SCIENCE CHINA
Information Sciences

February 2021, Vol. 64 120103:1–120103:13
https://doi.org/10.1007/s11432-020-2969-8

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

PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection

Jin XIE[1], Yanwei PANG[1*], Hisham CHOLAKKAL[2], Rao ANWER[2], Fahad KHAN[2] & Ling SHAO[2]

[1]*Tianjin Key Laboratory of Brain-Inspired Artificial Intelligence, School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072*, China;*
[2]*Inception Institute of Artificial Intelligence, Abu Dhabi* 999041*, UAE*

Received 10 March 2020/Revised 9 May 2020/Accepted 28 June 2020/Published online 19 November 2020

Abstract Detecting pedestrians, especially under heavy occlusion, is a challenging computer vision problem with numerous real-world applications. This paper introduces a novel approach, termed as PSC-Net, for occluded pedestrian detection. The proposed PSC-Net contains a dedicated module that is designed to explicitly capture both inter and intra-part co-occurrence information of different pedestrian body parts through a graph convolutional network (GCN). Both inter and intra-part co-occurrence information contribute towards improving the feature representation for handling varying level of occlusions, ranging from partial to severe occlusions. Our PSC-Net exploits the topological structure of pedestrian and does not require part-based annotations or additional visible bounding-box (VBB) information to learn part spatial co-occurrence. Comprehensive experiments are performed on three challenging datasets: CityPersons, Caltech, and CrowdHuman datasets. Particularly, in terms of log-average miss rates and with the same backbone and input scale as those of the state-of-the-art MGAN, the proposed PSC-Net achieves absolute gains of 4.0% and 3.4% over MGAN on the heavy occlusion subsets of CityPersons and Caltech test sets, respectively.

Keywords pedestrian detection, graph convolutional network, occlusion, object detection, feature extraction

Citation Xie J, Pang Y W, Cholakkal H, et al. PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection. Sci China Inf Sci, 2021, 64(2): 120103, https://doi.org/10.1007/s11432-020-2969-8

1 Introduction

Pedestrian detection is a challenging problem in computer vision with various real-world applications, e.g., robotics, autonomous driving, and visual surveillance [1, 2]. Recent years have witnessed significant progress in the field of pedestrian detection, mainly owing to the advances in deep convolutional neural networks (CNNs) [3, 4]. Modern pedestrian detection methods can be generally classified into single-stage [5–7] and two-stage [8–18] categories. Single-stage pedestrian detectors typically work by directly regressing the default anchors into pedestrian detection boxes. Different from single-stage pedestrian detectors, two-stage methods first produce a set of candidate pedestrian proposals in the first stage and then classify and regress these proposals in the second stage. Most existing two-stage pedestrian detectors [8, 13–15, 17, 18] are adapted from the popular Faster R-CNN generic detection framework [19].

Though promising results have been achieved by existing pedestrian detectors on standard non-occluded pedestrians, their performance on heavily occluded pedestrians is far from satisfactory. This is evident from the fact that the best reported log-average miss rates [13] on the reasonable (R) set (where visibility ratio is larger than 65%) of CityPersons test set [1] is 9.3% whereas it is 41.0% on the heavy occlusion (HO) set (where visibility ratio ranges from 20% to 65%) of the same dataset. Handling pedestrian occlusion is an open problem in computer vision and presents a great challenge for detecting pedestrians in real-world practical applications owing to its frequent occurrence. Therefore, a pedestrian detector

* Corresponding author (email: pyw@tju.edu.cn)

is desired to be accurate with respect to varying level of occlusions, ranging from slightly occluded to severely occluded pedestrians.

A lot of occluded pedestrian detection approaches have been proposed in the past years. A common strategy to address occlusion is based on learning and integrating a set of part detectors [20–25]. Earlier part-based pedestrian detection approaches utilize body part annotations [25, 26]. Recently, the deployment of pre-trained part models was investigated [14] to exploit part correlations, typically relying on part detection scores corresponding to the visible regions of the pedestrian. Other methods [20–23, 27] utilize the bounding-box of the pedestrian and train a large number of independently learned part detectors. Alternatively, the topological structure of the pedestrian was also exploited [15] to avoid the reliance on body part annotations, leading to promising detection performance. However, it predominantly relies on the detection scores of parts to highlight visible regions of the pedestrian and it neither considers spatial co-occurrence relations between different body parts (e.g., head and arms) nor spatial co-occurrence relations between different sub-regions (e.g., eyes and ears of a head region) within a body part. The part spatial co-occurrence information is expected to enrich the feature representation by exploiting the information about the spatially adjacent parts. Knowledge about the typical configuration of objects (e.g., humans) in a scene and its impact on recognition performance was extensively studied in the field of both psychology and computer vision [28–30]. To the best of our knowledge, modern two-stage CNN-based pedestrian detectors do not explicitly encode the part spatial co-occurrence information. In this study, we introduce a data driven approach to handle occlusion problem that goes beyond part detection scores by explicitly integrating the part spatial co-occurrence information not only between different body parts but also between different sub-regions within a body part. Figure 1 shows detection results with occlusion level varying from slight to severe. Our PSC-Net is able to more accurately detect pedestrians, compared with both the baseline (Subsection 3.1) and the state-of-the-art MGAN [13] even there are heavy occlusions.

The contributions and the characteristics of the proposed method are as follows.

(1) We propose to utilize the cue of not only the co-occurrence of parts of a pedestrian (inter-part co-occurrence) but also the co-occurrence of sub-parts of a part (intra-part co-occurrence) for detecting occluded pedestrians. One intuition of applying inter-part occurrence is that if two or more parts (e.g., a head and an arm) co-occur when other parts are occluded then one can also infer that there is a pedestrian. Another intuition is that two or more parts can mutually support existence of each other. The intuition of applying intra-part occurrence is analogous.

(2) Both the inter-part occurrence information and intra-part occurrence information are adaptively modeled by graph convolution networks (GCNs). The module is called part spatial co-occurrence (PSC) module. With GCNs, it is not necessary for our method to explicitly and deterministically decide whether or not a part/sub-part is occluded. Therefore, the method can avoid the risk of mistakenly classification of a part/sub-part. Moreover, this makes our method not requiring distinguishing occluded regions from visible ones in the process of annotating a pedestrian for training.

(3) Integrating the proposed PSC module with a baseline detection network (e.g., Faster R-CNN [19]) results in remarkable improvement for detecting heavily occluded pedestrians and at the same time is significantly beneficial for improving the performance of detecting pedestrians with light or no occlusion (a.k.a., reasonable subset). The computational cost of the PSC module is quietly small compared with that of the baseline network.

(4) The proposed method achieves the best detection accuracy on three challenging datasets compared with existing state-of-the-art methods in the sense of detecting occluded pedestrians.

2 Related work

Two-stage deep pedestrian detection. In recent years, two-stage pedestrian detection approaches [1, 8, 9, 12–15, 17] have shown superior performance on standard pedestrian benchmarks. Generally, in two-stage pedestrian detectors, a set of candidate pedestrian proposals is first generated. Then, these candidate object proposals are classified and regressed. Zhang et al. [1] proposed key adaptations in the popular Faster R-CNN [19] for pedestrian detection. Wang et al. [17] proposed an approach based on a bounding-box regression loss designed for crowded scenes. Zhang et al. [14] proposed to investigate several channel attention strategies for pedestrian detection. MS-CNN [31] was proposed to introduce a multi-scale pedestrian detection approach with layers having receptive fields similar to object scales. Zhang et

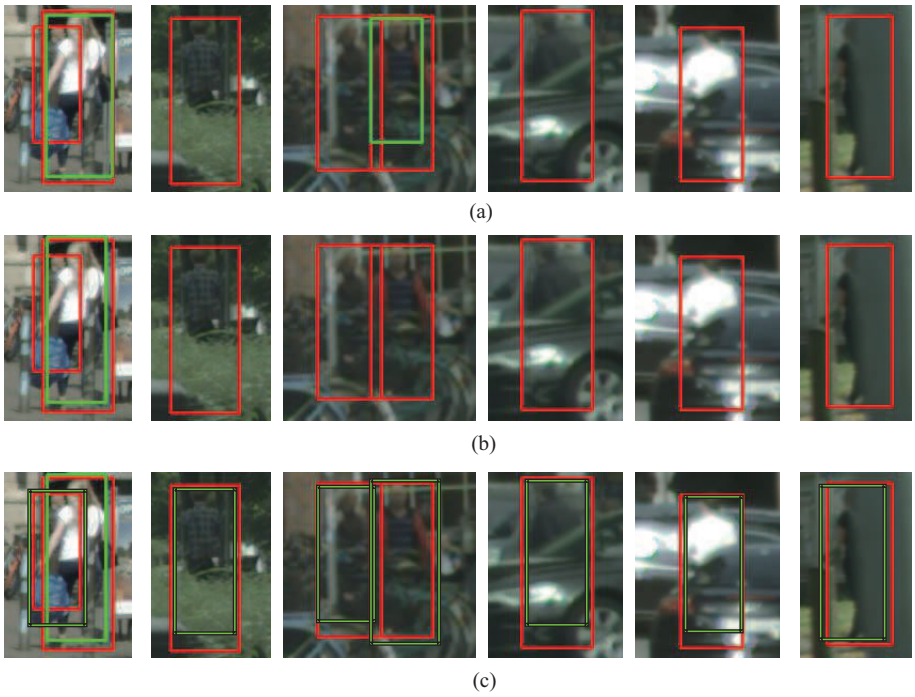

Figure 1 (Color online) Qualitative detection examples using (a) the state-of-the-art MGAN [13], (b) our baseline, and (c) our PSC-Net on CityPersons validation images. In these examples, red boxes denote the ground-truth and detector predictions are indicated by green boxes. Note that all detection results are obtained using the same false positive per image (FPPI) criterion. Our PSC-Net accurately detects pedestrians with varying level of occlusions.

al. [15] proposed a loss formulation to enforce candidate proposals to be close to the corresponding objects and proposed to integrate structural information with visibility predictions. AR-Ped [7] is a multi-phase autoregressive pedestrian detection approach which utilizes a stackable decoder-encoder module with convolutional re-sampling layers. In [8], an adaptive NMS strategy was introduced in order to apply a dynamic suppression threshold to an instance. Multiple pedestrian detectors stacked in a series were also investigated [5, 7] to improve the detection performance.

Towards occluded pedestrian detection. The problem of occluded pedestrian detection was well studied in the literature [9, 12–15, 17, 20, 21]. To handle occlusion problem, some of these pedestrian detection approaches [20,21] exploit part-based information by learning a set of body part detectors. Each part is designated to handle a specific type (pattern) of occlusions. Other approaches [15, 17] investigate novel loss formulations for detector training to improve pedestrian detection in crowded scenes under heavy occlusion.

Most recent approaches [9, 12–15] tackle the problem of occluded pedestrian detection by utilizing additional visible bounding-box (VBB) annotations together with the standard full body information. Zhang et al. [14] employed VBB along with a pre-trained body part prediction model to deal with occluded pedestrian detection. The work of [9] demonstrates that an additional task of visible-region bounding-box prediction can improve the full body pedestrian detection. Zhang et al. [15] proposed a novel loss that improves the localization, and a part occlusion-aware region of interest pooling integrating structure information with visibility predictions. Zhou et al. [12] proposed a discriminative feature transformation module that projects the features into a feature space, where the distance between occluded and non-occluded pedestrians is minimized. Such a transformation improves the robustness of the pedestrian detector. In their approach, the VBB is used to identify the occluded pedestrian. In [13], we proposed a mask-guided attention network (MGAN) which utilizes VBB annotation to emphasize the visible regions and at the same time suppress the occluded regions. To the best of our knowledge, MGAN achieves state-of-the-art results on several popular benchmarks.

Our approach. Contrary to above mentioned recent approaches that rely on additional VBB annotations, our proposed PSC-Net only requires the standard full body supervision to handle occluded pedestrian detection. The core of our approach is a PSC module which explicitly captures both inter and intra-part co-occurrence information of different body parts through a GCN [32]. To the best of

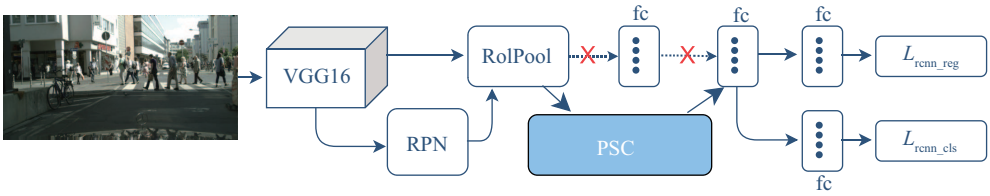

Figure 2 (Color online) Overall network architecture of our PSC-Net. It consists of a pedestrian detection (PD) branch and a part spatial co-occurrence (PSC) module. In contrast to the baseline standard PD branch where the RoI features are used for box regression and classification (X), the RoI features in our PSC-Net are fed in to the proposed PSC module to integrate both intra and inter-part spatial co-occurrence information. The resulting enriched features are then deployed for final bounding-box regression and classification.

our knowledge, the proposed approach is the first to capture both inter and intra-part co-occurrence information through a GCN to address the problem of occluded pedestrian detection.

3 Proposed method: PSC-Net

As discussed above, occlusion is one of the most challenging problems in pedestrian detection. The degree of occlusion ranges from slight (a very small fraction of a pedestrian is occluded), severe (a large fraction of a pedestrian is occluded), to complete (a pedestrian is completely occluded). Though it seems to be impossible to detect completely occluded pedestrians, there are cues for detecting slightly and severely occluded ones. In this paper, we propose to adopt not only the cue of existence of parts of a pedestrian but also the cue of spatial co-occurrence of several parts of a pedestrian.

We divide the spatial co-occurrence patterns into two types: (1) inter-part co-occurrence indicating the co-occurrence of parts of a pedestrian, and (2) intra-part co-occurrence meaning the co-occurrence of sub-parts of a part. The intuition of using inter-part co-occurrence is as follows. (1) In the situation of occlusion, some parts are occluded and other parts are visible. The co-occurrence of the some or all of the visible parts is a strong evidence of existence of a pedestrian. (2) Two or more visible parts can mutually support their existence. This intuition can be analogously extended to that of using intra-part co-occurrence.

The question is how to model the co-occurrence information. One classical way is to classify each part and then integrate the classification score to make the final decision. This way heavily relies on the classification accuracy of each part and also depends on accurately annotating visible parts and occluded parts in the training stage. Incorrect classification is harmful for the final decision and it is time-consuming for annotating visible parts when some parts are occluded. To overcome the drawbacks, we propose to employ GCNs to model both the inter-part and intra-part co-occurrence patterns in the process of feature extraction. The corresponding module is called PSC module.

Figure 2 shows the overall architecture of the proposed PSC-Net. It consists of a standard pedestrian detection (PD) branch (Subsection 3.1) and a PSC module (Subsection 3.2). The standard PD branch is based on Faster R-CNN [19] typically employed in existing pedestrian detection studies [1, 13]. The PSC module encodes both inter and intra-part co-occurrence information of different body parts. The PSC module comprises two components. In the first component, intra-part co-occurrence of a pedestrian body part is captured by utilizing the corresponding RoI features. As a result, an enhanced part feature representation is obtained. This representation is used as an input to the second component for capturing the inter-part co-occurrence between spatially adjacent body parts, leading to a final enhanced feature representation that encodes both intra and inter-part information. This final enhanced feature representation of a candidate proposal is then deployed as an input to the later part of the PD branch which implements final bounding-box regression and category classification.

In what follows, we briefly describe the standard PD branch, followed by a detailed presentation of our PSC module (Subsection 3.2).

3.1 Standard pedestrian detector

The standard PD branch is based on the popular Faster R-CNN framework [19] which is typically employed in several pedestrian detection methods [1,13]. The PD branch consists of a backbone network, a

region proposal network (RPN), region-of-interest (RoI) pooling layer, and a classification network for final bounding-box regression and classification. In the PD branch, an image is first feed into the backbone network and the RPN generates a set of candidate proposals for the input image. For each candidate proposal, a fixed-sized feature representation is obtained through an RoI pooling layer. Finally, this fixed-sized feature representation is passed through a classification network to output the classification score and the regressed bounding box locations for the corresponding proposal. The loss function L_f of the standard PD branch is given as follows:

$$L_f = L_{\text{rpn\_cls}} + L_{\text{rpn\_reg}} + L_{\text{rcnn\_cls}} + L_{\text{rcnn\_reg}}, \tag{1}$$

where $L_{\text{rpn\_cls}}$ and $L_{\text{rpn\_reg}}$ are respectively the classification loss and bounding box regression loss of RPN, and $L_{\text{rcnn\_cls}}$ and $L_{\text{rcnn\_reg}}$ are respectively the classification and bounding box regression loss of the classification network. Generally, Cross-Entropy loss is used as classification loss, and Smooth-L1 loss is used as bounding-box regression loss.

Limitations. To handle heavy occlusion, several recent two-stage pedestrian detection approaches [13–15] extend the PD branch by exploiting additional VBB annotations along with the standard full body information. However, this reliance on additional VBB information implies that two sets of annotations are required for pedestrian detection training.

In this study, we propose a two-stage pedestrian detection method, termed as PSC-Net, to address the problem of heavy occlusions. Our main contribution is the introduction of a PSC module that only requires standard full body supervision and explicitly captures inter-part spatial co-occurrence information of different sub-regions within a body part and intra-part spatial co-occurrence information of different body parts. Next, we describe the details of our PSC module.

3.2 Part spatial co-occurrence module

In pedestrian detection, the task is to accurately localize the full body of a pedestrian whether it is occluded or not. This task is relatively easier in the case of regular non-occluded pedestrians. However, it becomes particularly challenging in the case of slight or severe occlusions. Here, we introduce a PSC module that utilizes spatial co-occurrence of different body parts captured through a GCN [32]. In PSC module, the GCN is employed to capture intra and inter-part spatial co-occurrence by exploiting the topological structure of a pedestrian. The intra-part co-occurrence is expected to improve the feature representation in scenarios where a particular body part is partially occluded whereas the inter-part co-occurrence targets at the severe occlusion of a particular body part.

Our PSC module neither requires pedestrian body part annotations nor relies on the use of an external pre-trained part model. Instead, it divides the full body bounding-box of a pedestrian into five parts (F_{head}, F_{left}, F_{right}, F_{mid}, F_{foot}), based on empirical fixed ratio of human body (see Figure 3), as in [15]. The RoI pooling operation is performed on each body part as well as the full body (F_D), resulting in six RoI pooled features of each region proposal.

As described above, the RoI pooling is performed for five body parts as well as the full body, resulting in an increased feature dimension. Therefore, direct utilization of all these RoI features will drastically increase the computational complexity of our PSC module. Note that the Faster R-CNN and its pedestrian detection adaptations [1, 13, 15] commonly use a single RoI pooling layer on the conv5_3 features of VGG16, resulting in 512 channels. To maintain a similar number of channels as in Faster R-CNN and its pedestrian detection adaptations [13–15], we introduce an additional 1×1 convolutional layer in the RoI pooling strategy that significantly reduces the number of channels (572 in total). Consequently, the RoI pooled features of each body part and the full body have only 64 and 256 channels, respectively.

3.2.1 *Intra-part co-occurrence*

The RoI pooled feature representation of each body part is enhanced by considering their intra-part co-occurrence. For instance, consider a scenario where head part F_{head} is partially occluded, thereby making top-part of the head invisible. Our intra-part co-occurrence component aims to capture spatial relations between different sub-regions (e.g., eyes and ears) within an RoI feature $F_m \in \mathbb{R}^{H \times W \times C}$ of a body part (e.g., F_{head}) through a graph convolutional layer,

$$\tilde{F}_m = \sigma(A_s F_m W_s), \tag{2}$$

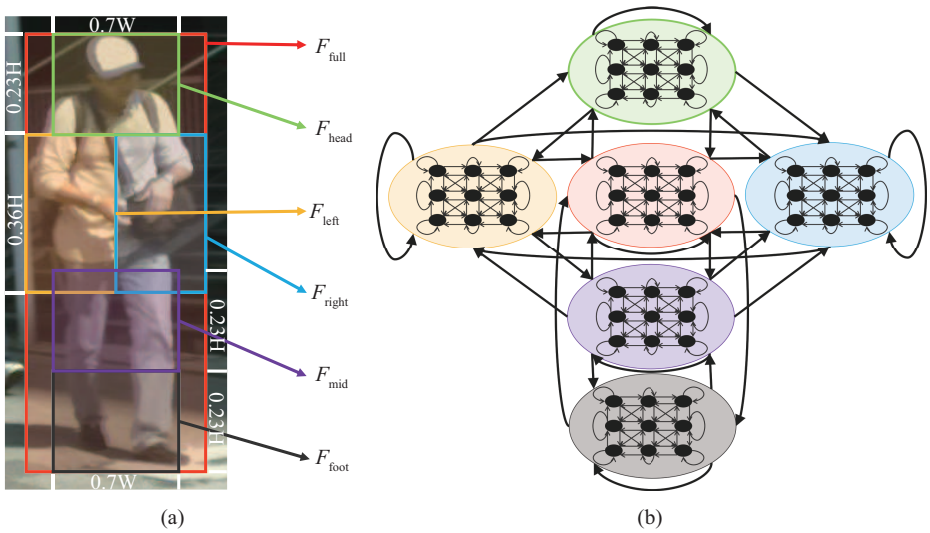

Figure 3 (Color online) (a) Full body pedestrian bounding-box is partitioned into five parts based on empirically fixed ratio of human body. Each body part is shown with a different color and full body is in red. (b) Illustration of intra and inter-part spatial adjacency, used within our PSC module, to capture the spatial co-occurrence information. Our intra-part co-occurrence component employs a graph convolutional layer to capture the spatial relation between different sub-regions of each body part. Differently, our inter-part component captures co-occurrence of spatially adjacent body parts using an additional graph convolutional layer. Note that node colors in (b) are identical to the corresponding body parts in (a).

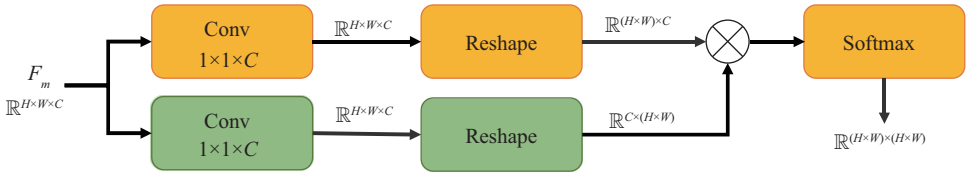

Figure 4 (Color online) Computation of the intra-part spatial adjacency matrix $\mathcal{A}_s$. Each body part RoI feature $F_m \in \mathbb{R}^{H\times W\times C}$ is first passed through two parallel convolutional layers. The resulting features are re-shaped to perform matrix multiplication, followed by a softmax operation.

where $\tilde{F}_m$ is the enhanced feature, σ is the ReLU activation, $W_s \in \mathbb{R}^{C\times C}$ is the learnable parameter matrix, C is the number of channels (64 for each body part and 256 for full body), $A_s \in \mathbb{R}^{(H\times W)\times(H\times W)}$ is the intra-part spatial adjacency matrix of a graph $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{A}_s)$, and W and H are the width and height of F_m. Here, each pixel within the RoI region is treated as a node in the graph. In total, there are $H \times W$ nodes $\mathcal{V}_s$ in the graph.

The intra-part spatial adjacency matrix A_s is computed as follows, and is also shown in Figure 4. We first pass the RoI feature $F_m \in \mathbb{R}^{H\times W\times C}$ through two parallel 1×1 convolutional layers. The resulting outputs are re-shaped prior for performing matrix multiplications, followed by a softmax operation to compute the intra-part spatial adjacency matrix A_s. It is noted that typical values of H and W are $H = 7$ and $W = 7$ and the size of A_s is $(H \times W) \times (H \times W) = 49 \times 49$. Therefore, A_s is usually a small matrix.

The output from the graph convolutional layer (Eq. 2) is denoted by $\tilde{F}_m \in \mathbb{R}^{H\times W\times C}$. This output $\tilde{F}_m$ is first added to its input F_m (original RoI features), followed by a fully connected layer to obtain a d dimensional enhanced part features. The enhanced part features F_e of all six parts (five parts and full body) are further used to capture inter-part co-occurrence described next.

3.2.2 *Inter-part co-occurrence*

Our inter-part co-occurrence component is designed to improve the feature representation, especially in the case of severe occlusion of a particular body part. The traditional convolutional layer only captures information from a small spatial neighborhood, defined by the kernel size (e.g., 3×3), and is therefore often ineffective to encode inter-part co-occurrence information of a pedestrian. To address this issue, we introduce an additional graph convolutional layer in our PSC module, so that the inter-part relationship of different body parts and the full body of a pedestrian is captured. We treat each part (including full

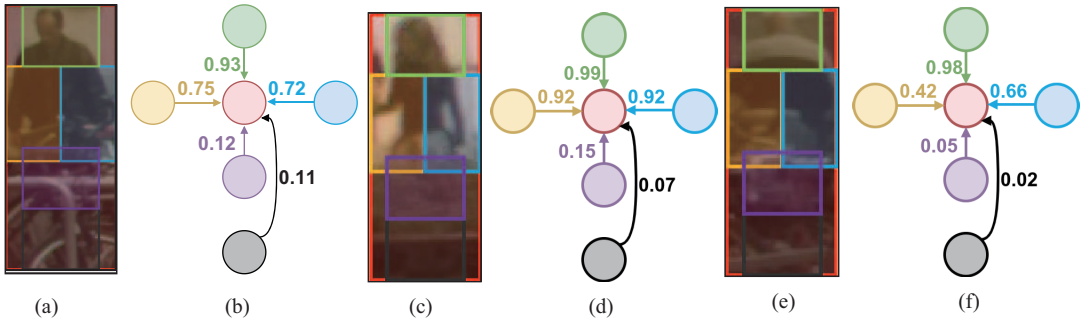

Figure 5 (Color online) The first row of predicted $\hat{\mathcal{A}}_p(i,j)$ on four example proposals. (a), (c), and (e) are three examples of pedestrian proposals, respectively. The full body proposal (bounding-box) in red is partitioned into five (body) parts (in different colors). (b), (d), and (f) are the first row of the predicted unnormalized spatial adjacency matrix $\hat{\mathcal{A}}_p(i,j)$ corresponding to (a), (c), and (e), respectively. It indicates the co-occurrence of full body with respect to specific body part. The co-occurrence of full boy with respect to the full body itself is omitted in this figure for brevity. Note that the colors in (b), (d), and (f) are identical to the corresponding body parts in (a), (c), and (e), respectively.

body) as separate node of $\mathcal{V}_p$ of a graph $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{A}_p)$, where $\mathcal{A}_p$ denotes the spatial adjacency matrix capturing the neighborhood relationship of different nodes. The graph convolutional operation is used to improve the node features F_e as follows:

$$\tilde{F}_e = \sigma((I - \mathcal{A}_p)F_e W_p), \tag{3}$$

where $\tilde{F}_e \in \mathbb{R}^{n\times d}$ is the enhanced feature of n parts, $\mathcal{A}_p$ is the spatial adjacency matrix, $W_p \in \mathbb{R}^{d\times d}$ is the learnable parameter matrix, I is the identity matrix, and σ is the ReLU activation.

Typical values of n (the number of nodes/parts) and d (the channel number) are $n = 6$ and $d = 1024$, respectively. The matrix $(I - \mathcal{A}_p)$ is used to conduct Laplacian smoothing [33] to propagate the node features over the graph.

The spatial adjacency matrix $\mathcal{A}_p$ encodes the relation between different body parts (i.e., graph nodes). When there are heavy occlusions, features of a particular part/node may not contain relevant body part information. Therefore, it is able to assign smaller weights to the edges linking such nodes in $\mathcal{A}_p$. To this end, we introduce a self-attention scheme for each edge which is assigned learnable weight a_{ij}. The input of the self attention is the concatenated features of nodes i and j. The self attention of each edge a_{ij} is computed by a fully-connected operations followed by a sigmoid activation. The unnormalized spatial adjacency matrix $\hat{\mathcal{A}}_p(i,j)$ is defined as

$$\hat{\mathcal{A}}_p(i,j) = \begin{cases} a_{ij}, & \text{if parts } i \text{ and } j \text{ are spatial adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The spatial adjacency matrix $\mathcal{A}_p(i,j)$ is computed by conducting normalization in each row of $\hat{\mathcal{A}}_p(i,j)$. Figure 5 shows the first row of predicted $\hat{\mathcal{A}}_p(i,j)$ on three example proposals. It indicates that the co-occurrence of full body with respect to specific body part. It can be found that the occluded body part contributes less than the visible parts. It would help GCN learn more discriminative features to detect occluded pedestrians. When occlusion occurs, observed features of an occluded body part will be different from its expected features estimated from its spatially adjacent body parts. During end-to-end training of our network containing GCN, both the spatial adjacency matrix and learnable parameter matrix are updated iteratively, enabling the self attention of each edge a_{ij} to reflect the visibility of each body part.

Afterwards, we employ fully connected layer to merge all the features $\tilde{F}_e \in \mathbb{R}^{n\times d}$ into a d-dimensional feature vector. The resulting enriched features are then utilized as an input to the classification network which predicts the final classification score and regresses the location of the bounding box.

4 Experimental results

Datasets. We perform experiments on three datasets: CityPersons [1], Caltech [34], and CrowdHuman [35]. The CityPersons dataset [1] consists of 2975 training, 500 validation, and 1525 test images. The Caltech dataset [34] contains 11 sets of videos. The first 6 sets (0–5) are used for training, and the

Table 1 Impact of integrating our intra-part (Intra-Part Co) and inter-part (Inter-Part Co) co-occurrence into the baseline on the CityPersons validation set[a)]

Baseline (PD)	Intra-Part Co (Subsection 3.2.1)	Inter-Part Co (Subsection 3.2.2)	R	HO
✓	–	–	13.8	56.8
✓	✓	–	11.8	53.1
✓	–	✓	12.5	52.1
✓	✓	✓	**10.6**	**50.2**

a) Performance is reported in terms of log-average miss rates (%) and the best results are boldfaced. Our final PSC-Net that integrates both the intra-part and inter-part co-occurrence achieves consistent improvement in performance, with gains of 3.2% and 6.6% on the R and HO sets, respectively, over the baseline.

last 5 sets (6–10) are used for test. To get a large amount of training data, we sample the videos with 10 Hz. Consequently, the training sets consist of 42782 images in total.

The CrowdHuman dataset [35] is one the challenging datasets in crowd scenarios. The training, validation, and test sets contain 15000, 3870, and 5000 images, respectively. On average, there are more than 20 persons in each image.

Evaluation metrics. We report the performance using log-average miss rates (MR) throughout our experiments. Here, MR is computed over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ [34]. On CityPersons, we report the results on two different occlusion degrees: reasonable (R) and heavy occlusion (HO). For both R and HO sets, the height of pedestrians is larger than 50 pixels. The visibility ratio in R set is larger than 65% whereas in HO it ranges from 20% to 65%. In addition to these sets, the results are reported on combined (R+HO) set on Caltech. On CrowdHuman, we report the results on all pedestrians.

Implementation details. For all the datasets, we train our network on one NVIDIA GPU with the mini-batch consisting of two images per GPU. Adam [36] solver is selected as the optimizer. For CityPersons, we fine-tune the pre-trained ImageNet VGG model [37] on the training set of the CityPersons. We follow the same experimental protocol as in [1] and employ two fully connected layers with 1024 instead of 4096 output dimension. The initial learning rate is set to 10^{-4} for the first 8 epochs, and is then decayed by a factor of 10 for another 3 epochs. For Caltech, we start with a model that is pre-trained on CityPersons. An initial learning rate of 10^{-5} is used for the first 2 training epochs and is then decayed to 10^{-6} for another 1 training epoch. For CrowdHuman, we choose the same backbone network (i.e., ResNet50+FPN) as in [8, 35] for fair comparison. In addition, we follow the same experimental protocol as in [8, 35].

4.1 CityPersons dataset

Baseline comparison. As stated in Section 3, the core of the proposed pedestrian detection method is the PSC module which explicitly captures both intra-part (Subsection 3.2.1) and inter-part (Subsection 3.2.2) co-occurrence information of different body parts. For fair comparison, all results in Table 1 are reported by using the same set of ground-truth pedestrian examples during training. All ground-truth pedestrian examples which are at least 50 pixels tall with visibility $\geqslant 65\%$ are utilized for training. Further, the input scale of $1.0\times$ is employed during this experiment. Table 1 shows that the integration of each component into the baseline results in consistent improvement in performance. Further, our final PSC-Net that integrates both the intra and inter-part co-occurrence achieves absolute gains of 3.2% and 6.6% on the R and HO sets, respectively, over the baseline. These results demonstrate that both components are required to obtain optimal performance.

We also conduct an experiment by replacing our PSC module with part occlusion-aware RoI (PORoI) pooling unit of [15] in our framework. Note that PORoI utilizes VBB information to obtain part labels. For fair comparison, results are reported using the same settings. Our PSC-Net achieves improved results (10.6% on the R set and 50.2% on the HO set) compared with the baseline with PORoI (12.4% on the R set and 54.5% on the HO set) in terms of log-average miss rates.

State-of-the-art comparison. Here, we perform a comparison of our PSC-Net with state-of-the-art pedestrian detection methods in the literature. Table 2 [5, 8, 9, 12–15, 17, 38, 39] shows the comparison results on the CityPersons validation set. Note that existing approaches utilize different sets of ground-truth pedestrian examples for training. For fair comparison, we therefore select the same set of ground-truth pedestrian examples (denoted as data (visibility) in Table 2) and input scale, when performing a comparison with each state-of-the-art method. One can see from Table 2 that our PSC-Net achieves

Table 2 Comparison with the state-of-the-art (in terms of log-average miss rates (%)) on the CityPersons validation set[a)]

Methods	Data (visibility) (%)	Input scales	R	HO
TLL [38]	–	1×	14.4	52.0
F.RCNN+ATT-part [14]	$\geqslant 65$	1×	16.0	56.7
F.RCNN+ATT-vbb [14]		1×	16.4	57.3
Repulsion loss [17]		1×	13.2	56.9
Adaptive-NMS [8]		1×	11.9	55.2
MGAN [13]		1×	11.5	51.7
PSC-Net (ours)		1×	**10.6**	**50.2**
OR-CNN [15]	$\geqslant 50$	1×	12.8	55.7
MGAN [13]		1×	10.8	46.7
PSC-Net (ours)		1×	**10.3**	**44.9**
ALFNet [5]	$\geqslant 0$	1×	12.0	52.0
CSP [39]		1×	11.0	49.3
MGAN [13]		1×	11.3	42.0
PSC-Net (ours)		1×	**10.5**	**39.5**
Repulsion Loss [17]	$\geqslant 65$	1.3×	11.5	55.3
Adaptive-NMS [8]		1.3×	10.8	54.0
MGAN [13]		1.3×	10.3	49.6
PSC-Net (ours)		1.3×	**9.8**	**48.3**
OR-CNN [15]	$\geqslant 50$	1.3×	11.0	51.3
MGAN [13]		1.3×	9.9	45.4
PSC-Net (ours)		1.3×	**9.6**	**43.6**
Bi-box [9]	$\geqslant 30$	1.3×	11.2	44.2
FRCN +A +DT [12]		1.3×	11.1	44.3
MGAN [13]		1.3×	10.5	39.4
PSC-Net (ours)		1.3×	**9.9**	**37.2**

a) In each case, the best results are boldfaced. Our PSC-Net achieves superior performance on both the R and HO sets, compared with existing methods. When using the same input scale (1.3×), training data visibility ($\geqslant 30\%$), and backbone (VGG), PSC-Net provides absolute gains of 7.1% and 2.2% over FRCN +A +DT [12] and MGAN [13], respectively, on the HO set.

the best performance on all these settings for both R and HO sets compared with the state-of-the-art methods.

Specifically, when using an input scale of 1× and data visibility ($\geqslant 65\%$), the attention-based approaches of F.RCNN+ATT-part [14] and F.RCNN+ATT-vbb [14], obtain log-average miss rates of (16.0%, 56.7%) and (16.4%, 57.3%) on the R and HO sets, respectively. The work of [17] based on Repulsion Loss obtains log-average miss rates of 13.2% and 56.9% on the R and HO sets, respectively. The Adaptive-NMS approach [8] that applies a dynamic suppression threshold and learns density scores obtains log-average miss rates of 11.9% and 55.2% on the R and HO sets, respectively. MGAN [13] learns a spatial attention mask using VBB information to modulate full body features and achieves log-average miss rates of 11.5% and 51.7% on the R and HO sets, respectively. Our PSC-Net outperforms MGAN, without using VBB supervision, on both sets with log-average miss rates of 10.6% and 50.2% on the R and HO sets, respectively. When using the same data visibility but 1.3× input scale, Adaptive-NMS [8] and MGAN [13] achieve log-average miss rates of 54.0% and 49.6% on the HO set, respectively. In addition, Adaptive-NMS and MGAN report 10.8% and 10.3% on the R set. PSC-Net achieves the best results with log-average miss rates of 9.8% and 48.3% on the R and HO sets, respectively. On this dataset, the best existing results of 39.4% are reported [13] on the HO sets, when using an input scale of 1.3× and data visibility ($\geqslant 30\%$). Our PSC-Net outperforms the state-of-the-art [13] with log-average miss rates of 37.2%.

Table 3 [1, 8, 13, 15, 17, 40] shows the comparison on CityPersons test set. Among existing methods, the multi-stage Cascade MS-CNN [40] consisting of a sequence of detectors trained with increasing IoU thresholds obtains log-average miss rates of 47.1% on the HO set. MGAN [13] obtains log-average miss rates of 41.0% on the same set. Our PSC-Net significantly reduces the error by 4.0% over MGAN on the HO set.

Computational complexity and inference speed. The computational complexity and test time of our proposed PSC-Net are reported in Table 4. For a fair comparison, the test time of both the baseline and our PSC-Net is measured on a single NVIDIA V100 GPU. For a 1024×2048 input, our PSC-Net

Table 3 State-of-the-art comparison (in terms of log-average miss rates (%)) on CityPersons test set[a)]

Method	R	HO
Adaptive faster RCNN [1]	13.0	50.5
MS-CNN [40]	13.3	51.9
Rep. loss [17]	11.5	52.6
OR-CNN [15]	11.3	51.4
Cascade MS-CNN [40]	11.6	47.1
Adaptive-NMS [8]	11.4	–
MGAN [13]	**9.3**	41.0
PSC-Net (ours)	**9.3**	**37.0**

a) The test set is withheld and results are obtained by sending our PSC-Net detection predictions for evaluation to the authors of CityPersons [1]. Our PSC-Net outperforms existing methods on both the R and HO sets. On the heavy occlusion HO set, PSC-Net achieves an absolute gain of 4.0% over the state-of-the-art [13]. In each case, the best results are boldfaced.

Table 4 Comparison of the proposed PSC-Net with the baseline in terms of the running time for detecting a 1024×2048 image

Method	Test time (s)	MR^{-2}	
		R	HO
Baseline	0.13	13.8	56.8
PSC-Net	0.17	10.6	50.2

Table 5 State-of-the-art comparison (in terms of log-average miss rates (%)) on Caltech test set[a)]

Detector	Occlusion handling	R	HO	R+HO
CompACT-Deep [41]	×	11.8	65.8	24.6
MCF [42]	×	10.4	66.7	22.9
ATT-vbb [14]	✓	10.3	45.2	18.2
MS-CNN [31]	×	10.0	59.9	21.5
SA-F.RCNN [43]	×	9.7	64.4	21.9
RPN+BF [44]	×	9.6	74.4	24.0
FRCN+A+DT [12]	✓	8.0	37.9	–
GDFL [45]	×	7.9	43.2	15.6
Bi-Box [9]	✓	7.6	44.4	16.1
SDS-RCNN [16]	×	7.4	58.6	19.7
MGAN [13]	✓	6.8	38.2	13.8
AR-Ped [7]	×	6.5	48.8	16.1
PSC-Net (ours)	✓	**6.4**	**34.8**	**12.7**

a) The best results are boldfaced in each case. Our PSC-Net provides consistent improvements (over) on all sets. On the HO set, PSC-Net outperforms the best reported results [13] by reducing the error from 37.9% to 34.8%.

takes 0.17 s whereas the baseline takes 0.13 s. The log-average miss rates of the proposed PSC-Net and the baseline are 50.2% and 56.8% on the HO sets, respectively. The proposed method outperforms the baseline by 6.6%. Compared with the baseline, our PSC-Net only has additional 0.04 s for one image in CityPersons validation set. Therefore, it can be concluded that the proposed PSC module results in much lower miss rate at the cost of slight computational burden. The computational complexity of the PSC-Net is limited owing to the following factors. (1) The spatial size of the RoI features is 7×7 and hence the size of the adjacency matrix for the intra-part co-occurrence is as small as 49×49. (2) A pedestrian is divided into six parts and so the adjacency matrix for the inter-part co-occurrence is only 6×6. (3) By an efficient 1×1 convolutional layer, the channel number of the RoI features is reduced to 256 for the full body and 64 for each part.

4.2 Caltech dataset

Table 5 [7, 9, 12–14, 16, 31, 41–45] shows the comparison on Caltech test set under three sets: R, HO, and R+HO. Among existing methods, ATT-vbb [14], Bi-Box [9], FRCN+A+DT [12], and MGAN [13], address the problem of occlusions by utilizing VBB information. On the R, HO, and R+HO subsets, AR-Ped [7], FRCN+A+DT [12], and MGAN [13] report the best existing performance, respectively. PSC-Net achieves superior detection performance on all three subsets with log average miss rates of 6.4%, 34.8%, and 12.7%, respectively. It is concluded from Table 5 that the proposed PSC-Net significantly outperforms

Figure 6 (Color online) Qualitative detection comparison of (c) PSC-Net with (a) AR-Ped [7] and (b) MGAN [13] under occlusions on Caltech test images. Here, all detection results are obtained using the same false positive per image criterion. The red boxes denote the ground-truth whereas the detector's predictions are marked in green.

Table 6 Comparison (in log-average miss rates (%)) on the CrowdHuman dataset[a)]

Method	MR
FPN [35]	50.4
FPN+Adaptive NMS [8]	49.7
FPN+PSC-Net (ours)	**45.9**

a) The best results are boldfaced.

existing state-of-the-art methods.

Figure 6 visualizes some examples of the proposed PSC-Net. It is observed that AR-Ped [7] and MGAN [13] do not detect some pedestrians when there are heavy occlusions whereas our method can successfully find the pedestrians (e.g., the column 4 of Figure 6). Moreover, the location precision of PSC-Net is better than that of AR-Ped and MGAN (e.g., the first column of Figure 6).

4.3 CrowdHuman dataset

Finally, the proposed method is compared with FPN [35] and FPN+Adaptive NMS [8] on the CrowdHuman dataset. For fair comparison, all the three methods including PSC-Net employ the same backbone network of ResNet50+FPN. The results are given in Table 6. The log-average miss rates of FPN [31], FPN+Adaptive NMS [29], and the proposed PSC-Net are 50.4%, 49.7%, and 45.9%, respectively. The proposed method outperforms FPN and FPN+Adaptive NMS by 4.5% and 3.8%, respectively. The results on the CrowdHuman dataset also demonstrate the superiority of the proposed method where intra-part and inter-part spatial co-occurrence is adopted with graph convolutional networks.

5 Conclusion

We have presented a two-stage approach, PSC-Net, for occluded pedestrian detection. The proposed PSC-Net consists of a standard pedestrian detection branch and a PSC module. The key of PSC-Net is that the PSC module is capable of capturing both intra-part and inter-part spatial co-occurrence of different body parts through GCN. The PSC module only requires standard full body supervision and exploits the topological structure of pedestrians. Experiments have been conducted on three popular datasets: CityPersons, Caltech, and CrowdHuman. The results clearly demonstrate that the proposed PSC-Net significantly outperforms the baseline in all cases. Further, the PSC-Net sets a new state-of-the-art on all the datasets.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61632018) and National Key R&D Program of China (Grant Nos. 2018AAA0102800, 2018AAA0102802).

References

1 Zhang S, Benenson R, Schiele B. Citypersons: a diverse dataset for pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
2 Zhang Z J, Pang Y W. CGNet: cross-guidance network for semantic segmentation. Sci China Inf Sci, 2020, 63: 120104
3 Sun H Q, Pang Y W. GlanceNets-efficient convolutional neural networks with adaptive hard example mining. Sci China Inf Sci, 2018, 61: 109101
4 Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63: 120106
5 Liu W, Liao S, Hu W, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of European Conference on Computer Vision, 2018
6 Noh J, Lee S, Kim B, et al. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
7 Brazil G, Liu X. Pedestrian detection with autoregressive network phases. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
8 Liu S, Huang D, Wang Y, et al. Adaptive NMS: refining pedestrian detection in a crowd. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
9 Zhou C, Yuan J. Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of European Conference on Computer Vision, 2018
10 Cao J, Pang Y, Han J, et al. Taking a look at small-scale pedestrians and occluded pedestrians. IEEE Trans Image Process, 2020, 29: 3143–3152
11 Cao J, Pang Y, Zhao S, et al. High-level semantic networks for multi-scale object detection. IEEE Trans Circuits Syst Video Technol, 2019. doi: 10.1109/TCSVT.2019.2950526
12 Zhou C, Yang M, Yuan J, et al. Discriminative feature transformation for occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019
13 Pang Y, Xie J, Khan M, et al. Mask-guided attention network for occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019
14 Zhang S, Yang J, Schiele B, et al. Occluded pedestrian detection through guided attention in CNNs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
15 Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: Proceedings of European Conference on Computer Vision, 2018
16 Brazil G, Xi Y, Liu X. Illuminating pedestrians via simultaneous detection and segmentation. In: Proceedings of IEEE International Conference on Computer Vision, 2017
17 Wang X, Xiao T, Jiang Y, et al. Repulsion loss: detecting pedestrians in a crowd. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
18 Mao J, Xiao T, Jiang Y, et al. What can help pedestrian detection? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
19 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Conference and Workshop on Neural Information Processing Systems, 2015
20 Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2015
21 Zhou C, Yuan J. Multi-label learning of part detectors for heavily occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2017
22 Ouyang W, Wang X. Joint deep learning for pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2013
23 Mathias M, Benenson R, Timofte R, et al. Handling occlusions with Franken-classifiers. In: Proceedings of IEEE International Conference on Computer Vision, 2013
24 Ouyang W, Zeng X, Wang X. Modeling mutual visibility relationship in pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013
25 Mikolajczyk K, Schmid C, Zisserman A. Human detection based on a probabilistic assembly of robust part detectors. In: Proceedings of European Conference on Computer Vision, 2004
26 Mohan A, Papageorgiou C, Poggio T. Example-based object detection in images by components. IEEE Trans Pattern Anal Machine Intell, 2001, 23: 349–361
27 Zhou C, Yuan J. Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Proceedings of Asian Conference on Computer Vision, 2016
28 Biederman I, Mezzanotte R J, Rabinowitz J C. Scene perception: detecting and judging objects undergoing relational violations. Cognitive Psychol, 1982, 14: 143–177
29 Bar M, Ullman S. Spatial context in recognition. Perception, 1996, 25: 343–352

30 Galleguillos C, Rabinovich A, Belongie S. Object categorization using co-occurrence, location and appearance. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008
31 Cai Z, Fan Q, Feris R, et al. A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of European Conference on Computer Vision, 2016
32 Kipf T, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of International Conference on Learning Representations, 2017
33 Li Q, Han Z, Wu X. Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018
34 Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 743–761
35 Shao S, Zhao Z, Li B. CrowdHuman: a benchmark for detecting human in a crowd. 2018. ArXiv: 1805.00123
36 Kingma D, Ba J. Adam: a method for stochastic optimization. In: Proceedings of International Conference on Learning Representations, 2014
37 Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv: 1409.1556
38 Song T, Sun L, Xie D, et al. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proceedings of European Conference on Computer Vision, 2018
39 Liu W, Liao S, Ren W, et al. High-level semantic feature detection: a new perspective for pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
40 Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. 2019. ArXiv: 1906.09756
41 Cai Z, Saberian M, Vasconcelos N. Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2015
42 Cao J, Pang Y, Li X. Learning multilayer channel features for pedestrian detection. IEEE Trans Image Process, 2017, 26: 3210–3220
43 Li J, Liang X, Shen S M, et al. Scale-aware fast R-CNN for pedestrian detection. IEEE Trans Multimedia, 2017. doi: 10.1109/TMM.2017.2759508
44 Zhang L, Lin L, Liang X, et al. Is faster R-CNN doing well for pedestrian detection? In: Proceedings of European Conference on Computer Vision, 2016
45 Lin C, Lu J, Wang G, et al. Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of European Conference on Computer Vision, 2018