IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 30, 2019, accepted October 23, 2019, date of publication November 4, 2019,
date of current version November 14, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2951182*

# A Part-Based Deep Neural Network Cascade Model for Human Parsing

**YANGHONG ZHOU[1], P. Y. MOK[1,2], (Member, IEEE), AND SHIJIE ZHOU[3]**

[1]Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong
[2]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen 518052, China
[3]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding author: P. Y. Mok (tracy.mok@polyu.edu.hk)

This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project152161/17E, in part by the Innovation and Technology Fund under Grant ITP/035/18TI, in part by The Hong Kong Polytechnic University under Grant G-YBRG and Grant G-UA9L, and in part by the Shenzhen Science and Technology Innovation Commission, China, under Project JCYJ20170303160155330.

**ABSTRACT** Human parsing is important for image-based human-centric and clothing analyses. With the development of deep neural networks, some deep human parsing methods were recently proposed, which substantially improve the parsing accuracy. However, some localized small regions (such as sunglasses) are not parsed well in these methods. In this paper, we propose a Part-based Human Parsing Cascade (PHPC) to segment human images, imitating the observational mechanism of how people, when first looking at a human image, quickly scan the entire photograph to first locate the face and then the body parts to see what clothing the person is wearing. The observational mechanism of human vision is used to establish a cascade relationship in designing our network, in which a head-parsing sub-network and a body-parsing sub-network are integrated to the cascade of human parsing networks. The head- and body-parsing sub-networks focus on the head and body classes, respectively, and add attention to the head and body in the final neural networks. Comprehensive evaluations on the ATR dataset have demonstrated the effectiveness of our method.

**INDEX TERMS** Human parsing, deep learning, fashion parsing, image segmentation, image understanding, convolutional neural networks.

## I. INTRODUCTION

Due to its importance to both human-centric and clothing analyses, human parsing has become an attractive subject for research over the past few years. Human parsing involves segmenting the person in a fashion image into regions according to their different body parts (e.g. face, left-arm, and right-leg) and the clothing (e.g. upper-clothing, dress, and trousers) that the person is wearing. Fig. 1 shows an example of human parsing. After human parsing, each pixel of the input image is given a label.

For human parsing research, researchers have mainly adopted one of two approaches: (1) a *bottom-up* approach [1], [2], where input images are first analysed using superpixel technique, and conditional random fields (CRFs) are then used to group and refine the initial superpixel results into larger segments and labels; and (2) a *top-down* approach [2], [3], where input images are first segmented into regions that are further classified into given labels.

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.


**FIGURE 1.** Example of human parsing: (a) Input image and (b) Parsing ground-truth.

Following the bottom-up approach, Yamaguchi *et al.* [1], [2] proposed to segment an image into superpixels and then predicted the clothing labels for each superpixel using a CRF model. This method performed quite well on the constrained parsing problem, where test images are parsed given user-provided tags that indicate the depicted clothing items. This

approach was less effective at unconstrained clothing parsing, however, where test images are parsed in the absence of any textual information.

Following the top-down approach, another group of researchers first aligned human parts by using the parselet representation as building blocks for a parsing model [3]. Parselets are groups of parsable segments that can generally be obtained from segmentation algorithms using low-level features. Dong *et al.* [4] built a deformable mixture-parsing model (DMPM) for human parsing to simultaneously handle the deformation and multimodalities of parselets. A DMPM seamlessly formulates the human parsing and pose estimation problem within a unified framework via a tailored And-Or graph, using parselets and a mixture of joint-group templates as the semantic components. Their work is limited by the suboptimal performance of many hand-designed intermediate components, such as handcrafted feature extraction and pose estimation [5].

Inspired by the remarkable improvement in accuracy introduced by the use of deep networks, deep human parsing methods have recently been proposed. Liang *et al.* [6] proposed a contextualised convolutional network, a fully convolutional network, to address the human parsing task. They integrated the cross-layer and global image-level contexts within the superpixel and cross superpixel neighbourhood contexts into a unified network. To increase the network capability, they incorporated Long Short-Term Memory (LSTM) layers into the convotutional neural networks (CNNs) in their extension work [7], which allowed memorisation of previous contextual interactions from local neighbouring positions and the whole image in previous LSTM layers. However, these parsing methods did not take into consideration of local and regional information, which are fundamental important because some items are so small that special attention must be drawn onto specific part regions to identify and describe such items.

In this paper, we propose a new human parsing network cascade that is inspired by the observational mechanism of how people, when first looking at a human photo, quickly scan the entire photograph to first locate the face and then the body parts to see what clothing the person is wearing. We propose in this paper a **P**art-based **H**uman **P**arsing **C**ascade (PHPC) of networks. To imitate human observation, we integrate a head-parsing sub-network and a body-parsing sub-network into a cascade of human parsing networks. The head- and body-parsing sub-networks focus on the head classes and body classes, respectively, and add the attention to the head and body in the final neural networks. We choose FCN-8s network [9] as our baseline network, as it is efficient and has shown great improvement on semantic segmentation. Semantic segment and human parsing are closely related, which we will discuss in next section. To evaluate the effectiveness of our PHPC method, we conducted several experiments on the ATR dataset [8], which we also train our PHPC model. For comparison purposes, we also trained a FCN-8s model using the method proposed by Long *et al.* [9] and a CRFasRNN model using the method proposed by Zheng *et al.* [10]. We will also discuss the effectiveness of super-pixel and CRF refinement in the discussion section. *The main contribution of our work is that we propose a novel PHPC model that mimics human vision.*

## II. RELATED WORK

We review the related work of this study, including human parsing and also some recent deep learning based development on semantic segmentation – a research area closely related to human parsing. Both semantic segmentation and human parsing attempt to assign a label to each pixel in an image.

### A. SEMANTIC SEGMENTATION

There have been a wide range of approaches using deep learning to tackle the semantic image segmentation. These approaches can be categorized into two main strategies. The first strategy to extract better meaningful features by improving mechanisms, such as using super pixels, multi-scale image size and optimized filters, etc. Mostajabi *et al.* [12] first obtained superpixels from the image and then used a feature extraction process on each of them. Chen *et al.* [13] combined CNN outputs from multiple scales image such that each feature vector represents a large contextual window around each pixel. Hariharan *et al.* [14] combine features from the intermediate layers to enhance the feature extraction. Yu and Koltun [15] proposed dilated convolutions to support exponential expansion of the receptive field without loss of resolution or coverage. Pinheiro and Collobert [16] employed an RNN to model the spatial decencies during scene parsing. Another strategy is to incorporate CRF into CNN to refine the result. Chen *et al.* [17] exploited a pre-trained CNN to generate deep features for CRF learning and illustrated that CRF learning with CNN features yields astounding results. Subsequent work [10], [18] have taken the idea further by incorporating a CRF as layers within a deep network and then learning parameters of both the CRF and CNN together via back propagation. For example, Zheng *et al.* [10] formulated a CRF as an RNN and then plugged into the network as a part of a CNN. However, these approaches have not employed higher order potentials, which have previously been shown to significantly improve segmentation performance. Arnab *et al.* [18] combined object-detection based potentials and superpixels based potentials into the CRF embedded within a deep network.

### B. HUMAN PARSING

Similar to semantic segmentation, human parsing is also to predict the label of each pixel in the image, but focus on human images, namely the segmentation of human body parts and clothing region from the background. Much research has devoted to human parsing in recent years. Most previous methods often rely on much complicated preprocessing, such as human pose estimation, bottom-up hypothesis and template dictionary learning. For example, Yamaguchi *et al.* [2] performed human pose estimation and attribute labeling

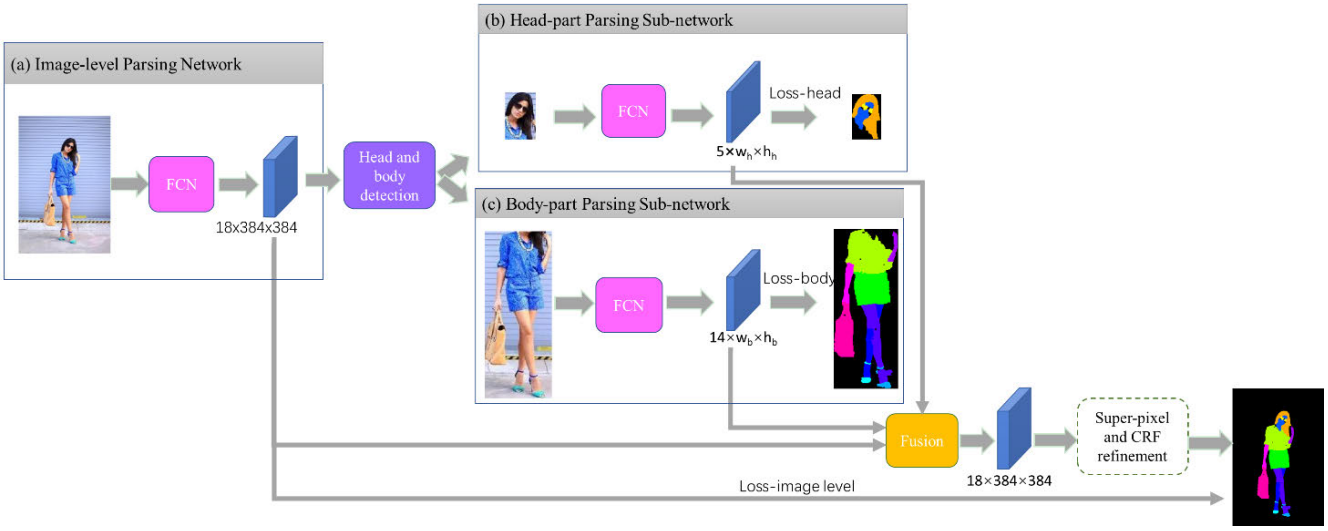



**FIGURE 2.** Part-based human parsing cascade of networks (PHPC): (a) image-level parsing network; (b) head-part parsing sub-network; and (c) body-part parsing sub-network.

sequentially and then used a retrieval-based approach to improve clothes parsing. For a query image, they found similar styles from a large database of tagged fashion images and used these examples to parse the query. Their approach combined parsing from pre-trained global clothing models, local clothing models learned on the fly from retrieved examples and transferred parse masks (paper doll items transfer) from retrieved examples. Similarly, Liu *et al.* [5] also based on the retrieval-based method. They retrieved the best matching clothing region of the test image from the annotated-parsed human image corpus and then used convolutional network to learn the inference and displacement coefficients. Dong *et al.* [4] proposed to use Parselet hypotheses to build the parsing model. Liang *et al.* [8] formulated the human parsing as an active template regression problem, where the template coefficients for each label mask and their corresponding locations were predicted using convolutional networks. But none of them is able to train in a fully end-to-end way over raw image pixels. In the recent past, a few methods have started using deep convolutional networks to train the network from end-to-end. Liang *et al.* [6] based on the fully convolutional network and proposed the contextualized convolutional network, which integrated the cross-layer context, global image-level context, within-superpixel context and cross-superpixel neighborhood context into a unified network. In their extension work [7], they incorporated short-distance and long-distance spatial dependencies into the feature learning by a Local-Global Long Short-Term Memory (LG-LSTM) layers. In [11], they split the feature map into several cells and only consider the local neighboring positions. So they proposed a Graph Long Short-Term Memory (Graph-LSTM) network, which is more naturally aligned with visual patterns in the image. However, these methods have not considered effect of the scale and localization of objects on parsing efficiency. Xia *et al.* [23] proposed to detect objects and parts regions based on the parsing results and then zoom into proper scales to refine the parsing. Gong *et al.* [24] proposed a Part Grouping Network (PGN), which jointly unify semantic part segmentation and instance-level human parsing, in which these two correlated task are able to mutually refine each other. Ruan *et al.* [25] conducted a great deal of rigorous experiments to clarify the properties affecting the performance of human parsing, including feature resolution, global context information and edge details, These methods actually has demonstrate the effectiveness of focusing on part regions, but all attention at the object level.

In sum, the existing methods of human parsing advance substantially in segmentation accuracy, but with a known drawback that not all fine-grained parts of the human are segmented well by a single classifier [26].

## III. THE PROPOSED PHPC NETWORKS

Fig. 2 shows the framework of the part-based human parsing cascade (PHPC). As shown, the PHPC consists of three networks: (1) an image-level parsing network, (2) a head-parsing sub-network, and (3) a body-parsing sub-network. These three networks were all built on the Fully Convolutional Neural network (FCN) and generate three feature maps. In Fig. 2, $w_h$ and $h_h$ indicate the width and height of head-part feature map, while $w_b$ and $h_b$ correspond to width and height of body-part feature map. Lastly, we combined these feature maps to refine the parsing results.

First, our image-level parsing network (a) generates an initial parsing result for the whole image. Second, based on the initial result, we detect the head and body regions of the image. Third, we input the head part and body part into our (b) head-parsing sub-network and (c) body-parsing sub-network, respectively. To capture the details for small items, the head and body sub-images are scaled up and double the original

image size for sub-networks (b) and (c). Finally, we combine all of the feature maps.

In sections III-A and III-B, we introduce in detail the image-level parsing network and the head- and body-parsing sub-networks. In section III-C, we introduce the fusion of all the networks.

### A. IMAGE-LEVEL PARSING NETWORK

For the image-level parsing network, we used the FCN proposed for semantic segmentation by Long *et al.* [9]. Compared with ordinary CNNs, FCN replaces all of the fully connected layers with convolutional layers. The FCN can therefore operate on an input of any size and produce an output of the same size, so it can be trained end-to-end, providing pixel-to-pixel labels from raw images.

In our method, we used the VGG 16-layer network [19] as our base network. There are 13 convolutional layers with Rectified Linear Units (ReLU), 5 pooling layers, and 3 fully connected layers in the VGG-16 network. To use the network for segmentation application, the 3 fully connected layers were converted to convolutional layers, resulting in feature maps that are 32 times smaller than the original size. The number of feature maps (also called channels) is the same as the number of class labels. Next, upsmapling and skip layers are added to convert and fuse feature maps from different convolutional layers to obtain the final feature maps with the same size of the original image. In our method, inputs to the image-level parsing network were 384 $\times$ 384 colour images, passing through a stack of convolutional and pooling layers.

We defined loss ($L$) by averaging the cross-entropy loss over all image pixels. More specifically, the loss function is defined as follows:

$$L = -\frac{1}{W \times H} \sum_{k=1}^{N} \sum_{j=1}^{W} \sum_{i=1}^{H} y_{i,j,k} \log \widehat{y}_{i,j,k} \tag{1}$$

where $\widehat{y}_{i,j,k}$ represents the likelihood for pixel $(i, j)$ to belong to label $k$, $y_{i,j,k}$ represents the ground-truth value if the label of pixel $(i, j)$ is $k$; $N$ is the total number of labels, $H$ and $W$ are the height and width of input image. The likelihood $\widehat{y}_{i,j,k}$ can be computed by the softmax function:

$$\hat{y}_{i,j,k} = \frac{\exp(z_{i,j,k})}{\sum_{i=1}^{N} \exp(z_{i,j,k})} \tag{2}$$

where $z$ is the output of the last layer of the network. The softmax function ensures a diffuse network output, so that the class with a high probability score is highlighted and the classes with lower scores are suppressed.

During training, the goal was to adjust the neural network weights so that the predictions matched the ground truth by updating the weights in the direction of a desceasing loss function value. We trained the parameters of the network to minimise the loss using Stochastic Gradient Descent (SGD), a common optimising algorithm in network training. SGD computes the gradient in every layer, updating the parameters layer by layer until the loss function (1) converges. The label prediction for pixel $(i, j)$ is obtained by $\arg\max_k y_{i,j,k}$.

**TABLE 1. Classes of image-level, head and body part networks.**

| Type | Classes |
|---|---|
| Image-level | background ($k$=0), hat ($k$=1), hair ($k$=2), sunglasses ($k$=3), upper-clothing ($k$=4), skirt ($k$=5), trousers ($k$=6), dress ($k$=7), belt ($k$=8), left-shoe ($k$=9), right-shoe ($k$=10), face ($k$=11), left-leg ($k$=12), right-leg ($k$=13), left-arm ($k$=14), right-arm ($k$=15), bag ($k$=16), and scarf ($k$=17) |
| Head | background ($k$=0), hat ($k$=1), hair ($k$=2), sunglasses ($k$=3), and face ($k$=4) |
| Body | background ($k$=0), upper-clothing ($k$=1), skirt ($k$=2), trousers ($k$=3), dress ($k$=4), belt ($k$=5), left-shoe ($k$=6), right-shoe ($k$=7), left-leg ($k$=8), right-leg ($k$=9), left-arm ($k$=10), right-arm ($k$=11), bag ($k$=12), and scarf ($k$=13) |

### B. HEAD AND BODY PARSING SUB-NETWORKS

By inputting an image into the image-level parsing network, we obtained an initial parsing result, which provided a label for each pixel. The feature maps generated by the image-parsing level network provided 18 labelled parts, so the head and body regions can be easily localised on the input image by combining some of the labelled parts together. The head and body parts of the image can be obtained using the smallest rectangle bounding boxes to crop the image with all corresponding classes. Table 1 shows the classes of each network.

The head and body part images are resized to double the original size, and then input to the corresponding sub-networks for training. The head- and body-parsing sub-networks have the same design as the image-level parsing network (discussed in section III-A), except that the output layer is different because of the different ground-truth.

### C. SUB-NETWORK COMBINATION

We obtain three feature maps from the image-level network and the head and body parsing sub-networks. Let $z^I_{k,i,j}, z^H_{k,i,j}, z^B_{k,i,j}$ denote the feature scores from the image-level FCN, head parsing sub-network and body parsing sub-network, respectively, where pixel $(i, j)$ belongs to the $k$-th label. According to the locations of the head and body regions of the images, recorded by the coordinates of bounding boxes $(x^H_{start}, x^H_{stop}, y^H_{start}, y^H_{stop})$ and $(x^B_{start}, x^B_{stop}, y^B_{start}, y^B_{stop})$, the head part feature maps $z^H$ and the body part feature maps $z^B$ can be merged to the feature maps of image-level FCN $z^I$ by the following formulae:

$$z^I_{x^H_{start}+i, x^H_{xstop}+j, k} = z^I_{x^H_{start}+i, x^H_{xstop}+j.k} + z^H_{i,j,k}, \quad k = 0, 1, 2, 3 \tag{3a}$$

$$z^I_{x^H_{start}+i, x^H_{xstop}+j, k+7} = z^I_{x^H_{start}+i, x^H_{xstop}+j, k+7} + z^H_{i,j,k}, \quad k = 4 \tag{3b}$$

$$z^I_{x^B_{start}+i, x^B_{xstop}+j, k} = z^I_{x^B_{start}+i, x^B_{xstop}+j, k} + z^H_{i,j,k},$$

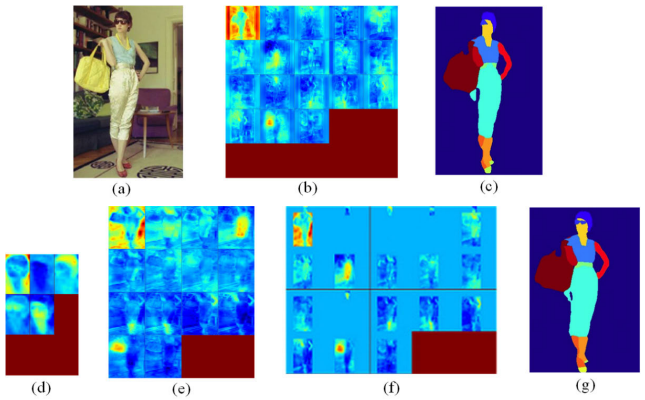



**FIGURE 3.** (a) Input image (b) image-level output score maps: The first score map is the background, and followed by hat, hair, sunglasses, upper-clothing, skirt, dress, belt, trousers, left-shoe, right-shoe, face, left-leg, right-leg, left-ram, right-arm, bag and scarf; (c) initial result of the image-level parsing network; (d) head-parsing sub-network output score maps: background, hat, hair, sunglasses and face; (e) body-parsing sub-network output score maps: background, upper-clothing, skirt, dress, belt, trousers, left-shoe, right-shoe, left-leg, right-leg, left-ram, right-arm, bag and scarf; (f) feature score maps after combination; and (g) final result of our method.

$$k = 0 \tag{3c}$$

$$z^{I}_{x^{B}_{start}+i,x^{B}_{xstop}+j,k} = z^{I}_{x^{B}_{start}+i,x^{B}_{xstop}+j,k} + z^{B}_{i,j,k+3}, \quad k = 1, \ldots, 8 \tag{3d}$$

$$z^{I}_{x^{B}_{start}+i,x^{B}_{xstop}+j,k} = z^{I}_{x^{B}_{start}+i,x^{B}_{xstop}+j,k} + z^{B}_{i,j,k+4}, \quad k = 9, \ldots, 13 \tag{3e}$$

Fig. 3 shows the comparison results of the before and after the combination of sub-networks. As shown in Fig. 3(b) and Fig. 3(c), the heat map of sunglasses is not very obvious, but after combination it becomes much more obvious (see Fig. 3(f)) because of the fusion of the head parsing sub-network (see Fig. 3(d)).

## IV. CONFIGURATION AND IMPLEMENTATION DETAILS

### A. DATA SET PARTITIONING AND PRE-PROCESSING

We trained the proposed PHPC on the ATR dataset [8]. This dataset contains 7702 images, where each image is paired with a ground-truth – a mask of pixels in 18 semantic labels. We split the available data into two sets (as shown in Fig. 4): the first set of 6898 images for training and the second set of 804 images for testing.

We augment the training data by randomly mirroring and cropping the images. The images are also normalised by subtracting the mean RGB value of all the training data. To balance computational efficiency and practicality (e.g., GPU memory), all images are resized to a resolution of 384x384 with 0 padding.

### B. TRAINING DETAILS

The final network models, namely the image-level parsing network (FCN), body and head sub-networks, were built based on caffe and trained with mini batch stochastic gradient descent with a momentum of 0.9, weight decay of 0.0005 and fixed learning rate of 0.0001. The setup for training PHPC networks is shown in Table 3.

To train the three sub-networks, we sped up the training processing by fine-tuning on parameters pre-trained on ImageNet dataset, and trained by phase, which is the same with fine-tune strategy in the work of Long *et al.* [9]. We trained the model on NVIDIA Tian X GPU for 30 epochs, and each epoch meant one pass of the full training set. We did not separate the images into batches for iterations. In other words, each iteration contained one single image in the training set.

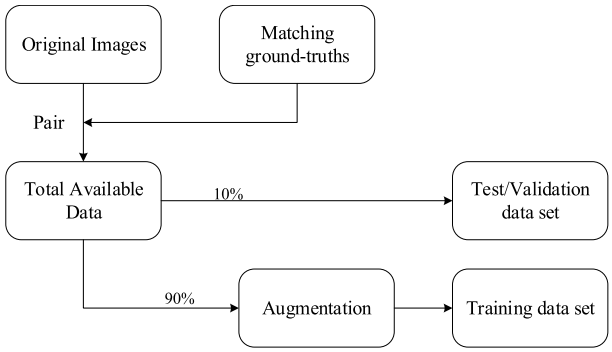



**FIGURE 4.** Method used for generating training and test/validation set.

**TABLE 2.** Dataset partitioning and preprocessing for PHPC.

| Dataset | ATR dataset [8] | No. of images [a] |
|---|---|---|
| | Training data (90%) | 6898 |
| | Test/validation data (10%) | 804 |
| | Total | 7702 |
| Pre-processing | - Subtracting mean RGB value of ImageNet training set<br>- Augmentation by mirror and fine cropping scheme [22]<br>- All resizing to 384×384 with 0 padding | 137960 |

**TABLE 3.** Training parameter setup for PHPC.

| Setup | Parameters/details |
|---|---|
| Weight initialisation | Parameters trained on ImageNet dataset |
| Optimiser | Stochastic gradient descent (SGD) |
| Momentum | 0.9 |
| Weight decay | 0.0005 |
| Learning rate | 0.0001 |
| Batch size | 1 image per batch |
| Number of epochs | 30 |

### C. EVALUATION METRICS

Evaluation metrics were defined to assess the performance of the network. We defined the Percentage of Correctly Localised Parts (PCP) metric to evaluate the location accuracy for part detection as follows:

$$\text{PCP} = \frac{\text{Number of correctly localised parts}}{\text{Number of all parts}} \tag{4}$$

where the detection part is correctly localised if and only if the overlap of the bounding box of the detection part and bounding box of the ground-truth is over 50%. More than 50% overlap is used because the input to sub-networks needs not to be very accurate and the sub-networks will parse the detection region again.

In addition, three metrics commonly used for evaluating semantic segmentation performance, including per-pixel accuracy, average per-class accuracy and mean accuracy on intersection over union region (IoU) are used in this study. The same metrics were also used in [9].

Let $n_{ij}$ be the number of pixels of class $i$ predicted as belonging to class $j$, and $n_{cl}$ be the number of classes. Therefore, $\sum_j n_{ij}$ is the total number of pixels in class $i$. Per-pixel accuracy is computed as follows:

$$Pixel\ acc = \sum_i n_{ii} / \sum_i \sum_j n_{ij} \tag{5}$$

Mean per-class accuracy is computed as follows:

$$\text{Mean } acc = (1/\text{n}_{cl}) \sum_i n_{ii} / \sum_j n_{ij} \tag{6}$$

Mean accuracy of IoU is computed as follows:

$$\text{mean IoU} = (1/\text{n}_{cl}) \sum_i n_{ii} / (\sum_j n_{ij} + \sum_j n_{ji} - n_{ii}) \tag{7}$$

We did not set aside a validation dataset in this study. Instead, we trained the models over a set of epochs by updating the networks weights. The network weights are saved after every epoch. Inference was carried out to yield pixelwise predictions for the test data using all models optimised during the training process. Fig. 5 shows the achieved mean IoU of the test data set over every epoch. The mean IoU accuracy appears to improve quickly during the first 5 epochs and becomes stable (converged) after 10 epochs.

### D. SUPERPIXEL AND CRFS REFINEMENT

As reviewed in section II, superpixel and CRF were reported as effective strategy to refine the segmentation results. We experiment to add superpixel and CRF as a post-processing step to the final results of PHPC, as shown in the dotted region of Fig. 1. To add superpixel and CRF refinement, we used the Simple Linear Iterative Clustering (SLIC) algorithm [20], a modified version of the K-means algorithm, to calculate superpixels in the input images. The SLIC algorithm segments image $I$ to a set of superpixels $R = \{R_1, R_2, \ldots, R_m\}$. Each superpixel $R_m$ is associated with the possible labels $L = \{l_1, l_2, \ldots, l_N\}$, $R_m \in L$. In the last section, we obtained 18 label probability score maps. Given image $I$, for every superpixel $R_m$, we computed the probability score of $R_m$ belong to label $l_k$ by averaging the feature score for all inner pixels of $R_m$ belong to label $l_k$, as follows:

$$P(R_m = l_k \,|I\,) = \frac{1}{c} \sum_{(i,j)} z_{i,j,l_k} \tag{8}$$

where $c$ is the number of pixels within superpixel $R_m$. We assigned superpixels to labels with the maximum probability score $P$.

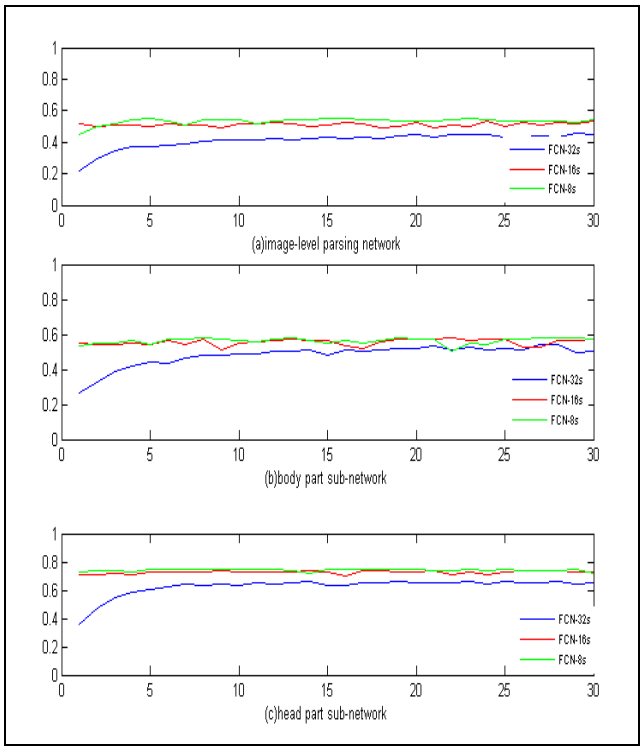



**FIGURE 5.** Evolution of the mean IoU accuracy of the test data set calculated over all training epochs.

Considering the relationship between neighbouring superpixels, we also adopted a CRF model based on the superpixels for better segmentation. Given an image $I$, our objective is to minimise the following energy function:

$$E(R, I) = \sum_m \psi_m(R_m = l_y \,|I\,) + \sum_{m,n} \psi_{m,n}(R_m = l_y, R_n = l_{y'} \,|I\,) \tag{9}$$

where $\psi_m$ is a unary energy involving the superpixel $R_m$, and $\psi_{i,j}$ is a pairwise energy involving a pair of superpixels $R_m$ and $R_n$, $l_y$ is the true label and $l_{y'}$ is the prediction label. The unary energy is compute by the following function:

$$\psi_m(R_m = l_y \,|I\,) = \frac{\exp(P(R_m = l_y \,|I\,))}{\sum_{k=1}^{N} \exp(P(R_m = l_k \,|I\,)} \tag{10}$$

The pairwise term $\psi_{m,n}$ models the similarity between two superpixels. We only adopted a pairwise term for adjacent superpixels and considered the similar appearance for adjacent superpixels. The pairwise term is defined as follows:

$$\psi_{m,n}(R_m = l_y, R_n = l_{y'} \,|I\,) = u(l_y, l_{y'}) k(f_m, f_n) \tag{11a}$$

$$k(f_m, f_n) = \omega \exp\left(-\frac{\|p(R_m) - p(R_n)\|^2}{\delta_c} - \frac{\|I(R_m) - I(R_n)\|^2}{\delta_t}\right) \tag{11b}$$

where $u(l_y, l_{y'}) = 1$ if $l_y \neq l_{y'}$ and otherwise $u(l_y, l_{y'}) = 0$. The appearance kernel is inspired by the observation that nearby pixels with similar colour are likely to have the same

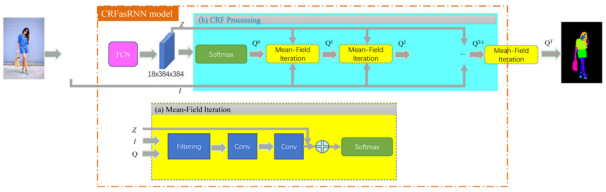



**FIGURE 6.** CRFasRNN network: with mean field iteration (a) within the CRF processing (b).

label. The degrees of nearness and similarity are controlled by the parameters $\omega$, $\delta_c$ and $\delta_t$.

We decided the most likely assignment $\hat{y} = \underset{R}{\text{argmax}}\, P(R\,|I\,)$, where $P(R\,|I\,) = E(R\,|I\,)$. We minimised the CRF energy (9) using alpha-expansion [21]. We trained the CRF model for post-progressing using Pystruct tool.

### E. COMPARATIVE STUDY

To evaluate the effectiveness of our proposed PHPC network, we conducted a comparative study. We compared the accuracy of PHPC (Fig. 2) with two other networks: (1) basic FCN-8s image-level parsing model (Fig. 2(a)) and the CRFasRNN network proposed by Zheng *et al.* [10], as shown in Fig. 6.

For comparative purposes, we describe the CRFastRNN model here. As shown in Fig. 6, CRFasRNN refines the coarse segmentation from the FCN by integrating CRF optimisation into the model for an end-to-end training.

Let $X_i$ be the variable associated with pixel $i$, where $X_i \in L$. Given an image I, the probability score of pixel $i$ being assigned to $l_i$ is initialised from the outputs of FCN, that is

$$P(X_i = l_i) = U_i(l_i) \tag{12}$$

where $U_i$ is the $i$-th feature map of the outputs from FCN. The best label assignment $l_i$ is obtained by minimising the total CRF energy function:

$$\sum\nolimits_i \psi_i(X_i = l_i\,|I\,) + \sum\nolimits_{i,j} \psi_{i,j}(X_i = l_i, X_j = l_j\,|I\,) \tag{13}$$

With reference to Equation (9), both energy functions (9) and (13) have the same formulation, including unary and pairwise energies. The key difference is that in our PHPC, CRF is based on the superpixel and modelled as a post-processing step (see section IV-D), while the CRFasRNN is based on every pixel and uses the mean-field approximation to minimise the CRF energy for an end-to-end model.

For pixel-level formulation, the computation of pairwise energy is very large. In view of this, mean-field inference is used to approximate the distribution of $P(X\,|I\,)$ as a simpler distribution $Q(X\,|I\,)$:

$$Q(X_i = l_i\,|I\,) = \frac{1}{z}\exp\left\{-\psi_i(X_i = l_i\,|I\,) - \sum\nolimits_{i=0}^{N} u(l_i, l_j) \times \sum\nolimits_{m=1}^{M} \omega^{(m)} \sum\nolimits_{i\neq j} k^{(m)}(f_i, f_j) Q(X_j = l_j)\,|I\right\} \tag{14}$$

The algorithm of the mean-field inference updating is shown below:

| | |
|---|---|
| Initialisation | $Q(X_i = l_i) \leftarrow \frac{\exp(P(X_i = l_i\|I\,))}{\sum_i \exp(P(X_i = l_i\|I\,))}$ |
| Do until converge | |
| Message passing | $\hat{Q}_m(X_i = l_i) \leftarrow \sum_{i\neq j} k(f_i, f_j)$ $Q(X_j = l_j\,\|I\,)$ for all $m$ |
| Weighting filter outputs | $\hat{Q}(X_i = l_i) \leftarrow \sum_{m=1}^{M} \omega^{(m)}$ $\hat{Q}_m(X_i = l_i)$ |
| Compatibility transform | $\hat{Q}(X_i = l_i) \leftarrow \sum_{i=1}^{N} u(l_i, l_j)$ $\hat{Q}(X_i = l_i)$ |
| Adding unary potentials | $\hat{Q}(X_i = l_i) \leftarrow \psi_i(X_i = l_i\,\|I\,)$ $-\hat{Q}(X_i = l_i)$ |
| Normalisation | $Q(X_i = l_i) \leftarrow \frac{\exp(\hat{Q}(X_i = l_i))}{\sum_i \exp(\hat{Q}(X_i = l_i))}$ |

The above algorithm shows that the updated equation of mean-field inference of a DenseCRF model can be broken into a series of small steps, as neural network operations. The message parsing step involves a bilateral filter, which can be viewed as convolutional. The weighting filter outputs and compatibility transform steps can be viewed as convolutions with $1 \times 1$ kernels. The adding unary potential step is a common operation in neural networks. The initialisation and normalisation steps are both equivalent to the softmax operation.

Fig. 6(a) illustrates the equivalent network layers of a mean-field iteration. By performing multiple mean-field iterations (Fig. 6(b)), where the output of one iteration becomes the input of the next iteration, the mean-field inference algorithm can be formulated as an RNN. Therefore, the model is called CRFasRNN network, and it is able to provide end-to-end training. We compared our PHPC networks with FCN-8s [9] and this CRFasRNN (Fig. 6), and the detail results are discussed in the next section.

## V. EXPERIMENTAL RESULTS AND DISCUSSIONS

We evaluate the proposed PHPC networks in this section. For comparative purposes, we also trained FCN-8s and CRFasRNN on the ART dataset using the same data partitioning and preprocessing, as outlined in section IV-A. We evaluated the overall effectiveness of PHPC networks in this section using the metrics defined in section IV-C.

The core idea of our PHPC lies in the detection of the head and body regions of the image, so we first evaluated the part detection performance on the test dataset in section V-A. In section V-B, we used the trained networks to inference pixelwise predictions on the test data. To do so, the weights determined in the training were first loaded into the network and then the inference was applied. Finally, the output prediction metrics were evaluated and compared. We evaluated the effectiveness of our PHPC without refinement and compared the results with the inference results of the FCN-8s

**TABLE 4. Localisation accuracy PCP of the head and body parts.**

| Parts \ Models | FCN-8s | PHPC |
|---|---|---|
| Head | 99.50% | 99.50% |
| Body | 91.79% | 93.16% |

**TABLE 5. Comparison of overall accuracy achieved (%) by our PHPC model and two state-of-the-arts paring models when testing on ATR dataset.**

| Models \ Metrics | Pixel acc Eq. (5) | Mean acc Eq. (6) | Mean IoU Eq. (7) |
|---|---|---|---|
| FCN-8s | 92.37 | 67.82 | 55.35 |
| CRFasRNN | 92.38 | 68.01 | 56.44 |
| PHPC | 93.32 | 71.03 | 59.93 |

and CRFastRNN networks. We discuss the effectiveness of superpixel and CRF refinement in section V-C.

### A. EVALUATION OF HEAD AND BODY PART DETECTIONS

We calculated the PCP for all the test data using Equation (4). Table 4 shows the localisation accuracy for the head and body parts of our PHPC model and that of FCN-8s. As shown, the head localisation accuracy of FCN-8s and our PHPC is close to 100% while body part localisation accuracy of PHPC is higher 1.37% than FCN-8s for test data. This illustrates that our PHPC method is effective in localising/ detecting the head and body regions of the input images.

### B. EVALUATION OF HUMAN PARSING RESULTS

The per-pixel accuracy (pixel acc), per-class accuracy (mean acc) and mean accuracy on intersection over union region (mean IoU) are calculated by inference the trained PHPC model on the test data and listed in Table 5, which also compares the metrics of the FCN-8s and CRFasCNN models. As shown, the proposed PHPC model achieves the best results in all performance metrics. In comparison to FCN-8s and CRFasRNN models, the mean IoU of our PHPC has improved by 4.58% and 3.49%, respectively, and the mean acc has improved by 3.21% and 3.02%, respectively.

Table 6 also shows per-class mean accuracy on intersection over union region (mean IoU) comparison of our PHOC model, FCN-8s and CRFasCNN. It is obvious that our PHPC significantly improves the IU score in every category, as compared to FCN-8s and CRFasRNN. In particularly, the accuracy of small items improves significantly. For example, the accuracy of sun-glasses is 11.433% greater than that of FCN, and 7.278% greater than that of CRFasRNN. For another example, the accuracy of belt is 3.430% greater than that of FCN-8s and 3.416% greater than that of CRFasRNN. Also, compared to the CRFasRNN, the accuracy of scarf has improved by 5.938%. The reason why our model performs better than FCN-8s and CRFasRNN is that our model strengthens the attention of head and body part, and combines the parsing result of subnetworks (focusing attention to local parts) into the image-level parsing result.

**TABLE 6. Comparison of per-class mean IoU accuracy (%) between proposed PHPC model and two state-of-the-arts models FCN-8s and CRFastRNN.**

| Models | Hat $k$=1 | Hair $k$=2 | S-gls $k$=3 | U-cloth $k$=4 | Skirt $k$=5 |
|---|---|---|---|---|---|
| FCN-8s | 57.337 | 71.435 | 41.557 | 65.431 | 47.204 |
| CRFasRNN | 61.473 | 72.284 | 45.712 | 68.268 | 46.932 |
| PHPC | **66.232** | **75.271** | **52.990** | **69.028** | **52.458** |
| Models | Trousers $k$=6 | Dress $k$=7 | Belt $k$=8 | L-shoe $k$=9 | R-shoe $k$=10 |
| FCN-8s | 59.761 | 50.513 | 21.829 | 45.595 | 43.804 |
| CRFasRNN | 63.015 | 50.073 | 21.843 | 46.923 | 45.592 |
| PHPC | **65.808** | **52.041** | **25.259** | **48.644** | **48.834** |
| Models | Face $k$=11 | L-leg $k$=12 | R-leg $k$=13 | L-arm $k$=14 | R-arm $k$=15 |
| FCN-8s | 74.925 | 61.396 | 59.780 | 62.371 | 60.678 |
| CRFasRNN | 75.619 | 63.886 | 60.858 | 64.310 | 62.166 |
| PHPC | **79.305** | **65.293** | **64.633** | **68.038** | **66.250** |
| Models | Bag $k$=16 | Scarf $k$=17 | Bkg $k$=0 | | |
| FCN-8s | 53.187 | 23.257 | 96.349 | | |
| CRFasRNN | 52.252 | 18.213 | 96.629 | | |
| PHPC | **57.873** | **24.151** | **96.651** | | |

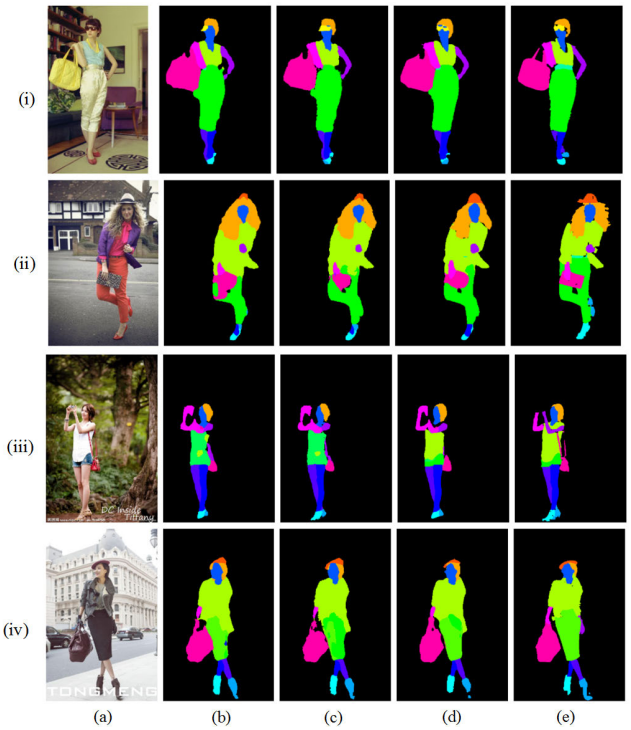



**FIGURE 7. Comparison of parsing results: (a) input images, (b) results from FCN-8s, (c) results from CRFasRNN, (d) results of our PHPC networks, and (e) ground-truths.**

**TABLE 7. Comparison of average running time per image.**

| Models | Average running time per image |
|---|---|
| FCN-8s | 296.91ms |
| CRFasRNN | 854.32ms |
| PHPC | 4.98s |

* System configuration: PC with Inter(R) Core (TM) i7-6700K, 32G RAM.

Fig. 7 shows some parsing results generated by the different models. As shown in Fig. 7 row (i), both FCN-8s

**TABLE 8. Comparison of overall accuracy (%) for PHPC models with and without refinements.**

| Models \ Metrics | Pixel acc Eq. (5) | Mean acc Eq. (6) | Mean IoU Eq. (7) |
|---|---|---|---|
| CRFasRNN | 92.38 | 68.01 | 56.44 |
| PHPC w/o refinement | 93.32 | 71.03 | 59.93 |
| PHPC refined w/Superpixel | 93.44 | 70.30 | 59.27 |
| PHPC refined w/ Superpixel and CRF | 92.83 | 68.01 | 562.45 |

**TABLE 9. Comparison of per-class mean IoU accuracy (%) for models with and without refinements.**

| Models | Hat $k$=1 | Hair $k$=2 | S-gls $k$=3 | U-cloth $k$=4 |
|---|---|---|---|---|
| CRFasRNN | 61.473 | 72.284 | 45.712 | 68.268 |
| PHPC w/o refinement | 66.232 | **75.271** | **52.990** | 69.028 |
| PHPC refined w/Superpixel | **66.831** | 73.711 | 44.835 | 69.890 |
| PHPC refined w/ Superpixel and CRF | 64.135 | 72.591 | 14.371 | **70.025** |
| Models | Skirt $k$=5 | Trousers $k$=6 | Dress $k$=7 | Belt $k$=8 |
| CRFasRNN | 46.932 | 63.015 | 50.073 | 21.843 |
| PHPC w/o refinement | 52.458 | 65.808 | 52.041 | **25.259** |
| PHPC refined w/Superpixel | **52.760** | **66.721** | 52.426 | 24.811 |
| PHPC refined w/ Superpixel and CRF | 53.313 | 65.895 | **52.488** | 6.130 |
| Models | L-shoe $k$=9 | R-shoe $k$=10 | Face $k$=11 | L-leg $k$=12 |
| CRFasRNN | 46.923 | 45.592 | 75.619 | 63.886 |
| PHPC w/o refinement | **48.644** | **48.834** | 79.305 | **65.293** |
| PHPC refined w/Superpixel | 47.943 | 48.309 | **76.389** | 64.216 |
| PHPC refined w/ Superpixel and CRF | 45.622 | 47.046 | 74.619 | 62.607 |
| Models | R-leg $k$=13 | L-arm $k$=14 | R-arm $k$=15 | Bag $k$=16 |
| CRFasRNN | 60.858 | 64.310 | 62.166 | 52.252 |
| PHPC w/o refinement | **64.633** | **68.038** | 66.250 | 57.873 |
| PHPC refined w/Superpixel | 64.047 | 67.635 | **66.405** | **58.772** |
| PHPC refined w/ Superpixel and CRF | 62.293 | 63.971 | 63.226 | 58.562 |
| Models | Scarf $k$=17 | Bkg $k$=0 | | |
| CRFasRNN | 18.213 | 96.629 | | |
| PHPC w/o refinement | 24.151 | 96.651 | | |
| PHPC refined w/Superpixel | **24.162** | 96.941 | | |
| PHPC refined w/ Superpixel and CRF | 20.853 | **97.179** | | |

and CRFasRNN do not detect the sunglasses, but our PHPC model parses them accurately. Fig. 7 row (ii) also shows that our PHPC model performs better on parsing the belt than FCN-8s and CRFasRNN. Fig. 7 row (iii) shows that FCN-8s and CRFasRNN confuse upper-clothing and dress, but our PHPC model segments the upper-clothing and trousers properly. In Fig. 7 row (iv), the skirt was not accurately parsed by FCN-8s and CRFasRNN models, but our PHPC model parsed the skirt well.

We also compare the speed of the three networks in Table 7. As shown, compared with FCN-8s and CRFasRNN, the running time needed to parse an image using our PHPC networks is much longer. This is because our PHPC networks contain three networks in total and these networks are arranged as a cascade and process the input image in turn. The output

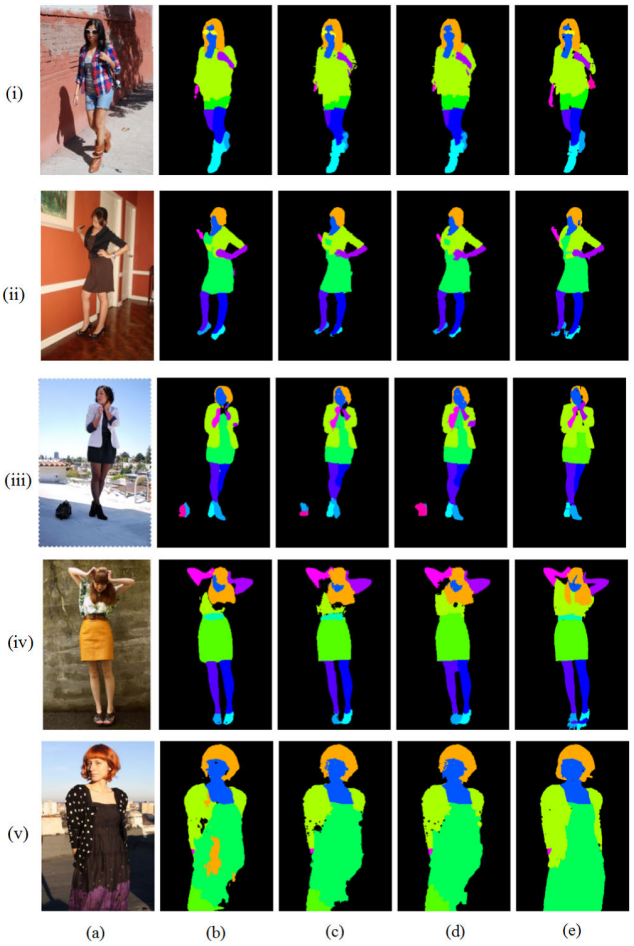



**FIGURE 8. Comparisons of parsing results with and without refinement: (a) input images, (b) results of PHPC without refinement, (c) results of PHPC with superpixel refinement, (d) results of PHPC with superpixel and CRF refinement, and (e) ground-truths.**

of the image-level parsing network is used to generate the input of the head- and body-parsing sub-networks, which substantially extends the running time needed.

### C. EVALUATION OF HUMAN PARSING RESULTS

As discussed in section IV-D, we experimented on using superpixel and CRF as a post-processing step to further refine the results. A comparison of the overall accuracy of our PHPC models with and without superpixel and/or CRF refinements is shown in Table 8. Although the literature reported that superpixel and CRF could improve the segmentation accuracy [1], [2], [10] our results do not show improvements in accuracy. Instead, the overall pixel accuracy, overall per-class accuracy and mean IoU accuracy all dropped after refinement process, while the CRF refinement results in a bigger drop than only using superpixel for refinement. Therefore, for PHPC model, we think superpixel and CRF as post-processing is not effective, so such refinement (dotted region) are not suggested in Fig. 2.

By looking at the per-class mean IoU comparison in Table 9, we can see that the accuracy of big items, such as upper-clothing, trousers and dress increased, but the accuracy for small items decreased. This demonstrates that the proposed superpixel and/or CRF refinement is only effective for big items. The accuracy of small items may have decreased because that the superpixel and CRF model we used is only a post-processing step (as discussed in section IV-D), not an end-to-end model like CRFasRNN. Fig. 8 shows some of the parsing results of the models with and without refinement. The post-processing step may incorrectly update or group small item pixels into the groups of large items.

To conclude, the CRF refinement based on superpixel as a post-processing step for the final results in the current model, it appears that this refinement does not improve the parsing accuracy for small items because the CRF model was not integrated or trained end-to-end. We will improve the refinement model and integrate all of the components of our method as a unified network in our future work.

## VI. CONCLUSION

In this paper, we proposed a novel part-based human parsing cascade (PHPC) of networks for the parsing of human images, which consisted of an image-level parsing network, two part-based paring networks and a combination module. The inputs to the part-based parsing networks are partial images detected based on the results from the image-level parsing network and then scaled up to double the original size. We have demonstrated this network design is beneficial for extracting more detailed features from input images. By doubling the partial raw image in localised area of interests, the part-based parsing sub-networks decreased the impact of the background and improved the parsing accuracy of small items. The experimental results have demonstrated the effectiveness of the proposed PHPC networks.

The proposed PHPC networks do, however, have some known limitations. First, the inputs for the part-based parsing sub-networks rely on the output of image-level parsing network: the image-level and part sub-networks do not share any parameters. As a result, the running speed of the PHPC is slow in comparison to other end-to-end parsing models.

## REFERENCES

[1] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3570–3577.

[2] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 3519–3526.

[3] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. CVPR*, Jun. 2011, pp. 1705–1712.

[4] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 3408–3415.

[5] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-CNN meets KNN: Quasi-parametric human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1419–1427.

[6] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1386–1394.

[7] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3185–3193.

[8] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1529–1537.

[11] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph LSTM," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 125–143.

[12] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3376–3385.

[13] L.-C. Chen, Y. Yang, J. Wang, X. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3640–3649.

[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 447–456.

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: https://arxiv.org/abs/1511.07122

[16] P. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1–9.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[18] A. Arnab, S. Jayasumana, S. Zheng, and P. H. S. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 524–540.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[21] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.

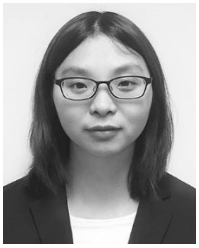
[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[23] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, "Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 648–663.

[24] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 770–785.

[25] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 4814–4821.

[26] J. Hu, Z. Sun, Y. Sun, and J. Shi, "Progressive refinement: A method of coarse-to-fine image parsing using stacked network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.



**YANGHONG ZHOU** received the bachelor's degree from Fujian Normal University, in 2011, and the master's degree from the University of Electronic Science and Technology of China, in 2014. She is currently pursuing the Ph.D. degree with The Hong Kong Polytechnic University. Her research interests include deep learning and image analysis.

**P. Y. MOK** received the B.Eng. degree (Hons.) majoring in industrial and manufacturing systems engineering and the Ph.D. degree from the University of Hong Kong, in 1998 and 2002, respectively. She is currently an Associate Professor with The Hong Kong Polytechnic University. Her current research interests include fashion pattern engineering, fashion 2D and 3D CAD, digital human modeling, 3D scanning and sizing, cloth simulation, deep learning, computer generated textile, sketch and pattern designs, computer vision and computer graphics in fashion applications, advanced data analysis, and artificial intelligent applications in the fashion industry.

**SHIJIE ZHOU** received the bachelor's degree from the Lanzhou University of Technology, in 1995, and the Ph.D. degree from the University of Electronic Science and Technology of China, in 2004. He is currently a Professor with the University of Electronic Science and Technology of China. His research interests include network security, traffic simulation, and artificial intelligence.