Blended Grammar Network for Human Parsing

Xiaomei Zhang[1,2], Yingying Chen[1,2,3], Bingke Zhu[1,2], Jinqiao Wang[1,2,4], and Ming Tang[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] ObjectEye Inc., Beijing, China
[4] NEXWISE Co., Ltd, Guangzhou, China
{xiaomei.zhang, yingying.chen, bingke.zhu, jqwang, tangm}@nlpr.ia.ac.cn

Abstract. Although human parsing has made great progress, it still faces a challenge, i.e., how to extract the whole foreground from similar or cluttered scenes effectively. In this paper, we propose a Blended Grammar Network (BGNet), to deal with the challenge. BGNet exploits the inherent hierarchical structure of a human body and the relationship of different human parts by means of grammar rules in both cascaded and paralleled manner. In this way, conspicuous parts, which are easily distinguished from the background, can amend the segmentation of inconspicuous ones, improving the foreground extraction. We also design a Part-aware Convolutional Recurrent Neural Network (PCRNN) to pass messages which are generated by grammar rules. To train PCRNNs effectively, we present a blended grammar loss to supervise the training of PCRNNs. We conduct extensive experiments to evaluate BGNet on PASCAL-Person-Part, LIP, and PPSS datasets. BGNet obtains state-of-the-art performance on these human parsing datasets.

1 Introduction

Human parsing aims to segment human images into multiple human parts of fine-grained semantics and benefits a detailed understanding of images. It has many applications, such as person re-identification [8], human behavior analysis [37], clothing style recognition and retrieval [40], clothing category classification [35], to name a few. With the rapid development of electronic commerce and online shopping, human parsing has been attracting much attention [34, 38, 39, 14, 13, 17, 19, 20, 45, 28, 23, 41]. However, it is still a challenging task and faces the difficulty in accurate extraction of the whole foreground from similar, cluttered scenes, and blurred images. Fig. 1 provides two examples where there exist similar appearances to the foreground in the background or cluttered scenes.

In order to deal with the problem, HAZA [38] and Joint [39] introduced human detection into their human parsing algorithms to abate the interference of the similar or cluttered background. Nevertheless, their methods strongly depend on the detectors, and still need to extract the whole body of a human from

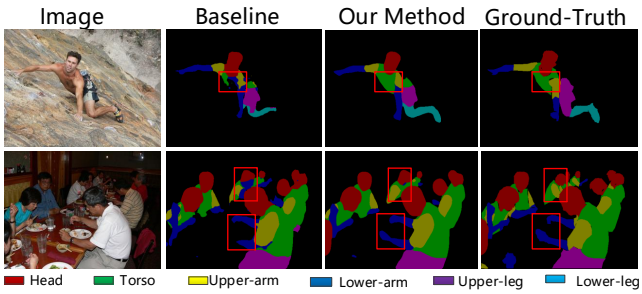

Fig. 1. Examples of the challenge in human parsing. The original images and ground-truth come from PASCAL-Person-Part dataset [5]. There are similar appearances to the foreground in the background. Second column: the baseline fails to extract the whole foreground. Third column: our method has better performance in extracting foreground from similar or cluttered scenes.

the cluttered background around the bounding boxes of detection, even if the detection is correct. If the detectors fail to locate a human body, that body will be treated as the cluttered background, or vice versa. This may in turn increase the background interference for human parsing. Additionally, MuLA [29] and LIP [18] combined human pose estimation and human parsing to improve the foreground extraction. Nevertheless, human pose estimation does not focus on assigning a unique label to a pixel. The assistance against the cluttered background by means of the human pose estimation is limited.

In this paper, we propose a Blended Grammar Network (BGNet) to deal with the above problem by exploiting the inherent hierarchical structure of a human body and the relationship of different human parts. We design a Part-aware Convolutional Recurrent Neural Network (PCRNN) to model grammar rules in our BGNet, which can adaptively extract features to improve accuracy. A blended grammar loss is designed to supervise the training of PCRNNs.

Our BGNet, which is developed to exploit the inherent hierarchical structure of a human body and capture the relationship of different human parts, is based on two insights. One comes from the visual psychology which claimed that humans often pay attention to conspicuous parts first, such as head, and then tend to other parts [11, 33] when observing a person. And the other is reported by [44], which verified that conspicuous parts, such as head, are relatively more separable and can more easily be distinguished from the background. According to these insights, the grammar rules of BGNet leverage conspicuous parts (e.g., torso, head) distinguished from the background easily to amend the segmentation of inconspicuous ones (e.g., low-leg, low-arm), improving the foreground extraction. And we use grammar rules to progressively explore the inherent hierarchical structure of a human body in both cascaded and paralleled manner, as shown in Fig. 2. Specifically, a latter grammar rule inherits the outputs of its former rule as one of its inputs in a cascaded manner, and two grammar rules take the outputs of the same grammar rule as their one input in a paralleled

manner. For example, $Rule_1^c$, $Rule_2^h$ and $Rule_3^h$ connect in a cascaded manner, meanwhile, $Rule_2^h$ and $Rule_4^v$ connect in parallel. In this way, BGNet combines the advantages of both cascaded and paralleled architectures, thus further improving the accuracy of foreground extraction.

We propose PCRNN to represent grammar rules and model messages passing, which can adaptively extract features to improve accuracy. PCRNN consists of two stages to effectively generate reliable features of parts, as shown in Fig. 4. The first stage of grammar rules (e.g., $Rule_2^h$) uses two inputs, the results of the former rule (e.g., the outputs of the $Rule_1^c$) and features of its corresponding part (e.g., the features of upper-arm), to model the relationship among human parts. The second stage extracts features by adaptively selecting the results (S_r) of the first stage and generates features of its corresponding parts (e.g., head, torso and upper-arm in $Rule_2^h$). To train PCRNNs effectively, we design a blended grammar loss which uses a less-to-more manner with increasing parts in PCRNNs. The supervision of each PCRNN is the ground truth of its corresponding human parts.

Extensive experiments show that our network achieves new state-of-the-art results consistently on three public benchmarks, PASCAL-Person-Part [5], LIP [14] and PPSS [27]. Our method outperforms the best competitors by 3.08%, 2.42%, and 4.59% on PASCAL-Person-Part, LIP, and PPSS in terms of mIoU, respectively. In summary, our contributions are in three folds:

1. We propose a novel Blended Grammar Network (BGNet) to improve the extraction accuracy of the foreground out of the cluttered background in both cascaded and paralleled manner. And grammar rules of BGNet use the conspicuous parts to amend the inconspicuous ones, improving the foreground extraction.
2. We design a Part-aware Convolutional Recurrent Neural Network (PCRNN) to pass messages across BGNet and a novel deep blended grammar loss to supervise the training of PCRNNs. With the grammar loss, the PCRNN effectively represents the relationship of human parts.
3. The proposed BGNet achieves new state-of-the-art results consistently on three public human parsing benchmarks, including PASCAL-Person-Part, LIP, and PPSS.

2 Related Work

Human Parsing. Many research efforts have been devoted to human parsing [38, 39, 14, 17, 19, 20, 45, 28, 23]. Chen *et al.* [4] proposed an attention mechanism that learns to weight the multi-scale features at each pixel location softly. Xia *et al.* [38] proposed HAZN for object part parsing, which adapted to the local scales of objects and parts by detection methods. Human pose estimation and semantic part segmentation were two complementary tasks [18], in which the former provided an object-level shape prior to regularize part segments while the latter constrained the variation of pose location. Ke *et al.* [12] proposed Graphonomy, which incorporated hierarchical graph transfer learning upon the

conventional parsing network to predict all labels. Wang *et al.* [36] combined neural networks with the compositional hierarchy of human bodies for efficient and complete human parsing. Different from the above methods, our BGNet exploits the inherent hierarchical structure of a human body and the relationship of different human parts by means of grammar rules.

Grammar Model. Grammar models are powerful tools for modeling high-level human knowledge in some tasks of human pose estimation [7], Clothing Category Classification [35] and so on. Qi *et al.* [31] presented the stochastic grammar to predict human activities. Grammar models allow an expert injects domain-specific knowledge into the algorithms, avoiding local ambiguities [1, 30]. Wang *et al.* [35] designed the fashion grammar to capture the relations among a few joints and focused on the local context through short grammar rule, in which grammar layers connected in parallel. Different from the above methods, first, our BGNet uses a blended architecture of both parallelled and cascaded connections among grammar rules, and latter rules inherit and refine results of former rules. Second, our grammar rules exploit the relations of different human parts to capture the local and global context because human parsing needs not only the local context but also the global context. Third, our PCRNN passes messages unidirectionally. The purpose of our PCRNNs is to carry out the message passing from conspicuous parts (e.g., torso, head) to the segmentation of inconspicuous parts (e.g., low-leg).

Supervision Mechanism. Various types of supervision approaches have been popular in the tasks of human parsing [18, 45], semantic segmentation [43, 21, 42, 10, 9] and human pose estimation [16]. Liang *et al.* [18] introduced a structure-sensitive loss to evaluate the quality of the predicted parsing results from a joint structure perspective. Zhu *et al.* [45] proposed a component-aware region convolution structure to segment human parts with hierarchical supervision. Zhao *et al.* [43] designed an auxiliary loss in the backbone. Zhao *et al.* [42] developed an image cascade network (ICNet) with cascade label guidance. Ke *et al.* [16] proposed multi-scale supervision to strengthen contextual feature learning in matching body keypoints by combining feature heatmaps across scales. Different from them, our deep blended grammar loss uses the less-to-more manner with an increasing number of parts in rules.

3 Blended Grammar Network

The overall architecture of BGNet is shown in Fig. 2. We the feature extractor to generate original features. Then, a convolutional layer is applied to features to generate the corresponding coarse predictions of human parts, which are learned under the supervision from the ground-truth segmentation. The coarse prediction is sent into grammar rules to exploit the inherent hierarchical structure of a human body and the relationship of different human parts. We design PCRNN to model grammar rules, and every grammar rule is represented by one PCRNN. Finally, the outputs of the feature extractor and PCRNN are concatenated to obtain the final fine prediction.

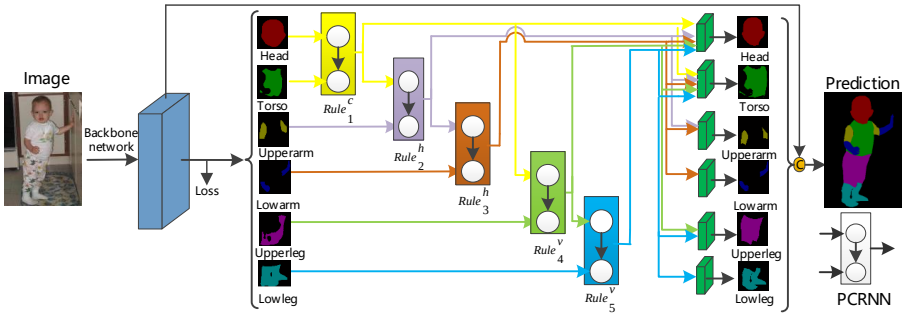

Fig. 2. Overview of the proposed Blended Grammar Network (BGNet). An input image goes through the backbone network to generate its original features and the corresponding coarse predictions of human parts. Then grammar rules are applied to exploit the inherent hierarchical structure of a human body and the relationship of different human parts. Outputs of grammar rules and original features are concatenated to get the final fine prediction. Every grammar rule is represented by one PCRNN. © indicates the concatenation operation.

3.1 Grammar

Dependency grammars [30] have been widely used in natural language processing for syntactic parsing. It has a root node S and set of n other nodes $\{A_1, ..., A_n\}$ with rules like

$$\begin{aligned} &S \rightarrow A_1|A_2 \cdots |A_n, \\ &A_i \rightarrow a_i|a_iA_j|A_ja_i, \forall i = 1, 2..., n, j{\neq}i, \end{aligned} \tag{1}$$

where "|" denotes "or" function, "$\rightarrow$" denotes the flow of information, root node S can transit to any other nodes once, and then each node A_i can terminate as a_i or transit to another node A_j to the left or right side. It is seen that S can also indirectly transmit to any other node through an intermediate node.

Our Grammar. Inspired by dependency grammars, we define a novel progressive grammar model of grammar rules, to segment human parts, as shown in Fig. 2. Grammar rules of BGNet set conspicuous parts as root node to amend the segmentation of inconspicuous parts by means of their relationship, improving the extraction accuracy of the foreground out of the cluttered background. The more conspicuous parts are used in more rules. For example, head is the most conspicuous part [44], thus head is integrated into all grammar rules.

We consider a total of six human parts to constitute grammar rules which are tried during preparing the paper, as shown in Tabel 1. According to the performance and human anatomical and anthropomorphic constraints, we use five grammar rules, $Rule_1^c$, $Rule_2^h$, $Rule_3^h$, $Rule_4^v$ and $Rule_5^v$, to progressively exploit the inherent hierarchical structure of a human body. These five grammar rules describe central, horizontal, and vertical relation, respectively. Due to different datasets with different kinds of labels of human parts, the grammar rules may be adjusted slightly. The five rules are:

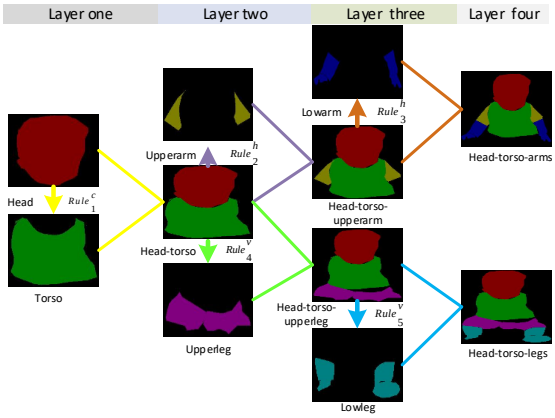

Fig. 3. Message across our BGNet. Lines and arrows with different colors and their connective human parts represent different grammar rules.

$$
\begin{aligned}
&Rule_1^c : head \rightarrow torso,\\
&Rule_2^h : head \rightarrow torso \rightarrow upperarm,\\
&Rule_3^h : head \rightarrow torso \rightarrow upperarm \rightarrow lowarm,\\
&Rule_4^v : head \rightarrow torso \rightarrow upperleg,\\
&Rule_5^v : head \rightarrow torso \rightarrow upperleg \rightarrow lowleg.
\end{aligned} \tag{2}
$$

There is a progressive relation among the five grammar rules. For example, $Rule_2^h$ is the growth of $Rule_1^c$, and $Rule_3^h$ is the growth of $Rule_2^h$. Thus, Eq.2 can be represented by the following expression,

$$
\begin{aligned}
&Rule_1^c : head \rightarrow torso,\\
&Rule_2^h : Rule_1^c \rightarrow upperarm,\\
&Rule_3^h : Rule_2^h \rightarrow lowarm,\\
&Rule_4^v : Rule_1^c \rightarrow upperleg,\\
&Rule_5^v : Rule_4^v \rightarrow lowleg.
\end{aligned} \tag{3}
$$

3.2 Blended Grammar Network Structure

As shown in Fig. 2, BGNet is a blended architecture among grammar rules. It combines the advantages of both cascaded and paralleled architectures. i.e., the latter rule inherits the results of its former rule in a cascaded way, meanwhile, $Rule_2^h$ and $Rule_4^v$ all take the results of $Rule_1^c$ as inputs in parallel because they have the tight relationship with $Rule_1^c$. In the cascaded way, the former results provide valuable context to the latter layers, meanwhile, the latter layers take

features of a corresponding part as input to further refine the results of the former layer. In the paralleled way, BGNet can make full use of the relationship of human parts.

Fig. 3 provides a more vivid representation of the message passing in the blended architecture. $Rule_1^c$ is on the first layer, which exploits the relation between the head and torso and generates the features of head-torso in the middle of the second layer. In the next layer, both $Rule_2^h$ and $Rule_4^v$ leverage the features of head-torso to further explore the relation of upper-arm and upper-leg, respectively. Similarly, $Rule_3^h$ and $Rule_5^v$ repeat the growth by inheriting results from $Rule_2^h$ and $Rule_4^v$, respectively.

Because every grammar rule generates features of its corresponding human parts, thus, different parts have the different number of features. In order to obtain the same number of features of each part, we concatenate the features generated by all relevant grammar rules of a part. Then we apply a convolutional layer on the concatenating features to generate its final prediction. These are

$$\begin{aligned} M^p &= concat(\{M_r^p\}_{r=1}^5), \\ P^p &= \mathcal{W}^p * M^p + b^p, \end{aligned} \tag{4}$$

where M_r^p denotes the feature of part p of the rule r, $concat(\cdot)$ denotes the concatenating function, M^p denote all features of part p and P^p denotes the feature of part p, $\mathcal{W}^p$ refers to weights, b^p refers to bias.

Note that the predictions of all parts and the outputs of the baseline F are concatenated, that is,

$$F^b = concat(\{P^p\}_{p=1}^n, F), \tag{5}$$

where n denotes the number of parts.

Different Architectures of Grammar Network. To prove the capability of our blended architecture, we introduce another architecture named Paralleled Grammar Network (PGNet), in which grammar rules connect in parallel. Two architectures of networks take the outputs of the baseline as their inputs, and use PCRNNs to pass messages. In PGNet, there are also five grammar rules and 5 PCRNNs to pass messages. However, the paralleled connection among grammar rules are designed in PGNet, and the results of the former rules can not be inherited by the latter rules. The results show that our blended architecture has better performance on human parsing than the paralleled architecture.

3.3 Part-aware CRNN

We design Part-aware CRNN (PCRNN) to pass messages which are generated by grammar rules in BGNet, as shown in Fig. 4. Each grammar rule of BGNet is represented by one PCRNN. PCRNN has the capacity to model the relationship among human parts. What is more, PCRNN can preserve the spatial semantic information via the convolutional operations.

PCRNN has two inputs, including the inherited results generated by former rule and features of its corresponding part. It consists of two stages. The first

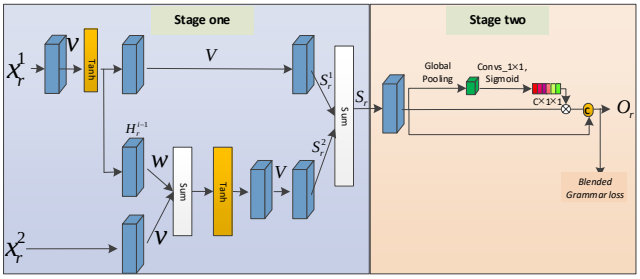

Fig. 4. Architecture of the proposed PCRNN. © indicates the concatenation operation.

stage models the relationship among human parts. The second stage extracts features by adaptively selecting the results of the first stage. And at the end of it, a convolutional layer is applied to generate the corresponding coarse predictions of human parts under the supervision of the blended grammar loss. Then we concatenate all features of one part. The functions are

$$\begin{aligned} H_r^i &= tanh(\mathcal{V} * \chi_r^i + \mathcal{W} * H_r^{i-1} + b), \\ S_r^i &= V * \mathcal{H}_r^i + c, \\ S_r &= \mathcal{W}_r * (S_r^1 + S_r^2) + b_r, \\ O_r &= concat(sigmoid(gp(S_r)) \circ S_r, S_r), \\ M_r^p &= S(O_r), \end{aligned} \tag{6}$$

where χ_r^i denotes the input i of the rule r, i refers to 1,2 and r is from 1 to 5, H_r^{i-1} denote the information of the input $i-1$ of the rule r, S_r^i denotes the middle result of the input i, S_r denotes the summation of S_r^i, O_r denotes the output of the rule r, S denotes the sliced operation, M_r^p denotes the feature of corresponding part p of the rule r, '$\circ$' denotes the channel-wise multiplication, $\mathcal{V}$, $\mathcal{W}$, $\mathcal{W}_r$ and V are weights, b, b_r and c are bias. When $i = 1$, H_r^{i-1} does not exist.

The first input of PCRNN is updated by the results of its former one,

$$\chi_{r+1}^1 \leftarrow O_r, \tag{7}$$

where χ_{r+1}^1 denotes the first input of the rule $r+1$.

Our PCRNN passes messages unidirectionally. The purpose of our PCRNNs is to carry out the message passing from conspicuous parts (e.g., torso, head) to the segmentation of inconspicuous parts (e.g., low-leg, low-arm). Exploiting inconspicuous parts to amend conspicuous ones is generally unreliable because distinguishing inconspicuous parts from their background is much harder than distinguishing conspicuous ones. We design B-PCRNN which passes messages back and forth to improve performance directionally. Experimental results on B-PCRNN and PCRNN in Table 1 and show that PCRNN has better performance.

Blended Grammar Loss. Every PCRNN has its corresponding blended grammar loss which locates at the end of it. The supervision is the ground truth of its corresponding human parts. For example, the supervision of $Rule_1^c$ is head and torso, the next layer, the supervision of $Rule_2^h$ is head, torso and upper-arm, etc. It is seen that our grammar loss uses a less-to-more manner with the increasing number of parts in rules.

$$L_r = -\frac{1}{MN}\sum_{i=1}^{MN}\sum_{k=1}^{K}(y_i = k)log(p_{i,k}), \tag{8}$$

where M and N is the height and width of the input image, and K is the number of the categories in rule r, y is a binary indicator (0 or 1) if categories label k is the correct classification for observation i, and p predicts probability observation i of categories k.

3.4 Loss Function

In our BGNet, we design a novel deep blended grammar loss to supervise the training of PCRNNs, termed L_r. For the deep blended grammar loss, we utilize softmax cross-entropy loss.

Following PSPNet [43], BGNet employs two deep auxiliary losses, one locates at the end of the baseline and the other is applied after the twenty-second block of the fourth stage of ResNet101, i.e., the res4b22 residue block, which is named as L_{aux1} and L_{aux2}, respectively. The loss at the end of our method is named as $L_{softmax}$. The total loss can be formulated as:

$$L = \lambda L_{softmax} + \lambda_1 L_{aux1} + \lambda_2 L_{aux2} + \sum_{r=1}^{n} L_r, \tag{9}$$

where we fix the hyper-parameters $\lambda = 0.6$, $\lambda_1 = 0.1$, and $\lambda_2 = 0.3$ in our experiments, r denotes rules in BGNet, n denotes the number of rules and it is 5 in our parsing network. We experiment with setting the auxiliary loss weight λ_1 and λ_2 between 0 and 1, respectively. Then, we set the loss at the end of our method λ between 0 and 1. $\lambda = 0.6$, $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$ yield the best results.

4 Experiments

4.1 Datasets

PASCAL-Person-Part. PASCAL-Person-Part [5] has multiple person 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).

LIP. LIP dataset [14] 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

Table 1. Ablation study for our network. The results are obtained on the validation set of PASCAL-Person-Part [5]. CCL denotes the common convolutional layers with an approximate quantity of parameters compared with 5 PCRNNs. R-PCRNN denotes the reversible order of grammar rules. B-PCRNN is proposed by [35]. PCRNN is proposed in our paper. BGL denotes our novel deep blended grammar loss. DA denotes data augmentation. MS denotes multi-scale testing.

Method	CCL	R-PCRNN	B-PCRNN	PCRNN	BGL	DA	MS	Ave.
Baseline								66.61
Baseline + CCL	✓							67.04
Baseline + R-PCRNN		✓						67.50
Baseline + B-PCRNN			✓					68.1
Baseline + PCRNN				✓				70.35
Baseline + PCRNN + BGL				✓	✓			72.46
Baseline + PCRNN + BGL + DA				✓	✓	✓		73.13
Baseline + PCRNN + BGL + DA + MS				✓	✓	✓	✓	74.42

(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 network.

PPSS. PPSS dataset [27] includes 3,673 annotated samples, which are divided into a training set of 1,781 images and a testing set of 1,892 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.

Evaluation Metrics. We evaluate the mean pixel Intersection-over-Union (mIoU) of our network in experiments.

4.2 Implementation Details

As for the baseline, we use the FCN-like ResNet-101 [15] (pre-trained on ImageNet [32]). In addition, the PPM module [43] is applied for extracting more effective features with multi-scale context. Following PSPNet [43], the classification layer and last two pooling layers are removed and the dilation rate of the convolution layers after the removed pooling layers are set to 2 and 4 respectively. Thus, the output feature is $8\times$ smaller than the input image.

We train all the models using stochastic gradient descent (SGD) solver, momentum is 0.9 and weight decay is 0.0005. As for these three datasets (PASCAL-Person-Part, LIP and PPSS), we resize images to 512×512, 473×473, and 512×512 as the input size, respectively, the batch sizes are 8, 12, and 8, respectively, the epochs of three datasets are 100, 120, 120, respectively. We do not use OHEM. 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 a multi-scale context.

Table 2. These five grammar rules influence each part category. From the numbers marked with different colors, we show that each grammar rule can improve the accuracy of human parts. The results are obtained on the validation set of PASCAL-Person-Part.

Method	Rule-1	Rule-2	Rule-3	Rule-4	Rule-5	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
Baseline						86.95	70.72	58.22	55.17	51.22	47.09	96.11	66.61
(a)	✓					87.98	72.93	60.03	57.57	54.3	50.71	96.12	68.52
(b)	✓	✓				88.02	73.65	64.12	62.72	56.15	53.73	96.1	70.64
(c)	✓	✓	✓			88.11	73.58	64.51	63.22	57.39	55.6	96.13	71.22
(d)	✓	✓	✓	✓		88.67	73.8	66.73	66.08	58.15	55	96.18	71.93
(e)	✓	✓	✓	✓	✓	88.74	75.68	67.09	64.99	58.14	56.18	96.37	72.46

4.3 Ablation Study

In this section, we conduct several experiments to analyze the effect of each component in our BGNet on PASCAL-Person-Part [5]. The grammar rules for PASCAL-Person-Part are represented in Eq.3.

Grammar Rule and PCRNN. To show some empirical details of designing the five grammar rules and evaluate the performance of the PCRNN, we conduct experiments by four settings without blended grammar loss in BGNet, as illustrated in Tabel 1. First, the common convolutional layers have an approximate quantity of parameters compared with 5 PCRNNs, named Baseline + CCL. Second, we reverse the order of all the five grammar rules, such as changing the $Rule_1^c$ as $torso \rightarrow head$, named Baseline + R-PCRNN. Third, we also adopt B-PCRNN [35] to express the grammar rules, named Baseline + B-PCRNN. Fourth, PCRNN is proposed in our paper, named Baseline + PCRNN. We have tried more grammar rules to parse a human body while preparing the paper. For example, $torso \rightarrow upper-arm$,$upper-arm \rightarrow low-arm$, and so on. But the improvement over these five grammar rules can be negligible. Therefore, these five grammar rules are the best choice.

Because the conspicuous parts, which are easily distinguished from the background, can amend the inconspicuous ones, improving the foreground extraction. Exploiting inconspicuous parts to amend conspicuous ones is generally unreliable because distinguishing inconspicuous parts from their background is much harder than distinguishing conspicuous ones. Thus, PCRNN is better than B-PCRNN [35] and has the best performance compared with other settings.

In Table 2, (a)-(e) presents the performance of BGNet with adding grammar rules. It is seen that each grammar rule can improve the accuracy of the grammar model. The five grammar rules improve the accuracy of every part, as shown in Table 2. From the numbers marked with different colors, it can be seen that $Rule_1$ improves head by 1.03% and torso by 2.21%, compared with the baseline, $Rule_2$ improves upper-arm by 4.09%, compared with (a), $Rule_3$ improves low-arm by 0.5%, compared with (b), $Rule_4$ improves upper-legs by 0.76%, compared with (c), and $Rule_5$ improves low-legs by 1.18%, compared with (d).

Computation Comparison. We experiment by adding common convolutional layers on the top of the baseline, named Baseline + CCL. The computation

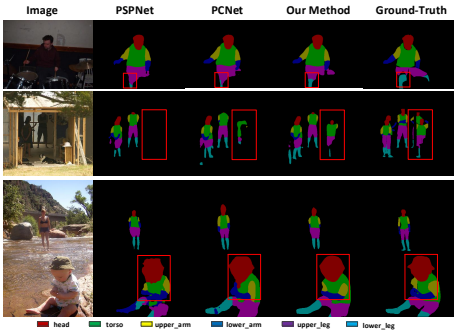

Fig. 5. Qualitative comparison among our method and state-of-the-art approaches on PASCAL-Person-Part[5] dataset. In the first two rows, our method extracts more complete foregrounds from cluttered scenes. And in the last two rows, our method segments different human parts more accurately, such as head and upper-arm.

of Baseline + CCL is similar to our framework. Our framework yields a result of 72.46% on PASCAL-Person-Part, exceeding Baseline + CCL 67.04% by 4.42%. The result shows that our method improves performance due to our algorithm rather than extra computational overhead.

Different Architectures. Blended Grammar Network (BGNet) and Paralleled Grammar Network (PGNet) are two architectures of the grammar model. BGNet yields a result of 74.42% and exceeds paralleled PGNet 71.02% by 3.4%, showing our blended architecture has better performance than the paralleled architecture.

4.4 Comparison with State-of-the-Art

Results on PASCAL-Person-Part Dataset. To further demonstrate the effectiveness of BGNet, we compare it with state-of-the-art methods on the validation set of PASCAL-Person-Part. As shown in Table 3, our BGNet outperforms other methods on all categories. Furthermore, BGNet achieves the state-of-the-art performance, i.e., 74.42%, and outperforms the previous best one by 3.08%.

Qualitative Comparison. The qualitative comparison of results on PASCAL-Person-Part [5] is visualized in Fig. 5. From the first row, we find that our method has better performance in extracting the foreground from cluttered scenes compared with the PSPNet [43] and PCNet [45]. In the second row, PSPNet misses some parts of human bodies, and PCNet only can segment a few parts. However, most of the parts can be segmented by our network. For the head, upper-arm in the last row, ours performs well on these small parts and large parts in the image compared with the other methods.

Results on LIP dataset. According to PASCAL-Person-Part, we define 6 human parts to constitute grammar rules, which are head, upper-clothes, arm, upper-leg, low-leg and shoes. The region of the head is generated by merging

Table 3. Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with the state-of-the-art methods on PASCAL-Person-Part [5].

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
HAZA [38]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
LIP [14]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
MMAN [28]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
MuLA [29]	-	-	-	-	-	-	-	65.1
PCNet [45]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [17]	-	-	-	-	-	-	-	66.3
WSHP [6]	87.15	72.28	57.07	56.21	52.43	50.36	97.72	67.60
PGN [13]	90.89	75.12	55.83	64.61	55.42	41.47	95.33	68.40
RefineNet [21]	-	-	-	-	-	-	-	68.6
Learning [36]	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
Graphonomy [12]	-	-	-	-	-	-	-	71.14
DPC [2]	88.81	74.54	63.85	63.73	57.24	54.55	96.66	71.34
CDCL [22]	86.39	74.70	68.32	65.98	59.86	58.70	95.79	72.82
BGNet (ours)	**90.18**	**77.44**	**68.93**	**67.15**	**60.79**	**59.27**	**97.12**	**74.42**

Table 4. Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with state-of-the-art methods on LIP [14].

Method	hat	hair	glov	sung	clot	dress	coat	sock	pant	suit	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-sh	r-sh	bkg	Ave.
FCN-8s [25]	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
DeepLabV2 [3]	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	41.64
Attention [4]	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[3]	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 [14]	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
ASN [26]	56.92	64.34	28.07	17.78	64.90	30.85	51.90	39.75	71.78	25.57	7.97	17.63	70.77	53.53	56.70	49.58	48.21	34.57	33.31	84.01	45.41
MMAN [28]	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 [18]	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 [23]	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
BraidNet [24]	66.8	72.0	42.5	32.1	69.8	33.7	57.4	49.0	74.9	32.4	19.3	27.2	74.9	65.5	67.9	**60.2**	59.6	47.4	47.9	88.0	54.4
BGNet (ours)	**69.18**	**73.14**	**44.27**	**34.16**	**72.32**	**36.13**	**60.69**	**50.93**	**76.56**	**38.78**	**31.38**	**33.86**	**76.21**	**66.53**	**68.04**	59.83	**60.06**	**47.96**	**48.01**	**88.30**	**56.82**

parsing labels of hat, hair, sunglasses and face. Similarly, upper-clothes, coat, dress, jumpsuits and scarf are merged to be upper-clothes, right-arm, left-arm and gloves for arm, pants and skirt for upper-leg. The rest regions can also be obtained by correcting labels. Due to different datasets with different kinds of labels of human parts, the grammar rules may be adjusted slightly. The five grammar rules for LIP dataset are:

$$
\begin{aligned}
&Rule_1^c : head \rightarrow upperclothes,\\
&Rule_2^h : Rule_1^c \rightarrow arm,\\
&Rule_3^v : Rule_1^c \rightarrow upperleg,\\
&Rule_4^v : Rule_3^v \rightarrow lowleg,\\
&Rule_5^v : Rule_4^v \rightarrow shoes.
\end{aligned}
$$

We compare our method with previous networks on the validation set, which are FCN-8s [25], Attention [4], LIP [14], BraidNet [24] and so on. As shown in Table 4, our method outperforms all priors. Our proposed framework yields 56.82% in terms of mIoU on the LIP. Compared with the best methods, ours exceeds it 2.42%.

Table 5. In the first two rows, performance comparison of model trained on LIP to test the PPSS [27]. In the bottom half, performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with the state-of-the-art methods on PPSS.

Method	Hair	Face	U-cloth	arms	L-cloth	Legs	Background	Ave.
MMAN [28]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
BGNet (ours)	59.36	57.15	63.94	42.68	59.96	27.68	86.09	56.69
DDN [27]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [26]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
BGNet (ours)	**70.67**	**62.31**	**82.59**	**48.12**	**72.61**	**29.82**	**92.97**	**65.44**

Results on PPSS dataset. Similar to LIP [14], we merge hair and face into the head and the grammar rules may be adjusted slightly. The five grammar rules for PPSS dataset are:

$$
\begin{aligned}
&Rule_1^c : head \rightarrow upperclothes,\\
&Rule_2^h : Rule_1^c \rightarrow arms,\\
&Rule_3^v : Rule_1^c \rightarrow lowcloth,\\
&Rule_4^v : Rule_3^v \rightarrow legs,\\
&Rule_5^v : Rule_4^v \rightarrow shoes.
\end{aligned}
$$

We compare our method with some methods on the testing set, DDN [4], ASN [14] and MMAN [28]. In the first two rows of Table 5, we deploy the model trained on LIP [14] to the testing set of the PPSS [27] without any fine-tuning, to evaluate the generalization ability of we proposed model, which is similar to MMAN. We merge the fine-grained labels of LIP into coarse-grained human parts defined in PPSS. From Table 5, our method outperforms MMAN by 4.59%. We also train our method on the training set of PPSS dataset, whose results in segmentation on the testing set achieve further improvement. Our proposed framework achieves 65.44% in terms of Mean IoU on PPSS dataset. Compared with ASN, our method exceeds 14.74%.

5 Conclusion

In this work, we propose a Blended Grammar Network (BGNet) to improve the extraction accuracy of the foreground from the cluttered background. BGNet exploits the inherent hierarchical structure of a human body and the relationship of human parts by means of grammar rules in both cascaded and paralleled manner. Then, we design the Part-aware Convolutional Recurrent Neural Network (PCRNN) to pass messages across BGNet which can adaptively extract features to improve accuracy. To train PCRNN effectively, we develop a blended grammar loss to supervise the training of PCRNNs, which uses a less-to-more manner with increasing parts in grammar rules. Finally, extensive experiments show that BGNet improves the performance of the baseline models on three datasets significantly. These results on three datasets prove that our framework works well on different kinds of datasets.

Acknowledgement. This work was supported by Research and Development Projects in the Key Areas of Guangdong Province (No.2019B010153001), and National Natural Science Foundation of China (No.61772527, 61976210, 61806200, 61702510 and 61876086). Thanks Prof. Si Liu and Wenkai Dong for their help on paper writing.

References

1. Amit, Y., Trouvé, A.: Pop: Patchwork of parts models for object recognition. In IJCV **75**(2), 267–282 (2007)
2. Chen, L.C., Collins, M., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J.: Searching for efficient multi-scale architectures for dense image prediction. In: NeurIPS. pp. 8699–8710 (2018)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In IEEE TPAMI **40**(4), 834–848 (2018)
4. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR (June 2016)
5. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: CVPR. pp. 1979–1986 (2014)
6. Fang, H.S., Lu, G., Fang, X., Xie, J., Tai, Y.W., Lu, C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. arXiv preprint arXiv:1805.04310 (2018)
7. Fang, H., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. AAAI (2018)
8. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. pp. 2360–2367 (2010)
9. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
10. Fu, J., Liu, J., Wang, Y., Lu, H.: Densely connected deconvolutional network for semantic segmentation. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3085–3089. IEEE (2017)
11. Garland-Thomson, R.: Staring: How we look. Oxford University Press (2009)
12. Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., Lin, L.: Graphonomy: Universal human parsing via graph transfer learning. In: In CVPR (2019)
13. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 770–785 (2018)
14. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR. vol. 2, p. 6 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: ECCV. pp. 713–728 (2018)
17. Li, Q., Arnab, A., Torr, P.H.: Holistic, instance-level human parsing. arXiv preprint arXiv:1709.03612 (2017)
18. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing & pose estimation network and a new benchmark. In IEEE TPAMI (2018)
19. Liang, X., Lin, L., Shen, X., Feng, J., Yan, S., Xing, E.P.: Interpretable structure-evolving lstm. In: CVPR. pp. 2175–2184 (2017)
20. Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., Yan, S.: Semantic object parsing with local-global long short-term memory. In: CVPR. pp. 3185–3193 (2016)

21. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR. pp. 1925–1934 (2017)
22. Lin, K., Wang, L., Luo, K., Chen, Y., Liu, Z., Sun, M.T.: Cross-domain complementary learning with synthetic data for multi-person part segmentation. arXiv preprint arXiv:1907.05193 (2019)
23. Liu, T., Ruan, T., Huang, Z., Wei, Y., Wei, S., Zhao, Y., Thomas, H.: Devil in the details: Towards accurate single and multiple human parsing. AAAI (2019)
24. Liu, X., Zhang, M., Liu, W., Song, J., Mei, T.: Braidnet: Braiding semantics and details for accurate human parsing. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 338–346. ACM (2019)
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
26. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
27. Luo, P., Wang, X., Tang, X.: Pedestrian parsing via deep decompositional network. In: CVPR. pp. 2648–2655 (2014)
28. Luo, Y., Zheng, Z., Zheng, L., Guan, T., Yu, J., Yang, Y.: Macro-micro adversarial network for human parsing. In: ECCV (2018)
29. Nie, X., Feng, J., Yan, S.: Mutual learning to adapt for joint human parsing and pose estimation. In: ECCV. pp. 502–517 (2018)
30. Park, S., Nie, B.X., Zhu, S.C.: Attribute and-or grammar for joint parsing of human pose, parts and attributes. In IEEE TPAMI **40**(7), 1555–1569 (2018)
31. Qi, S., Huang, S., Wei, P., Zhu, S.C.: Predicting human activities using stochastic grammar. In: CVPR (2017)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
33. Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. nature **381**(6582), 520 (1996)
34. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.: Joint object and part segmentation using deep learned potentials. In: ICCV. pp. 1573–1581 (2015)
35. Wang, W., Xu, Y., Shen, J., Zhu, S.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: CVPR (2018)
36. Wang, W., Zhang, Z., Qi, S., Shen, J., Pang, Y., Shao, L.: Learning compositional neural information fusion for human parsing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
37. Wang, Y., Duan, T., Liao, Z., Forsyth, D.: Discriminative hierarchical part-based models for human parsing and action recognition. Journal of Machine Learning Research **13**(1), 3075–3102 (2012)
38. Xia, F., Wang, P., Chen, L.C., Yuille, A.L.: Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In: ICCV. pp. 648–663 (2015)
39. Xia, F., Wang, P., Chen, X., Yuille, A.: Joint multi-person pose estimation and semantic part segmentation. In: CVPRW. pp. 6080–6089 (2017)
40. Yamaguchi, K., Kiapour, M.H., Berg, T.L.: Paper doll parsing: Retrieving similar styles to parse clothing items. In: ICCV. pp. 3519–3526 (2013)
41. Zhang, X., Chen, Y., Zhu, B., Wang, J., Tang, M.: Part-aware context network for human parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8971–8980 (2020)
42. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV. pp. 405–420 (2018)

43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
44. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
45. Zhu, B., Chen, Y., Tang, M., Wang, J.: Progressive cognitive human parsing. In AAAI (2018)