SCIENCE CHINA
Information Sciences

December 2019, Vol. 62 220101:1–220101:14
https://doi.org/10.1007/s11432-019-9935-6

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Feature context learning for human parsing

Tengteng HUANG[1], Yongchao XU[1*], Song BAI[1], Yongpan WANG[2] & Xiang BAI[1]

[1]*School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[2]*Alibaba Group, Hangzhou* 311121*, China*

Received 3 June 2019/Accepted 13 June 2019/Published online 11 November 2019

Abstract Parsing inconsistency, referring to the scatters and speckles in the parsing results as well as imprecise contours, is a long-standing problem in human parsing. It results from the fact that the pixel-wise classification loss independently considers each pixel. To address the inconsistency issue, we propose in this paper an end-to-end trainable, highly flexible and generic module called feature context module (FCM). FCM explores the correlation of adjacent pixels and aggregates the contextual information embedded in the real topology of the human body. Therefore, the feature representations are enhanced and thus quite robust in distinguishing semantically related parts. Extensive experiments are done with three different backbone models and four benchmark datasets, suggesting that FCM can be an effective and efficient plug-in to consistently improve the performance of existing algorithms without sacrificing the inference speed too much.

Keywords human parsing, context learning, fully convolutional networks, graph convolutional network, semantic segmentation

Citation Huang T T, Xu Y C, Bai S, et al. Feature context learning for human parsing. Sci China Inf Sci, 2019, 62(12): 220101, https://doi.org/10.1007/s11432-019-9935-6

1 Introduction

Human parsing is a particular segmentation task, aiming to segment a human body into fine-grained semantic parts. In recent years, human parsing has received growing interests in computer vision community due to its potential applications in human behavior analysis [1], human fashion [2], and person re-identification [3]. Benefiting from the success of fully convolutional networks [4] in semantic segmentation [5,6], significant progresses have been achieved by adapting convolutional neural networks to human parsing.

However, unlike general semantic segmentation or instance segmentation, human parsing is essentially a fine-grained segmentation task. Because the semantically related body parts are quite difficult to be distinguished, the parsing results usually suffer from the problem of inconsistency, that is, there exist scatters, speckles and imprecise boundaries as illustrated in Figure 1. The torso should be predicted as upper clothes but partially misclassified as coat. Moreover, the body contours and boundaries of semantic parts are too sensitive to be accurately predicted, resulting in non-smooth contours of the hands, as well as the imprecise boundary between socks and shoes.

In literature, Luo et al. [7] has suggested that the pixel-wise classification loss is responsible for such an inconsistency. To remedy this, previous work can be coarsely divided into three categories, including (1) methods using conditional random fields (CRFs) (e.g., [8]), (2) methods using adversarial learning

* Corresponding author (email: yongchaoxu@hust.edu.cn)

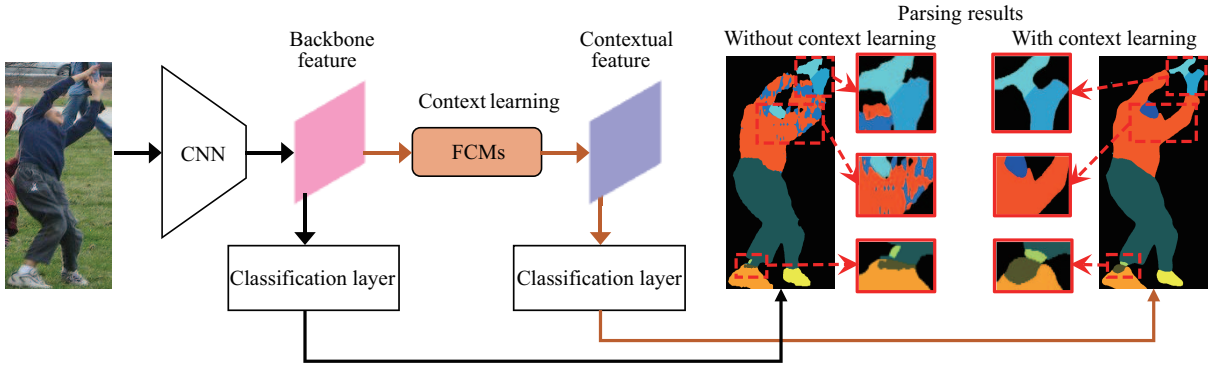


Figure 1 (Color online) Overall pipeline of the proposed method. FCMs are inserted into a human parsing model (solid black path), resulting in a more robust contextual feature and a more consistent parsing result.

(e.g., [7]), and (3) methods using extra information such as joints [9] and edges [10, 11] as a guidance for the parsing procedure. As indicated in their papers, the topology of the human body, which can be roughly represented by joints or edges, is of great importance for alleviating the inconsistency issue.

Inspired by the success of graph convolutional network [12, 13] in modeling the topology of graphs by exploiting the connection relations of adjacent nodes, we propose to represent the topology of human body as graphs. For a 2D image containing human bodies, we regard each pixel as a node and leverage a learning module to predict the correlations between neighboring nodes. These functions are implemented inside a network component named feature context module (FCM). FCM does not require extra annotations and only slightly increases inference time, different from [9, 10]. As shown in Figure 1, our model can generate consistent parsing results within a certain body part (e.g., the torso) and successfully corrects the unreasonable spatial relation of body parts (e.g., the head), demonstrating the superiority of FCM in correlating neighboring nodes and improving the human parsing results.

We give a thorough evaluation of FCM on four challenging benchmark datasets with three different backbone models, and witness a remarkable performance improvement with a minor increase of inference time. The consistent performance improvements suggest FCM is effective in addressing the inconsistency issue in human parsing. Moreover, FCM possesses three distinct advantages over the competing approaches, including (1) easy to implement: it requires no extra annotations (e.g., joints) as opposed to [9], and is end-to-end trainable compared to the CRF-based method [8]; (2) efficiency: FCM is a highly flexible model which can be plugged into most existing CNN-based parsing models to further improve the performance with negligible extra time overhead; (3) generalization: FCM shows favorable generalization capability on a challenging video surveillance dataset and is robust to heavy occlusions.

2 Related work

Fully convolutional networks (FCNs) [4] have demonstrated excellent performances in dense prediction tasks, and greatly promoted the development of segmentation approaches, such as PSPNet [5] and DeepLab [8]. Human parsing [14–18] is a fine-grained segmentation task [19–21], which targets at segmenting human bodies into semantic parts. It is receiving more and more attention owing to its potential importance in various human analysis tasks. In this section, we shortly review some human parsing methods and related methods.

Conditional random field (CRF) [22] is widely adopted as a post-processing procedure [8]. However, CRFs are usually time-consuming and sensitive to visual appearance changes as pointed out in [23]. Inspired by the success of generative adversarial networks (GAN) [24], Luo et al. [7] introduced adversarial loss into the human parsing task to assess whether a parsing result is reasonable. They employed two separate discriminators to correct the inconsistencies from global and local perspectives. However, large varieties in poses, occlusions and backgrounds bring great difficulties to judge the rationality of the parsing result, especially when multiple person instances are crowded in an image.

Some efforts are made to take advantage of extra guidance, such as joints and edges. Gong et al. [25] proposed a self-supervised strategy to align parsing map with human joint structures. Nie et al. [9] introduced a mutual learning to adapt (MuLA) scheme for boosting the performance of human parsing and pose estimation simultaneously in a multi-task learning manner. Gong et al. [10] employed an edge prediction branch to help distinguishing crowded person instances. However, these methods either require extra annotations or significantly increase the inference time.

Graph convolutional network (GCN) [12, 13, 26] is widely applied to deal with data that cannot be represented in a regular grid-like structure. Examples are social networks and 3D meshes. Kipf et al. [12] proposed a classic convolutional architecture for graph convolutional network via localized first-order approximation of spectral graph convolutions, which can effectively encode the graph structure. Graph convolutional network takes the node features and corresponding adjacent matrix as input, and outputs enriched node features by aggregating the contextual information of connected nodes. The proposed FCM mainly follows the design of GCN. We view each pixel in the 2D image as a node in a graph and employ convolution layers to estimate the affinity of adjacent nodes, forming the adjacent matrix. We then aggregate the contextual information in a weighted sum manner guided by the estimated adjacent matrix.

Our method is similar to [27] in modelling the topology of human body. They used a recurrent neural network (RNN) to learn the dependency of adjacent superpixels. The main difference is that we estimate the affinity of adjacent neighbors in an efficient convolutional manner on the feature maps instead of superpixels. FCM is also related to previous work aiming at making rational use of context information. Zhang et al. [28] proposed a context encoding module to selectively highlight the channels of feature maps. In comparison, FCM incorporates neighboring context from spatially adjacent pixels. Wang et al. [29] proposed a non-local operation to capture global long-range dependencies. Huang et al. [30] proposed criss-cross attention which was a smart variant of the non-local operation and aimed at aggregating global long-range contextual information in a more efficient way. Criss-cross attention can harvest the context information along the criss-cross path for each pixel. By stacking two criss-cross attention models, global contextual information from all the spatial positions can be aggregated, which is similar to the non-local operation [29]. FCM differently estimates the correlations between a pixel and its adjacent pixels in a local neighborhood instead of aggregating context from all positions, avoiding the intensive matrix operations (e.g., inner product) involved in the non-local operation. Besides, experimental results also indicate that a large context might be harmful to human parsing results.

3 Proposed approach

As illustrated in Figure 1, most human parsing methods build upon the powerful CNN model to learn a backbone feature map F_b, which is followed by several differentiable operations (e.g., upsampling, convolution) to generate the pixel-wise prediction. However, such an individual pixel-wise prediction ignores the correlation between adjacent pixels, leading to an inconsistent parsing result.

As humans share a common topological structure of the body, the contextual information along the topology of the human body is greatly helpful to more accurately parse each pixel and enforce the consistency of the parsing result. Inspired by GCN [13], we propose the FCM that outputs a contextual feature map F_c incorporating the aforementioned context. Note that FCMs can be plugged into most existing CNN-based human parsing models, and trained in an end-to-end manner.

In the following, we first describe the detail of FCM in Subsection 3.1, then analyze the effect of stacking multiple FCMs in Subsection 3.2. The training objective is given in Subsection 3.3.

3.1 Feature context module

The motivation of our work is that aggregating contextual information embedded in the topology of the human body is helpful for detecting and correcting inconsistency in parsing results. Inspired by graph convolutional network [13], we view a 2D image as a graph whose nodes are composed of all pixels in the

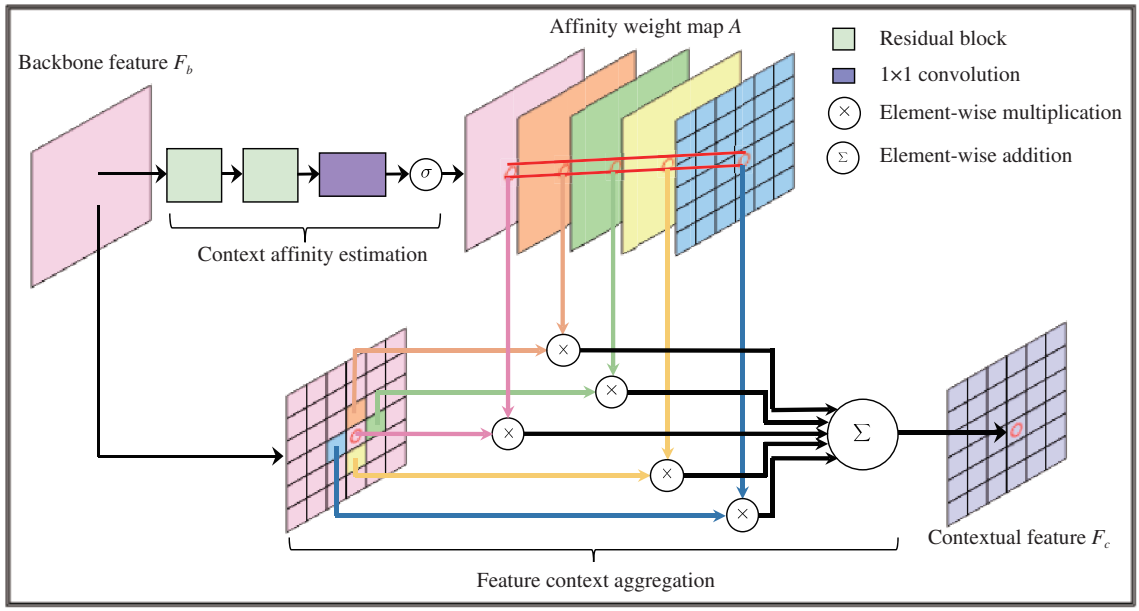


Figure 2 (Color online) Illustration of feature context module (FCM). FCM consists of two parts: (1) context affinity estimation which computes the correlation of each pixel with its adjacent pixels (e.g., 4 adjacent pixels under 4-connectivity and the pixel itself), resulting in a 5-channel affinity weight map; (2) feature context aggregation which enriches the input feature with the contextual feature in a weighted sum manner guided by the affinity weight map.

corresponding image and model the affinity of neighboring nodes[1)]. Concretely, we design an FCM which consists of two parts, i.e., context affinity estimation and feature context aggregation. Context affinity estimation dynamically estimates the correlation between each node and its adjacent node in the context, which reveals the cues of the topological structure. Based on the estimated affinity values, feature context aggregation enriches the feature of each node with the contextual features of its neighbors. The pipeline of FCM is given in Figure 2 and detailed as follows.

Context affinity estimation. Suppose we have an input backbone feature map F_b of size $D\times H\times W$, where D is the channel dimension, H is the height, and W is the width of the feature map, respectively. That is, there are $H\times W$ nodes in total. Context affinity estimation propagates the visual features in a convolutional manner to estimate the correlation between each node and its adjacent nodes.

In more detail, the context affinity estimation outputs an affinity weight map A of size $(K+1)\times H\times W$, where K denotes the number of nodes in the considered context. The first channel of A measures the self-affinity of each node and the rest K channels correspond to the estimated correlation of each node to the K adjacent nodes. Adjusting the self-affinity helps to correct the inconsistency when the node itself shows distinct features compared to its neighboring nodes. A natural choice to define the adjacent nodes in the context is the standard 4-connectivity or 8-connectivity. For a given node, 4-connectivity refers to its adjacent nodes in the directions of up, down, left and right. As for the 8-connectivity, four extra nodes in the directions of upper left, upper right, lower left and lower right are taken into consideration. Different channels and neighboring nodes form a bijection, ensuring the affinity weight learning for different neighbors. In this paper, unless explicitly stated, we use the 4-connectivity (i.e., K = 4) to define the context in our experiments.

To dynamically compute the affinity map A, we adopt a simple network consisting of two residual blocks, a 1×1 convolution and a sigmoid activation function. The residual blocks provide the local cues to identify the potential positions of inconsistency for robust context affinity estimation. The following 1×1 convolution outputs the correlations between each node and its adjacent nodes by mapping the feature maps to $K+1$ channels. We then adopt the sigmoid function to scale the affinity weight into the range of [0, 1]. As for the architecture of residual block, we employ the bottleneck architecture [31].

1) We do not differentiate node and pixel in the following since they have identical meaning in our case.

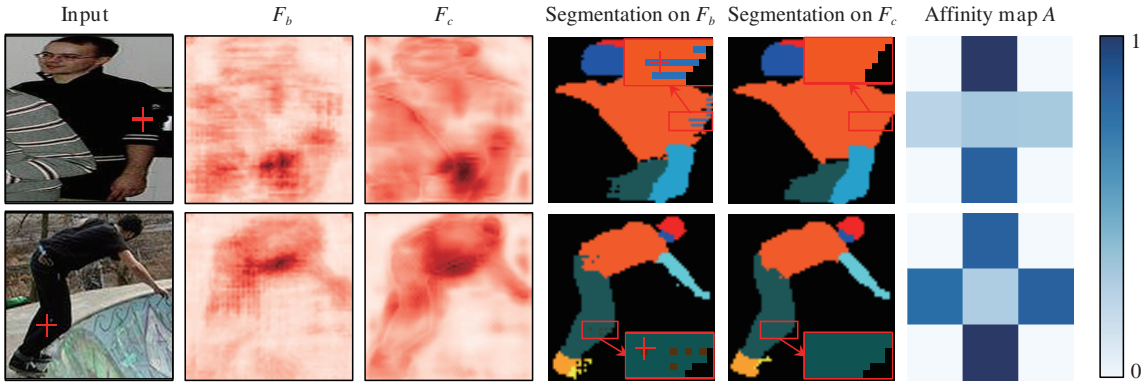


Figure 3 (Color online) Visualization of the effect of FCM on the feature maps. The right column shows the learned affinity weight map ($K = 4$) for the positions marked by red crisscross in input images. Parsing result on F_b exhibits inconsistency on the marked positions. The proposed FCM tends to assign higher weights to correctly parsed context pixels of the same semantics (e.g., up and down neighbors in the first row, and all four neighbors except the pixel itself in the second row), alleviating the inconsistency problem.

Feature context aggregation. With the affinity map A generated by the context affinity estimation part, feature context aggregation aims to enrich the input feature map F_b with the contextual information parameterized by A. Concretely, for each spatial position p in the input feature map, the output feature map F_c on p is given by

$$F_c(p) = \sum_{k=1}^{K+1} A_p(k) \times F_b\left(\mathcal{N}_p(k)\right), \tag{1}$$

where $\mathcal{N}_p$ is the set of adjacent nodes of p in the context, and $A_p(k)$ is the learned affinity between p and its k-th adjacent node $\mathcal{N}_p(k)$. Recall that the cardinality of the set $\|\mathcal{N}_p(k)\| = K+1$, and the first element in $\mathcal{N}_p$ is the node p itself. For boundary pixels, zero-padding is applied so that each pixel possesses K neighbors. Note that padded zeros do not contribute to F_c in the weighted sum process (Eq. (1)).

The design of our feature context module is similar to graph convolutional network. Each pixel in the feature map can be analogized to a node in the graph, and the estimated affinity weight map can be analogized to the adjacent matrix of the corresponding graph. The difference is that our affinity map is dynamically estimated while the adjacent matrix is fixed for GCN [12]. By considering the correlation with context, the feature context module provides a more robust image feature for parsing human body and encourages highly-correlated nodes to exhibit similar features, thus alleviating the inconsistency issue in human parsing.

Although each pixel is only correlated to local adjacent neighbors, distant pixels are also implicitly connected by propagating through a chain of edges, which encode the human topology in this sense. We present visualizations of learned feature maps and affinity maps in Figure 3. Compared with the input backbone feature map F_b, the resulted feature map F_c possesses a clear delimitation between the foreground person and the background, as well as smooth boundaries (better visualized by zooming in the electronic version of Figure 3) of different body parts, demonstrating that FCM effectively explores the context for understanding the topology of the human body. Besides, the learned affinity maps further shed light on how FCM works. The context affinity estimation works as a gating mechanism assigning high weight for closely correlated neighbors which are correctly classified and low weight for noising ones.

3.2 Stacking feature context

With a single feature context module, each node only incorporates the context from its four closest adjacent nodes when $K = 4$. Considering a broader context may be more helpful for revealing the topology of the human body. To this end, similar to classical deep CNNs which stack convolutional layers with small convolutional kernels to enlarge the perceptive field [31], we also stack M feature context modules to widen the context to be considered. With two FCMs, the correlation of each node

with its surrounding 12 adjacent nodes is considered. We will discuss how the size of context region influences the parsing performance in Subsection 4.8.

3.3 Training objective

In the human parsing task, the pixel-wise classification loss based on cross-entropy is widely-used, which is formulated as

$$\mathcal{L}_b = -\sum_{i=1}^{H\times W}\sum_{k=1}^{C} y_{ik}\log \bar{y}_{ik}, \tag{2}$$

where C is the total number of classes, and $\bar{y}_{ik}$ denotes the learned probability of the i-th pixel classified to the k-th category. y_{ik} is the label of the i-th pixel. If the i-th pixel belongs to the k-th category, then $y_{ik} = 1$, otherwise 0.

In addition to the standard pixel-wise classification loss based on the backbone feature F_b, we employ another supervision on the prediction from the contextual feature F_c given by FCMs. Similar to (2), we can define the objective function $\mathcal{L}_c$ via the cross-entropy loss. The final training objective of the proposed method is

$$\mathcal{L} = \lambda_b \times \mathcal{L}_b + \lambda_c \times \mathcal{L}_c, \tag{3}$$

where $\mathcal{L}_b$ and $\mathcal{L}_c$ represent losses corresponding to the prediction on the backbone feature map F_b and the contextual feature map F_c, respectively. λ_b and λ_c are weight constants to balance the contribution of the two losses.

4 Experiments

To validate the effectiveness of the proposed method, we conduct experiments on four widely-used datasets for human parsing, including LIP, PASCAL-Person-Part, CIHP and PPSS. We compare the performances with other state-of-the-art algorithms on the first three datasets and further assess the generalizability on the PPSS dataset.

4.1 Datasets and evaluation metric

LIP [25] is a large-scale benchmark dataset for single-person semantic parsing, which consists of 30462 training images, 10000 validation, and 10000 testing images. All images are parsed into a background class and 19 human parts: hat, hair, sunglasses, upper-clothes, dress, coat, socks, pants, gloves, scarf, skirt, jumpsuits, face, right arm, left arm, right leg, left leg, right shoe, left shoe. LIP is quite challenging due to the various poses and severe occlusions. Besides, many categories are semantically related (e.g., coat and upper-clothes), making it more difficult.

PASCAL-Person-Part [32] is a relatively small multi-instance dataset with 2.2 persons per image in average. It has 1716 training images and 1817 testing images annotated into 6 human parts: head, torso, upper arms, lower arms, upper legs, lower legs, and one background class. Scarce image samples and crowded persons in various scales pose challenges for human parsing on this dataset.

CIHP [10] is a recent large-scale multi-instance dataset, containing 28280 training images, 5000 validation and 5000 testing images. All images are annotated in the same way as the LIP dataset does. CIHP is more challenging than LIP as many images therein have multiple crowded persons. This dataset has 3.4 instances per image in average.

PPSS [33] provides 3673 images annotated into seven human parts and a background class. The images are collected from 171 surveillance videos of different scenes varied in background, occlusion, and illumination. Hence, it can serve as a benchmark dataset to evaluate the generalization ability of human parsing algorithms.

Metric. We employ intersection-over-union (IoU) as the evaluation metric, and report both mean intersection-over-union (mIoU) and IoU for each class.

4.2 Implementation details

Backbone networks. Three backbone networks are used, including (1) DeepLab-v2 [8] framework based on ResNet-101 [31] pre-trained on the MS-COCO dataset [34], (2) MMAN [7] to test the complementarity of FCMs with adversarial, and (3) PGN [10] to further validate the complementarity of FCMs to methods utilizing extra guidance in a multi-task training manner. We remove the image pyramid scheme in the original DeepLab-v2 by feeding only images at the original scale. For all these three backbone networks, we plug two feature context modules before their prediction layers and take the parsing results based on F_c as the output.

Training and testing details. In the implementation of FCM upon the backbone DeepLab-v2, we first resize the training images to 572×572, and then sample an image patch of size 512×512 for data augmentation. Random horizontal flip is also applied. We use SGD optimizer, set momentum to 0.9 and weight decay to 0.0005, respectively. The initial learning rate is set to 0.01. We train the model on the LIP and CIHP dataset for 30 epochs, and the learning rate is divided by 10 after 15 epochs. On the PASCAL-Person-Part dataset, we train the model for 50 epochs, and the learning rate is divided by 10 after 25 epochs. During inference, multi-scale predictions by feeding the network with images at scales of {0.8, 1.0, 1.2} are combined following [7].

In the implementation of FCM upon the backbone MMAN and PGN, we use the publicly released codes of MMAN[2)] and PGN[3)], and follow their default training and testing settings. In all the experiments, we set both λ_b and λ_c to 1.0 in (3).

4.3 Experiment on LIP

We first evaluate our method on the LIP dataset [25], which is a single-person semantic parsing benchmark. The quantitative comparison with other state-of-the-art methods is presented in Table 1 [4, 7–11, 25,35,36]. By inserting FCM to DeepLab-v2, we achieve mIoU 51.23%, outperforming the state-of-the-art methods SSL [25] by 6.50%, MMAN [7] by 4.42%, PGN [10] by 2.72%, and MuLA [9] by 1.93%, respectively. The performance improvement over MuLA is more valuable as MuLA makes use of extra joint annotations while FCM does not. As PGN [10] is specifically designed for parsing multiple persons, the authors did not report experimental results on the LIP dataset. The results of PGN are reproduced using their released source codes. Compared with the baseline DeebLab-v2, FCM leads to an improvement of mIoU 3.32%. We also evaluate the performance of plugging FCM into MMAN, PGN or CE2P [11] before their final label prediction layers and observe a consistent performance improvement. Specifically, FCM boosts the performance of MMAN by 1.03%, that of PGN by 2.54%, and that of CE2P by 1.26%, respectively. Note that MMAN, PGN and CE2P have exploited other mechanisms to improve the performance. Concretely, MMAN proposes to alleviate the inconsistency by using adversarial learning. PGN and CE2P rely on an extra edge branch to refine the parsing result. Nevertheless, the performance improvement still demonstrates that FCM is complementary to these methods in alleviating the inconsistency issue.

In Figure 4, we exhibit some qualitative results of DeepLab-v2 and DeepLab-v2+FCM. As can be seen clearly, the inconsistency in a semantic part is alleviated with the aid of the proposed FCM. Furthermore, the boundary between different semantic parts is more precise, which firmly demonstrates the efficacy of FCM.

4.4 Experiment on PASCAL-Person-Part

We then evaluate the proposed FCM on the PASCAL-Person-Part dataset to assess the performance in parsing multiple persons. Table 2 [7–10, 18, 25, 27, 36] shows the quantitative comparison of different methods. Compared with the backbone DeepLab-V2, FCM achieves mIoU 67.05% with an improvement of 3.19%. Meanwhile, when using the backbone PGN, FCM reports mIoU 69.54% with an improvement of 1.14%. As PGN is particularly designed for parsing multiple-persons using contours between different

2) https://github.com/RoyalVane/MMAN.
3) https://github.com/Engineering-Course/CIHP_PGN.

Table 1 Quantitative comparison (IoU (%) for each class and mIoU (%)) with state-of-the-art methods on the LIP validation set[a)b)]

Method	Hat	Hair	Glov	Sung	Clot	Dress	Coat	Sock	Pant	Suit
SegNet [35]	26.60	44.01	0.01	0.00	34.46	0.00	15.97	3.59	33.56	0.01
FCN-8s [4]	39.79	58.96	5.32	3.08	49.08	12.36	26.82	15.66	49.41	6.48
Attention [36]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73
SSL [25]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48
MuLA [9][c)]	–	–	–	–	–	–	–	–	–	–
MMAN [7]	57.66	65.63	30.07	20.02	64.15	28.39	51.98	41.46	71.03	23.61
MMAN+FCM	60.58	68.98	30.78	25.00	65.01	29.40	51.57	43.09	70.19	21.77
PGN [10][d)]	61.53	69.13	34.13	26.99	68.17	**34.93**	55.78	42.50	70.69	25.30
PGN+FCM	64.00	70.61	36.74	30.88	68.66	33.42	55.92	46.67	71.99	27.54
DeepLab [8]	59.46	67.54	32.62	25.49	65.78	31.94	55.43	39.80	70.45	24.70
DeepLab+FCM	65.70	71.32	37.96	33.37	68.26	33.74	54.96	47.79	72.58	28.43
CE2P [11]	65.29	72.54	39.09	32.73	69.46	32.52	**56.28**	49.67	74.11	27.23
CE2P+FCM	**66.31**	**73.58**	**40.21**	**34.03**	**70.69**	33.27	55.63	**50.62**	**75.32**	**29.83**

Method	Scarf	Skirt	Face	l-arm	r-arm	l-leg	r-leg	l-sh	r-sh	bkg	mIoU
SegNet [35]	0.00	0.00	52.38	15.30	24.23	13.82	13.17	9.26	6.47	70.62	18.17
FCN-8s [4]	0.00	2.16	62.65	29.78	36.63	28.12	26.05	17.76	17.70	78.02	28.29
Attention [36]	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
SSL [25]	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.73
MuLA [9][c)]	–	–	–	–	–	–	–	–	–	–	49.30
MMAN [7]	9.65	23.20	69.54	55.30	58.13	51.90	52.17	38.58	39.05	85.75	46.81
MMAN+FCM	10.63	20.41	72.56	58.02	60.75	52.13	51.61	39.25	39.80	85.28	47.84
PGN [10][d)]	16.05	24.79	73.74	59.33	60.78	47.47	46.62	32.74	33.75	85.67	48.51
PGN+FCM	21.60	24.42	73.49	61.76	63.14	52.13	50.93	40.00	40.45	86.58	51.05
DeepLab [8]	15.51	**28.13**	70.53	55.76	58.56	48.99	49.49	36.76	36.79	85.49	47.91
DeepLab+FCM	**23.53**	26.16	74.38	60.11	62.71	50.01	49.46	38.41	38.90	86.77	51.23
CE2P [11]	14.19	22.51	75.50	65.14	66.59	60.10	**58.59**	46.63	46.12	87.67	53.10
CE2P+FCM	20.66	21.93	**76.32**	**67.73**	**68.17**	**61.09**	58.27	**47.52**	**47.38**	**88.56**	**54.36**

a) Glov, Sung, Clot, Sock, Pant, Suit, l-arm, r-arm, l-leg, r-leg, l-sh, r-sh, bkg represent gloves, sunglasses, clothes, socks, pants, jumpsuits, left arm, right arm, left leg, right leg, left shoe, right shoe, background class, respectively.

b) Bold typeface indicates the best performance for each category.

c) Extra annotations are used.

d) The results are reproduced using the source code released by the authors.

instances, the improvement of FCM over PGN is less dramatic than that over DeepLab-v2, but it still reveals that FCM is complementary to PGN in parsing multiple humans. It can be observed that MMAN does not work well on this dataset in Table 2. A possible interpretation is that adversarial training cannot effectively align multiple persons simultaneously. Therefore, we do not evaluate the performance of combining FCM with MMAN. Moreover, compared with other state-of-the-art methods, FCM with PGN outperforms SSL [25] by 10.18%, MMAN [7] by 9.63%, Graph LSTM [37] by 9.38%, structure-evolving LSTM [27] by 5.97%, and MuLA [9] using extra joint annotations by 4.44%, respectively. Compared with superpixel-based methods like Graph LSTM [37] and structure-evolving LSTM [27], FCM performs directly on the pixels in the feature space with a convolutional manner, thus making it more efficient for inference.

Some parsing results using DeepLab-v2+FCM are depicted in Figure 4. Qualitatively, the observations on the LIP dataset also hold for the PASCAL-Person-Part dataset. By applying FCM, the inconsistency problems are considerably mitigated in parsing multiple humans. Besides, the boundary between different parts is more precise.

4.5 Experiment on CIHP

We then further evaluate the proposed method on a recent released multi-person dataset CIHP [10] con-

Figure 4 (Color online) Some qualitative illustrations on LIP in the red box, PASCAL-Person-Part in the green box, and CIHP in the blue box. For each example, from left to right: input image, groundtruth, result given by DeepLab-v2 baseline, and result of DeepLab-v2+FCM. The red rectangle and oval markers refer to inconsistency regions containing scatters and imprecise boundaries, respectively.

Table 2 Quantitative evaluation (IoU (%) for each class and mIoU (%)) on the PASCAL-Person-Part test set[a)b)]

Method	Head	Torso	u-arm	l-arm	u-leg	l-leg	bkg	mIoU
Attention [36]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
SSL [25]	83.26	62.40	47.80	45.58	42.32	39.48	94.68	59.36
MMAN [7]	82.58	62.83	48.49	47.37	42.80	40.40	94.92	59.91
Structure-evolving LSTM [27]	82.89	67.15	51.42	48.72	51.72	45.91	**97.18**	63.57
DeepLab-ASPP [8]	–	–	–	–	–	–	–	64.94
MuLA [9][c)]	–	–	–	–	–	–	–	65.10
PCNet [18]	86.81	69.06	55.35	55.27	50.21	**48.54**	96.07	65.90
DeepLab [8]	85.67	67.12	54.00	54.41	47.06	43.63	95.16	63.86
DeepLab+FCM	86.42	70.37	**59.57**	59.35	51.22	47.12	95.33	67.05
PGN [10]	90.89	75.12	55.83	64.61	55.42	41.57	95.33	68.40
PGN+FCM	**91.16**	**76.45**	57.77	**66.23**	**56.58**	43.21	95.41	**69.54**

a) u-arm, l-arm, u-leg, l-leg, bkg represent upper arms, lower arms, upper legs, lower legs, background class, respectively.
b) Bold typeface indicates the best performance for each category.
c) Extra annotations are used.

taining crowed persons, which is much more challenging for human parsing. The quantitative comparison is given in Table 3 [8,10]. As it shows, consistent improvements have been achieved for both two baseline methods, i.e., DeepLab-v2 and PGN. It should be mentioned that PGN is the baseline method proposed along with the release of the dataset [10], which is customized to fit for the multi-instance case. They employ an extra edge prediction branch to train together with the parsing branch. The edge information provides the precise delimitations of different persons, which is helpful for distinguishing multiple instances. Besides, PGN also combines feature maps from multiple layers. With those improvements, PGN achieves superior performance on the CIHP dataset at the cost of significantly increasing the inference time. Even so, FCM still yields a performance gain of 1.30% using PGN [10] as the backbone,

Table 3 Quantitative evaluation (IoU (%) for each class and mIoU (%)) on the CIHP validation set[a)b)c)]

Method	Hat	Hair	Glov	Sung	Clot	Dress	Coat	Sock	Pant	Suit	
DeepLab [8]	65.70	78.89	22.33	50.81	64.18	52.28	62.57	30.95	70.60	70.23	
DeepLab+FCM	**70.11**	**81.71**	25.04	**56.92**	**68.95**	**56.76**	**65.61**	**33.32**	**73.75**	**73.99**	
PGN [10]	68.78	78.61	23.07	47.14	67.42	55.11	63.97	26.68	72.16	71.56	
PGN+FCM	65.41	79.04	**30.57**	54.19	67.16	54.36	63.72	30.74	73.04	71.93	
Method	Scarf	Skirt	Face	l-arm	r-arm	l-leg	r-leg	l-sh	r-sh	bkg	mIoU
DeepLab [8]	28.42	37.53	86.60	24.69	31.88	25.91	22.66	19.15	18.56	92.01	47.80
DeepLab+FCM	30.64	37.98	**88.57**	29.98	34.06	30.73	21.13	21.80	20.13	**93.82**	50.75
PGN [10]	29.55	**38.56**	86.54	67.88	68.94	54.21	54.62	38.60	38.74	93.21	57.27
PGN+FCM	**33.41**	37.74	86.86	**68.20**	**69.19**	**55.75**	**56.07**	**40.00**	**40.66**	93.37	**58.57**

a) Glov, Sung, Clot, Sock, Pant, Suit, l-arm, r-arm, l-leg, r-leg, l-sh, r-sh, bkg represent gloves, sunglasses, clothes, socks, pants, jumpsuits, left arm, right arm, left leg, right leg, left shoe, right shoe, background class, respectively.

b) Bold typeface indicates the best performance for each category.

c) Results for PGN are evaluated with the model released by the authors.

Table 4 Analysis of inference time (s) on the LIP dataset

Backbone	Inference time	+FCM	Increase time
MMAN [7]	0.048	0.051	↑0.003
PGN [10]	0.231	0.243	↑0.012
DeepLab [8]	0.021	0.025	↑0.004

further demonstrating the complementarity of FCM to methods utilizing extra guidance for multi-task training. Besides, FCM achieves a performance improvement of 2.95% in terms of mIoU compared with the backbone DeepLab-v2.

Figure 4 depicts some qualitative parsing results based on DeepLab-v2. Similar to the observation on LIP and PASCAL-Person-Part datasets, the proposed FCM alleviates the inconsistency problem and results in more precise part delimitations for crowded human parsing.

4.6 Inference speed

As can be concluded above, FCM achieves consistent performance gain on three benchmark datasets for human parsing using three different backbone networks. We demonstrate here that FCM is a rather computational-cheap plug-in and requires a minor increase of inference time. As different image preprocessing strategies lead to different inference speed, the evaluation is done using the images of the LIP dataset [25] in the original size for fair comparison.

The analysis of average inference speed is presented in Table 4 [7, 8, 10]. The extra inference cost brought by FCM is almost ignorable compared with that of the backbone networks. This suggests that FCM can be used as a quite efficient module to further improve the performance of various human parsing algorithms. Meanwhile, it is worth noting that DeepLab-v2+FCM is a good compromise between efficacy and efficiency. It achieves superior performance in most experimental settings and can be executed in a real-time manner.

4.7 Cross-dataset experiment on PPSS

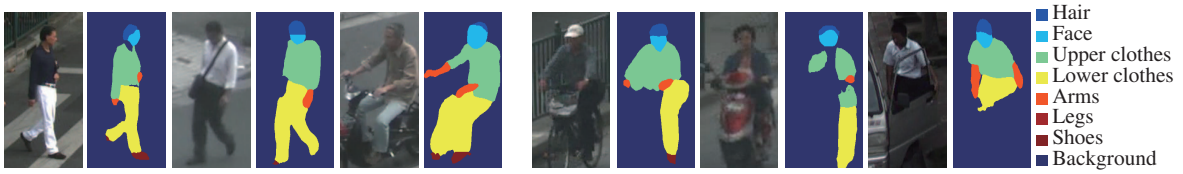
We evaluate the generalization ability of our method by directly applying the model trained on the LIP dataset to the PPSS testing dataset without any fine-tuning. Following [7], we merge semantically related categories of the LIP dataset to ensure that it has the same category definition as the PPSS dataset.

The quantitative evaluation is given in Table 5 [7, 8, 10, 33, 38]. FCM significantly improves the generalization ability of the trained model to unseen images. Specifically, FCM improves MMAN by 12.0%, PGN by 6.1% and DeepLab-v2 by 4.3% in terms of mean accuracy, respectively. Several representative parsing results are shown in Figure 5. Even without fine-tuning on the PPSS dataset, DeepLab-v2+FCM correctly parses person images. Especially for those under heavy occlusion conditions, such as bikes and

Table 5 Cross-dataset evaluation (IoU (%) for each class and mIoU (%)) on the PPSS test set using the model trained on the LIP dataset[a)b)]

Method	Hair	Face	u-c	Arms	l-c	Legs	bkg	mIoU
DL [33]	22.0	29.1	57.3	10.6	46.1	12.9	68.6	35.2
DDN [33]	35.5	44.1	68.4	17.0	61.7	23.8	80.0	47.2
ASN [38]	51.7	51.0	65.9	29.5	52.8	20.3	83.8	50.7
MMAN [7]	53.1	50.2	69.0	29.4	55.9	21.4	85.7	52.1
MMAN+FCM	60.0	70.7	75.5	62.6	43.0	**42.7**	94.4	64.1
PGN [10]	55.5	62.4	70.3	56.3	29.3	24.4	97.9	56.6
PGN+FCM	62.0	67.4	74.0	64.3	39.2	35.1	96.8	62.7
DeepLab [8]	**65.8**	59.5	**84.5**	**76.3**	35.0	25.6	90.4	62.4
DeepLab+FCM	64.1	**72.6**	80.6	67.3	**48.7**	39.1	**94.8**	**66.7**

a) u-c, l-c and bkg represent upper-clothes, lower-clothes and background class, respectively.
b) Bold typeface indicates the best performance for each category.



(a) (b)

Figure 5 (Color online) Examples of parsing results on the PPSS testing dataset of DeepLab-v2+FCM. The model is trained on the LIP dataset without further finetuning. Severe occlusion is present for images in (b). (a) Normal cases; (b) cases with severe occlusion.

Table 6 Ablation study on the LIP dataset[a)]

Method	mIoU (%)
DeepLab	47.91
DeepLab + extra resblocks	48.07
DeepLab + CRF [22]	48.53
DeepLab + Non-local [29]	49.48
DeepLab + FCM ($\lambda_b = 0$)	50.76
DeepLab + FCM	51.23
DeepLab + FCM [b)]	**51.65**

a) Bold typeface indicates the best performance.
b) This method employs FCM in both the last and the penultimate stage of DeepLab-v2.

cars in Figure 5(b), DeepLab-v2+FCM still outputs clean segmentation maps. The qualitative and quantitative results consistently demonstrate that FCM is robust to complex backgrounds, occlusions, and various illuminations, showing its potential capability of handling diverse real video surveillance scenes.

4.8 Ablation study

In this subsection, we present some ablation experiments of the proposed method. All experiments are conducted on the LIP dataset with DeepLab-v2 as the backbone network. We first remove the auxiliary loss by setting $\lambda_b = 0$ to show its affects, and then discuss how the size of contextual region for each node influences the model performance. We also further explore the effectiveness of FCM in shallower layers of the backbone network. Finally, we compare our method to other alternative methods, further validating the superiority of FCM.

Auxiliary loss. As depicted in Table 6 [22, 29], setting λ_b to zero only leads to a minor performance decrease of mIoU by 0.47%. This comparison suggests the main performance improvement with respect to the baseline is brought by FCM.

Size of contextual region. The size of contextual region involves two hyper-parameters, i.e., the

Table 7 Quantitative analysis of the contextual region incorporated for the parsing performance on the LIP dataset[a)b)]

M	K	Contextual region	mIoU (%)
1	4	4	49.13
1	8	8	50.04
2	4	12	**51.23**
2	8	24	49.95
3	4	24	50.16

a) For networks with stride 8 (e.g., DeepLab-v2), each pixel in the feature map corresponds to an 8×8 patch of the original image.

b) Bold typeface indicates the best performance.

number of stacked FCMs M and the number of nodes in the context K. We conduct experiments by varying M and K as shown in Table 7. It can be observed that the parsing performance achieves consistent improvement by varying the contextual region size from 4 to 12. However, taking more adjacent nodes into consideration cannot further bring performance gain. This is probably because incorporating too much context for some relatively small human body parts might involve irrelevant information, thus harming the performance. It should be noted that experiment settings with $M = 2, K = 8$ and $M = 3, K = 4$ share the same size of contextual region and achieve similar mIoU (49.95% vs. 50.16%), demonstrating that size of contextual region plays the key role for FCM.

FCM in shallower layers. We conduct experiments by integrating FCM into the penultimate stage of DeepLab-v2 (in addition to the last stage), and obtain a further improvement of 0.42% in mIoU (from 51.23% to 51.65%) on the LIP dataset as shown in Table 6. We also try to plug FCMs into shallower layers but do not observe obvious performance gain. A possible reason is that shallow layers mainly encode low-level cues which cannot provide semantic information for effectively exploring the correlation of adjacent pixels.

Comparison to alternatives. It is known that adding a sequence of convolutional layers will give the model a certain enhanced ability to use context information. To investigate the performance of such an alternative, we add more residual blocks to DeepLab-v2 as a baseline model which shares approximately the same parameter numbers as DeepLab-v2+FCM. From Table 6, adding more residual blocks only yields a marginal gain of mIoU (0.16%). Compared to DeepLab-v2+FCM, this method is exceptionally less effective. We then adopt CRF [22] as a post-procedure to refine the parsing results generated by DeepLab-v2. Besides, we replace FCM with the non-local module [29] which can aggregate global context information from all the positions. From the results presented in Table 6, FCM shows superior performance compared to CRF (51.23% vs. 48.53%) and the non-local module (51.23% vs. 49.48%). These results firmly demonstrate the superiority of FCM over other state-of-the-art methods in leveraging context information in the human parsing task.

Boundary precision. As shown in Figure 4, qualitatively, FCM leads to more precise boundaries between different semantic parts. We also propose a new metric named edge-masked mIoU to quantitatively measure the boundary precision. For that, we first generate edge labels for each image from its semantic annotation following CE2P [11]. We then evaluate the edge-masked mIoU by only considering pixels which fall on the edges and ignoring all others. The proposed method yields an improvement of 2.42% (33.39% vs. 30.97%) in terms of edge-masked mIoU compared to DeepLab-v2. This further demonstrates the effectiveness of FCM.

5 Conclusion

In this paper, we propose a feature context module to incorporate the context along the topology of the human body to improve the performance of human parsing. The learned features effectively alleviate the inconsistency issue and smooth the contour in human parsing. FCM is independent of the backbone network architecture and only incurs a slight increase of the inference time. Experiments with three different backbone networks demonstrate that FCM achieves consistent improvements over the backbone

models and very competitive performances on three benchmark datasets. Furthermore, FCM also shows the noteworthy generalization ability to unseen images collected from surveillance videos of multiple different real scenes.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2018YFB1004600), National Natural Science Foundation of China (Grant No. 61703171), and Natural Science Foundation of Hubei Province of China (Grant No. 2018CFB199). This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) Program. The work of Yongchao XU was supported by Young Elite Scientists Sponsorship Program by CAST. The work of Xiang BAI was supported by National Program for Support of Top-Notch Young Professionals and in part by Program for HUST Academic Frontier Youth Team.

References

1 Gan C, Lin M, Yang Y, et al. Concepts not alone: exploring pairwise relationships for zero-shot video activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2016. 3487–3493
2 Han X, Wu Z X, Wu Z, et al. Viton: an image-based virtual try-on network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2018. 7543–7552
3 Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2018. 1062–1071
4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440
5 Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017. 2881–2890
6 Zhou Y Y, Wang Y, Tang P, et al. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: Proceedings of 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 2019. 121–140
7 Luo Y W, Zheng Z D, Zheng L, et al. Macro-micro adversarial network for human parsing. In: Proceedings of European Conference on Computer Vision, 2018. 418–434
8 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 834–848
9 Nie X C, Feng J S, Yan S C. Mutual learning to adapt for joint human parsing and pose estimation. In: Proceedings of European Conference on Computer Vision, 2018. 502–517
10 Gong K, Liang X D, Li Y C, et al. Instance-level human parsing via part grouping network. In: Proceedings of European Conference on Computer Vision, 2018. 770–785
11 Liu T, Ruan T, Huang Z, et al. Devil in the details: towards accurate single and multiple human parsing. 2018. ArXiv: 1809.05996
12 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. 2016. ArXiv: 1609.02907
13 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2017. ArXiv: 1710.10903
14 Xia F T, Wang P, Chen X J, et al. Joint multi-person pose estimation and semantic part segmentation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017. 6769–6778
15 Fang H-S, Lu G S, Fang X L, et al. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2018
16 Liu S, Sun Y, Zhu D F, et al. Cross-domain human parsing via adversarial feature and label adaptation. 2018. ArXiv: 1801.01260
17 Liang X, Gong K, Shen X, et al. Look into person: joint body parsing & pose estimation network and a new benchmark. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 871–885
18 Zhu B K, Chen Y Y, Tang M, et al. Progressive cognitive human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018
19 Guo L H, Guo C G, Li L, et al. Two-stage local constrained sparse coding for fine-grained visual categorization. Sci China Inf Sci, 2018, 61: 018104
20 Sun H Q, Pang Y W. GlanceNets - efficient convolutional neural networks with adaptive hard example mining. Sci China Inf Sci, 2018, 61: 109101
21 Xu Y, Wang Y, Zhou W, et al. TextField: learning a deep direction field for irregular scene text detection. IEEE Trans Image Process, 2019. doi: 10.1109/TIP.2019.2900589
22 Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. In: Proceedings of Advances in Neural Information Processing Systems, 2011
23 Ke T-W, Hwang J-J, Liu Z W, et al. Adaptive affinity field for semantic segmentation. In: Proceedings of 2018 European Conference on Computer Vision. Berlin: Springer, 2018. 605–621
24 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, 2014
25 Gong K, Liang X D, Zhang D Y, et al. Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017. 932–940

26 Jin J W, Liu Z L, Chen C L P. Discriminative graph regularized broad learning system for image recognition. Sci China Inf Sci, 2018, 61: 112209

27 Liang X D, Lin L, Shen X H, et al. Interpretable structure-evolving LSTM. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2017. 1010–1019

28 Zhang H, Dana K, Shi J P, et al. Context encoding for semantic segmentation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2018. 7151–7160

29 Wang X L, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 7794–7803

30 Huang Z, Wang X, Huang L, et al. Ccnet: criss-cross attention for semantic segmentation. 2018. ArXiv: 1811.11721

31 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016. 770–778

32 Chen X J, Mottaghi R, Liu X B, et al. Detect what you can: detecting and representing objects using holistic models and body parts. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2014. 1971–1978

33 Luo P, Wang X G, Tang X O. Pedestrian parsing via deep decompositional network. In: Proceedings of IEEE International Conference on Computer Vision, 2013. 2648–2655

34 Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context. In: Proceedings of European Conference on Computer Vision, 2014

35 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495

36 Chen L-C, Yang Y, Wang J, et al. Attention to scale: scale-aware semantic image segmentation. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016. 3640–3649

37 Liang X, Shen X, Feng J, et al. Semantic object parsing with graph LSTM. In: Proceedings of European Conference on Computer Vision, 2016

38 Luc P, Couprie C, Chintala S, et al. Semantic segmentation using adversarial networks. In: Proceedings of NIPS Workshop, 2016