Part-aware Context Network for Human Parsing

Xiaomei Zhang Yingying Chen Bingke Zhu Jinqiao Wang Ming Tang

National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

{xiaomei.zhang, yingying.chen, bingke.zhu, jqwang, tangm}@nlpr.ia.ac.cn

Abstract

Recent works have made significant progress in human parsing by exploiting rich contexts. However, human parsing still faces a challenge of how to generate adaptive contextual features for the various sizes and shapes of human parts. In this work, we propose a Part-aware Context Network (PCNet), a novel and effective algorithm to deal with the challenge. PCNet mainly consists of three modules, including a part class module, a relational aggregation module, and a relational dispersion module. The part class module extracts the high-level representations of every human part from a categorical perspective. We design a relational aggregation module to capture the representative global context by mining associated semantics of human parts, which adaptively augments the context for human parts. We propose a relational dispersion module to generate the discriminative and effective local context and neglect disturbing one by making the affinity of human parts dispersed. The relational dispersion module ensures that features in the same class will be close to each other and away from those of different classes. By fusing the outputs of the relational aggregation module, the relational dispersion module and the backbone network, our PCNet generates adaptive contextual features for various sizes of human parts, improving the parsing accuracy. We achieve a new state-of-the-art segmentation performance on three challenging human parsing datasets, i.e., PASCAL-Person-Part, LIP, and CIHP.

1. Introduction

Human parsing aims at classifying every pixel in images to one of the predefined categories of parts or clothes (e.g., head, torso, etc.). It has been widely used in various challenging fields such as person re-identification [9], human behavior analysis [40], clothing style recognition and retrieval [44], clothing category classification [38]. With the rapid development of electronic commerce and online shopping, human parsing has attracted much attention [37, 42, 43, 14, 20, 22, 24, 15, 47, 31, 27]. However, it is still a challenging task to generate adaptive contextual features for different human parts, due to various sizes and shapes of human parts.

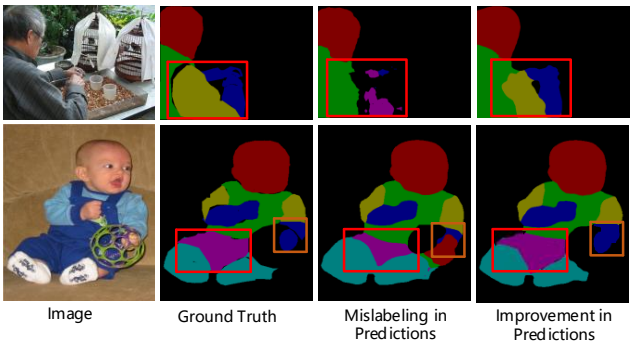

Figure 1. Examples of the challenge in human parsing. The original images and ground-truth come from PASCAL-Person-Part dataset [5]. There are various scales and shapes of human parts. Third column: CE2P [27] fails to extract the lower-arms, upper-arms, and upper-legs with the predefined fixed context. Fourth column: our method has better performance in extracting the three parts by the adaptive context.

The most widely used method [25, 31, 14, 27] to solve the above problem is aggregating various contexts from predefined fixed regions. However, the semantic context represented by predefined fixed contexts cannot meet the requirement of dense prediction tasks. Another way is applying attention mechanisms [4, 41] that assigning different weights on different channels or positions. However, attention models have a major drawback that they focus on local patterns of parts and ignore the relation between parts, which restricts their capability to capture the global context. As shown in Figure 1, lower-arms, upper-arms and upper-legs cannot be categorized correctly due to the lack of adaptive contextual features for human parts with various scales and shapes. Different from the above methods, we argue that different human parts need an adaptive context. In this

paper, we propose a Part-aware Context Network (PCNet) to leverage the global and local context and adaptively generate proper contextual features for human parts with various scales and shapes.

To exploit the high-level representations of every human part, we propose a part class module which describes the overall representations of each category, as shown in Figure 2(a). Specifically, the high-level representation for one category is extracted by clustering the features of all the positions belonging to this class. Moreover, we design a relational aggregation module to generate global features by leveraging a dynamic convolution kernel with the global context, as shown in Figure 2(b). The global context is generated by progressively refining graph representations within the same graph structure of all human parts. Because neurons processing information reveals that neurons are adaptive processors, changing their function according to behavioral context [12] and some methods [16, 7] verified that outputs of a specific convolutional layer are guided by its kernel content. Thus, the dynamic convolution kernel with the global context is applied to the original features generated from the backbone network to adaptively generate the global features for human parts.

To alleviate the negative influence of redundant or interfering information in the global context, we develop a relational dispersion module which generates the discriminative and effective local features by applying a dynamic convolution kernel with the discriminative local context, as shown in Figure 2(c). Specifically, the discriminative local context (affinity dispersion) is generated by making the affinity of human parts dispersed and the discrimination of every human part enhanced. By measuring the similarity among the features of human parts, we can obtain the affinity coefficient of parts. The smaller coefficient indicates that the two parts have a closer relation. Then the features of the part multiply with its affinity coefficient to make the affinity disperse and itself remain unchanged, called affinity dispersion. We concatenate affinity dispersion of all human parts as a dynamic convolution kernel. The kernel applied to the original features from the backbone network generates the discriminative and effective local features for human parts.

We fuse the outputs of the relational aggregation module, the relational dispersion module and the backbone network to obtain proper contextual features for different human parts. The overall structure of our PCNet is shown in Figure 2. Extensive experiments on three popular benchmarks show that our network achieves a new state-of-the-art consistently on three public benchmarks, PASCAL-Person-Part [5], LIP [15] and CIHP [14]. In summary, our contributions are in three folds:

1. A novel Part-aware Context Network (PCNet) is designed specially to generate adaptive contextual features for human parts with various scales and shapes.

2. The proposed PCNet consists of three modules, including a part class module exploits the high-level representations of every human part, a relational aggregation module captures representative global context, a relational dispersion module generates the discriminative and effective local context and neglects disturbing one.

3. The proposed PCNet achieves new state-of-the-art results consistently on three public human parsing benchmarks. Specifically, our method outperforms the best competitor by 3.25%, 2.63%, and 0.43% on PASCAL-Person-Part, LIP, and CIHP in terms of mIoU, respectively.

2. Related Work

Human Parsing. Many research efforts have been devoted to human parsing [42, 43, 15, 20, 22, 24, 47, 45, 31, 27]. Chen *et al*. [4] proposed an attention mechanism that learned to weigh the multi-scale features at each pixel location softly. Xia *et al*. [42] proposed HAZN for object part parsing, which adapted to the local scales of objects and parts by detection methods. However, [4, 42] ignored the relation of human parts. In this paper, our PCNet captures the global context of human parts by mining their associated semantic.

Human pose estimation and semantic part segmentation are two complementary tasks [21], 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*. [13] proposed Graphonomy, which incorporated hierarchical graph transfer learning upon the conventional parsing network to predict all labels. Wang *et al*. [39] combined neural networks with the compositional hierarchy of the human body for complete human parsing. All the above methods focus on how to capture the relation of different human parts and ignore how to generate discriminative contextual representation. Different from the above methods, our PCNet can generate a remarkable contextual representation for human parsing.

Contextual Modeling. There are two mainstreams to enhance contextual aggregation in parsing networks. One is to concatenate or sum multi-scale features to segment human parts [22, 14, 27, 10], Liang *et al*. [22] proposed a contextualized convolutional neural network to integrate the cross-layer context of an image by summing features of a down-sampling process for human parsing. Gong *et al*. [14] and Liu *et al*. [27] used a pyramid pooling module to concatenate multi-scale features. The other is that [4, 11] proposed an attention model to output a weight map which weights features pixel by pixel for each scale and is shared across all the channels of the same scale. However, the kernels of all the above methods are fixed after training. D-

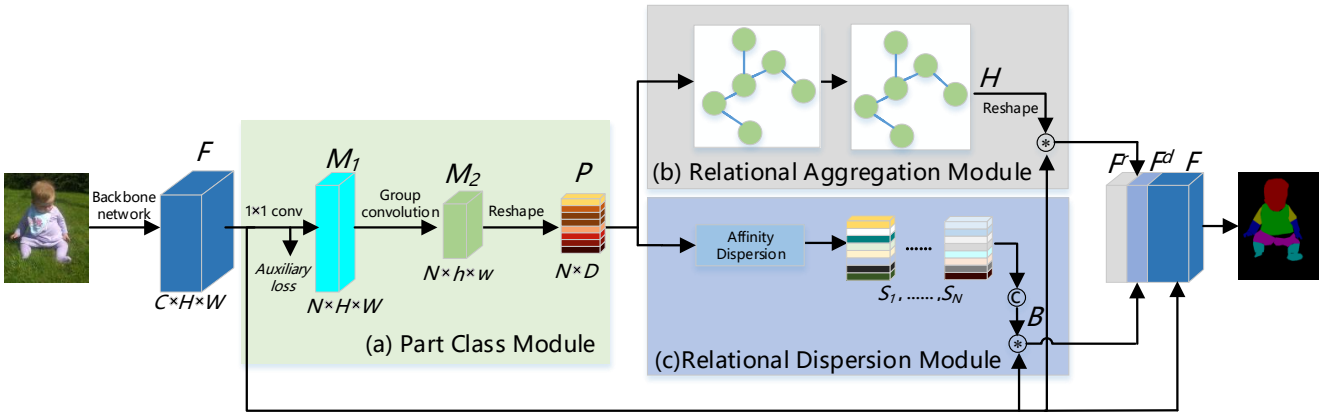

Figure 2. Overview of the proposed Part-aware Context Network (PCNet). An input image goes through the backbone network to generate its original features. The part class module is appended to extract the high-level representations of every human part. Then the relational aggregation module and the relational dispersion module adaptively generate the global context and the discriminative local context, respectively. By fusing the outputs of the relational aggregation module, the relational dispersion module and the backbone network, PCNet can adaptively generate features with rich contexts for human parts with various scales. ⊛ denotes the convolution operation, ⓒ indicates the concatenation operation.

ifferent from the above methods, our kernels can generate dynamically conditioned on an input.

Additionally, Ding *et al*. [7] proposed a shape-variant context to model the diverse shapes and scales of contexts, which greatly enhanced the modeling ability of the network. Wang *et al*. [36] reassembled the features inside a predefined region centered at each location via a weighted combination, where the weights were generated in a content-aware manner. Although [7, 36] have rich contexts, they cannot spotlight discriminative and effective local features. In this paper, our relational dispersion module can enhance discriminative and effective local features.

3. Part-aware Context Network

3.1. Overall Framework

The overall framework of our network, Part-aware Context Network (PCNet), is shown in Figure 2. Our backbone network is ResNet-101 [17] (pre-trained on ImageNet [34]). Following PSPNet [46], 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 stride of the network is set to 8. Our PCNet consists of three modules, including a part class module, a relational aggregation module, and a relational dispersion module. The part class module takes the original features of the backbone network as inputs and exploits the high-level representations for human parts from a categorical perspective. Then the outputs of the part class module and the backbone network are sent into the relational aggregation module and the relational dispersion module, respectively. The relational aggregation module adaptively generates global features depending on the associated semantics of human parts. The relational dispersion module adaptively generates discriminative local features. Finally, we aggregate the representative global features, discriminative local features and original features to obtain adaptive contextual features for human parts.

3.2. Part Class Module

The proposed part class module exploits the high-level representations of every human part from a categorical perspective. As shown in Figure 2(a), the part class module applies a channel reduction operation to original features through a 1×1 $conv$ to reduce the channel number to obtain $M_1 \in R^{N \times H \times W}$, which are learned under the supervision from the ground-truth segmentation.

In order to decrease the computation, we compress $M_1 \in R^{N \times H \times W}$ to $M_2 \in R^{N \times h \times w}$ by a group convolutional layer. We reshape $M_2 \in R^{N \times h \times w}$ to $P \in R^{N \times D}$, where N is the number of categories, $D = hw$ and D is the feature dimension for each category. The number of nodes N typically corresponds to the number of target labels of a dataset.

3.3. Relational Aggregation Module

The relational aggregation module is proposed to adaptively generate global features by capturing high-order associated semantics among human parts. The relational aggregation module is proposed to exploit the graph representation of human parts to generate dynamic global-aware convolutional kernels. Based on the high-level representations $P \in R^{N \times D}$, we leverage semantic constraints from the human body structure knowledge to evolve the global context by graph reasoning. One node denotes one category of a dataset. We introduce the connections between the human parts to encode the relation between two nodes, as

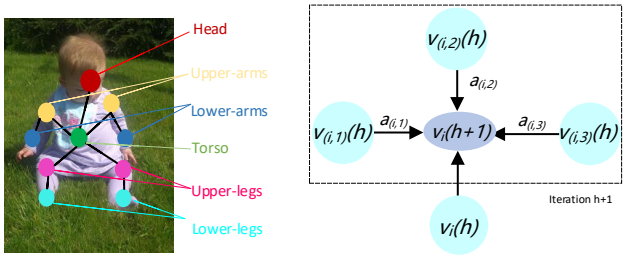

Figure 3. (a) Examples of the definited connections between each two human body parts, which is the foundation to encode the relations between two semantic nodes in the graph for reasoning. Two nodes are related if they are connected by a black line. (b)The framework of GCN.

shown in Figure 3(a). For example, $head$ usually appears with $torso$ so these two nodes are linked while the $head$ node and the $Lower-legs$ are disconnected because they have nothing related.

Following Graph Convolution [19], we perform graph propagation over representation $P \in R^{N\times D}$ of all nodes. As shown in Figure 3(b), we represent the associated semantics as a graph $G(V, E)$, where V is the vertex set, E is the edge set. In particular, V is the human part set, E is the relation set. The adjacency matrix $A \in R^{N\times N}$ of G is defined as:

$$A_{i,j} = \frac{\exp(r_{i,j})}{\sum_{i=1}^{m} \exp(r_{i,j})}, \tag{1}$$

where $r_{i,j}$ computes the score of the semantic relation between two human parts by a function p, i.e., inner product:

$$r_{i,j} = p(\nu_i, \nu_{(i,j)}), \tag{2}$$

where $\nu_i \in R^{1\times D}$ is a human part, $\nu_{(i,j)} \in R^{1\times D}$ is a human part directly connected to ν_i, $j \in [1, 2, ..., m]$, m is the number of ν_i's neighbors. In general, $r_{i,j} \in R$ characterizes the importance of the semantic relation between human part ν_i and ν_j. For example, a part may have more potential interesting in the part to share inherent structures of the human body, whereas another part may be more concerned about the appearance of parts.

The next step in the G is to aggregate the human part ν_i and its neighbors:

$$\widetilde{A}_{i,j} = A_{i,j} + I_n, \tag{3}$$

where I_n is the identity matrix. Finally, the output of the graph convolution layer is computed as:

$$H = \sigma(\widetilde{A}PW), \tag{4}$$

where W is a trainable parametric matrix, and σ is the nonlinear function such as ReLU.

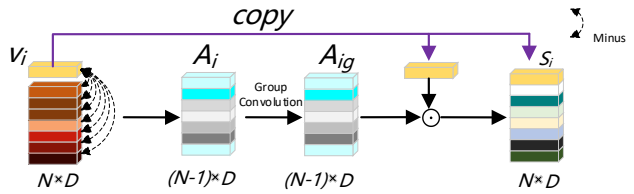

Figure 4. Examples of affinity dispersion of human parts. $\odot$ indicates the element-wise multiplication operation.

Following [35], we convert $H \in R^{N\times D}$ (the part-aware global context) to $H' \in R^{D\times h\times h}$ by using a learnable projection matrix. Then, H' goes through a convolutional layer to increase the channel number to C, obtaining $\theta' \in R^{c_i\times h\times h}$. Note that $c_i = C$, C is the number of channels of the original features. Following [33], here h=7 is the convolution kernel size. Following [1], we perform the following operations on θ' to generate the convolution kernels $\theta \in R^{h\times h\times c_i\times c_o}$:

$$\theta = U * \theta' *_c V, \tag{5}$$

where $*$ denotes the convolution operation, $*_c$ denotes the channel-wise convolution operation, $U \in R^{1\times 1\times c_i\times c_o}$ and $V \in R^{1\times 1\times c_i\times c_i}$ are auxiliary parameters to learn for the convolution kernels.

With the convolution kernels θ, we adopt a convolutional layer, different from the traditional convolutional layer where kernel weights are generated dynamically conditioned on an input.

$$F^r = \sigma(\theta * F), \tag{6}$$

where F is the output of the backbone network, F^r is the output of the relational aggregation module.

It can be seen from the above formulation that parts with different sizes can automatically augment the representation of the context of human parts with the part relation presentation. With our design, the relational aggregation module can generate global features according to the global context.

3.4. Relational Dispersion Module

The global features are essential for human parsing, but it may bring the negative influence of redundant or interfering information because it is difficult for the global features to classify adjacent patches with similar appearances but different semantic labels. In order to learn the discriminative and effective local features, we propose the relational dispersion module. The module can ensure that features in the same class will lie close to each other and away from those of different classes, as shown in Figure 2(c).

In order to adaptively enhance local features, we disperse the affinity of human parts, as shown in Figure 4. Specifically, we first measure the similarity among features of human

parts. For example, the similarity of a human part ν_i and other human parts ν_p is calculated by matrix minus, which gets a_{ip}:

$$a_{ip} = \nu_i - \nu_p, \tag{7}$$

where $\nu_i \in R^{1\times D}$, $\nu_p \in R^{1\times D}$, $p \in [1, 2, ..., N-1]$, and N is the number of categories. Noted that the smaller a_{ip} indicates that the features of the human part ν_i is more similar to the human part ν_p's, $a_{ip} \in A_i$.

Then a group convolution layer is applies to $A_i \in R^{(N-1)\times D}$ to generate $A_{ig} \in R^{(N-1)\times D}$. ν_i made product with $A_{ig} \in R^{(N-1)\times D}$ to generate $S_i' \in R^{(N-1)\times D}$. According to the original location of ν_i, we put it to the S_i', which generates $S_i \in R^{N\times D}$. We concatenate all the S_i, that is:

$$\begin{aligned} S_i' &= A_{ig} \odot \nu_i, \\ B &= cat(\{S_i\}_{i=1}^{N}), \end{aligned} \tag{8}$$

where $\odot$ denotes the element-wise multiplication operation, $cat(\cdot)$ denotes the concatenation function, B denotes the results by concatenating S_i.

We apply a convolution layer on $B \in R^{N\times N\times D}$ to generate $\omega' \in R^{c\times h\times h}$, and use Eq. 5 on the ω' to generate $\omega \in R^{h\times h\times c\times c_o}$.

With kernel ω, we adopt a convolutional layer, different from the traditional convolutional layer. The kernels are generated dynamically conditioned on an input:

$$F^d = \sigma(\omega * F) \tag{9}$$

where F^d is the outputs of the relational dispersion module.

It can be seen that our relational dispersion module can learn the discriminative and effective local features by class-specific cohesion and separation of sample features of inter-class.

3.5. Loss Function

Following PSPNet [46], PCNet employs two deep auxiliary losses that are softmax cross-entropy loss, one locates at the part class part as shown in Figure 2 and the other is applied after the twenty-second module of the fourth stage of ResNet101, i.e., the res4b22 residue module, they are 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:

$$L = \lambda L_{softmax} + \lambda_1 L_{aux1} + \lambda_2 L_{aux2}, \tag{10}$$

where we fix the hyper-parameters λ = 0.6, λ_1 = 0.1, and λ_2 = 0.3 in our experiments.

We experiment with setting the auxiliary loss weight λ_1 and λ_2 between 0 and 1, respectively. Then, we experiment with setting the loss at the end of our method λ between 0 and 1. λ = 0.6, λ_1 = 0.1 and λ_2 = 0.3 yield the best results.

4. Discussion

Here we discuss the relations between PCNet and dynamic filter [18], and deformable convolution [6], which share similar design philosophy but with different focuses.

Dynamic Filter. Dynamic filter generates instance-specific convolutional filters conditioned on the input of the network, and then applies the predicted filter on the input. Both dynamic filter and PCNer are content-aware operators, but a fundamental difference between them lies in their kernel generation process. Specifically, the contents of the dynamic filter are not related to each other, whereas the contents of our PCNet are the global-aware content or local-aware content.

Deformable Convolutional Networks (DCN). DCN also adopts the idea of learning geometric transformation and combines it with the regular convolution layers. It predicts kernel offsets other than using grid convolution kernels. It is also known to be sensitive to parameter initialization.

5. Experiments

5.1. Datasets

PASCAL-Person-Part dataset [5], there are multiple person appearances in an unconstrained environment. Each image has 7 labels: background, head, torso, upper-arm, lower-arm, upper-leg and lower-leg. Originally, we only use the images containing human for training (1716 images) and validation (1817 images).

LIP dataset [15] 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 network.

CIHP dataset [14] is a new large-scale benchmark for human parsing task, including 38,280 images with pixel-wise annotations on 19 semantic part labels. The images are collected from the real-world scenarios, containing persons appearing with challenging poses and viewpoints, heavy occlusions, and in a wide range of resolutions. Following the benchmark, we use 28,280 images for training, 5,000 images for validation and 5,000 images for testing.

Evaluation Metrics We evaluate the mean pixel Intersection-over-Union (mIoU) of our network in experiments on these three datasets. Additionally, we adopt mean accuracy (Mean Acc) for CIHP dataset.

5.2. Implementation Details

As for the baseline, we use the FCN-like ResNet-101 [17] (pre-trained on ImageNet [34]). Following PSPNet [46], the classification layer and last two pooling layers

#	Baseline	Part-class (ours)	RAM-mult (ours)	RAM-conv (ours)	D-gcn (ours)	T-gcn (ours)	RDM-concat (ours)	RDM-conv (ours)	DA	MS	mIoU(%)
1	✓										66.61
2	✓	✓									67.40
3	✓	✓	✓								68.53
4	✓	✓		✓							69.41
5	✓	✓		✓	✓						70.12
6	✓	✓		✓		✓					70.52
7	✓	✓		✓		✓	✓				71.08
8	✓	✓		✓		✓		✓			72.26
9	✓	✓		✓		✓		✓	✓		73.32
10	✓	✓		✓		✓		✓	✓	✓	74.59

Table 1. Ablation study for our network PCNet. The results are obtained on the validation set of PASCAL-Person-Part [5]. The baseline is ResNet-101. Part-class denotes our part class module. RAM denotes our relational aggregation module. RDM denotes our relational dispersion module. RAM-mult denotes that the global context H and original features from the backbone are multiplied in a channel-wise way in our relational aggregation module. RAM-conv denotes that the global context H as kernels applied to original features in our relational aggregation module. D-gcn denotes that dual graph convolution layers are applied in our relational aggregation module. T-gcn denotes that three graph convolution layers are applied in our relational aggregation module. RDM-concat denotes that the B and original features are concatenating in our relational dispersion module. RDM-conv denotes that the B as kernels applied to original features in our relational dispersion module. DA denotes data augmentation with multi-scale input during the training phase, MS denotes multi-scale testing.

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× smaller than the input image if not specified. Additionally, we train the method in an end-to-end manner. The number of nodes in the relational aggregation module is set according to the number of categories of human parts, i.e., $N = 7$ for Pascal-Person-Part dataset, $N = 20$ for LIP dataset, $N = 20$ for CIHP dataset. The feature dimension D of each semantic node is 128. The relational aggregation module has two graph convolution layers with the ReLU activate function.

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 CIHP), 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. All networks are trained on NVIDIA GTX TITAN X GPU with 12 GB memory.

5.3. Ablation Study

We conduct all of our ablation study experiments with ResNet-101 as our backbone network and report all the performance with only single scale testing on the PASCAL-Person-Part validation set. For starters, we evaluate the performance of the baseline, as the #1 result in Table 1. It should be noted that all our experiments use auxiliary supervision.

Ablation for the Part Class Module. To verify the effect of the part class module, we first remove the relational aggregation module and the relational dispersion module in Figure 2. The $P \in R^{N \times D}$ is upsampled and reshaped to $R^{N \times H \times W}$. Then the upsampled $P \in R^{N \times H \times W}$ and the original features from the backbone network are concatenated to get the fine segmentation result. The experiment result is shown in Table 1 (#2). This modification improves the performance to 66.61%(0.79↑) on the PASCAL-Person-Part validation set with negligible additional parameters.

Ablation for the Relational Aggregation Module. We further evaluate the role of the relational aggregation module. As for the relational aggregation module, we replace its convolutional layer named $RAM - conv$ (described in Eq. 6) with channel-wise multiplication named $RAM - mult$, the operation can be formulated as follows:

$$F^{r'} = F \cdot H, \tag{11}$$

where F is the outputs of the backbone network, $F^{r'}$ is the outputs of the relational aggregation module. The experiment result is shown in Table 1 (#3). Compared with the experiment of $baseline + partclass$ (#2), the #3 improves the performance of the PASCAL-Person-Part validation set from 67.40% to 68.53%, whereas the #4 with $RAM - conv$ achieves a performance of 69.41%. Compared with the baseline, the improvement is significant. In the following experiments, we use the relational aggregation module with a convolutional layer strategy $RAM - concat$ as default. Note that #4 and #5 apply one graph convolution layer.

Method	mIoU(%)	FLOPs	Memory	Params
ResNet-101	66.61	190.3G	2.601G	42.4M
DeeplabV3	68.32	+62.9G	+64M	+15.2M
CE2P	69.22	+60.9G	+59M	+20M
PCNet (ours)	74.59	+42.2G	+30.1M	+11.4M

Table 2. Detailed comparisons on PASCAL VOC with the baseline (ResNet-101) and DeeplabV3 in mIoU (%). All results are achieved with the backbone ResNet-101 and output stride 8. The FLOPs and memory are computed with the input size 512×512.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	mIoU(%)
HAZA [42]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
LIP [15]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
MMAN [31]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
Graph LSTM [23]	82.69	62.68	46.88	47.71	45.66	40.93	94.59	60.61
SE LSTM [22]	82.89	67.15	51.42	48.72	51.72	45.91	97.18	63.57
Joint [43]	85.50	67.87	54.72	54.30	48.25	44.78	95.32	64.39
MuLA [32]	-	-	-	-	-	-	-	65.1
PCNet [47]	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
Holistic [20]	-	-	-	-	-	-	-	66.3
WSHP [8]	87.15	72.28	57.07	56.21	52.43	50.36	97.72	67.60
PGN [14]	90.89	75.12	55.83	64.61	55.42	41.47	95.33	68.40
RefineNet [26]	-	-	-	-	-	-	-	68.6
Learning [39]	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
Graphonomy [13]	-	-	-	-	-	-	-	71.14
DPC [2]	88.81	74.54	63.85	63.73	57.24	54.55	96.66	71.34
PCNet (ours)	**90.04**	**76.89**	**69.11**	**68.4**	**60.78**	**60.14**	**96.78**	**74.59**

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].

To sufficiently propagate the global context, we employ a different number of such graph convolution. From the #4, #5 and #6 results in Table 1, we can find that three graph convolutions have the best trade-off.

Ablation for the Relational Dispersion Module. We also evaluate the role of the relational dispersion module. We replace the convolutional layer named $RDM - conv$ (described in Eq. 9) with the concatenation operation $RDM - concat$, the operation can be formulated as follows,

$$F^{d'} = concat(F, B), \tag{12}$$

where F is the outputs of the backbone network, B is upsampled to the size of F, $F^{d'}$ is the output of the relational aggregation module. The experiment results are shown in Table 1. Compared with the experiment of #6, the #7 improves the performance of the PASCAL-Person-Part validation set from 70.52% to 71.08%, whereas the #8 achieves a performance of 72.26%. When comparing with the baseline, the improvement is significant. In the following experiments, we use the $RDM - conv$ strategy as default.

We adopt data augmentation with multi-scale input during the training phase and multi-scale testing during the testing. Our PCNet achieves a performance of 74.59%.

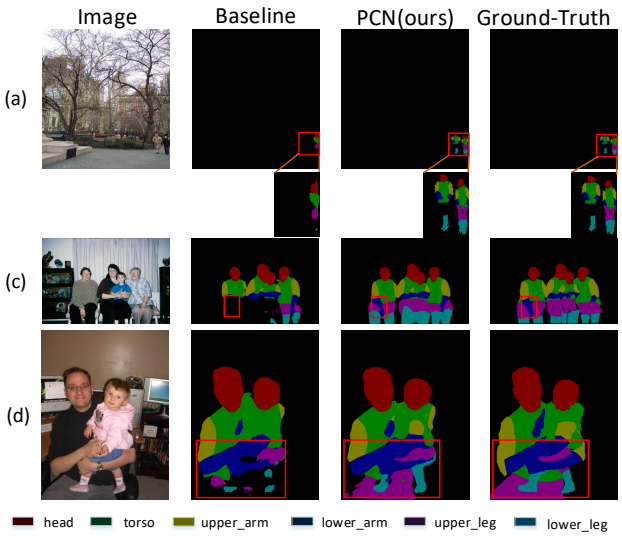

Figure 5. Qualitative comparison of PCNet results on PASCAL-Person-Part[5] dataset. In the first row, our method extracts more complete foregrounds from cluttered scenes. And in the last two rows, our method segments different human parts more accurately, such as, upper-arm and lower-arm.

Ablation for the Computation. We thoroughly compare PCNet with two models, including DeeplabV3 and CE2P [27] on the validation set of PASCAL-Person-Part. We report mIoU, FLOPS, memory cost and parameter number in Table 2. Our PCNet achieves the best performance of 74.59% among networks with the ResNet-101 and outperforms the strong competitiveness one (CE2P) by 5.37%, which is significant due to the fact that this benchmark is very competitive. Moreover, it achieves the performance that is comparable with the method based on some larger backbones. We can see that PCNet outperforms two methods by a large margin. Moreover, PCNet is much lighter in computation and memory.

Qualitative Comparison. The qualitative comparison of results on PASCAL-Person-Part [5] are visualized in Figure 5. From the first row, we find that our method has better performance in extracting foreground from cluttered scenes compared with the baseline because our relational aggregation module can generate representative global features to distinguish foreground and background and our relational dispersion module discriminative local features to segment human parts. In the second row, the baseline misses some parts of a human body, and only can segment a few parts, whereas most of parts can be segmented by our network. For upper-leg and lower-leg in the last row, ours performs well on these small parts and large parts in the image compared with the baseline.

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	mIoU(%)
FCN-8s [29]	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 [15]	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 [30]	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 [31]	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 [21]	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 [27]	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 [28]	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
PCNet (ours)	**69.32**	**73.08**	**44.72**	**34.21**	**72.59**	**36.02**	**60.84**	**51.03**	**76.66**	**38.78**	**31.60**	**33.94**	**76.65**	**67.07**	**68.74**	**60.22**	**60.16**	**47.65**	**48.67**	**88.68**	**57.03**

Table 4. Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with state-of-the-art methods on LIP [15].

Method	Mean accuracy(%)	mIoU(%)
PGN [14]	64.22	55.80
Graphonomy [13]	66.65	58.58
M-CE2P [27]	-	59.50
BraidNet [28]	-	60.62
PCNet(ours)	67.05	61.05

Table 5. Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) (%) with state-of-the-art methods on CIHP [14].

5.4. Comparison with the State-of-the-Art

Results on PASCAL-Person-Part Dataset. We compare our method with several human parsing methods including HAZA [42], LIP [15], MMAN [31], Graph LSTM [23], SE LSTM [22], Joint [43], PCNet [47], MuLA [32], Holistic [20], WSHP [8], PGN [14], RefineNet [26], Learning [39], Graphonomy [13], and DPC [2]. As shown in Table 3, we can observe that our PCNet outperforms other methods on all categories of PASCAL-Person-Part. Furthermore, our PCNet achieves state-of-the-art performance, i.e., 74.59%, and outperforms the previous best one by 3.25%.

Results on LIP dataset. In this subsection, we conduct experiments on the LIP dataset to validate the effectiveness of our method. Following previous works [15, 31, 21, 21], data augmentation with multi-scale input and multi-scale testing are used.

We compare our method with previous networks on the validation set, which are FCN-8s [29], Attention [4], DeepLab-ASPP[3], LIP [15], MMAN [31], JPPNet [21] CE2P[21], and BraidNet [27]. As shown in Table 4, our method outperforms all priors. Our proposed framework yields 57.03% in terms of mIoU on the LIP. Compared with the best methods, ours exceeds it 2.63%.

Results on CIHP dataset. The human parsing results evaluated on CIHP dataset are reported in Table 5. The previous work achieves high performance with 60.62% in mIoU this challenging dataset. Our PCNet improves the results up to 61.05%, which demonstrates its superiority and capability to takes full advantage of the global features and the discriminative local features to boost the human parsing performance.

Overall, our PCNet consistently obtains promising results over different datasets, which clearly demonstrates its superior performance and strong general. This also distinguishes our model from several previous state-of-the-art deep human parsers, such as [15, 32, 43], since it does not use extra pose annotations during training.

6. Conclusion

In this work, we propose a Part-aware Context Network (PCNet), which significantly improves performance on human parsing. PCNet consists of three modules, a part class module, a relational aggregation module, and a relational dispersion module. The part class module extracts the high-level representations of every human part from the original features from a categorical perspective. Based on the outputs of the part class module and the backbone network, our relational aggregation module can capture the global features of human parts by mining their associated semantics, and our relational dispersion module can select discriminative and effective local contexts and neglect disturbing ones for one human part. Finally, extensive experiments show that our method 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, including a simple dataset with a small number of body parts and complex datasets with a variety of human body parts.

Acknowledgement. This work was supported by National Natural Science Foundation of China (No.61976210, 61772527, 61806200, 61702510 and 61876086), and Research and Development Projects in the Key Areas of Guangdong Province (No.2020B010165001, 2019B010153001).

References

[1] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in Neural Information Processing Systems*, pages 523–531, 2016.

[2] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *NeurIPS*, pages 8699–8710, 2018.

[3] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *In IEEE TPAMI*, 40(4):834–848, 2018.

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1979–1986, 2014.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[7] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Semantic correlation promoted shape-variant context for segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8885–8894, 2019.

[8] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[10] Jun Fu, Jing Liu, Yong Li, Yongjun Bao, Weipeng Yan, Zhiwei Fang, and Hanqing Lu. Contextual deconvolution network for semantic segmentation. *Pattern Recognition*, page 107152, 2020.

[11] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE international conference on computer vision*, pages 6748–6757, 2019.

[12] Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013.

[13] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohu Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *In CVPR*, pages 7450–7459, 2019.

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018.

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, volume 2, page 6, 2017.

[16] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3562–3572, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[20] Qizhu Li, Anurag Arnab, and Philip HS Torr. Holistic, instance-level human parsing. *arXiv preprint arXiv:1709.03612*, 2017.

[21] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *In IEEE TPAMI*, 41(4):871–885, 2018.

[22] Xiaodan Liang, Liang Lin, Xiaohui Shen, Jiashi Feng, Shuicheng Yan, and Eric P. Xing. Interpretable structure-evolving lstm. In *CVPR*, pages 2175–2184, 2017.

[23] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, pages 125–143. Springer, 2016.

[24] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, pages 3185–3193, 2016.

[25] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. *In IEEE TPAMI*, 39(1):115–127, 2016.

[26] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, pages 1925–1934, 2017.

[27] Ting Liu, Tao Ruan, Zilong Huang, Yunchao Wei, Shikui Wei, Yao Zhao, and Huang Thomas. Devil in the details: Towards accurate single and multiple human parsing. *AAAI*, pages 4814–4821, 2019.

[28] Xinchen Liu, Meng Zhang, Wu Liu, Jingkuan Song, and Tao Mei. Braidnet: Braiding semantics and details for accurate human parsing. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 338–346. ACM, 2019.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[30] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[31] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *ECCV*, 2018.

[32] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, pages 502–517, 2018.

[33] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[35] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3126–3135, 2019.

[36] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. *arXiv preprint arXiv:1905.02188*, 2019.

[37] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan Yuille. Joint object and part segmentation using deep learned potentials. In *ICCV*, pages 1573–1581, 2015.

[38] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Songchun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018.

[39] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[40] Yang Wang, Tran Duan, Zicheng Liao, and David Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research*, 13(1):3075–3102, 2012.

[41] Zhen Wei, Yao Sun, Jinqiao Wang, Hanjiang Lai, and Si Liu. Learning adaptive receptive fields for deep image parsing network. In *CVPR*, pages 3947–3955, 2017.

[42] Fangting Xia, Peng Wang, Liang Chieh Chen, and Alan L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ICCV*, pages 648–663, 2015.

[43] Fangting Xia, Peng Wang, Xianjie Chen, and Alan Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPRW*, pages 6080–6089, 2017.

[44] Kota Yamaguchi, M. Hadi Kiapour, and Tamara L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, pages 3519–3526, 2013.

[45] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, Ming Tang, and Hanqing Lu. Tree hierarchical cnns for object parsing. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1588–1592. IEEE, 2018.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[47] Bingke Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Progressive cognitive human parsing. *In AAAI*, 2018.