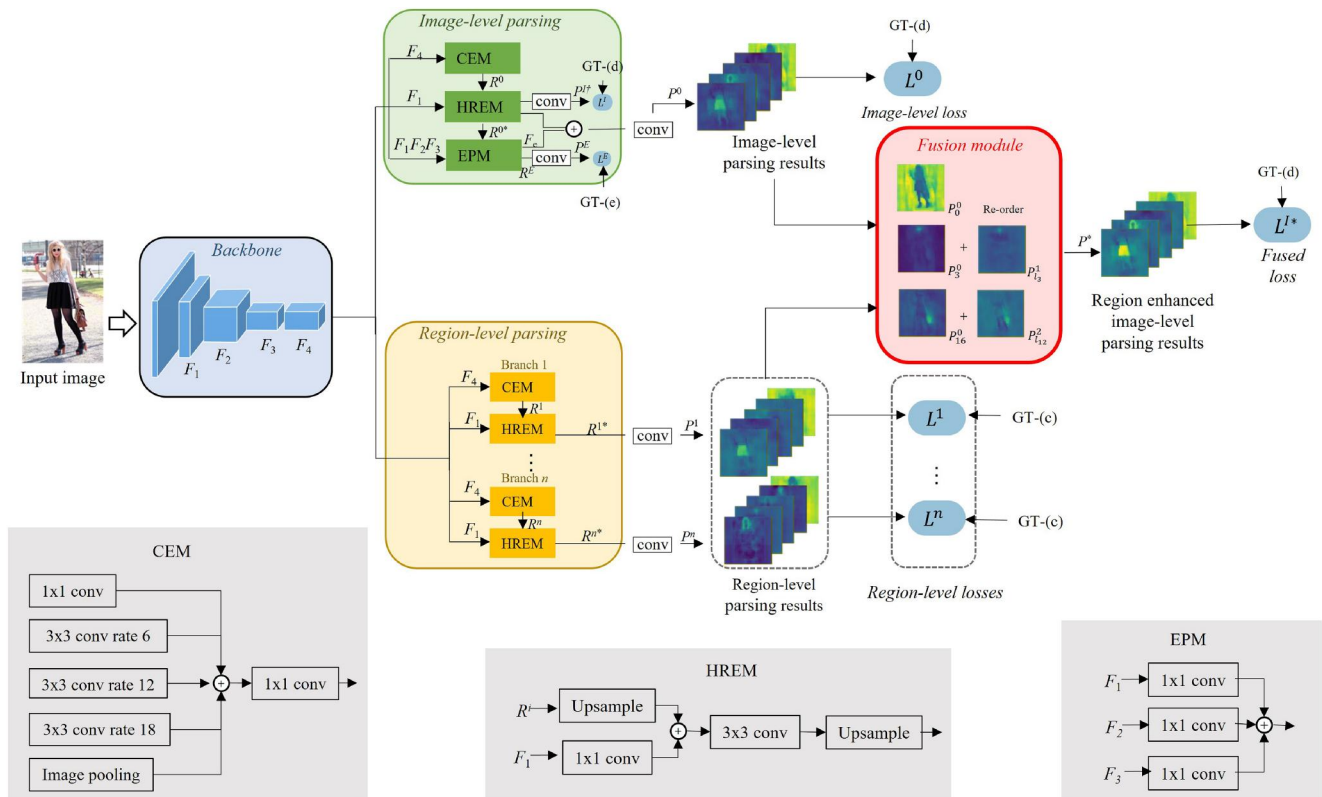


FIGURE 1 Architecture of the proposed Region-level Parsing Refiner (RPR) on a CE2P baseline.

the RPRs are applied on top of these as additional parsing branches. We present the baseline model, the construction of the RPR, the fusion module, and the loss function one by one.

### 3.1 | Baseline

A typical parsing model consists of a backbone and an image-level parser, this type of two-stage structure design is found in most existing human parsing methods [21, 22].

The backbone network is usually a convolutional neural network (CNN) that is good for the extraction of visual representations. Typical backbones include Visual Geometry Group [51], ResNet [52] and HRNet [53]. Since ResNet has been widely used in many human parsing models and has yielded superior performance [21, 22], we employ ResNet-101 [52] here as our backbone network. The outputs from the first to fourth residual stages of ResNet-101 are extracted as low-level features for further processing.

The image-level parser takes the low-level visual features as input, and outputs the pixel-level classification results in the form of a map of parsing scores. Different structures have been proposed in the literature for building human parsers. Of the existing designs for human parser, CE2P [21] is a representative and effective choice, and we thus adopt it as our image-level parser. The key structure of CE2P is shown as green squares in Figure 1, including Context Embedding Module (CEM), High-Resolution Embedding Module (HREM), and Edge Perceiving Module (EPM). It is important to note that the structure of the parser is not fixed, and our method is flexible enough to work with other parser structures. Since we do not focus this paper on identifying the best design for an image-level parser, we choose CE2P as a baseline to test our concept of region-level parsing learning. In the experimental section, we compare our method with different image-level parser designs.

In our model, an input image  $I$  is first processed by ResNet-101 to generate four sets of low-level features,  $F_1$ ,  $F_2$ ,  $F_3$  and  $F_4$ , where the subscript denotes the residual stage. Next, the low-level features are input to the image-level parser to generate the image-level score maps ( $P^0$ ). The data processing at the image-level parsing stage is similar to that for region-level parsing, and this is explained in detail below.

### 3.2 | Region-level parsing branches

Context Embedding Module consists of four average pooling layers with different scales, a  $1 \times 1$  convolution layer to reduce the feature channel after each stage of average pooling, and a concatenation operation to fuse the features upsampled from low-level features.

$$R^t = f(F_4, \theta_{\text{CEM}}^t) \quad t = 0, 1, \dots, n \quad (1)$$

where  $\theta_{\text{CEM}}^t$  are the parameters of the CEM of the  $t$ th parser. A value of  $t = 0$  refers to the image-level parser, while values of  $t = 1, \dots, n$  refer to the region-level parsers. Context

Embedding Module is designed to learn different sub-region representations [20]; it not only learns better global representations for local regions but also provides additional contextual information.

The low-level feature  $F_1$  and the high-level feature  $R^t$  from CEM each undergo a  $1 \times 1$  convolution to transform the features, and then unsample the transformed feature of  $R^t$  to the size of feature  $F_1$  with bilinear interpolation to form a HREM feature:

$$R^{t*} = \text{conv}(F_1) + \text{upsample}(\text{conv}(R^t)) \quad (2)$$

where  $\text{conv}(\cdot)$  represents a  $1 \times 1$  convolution and  $\text{upsample}(\cdot)$  represents bilinear interpolation. The HREM feature  $R^{t*}$  is followed by two  $1 \times 1$  convolutions to output score map  $P^t$ . Both CEM and HREM learn specific features for local regions, and therefore are included in region-level parser learning.

For image-level parser learning, the low-level image features  $F_1$ ,  $F_2$ , and  $F_3$  are input to learn two feature maps,  $F_e$  and  $R^E$ , as follows:

$$F_e = \text{upsample}(\text{conv}(F_1)) \oplus \dots \oplus \text{upsample}(\text{conv}(F_3)) \quad (3)$$

$$R^E = \text{upsample}(\text{conv}(F_{e1})) \oplus \dots \oplus \text{upsample}(\text{conv}(F_{e3})) \quad (4)$$

where  $\oplus$  denotes concatenation.

As shown in Equation (3),  $F_1$ ,  $F_2$  and  $F_3$  are first transformed as edge features  $F_{e1}$ ,  $F_{e2}$  and  $F_{e3}$ , respectively, via convolution operations. These edge features are upsampled and concatenated to form  $F_e$ . Similarly,  $F_{e1}$ ,  $F_{e2}$  and  $F_{e3}$  are then transformed by convolutions, upsampled and concatenated to form the feature  $R^E$  in Equation (4). The resulting  $R^{0*}$  from Equation (2) and  $F_e$  from Equation (3) are processed by two  $1 \times 1$  convolution operations to generate the image-level parsing results  $P^0$ .

### 3.3 | Fusion module

Based on the image-level parsing score maps  $P^0$  and the region-level parsing score maps  $P^1, P^2, \dots, P^n$ , we can refine the image-level parsing results using the region-level predictions as follows:

$$P_k^* = \begin{cases} P_k^0, & k = 0 \\ P_k^0 + P_{l_k}^t & k = 1, \dots, K, \text{ and } l_k \in \mathcal{L}^t, \end{cases} \quad (5)$$

where  $k$  represents the  $k$ th class in the image-level parser,  $K$  denotes the total number of parsing labels,  $l_k$  represents the index of the  $k$ th class in the  $t$ th region, and  $\mathcal{L}^t$  is the set of the labels covered in the  $t$ th region. As shown in Equation (5), for each class  $k$  in the image-level branch, except for the background ( $k = 0$ ), there is a score map in the prediction of region-level parsing branch  $P_{l_k}^t$  that corresponds to that class. We therefore extract the score maps for all the region-level branches and

re-order these score maps according to the class labels of the image-level branch. We then sum the re-ordered score maps with the score maps from the image-level parsing to give a refined prediction  $P^*$ . We refer Equation (5) as *late fusion*, because the final prediction results of all the branches in our model are fused.

Figure 2 gives an example of the qualitative score maps for the head-region, body-region and image-level parsers before fusion. The score maps after fusion are shown in Figure 3. By comparing Figure 2c and Figure 3, we can see that some missing and inaccurate regions in the image-level parsing results, such as the sunglasses, left shoe and right shoe, have been corrected by incorporating the region-level parsing results using the proposed fusion operation in Equation (5).

In addition to *late fusion*, we also introduce a *mid-fusion* scheme, in which we fuse the intermediate features to refine the parsing results. The *mid-fusion* scheme concatenates the extracted high-level features from the image-level parser  $R^{0*}$ , the high-level features of  $R^{t*}$  obtained from Equation (2) and the extracted edge features  $F_e$  from Equation (3), as follows:

$$R^* = R^{0*} \oplus R^{1*} \oplus \dots \oplus R^{n*} \oplus F_e \quad (6)$$

The concatenated features  $R^*$  are then input to two convolution layers to generate the refined image-level prediction  $P^*$ .

### 3.4 | Loss function

To train the model, we adopt a cross-entropy loss averaged over all pixel positions for all the network branches, including the

image-level parsing branch (shown in green in Figure 1), the region-level parsing branches (shown in orange in Figure 1), and the fusion parsing results (shown in red in Figure 1). The overall loss function is as follows:

$$L = L^*(Y, P^*) + L^{I^t}(Y, P^t) + L^I(Y, P^0) + L^E(E, P^E) + \sum_{t=1}^n \lambda_t \cdot L^t(Y^t, P^t) \quad (7)$$

where  $L^*$ ,  $L^I$ , and  $L^t$  denote the fusion, image-level and region-level parsing losses, and  $L^{I^t}$  is the auxiliary loss for image-level parsing and  $L^E$  is the edge loss.  $Y$  is the image-level parsing ground truth,  $\{Y^1, Y^2, \dots, Y^m\}$  is the set of region-level ground truths for all the regions,  $E$  is the edge ground truth, and  $\lambda_t$  is the weight of the  $t$ th region-level parsing loss.

As the number of classes covered in different region-level parsers varies, we alleviate the influence of data imbalance by setting the weight of  $t$ th region-level parser as follows:

$$\lambda_t = \frac{K}{|\mathcal{L}^t|}, \quad (8)$$

which is the ratio of the total number of classes to the number of classes covered in the  $t$ th region.

To generate the ground truths for each region-level parsing branch, we relabel the pixels belonging to the region with the labels corresponding to the definitions of the region-level branches, and relabel any other pixels as zero (i.e. background). Figure 4 shows the parsing ground truths for an example image.

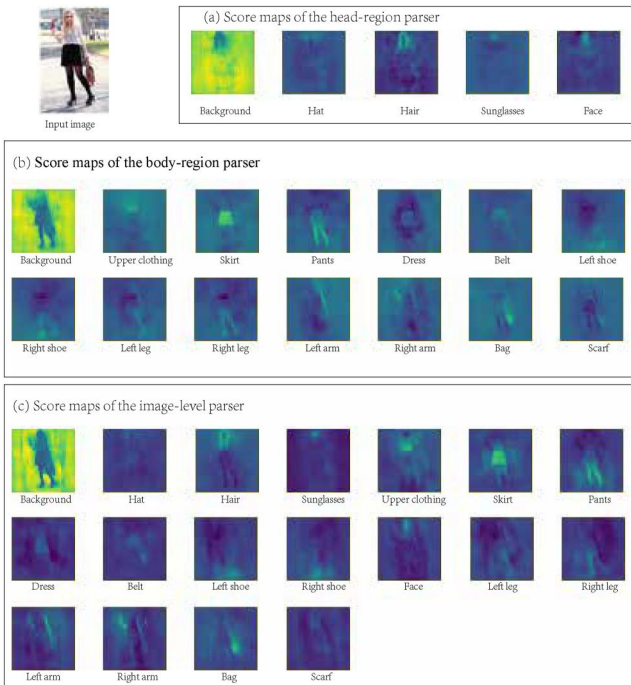


FIGURE 2 Example score maps from (a) the head-region parser, (b) the body-region parser, and (c) the image-level parser.

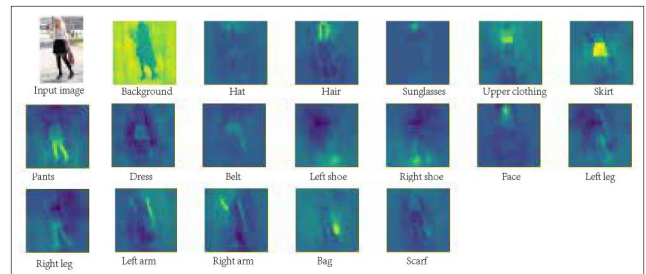


FIGURE 3 Score maps after fusion.

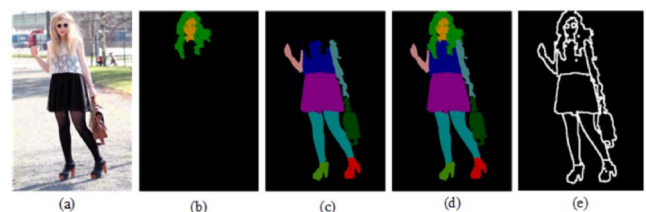


FIGURE 4 Parsing ground truths: (a) original image; (b) ground truth for the head-region; (c) ground truth for the body-region; (d) ground truth for the image-level branch; and (e) ground-truth for the edge features.

## 4 | EXPERIMENTAL RESULTS

In this section, we report the results of extensive experiments carried out on two benchmark human parsing datasets.

### 4.1 | Experimental settings

#### 4.1.1 | Datasets

Experiments were conducted on two benchmark datasets as follows. *The ATR dataset* [23] contains 7700 human images, each of which is annotated with a pixel mask with 18 semantic labels. We split the available data into three sets: 6000 images were used for training, 700 for validation and the remaining 1000 for testing. *The LIP dataset* [24] is a widely used human parsing dataset that is partitioned into training, validation and testing sets, containing 30,462, 10,000 and 10,000 images respectively. The LIP dataset contains images of people labelled in 20 body-part and clothing classes, which makes the parsing task more challenging.

#### 4.1.2 | Evaluation metrics

To facilitate comparative studies with other reported work using the two datasets, we calculated the pixel accuracy, foreground (denoted as ‘F.G.’) accuracy and mean accuracy (precision, recall score, and F1-score per pixel) and used as evaluation metrics [54] on the ATR dataset. For the LIP dataset, we calculated the pixel accuracy, mean accuracy and mean accuracy for the intersection over union (IoU) region [25]. We scaled the outputs of pixelwise prediction back to the size of the original ground-truth labels before calculating these metrics.

#### 4.1.3 | Region-level Parsing Refiner settings and implementation details

We defined two region-level parsers for the head and body regions. The head-region parser included classes for the hat, face, sunglasses and hair, whereas the body-region parser included classes of the gloves, upper clothing, dress, coat, socks, pants, jumpsuit, scarf, skirt, face, left arm, right arm, left leg, right leg, left shoe and right shoe. We used CE2P as baseline to evaluate the proposed RPR module, except for the results reported in Section 4.3.

We implemented our method using PyTorch. The input size of each image was  $473 \times 473$  for the ATR dataset and  $384 \times 384$  for the LIP dataset during training and testing, and we augmented the number of training images by mirroring, random cropping, and normalisation. We trained the RPR by fine-tuning the parameters from a model that was pretrained on the ImageNet dataset. The model was trained using a min-batch stochastic gradient descent with a momentum of 0.9, a weight decay of 0.0005 and an initial learning rate of 0.0001.

The learning rate was updated by multiplying the initial rate by  $\left(1 - \frac{it}{it_{max}}\right)^{0.9}$  after each iteration. We trained the model on two NVIDIA Giga Texel Shader eXtreme1080Ti graphics processing units for 180 epochs. The batch sizes were set to 8 and 16 for training on the ATR and LIP datasets respectively.

### 4.2 | Comparison with state-of-the-art models

#### 4.2.1 | Performance on the ATR dataset

We compared our proposed RPR with other state-of-the-art methods based on the defined metrics in Table 1. To ensure a fair comparison, we added another Chictopia10K dataset to the training data, in a way similar to the studies in refs. [20, 22, 42, 55, 56]. We applied RPR with two region-level parsers for the head and body regions on a CE2P baseline [21].

It can be seen from Table 1 that the value of the F1-score, the most important metric, showed significant improvement for our method compared to all other state-of-the-arts methods. For example, our model achieved F1-score value of 85.11%, compared to the values of 80.14% yielded by Co-CNN [42], 81.00% yielded by CPNet [47] and 81.76% by TGPNet [56]. As shown in Table 1, the pixel accuracy value of our RPR-CE2P is only slightly lower than that of CPNet [47] and TGPNet [56] by 0.2%. Compared to CPNet, the foreground accuracy (F.G. acc) of our RPR-CE2P is lower, because CPNet uses the edge-preserving filter as a pre-processor to denoise the training set's label and improve the annotation quality, particularly the foreground. However, CPNet's precision value is 0.91% lower than that of our RPR-CE2P. Compared to Co-CNN [42], the precision value of our RPR-CE2P is lower by 0.77%. This is because that Co-CNN uses the human detection algorithm [57] to detect the human body and uses the detected human body as the input, which avoids the inference of background and human scale. Comparatively,

**TABLE 1** Comparison of human parsing performance metrics (%) on the ATR test set.

Models	Pixel acc	F.G. acc	Precision	Recall	F1-score
Yamaguchi [58]	84.38	55.59	37.54	51.05	41.80
Paperdoll [1]	88.96	62.18	52.75	49.43	44.76
ATR [23]	91.11	71.04	71.69	60.25	64.38
DeepLab2 [22]	94.42	82.93	78.48	69.24	73.53
PSPNet [20]	95.2	80.23	79.66	73.79	75.84
DeepLab3+ [59]	95.96	83.04	80.41	78.79	79.49
Co-CNN [42]	96.02	83.57	84.95	77.66	80.14
CPNet [47]	96.46	90.38	83.27	80.03	81.00
TGPNet [56]	96.45	87.91	83.36	80.22	81.76
RPR-CE2P	96.25	88.23	84.18	86.05	85.11

our RPR-CE2P directly uses the original image as the input while the network is designed to automatically focus on the foreground pixels, achieving an impressive F1-score.

We compare the F1-scores for different foreground labels in Table 2, where the best performance in each label is highlighted in bold. Our method achieved better performance than the other methods on eight out of 17 labels. For the classes with few samples, such as hat, belt, bag and scarf, our method

**TABLE 2** Comparison of F1-scores (%) of our Region-level Parsing Refiner (RPR) module and other benchmark methods on the ATR testing set.

Models	Hat	Hair	Sgls	Uclo	Skirt	Pants
Yamaguchi [58]	8.44	59.96	12.09	56.07	17.57	55.42
PaperDoll [1]	1.72	63.58	0.23	71.87	40.20	69.35
ATR [23]	77.97	68.18	29.20	79.39	80.36	79.77
DeepLab2 [22]	72.25	82.58	44.61	87.12	80.91	85.80
PSPNet [20]	74.30	86.51	67.78	88.53	79.04	86.73
DeepLab3+ [59]	77.22	87.44	73.06	89.64	85.15	90.11
Co-CNN [42]	75.88	<b>89.97</b>	<b>81.26</b>	87.38	71.94	84.89
CPNet [47]	79.38	88.59	71.96	90.73	85.03	90.48
TGPNNet [56]	80.18	87.13	70.93	<b>91.01</b>	<b>88.95</b>	<b>90.72</b>
RPR-CE2P	<b>86.93</b>	89.59	79.27	90.40	82.95	90.64

Models	Dress	Belt	L-shoe	R-shoe	Face	L-leg
Yamaguchi [58]	40.94	14.68	38.24	38.33	72.10	58.52
PaperDoll [1]	59.49	16.94	45.79	44.47	61.63	52.19
ATR [23]	82.02	22.88	53.51	50.26	74.71	69.07
DeepLab2 [23]	79.05	24.96	65.44	65.70	85.33	80.21
PSPNet [20]	77.14	41.76	64.53	62.94	89.45	82.55
DeepLab3+ [59]	79.99	44.48	70.08	71.13	90.53	85.60
Co-CNN [42]	71.03	40.14	<b>81.43</b>	<b>81.49</b>	<b>92.73</b>	88.77
CPNet [47]	79.94	38.16	76.31	75.10	91.00	88.38
TGPNNet [56]	<b>87.42</b>	51.73	75.13	75.36	89.78	89.06
RPR-CE2P	81.87	<b>61.75</b>	79.88	79.83	91.88	<b>90.97</b>

Models	Rleg	Larm	Rarm	Bag	Scarf
Yamaguchi [58]	57.03	45.33	46.65	24.53	11.43
PaperDoll [1]	55.60	45.23	46.75	30.52	2.95
ATR [23]	71.69	53.79	58.57	53.66	57.07
DeepLab2 [22]	80.34	73.04	74.49	78.33	46.99
PSPNet [20]	81.92	77.68	78.01	77.69	49.83
DeepLab3+ [59]	85.25	81.96	82.48	81.73	53.46
Co-CNN [42]	88.48	89.00	88.71	83.81	46.24
CPNet [47]	88.19	84.67	85.55	83.58	61.79
TGPNNet [56]	88.73	83.91	83.96	84.72	52.86
RPR-CE2P	<b>90.36</b>	<b>90.23</b>	<b>90.11</b>	<b>88.11</b>	<b>66.37</b>

yielded larger gains, for example, 86.93% versus 80.18% (TGPNNet) for the hat class, 61.75% versus 51.73% (TGPNNet) for belt, 88.11% versus 84.72% (TGPNNet) for bag, and 66.37% versus 61.79% (CPNet) for scarf. These results demonstrate that the RPR performs particularly well on rare labels, and this is one of main advantages of our method.

#### 4.2.2 | Performance on the LIP dataset

In this section, we compare our RPR-CE2P with nine other methods on the LIP dataset. The overall results are presented in Table 3. It can be seen that our PRP-CE2P outperformed the baseline model of CE2P [21] by 2.61 in terms of the mean IoU score, again the most important metric. Our RPR-CE2P achieved a score 0.75 higher than that of BraidNet [64], which is similar to CE2P except that the EPM module is replaced by a separated sub-net to preserve local details and a pairwise hard region embedding strategy is used. Compared to SNT [12], which exploited the hierarchical structure of human body in the network design, our RPR-CE2P model achieved a score that was higher by 0.31. These results demonstrate that our model can learn features for human parsing more effectively than methods that do not use region-level parsers.

Table 4 shows a comparison of per-class mean IoU results, where the best performance in each class is highlighted in bold. As shown, the performance of other image-level parsing methods was very poor on classes with scarce samples (e.g. scarf, jumpsuit, skirt, and dress), since these methods suffer from data imbalance issues. In comparison, our method yielded significant improvements on these challenging classes. For example, the mean IoU for the dress, skirt, glove, and scarf classes were increased by 3.88, 2.38, 1.63, and 0.77, respectively, compared to SNT [12]. Our RPR module is shown also effective on the LIP dataset and especially good on the rare classes.

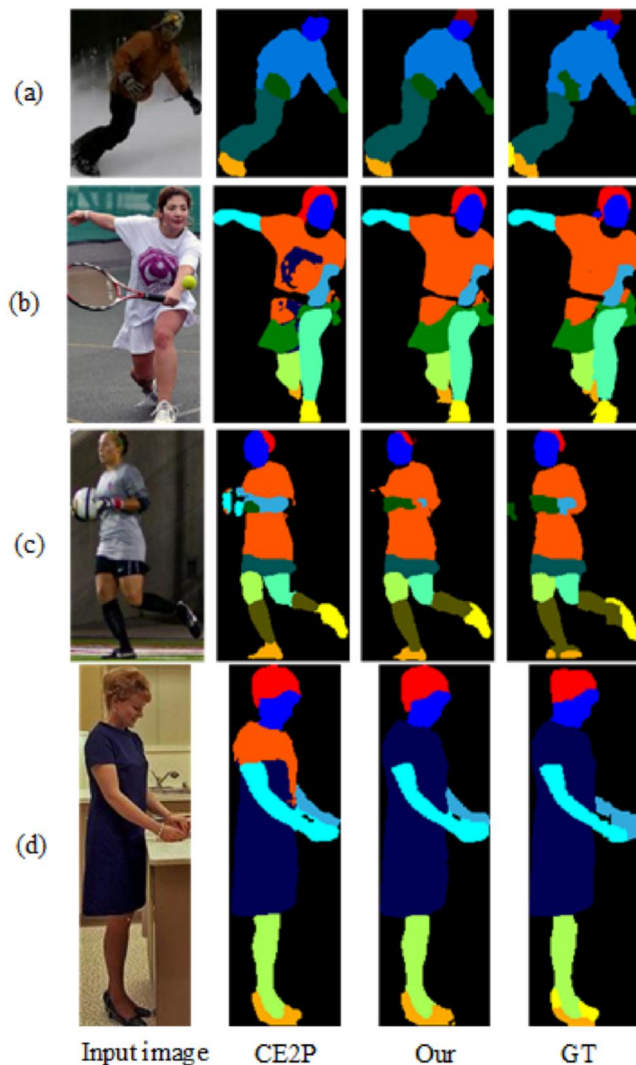
Some qualitative results are compared in Figure 5. In the example in Figure 5a, baseline CE2P did not recognise the hat due to low level of illumination, but our model correctly

**TABLE 3** Overall accuracy (%) achieved by our RPR-CE2P model and other benchmark methods on the LIP validation set.

Models	Pixel acc	Mean acc	Mean IoU
FCN-8s [25]	76.06	36.75	28.29
DeepLab2 [60]	82.66	51.64	41.64
MMAN [61] (ECCV'18)	-	-	46.81
SS-NAN [62]	87.59	56.03	47.92
JPPNet [63] (TPAMI'18)	86.39	62.32	51.37
CE2P [21] (AAAI'19)	-	-	52.56
BraidNet [64] (MM'19)	87.60	66.09	54.42
SNT [12] (ECCV'20)	88.10	70.41	54.86
RPR-CE2P	87.62	66.88	55.17

**TABLE 4** Comparison of per-class mean IoU accuracy (%) for our RPR-CE2P model and other benchmark methods on the LIP validation set.

Models	Hat	Hair	Glove	Sgl	Uclo
FCN-8s [25]	39.79	58.96	5.32	3.08	49.08
DeepLab2 [60]	57.94	66.11	28.50	18.40	60.94
MMAN [61]	57.66	65.63	30.07	20.02	64.15
SS-NAN [62]	63.86	70.12	30.63	23.92	70.27
JPPNet [63]	63.55	70.20	36.16	23.48	68.15
CE2P [21]	64.62	72.07	38.36	32.20	68.92
BraidNet [64]	66.80	72.00	42.50	32.10	69.80
SNT [12]	66.9	<b>72.20</b>	42.7	32.30	<b>70.10</b>
RPR-CE2P	<b>67.00</b>	71.95	<b>44.33</b>	<b>30.45</b>	69.93
Models	Dress	Coat	Socks	Pants	Jps
FCN-8s [25]	12.36	26.82	15.66	49.41	6.48
DeepLab2 [60]	23.17	47.03	34.51	64.00	22.38
MMAN [61]	28.39	51.98	41.46	71.03	23.61
SS-NAN [62]	33.51	56.75	40.18	72.19	27.68
JPPNet [63]	31.42	55.65	44.56	72.19	28.39
CE2P [21]	32.15	55.61	48.75	73.54	27.24
BraidNet [64]	33.70	57.40	49.00	74.90	32.40
SNT [12]	35.60	<b>57.50</b>	48.9	75.20	33.4
RPR-CE2P	<b>39.48</b>	56.64	<b>50.91</b>	<b>75.56</b>	<b>33.58</b>
Models	Scarf	Skirt	Face	Larm	Rarm
FCN-8s [25]	0.00	2.16	62.65	29.78	36.63
DeepLab2 [60]	14.29	18.74	69.70	49.44	51.66
MMAN [61]	9.65	23.20	69.54	55.30	58.13
SS-NAN [62]	16.98	26.41	75.33	55.24	58.93
JPPNet [63]	18.76	25.14	73.36	61.97	63.88
CE2P [21]	13.84	22.69	<b>74.91</b>	64.00	65.87
BraidNet [64]	19.30	27.20	74.90	65.50	67.90
SNT [12]	21.40	27.40	74.90	<b>66.80</b>	68.10
RPR-CE2P	<b>22.17</b>	<b>29.78</b>	74.66	66.21	<b>68.47</b>
Models	Lleg	Rleg	Lshoe	Rshoe	Bkg
FCN-8s [25]	28.12	26.05	17.76	17.70	78.02
DeepLab2 [60]	37.49	34.60	28.22	22.41	83.25
MMAN [61]	51.90	52.17	38.58	39.05	84.75
SS-NAN [62]	44.01	41.87	29.15	32.64	<b>88.67</b>
JPPNet [63]	58.21	57.99	44.02	44.09	86.26
CE2P [21]	59.66	58.02	45.70	45.63	87.41
BraidNet [64]	60.20	59.60	47.40	47.90	88.00
SNT [12]	<b>60.30</b>	<b>59.80</b>	<b>47.60</b>	<b>48.10</b>	88.20
RPR-CE2P	60.25	59.56	47.10	48.05	87.74

**FIGURE 5** Qualitative comparison of results: The first column shows the input images, the second shows the parsing results from CE2P, the third shows the parsing results from our RPR-CE2P method, and the last shows the ground truths.

segmented the hat. In Figure 5b and Figure 5c, the segmentation of the dress and upper clothing was confused by CE2P, but these two regions were well distinguished by our method. For the segmentation of rare classes, such as glove and scarf, our model also performed better than CE2P (Figure 5c). These results demonstrate that our proposed RPR has both good robustness and strong feature discrimination ability.

### 4.3 | Ablation study

To evaluate the effects of each of the different components in our RPR module, including the region-level parser, fusion module, loss weight settings, and the different baselines, we conducted a set of ablation experiments on the ATR dataset.



### 4.3.1 | Effects of the region-level parser

As discussed in Section 3.2, the proposed RPR module allows the additional region-level parser to reinforce the learning of specific classes. To evaluate the effectiveness of region-level parsing, we assembled different ablation models that incorporated different RPRs. In particular, we compare the performance of five different models, as follows: *CE2P* (baseline model without a region-level parser); *RPR-head* (baseline model with only a head-region parser); *RPR-body* (baseline model with only a body-region parser); *RPR2* (baseline model with both head- and body-region parsers); and *RPR3* (the RPR2 model with one additional sunglasses-region parser). The overall and per-class results from each of these models are compared in Tables 5 and 6.

We can drawing the following observations from the experimental results. First, both the head and body parsers not

**TABLE 5** Comparison of overall accuracy (%) achieved by our Region-level Parsing Refiner (RPR) with different region-level parser settings.

Models	Pixel acc	F.G. acc	Precision	Recall	F1-score
CE2P	95.45	85.10	79.85	82.37	81.09
RPR-head	95.63	85.79	81.01	83.62	82.29
RPR-body	95.74	86.21	81.53	83.40	82.45
RPR2	95.80	86.52	82.05	83.73	82.88
RPR3	95.85	86.65	82.00	83.86	82.92

**TABLE 6** Comparison of F1-score (%) of our model with different Region-level Parsing Refiner (RPR) settings.

Models	Hat	Hair	Sgls	Ucloth	Skirt	Pants
CE2P	82.78	88.39	74.57	88.37	78.78	88.09
RPR-head	85.44	89.13	76.97	88.73	78.51	88.93
RPR-body	84.88	89.07	76.41	89.04	79.94	88.93
RPR2	85.75	88.97	77.18	89.28	80.20	89.46
RPR3	85.41	88.99	77.73	89.46	81.32	89.28

Models	Dress	Belt	Lshoe	Rshoe	Face	Lleg
CE2P	76.85	52.16	74.47	74.61	91.13	87.44
RPR-head	77.38	54.89	76.41	76.49	91.62	88.54
RPR-body	78.61	55.54	76.55	76.72	91.47	88.63
RPR2	79.06	54.60	76.84	76.58	91.68	89.72
RPR3	80.44	54.73	76.87	76.78	91.76	89.26

Models	Rleg	Larm	Rarm	Bag	Scarf	Bkg
CE2P	87.58	87.40	87.18	84.23	54.94	98.88
RPR-head	88.53	87.72	87.90	85.63	57.54	98.92
RPR-body	88.71	88.15	88.21	85.80	57.69	98.94
RPR2	89.50	88.75	88.78	86.00	59.50	98.90
RPR3	89.13	88.28	87.92	85.62	59.55	98.90

only improved the parsing performance on their specific regions, but also improved the performance on other regions. Second, the performance of the classes covered by a region-level parser would improve more than the classes not being covered by the region-level parser. Third, the simultaneous use of two region-level parsers (i.e. for the head and body regions) generally achieved better performance than using a single parser, since the two parsers cover all classes of the body parts and clothing (foreground). These results clearly show that the use of regional parsing learning can significantly improve the parsing results on the corresponding regions, and further support the idea of leveraging region-level parsing learning to enhance human parsing performance.

### 4.3.2 | Effects of the fusion module

We investigated the fusion module of our RPR network by comparing the performance of different fusion schemes. Table 7 shows the results of the *mid-fusion* scheme defined in Equation (6) and that of the *late fusion* scheme defined in Equation (5). It is shown that the late fusion scheme was consistently better than the mid-fusion scheme for all evaluation metrics, thus demonstrating the superiority of our design.

### 4.3.3 | Effects of weight loss

As discussed in Section 3.4, different weights can be set in the loss function Equation (7) to balance the learning between the region-level and image-level parsers. The specific way in which we set the loss weights is given in Equation (8). In this section, we report the effectiveness of different loss weight settings. Table 7 gives the performance of the model for two different values of  $\lambda$ . It can be seen that compared to the RPR with a loss weight  $\lambda = 1$ , the precision of the RPR model with  $\lambda_t$  calculated using Equation (8) improved by 0.64%, which demonstrates the effectiveness of our loss weight setting.

### 4.3.4 | Different baselines

As discussed in the Introduction, one of the key characteristics of the proposed RPR model is that its network structure is flexible, meaning that it can easily work with different baseline models. We demonstrate this advantage by applying RPR to three different baseline models: DeepLab3 [22], CE2P [21]

**TABLE 7** Comparison of overall accuracy (%) achieved by our Region-level Parsing Refiner (RPR) with different fusion schemes and loss weight settings.

Fusion	Loss weight	Pixel acc	F.G. acc	Precision	Recall	F1-score
Mid-fusion	$\lambda$ by (8)	95.74	86.20	81.53	83.40	82.45
Late fusion	$\lambda = 1$	95.70	86.01	81.41	83.77	82.57
Late fusion	$\lambda$ by (8)	95.80	86.52	82.05	83.73	82.88

and HRNet-OCR [29]. All selected baselines are representative network models for human parsing, DeepLab3 and CE2P use ResNet-101 as a backbone while HRNet-OCR based on HRNet. DeepLab3 and HRNet-OCR do not exploit edge information, whereas CE2P uses edge information to improve the parsing results. Table 8 presents the pixel accuracy, foreground accuracy, precision, recall and F1-scores of different baselines and their corresponding RPR-enhanced models. For all three baselines, our RPR module improved all metrics. In terms of the F1-score, our RPR-enhanced models outperformed DeepLab3 [22] by 1.20%, CE2P [21] by 2.24%, and HRNet-OCR [29] by 0.41%. This demonstrates both the effectiveness and the adaptability of our method.

#### 4.4 | Application

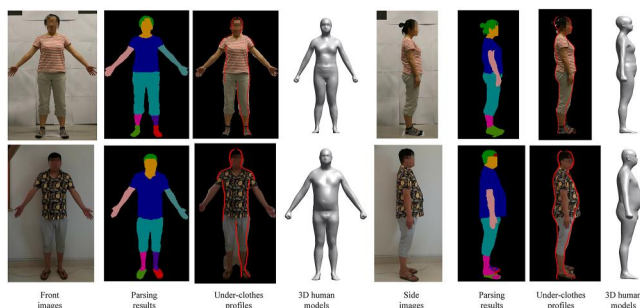
By applying this RPR method, we developed a mobile application called “1Measure” that allows users to obtain their accurate body measurements and shape information by taking two (front and back view) photographs, as shown in Figure 6. The RPR human parsing method is very suitable for 3D human shape modelling and size extraction. Our app can be downloaded from Apple App Store and Google Play Store and has been widely used in online shopping platforms for size recommendations [65].

#### 4.5 | Efficiency and resource consumption

Table 9 reports the model parameter numbers as well as speed based on the LIP validation set. It can be found that the

**TABLE 8** Comparison of overall accuracy (%) achieved by our Region-level Parsing Refiner (RPR) with different baseline models.

Models	Pixel acc	F.G. acc	Precision	Recall	F1-score
DeepLab3	95.41	85.08	79.79	81.51	80.64
RPR-DeepLab3	95.65	86.02	81.08	82.61	81.84
CE2P	95.45	85.10	79.85	82.37	81.09
RPR-CE2P	95.80	86.52	82.05	83.73	82.88
HRNet-OCR	95.78	86.49	82.49	84.12	83.29
RPR-HRNet-OCR	95.93	86.96	82.53	84.91	83.70



**FIGURE 6** Two examples of size and shape extraction.

frequency per second (FPS) value of the proposed model is lower than that of the baseline CE2P model by 6.7 FPS, because the region-level parsing results are learnt by extra parameters and then directly fused with image-level parsing results so as to enhance parsing results. To solve this problem, online knowledge distillation [66] can be used to distill the knowledge from region enhanced image-level parsing to the image-level parsing in the future work.

#### 4.6 | Failure analysis and limitations

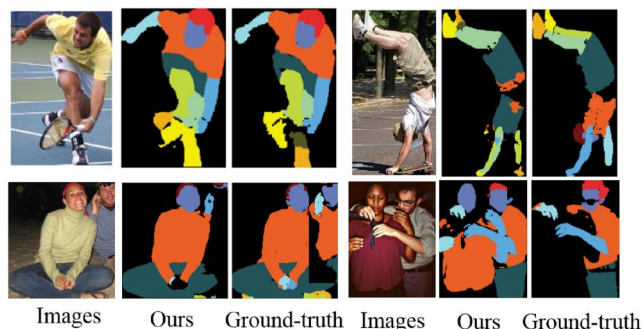
Figure 7 shows some failure examples of the proposed RPR-model on LIP validation set. As the first row of Figure 7 shown, when the pose of the human is very complex, our model could not segment well for certain regions. This is because that the complex poses with self-occlusion and upside-down orientation increase the difficulty of segmentation. In the future, we will use pose estimation to assist human parsing. In addition, our model tends to segment human parts of all persons in the image and may confuse with segmentation of different persons, resulting in failure cases. This may be related to ground-truth annotation problems. As shown in the second row of Figure 7, the part segmentation of all persons is annotated for the first image while only one person's part segmentation is annotated in the second image. Inconsistent data annotations may affect model training. We can alleviate this problem by re-annotating these noisy data.

### 5 | CONCLUSION

In this paper, we have proposed a novel PRP model for human parsing. Unlike existing methods that learn shared features for all labels, we apply region-level parsing learning to

**TABLE 9** Model parameters and the mean frequency per second (FPS) that was tested on a Giga Texel Shader eXtreme (GTX) 1080Ti GPU based on the LIP validation set.

Model	Model parameters	FPS
CE2P [21]	66.61M	26.88
RPR-CE2P (ours)	113.65M	20.18



**FIGURE 7** Failure examples of the LIP validation set.

enhance the representation of different parsing labels. Experiments show that compared with other state-of-the-art models, our proposed PRP model is more effective and is particularly superior for rare parsing labels. We have also carried out experiments on our method using a variety of baseline models, and have shown that our approach can work with all of them. Our PRP network allows for the flexible definition of region-level parsers, which can improve the overall network performance. In future work, we intend to explore different region-level parser settings, as well as use pose estimation and knowledge distillation, to optimise the network design.

## AUTHOR CONTRIBUTIONS

**Yanghong Zhou:** Conceptualisation; formal analysis; investigation; methodology; validation; writing - original draft. **P. Y. Mok:** Funding acquisition; investigation; project administration; resources; supervision; visualisation; writing - review & editing.

## ACKNOWLEDGEMENTS

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant Numbers 152161/17E & 152112/19E). This research is funded by the Laboratory for Artificial Intelligence in Design (Project Code: RP1-1), Innovation and Technology Fund, Hong Kong.

## CONFLICT OF INTEREST STATEMENT

Yanghong Zhou declares no conflict of interests; P.Y. Mok received grants disclosed in the funding information.

## DATA AVAILABILITY STATEMENT

The code and dataset are released at this link <https://github.com/applezhou/PRP>.

## ORCID

P. Y. Mok  <https://orcid.org/0000-0002-0635-5318>

## REFERENCES

1. Yamaguchi, K., et al.: Retrieving similar styles to parse clothing. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(5), 1028–1040 (2015). <https://doi.org/10.1109/tpami.2014.2353624>
2. Zhou, W., et al.: Fashion recommendations through cross-media information retrieval. *J. Vis. Commun. Image Represent.* 61, 112–120 (2019). <https://doi.org/10.1016/j.jvcir.2019.03.003>
3. Moghaddam, M., Charmi, M., Hassanpoor, H.: Jointly human semantic parsing and attribute recognition with feature pyramid structure in efficientnets. *IET Image Process.* 15(10), 2281–2291 (2021). <https://doi.org/10.1049/ipr2.12195>
4. Gan, C., et al.: Concepts not alone: exploring pairwise relationships for zero-shot video activity recognition. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
5. Liang, X., et al.: Proposal-free network for instance-level object segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(12), 2978–2991 (2018). <https://doi.org/10.1109/tpami.2017.2775623>
6. Goyal, M., et al.: Fully convolutional networks for diabetic foot ulcer segmentation. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 618–623. IEEE (2017)
7. Goyal, M., Yap, M.H., Hassanpour, S.: Multi-class Semantic Segmentation of Skin Lesions via Fully Convolutional Networks (2017). *arXiv preprint arXiv:1711.10449*
8. Wang, L., et al.: Deformable part model based multiple pedestrian detection for video surveillance in crowded scenes. In: 2014 International Conference on Computer Vision Theory and Applications (VISAPP), vol. 2, pp. 599–604. IEEE (2014)
9. Zhou, X., et al.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4966–4975 (2016)
10. Hu, J., et al.: Progressive refinement: a method of coarse-to-fine image parsing using stacked network. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
11. Zhu, B., et al.: Progressive cognitive human parsing. In: AAAI, pp. 7607–7614 (2018)
12. Ji, R., et al.: Learning Semantic Neural Tree for Human Parsing (2019). *arXiv preprint arXiv:1912.09622*
13. Wang, W., et al.: Hierarchical human parsing with typed part-relation reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8929–8939 (2020)
14. Li, T., et al.: Self-learning with rectification strategy for human parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9263–9272 (2020)
15. Caruana, R.: Multitask learning. *Mach. Learn.* 28(1), 41–75 (1997). <https://doi.org/10.1023/a:1007379606734>
16. Ruder, S.: An Overview of Multi-Task Learning in Deep Neural Networks (2017). *arXiv preprint arXiv:1706.05098*
17. Li, M., Gao, Y., Sang, N.: Exploiting Learnable Joint Groups for Hand Pose Estimation (2020). *arXiv preprint arXiv:2012.09496*
18. Huang, E., et al.: Learning rebalanced human parsing model from imbalanced datasets. *Image Vis Comput.* 99, 103928 (2020). <https://doi.org/10.1016/j.imavis.2020.103928>
19. Zhao, Y., Liu, S., Hu, Z.: Focal learning on stranger for imbalanced image segmentation. *IET Image Process.* 16(5), 1305–1323 (2022). <https://doi.org/10.1049/ipr2.12410>
20. Zhao, H., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
21. Ruan, T., et al.: Devil in the details: towards accurate single and multiple human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4814–4821 (2019)
22. Chen, L., et al.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4), 834–848 (2017). <https://doi.org/10.1109/tpami.2017.2699184>
23. Liang, X., et al.: Deep human parsing with active template regression. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(12), 2402–2414 (2015). <https://doi.org/10.1109/tpami.2015.2408360>
24. Gong, K., et al.: Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 932–940 (2017)
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
26. Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-wise Labelling (2015). *arXiv preprint arXiv:1505.07293*
27. Lin, G., et al.: Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
29. Yuan, Y., Chen, X., Wang, J.: Object-contextual Representations for Semantic Segmentation (2020)

30. Chen, L.-C., et al.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected Crfs, pp. 12 (2015)
31. Zheng, S., et al.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)
32. Arnab, A., et al.: Higher order conditional random fields in deep neural networks. In: European Conference on Computer Vision, pp. 524–540. Springer (2016)
33. Chen, L.-C., et al.: Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4545–4554 (2016)
34. Takikawa, T., et al.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5229–5238 (2019)
35. Ding, H., et al.: Boundary-aware feature propagation for scene segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6819–6829 (2019)
36. Li, X., et al.: Improving Semantic Segmentation via Decoupled Body and Edge Supervision (2020). *arXiv preprint arXiv:2007.10035*
37. Bai, X., Zhou, J.: Parallel global convolutional network for semantic image segmentation. *IET Image Process.* 15(1), 252–259 (2021). <https://doi.org/10.1049/ipr2.12025>
38. Wang, D., et al.: Bilateral attention network for semantic segmentation. *IET Image Process.* 15(8), 1607–1616 (2021). <https://doi.org/10.1049/ipr2.12129>
39. Zhou, Y., Mok, P.Y.: A pose-aware global representation network for human parsing. *IEEE Trans. Circ. Syst. Video Technol.* 33(4), 1710–1724 (2022). 2023. <https://doi.org/10.1109/tcsvt.2022.3213270>
40. Zhao, J., et al.: Understanding humans in crowded scenes: deep nested adversarial learning and a new benchmark for multi-human parsing. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 792–800 (2018)
41. Zhou, T., et al.: Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1622–1631 (2021)
42. Liang, X., et al.: Human parsing with contextualized convolutional neural network. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(1), 115–127 (2017). <https://doi.org/10.1109/tpami.2016.2537339>
43. Liang, X., et al.: Semantic object parsing with local-global long short-term memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3185–3193 (2016)
44. Liang, X., et al.: Semantic object parsing with graph lstm. In: European Conference on Computer Vision, pp. 125–143. Springer (2016)
45. Fan, W., et al.: Parsing fashion image into mid-level semantic parts based on chain-conditional random fields. *IET Image Process.* 10(6), 456–463 (2016). <https://doi.org/10.1049/iet-ipr.2015.0507>
46. Gong, K., et al.: Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–785 (2018)
47. Su, Z., et al.: Conditional progressive network for clothing parsing. *IET Image Process.* 13(4), 556–565 (2019). <https://doi.org/10.1049/iet-ipr.2018.5494>
48. Zhang, Z., et al.: Correlating edge, pose with parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8900–8909 (2020)
49. Zeng, D., et al.: Neural architecture search for joint human parsing and pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11.385–11.394 (2021)
50. Yang, B., et al.: Pose-guided hierarchical semantic decomposition and composition for human parsing. *IEEE Trans. Cybern.* 53(3), 1641–1652 (2023). <https://doi.org/10.1109/tycb.2021.3107544>
51. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2014). *arXiv 1409.1556*
52. He, K., et al.: Deep Residual Learning for Image Recognition (2015). *arXiv preprint arXiv:1512.03385*
53. Wang, J., et al.: Deep High-Resolution Representation Learning for Visual Recognition. TPAMI (2019)
54. Liu, S., et al.: Matching-cnn meets knn: quasi-parametric human parsing. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1419–1427 (2015)
55. Chen, L.-C., et al.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
56. Luo, X., et al.: Trusted guidance pyramid network for human parsing. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 654–662 (2018)
57. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
58. Yamaguchi, K., et al.: Parsing clothing in fashion photographs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3570–3577 (2012)
59. Chen, L.-C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
60. Chen, L.-C., et al.: Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs (2016). *arXiv:1606.00915*
61. Luo, Y., et al.: Macro-micro adversarial network for human parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434 (2018)
62. Zhao, J., et al.: Self-supervised neural aggregation networks for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 7–15 (2017)
63. Liang, X., et al.: Look into person: joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(4), 871–885 (2018). <https://doi.org/10.1109/tpami.2018.2820063>
64. Liu, X., et al.: Braidnet: braiding semantics and details for accurate human parsing. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 338–346 (2019)
65. Mok, P.Y. and Zhu, S.: Method And/or System for Reconstructing from Images a Personalized 3d Human Body Model and Thereof, uS Patent 10, 832 (2020) 472. Mobile App 1measure. <https://apps.apple.com/us/app/1measure/id1234853015>
66. Wang, L., Yoon, K.-J.: Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)

**How to cite this article:** Zhou, Y., Mok, P.Y.: Enhancing human parsing with region-level learning. *IET Comput. Vis.* 18(1), 60–71 (2024). <https://doi.org/10.1049/cvi2.12222>