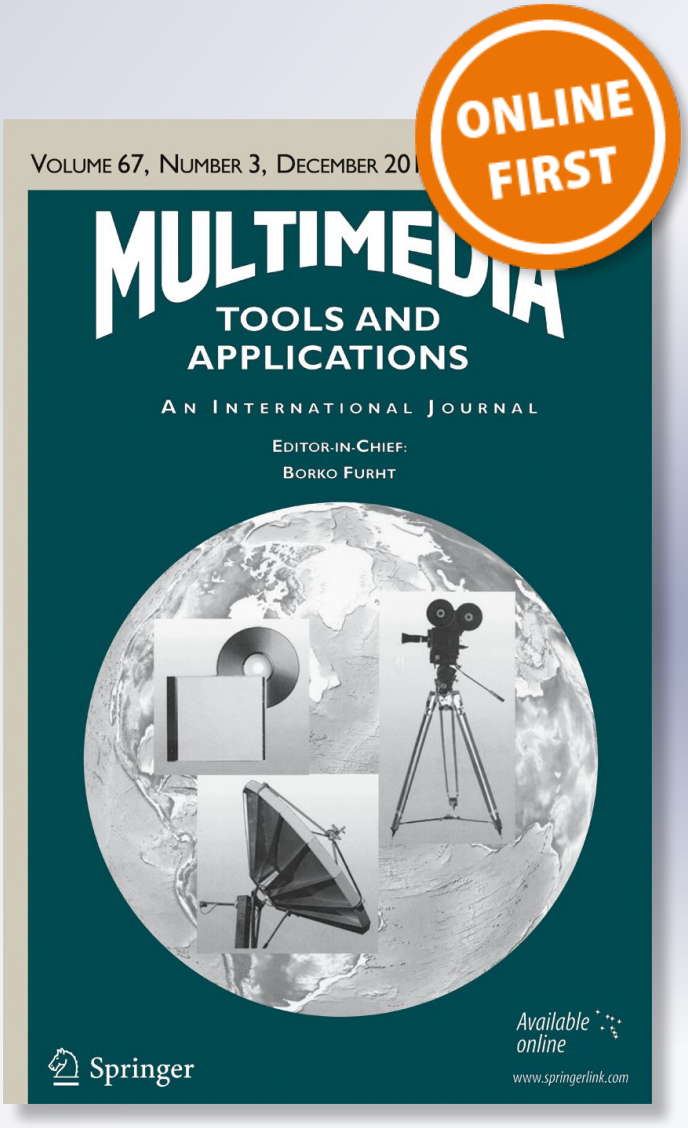
Human parsing by weak structural label

Zhiyong Chen, Si Liu, Yanlong Zhai, Jia Lin, Xiaochun Cao & Liang Yang

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-017-5368-4



Springer

Multimed Tools Appl
https://doi.org/10.1007/s11042-017-5368-4

Human parsing by weak structural label

Zhiyong Chen[1] · Si Liu[2] · Yanlong Zhai[3] · Jia Lin[4] · Xiaochun Cao[2] · Liang Yang[5]

Received: 2 January 2017 / Revised: 1 September 2017 / Accepted: 30 October 2017

Abstract Human parsing, which decomposes a human centric image into several semantic labels, e.g., face, skin etc, is an active topic in recent years. Traditional human parsing methods are always conducted on a supervised setting, i.e., the pixel-wise labels are available during the training process, which require tedious human labeling efforts. In this paper,

✉ Yanlong Zhai
ylzhai@bit.edu.cn

Zhiyong Chen
changingivan@gmail.com

Si Liu
liusi@iie.ac.cn

Jia Lin
linjia1@jd.com

Xiaochun Cao
caoxiaochun@iie.ac.cn

Liang Yang
yangliang@vip.qq.com

[1] School of Information Science and Engineering, Lanzhou University, No. 222, TianShui south Road, ChengGuan District, Lanzhou, Gansu, China

[2] State Key Laboratory of Information Security, Institute of Information Engineering, CAS, No. 89, Minzhuang Road, Haidian District, Beijing, China

[3] School of Computer Science and Technology, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing, China

[4] JD.com, North Star Century Center 10 floor, Beijng, China

[5] School of Information Engineering, Tianjin University of Commerce, No. 409 Guangrong Road, Beichen District, Tianjin, China

Published online: 25 November 2017

we propose a weakly supervised deep parsing method to alleviate the human from the time-consuming labeling. More specifically, we resort to train a robust human parser with the structural image-level labels, e.g., "red jeans" etc. The structural label contains an attribute, e.g., "red", as well as a class label, e.g., "jeans". Our framework is based on the Fully Convolution Network (FCN) (Pathak et al. 2014) with two critical differences. First, the loss function defined on the pixel by FCN (Pathak et al. 2014) is modified to the image-level loss by aggregating the pixel-wise prediction of the whole image into a multiple instance learning manner. Besides, we develop a novel logistic pooling layer to constrain that the pixels responding to the color and corresponding category labels are the same to interpret the structural label. Extensive experiments in the publicly available dataset (Liu et al. IEEE Trans Multimedia 16(1):253–265, 2014) show the effectiveness of the proposed method.

Keywords Human parsing · Deep learning

1 Introduction

Human parsing, which aims to segment the human image into multiple components, i.e., face, hair and bags etc., has attracted a lot of attention these years. It can be used for many real applications such as clothes recommendation [14, 35], clothes retrieval [1, 3, 15, 34], visual relation detection [36], and online virtual fitting room. Many efforts have been made to improve the performance during the past few years. However, the performance of human parsing is not satisfactory for many reasons. Firstly, hand-crafted image feature cannot fully describe the image well. Secondly, many existing algorithms need numerous training images with pixel-wise labels such as fully convolutional network [20], hyper-column based image segmentation [8]. In fact, obtaining pixel-wise labels is very difficult and expensive. When the training data is insufficient, the model may overfit the data. Thus, how to use the limited pixel-wise label images to learn semantic features and clothes model is the key to improve the performance. Although, the pixel-wise images are limited, images with attributes are very common on website with the popularity of social networks and online shopping, such as chictopia.[1] For example, on many clothes shopping websites, sellers provide many information about each piece of clothing, such as the clothing style, color and material [15]. Besides, in order to enhance the user experience, many real photos of clothing are provided as well. These photos can be regarded as images with the same attributes of the clothing. As a result, making full use of the structural attributes becomes the most practical solutions to improve the human parsing results. A lot of algorithms are designed to handle the weakly supervised image parsing problem [22, 23]. But all of them treat each attribute individually. By analyzing the image-level tag in many fashion websites, we observe that there are usually multiple components in the structural label. For example, the label "navy dress" contains a color label "navy" and a category label "dress". Both the color component and the category component of the label can be used as the weakly supervised information to enhance the original pixel-wise label based human parsing. However, since both color and category label components describe the same object in the images, just as Fig. 1 shows, how to make use of the relationships between the two components, is the key to solve the problem. In this paper, we target at solving the human parsing problem using the multiple weakly supervised structural information. We have the publicly available

[1] http://www.chictopia.com/

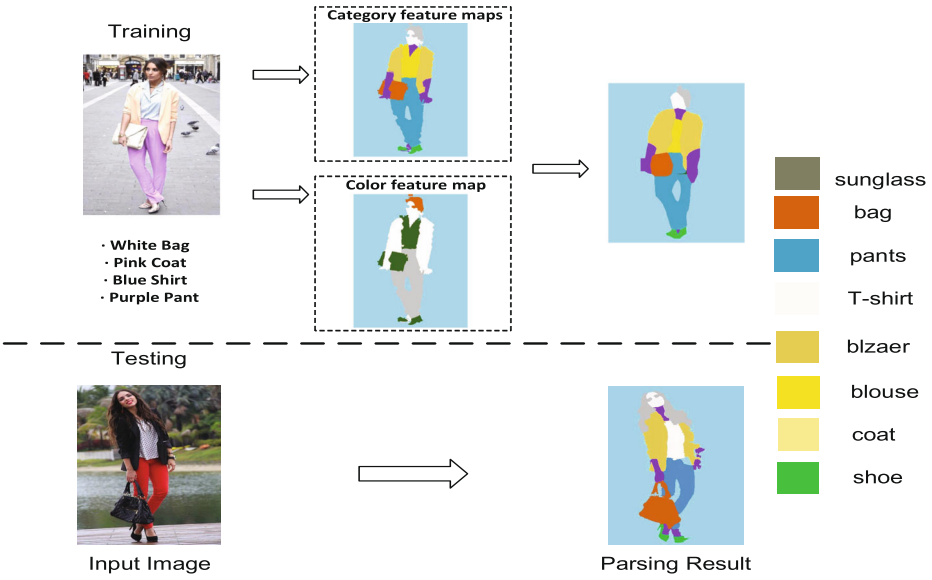


Fig. 1 For one image, the feature map for one color and the corresponding category should be turned on at the same location

dataset [16] to valid the effectiveness of our method. Recently, deep learning has achieved great success in many image semantic understanding tasks [2, 6, 7, 10, 12, 25–27, 29, 30, 37, 38], which motivates us to develop a deep framework as shown in Fig. 2. by feeding the images into the VGG16 [26] network to extract very deep features. The fully connected layers are converted to convolution layers as [23], and we produce both the color (13 dimensions for each pixel) and category (23 dimensions for each pixel) response maps and obtain the final response maps by logically combining the values of them at each pixel location. Finally, we use the Multi-Instance Learning (MIL) loss computed at maximum predictions as the loss function to be minimized. The contributions of this paper can be summarized as follows:

- We propose a weakly- and semi-supervised deep human parsing framework to learn from both pixel-level and image-level structural labels and predict each pixel's label in any testing image.
- The deep framework contains a logistic pooling layer to fully explore the relationships between the color and category components inside the structure label. More specifically, we constrain that the responses of the color and category should be turned on/off in the same location.
- Our proposed deep framework contains a label aggregation layer which can summarize the pixel-wise label to image label in a MIL fashion. The MIL loss is designed upon the aggregated layers.

The rest of the paper is organized as follows. We review the most related work in Section 2. Then the details of our method are illustrated in Section 3. Section 4 will validate the effectiveness of the proposed method. Section 5 summarizes our method.

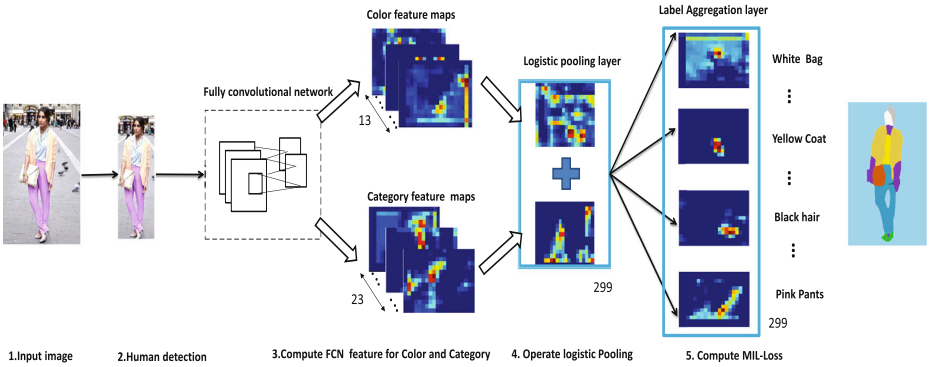


Fig. 2 The proposed deep weakly supervised human parsing structure overview. Our method (1) take an input image, (2) use FASTER-RCNN to detect the human (3) go forward through FCN to compute both color and category feature maps, (4) compute the logistic value of color and category at the same location for each combination, (5) compute MIL-loss between groundtruth color-category labels and predicted structured labels

2 Related work

2.1 Human parsing

Human parsing has attracted much attention these years. A lot of work have been proposed. Active Template Regression [13] expresses the normalized mask of each label as the linear combination of the learned mask templates, and then morphed to a more precise mask with the active shape parameters estimated by CNN, including position, scale and visibility of each semantic region. Matching-CNN [18] is a quasi-parametric deep human parsing model to predict the matching confidence and displacements of the best matched region in the testing image for a particular semantic region in one KNN image. Deng et al. [5] present a approach to infer the attribute by extracting the foreground segments of pedestrian through deep learning-based parsing. Luo et al. [21] propose a new Deep Decompositional Network (DDN) to directly maps low-level visual features to the label maps of body parts with DDN, which is able to accurately estimate complex pose variations with good robustness to occlusions and background clutters. Liu et al. [19] use only one labeled frame per video as supervision information, as well as estimated optical flow between frames to obtain parsing result in surveillance video. Yang et al. [33] propose a data-driven framework of clothing co-parsing, in order to jointly parse a set of clothing images (unsegmented but annotated with tags) into semantic configurations. To solve the data limitation problem, Liu et al. [17] propose a semi-supervised learning strategy to harness the rich contexts in those easily available web videos to boost any existing human parser. However, these methods all rely on the extensively labeled dataset, which prohibit the large-scale training. To the contrary, our method can effectively make full use of the big amount of weakly supervised data.

2.2 Weakly supervised semantic segmentation

To solve the weakly supervised problem, Pathak et al. [23] propose a MIL formulation of multi-class semantic segmentation learning by a fully convolutional network. They seek to learn a semantic segmentation model from just weak image-level labels. Papandreou et al.

[22] study the problem of learning DCNNs for semantic image segmentation from either weakly annotated training data such as bounding boxes, image-level labels or a combination of few strongly labeled and many weakly labeled images, sourced from one or multiple datasets. They develop Expectation-Maximization (EM) methods for semantic image segmentation model training under these weakly supervised and semi-supervised settings. Liu et al. [16] propose a weakly supervised human parsing methods with weak supervision from the user-generated color-category tags such as "red jean" and "white T-shirt". They propose to combine the human pose estimation module, the MRF-based color and category inference module and the (super) pixel-level category classifier learning module to generate multiple well-performing category classifiers, which can be directly applied to parse the fashion items in the images. Recently, Hong et al. [9] proposes a deep neural network which decouples classification and segmentation, and learns a separate network for each task. In this architecture, labels associated with an image are identified by classification network, and binary segmentation is subsequently performed for each identified label in segmentation network. However, they are either designed upon the hand-crafted features or cannot directly applied to handle the structural model.

3 Methods

In this section, we sequentially introduce the modules used in our methods. Firstly, human detection is used to extract the human regions (Section 3.1). Then the clothes region is fed into our framework with multiple layers, including convolutional layers(Section 3.2), logistic pooling layers (Section 3.3) and label aggregation layers (Section 3.4).

3.1 Human detection

Before training the Color-Category Network, we first train a human detector to crop the target person from the image. We follow the state-of-the-art Faster R-CNN [25] as its simple and efficient. Faster R-CNN outputs four-dimension bounding boxes of every object class and their corresponding confidence. Here, the object classes only contains two classes, namely human and background. In our implementation, we use the default configuration of Faster R-CNN and fine-tune the detector on the pre-trained VGG16 network by the same training data in the parsing task below. The final detection average precision in test set is 0.91, and it's sufficient for the further processing. Then we apply the detector to obtain human bounding boxes. As we care more about the detection precision than recall, we only keep those bounding boxes according to object confidence larger than 0.9. In this way, most false negative are removed. Several exemplars are shown in Fig. 3.

3.2 Convolutional layers

The convolution operation is expressed as

$$y^j = \max(0, b^j + \sum_i k^{ij} * x^i) \tag{1}$$

where x^i and y^j are the i-th input map and the j-th output map, respectively. k^{ij} is the convolution kernel between the i-th input map and the j-th output map. $*$ denotes convolution. b^j is the bias of the j-th output map. We use ReLU nonlinearity for hidden neurons, which is

Fig. 3 Example of four images and the human detection results of Faster R-CNN. The first row are the original images, and the second row are the corresponding results

shown to have better fitting abilities than the sigmoid function [27]. We have convolutional layers with increasingly larger receptive field.

3.3 Logistic pooling layers

In this section, we propose a Logistic Pooling Layer to combine color and category. For one structured label, such as "red jeans", the final parsing result should have consistent response at the same location both for color and category. Compared with using category information only, utilizing multiple information could help improve the performance. As we describe above, in order to enforce the color and the corresponding category labels fire at the same position, we define this layer by:

$$Score(x, y, s) = S_{color}(x, y, r) \otimes S_{category}(x, y, l) \tag{2}$$

where s denotes the structured channel, such as "red jeans", r is the color channel, such as "red", and l is the category channel, such as "jeans". The number of structural channels is the multiplication of the number of colors with the number of categories. In our implementation, we use two Fully Convolutional Networks [20] to obtain S_{color} and $S_{category}$ respectively. $\otimes$ denotes the logistic operation between these two scores, in experiments we use add operation. Similar with Siamese network [4], both FCN share parameters in previous convolutional layers and initial input images. The output of FCN is color or category feature maps with same size as initial input image. Note that, unlimited to color and category, our Logistic Pooling Layers could be easily expanded to cover more information. These multiple information always describe the different attributes for the same object. For example, one dress may have various value on color, length, style, category, and so on, such as "a red long tight sweater".

3.4 Label aggregation layers

The output of the logistic pooling layers are the response of color and category. Since the labels are defined in the image-level, thus we define a multi-instance loss as the MIL loss

Table 1 Comparison among PaperDoll, FCN, FCN+category, FCN+structure in testing set

Method	Accuracy	Avg.precision	Avg.recall	Avg. F-1 score
PaperDoll	0.847	0.359	0.382	0.341
CCNN	0.807	0.419	0.471	0.364
FCN	0.891	**0.54**	0.448	0.467
FCN+category	**0.892**	0.531	0.469	0.478
FCN+structure	0.889	0.505	**0.511**	**0.483**

computed at maximum predictions. We identify the max scoring pixel in the coarse heat-maps of classes present in image and background. The background class is analogous to the negative instances by competing against the positive object classes. Let the input image be I, its label set be L_I and $\hat{p}_l(x, y)$ be the output heat-map for the l-th label at location (x, y). The loss is defined as:

$$(x_l, y_l) = \arg\max_{\forall(x,y)} \hat{p}_l(x, y) \quad \forall l \in L_I \tag{3}$$

$$MILLOSS = \frac{-1}{|L_I|} \sum_{l \in L_I} \log \hat{p}_l(x_l, y_l) \tag{4}$$

Simultaneous training exploits the context among different labels to help refine the parsing accuracy. At inference phase, the MIL-FCN takes the top class prediction at every point in the coarse prediction and bilinearly interpolates into image resolution to obtain a pixel-wise segmentation. In order to constrain pixel-score into the scope of [0, 1] , we normalize output score according to channel axis just like softmax layer:

$$Score_l(x, y) = \frac{exp(Score_l(x, y))}{\sum_{l=1}^{L} exp(Score_l(x, y))} \quad \forall l \in L_I \tag{5}$$

Note that, only those channels appeared in L_I will take part in the normalization. Correspondingly, in back propagation process we calculate gradient only for those channels, which achieves the max score, we calculate the gradient by:

$$Diff = \frac{exp(Score_l(x, y)) * (A - exp(Score_l(x, y))}{A^2} = Score_l * (1 - Score_l) \quad \forall l \in L_I \tag{6}$$

where $A = \sum_{l=1}^{L} exp(Score_l(x, y))$. For channels, which do not contain the current max score, we calculate its gradient by:

$$Diff = \frac{-A * b}{A^2} = -Score_l * Score_{max-l} \quad \forall l \in L_I \tag{7}$$

where $b = exp(Score_{max-l}(x, y))$, as these elements contribute to the normalization, these should be also updated in each iteration.

Table 2 Comparison among PaperDoll, FCN, FCN+category, FCN+structure in testing set

Category	T-shirt	bag	belt	blazer	blouse	coat
PaperDoll	0.199	0.572	0.169	0.196	0.197	0.226
Fashion-Parsing	0.426	0.233	0.257	**0.497**	0.440	**0.473**
CCNN	0.310	0.504	0.114	0.407	**0.451**	0.317
FCN	0.451	0.662	0.224	0.390	0.445	0.255
FCN+category	0.452	**0.667**	0.278	0.393	0.445	0.265
FCN+structure	**0.453**	0.663	**0.281**	0.391	0.437	0.274
Category	dress	face	hair	hat	jeans	legging
PaperDoll	0.210	0.651	0.655	0.472	0.160	0.158
Fashion-Parsing	0.602	0.252	0.402	0.273	**0.623**	**0.641**
CCNN	**0.659**	0.333	0.268	0.161	0.532	0.342
FCN	0.482	0.705	0.702	0.491	0.460	0.199
FCN+category	0.483	**0.707**	**0.708**	0.525	0.465	0.229
FCN+structure	0.465	0.689	0.694	**0.538**	0.473	0.238
Category	pants	scarf	shoe	shorts	skin	skirt
PaperDoll	0.253	0.226	0.517	0.457	0.654	0.285
Fashion-Parsing	0.587	0.318	0.290	0.525	0.367	0.593
CCNN	**0.721**	**0.427**	0.313	0.331	0.376	0.614
FCN	0.513	0.226	0.600	0.679	0.714	0.651
FCN+category	0.517	0.240	0.608	**0.684**	**0.716**	0.655
FCN+structure	0.514	0.308	**0.615**	0.679	**0.716**	**0.657**
Category	socks	stocking	glass	sweater	bk	mean
PaperDoll	0.019	0.256	0.245	0.104	0.970	0.341
Fashion-Parsing	**0.181**	0.496	0.110	0.511	0.676	0.421
CCNN	0.129	0.479	0.044	**0.529**	**0.987**	0.419
FCN	0.029	0.508	0.318	0.061	0.973	0.467
FCN+category	0.037	0.509	**0.367**	0.078	0.973	0.478
FCN+structure	0.085	**0.537**	0.316	0.116	0.974	**0.483**

4 Experiments

4.1 Experiment setting

We test the performance in the publicly available dataset Colorful-Fashion [16], which contains all 2,682 images are labeled with pixel-level color-category labels. All images have good visibility of the full body. There are 13 colors by referring to color naming research [28]. Note that the dataset contains both solid color clothes and multiple-color clothes. There are 23 categories of tags. We use the same metric as PaperDoll [16, 31, 32] to evaluate the performance, including several standard metrics: accuracy, average precision, average recall and average F-1 over pixels. The Colorful-Fashion dataset is divided into three sets randomly: one training set, which contains 300 images, training initial FCN network. One training set for training Label Aggregation layers and Logistic Pooling Layers, which

Fig. 4 Parsing results of 16 example images in Testing set: initial image, ground truth and our parsing result

contains 1,488 images only armed with image-level label information. And the left 894 images are given testing set.

4.2 Implementation details

We train and test our models based on Caffe [11] on a single NVIDIA TITAN-X. The base learning rate is set as 1e-7 and fixed through all training process. We use SGD to optimize our model with momentum of 0.9, weight decay of 0.0005. The batch size is set as 1, with the limitation of GPU memory. Note that, as Fully Convolutional Network could cover any size images, we do not need to resize input image, which could keep the spatial information of initial image. The forward of the model is efficient. It costs about 0.1 second to predict the category labels for an image (Fig. 4).

4.3 Quantitative fashion parsing results

We compare PaperDoll, FCN, FCN+category, CCNN [24], Fashion-Parsing [16], FCN+structure performance. FCN+category means we use MIL-loss on the category label only to train the parsing network. FCN+structure means we both use color and category information with MIL-loss to predict pixel-level labels. We report the comparison results among these methods in Table 1 on test set among overall metrics respectively. The comparison of F-1 value of all 23 categories and their mean value are showed in Table 2. Note that bold entries mean the best performance among comparison methods for special metric (Table 1) or for special category (Table 2). From the results we can see that FCN+structure could obtain better performance on most categories. This is because color information will help category pixel-level classification task when they locate at the same position, such

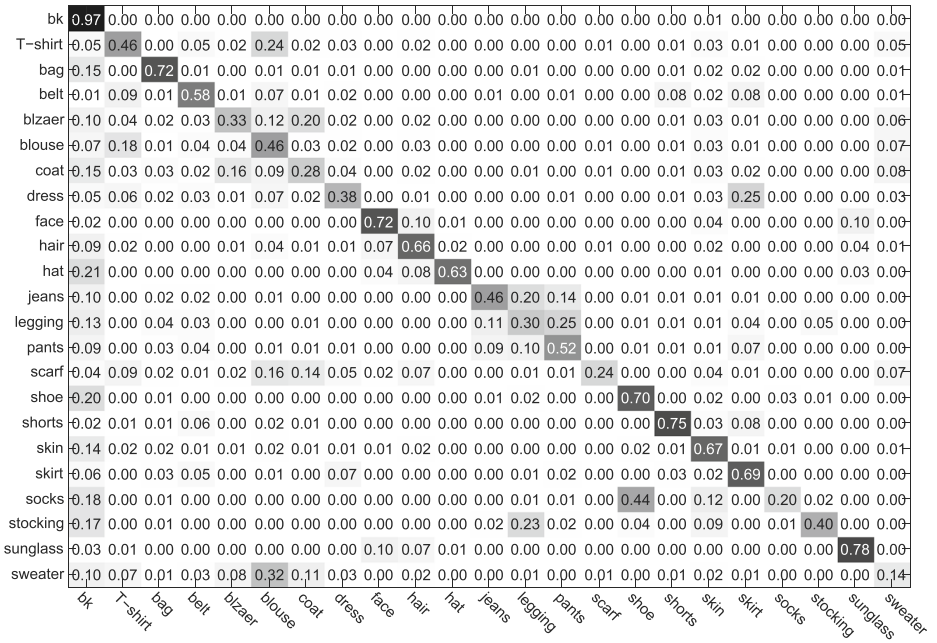


Fig. 5 The confusion matrix of the testing set

as "scarf", "stocking", "hat" and so on. Those labels ,such as "hat", "bag" and "belt" are always too small in human bodies to be classify, however our method still correctly predicts the labels. Besides, using MIL-loss would also encourage those classes appearing in images obtain a higher corresponding score compared with those does not appeared classes. In order to analyze the deeper properties, the confusion matrix of FCN+structure model on testing set is drawn in Fig. 5 respectively. We can see that almost the largest values from each class distribute in the diagonal elements, which shows the effectiveness of our model. Background usually confuses with other classes as it always own various patterns. Those neighboring lables, such as "socks" and "shoe", always confuses with each other, because their adjacence and co-occurance. The "skin" is easily misclassified with almost all the others, since in human parsing, skin could appear in every where of the human. For example, skin may appears in neck as well as scarf does, it may also appears on the ankle as well as socks do.

4.4 Qualitative fashion paring results

Figure 4 show the final parsing results of all 23 category classes. From these examples we can see that our method obtains good performance in most instances. In some challenging situations, our method still give a better parsing result. For example, in Fig. 4, the second image in the first row, the clothes' color is similar with the girl's skin, our model still distinguish them successfully. Beside that, even after previous detection, the detected human box still have complicated background, our model can also handle with these situations and marks them as background. Another advantage of our model is that it is robust toward pose variation and view change. This robustness owes to the convolution and logistic pooling operation of model.Some classes, which does not appears in image,are still assigned in final

result. This may because our model only encourages classes that appeared in image, but not suppress classes which does not appear due to the loss defined by the (4). We will solve these problems in the future.

5 Conclusion

In this paper, we propose a weakly- and semi-supervised human parsing framework. We design a new kind of Logistic Pooling Layer for encouraging both color and category information fire/turned off at the same location to infer the fashion parsing result. Label Aggregation Layer with MIL-loss is also used to guide multi-instance learning. Experiments validate the effectiveness of the proposed framework. In the future, we will explore more information besides color and category, then we will explore other kinds of loss to obtain better fashion parsing result.

Acknowledgements This work was supported by National Natural Science Foundation of China (No.61572493, No.61503281, Grant U1536203, Grant 61602037).

References

1. Chen H, Gallagher A, Girod B (2012) Describing clothing by semantic attributes. In: Computer vision–ECCV 2012, pp 609–623
2. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv:1412.7062
3. Chen Q, Huang J, Feris R, Brown LM, Dong J, Yan S (2015) Deep domain adaptation for describing people based on fine-grained clothing attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5315–5324
4. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005, vol 1. IEEE, pp 539–546
5. Deng Y, Luo P, Loy CC, Tang X (2015) Learning to recognize pedestrian attribute. arXiv:1501.00901
6. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
7. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
8. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 447–456
9. Hong S, Noh H, Han B (2015) Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in neural information processing systems, pp 1495–1503
10. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456
11. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. ACM, pp 675–678
12. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137
13. Liang X, Liu S, Shen X, Yang J, Liu L, Dong J, Lin L, Yan S (2015) Deep human parsing with active template regression. IEEE Trans Pattern Anal Mach Intell 37(12):2402–2414
14. Liu S, Feng J, Song Z, Zhang T, Lu H, Xu C, Yan S (2012) Hi, magic closet, tell me what to wear! In: Proceedings of the 20th ACM international conference on multimedia. ACM, pp 619–628

15. Liu S, Song Z, Liu G, Xu C, Lu H, Yan S (2012) Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012. IEEE, pp 3330–3337
16. Liu S, Feng J, Domokos C, Xu H, Huang J, Hu Z, Yan S (2014) Fashion parsing with weak color-category labels. IEEE Trans Multimedia 16(1):253–265
17. Liu S, Liang X, Liu L, Lu K, Lin L, Cao X, Yan S (2015) Fashion parsing with video context. IEEE Trans Multimedia 17(8):1347–1358
18. Liu S, Liang X, Liu L, Shen X, Yang J, Xu C, Lin L, Cao X, Yan S (2015) Matching-cnn meets knn: quasi-parametric human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1419–1427
19. Liu S, Wang C, Qian R, Yu H, Bao R (2016) Surveillance video parsing with single frame supervision. arXiv:1611.09587
20. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
21. Luo P, Wang X, Tang X (2013) Pedestrian parsing via deep decompositional network. In: Proceedings of the IEEE international conference on computer vision, pp 2648–2655
22. Papandreou G, Chen LC, Murphy KP, Yuille AL (2015) Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1742–1750
23. Pathak D, Shelhamer E, Long J, Darrell T (2014) Fully convolutional multi-class multiple instance learning. arXiv:1412.7144
24. Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1796–1804
25. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
26. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
27. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
28. Van De Weijer J, Schmid C, Verbeek J (2007) Learning color names from real-world images. In: IEEE conference on computer vision and pattern recognition, 2007. CVPR'07. IEEE, pp 1–8
29. Wang C, Yang H, Meinel C (2016) A deep semantic framework for multimodal representation learning. Multimedia Tools Appl 75(15):9255–9276
30. Wang H, Cai Y, Chen X, Chen L (2016) Occluded vehicle detection with local connected deep model. Multimedia Tools Appl 75(15):9277–9293
31. Yamaguchi K, Kiapour MH, Ortiz LE, Berg TL (2012) Parsing clothing in fashion photographs. In: IEEE conference on computer vision and pattern recognition (CVPR), 2012. IEEE, pp 3570–3577
32. Yamaguchi K, Hadi Kiapour M, Berg TL (2013) Paper doll parsing: retrieving similar styles to parse clothing items. In: Proceedings of the IEEE international conference on computer vision, pp 3519–3526
33. Yang W, Luo P, Lin L (2014) Clothing co-parsing by joint image segmentation and labeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3182–3189
34. Zhang H, Zha ZJ, Yang Y, Yan S, Gao Y, Chua TS (2013) Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In: Proceedings of the 21st ACM international conference on multimedia. ACM, pp 33–42
35. Zhang H, Shen F, Liu W, He X, Luan H, Chua TS (2016) Discrete collaborative filtering. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 325–334
36. Zhang H, Kyaw Z, Chang SF, Chua TS (2017) Visual translation embedding network for visual relation detection. arXiv:1702.08319
37. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2014) Object detectors emerge in deep scene cnns. arXiv:1412.6856
38. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. In: Advances in neural information processing systems, pp 487–495

Zhiyong Chen received the BE degree from the School of Information Science and Engineering, Lanzhou University, in 2013. He is a senior master student of the School of Information Science and Engineering, Lanzhou University, China. His current research interests include Computer Vison, Deep learning, Machine Learning and Data Mining.

Si Liu is an Associate Professor in Institute of Information Engineering, Chinese Academy of Sciences (CAS). She was a Research Fellow in Learning and Vision Research Group at National University of Singapore. She obtained Ph.D. degree from Institute of Automation, CAS. Her research interests include object categorization, object detection, image parsing and human pose estimation.

Yanlong Zhai is an Assistant Professor in the School of Computer Science, Beijing Institute of Technology. He was a Visiting Scholar in the Department of Electrical Engineering and Computer Science, University of California, Irvine. He received Ph.D. degree from Beijing Institute of Technology. His research interests include distributed and parallel computing.

Jia Lin is a product manager in the AI and Big Data Division, JD.com. She was a master in Institute of Information Engineering, Chinese Academy of Sciences (CAS). Her research interests include image forensics, object detection, face recognition.

Xiaochun Cao received the BE and ME degrees both in computer science from Beihang University, China, and the PhD degree in computer science from the University of Central Florida, with his dissertation nominated for the university level Outstanding Dissertation Award. He is a professor at the Institute of Information Engineering, Chinese Academy of Sciences. He is also a visiting professor in the School of Information Science and Engineering, Lanzhou University. After graduation, he spent about three years at Object- Video Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. He has authored and coauthored more than 80 journal and conference papers. In 2004 and 2010, he received the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He is a senior member in IEEE.

Liang Yang received the B.E. and M.E. degrees from Nankai University, Tianjin, China, in 2004 and 2007, respectively, both in computational mathematics. He received the Ph.D. degree from Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He is an Assistant Professor with the School of Information Engineering, Tianjin University of Commerce, Tianjin. His current research interests include community detection, semi-supervised learning, low-rank modeling, and deep learning.