The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Progressive Cognitive Human Parsing

Bingke Zhu,[1,2] **Yingying Chen,**[1,2] **Ming Tang,**[1,2] **Jinqiao Wang**[1,2]
[1]National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{bingke.zhu, yingying.chen, tangm, jqwang}@nlpr.ia.ac.cn

Abstract

Human parsing is an important task for human-centric understanding. Generally, two mainstreams are used to deal with this challenging and fundamental problem. The first one is employing extra human pose information to generate hierarchical parse graph to deal with human parsing task. Another one is training an end-to-end network with the semantic information in image level. In this paper, we develop an end-to-end progressive cognitive network to segment human parts. In order to establish a hierarchical relationship, a novel component-aware region convolution structure is proposed. With this structure, latter layers inherit prior component information from former layers and pay its attention to a finer component. In this way, we deal with human parsing as a progressive recognition task, that is, we first locate the whole human and then segment the hierarchical components gradually. The experiments indicate that our method has a better location capacity for the small objects and a better classification capacity for the large objects. Moreover, our framework can be embedded into any fully convolutional network to enhance the performance significantly.

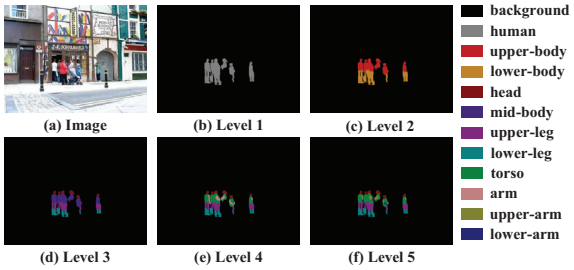

Figure 1: A sketch of progressive cognitive human parsing. (a) Input image. (b) Level 1 segmentation mask for human level. (c) Level 2 segmentation mask for {upper-body, lower-body} level. (d) Level 3 segmentation mask for {head, mid-body, upper-leg, lower-leg} level. (e) Level 4 segmentation mask for {head, torso, arm, upper-leg, lower-leg} level. (f) Level 5 segmentation mask for {head, torso, upper-arm, lower-arm, upper-leg, lower-leg} level. Note that the level division can be altered for different dataset.

Introduction

The goal of human parsing is to segment a human image into different fine-grained semantic parts such as head, torso, arms and legs. Human parsing can provide more detailed understanding of image contents and the human parts. It has many high-level computer vision applications such as image/video retrieval (Yamaguchi et al. 2012), person re-identification (Zhao, Ouyang, and Wang 2013), virtual fitting (Liu et al. 2012), fine-grained recognition (Fu, Zheng, and Tao 2017), action recognition (Wang, Wang, and Yuille 2013) as well as video surveillance (Lu et al. 2014). However, human parsing is still a challenging computer vision problem due to the complicated and various human appearance, size, shape, clothes, occlusion, illumination and semantic ambiguity.

Generally, there are two mainstreams to deal with the human parsing task. The first one is employing the graph models based on the hierarchical prior information (Xia et al. 2016b; Park, Nie, and Zhu 2017). These methods segment human parts with the parse graph models which are built by poses, locations and shapes of human. As a result, these methods can use the relation of human parts for better human parsing. However, these methods often need extra human pose information and cannot be trained in an end-to-end way.

In contrast, another mainstream methods build their models with the end-to-end networks based on the semantic information (Xia et al. 2017; Chen et al. 2017; Lin et al. 2017), which benefit from the powerful performance of fully convolutional networks in semantic segmentation task. Although such approaches demonstrate the good performance on the semantic segmentation, it is still hard to segment all the fine-grained parts of the human with a single classifier because of the complex relationships among human parts and the unbalanced classes among different human parts and the background.

In this paper, we develop an end-to-end progressive cognitive network to recognize different human parts from the whole image gradually. This recognition manner is similar to human vision habits. In fact, the human vision system recognizes an object and its constituent parts in a quite different way. The visual cortex extracts features of an object in the

bottom-up way and classifies the object and its fine-grained parts in the top-down way (Epshtein, Lifshitz, and Ullman 2008). This means that the object parts can be better recognized with the prior detection and localization of the entire object. Inspired by the recognition process of human vision system, we propose a novel network, progressive cognitive network (PCNet).

In order to achieve the progressive networks for human parsing, a novel component-aware region convolution structure is proposed. In this structure, the latter layers can take advantage of the prior information from the former layers. The prior information makes the latter layers can focus on the region of interest and ignore the irrelevant information. The region of interest is sent into the next component-aware region convolutional networks for further analysis so that the part recognition stress of the networks can be decomposed into stacked hierarchical networks.

The component-aware region convolution can be regarded as an irregular convolution, which only operates the convolution on the specific regions. The component-aware region loss is used to provide a prior shape of human parts to the feature maps. For example, we can separate the human-level feature maps and the background regions from the image-level feature maps. In this way, the former layers can provide a prior shape and location information to the latter layers.

As shown in Fig. 1, we divide the labels of PASCAL-Person-Part dataset into five levels. The progressive cognitive networks distinguish the human from the background first in the level 1. And then the component-aware region convolutions only perform on the human feature maps. We convolve the human feature maps into upper-body feature maps through upper-body aware convolutions, and lower-body feature maps through lower-body aware convolutions, respectively. The upper-body feature maps and lower-body feature maps are concatenated into {upper-body, lower-body} level feature maps. The similar processes are performed from level 2 to level 5 until all parts are segmented. More details of the framework are illustrated in Fig. 2, and we will elaborate it in the following sections.

Our progressive cognitive networks fuse the ideas of hierarchical structures and progressive recognition of human parts. The proposed framework has five advantages. Firstly, the progressive cognitive networks for human parsing can be embedded into any fully convolutional network to enhance the performance significantly. Secondly, our progressive cognitive networks inherit the idea of hierarchical structure, which is similar to the human vision habits and decomposes recognition stress of fine-grained task into the whole hierarchical networks rather than makes all decisions in the final classifier. Thirdly, our progressive cognitive networks fully exploit fully convolutional networks for powerful semantic understanding and can be trained in an end-to-end way. Fourthly, compared to the hierarchical optimization which is labeled from coarse to fine (Munoz, Bagnell, and Hebert 2010; Li et al. 2017b), our progressive cognitive optimization is more explicit and stable. Fifthly, not limited to the human parsing task, our progressive cognitive idea can be extended to other tasks such as scene parsing, object detection and human pose estimation.

In summary, compared to the existing methods, our work has three major contributions.

- We fuse the stacked hierarchical labeling structure into the fully convolutional framework, so that the whole framework associates progressive recognition of human parts with semantic understanding of the whole human.
- We propose a component-aware region convolution structure to transfer the prior shape and location information from the former layers to the latter layers thereby gradually segmenting the hierarchical human parts with the end-to-end progressive cognitive network.
- Experiments show that our framework can be embedded into any fully convolutional network to enhance the performance significantly.

Related Work

Semantic Segmentation Methods: Several popular semantic segmentation works are used to deal with the human parsing problem. In recent years, fully convolutional networks are popular configurations for semantic segmentation (Shelhamer, Long, and Darrell 2015). DeepLab proposed atrous convolution (a.k.a. dilated convolution) to dilate the receptive fields in fully convolutional networks, and obtained a high performance in human parsing task (Chen et al. 2017). In order to learn the multi-scale contextual information, the spatial pyramid structures are popular in the semantic segmentation networks (Chen et al. 2017; Zhao et al. 2017; Lin et al. 2017). By leveraging the multi-scale feature maps and refining the feature maps across scales, RefineNet achieved the best performance on PASCAL-Person-Part dataset.

Though the semantic segmentation methods have a powerful semantic understanding capacity, and have a high performance on human parsing task, these methods do not make full use of the characteristic of human parsing task. In this paper, we are trying to combine the semantic segmentation networks with our progressive cognitive structures, which are more similar to human vision. And the experiments show that the combination makes a significant improvement.

Human Parsing Methods: Contextualized convolutional neural network (a.k.a. Co-CNN) well integrates the cross-layer context, global image-level context, semantic edge context, cross super-pixel neighborhood context and within-super-pixel context into a unified framework (Liang et al. 2015). An attention-based model is proposed to softly weight the multi-scale features at each pixel location and deal with the multi-scale problem in the human parsing task (Chen et al. 2016). A novel self-supervised structure-sensitive learning approach is proposed to impose human pose structures on the parsing results (Gong et al. 2017). Human pose estimation and semantic part segmentation can be jointed to improve each other task with the external labels in natural multi-person images, in which the estimated pose provides object-level shape prior to regularize part segments and the part-level segments constrain the variation of pose locations (Xia et al. 2017).

Despite these human parsing methods have considered the characteristic of human parsing, they are still limited in various ways. Co-CNN and attention-to-scale model are more

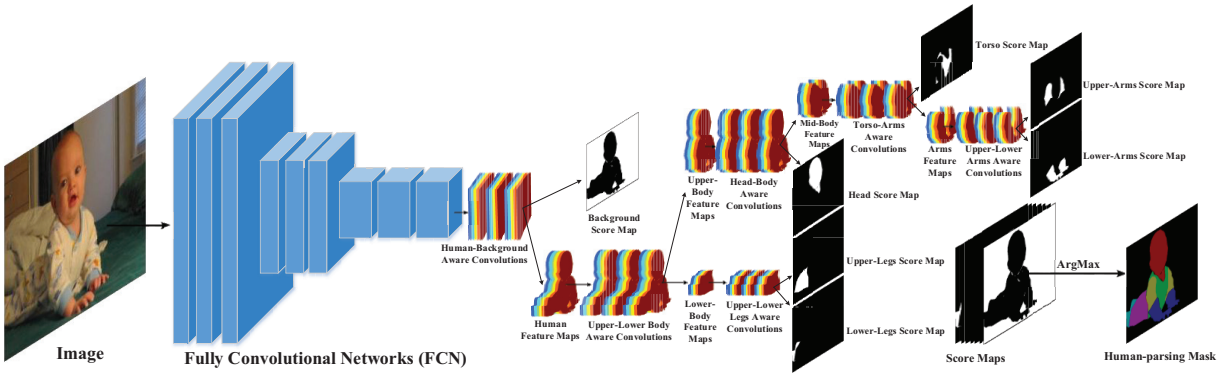

Figure 2: The framework of our progressive cognitive network. The input image is sent into the fully convolutional networks (FCN) to extract feature maps. And then the image-level feature maps are sent into the human-background aware convolutions to obtain a background score map and the human-level feature maps. The human-level feature maps are sent into the upper-lower body aware convolutions to obtain the upper-body feature maps and lower-body feature maps, respectively. The lower-body feature maps are sent into the upper-lower legs aware convolutions to obtain an upper-legs score map as well as a lower-legs score map. And the upper-body feature maps are sent into the head-body aware convolutions to obtain a head score map and the mid-body feature maps. The mid-body are then sent into the torso-arms aware convolutions to obtain a torso score map and the arms feature maps. The arms feature maps are sent into the upper-lower arms aware convolutions to obtain an upper-arms score map and a lower-arms score map. Note that the level division can be altered in various ways.

like the semantic segmentation methods because they only consider the contextual information and do not consider the human structure information. The jointed method makes full use of the human structure information, but it needs the extra labeled data and two tasks may cause the coupled optimization problems.

Hierarchical Models: The stacked hierarchical labeling method is proposed to segment the objects in an image from coarse to fine with the graphical model (Munoz, Bagnell, and Hebert 2010). Similarly, the cascade convolution networks are proposed to segment objects from hard to easy labels in an end-to-end way with the help of region convolution (Li et al. 2017b). The parsing methods are also combined with detection methods to deal with the human parsing problem, which can be also regarded as hierarchical structures because these methods detect the objects first and then parse the components with the prior information from the candidate bounding boxes (Li et al. 2017a; Xia et al. 2016a). The hierarchical human parsing methods have been proposed to parse the human with the help of And-Or graph, which is more intuitional and more similar to human vision (Xia et al. 2016b; Park, Nie, and Zhu 2017).

The stacked hierarchical labeling method and the cascade convolution method segment the object from coarse to fine, which is implicit and ambiguous. As a result, these methods are hard to train and rely on the parameters setting severely. The hierarchical methods based on detection rely on the detection networks, and once the candidate boxes lose the objects, the parsing task may be failed. Moreover, the detection feature maps and the human parsing networks may cause the coupled optimization problems. The hierarchical human parsing methods based on And-Or graph are more similar to our method, but these methods rely on the graph model and sometimes need to design features manually. As a result, they cannot be optimized in an end-to-end way.

Progressive Cognitive Networks (PCNet)

In this paper, we propose the PCNet for the human parsing task. Our PCNet parses the human step by step and utilizes the prior information from the former layers. Additionally, we develop the component-aware region convolutions to train our PCNet in an end-to-end way. We will illustrate the overall architecture of PCNet, component-aware region convolutions, and the implementation details as following.

Overall Architecture

As shown in Fig. 2, a raw image inputs to our our framework and the framework outputs the corresponding human parsing mask in an end-to-end way. First, the input image is sent into a fully convolutional network to extract the feature maps. It is worth to note that we can choose any full convolutional network. Here, we utilize the powerful semantic segmentation network pyramid scene parsing network (PSPNet) (Zhao et al. 2017) as our baseline model. The fully convolutional network outputs the image-level feature maps, which contain the rich category information at the image level. And then the image-level feature maps are sent into three human-background aware convolutions to obtain a background score map and the human-level feature maps. In this way, the image-level feature maps are transformed into the human-level feature maps, which are sent into the latter layers for further classification. Note that the human-level feature maps only contain the region of human without considering the background information, so it can focus on the human features and eliminate the distraction from the background. Additionally, the human-level feature maps can

provide the shape information and the location information for the latter layers. And then the human-level feature maps are sent into the upper-lower body aware convolutions to obtain the upper-body feature maps and the lower-body feature maps, respectively. The upper-body feature maps can only pay attention to the upper-body location. And the low-body feature maps possess the similar properties. Next, the lower-body feature maps are sent into the upper-lower legs aware convolutions to obtain an upper-legs score map as well as a lower-legs score map. These two score maps have a higher performance than that of the methods which segment the legs from the image-level feature maps. This is because that the feature maps in our framework have less distraction from the irrelevant regions. Meanwhile, the upper-body feature maps are sent into the head-body aware convolutions to obtain a head score map and the mid-body feature maps. Similarly, the head score map can have less distraction for better performance. And the mid-body feature maps are then sent into the torso-arms aware convolutions to obtain a torso score map and the arms feature maps. The arms feature maps are sent into the upper-lower arms aware convolutions to obtain an upper-arms score map and a lower-arms score map. Finally, the largest responses of the score maps are set as the corresponding categories for all locations. Note that we only illustrate a kind of level division for human part relation, and the level division can be altered in various ways based on the definitions of the users. With the powerful progressive cognitive structures, we can obtain the more explicit feature maps, which ignore distraction from the irrelevant regions. In this way, the latter layers pay more attention to the relevant regions thereby relieving the stress for multi-category classification, so that we can obtain the more accurate score maps and human parsing results.

Component-Aware Region Convolution (CAR-Conv)

We illustrate CAR-Conv on image-level in Fig. 3. The image-level feature maps from fully convolution network are sent into three convolutions to obtain further feature maps. Similar to ResNet (He et al. 2016), the kernel size of the first convolution is 1×1, the number of output channels is 256, and the stride is 1. For the second convolution, the kernel size is 3×3, the number of output channels is 256, and the stride is 1. For the third convolution, the kernel size is 1×1, the number of output channels is 512, and the stride is 1. The batch normalization (Ioffe and Szegedy 2015) and rectified linear units (Krizhevsky, Sutskever, and Hinton 2012) are following the convolutions. We sum up the convolutional feature maps and the FCN feature maps as the further feature maps. A 1×1 convolution with two channels output is operated on the further feature maps to obtain the human score maps. The opacities (a.k.a. alpha mattes) of the human and the background can be obtained through the sigmoid function on the human score maps. Specially, inspired by the edge detection work (Xie and Tu 2015), we compute the weighted cross-entropy loss at every pixel on the alpha

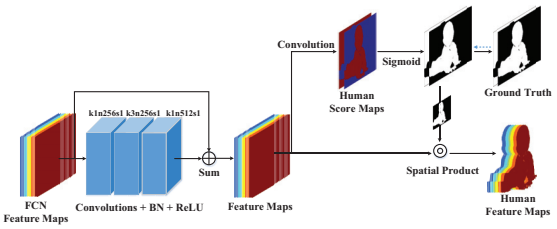

Figure 3: An illustration of component-aware region convolutions. The image-level feature maps from fully convolution networks are sent into three convolutions to obtain further feature maps. And then the feature maps are sent to obtain human score maps. We can obtain the opacities (a.k.a. alpha mattes) of the human and the background through a sigmoid function on the human score maps. A weighted cross-entropy loss is performed on the alpha mattes. Finally, the spatial product is operated on the feature maps and the human alpha matte to obtain the human feature maps.

mattes with respect to the ground truth as

$$L_c\left(X_i;W\right)=\begin{cases}\alpha\log\left(1-P\left(X_i;W\right)\right) & if\ y_i=0\\ \beta\log\left(P\left(X_i;W\right)\right) & if\ y_i=1\\ 0 & otherwise,\end{cases}$$

in which

$$\alpha=\lambda\frac{|Y^+|}{|Y^+|+|Y^-|},\ \beta=\frac{|Y^-|}{|Y^+|+|Y^-|},$$

and c represents each components. The hyper-parameter λ is used to balance the positive and negative samples.

Next, the spatial product is operated on the feature maps and the human alpha matte across channels, so that we can obtain the human feature maps. In this way, only the human region of the feature maps are nonzero. And the background region can be regarded as an irrelevant region for the latter layers because the convolutions on this region become invalid. And the latter layers can only operate on the relevant region. The similar operations are performed on {human, upper-body, lower-body, mid-body, arms} regions.

Implementation Details

Loss Function Similar to the general semantic segmentation methods, the proposed method uses the softmax loss function on the score maps, which is the sum of per-pixel multinomial logistic loss as follows,

$$L_{soft\max}\left(I\right)=-\frac{1}{MN}\sum_{i=1}^{MN}\sum_{k=0}^{K}1\left\{y_i=k\right\}\log\left(p_{i,k}\right),$$

where M and N is height and width of the input image and K is the number of the categories.

Specially, we utilize the weighted cross-entropy loss at every pixel on each alpha matte as follows,

$$L_{mattes}=\sum_{i=1}^{MN}\sum_{c\in components}L_c\left(X_i;W\right).$$

Additionally, similar to PSPNet (Zhao et al. 2017), we add an auxiliary loss L_{aux} at the end of the fully convolutional network for semantic segmentation.

Finally, the loss function of our whole framework is

$$L = L_{soft\max} + \lambda_1 L_{mattes} + \lambda_2 L_{aux}.$$

For our experiments on PASCAL-Person-Part dataset, we fix the hyper-parameters $\lambda = 1.1$, $\lambda_1 = 1e-5$, and $\lambda_2 = 0.4$.

Training and Inference In training process, we train the fully convolutional network for semantic segmentation first. And then we remove the last layer of the fully convolutional network, and replace it with our PCNet. Next, we conduct the end-to-end training of entire network integrally. We utilize the stochastic gradient descent (SGD) solver with batch size 8, momentum 0.9 and weight decay 0.0005. Inspired by the semantic segmentation optimization (Chen et al. 2017; Zhao et al. 2017), we use the "poly" learning rate policy $\left(1 - \frac{iter}{\max iter}\right)^{power}$. We set the base learning rate as 0.001 and the power as 0.9. As for the input image size, we resize it to 473×473. For data augmentation, we add random gaussian blur to the images and rotate the images in random degrees from -20 to 20. We setup our model training experiments on Caffe platform (Jia et al. 2014). Due to the limitation of physical memory on GPU cards, we modify the Caffe version to make it support batch normalization on data gathered from multiple GPUs based on OpenMPI. For the inference process, we only test on the single scale due to the low speed for multi-scale inference process utilized in semantic segmentation (Chen et al. 2017; Zhao et al. 2017). All of our experiments are implemented on a system of Core E5-2660 @2.60GHz CPU and four NVIDIA GeForce GTX TITAN X GPUs with 12GB memory.

Experiments

Data

We evaluate our algorithm on the public human parsing dataset, PASCAL-Person-Part (Chen et al. 2014), which contains a large number of part segment annotations for PASCAL person images with various poses and scales. There are 7 types of annotation in this dataset, i.e. background, head, torso, upper arm, lower arm, upper leg and lower leg. We only use the images containing human for training (1716 images) and validation (1817 images). In this dataset, we set the level division as shown in Fig. 1 and Fig. 2. The human level mask and feature maps are set as level 1, {upper-body, lower-body} level mask and feature maps are set as level 2, {head, mid-body, upper-leg, lower-leg} level mask and feature maps are set as level 3, {head, torso, arm, upper-leg, lower-leg} level mask and feature maps are set as level 4, {head, torso, upper-arm, lower-arm, upper-leg, lower-leg} level mask and feature maps are set as level 5.

Quantitative Evaluation

Ablation Estimation We utilize the pyramid scene parsing network (PSPNet) (Zhao et al. 2017) as our baseline fully convolutional network for feature extraction. In Tab. 1, we evaluate the mean pixel Intersection-over-Union (mIoU) of our PCNet with PSPNet at various depths for ablation experiment. We reimplement the PSPNet based on ResNet-50 and ResNet-101 (He et al. 2016), and name it as PSPNet-50 and PSPNet-101, respectively. We add our progressive cognitive structures at the end of fully convolutional networks, i.e., PSPNet-50 and PSPNet-101. Since the depths of the whole networks would be increased, we name the novel progressive cognitive networks as PCNet-75 and PCNet-126, respectively. Furthermore, in order to ablate the effect of depths, we add the extra layers to the end of PSPNet corresponding to the depths of PCNet, and name them as PSPNet-75 and PSPNet-126, respectively.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
PSPNet-50	80.59	60.12	44.70	44.96	37.85	34.54	93.87	56.66
PSPNet-101	83.82	64.56	49.40	49.04	41.48	39.37	94.44	60.30
PSPNet-75	80.73	60.18	45.26	45.53	38.62	36.71	93.98	57.29
PSPNet-126	83.88	64.60	50.04	49.48	41.77	40.04	94.26	60.58
PCNet-75	83.91	64.78	51.81	50.82	42.74	40.59	94.58	61.31
PCNet-126	**86.81**	**69.06**	**55.35**	**55.27**	**50.21**	**48.54**	**96.07**	**65.90**

Table 1: Mean Pixel Intersection-over-Union (mIoU) (%) of Human Semantic Part Segmentation on PASCAL-Person-Part. We compare our PCNet with PSPNet at various depths for ablation experiment.

As shown in Tab. 1, our PCNet-126 model outperforms all the baseline models for all metrics. Comparing our PCNet-75 and PCNet-126 with PSPNet-50 and PSPNet-101, we can observe that the progressive cognitive structures improve 4.65% and 5.6% on the baseline models, respectively. It can be declared that our PCNet is effective and reasonable. Furthermore, we compare our PCNet with PSPNet in the same layer depths. PCNets (PCNet-75 and PCNet-126) improve the two baseline models for 4.65% and 5.6%, while the deeper models (PSPNet-75 and PSPNet-126) only improve for 0.63% and 0.28%. Note that our PCNet-75 have better performance than that of PSPNet-101 and PSPNet-126 for all metrics. This proves that the significant improvements rely on our proposed structure and loss function, rather than the network depths. And our structures are more effective than the deeper structures. Finally, our PCNet-126 achieves the best performance for 65.90%. We believe that our progressive cognitive network can be combined with more powerful networks such as ResNeXt (Xie et al. 2017), DenseNet (Huang, Liu, and Weinberger 2017), RefineNet (Lin et al. 2017) and SENet (Hu, Shen, and Sun 2017) for better human parsing performance.

Comparison Estimation We compare our proposed PCNet with several state-of-the-art human parsing methods in Tab. 2. We evaluate the part segmentation results in terms of mIoU as the semantic segmentation tasks.

As shown in Tab. 2, our PCNet outperforms these three human parsing methods for all metrics. Though our PCNet does not utilize the attention-to-scale mechanism like Attention (Chen et al. 2016), or the detected bounding boxes mechanism like HAZN (Xia et al. 2016a) and Joint (Xia et al. 2017), nor utilize the extra dataset like Joint, PCNet has

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
Attention	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HAZN	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
Joint	85.50	67.87	54.72	54.30	48.25	44.76	95.32	64.39
Ours (PCNet-126)	**86.81**	**69.06**	**55.35**	**55.27**	**50.21**	**48.54**	**96.07**	**65.90**

Table 2: Mean Pixel Intersection-over-Union (mIOU) (%) of Human Semantic Part Segmentation on PASCAL-Person-Part. We compare our PCNet with the state of the art human parsing methods.

the similar properties and common advantages like these excellent human parsing methods. Firstly, our CAR-Conv at each level can be regarded as an attention mechanism, i.e., we ignore the irrelevant information and only pay attention to the relevant information. Benefit from the attention mechanism, the latter layers of PCNet cannot be disturbed by the irrelevant information so that the stress of the latter classifiers can be relieved. Secondly, similar to the methods based on detection bounding boxes, our progressive cognitive parsing can also be regarded as a hierarchical detected process, i.e., we detect the foreground human in pixel-level first, and then parse the components with the candidate human feature maps. Benefit from the detection-like mechanism, our PCNet has a better capacity to locate the small objects, which may be failed to parse in the semantic segmentation networks. Thirdly, similar to Joint which utilizes the pose information for human parsing, the latter layers possess the shape information of the foreground objects with the help of CAR-Conv. Benefit from the shape information, PCNet has a better capacity to classify each component and a better performance on human parsing task than that of other three human parsing methods.

Qualitative Evaluation

Score Maps Estimation As shown in Fig. 4, we output the qualitative results of the alpha mattes for HAZN (Xia et al. 2016a), PSPNet-101 (Zhao et al. 2017), and our PCNet-126, i.e., the transformed score maps by the sigmoid function. It can be observed that the alpha mattes of HAZN and PSPNet-101 are fuzzy and ambiguous, while the alpha mattes of our PCNet-126 are more precise for visualization. It is because that loss functions of HAZN and PSPNet are only calculated on the max likelihood for the ground-truth category and ignore the non-max probability of other categories. In contrast, our PCNet model calculates the cost for all score maps. As a result, the former two methods are easy to be disturbed by other components and hard to parse the human parts precisely. In contrast, PCNet can be less disturbed by the irrelevant regions because our proposed CAR-Conv only pay attention to the relevant regions. And we find that eliminating the irrelevant region can improve the human parsing performance significantly.

Visual Comparison We show qualitative comparison among three human parsing methods in Fig. 5. From the first three rows (a)-(c), we can draw a conclusion that PCNet-126 has better performance in locating the small objects. Though HAZN utilizes Faster R-CNN to detect the human first, it fails to locate the objects in Fig. 5(a), so it cannot continue

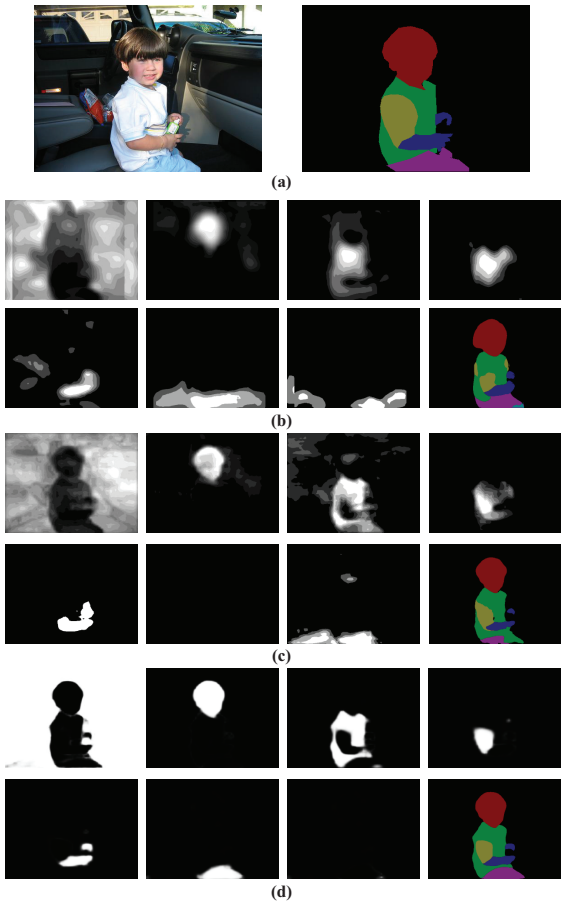

Figure 4: Qualitative results of the alpha mattes for three methods, i.e., the score maps through sigmoid transformation. (a) Input image and ground-truth mask. (b) Seven alpha mattes and the mask output by HAZN. (c) Seven alpha mattes and the mask output by PSPNet-101. (d) Seven alpha mattes and the mask output by our PCNet-126.

to segment human parts without human location. In contrast, PSPNet-101 can locate the objects in Fig. 5(a) but it cannot parse a large number of human parts. Due to the progressive cognitive structures, our PCNet-126 outperforms these two methods in parsing small-scale human tasks. PCNet-126 not only locates the small-scale human precisely, but also has a better performance in parsing the smaller human parts. The similar situations can be seen in Fig. 5(b) and Fig. 5(c). Moreover, PCNet-126 can also work well in large-scale human parsing tasks, which can be seen in Fig. 5(d) and Fig. 5(e). Large-scale human parsing is easier to work well. However, HAZN and PSPNet-101 lose a large number of detailed segmentation in this situation and classify the pixels into false categories. For instance, HAZN fails to segment the lower-arm in Fig. 5(d) and Fig. 5(e), and PSPNet-101 fails to segment the upper-arm and torso in Fig. 5(d) and Fig. 5(e). Compared to these two methods, our PCNet-126 has better performance in these details. Complicated human appearance and occlusion are common in human parsing

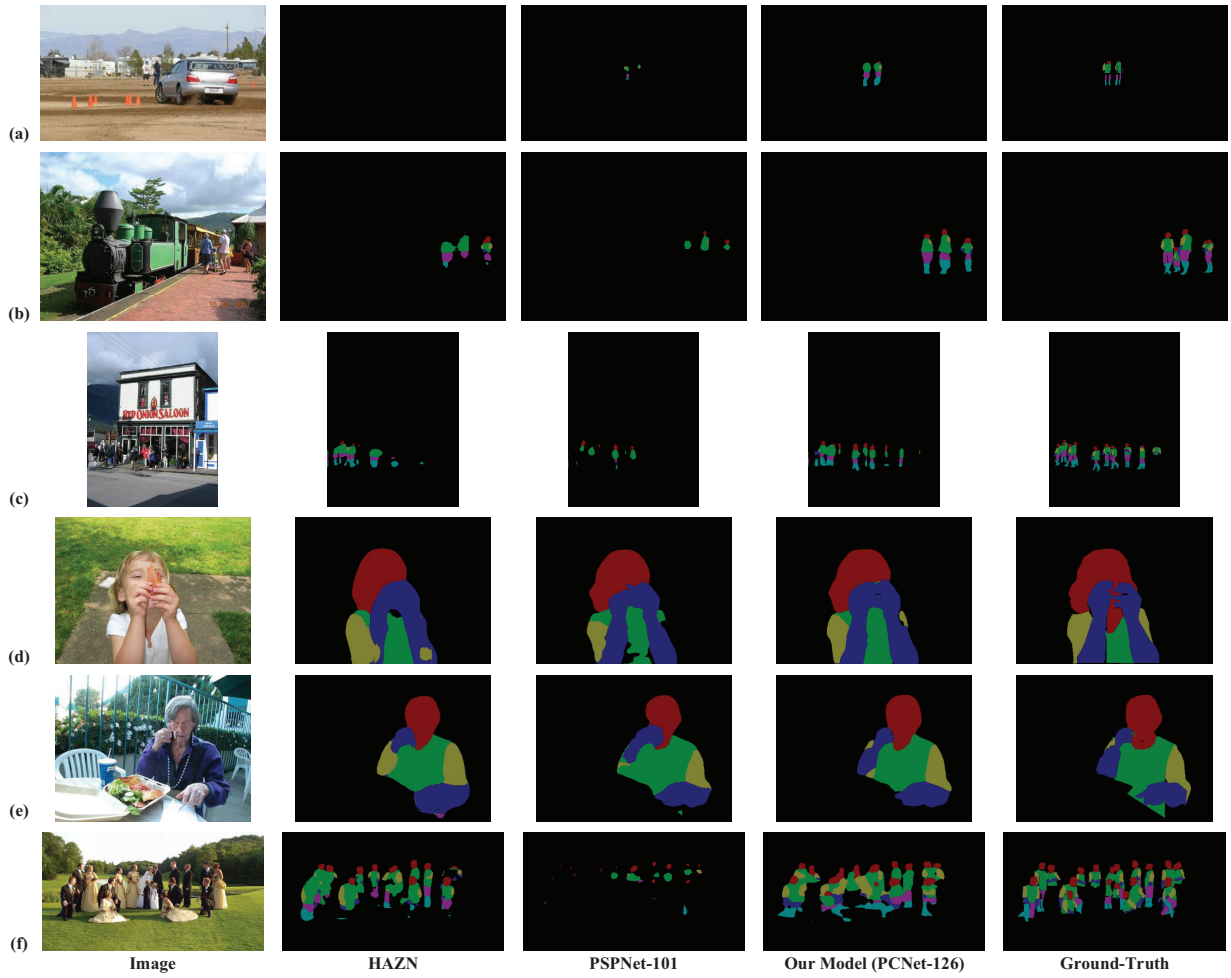

Figure 5: Visual comparison among three human parsing methods.

task, and Fig. 5(f) is an example of this situation. The semantic segmentation method PSPNet-101 fails in this image and classifies a large number of pixel as the background category. The human parsing methods based on detection have better performance because the candidate bounding boxes relieve the stress for human parsing, but it is still hard to segment the human. We can see that HAZN mixes up many humans together because the candidate proposals are fused together. In contrast, our PCNet-126 has better performance than these two methods. PCNet model segments the objects hierarchically so that our model does not confuse the different objects together and has a more precise location and classification capacity.

Failure Modes Similar to other human parsing methods, our method may fail when it comes to small human parts and complicated human occlusions. For examples, PCNet fails to segment the small upper-arm in Fig. 5(a) and the complicated occlusions in Fig. 5(f).

Conclusion

We proposed Progressive Cognitive Networks (PCNet) for human parsing. Our approach employs the proposed component-aware region convolution (CAR-Conv) to conduct the end-to-end training on the entire networks. With the hierarchical structures, the hierarchical prior information can be inherited to the latter layers. Our experiments show that our method improves the performance of the baseline models significantly.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant 61375035, 61772527).

References

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1979–1986.

Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; and Yuille, A. L. 2016. Attention to scale: Scale-aware semantic image segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3640–3649.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*.

Epshtein, B.; Lifshitz, I.; and Ullman, S. 2008. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences* 105(38):14298–14303.

Fu, J.; Zheng, H.; and Tao, M. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2017*. IEEE.

Gong, K.; Liang, X.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Computer Vision and Pattern Recognition (CVPR), 2017*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-excitation networks. *CoRR* abs/1709.01507.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2017*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R. B.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60:84–90.

Li, J.; Zhao, J.; Wei, Y.; Lang, C.; Li, Y.; and Feng, J. 2017a. Towards real world human parsing: Multiple-human parsing in the wild. *CoRR* abs/1705.07206.

Li, X.; Liu, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2017b. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Computer Vision and Pattern Recognition (CVPR), 2017*. IEEE.

Liang, X.; Xu, C.; Shen, X.; Yang, J.; Tang, J.; Lin, L.; and Yan, S. 2015. Human parsing with contextualized convolutional neural network. *2015 IEEE International Conference on Computer Vision (ICCV)* 1386–1394.

Lin, G.; Milan, A.; Shen, C.; and Reid, I. D. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2017*. IEEE.

Liu, S.; Feng, J.; Song, Z.; Zhang, T.; Lu, H.; Xu, C.; and Yan, S. 2012. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, 619–628. ACM.

Lu, Y.; Boukharouba, K.; Boonaert, J.; Fleury, A.; and Lecoeuche, S. 2014. Application of an incremental svm algorithm for on-line human recognition from video surveillance using texture and color features. *Neurocomputing* 126:132–140.

Munoz, D.; Bagnell, J. A.; and Hebert, M. 2010. Stacked hierarchical labeling. In *ECCV*.

Park, S.; Nie, X.; and Zhu, S.-C. 2017. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE transactions on pattern analysis and machine intelligence*.

Shelhamer, E.; Long, J.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39:640–651.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An approach to pose-based action recognition. *2013 IEEE Conference on Computer Vision and Pattern Recognition* 915–922.

Xia, F.; Wang, P.; Chen, L.-C.; and Yuille, A. L. 2016a. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*.

Xia, F.; Zhu, J.; Wang, P.; and Yuille, A. L. 2016b. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI*.

Xia, F.; Wang, P.; Chen, X.; and Yuille, A. 2017. Joint multi-person pose estimation and semantic part segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2017*. IEEE.

Xie, S., and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017*.

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2012. Parsing clothing in fashion photographs. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 3570–3577. IEEE.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Computer Vision and Pattern Recognition (CVPR), 2017*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3586–3593.