Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semantic-spatial fusion network for human parsing

Xiaomei Zhang[a,b], Yingying Chen[a,b,*], Bingke Zhu[a,b], Jinqiao Wang[a,b], Ming Tang[a,b]

[a] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No.95, Zhongguancun East Road, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 31 July 2019
Revised 21 March 2020
Accepted 28 March 2020
Available online xxx

Communicated by Dr Fenza Giuseppe

Keywords:
SSFNet
Semantic modulation model
Resolution-aware model
Human parsing

ABSTRACT

Recently, many methods have united low-level and high-level features to generate the desired accurate high-resolution prediction for human parsing. Nevertheless, there exists a semantic-spatial gap between low-level and high-level features in some methods, i.e., high-level features represent more semantics and less spatial details, while low-level ones have less semantics and more spatial details. In this paper, we propose a Semantic-Spatial Fusion Network (SSFNet) for human parsing to shrink the gap, which generates the accurate high-resolution prediction by aggregating multi-resolution features. SSFNet includes two models, a semantic modulation model and a resolution-aware model. The semantic modulation model guides spatial details with semantics and then effectively facilitates the feature fusion, narrowing the gap. The resolution-aware model sufficiently boosts the feature fusion and obtains multi-receptive-fields, which generates reliable and fine-grained high-resolution features for each branch, in bottom-up and top-down processes. Extensive experiments on three public datasets, PASCAL-Person-Part, LIP and PPSS, show that SSFNet achieves significant improvements over state-of-the-art methods.

1. Introduction

Human parsing [1–11] is a fine-grained semantic segmentation task. It aims to predict the pixel-wise mask of body parts or clothing items for human images. Understanding the details of human images makes sense in some applications, for example, person re-identification [12], human behavior analysis [13], clothing style recognition and retrieval [14], clothing category classification [15], to name a few. However, repeated down-sampling operations of pooling and convolution strides make the prediction lose some details compared to initial images. There are two mainstreams of the low-level and high-level feature fusion networks to obtain the high-resolution prediction. One type of methods [10,11,16] employs the "U-net [17]" structure, which fuses high-level and low-level features with skip connections. The other type of methods [18,19] fuses features by residual connections. The drawback of these above methods is that there is a semantic-spatial gap between features of two different levels [20].

The semantic-spatial gap in the feature fusion is that deep features represent more semantics and less spatial details compared with low features, and vice versa. Consider the extreme case that low-level features only have the capacity to distinguish shallow concepts such as points, lines or edges. Intuitively, it is difficult to fuse high-level features with low-level ones because low-level features are too noisy to provide high-resolution semantic guidance. Similarly, high-level features have little spatial details, and thus, low-level features may not take advantage of the semantics of the high-level. As shown in Fig. 1, some examples can verify the above insights, i.e., there exists the semantic-spatial gap of these predictions generated by MMAN [10] in the second column. Some parts have less spatial details or some spatial details with wrong semantic labels in the second column.

In this paper, we propose a Semantic-Spatial Fusion Network (SSFNet) for human parsing to shrink the gap, which generates an accurate high-resolution prediction. SSFNet mainly includes two models, a semantic modulation model and a resolution-aware model. There are fine-grained gaps compared with semantic segmentation because human parsing segments human bodies into small parts rather than the whole body as done in semantic segmentation. Thus, SSFNet gradually shrinks the fine-grained gap by importing two models in different branches to obtain a coarse-to-fine prediction. Especially, SSFNet fuses multi-resolution features to obtain the desired high-resolution prediction.

The semantic modulation model effectively facilitates the feature fusion between low-level and high-level features, which shrinks the semantic-spatial gap, as shown in Fig. 2 (b). Specifically, our semantic modulation model takes features of two

* Corresponding author at: National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, No.95, Zhongguancun East Road, Beijing 100190, China.
E-mail address: yingying.chen@nlpr.ia.ac.cn (Y. Chen).

https://doi.org/10.1016/j.neucom.2020.03.096
0925-2312/

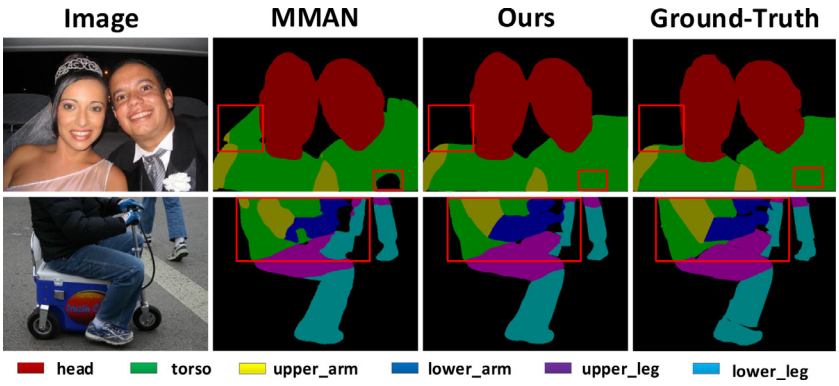


Fig. 1. Examples of predictions. The original images and ground truth come from PASCAL-Person-Part dataset [21]. Predictions are generated by MMAN [10] in the second column. Predictions are generated by our SSFNet in the third column.

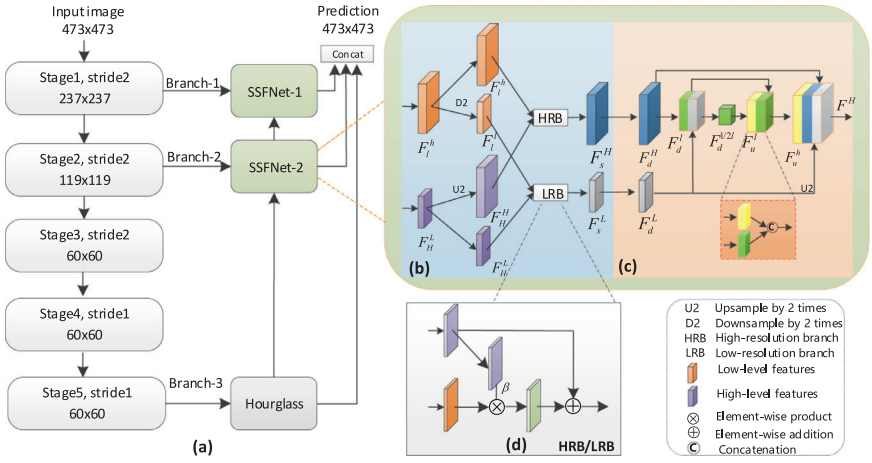


Fig. 2. Overview of the proposed SSFNet (a), it consists of two models: the semantic modulation model (b), and the resolution-aware model (c). (d) is the branch of the semantic modulation model.

different levels as inputs, and generates features with more semantics and spatial details, in a dual branch structure. As shown in Fig. 2 (d), in each branch, a convolutional layer applies to the high-level features to generate a modulation tensor, which guides low-level spatial details with high-level semantics and makes spatial details have related semantic labels. For example, the spatial details of the head position(i.e., edges, corners) have the head labels. Then, high-level features may have a chance to fuse themselves with semantic spatial details. Hence, the model can alleviate the semantic-spatial gap between low-level and high-level features by effective feature fusion. Moreover, this model not only up-samples high-level features but also down-samples low-level features, which is more robust and accurate than these methods [10,11,16] only up-sampling high-level features without regard to low-level features.

In order to obtain more reliable and fine-grained high-resolution features, we present the resolution-aware model, as shown in Fig. 2 (c). This model sufficiently boosts the feature fusion and further shrinks the gap, which can achieve multi-scales and multi-receptive-fields fusion to parse human parts. The ordinary hourglass network [22] centers its attention on one input, whereas our resolution-aware model is different from it. Our model takes two inputs to remedy the missing details along with a series of convolutional operations. Thus, this model has the capacity of extracting deep semantics and keeping the shallow details, and then generates reliable and fine-grained features with different resolutions in different branches, in bottom-up and top-down processes.

Extensive experiments show that SSFNet achieves a new state-of-the-art consistently on three public benchmarks, including PASCAL-Person-Part [21], LIP [8] and PPSS [23]. And LIP can be well generalized on a relatively small dataset PPSS. Specifically, SSFNet outperforms the competing methods by 1.42%, 1.43%, and 5.16% on PASCAL-Person-Part, LIP, and PPSS in terms of mIoU, respectively.

In summary, our contributions can be summarized in three folds:

1. We present a Semantic-Spatial Fusion Network (SSFNet) which shrinks the semantic-spatial gap and achieves new state-of-the-art results on three benchmark datasets.
2. We propose a semantic modulation model which guides spatial details with semantics and then effectively facilitates feature fusion to narrow the semantic-spatial gap between low-level and high-level features.
3. We develop a resolution-aware model which achieves multi-scales and multi-receptive-fields fusion to generate reliable and fine-grained high-resolution features, in bottom-up and top-down processes.

2. Related work

2.1. Human parsing

Many research efforts have been devoted to human parsing [4,5,10,11,16,19]. Gong et al. [4] presented PGN to fuse semantic features and edge features, which generated the prediction with

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096

accurate boundaries. Li et al. [5] proposed a network that fused detectable features with semantic features for human parsing. Nie et al. [19] introduced MuLA network to joint features of human parsing and pose estimation. Liu et al. [11] fused multi-scale features to leverage the useful properties to conduct human parsing. Different from the above methods, our SSFNet can shrink the semantic-spatial gap between low-level and high-level features, which introduces more semantic information into low-level features and more spatial high-resolution information into high-level features.

2.2. High-resolution prediction

Most deep convolutional networks [2–6,9,11,24–27] adopted to up-sample final features for obtaining the desired pixel-wise prediction. The overall strides may hinder the accuracy of results. Vijay et al. [28] proposed SegNet to generate the pixel-wise semantic segmentation by skip connections. Lin et al. [18] presented Refinenet, which gradually recovered the size of features. However, we not only gradually recover the features by some branches but also fuses the outputs of these branches, which combines the advantages of both cascaded and parallel architectures. Zhang et al. [20] proposed ExFuse to bridge the semantic-spatial gap of semantic segmentation. Different from ExFuse, we import some models to shrink the gap gradually because human parsing segments the human body into small parts rather than the whole body as done in semantic segmentation, and we propose a dual structure to make full use of the superiority of high-level semantics and low-level spatial details in our semantic modulation model. Thus our SSFNet is more effective than ExFuse. Chen et al. [29] designed DeepLabv3+ which took advantages of multi-scale features in the final stage to generate the prediction. However, our SSFNet uses a resolution-aware model to obtain multi-scale features in many stages (branches) to obtain more multi-scale features.

2.3. Feature fusion

Newell et al. [22] proposed a stacked hourglass network by repeating bottom-up, top-down processes. Ke et al. [30] introduced a multi-scale structure-aware network. Yang et al. [31] designed a Pyramid Residual Module (PRMs) to enhance the invariance in scales of DCNNs. Nie et al. [32] presented PPN to address the challenging multi-person pose estimation problem. Tang et al. [33] presented DLCM to exploit deep neural networks to learn the compositionality of human bodies. All the above methods take features of one resolution as inputs. In contrast, our resolution-aware model takes features of different resolutions as inputs in a dual structure. In addition, our semantic modulation model can guide spatial details with semantics to facilitate the feature fusion.

3. Proposed network

In this section, we elaborate on our proposed SSFNet including its overall structure and individual components, as shown in Fig. 2. We first introduce the whole network, then the semantic modulation model, finally the resolution-aware model.

3.1. Semantic-spatial fusion network

We aim to alleviate the semantic-spatial gap between high-level and low-level features and obtain the accurate high-resolution prediction. As shown in Fig. 2 (a), our SSFNet is based on the PSPNet [34] framework, and then we divide the framework into three branches on the basis of its resolution, i.e., 1/8, 1/4 and 1/2 size of the initial, respectively. We employ an hourglass network [22] at the top of PSPNet (branch-3) and two SSFNet blocks at branch-1 and branch-2, respectively. Note that each SSFNet block is composed of the semantic modulation model and the resolution-aware model, as shown in Fig. 2 (b)–(c).

We denote SSFNet-l as the SSFNet block that connects to the outputs of branch-l in PSPNet. In practice, the output of each branch is passed through one convolutional layer to adapt the dimensionality. Parameters of SSFNet block are not tied, allowing for a more flexible adaptation for an individual branch. Following the illustration in Fig. 2(a) bottom-up, we start from the bottom branch (branch-3) of PSPNet and connect the output of branch-3 to a common hourglass network. In the next stage, the outputs of the hourglass network and the PSPNet branch-2 are fed into SSFNet-2. SSFNet-2 combines the advantages of high-level semantics and low-level spatial details and facilitates the feature fusion to shrink the semantic-spatial gap, generating high-resolution features. Similarly, SSFNet-1 repeats operations of SSFNet-2.

The overall architecture of the network has two strengths. On the one hand, our network cascades multi-resolution branches by the two SSFNet blocks to gradually shrink the fine-grained gap. On the other hand, our network provides a generic means to fuse features, improving performance. As shown in Fig. 2(a), the network can capture features of three resolutions. We up-sample these features bottom-up by 8 times, 4 times and 2 times, respectively. In this way, our SSFNet can generate an accurate high-resolution prediction.

3.2. Semantic modulation model

Low-level and high-level features are complementary by nature, where low-level features are rich in spatial details but lack semantic information and vice versa [20]. However, as above mentioned, there is a semantic-spatial gap between high-level and low-level features when they are fused. To alleviate the gap between features of two different levels and facilitate feature fusion, we design the semantic modulation model, as shown in Fig. 2(b). Our semantic modulation model takes features of two different levels as inputs, that is, high-level but low-resolution features and low-level high-resolution features. Note that the high-resolution features are 2 times larger than the low-resolution ones in the semantic modulation model.

Our semantic modulation model has two branches which are named low-resolution branch and high-resolution branch, respectively. Specifically, each branch of the model takes features of two different levels as inputs. In the low-resolution branch, we down-sample low-level high-resolution features F_l^h to the size of low-resolution features to generate low-level low-resolution features F_l^l, where F_l^h are passed through a convolutional layer with stride 2. In the high-resolution branch, we up-sample high-level low-resolution features F_H^L to the size of high-resolution features to obtain high-level high-resolution features F_H^H, where F_H^L are passed through a bilinear up-sampling with 2 times. That is,

$$\begin{array}{lcl} F_l^h & \rightarrow & F_l^h \\ & \searrow & F_l^l, \\ F_H^L & \rightarrow & F_H^H \\ & \searrow & F_H^L. \end{array} \tag{1}$$

In each branch, one convolutional layer applies to the high-level features to produce a modulation parameter β, which guides spatial details with semantics. In this way, spatial details have related semantic labels. Unlike other methods, β is not the vector, but tensors with spatial dimensions. Then, high-level features have a chance to fuse themselves with low-level details, which can alleviate the semantic-spatial gap between low-level and high-level features by effective feature fusion. Finally, the model generates features with semantics and spatial details, as shown in Fig. 2 (d).

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096

That is,

$$\begin{aligned}\beta^{lL} &= W^{lL} * \mathcal{F}_H^L + b^{lL},\\ \beta^{hH} &= W^{hH} * \mathcal{F}_H^H + b^{hH},\\ \mathcal{F}_s^L &= sum((\mathcal{F}_l^l \times \beta^{lL}), \mathcal{F}_H^L),\\ \mathcal{F}_s^H &= sum((\mathcal{F}_l^h \times \beta^{hH}), \mathcal{F}_H^H),\end{aligned} \tag{2}$$

where $*$ denotes the convolutional operation, β^{lL} and β^{hH} denote the modulation parameters of the low-resolution and high-resolution branch, respectively, W^{lL} and W^{hH} refer to weights, b^{lL} and b^{hH} refer to biases, $\times$ denotes the element-wise multiplication, *sum* denotes the element-wise sum, $\mathcal{F}_s^L$ denotes the output of the low-resolution branch, and $\mathcal{F}_s^H$ denotes the output of the high-resolution branch.

In this way, low-level and high-level features can be fused in a dual branch structure, which can afford more choices for the network to accommodate the importance of inputs. Meanwhile, the features, the outputs of the semantic modulation model, can be learned to the features from early convolutional layers which encode low-level spatial visual information like edges, corners, circles, etc., and also learned to high-level features from deeper layers which encode high-level semantic information, including object- or category-level evidence.

3.3. Resolution-aware model

To acquire more reliable and fine-grained high-resolution features, we present the resolution-aware model that takes outputs of the semantic modulation model as inputs, as shown in Fig. 2 (c). Here, high-resolution features are 2 times larger than low-resolution features, which is the same as our semantic modulation model. Hourglass network [22] has one input, so its sequential operations take this as the source of information. Different from [22], our resolution-aware model has two inputs, which not only takes the features of high-resolution as the input but also absorbs the features of low-resolution during the single pipeline. This way preserves fine-grained semantic-spatial features in bottom-up and top-down processes. Every resolution-aware model reaches its lowest resolution at the one forth of high-resolution features and generates features with high-resolution.

In the bottom-up process, high-resolution features $\mathcal{F}_s^H$ and low-resolution features $\mathcal{F}_s^L$ are operated by convolutional layers, generating $\mathcal{F}_d^H$ and $\mathcal{F}_d^L$, respectively. Then $\mathcal{F}_d^H$ is down-sampled 2 times by a convolutional layer and concatenated with $\mathcal{F}_d^L$. Finally, they are sent into a convolutional layer with stride 2 to generate the smallest features to obtain $\mathcal{F}_d^{1/2l}$, that is,

$$\begin{aligned}&\mathcal{F}_s^H \rightarrow \mathcal{F}_d^H,\\ &\mathcal{F}_s^L \rightarrow \mathcal{F}_d^L,\\ &\mathcal{F}_d^H \xrightarrow{d,1/2} \mathcal{F}_d^l,\\ &concat(\mathcal{F}_d^l, \mathcal{F}_d^L) \rightarrow \mathcal{F}_d^l,\\ &\mathcal{F}_d^l \xrightarrow{d,1/2} \mathcal{F}_d^{1/2l}.\end{aligned} \tag{3}$$

After reaching the lowest resolution, the model begins the top-down sequence. The smallest features $\mathcal{F}_d^{1/2l}$ are zoomed 2 times by bilinear up-sampling. Meanwhile, the low-resolution semantic-spatial features $\mathcal{F}_s^L$ also are up-sampled 2 times. Then, these features are pass through a series of operations, that is,

$$\begin{aligned}&\mathcal{F}_d^{1/2l} \xrightarrow{u,2} \mathcal{F}_u^l,\\ &concat(\mathcal{F}_u^l, \mathcal{F}_d^L) \xrightarrow{u,2} \mathcal{F}_u^h,\\ &\mathcal{F}_d^L \xrightarrow{u,2} \mathcal{F}_u^H,\\ &concat(\mathcal{F}_d^H, \mathcal{F}_u^h, \mathcal{F}_u^H) \rightarrow \mathcal{F}^H,\end{aligned} \tag{4}$$

where $\mathcal{F}_u^l$ and $\mathcal{F}_u^h$ denote features up-sampled 2 times by bilinear interpolation, $\mathcal{F}_u^H$ denotes the $\mathcal{F}_d^L$ features up-sampled 2 times, $\mathcal{F}^H$ denotes the outputs of the resolution-aware model.

Finally, reaching the output resolution of the branch, two consecutive rounds of 1x1 convolutional layers are applied to produce the high-resolution features of this model.

3.4. Loss function

Following PSPNet [34], our SSFNet employs two deep auxiliary losses, which are named as L_{aux1} and L_{aux2}, respectively. L_{aux1} locates at the end of our baseline and L_{aux2} is applied after the twenty-second block of the fourth stage of ResNet101, i.e., the res4b22 residue block. In addition, there is a loss at the end of SSFNet named as $L_{softmax}$. The total loss can be formulated as:

$$L = \lambda L_{softmax} + \lambda_1 L_{aux1} + \lambda_2 L_{aux2}, \tag{5}$$

where we fix the hyper-parameters $\lambda = 0.6$, $\lambda_1 = 0.5$, and $\lambda_2 = 0.4$ in our experiments. Following PSPNet, the auxiliary loss weight λ_2 is the same as PSPNet. We experiment with setting hyper-parameters λ and λ_1 from 0 to 1, respectively. And then $\lambda = 0.6$ and $\lambda_1 = 0.5$ yield the best performance.

4. Experiments

To evaluate the performance of the proposed SSFNet, we perform the experiments on three datasets, including PASCAL-Person-Part [21], LIP [8] and PPSS [23]. The first is a single-person and multiple-person dataset, the last two are single-person datasets. The accuracy of each part (clothes) is measured by pixel Intersection-over-Union (IoU) in human parsing. The mean pixel Intersection-over-Union (mIoU) is computed by averaging the IoU across all parts. We use both IoU and mIoU as evaluation metrics for these three datasets.

4.1. Implementation details

We take the PSPNet [34] followed an hourglass network [22] as the baseline. Additionally, we train SSFNet in an end-to-end manner. As for the input size, we resize it to 473 $\times$ 473. We train all the networks using stochastic gradient descent (SGD) solver, momentum 0.9 and weight decay 0.0005. The batch sizes are 8 on the three different datasets. The epochs of three datasets are 100. For data augmentation, we apply the random scaling (from 0.5 to 1.5), and left-right flipping during training. In the inference process, we test images on the multi-scale to acquire multi-scale contexts as similar to most semantic segmentation tasks. All networks are trained on NVIDIA GTX TITAN X GPU with 12 GB memory.

4.2. Results on PASCAL-Person-Part

In PASCAL-Person-Part [21], there are multiple personal appearances in an unconstrained environment. Each image has 7 labels: background, head, torso, upper-arm, lower-arm, upper-leg and lower-leg. We use the images containing human for training (1716 images) and validation (1817 images).

Discussion about Baseline. Because the hourglass network [22] has the capacity of processing across all scales features to capture the various spatial relationships, so we use it to improve the performance of the baseline. In order to exploit the suitable baseline without any proposed models, we follow [22] to repeat different quantities of hourglass networks from 1 to 3 in a cascaded way. They are named as Baseline, Baseline-D, Baseline-T, Baseline-T-A, respectively. In Table 1, we use the same hyper-parameters to train Baseline, Baseline-D, and Baseline-T, i.e., learning rate, input size, batch size, epochs, and so on. These hyper-parameters

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096

Table 1
Discussion about baselines. The results are obtained on the validation set of PASCAL-Person-Part [21]. Baseline is PSPNet with an hourglass network [22]. Baseline-D is PSPNet with dual hourglass networks. Baseline-T is PSPNet with three hourglass networks. Baseline-T-A is Baseline-T with appropriate hyper-parameters.

Method	one	two	three	Ave.
Baseline	✓			65.67
Baseline-D		✓		65.85
Baseline-T			✓	65.46
Baseline-T-A			✓	65.98

Table 2
Ablation study of our SSFNet. H denotes the high-level features up-sampled. HL denotes the high-level features up-sampled and the low-level features down-sampled. SMM is our semantic modulation model. CH denotes the common hourglass network. L denotes the low-resolution semantic-spatial features to employ at the first concatenation of our models. RAM is the resolution-aware model. SSFNet-3D is the network adding the third branch with the smallest features. SSFNet-3U is the network adding the third branch with the largest features.

Method	Ave.
Semantic Modulation Model	
Baseline+H	65.80
Baseline+HL	66.09
Baseline+SMM	67.19
Resolution-Aware Model	
Baseline+SMM+CH	67.73
Baseline+SMM+L	68.55
Baseline+SMM+RAM	69.62
Semantic-Spatial Fusion Network	
SSFNet-3D	69.76
SSFNet-3U	69.83
SSFNet	**70.08**

are appropriate for Baseline-D that achieves the best performance, but not appropriate for Baseline-T. We further adjust these hyper-parameters to Baseline-T named Baseline-T-A, i.e., increasing the epochs from 100 to 120. The performance of Baseline- T-A is improved and exceeds Baseline-D. From Table 1, we find that increasing the number of hourglass networks from 1 to 3 improves little accuracy. We conclude that PSPNet with an hourglass network (the baseline) yields the best trade-off between performance and cost, with a performance of 65.67% in terms of mIoU. And the number of parameters for the baseline is the smallest in these baselines.

Ablation study for semantic modulation model. In order to investigate the effectiveness of our semantic modulation model, we conduct experiments with three settings. The first is to only up-sample high-level features 2 times to the resolution of the high-resolution features and then fuse them, which is similar to most methods [17,35] and named Baseline+H. The second is named as Baseline+HL, which not only up-samples high-level features but also down-samples low-level features in a dual branch structure, and then we fuse them. The third is to use our modulation mechanism to guide the low-level features with semantics, named as Baseline+SMM. As shown at the top of Table 2, the performance of Baseline+HL outperforms Baseline+H by 0.29% in terms of mIoU. The performance is further improved by our modulation mechanism, achieving 67.19%.

Ablation study for resolution-aware model. To evaluate the effectiveness of absorbing the other input in the resolution-aware model, we also conduct experiments with three settings. First, we add a common hourglass network [22] at the top of each semantic modulation model, which is named as Baseline+SMM+CH. Then, in the resolution-aware model, the low-resolution features are employed at the first concatenation, named Baseline+SMM+L. Third, based on Baseline+SMM+L, we up-sample low-resolution features to high-resolution features and concatenate them, to compose our resolution-aware model, the whole network is named Baseline+SMM+RAM. As shown in the middle of Table 2, our performance makes further improvements by our resolution-aware model, achieving 69.62%.

Ablation study for semantic-spatial fusion network. We try more features with different resolutions to structure our SSFNet block, i,e., the semantic modulation model has more branches and generates outputs of different resolutions, so the resolution-aware model has more inputs. Such as, we add the third branches in the semantic modulation model, and it corresponds to 1/4 time smaller or 2 times larger than the high-resolution features. They are named as SSFNet-3D and SSFNet-3U, respectively. The performance of those branches does not improve. Thus, we conclude that the semantic modulation model with two branches has the best trade-off between performance and cost.

Comparison with state-of-the-art methods. We compare SSFNet with state-of-the-art human parsing methods including HAZA [2], LIP [8], MMAN [10], Graph LSTM [36], SE LSTM [6], Joint [3], PCNet [9], Holistic [5] and RefineNet [18]. As shown in Table 3, the proposed SSFNet outperforms those human parsing methods for all metrics and exceeds them by 12.54%, 10.72%, 10.17%, 9.47%, 6.51%, 5.69%, 4.18%, 3.78% and 1.48%, respectively.

We propose a semantic modulation model to alleviate the semantic-spatial gap between high-level and low-level features, which combines high-level semantics and low-level spatial details. Our SSFNet has similar properties and common advantages like these methods in Table 3 and further improves the properties to improve accuracy. Firstly, we gradually refine the features like RefineNet [18], whereas we further fuse the features of multi-scales in parallel to obtain the rich contextual information. Secondly, Holistic [5], MMAN [10] and LIP [8] only fuse features of high-level semantics in one stage. However, our SSFNet fuses features in many stages to improve accuracy. Thirdly, PCNet [9] and HAZA [2] hierarchically segment the human parts to reduce the irrelevant information. In our SSFNet, the latter SSFNet-blocks inherit the results of formers to reduce the irrelevant information. Fourthly, Joint [3], SE LSTM [6] and Graph LSTM [36] use the global

Table 3
Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with state-of-the-art methods on PASCAL-Person-Part.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
HAZA [2]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
LIP [8]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
MMAN [10]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
Graph LSTM [36]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.61
SE LSTM [6]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [3]	85.50	67.87	54.72	54.30	48.25	44.78	95.32	64.39
PCNet [9]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [5]	–	–	–	–	–	–	–	66.3
RefineNet [18]	–	–	–	–	–	–	–	68.6
SSFNet(ours)	**87.82**	**73.22**	**62.75**	**61.54**	**55.54**	**53.41**	**96.29**	**70.08**

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096

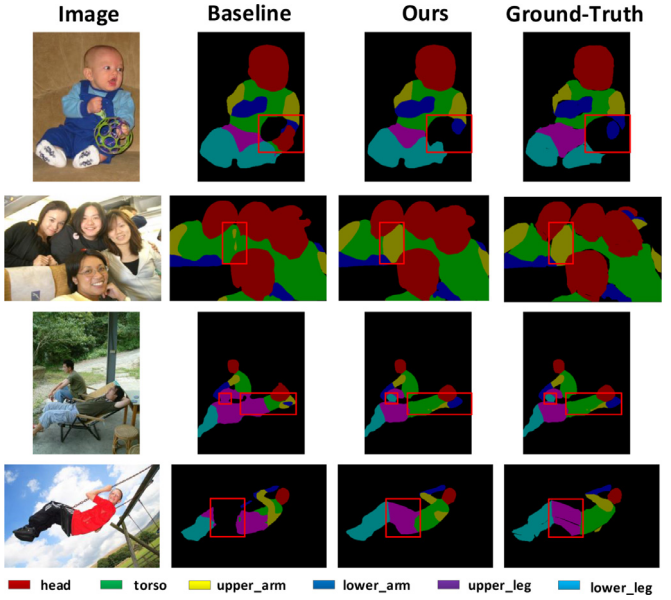


Fig. 3. Qualitative comparison between our SSFNet and the baseline on PASCAL-Person-Part [21] dataset. In the first two rows, SSFNet can correctly parse human parts and extract more spatial details. In the third row, SSFNet segments different human parts more accurately, such as torso and lower-leg. In the last row, our SSFNet accurately parses the whole upper-leg compared to the baseline.

semantic to improve the performance, our resolution-aware model can obtain multi-scale features to generate multi-receptive-fields to improve the performance. Benefit from the above properties, SSFNet has a better capacity to classify each part and a better performance on human parsing task than other human parsing methods.

Qualitative Comparison. The qualitative comparison of human parsing results on PASCAL-Person-Part [21] is visualized in Fig. 3. There are some failure cases of the baseline, nevertheless, our SSFNet can emend them effectively. In the first row, we find that our SSFNet has better performance in extracting spatial details and producing correct prediction from a complex scene, compared with the baseline. It is because that our semantic modulation model can guide spatial details with semantics to facilitate the feature fusion and extract more accurate details. In the second row, ours performs well on the upper-arm with salient and correct boundary in the image compared with the baseline because our resolution-aware model can obtain multi-receptive-fields to parse small parts effectively. In the third row, the baseline mistakes torso, upper-arm and lower-leg, however, but these parts can be well segmented by our network gradually refining the features. For upper-leg in the last row, SSFNet can accurately parse it due to the effective semantic-spatial fusion and the rich contextual information, whereas the baseline misses it entirely.

4.3. Results on LIP

LIP [8] contains 50,462 images in total, including 30,362 for training, 10,000 for testing and 10,000 for validation. LIP defines 19 human parts (clothes) labels, including hat, hair, sunglasses, upper-clothes, dress, coat, socks, pants, gloves, scarf, skirt, jumpsuits, face, right-arm, left-arm, right-leg, left-leg, right-shoe and left-shoe, and a background class. We use its training set to train our network and its validated set to test our SSFNet.

We compare SSFNet with state-of-the-art networks on the validation set, which are SegNet [28], FCN-8s [37], Attention [24], DeepLab-ASPP[25], LIP [8], MMAN [10], JPPNet [38] and CE2P [39]. As shown in Table 4, our SSFNet outperforms all prior methods.

Table 4
Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with the state-of-the-art methods on LIP [8]. The p-values of CE2P [39] and SSFNet (ours) come from two-tail t-tests using paired sample (α = 0.05).

Method	hat	hair	glov	sung	clot	dress	coat	sock	pants	suit	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-sh	r-sh	bkg	Ave.
SegNet [28]	26.60	44.01	0.01	0.00	34.46	0.00	15.97	3.59	33.56	0.01	0.00	0.00	52.38	15.30	24.23	13.82	13.17	9.26	6.47	70.62	18.17
FCN-8s [37]	39.79	58.96	5.32	3.08	49.08	12.36	26.82	15.66	49.41	6.48	0.00	2.16	62.65	29.78	36.63	28.12	26.05	17.76	17.70	78.02	28.29
Attention [24]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
DeepLab-ASPP [25]	56.48	65.33	29.98	19.67	62.44	30.33	51.03	40.51	69.00	22.38	11.29	20.56	70.11	49.25	52.88	42.37	35.78	33.81	32.89	84.53	44.03
LIP [8]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.73
MMAN [10]	57.66	66.63	30.70	20.02	64.15	28.39	51.98	41.46	71.03	23.61	9.65	23.20	68.54	55.30	58.13	51.90	52.17	38.58	39.05	84.75	46.81
JPPNet [38]	63.55	70.20	36.16	23.48	68.15	31.42	55.65	44.56	72.19	28.39	18.76	25.14	73.36	61.97	63.88	58.21	57.99	44.02	44.09	86.26	51.37
CE2P [39]	65.29	72.54	39.09	32.73	69.46	32.52	56.28	**49.67**	74.11	27.23	14.19	22.51	75.50	**65.14**	**66.59**	**60.10**	**58.59**	46.63	46.12	87.67	53.10
SSFNet (ours)	**68.60**	**73.14**	**40.02**	**33.57**	**70.51**	**34.89**	**57.38**	49.30	**74.87**	**33.16**	**21.30**	**29.11**	**75.74**	64.85	66.52	57.41	57.04	**47.43**	**47.74**	**87.87**	**54.53**
p-value(CE2P,SSFNet)	0.02	0.01	0.05	0.01	0.01	0.02	0.02	0.01	0.01	0.05	0.02	0.02	0.05	0.01	0.02	0.05	0.01	0.05	0.02	0.01	0.02

Table 5
Top: performance comparison of the model trained on LIP to test the PPSS [23]. Bottom: performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with the state-of-the-art methods on PPSS. The p-values of MMAN [10] and SSFNet (ours) come from two-tail *t*-tests using paired sample (α=0.05).

Method	Hair	Face	U-cloth	arms	L-cloth	Legs	Background	Ave.
MMAN [10]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
SSFNet (ours)	59.67	58.15	64.54	43.66	59.75	27.93	87.09	57.26
p-value(MMAN,SSFNet)	0.02	0.01	0.05	0.05	0.01	0.01	0.02	0.01
DDN [23]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [40]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
SSFNet (ours)	**70.95**	**62.43**	**83.61**	**48.08**	**73.36**	**29.93**	**92.60**	**65.85**

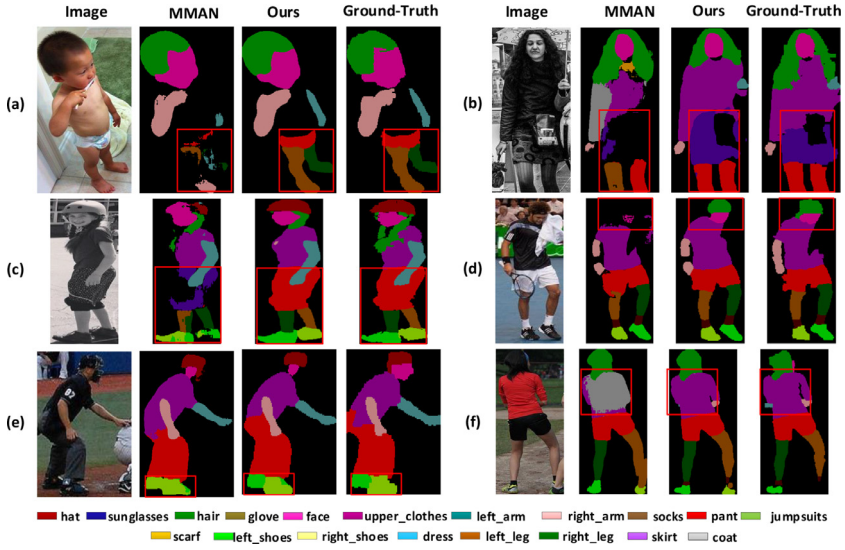


Fig. 4. Qualitative comparison between our SSFNet and state-of-the-art method MMAN on LIP [8] dataset. In the first row, because of most parts or clothes with similar colors, MMAN fails to segment pants, legs in the left panel and skirt, pants in the right panel. Yet our SSFNet can generate those parts accurately. In the second row, our SSFNet distinguishes the most parts compared with MMAN, such as pants, left-leg, right-leg, left-shoes and right-shoes. In the last row, our SSFNet corrects the prediction for shoes and skirt.

Fig. 5. Comparison of our method and state-of-the-art methods on PPSS [23]. In the second and third columns, these models only train on LIP and test on PPSS, the second is MMAN and the third is ours. The last two columns show that ours is able to segment parts correctly.

The proposed method yields the result of 54.53% in terms of mIoU on the LIP. Compared with other methods, ours exceeds 36.36%, 26.24%, 11.61%, 10.80%, 9.80%, 7.72%, 3.16% and 1.43%, respectively. The comparison of some parts is slightly inferior to CE2P, such as sock, right-leg. This is probably due to the human scale variance in this dataset.

Four examples are illustrated in Fig. 4. In the first row, due to most parts with similar colors, MMAN fails to segment the pants and legs, but SSFNet accurately segments all of human parts with specific details. In the second row, MMAN generates the wrong boundaries of a skirt and regards pants as right-leg, while our SSFNet corrects the errors. In the next row, MMAN regards pants as skirt and mistakes left-legs, right-legs and shoes, whereas our SSFNet corrects the errors. In the last row, for head and face, although MMAN almost misses them, ours correctly produces the results.

4.4. Results on PPSS

PPSS [23] includes 3673 annotated samples, which are divided into a training set of 1781 images and a testing set of 1892 images. It defines seven human parts, including hair, face, upper-clothes, low-clothes, arms, legs and shoes. Collected from 171 surveillance videos, the dataset can reflect the occlusion and illumination variation in the real scene.

To evaluate the generalization ability of we proposed SSFNet, we deploy the SSFNet trained on LIP [8] to the testing set of the PPSS [23] without any fine-tuning, which is similar to MMAN. We merge the fine-grained labels of LIP into coarse-grained human

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096

Table 6
Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) on the cityscapes test set. The p-values of Multitask Learning [44] and SSFNet(ours) come from two-tail t-tests using paired sample (α=0.05).

Method	mIoU(%)
DeepLabv2 [25]	70.4
LC [41]	71.1
Adelaide [42]	71.6
FRRN [43]	71.8
RefineNet [18]	73.6
PSPNet [34]	78.4
Multitask Learning [44]	78.5
SSFNet (ours)	79.7
p-value(Multitask Learning,SSFNet)	0.01

parts defined in PPSS. As shown in the first two rows of Table 5, our SSFNet outperforms MMAN by 4.59%.

We train SSFNet on PPSS [23] and compare our SSFNet with some methods on the testing set, DDN [24], ASN [8] and MMAN [10]. The results of the segmentation of SSFNet on the PPSS achieve further improvement. Our proposed framework achieves 65.85% in terms of Mean IoU on the PPSS dataset. Compared with the DDN [23] and ASN [40], our SSFNet exceeds them by 18.65% and 15.15%, respectively. Several examples are shown in Fig. 5.

4.5. SSFNet For semantic segmentation in general scenarios

To verify the efficacy of SSFNet for semantic segmentation in general scenarios, we evaluate the proposed SSFNet on CityScapes [45]. Cityscapes has 5000 images captured from 50 different cities. Each image has 2048 $\times$ 1024 pixels, which have high quality pixel-level labels of 19 semantic classes. There are 2979 images in the training set, 500 images in the validation set, and 1525 images in the test set. We do not use coarse data in our experiments.

Compared with the baseline (76.4%), our SSFNet improves the performance to 79.7%. In Table 6, compared with other methods, SSFNet achieves the best performance on CityScapes, which verifies its effectiveness for semantic segmentation in general scenarios.

5. Conclusion

In this paper, we propose a novel CNN architecture for human parsing, Semantic-Spatial Fusion Network (SSFNet), to alleviate the semantic-spatial gap and generate the precise prediction. SSFNet includes two models, a semantic modulation model, and a resolution-aware model. The semantic modulation model narrows the semantic-spatial gap between the high-level and low-level features by exploring the mutual information and outputs semantic-spatial features of two resolutions, where these maps learn and teach each other in a dual branch structure. In order to obtain reliable and fine-grained high-resolution features, the resolution-aware model achieves multi-scales and multi- receptive-fields fusion, in a bottom-up and top-down process. Moreover, we introduce a path aggregate architecture to fuse the advantages of features on different resolutions. Extensive experiments on three public datasets, PASCAL-Person-Part, LIP, and PPSS, show that our SSFNet achieves significant improvements over state-of-the-art methods.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xiaomei Zhang: Conceptualization, Methodology, Software, Validation, Writing - original draft. **Yingying Chen:** Writing - review & editing. **Bingke Zhu:** Data curation, Visualization, Writing - review & editing. **Jinqiao Wang:** Writing - review & editing. **Ming Tang:** Writing - review & editing.

Acknowledgments

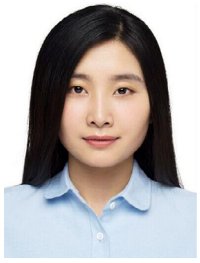
This work was supported by National Natural Science Foundation of China 61976210, 61772527.

References

[1] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille, Joint object and part segmentation using deep learned potentials, in: IEEE International Conference on Computer Vision, 2015, pp. 1573–1581.
[2] F. Xia, P. Wang, L.C. Chen, A.L. Yuille, Zoom better to see clearer: human and object parsing with hierarchical auto-zoom net, in: IEEE International Conference on Computer Vision, 2015, pp. 648–663.
[3] F. Xia, P. Wang, X. Chen, A. Yuille, Joint multi-person pose estimation and semantic part segmentation, in: IEEE International on Computer Vision and Pattern Recognition Workshops, 2017, pp. 6080–6089.
[4] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, L. Lin, Instance-level human parsing via part grouping network, in: European Conference on Computer Vision, 2018, pp. 57–70.
[5] Q. Li, A. Arnab, P.H. Torr, Holistic, instance-level human parsing, British Machine Vision Conference, (2017).
[6] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, E.P. Xing, Interpretable structure-evolving LSTM, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2175–2184.
[7] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, S. Yan, Semantic object parsing with local-global long short-term memory, in: IEEE International on Computer Vision and Pattern Recognition, 2016, pp. 3185–3193.
[8] K. Gong, X. Liang, D. Zhang, X. Shen, L. Lin, Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing., in: CVPR, 2, 2017, p. 6.
[9] B. Zhu, Y. Chen, M. Tang, J. Wang, in: Progressive cognitive human parsing, 2018.
[10] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, Y. Yang, Macro-micro adversarial network for human parsing, in: European Conference on Computer Vision, 2018.
[11] T. Liu, T. Ruan, Z. Huang, Y. Wei, S. Wei, Y. Zhao, H. Thomas, Devil in the details: towards accurate single and multiple human parsing, in: AAAI Conference on Artificial Intelligence, 2019.
[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE International on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[13] Y. Wang, T. Duan, Z. Liao, D. Forsyth, Discriminative hierarchical part-based models for human parsing and action recognition, J. Mach. Learn. Res. 13 (1) (2012) 3075–3102.
[14] K. Yamaguchi, M.H. Kiapour, T.L. Berg, Paper doll parsing: retrieving similar styles to parse clothing items, in: IEEE International Conference on Computer Vision, 2014, pp. 3519–3526.
[15] W. Wang, Y. Xu, J. Shen, S. Zhu, Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: IEEE International on Computer Vision and Pattern Recognition, 2018.
[16] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, S. Yan, Human parsing with contextualized convolutional neural network, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2016) 115–127.
[17] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[18] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.
[19] X. Nie, J. Feng, S. Yan, Mutual learning to adapt for joint human parsing and pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 502–517.
[20] Z. Zhang, X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284.
[21] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille, Detect what you can: detecting and representing objects using holistic models and body parts, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1979–1986.
[22] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: European Conference on Computer Vision, Springer, 2016, pp. 483–499.
[23] P. Luo, X. Wang, X. Tang, Pedestrian parsing via deep decompositional network, in: IEEE International Conference on Computer Vision, 2014, pp. 2648–2655.

[24] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: scale-aware semantic image segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[25] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[26] F. Cheng, H. Zhang, D. Yuan, M. Sun, Leveraging semantic segmentation with learning-based confidence measure, Neurocomputing 329 (2019) 21–31.
[27] F. Shen, G. Zeng, Semantic image segmentation via guidance of image classification, Neurocomputing 330 (2019) 259–266.
[28] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for scene segmentation, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017) 2481–2495.
[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder–decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.
[30] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 713–728.
[31] W. Yang, S. Li, W. Ouyang, H. Li, X. Wang, Learning feature pyramids for human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1281–1290.
[32] X. Nie, J. Feng, J. Xing, S. Yan, Pose partition networks for multi-person pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 684–699.
[33] W. Tang, P. Yu, Y. Wu, Deeply learned compositional models for human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 190–206.
[34] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: IEEE International on Computer Vision and Pattern Recognition, 2017.
[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[36] X. Liang, X. Shen, J. Feng, L. Lin, S. Yan, Semantic object parsing with graph LSTM, in: European Conference on Computer Vision, Springer, 2016, pp. 125–143.
[37] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[38] X. Liang, K. Gong, X. Shen, L. Lin, Look into person: joint body parsing & pose estimation network and a new benchmark, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2018) 871–885.
[39] T. Liu, T. Ruan, Z. Huang, Y. Wei, S. Wei, Y. Zhao, T. Huang, Devil in the details: towards accurate single and multiple human parsing, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 4814–4821.
[40] P. Luc, C. Couprie, S. Chintala, J. Verbeek, Semantic segmentation using adversarial networks, in: NIPS Workshop on Adversarial Training, 2016.
[41] X. Li, Z. Liu, P. Luo, C. Change Loy, X. Tang, Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3193–3202.
[42] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.
[43] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4151–4160.
[44] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491.
[45] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

Xiaomei Zhang received her B.S. degree from North China Electric Power University, China, in 2016. She is currently working toward the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. Her current research interests include pattern recognition and machine learning, image processing, and semantic segmentation.



Yingying Chen received her B.S. degree in 2013 from Communication University of China, and Ph.D. degree in 2018 from University of Chinese Academy of Sciences. She is currently an assistant professor in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. Her current research interests include pattern recognition and machine learning, image and video processing, and intelligent video surveillance.

Bingke Zhu received his B.S. degree from Beijing University of Chemical Technology, China, in 2016. He is currently working toward the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. His current research interests include pattern recognition and machine learning, alpha matting, semantic segmentation, and instance segmentation.

Jinqiao Wang received the B.E. degree in 2001 from Hebei University of Technology, China, and the M.S. degree in 2004 from Tianjin University, China. He received the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2008. He is currently a Professor with Chinese Academy of Sciences. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Ming Tang received the B.S. degree in computer science and engineering and M.S. degree in artificial intelligence from Zhejiang University, Hangzhou, China, in 1984 and 1987, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Chinese Academy of Sciences, Beijing, China, in 2002. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and machine learning.

Please cite this article as: X. Zhang, Y. Chen and B. Zhu et al., Semantic-spatial fusion network for human parsing, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.096