Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd

Shifeng Zhang[1,2][0000−0003−3109−5770], Longyin Wen[3][0000−0001−5525−492X], Xiao Bian[3][0000−0001−5447−6045], Zhen Lei[1,2⋆][0000−0002−0791−189X], and Stan Z. Li[4,1,2][0000−0002−2961−8096]

[1] Center for Biometrics and Security Research, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] GE Global Research, Niskayuna, NY
[4] Macau University of Science and Technology, Macau, China
{shifeng.zhang,zlei,szli}@nlpr.ia.ac.cn, {longyin.wen,xiao.bian}@ge.com

Abstract. Pedestrian detection in crowded scenes is a challenging problem since the pedestrians often gather together and occlude each other. In this paper, we propose a new occlusion-aware R-CNN (OR-CNN) to improve the detection accuracy in the crowd. Specifically, we design a new *aggregation loss* to enforce proposals to be close and locate compactly to the corresponding objects. Meanwhile, we use a new part occlusion-aware region of interest (PORoI) pooling unit to replace the RoI pooling layer in order to integrate the prior structure information of human body with visibility prediction into the network to handle occlusion. Our detector is trained in an end-to-end fashion, which achieves state-of-the-art results on three pedestrian detection datasets, *i.e.*, CityPersons, ETH, and INRIA, and performs on-pair with the state-of-the-arts on Caltech.

Keywords: Pedestrian detection · Occlusion-aware · Convolutional network · Structure information · Visibility prediction

1 Introduction

Pedestrian detection is an important research topic in computer vision field with various applications, such as autonomous driving, video surveillance, and robotics, which aims to predict a series of bounding boxes enclosing pedestrian instances in an image. Recent advances in object detection [20,43,10,27,68,57] are driven by the success of deep convolutional neural networks (CNNs), which uses the bounding box regression techniques to accurately localize the objects based on the deep features.

Actually, in real life complex scenarios, occlusion is one of the most significant challenges in detecting pedestrian, especially in the crowded scenes. For example, as pointed out in [55], 48.8% annotated pedestrians are occluded by

⋆ Corresponding author

other pedestrians in the CityPersons dataset [67]. Previous methods only require each predicted bounding box to be close to its designated ground truth, without considering the relations among them. Thus, they make the detectors sensitive to the threshold of non-maximum suppression (NMS) in the crowded scenes, wherein filling with occlusions. To that end, Wang *et al.* [55] design a repulsion loss, which not only pushes each proposal to approach its designated target, but also to keep it away from the other ground truth objects and their corresponding designated proposals. However, it is difficult to control the balance between the repulsion and attraction terms in the loss function to handle the overlapping pedestrians.

In this paper, we propose a new occlusion-aware R-CNN (OR-CNN) based on the Faster R-CNN detection framework [43] to mitigate the impact of occlusion challenge. Specifically, to reduce the false detections of the adjacent overlapping pedestrians, we expect the proposals to be close and locate compactly to the corresponding objects. Thus, inspired by the herd behavior in psychology, we design a new loss function, called *aggregation loss* (AggLoss), not only to enforce proposals to be close to the corresponding objects, but also to minimize the internal region distances of proposals associated with the same objects. Meanwhile, to effectively handle partial occlusion, we propose a new part occlusion-aware region of interest (PORoI) pooling unit to replace the original RoI pooling layer in the second stage Fast R-CNN module of the detector, which integrates the prior structure information of human body with visibility prediction into the network. That is, we first partition the pedestrian region into five parts, and pool the features under each part's projection as well as the whole proposal's projection onto the feature map into fixed-length feature vectors by adaptively-sized pooling bins. After that, we use the learned sub-network to predict the visibility score of each part to combine the extracted features for pedestrian detection.

Several experiments are carried out on four pedestrian detection datasets, *i.e.*, CityPersons [67], Caltech [14], ETH [18] and INRIA [11], to demonstrate the superiority of the proposed method, especially for the crowded scenes. Notably, the proposed OR-CNN method achieves the state-of-the-art results with 11.3% MR^{-2} on the CityPersons dataset, 24.5% MR^{-2} on the ETH dataset, and 6.4% MR^{-2} on the INRIA dataset. The main contributions of this work are summarized as follows.

- We propose a new occlusion-aware R-CNN method, which uses a new designed AggLoss to enforce proposals to be close to the corresponding objects, as well as minimize the internal region distances of proposals associated with the same objects.
- We design a new PORoI pooling unit to replace the RoI pooling layer in the second Fast R-CNN module to integrate the prior structure information of human body with visibility prediction into the network.
- Several experiments are carried out on four challenging pedestrian detection datasets, *i.e.*, CityPersons [67], Caltech [14], ETH [18], and INRIA [11], to demonstrate the superiority of the proposed method.

2 Related Work

Generic Object Detection. Early generic object detectors [53,19,12,40] rely on the sliding window paradigm based on the hand-crafted features and classifiers to find the objects of interest. In recent years, with the advent of deep convolutional neural network (CNN), a new generation of more effective object detection methods based on CNN significantly improve the state-of-the-art performances, which can be roughly divided into two categories, *i.e.*, the one-stage approach and the two-stage approach. The one-stage approach [28,42] directly predicts object class label and regresses object bounding box based on the pre-tiled anchor boxes using deep CNNs. The main advantage of the one-stage approach is its high computational efficiency. In contrast to the one-stage approach, the two-stage approach [43,10,27] always achieves top accuracy on several benchmarks, which first generates a pool of object proposals by a separated proposal generator (*e.g.*, Selective Search [52], EdgeBoxes [74], and RPN [43]), and then predicts the class label and accurate location and size of each proposal.

Pedestrian Detection. Even as one of the long-standing problems in computer vision field with an extensive literature, pedestrian detection still receives considerable interests with a wide range of applications. A common paradigm [13,58,3,59,64] to address this problem is to train a pedestrian detector that exhaustively operates on the sub-images across all locations and scales. Dalal and Triggs [11] design the histograms of oriented gradient (HOG) descriptors and support vector machine (SVM) classifier for human detection. Dollár *et al.* [12] demonstrate that using features from multiple channels can significantly improve the performance. Zhang *et al.* [66] provide a systematic analysis for the filtered channel features, and find that with the proper filter bank, filtered channel features can reach top detection quality. Paisitkriangkrai *et al.* [39] design a new features built on the basis of low-level visual features and spatial pooling, and directly optimize the partial area under the ROC curve for better performance.

Recently, pedestrian detection is dominated by the CNN-based methods (*e.g.*, [44,22,50,5,60,4]). Sermanet *et al.* [44] present an unsupervised method using the convolutional sparse coding to pre-train CNN for pedestrian detection. In [6], a complexity-aware cascaded detector is proposed for an optimal trade-off between accuracy and speed. Angelova *et al.* [1] combine the ideas of fast cascade and a deep network to detect pedestrian. Yang *et al.* [61] use scale-dependent pooling and layer-wise cascaded rejection classifiers to detect objects efficiently. Zhang *et al.* [63] present an effective pipeline for pedestrian detection via using RPN followed by boosted forests. To jointly learn pedestrian detection with the given extra features, a novel network architecture is presented in [30]. Li *et al.* [25] use multiple built-in sub-networks to adaptively detect pedestrians across scales. Brazil *et al.* [4] exploit weakly annotated boxes via a segmentation infusion network to achieve considerable performance gains.

However, occlusion still remains one of the most significant challenges in pedestrian detection, which increases the difficulty in pedestrian localization. Several methods [35,36,49,32,72,56,46,17,16] use part-based model to describe

the pedestrian in occlusion handling, which learn a series of part detectors and design some mechanisms to fuse the part detection results to localize partially occluded pedestrians. Besides the part-based model, Leibe *et al.* [24] propose an implicit shape model to generate a set of pedestrian hypotheses that are further refined to obtain the visible regions. Wang *et al.* [54] divide the template of pedestrian into a set of blocks and conduct occlusion reasoning by estimating the visibility status of each block. Ouyang *et al.* [37] exploit multi-pedestrian detectors to aid single-pedestrian detectors to handle partial occlusions, especially when the pedestrians gather together and occlude each other in real-world scenarios. In [48,41], a set of occlusion patterns of pedestrians are discovered to learn a mixture of occlusion-specific detectors. Zhou *et al.* [73] propose to jointly learn part detectors so as to exploit part correlations and reduce the computational cost. Wang *et al.* [55] introduce a novel bounding box regression loss to detect pedestrians in the crowd scenes. Although numerous pedestrian detection methods are presented in literature, how to robustly detect each individual pedestrian in crowded scenarios is still one of the most critical issues for pedestrian detectors.

3 Occlusion-aware R-CNN

Our occlusion-aware R-CNN detector follows the adaptive Faster R-CNN detection framework [67] for pedestrian detection, with the new designed aggregation loss (Section 3.1), and the PORoI pooling unit (Section 3.2). Specifically, Faster R-CNN [43] consists of two modules, *i.e.*, the first region proposal network (RPN) module and the second Fast R-CNN module. The RPN module is designed to generate high-quality region proposals, and the Fast R-CNN module is used to classify and regress the accurate locations and sizes of objects, based on the generated proposals.

To effectively generate accurate region proposals in the first RPN module, we design the AggLoss term to enforce the proposals locate closely and compactly to the ground-truth object, which is defined as

$$\mathbb{L}_{\text{rpn}}(\{p_i\},\{t_i\},\{p_i^*\},\{t_i^*\}) = \mathcal{L}_{\text{cls}}(\{p_i\},\{p_i^*\}) + \alpha \cdot \mathcal{L}_{\text{agg}}(\{p_i^*\},\{t_i\},\{t_i^*\}), \quad (1)$$

where i is the index of anchor in a mini-batch, p_i and t_i are the predicted confidence of the i-th anchor being a pedestrian and the predicted coordinates of the pedestrian, p_i^* and t_i^* are the associated ground truth class label and coordinates of the i-th anchor, α is the hyperparameters used to balance the two loss terms, $\mathcal{L}_{\text{cls}}(\{p_i\},\{p_i^*\})$ is the classification loss, and $\mathcal{L}_{\text{agg}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ is the AggLoss (see Section 3.1). We use the log loss to calculate the classification loss over two classes (pedestrian $p_i^* = 1$ *vs.* background $p_i^* = 0$), *i.e.*,

$$\mathcal{L}_{\text{cls}}(\{p_i\},\{p_i^*\}) = \frac{1}{N_{\text{cls}}} \sum_i -\Big(p_i^* \log p_i + (1-p_i^*)\log(1-p_i)\Big), \quad (2)$$

where N_{cls} is the total number of anchors in classification.

3.1 Aggregation Loss

To reduce the false detections of the adjacent overlapping pedestrians, we enforce proposals to be close and locate compactly to the corresponding ground truth objects. To that end, we design a new aggregation loss (AggLoss) for both the region proposal network (RPN) and Fast R-CNN [20] modules in the Faster R-CNN algorithm, which is a multi-task loss pushing proposals to be close to the corresponding ground truth object, while minimizing the internal region distances of proposals associated with the same objects, *i.e.*,

$$\mathcal{L}_{\text{agg}}(\{p_i^*\},\{\{t_i\},\{t_i^*\}) = \mathcal{L}_{\text{reg}}(\{p_i^*\},\{t_i\},\{t_i^*\}) + \beta \cdot \mathcal{L}_{\text{com}}(\{p_i^*\},\{t_i\},\{t_i^*\}), \quad (3)$$

where $\mathcal{L}_{\text{reg}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ is the regression loss which requires each proposal to approach the designated ground truth, and $\mathcal{L}_{\text{com}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ is the compactness loss which enforces proposals locate compactly to the designated ground truth object, and β is the hyper-parameters used to balance the two loss terms.

Similar to Fast R-CNN [20], we use the smooth L1 loss as the regression loss $\mathcal{L}_{\text{reg}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ to measure the accuracy of predicted bounding boxes, *i.e.*,

$$\mathcal{L}_{\text{reg}}(\{p_i^*\},\{t_i\},\{t_i^*\}) = \frac{1}{N_{\text{reg}}} \sum_i p_i^* \Delta(t_i - t_i^*), \quad (4)$$

where N_{reg} is the total number of anchors in regression, and $\Delta(t_i - t_i^*)$ is the smooth L1 loss of the predicted bounding box t_i.

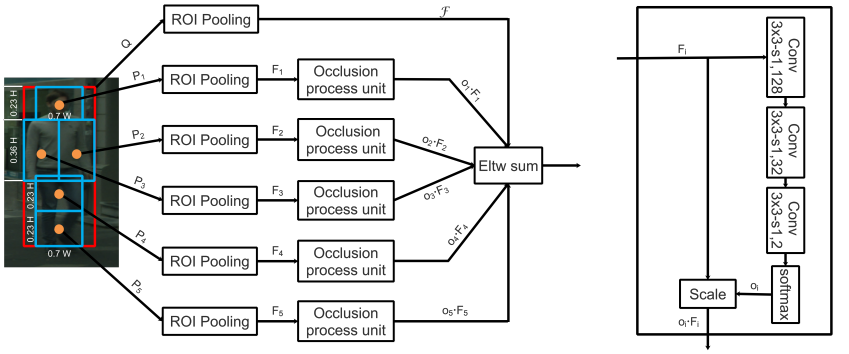
The compactness term $\mathcal{L}_{\text{com}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ is designed to consider the attractiveness among proposals associated with the same ground truth object. In this way, we can make the proposals to locate compactly around the ground truth to reduce the false detections of adjacent overlapping objects. Specifically, we set $\{\tilde{t}_1^*,\cdots,\tilde{t}_\rho^*\}$ to be the ground truth set associated with more than one anchor, and $\{\Phi_1,\cdots,\Phi_\rho\}$ to be the index sets of the associated anchors corresponding to the ground truth objects, *i.e.*, the anchors indexed by Φ_k are associated to the ground truth $\tilde{t}_k^*$, where ρ is the total number of ground-truth object associated with more than one anchor. Thus, we have $\tilde{t}_k^* \in \{t_i^*\}$, for $k = 1,\cdots,\rho$, and $\Phi_i \cap \Phi_j = \emptyset$. We use the smooth L1 loss to measure the difference between the average predictions of the anchors indexed by each set in $\{\Phi_1,\cdots,\Phi_\rho\}$ and the corresponding ground truth object, describing the compactness of predicted bounding boxes with respect to the ground truth object, *i.e.*,

$$\mathcal{L}_{\text{com}}(\{p_i^*\},\{t_i\},\{t_i^*\}) = \frac{1}{N_{\text{com}}} \sum_{i=1}^{\rho} \Delta(\tilde{t}_i^* - \frac{1}{|\Phi_i|}\sum_{j\in\Phi_i} t_j), \quad (5)$$

where N_{com} is the total number of ground truth object associated with more than one anchor (*i.e.*, $N_{\text{com}} = \rho$), and $|\Phi_i|$ is the number of anchors associated with the i-th ground truth object.

3.2 Part Occlusion-aware RoI Pooling Unit

In real life complex scenarios, occlusion is ubiquitous challenging the accuracy of detectors, especially in crowded scenes. As indicated in [35,49,73], the part-based



(a) Part Occlusion-aware RoI pooling unit (b) Occlusion process unit

Fig. 1. For each proposal Q, we divide it into 5 parts $(P_1, \cdots, P_5)$ and use RoIPooling to get the features $(F_1, \cdots, F_5)$, then feed them into the occlusion process unit to predict the visibility scores $(o_1, \cdots, o_5)$. We also apply RoIPooling on Q to generate the holistic feature $\mathcal{F}$. The final features is computed as $\mathcal{F} \oplus (o_1 \cdot F_1) \oplus (o_2 \cdot F_2) \oplus (o_3 \cdot F_3) \oplus (o_4 \cdot F_4) \oplus (o_5 \cdot F_5)$ for subsequent classification and regression.

model is effective in handling occluded pedestrians. In contrast to the aforementioned methods, we design a new part occlusion-aware RoI pooling unit to integrate the prior structure information of human body with visibility prediction into the Fast R-CNN module of the detector, which assembles a micro neural network to estimate the part occlusion status. As shown in Figure 1 (a), we first divide the pedestrian region into five parts with the empirical ratio in [19]. For each part, we use the RoI pooling layer [20] to pool the features into a small feature map with a fixed spatial extent of $H \times W$ (*e.g.*, 7×7).

We introduce an occlusion process unit, shown in Figure 1 (b), to predict the visibility score of the corresponding part based on the pooled features. Specifically, the occlusion process unit is constructed by three convolutional layers followed by a softmax layer with the log loss in training. Symbolically, $c_{i,j}$ indicates the j-th part of the i-th proposal, $o_{i,j}$ represents its predicted visibility score, and $o^*_{i,j}$ is the corresponding ground truth visibility score. If half of the part $c_{i,j}$ is visible, $o^*_{i,j} = 1$, otherwise $o^*_{i,j} = 0$. Mathematically, if the intersection between $c_{i,j}$ and the visible region of ground truth object divided by the area of $c_{i,j}$ is larger than the threshold 0.5, $o^*_{i,j} = 1$, otherwise $o^*_{i,j} = 0$. That is

$$o^*_{i,j} = \begin{cases} 1 & \dfrac{\Omega\big(U(c_{i,j}) \cap V(t^*_i)\big)}{\Omega\big(U(c_{i,j})\big)} > \theta, \\ 0 & \dfrac{\Omega\big(U(c_{i,j}) \cap V(t^*_i)\big)}{\Omega\big(U(c_{i,j})\big)} \leq \theta, \end{cases} \tag{6}$$

where $\Omega(\cdot)$ is the area computing function, $U(c_{i,j})$ is the region of $c_{i,j}$, $V(t^*_i)$ is the visible region of the ground truth object t^*_i, and $\cap$ is the intersection oper-

ation between two regions. Then, the loss function of the occlusion process unit is calculated as $\mathcal{L}_{\text{occ}}(\{t_i\},\{t_i^*\}) = \sum_{j=1}^{5} -(o_{i,j}^* \log o_{i,j} + (1-o_{i,j}^*)\log(1-o_{i,j}))$.

After that, we apply the element-wise multiplication operator to multiply the pooled features of each part and the corresponding predicted visibility score to generate the final features with the dimensions $512 \times 7 \times 7$. The element-wise summation operation is further used to combine the extracted features of the five parts and the whole proposal for classification and regression in the Fast R-CNN module (see Figure 1).

To further improve the regression accuracy, we also use AggLoss in the Fast R-CNN module, which is defined as:

$$\begin{aligned}\mathbb{L}_{\text{frc}}(\{p_i\},\{t_i\},\{p_i^*\},\{t_i^*\}) = \mathcal{L}_{\text{cls}}(\{p_i\},\{p_i^*\}) + \alpha \cdot \mathcal{L}_{\text{agg}}(\{p_i^*\},\{t_i\},\{t_i^*\}) \\ +\lambda \cdot \mathcal{L}_{\text{occ}}(\{t_i\},\{t_i^*\}),\end{aligned} \tag{7}$$

where α and λ are used to balance the three loss terms, $\mathcal{L}_{\text{cls}}(\{p_i\},\{p_i^*\})$ and $\mathcal{L}_{\text{agg}}(\{p_i^*\},\{t_i\},\{t_i^*\})$ are the classification and aggregation losses, defined the same as that in the RPN module, and $\mathcal{L}_{\text{occ}}(\{t_i\},\{t_i^*\})$ is the occlusion process loss.

4 Experiments

Several experiments are conducted on four datasets: CityPersons [67], Caltech-USA [14], ETH [18], and INRIA [11], to demonstrate the performance of the proposed OR-CNN method.

4.1 Experimental Setup

Our OR-CNN detector follows the adaptive Faster R-CNN framework [67] and uses VGG-16 [47] as the backbone network, pre-trained on the ILSVRC CLS-LOC dataset [23]. To improve the detection accuracy of pedestrians with small scale, we use the method presented in [70,69] to dense the anchor boxes with the height less than 100 pixels two times, and use the matching strategy in [71] to associate the anchors and the ground truth objects.

All the parameters in the newly added convolutional layers are randomly initialized by the "xavier" method [21]. We optimize the OR-CNN detector using the Stochastic Gradient Descent (SGD) algorithm with 0.9 momentum and 0.0005 weight decay, which is trained on 2 Titan X GPUs with the mini-batch involving 1 image per GPU. For the Citypersons dataset, we set the learning rate to 10^{-3} for the first $40k$ iterations, and decay it to 10^{-4} for another $20k$ iterations. For the Caltech-USA dataset, we train the network for $120k$ iterations with the initial learning rate 10^{-3} and decrease it by a factor of 10 after the first $80k$ iterations. All the hyperparameters α, β and λ are empirically set to 1.

4.2 CityPersons Dataset

The CityPersons dataset [67] is built upon the semantic segmentation dataset Cityscapes [7] to provide a new dataset of interest for pedestrian detection. It is

Table 1. Pedestrian detection results on the CityPersons validation set. All models are trained on the training set. The scale indicates the enlarge number of original images in training and testing. MR^{-2} is used to compare the performance of detectors (lower score indicates better performance). The top three results are highlighted in red, blue and green, respectively.

Method			Scale	Backbone	*Reasonable*	*Heavy*	*Partial*	*Bare*
Adapted Faster RCNN [67]			×1	VGG-16	15.4	-	-	-
			×1.3	VGG-16	12.8	-	-	-
Repulsion Loss [55]			×1	ResNet-50	13.2	56.9	16.8	7.6
			×1.3	ResNet-50	11.6	55.3	14.8	7.0
OR-CNN	AggLoss	PORoI						
			×1	VGG-16	14.4	59.4	18.4	7.9
	√	√	×1	VGG-16	12.8	55.7	15.3	6.7
			×1.3	VGG-16	12.5	54.5	16.8	6.8
	√		×1.3	VGG-16	11.4	52.6	13.8	6.2
		√	×1.3	VGG-16	11.7	53.0	14.8	6.6
	√	√	×1.3	VGG-16	11.0	51.3	13.7	5.9

recorded across 18 different cities in Germany with 3 different seasons and various weather conditions. The dataset includes $5,000$ images ($2,975$ for training, 500 for validation, and $1,525$ for testing) with $\sim 35,000$ manually annotated persons plus $\sim 13,000$ ignore region annotations. Both the bounding boxes and visible parts of pedestrians are provided and there are approximately 7 pedestrians in average per image.

Following the evaluation protocol in CityPersons, we train our OR-CNN detector on the training set, and evaluate it on both the validation and the testing sets. The log miss rate averaged over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ (MR^{-2}) is used to measure the detection performance (lower score indicates better performance). We use the adaptive Faster R-CNN method [67] trained by ourselves as the baseline detector, which achieves 12.5 MR^{-2} on the validation set with ×1.3 scale, sightly better than the reported result (12.8 MR^{-2}) in [67].

Ablation Study on AggLoss To demonstrate the effectiveness of AggLoss, we construct a detector, denoted as OR-CNN-A, that use AggLoss instead of the original regression loss in the baseline detector [67], and evaluate it on the validation set of CityPersons in Table 1. For a fair comparison, we use the same setting of parameters of OR-CNN-A and our OR-CNN detector in both training and testing. All of the experiments are conducted on the reasonable train/validation sets for training and testing.

Comparing the detection results between the baseline and OR-CNN-A in Table 1, we find that using the newly proposed AggLoss can reduce the MR^{-2} by 1.1% (*i.e.*, 11.4% MR^{-2} *vs.* 12.5% MR^{-2}) with ×1.3 scale. It is worth noting that the OR-CNN-A detector achieves 11.4% MR^{-2} with ×1.3 scale, surpassing the

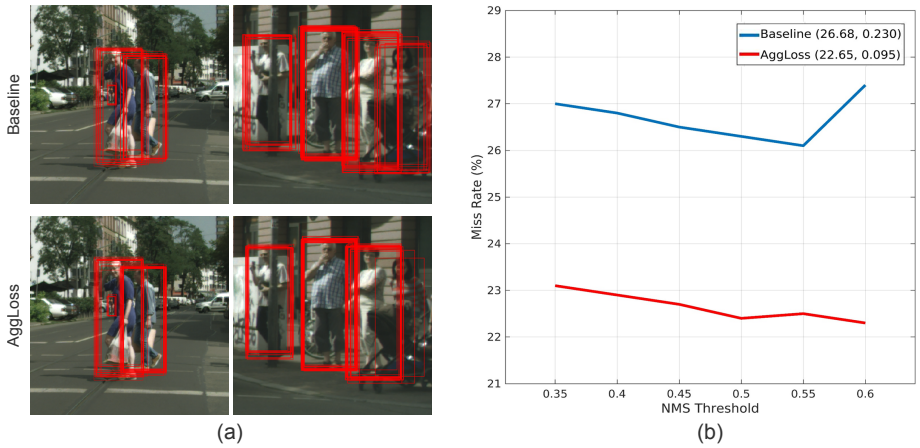


Fig. 2. (a) Visual comparisons of the predicted bounding boxes before NMS of the baseline and OR-CNN-A detectors. The predictions of OR-CNN-A locate more compactly than that of the baseline detector. (b) Results with AggLoss across various NMS thresholds at FPPI = 10^{-2}. The curve of AggLoss is smoother than that of the baseline detector, which indicates that it is less sensitive to the NMS threshold. The scores in the parentheses of the legend are the mean and variance of the miss rate on the curve.

state-of-the-art method using Repulsion Loss [55] (11.6% MR^{-2}), which demonstrates that AggLoss is more effective than Repulsion Loss [55] for detecting the pedestrians in a crowd.

In addition, we also show some visual comparison results of the predicted bounding boxes before NMS of the baseline and OR-CNN-A detectors in Figure 2(a). As shown in Figure 2(a), the predictions of OR-CNN-A locate more compactly than that of the baseline detector, and there are fewer predictions of OR-CNN-A lying in between two adjacent ground-truth objects than the baseline detector. This phenomenon demonstrates that AggLoss can push the predictions lying compactly to the ground-truth objects, making the detector less sensitive to the NMS threshold with better performance in the crowd scene. To further validate this point, we also present the results with AggLoss across various NMS threshold at FPPI = 10^{-2} in Figure 2(b). A high NMS threshold may lead to more false positives, while a low NMS threshold may lead to more false negatives. As shown in Figure 2(b), we find that the curve of OR-CNN-A is smoother than that of baseline (*i.e.*, the variances of the miss rates are 0.095 *vs.* 0.230), which indicates that the former is less sensitive to the NMS threshold. It is worth noting that across various NMS thresholds at FPPI = 10^{-2}, the OR-CNN-A method always produces lower miss rate, which is due to the NMS operation filtering out more false positives in the predictions of OR-CNN-A than that of baseline, implying that the predicted bounding boxes of OR-CNN-A locate compactly than baseline.

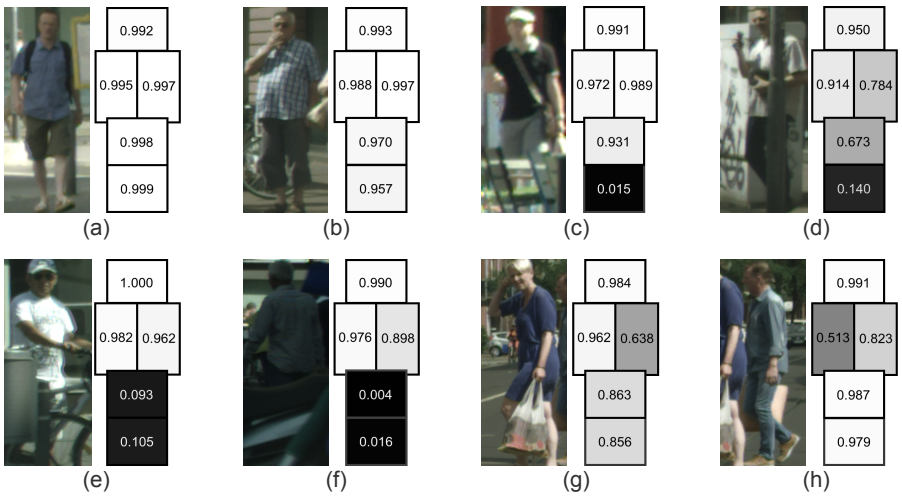


Fig. 3. Some examples of the predicted visibility scores of the pedestrian parts using the proposed PORoI pooling unit.

Ablation Study on PORoI Pooling To validate the effectiveness of the PORoI pooling unit, we construct a detector, denoted as OR-CNN-P, that use the PORoI pooling unit instead of the RoI pooling layer in baseline [67], and evaluate it on the validation set of CityPersons in Table 1. For a fair comparison, we use the same parameter settings of OR-CNN-P and our OR-CNN detector in both training and testing. All of the ablation experiments involved CityPersons are conducted on the reasonable train/validation sets for training and testing.

As shown in Table 1, comparing to baseline, OR-CNN-P reduces 0.8% MR^{-2} with $\times 1.3$ scale (*i.e.*, 11.7% *vs.* 12.5%), which demonstrates the effectiveness of the PORoI pooling unit in pedestrian detection. Meanwhile, we also present some qualitative results of the predictions with the visibility scores of the corresponding parts in Figure 3. Notably, we find that the visibility scores predicted by the PORoI pooling unit are in accordance with the human visual system. As shown in Figure 3(a) and (b), if the pedestrian is not occluded, the visibility score of each part of the pedestrian approaches 1. However, if some parts of the pedestrians are occluded by the background obstacles or other pedestrians, the scores of the corresponding parts decrease, such as the occluded thigh and calf in Figure 3(c)-(f). Besides, if two pedestrians gather together and occlude each other, our PORoI pooling unit successfully detects the occluded human parts that can help lower the contributions of the occluded parts in pedestrian detection, see Figure 3(g) and (h). Notably, the detection accuracy of the OR-CNN detector can not be improved if we fix the visibility score of each part to 1 instead of using the predictions of the occlusion process unit (see Figure 1). Thus, the occlusion process unit is the key component to detection accuracy, since it

Table 2. Pedestrian detection results of the proposed OR-CNN method and other state-of-the-art methods on the CityPersons testing set. The scale indicates the enlarge number of original images in training and testing. MR^{-2} is used to compare of the performance of detectors (lower score indicates better performance).

Method	Backbone	Scale	*Reasonable*	*Reasonable-Small*
Adapted FasterRCNN [67]	VGG-16	×1.3	12.97	37.24
Repulsion Loss [55]	ResNet-50	×1.5	11.48	15.67
OR-CNN	VGG-16	**×1.3**	**11.32**	**14.19**

enables our PORoI pooling unit to detect the occluded parts of pedestrians, which is useful to help extract effective features for detection.

Evaluation Results We compare the proposed OR-CNN method[5] with the state-of-the-art detectors [55,67] on both the validation and testing sets of CityPersons in Table 1 and Table 2, respectively. Our OR-CNN achieves the state-of-the-art results on the validation set of CityPersons by reducing 0.6% MR^{-2} (*i.e.*, 11.0% *vs.* 11.6% of [55]) with ×1.3 scale and 0.4% MR^{-2} (*i.e.*, 12.8% *vs.* 13.2% of [55]) with ×1 scale, surpassing all published approaches [55,67], which demonstrates the superiority of the proposed method in pedestrian detection.

To demonstrate the effectiveness of OR-CNN under various occlusion levels, we follow the strategy in [55] to divide the *Reasonable* subset in the validation set (occlusion $< 35\%$) into the *Reasonable-Partial* subset ($10\% <$ occlusion $\leq 35\%$), denoted as *Partial* subset, and the *Reasonable-Bare* subset (occlusion $\leq 10\%$), denoted as *Bare* subset. Meanwhile, we denote the annotated pedestrians with the occlusion ratio larger than 35% (that are not included in the *Reasonable* set) as *Heavy* subset. We report the results of the proposed OR-CNN method and other state-of-the-art methods [55,67] on these three subsets in Table 1. As shown in Table 1, OR-CNN outperforms the state-of-the-art methods consistently across all three subsets, *i.e.*, reduces 1.1% MR^{-2} on the *Bare* subset, 1.1% MR^{-2} on the *Partial* subset, and 4.0% MR^{-2} on the *Heavy* subset. Notably, when the occlusion becomes severely (*i.e.*, from *Bare* subset to *Heavy* subset), the performance improvement of our OR-CNN is more obvious compared to the state-of-the-art methods [55,67], which demonstrates that the AggLoss and PORoI pooling unit are extremely effective to address the occlusion challenge.

In addition, we also evaluate the proposed OR-CNN method on the testing set of CityPersons [67]. Following its evaluation protocol, we submit the detection results of OR-CNN to the authors for evaluation and report the results in Table 2. The proposed OR-CNN method achieves the top accuracy with only

[5] Due to the shortage of computational resources and the memory issue, we only train OR-CNN with two kinds of input sizes, *i.e.*, ×1 and ×1.3 scale. We believe the accuracy of OR-CNN can be further improved using larger input images. Thus, we only compare the proposed method with the state-of-the-art detectors using ×1 and ×1.3 input scales.

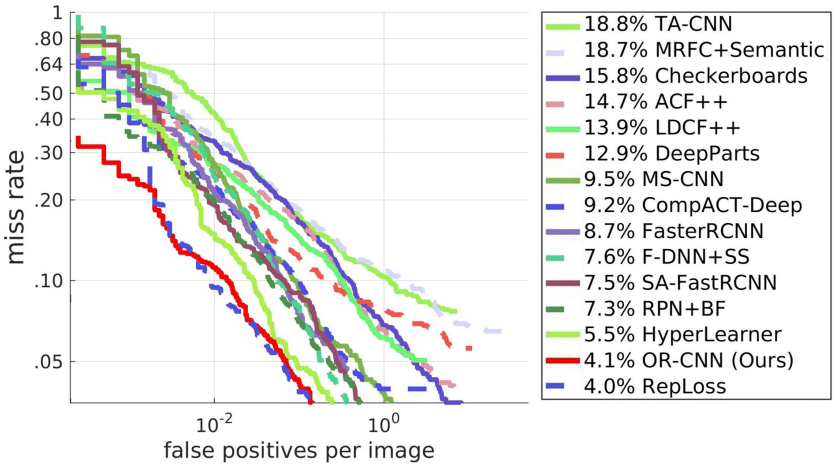


Fig. 4. Comparisons with the state-of-the-art methods on the Caltech-USA dataset. The scores in the legend are the MR^{-2} scores of the corresponding methods.

$\times 1.3$ scale. Although the second best detector Repulsion Loss [55] uses much bigger input images (*i.e.*, $\times 1.5$ scale of [55] *vs.* $\times 1.3$ scale of OR-CNN) and stronger backbone network (*i.e.*, ResNet-50 of [55] *vs.* VGG-16 of OR-CNN), it still produces 0.16% higher MR^{-2} on the *Reasonable* subset and 1.48% higher MR^{-2} on the *Reasonable-Small* subset. We believe the performance of OR-CNN can be further improved by using bigger input images and stronger backbone network.

4.3 Caltech-USA Dataset

The Caltech-USA dataset [14] is one of the most popular and challenging datasets for pedestrian detection, which comes from approximately 10 hours 30Hz VGA video recorded by a car traversing the streets in the greater Los Angeles metropolitan area. We use the new high quality annotations provided by [65] to evaluate the proposed OR-CNN method. The training and testing sets contains $42,782$ and $4,024$ frames, respectively. Following [14], the log-average miss rate over 9 points ranging from 10^{-2} to 10^{0} FPPI is used to evaluate the performance of the detectors.

We directly fine-tune the detection models pre-trained on CityPersons [67] of the proposed OR-CNN method on the training set in Caltech-USA. Similar to [55], we evaluate the OR-CNN method on the *Reasonable* subset of the Caltech-USA dataset, and compare it to other state-of-the-art methods (*e.g.*, [55,66,50,5,6,63,30,25,49,9,34,15]) in Figure 4. Notably, the *Reasonable* subset (occlusion $< 35\%$) only includes the pedestrians with at least 50 pixels tall, which is widely used to evaluate the pedestrian detectors. As shown in Figure 4, the OR-CNN method performs competitively with the state-of-the-art method [55] by producing 4.1% MR^{-2}.

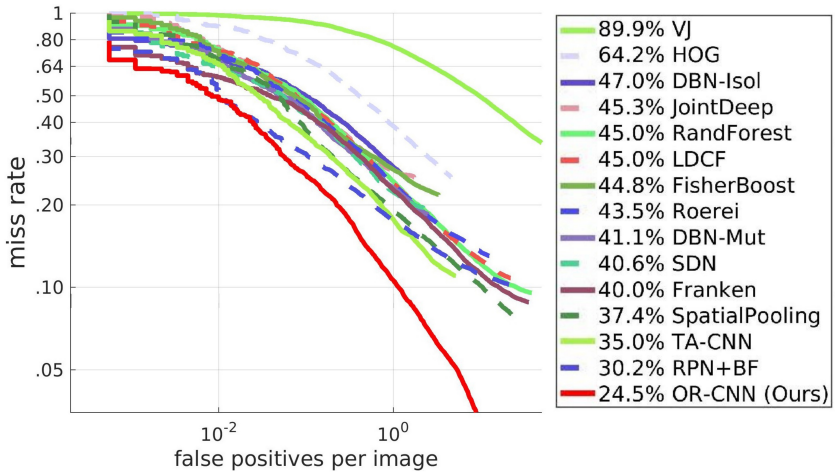


Fig. 5. Comparisons with the state-of-the-art methods on the ETH dataset. The scores in the legend are the MR^{-2} scores of the corresponding methods.

4.4 ETH Dataset

To verify the generalization capacity of the proposed OR-CNN detector, we directly use the model trained on the CityPersons [67] dataset to detect the pedestrians in the ETH dataset [18] without fine-tuning. That is, all 1,804 frames in three video clips of the ETH dataset [18] are used to evaluate the performance of the OR-CNN detector. We use MR^{-2} to evaluate the performance of the detectors, and compare the proposed OR-CNN method with other state-of-the-art methods (*i.e.*, [11,53,3,39,50,63,35,36,32,45,38,31,33,29]) in Figure 5. Our OR-CNN detector achieves the top accuracy by reducing 5.7% MR^{-2} comparing to the state-of-the-art results (*i.e.*, 24.5% of OR-CNN *vs.* 30.2% RFN-BF [63]). The results on the ETH dataset not only demonstrates the superiority of the proposed OR-CNN method in pedestrian detection, but also verifies its generalization capacity to other scenarios.

4.5 INRIA Dataset

The INRIA dataset [11] contains images of high resolution pedestrians collected mostly from holiday photos, which consists of 2,120 images, including 1,832 images for training and 288 images. Specifically, there are 614 positive images and 1,218 negative images in the training set. We use the 614 positive images in the training set to fine-tune our model pre-trained on CityPersons for $5k$ iterations, and test it on the 288 testing images. Figure 6 shows that our OR-CNN method achieves an MR^{-2} of 6.4%, better than the other available competitors (*i.e.*, [11,53,3,64,39,63,32,31,33,26,62,51,8,2]), which demonstrates the effectiveness of the proposed method in pedestrian detection.

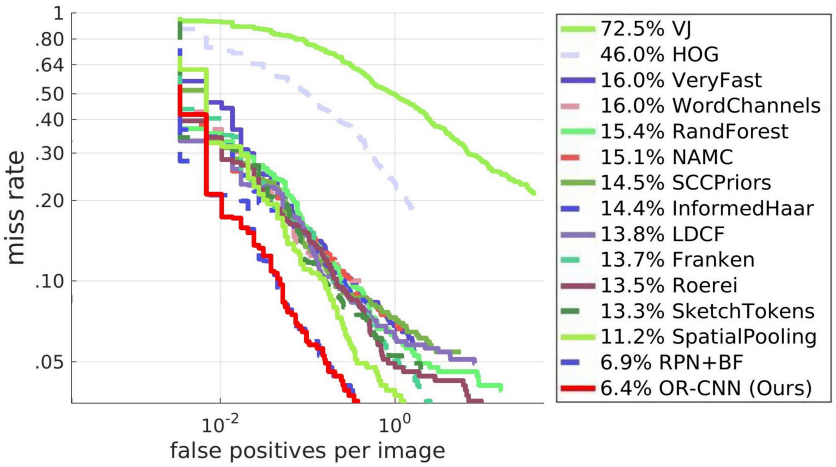


Fig. 6. Comparisons with the state-of-the-art methods on the INRIA dataset. The scores in the legend are the MR^{-2} scores of the corresponding methods.

5 Conclusions

In this paper, we present a new occlusion-aware R-CNN method to improve the pedestrian detection accuracy in crowded scenes. Specifically, we design a new aggregation loss to reduce the false detections of the adjacent overlapping pedestrians, by simultaneously enforcing the proposals to be close to the associated objects, and locate compactly. Meanwhile, to effectively handle partial occlusion, we propose a new part occlusion-aware RoI pooling unit to replace the RoI pooling layer in the Fast R-CNN module of the detector, which integrates the prior structure information of human body with visibility prediction into the network to handle occlusion. Our method is trained in an end-to-end fashion and achieves the state-of-the-art accuracy on three pedestrian detection datasets, *i.e.*, CityPersons, ETH, and INRIA, and performs on-pair with the state-of-the-arts on Caltech. In the future, we plan to improve the method in two aspects. First, we would like to redesign the PORoI pooling unit to jointly estimate the location, size, and occlusion status of the object parts in the network, instead of using the empirical ratio. And then, we plan to extend the proposed method to detect other kinds of objects, *e.g.*, car, bicycle, tricycle, etc.

Acknowledgments

This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects #61473291, #61572501, #61502491, #61572536, the Science and Technology Development Fund of Macau (No. 0025/2018/A1, 151/2017/A, 152/2017/A), and AuthenMetric R&D Funds. We also thank NVIDIA for GPU donations through their academic program.

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A.S., Ferguson, D.: Real-time pedestrian detection with deep network cascades. In: BMVC. pp. 32.1–32.12 (2015)
2. Benenson, R., Mathias, M., Timofte, R., Gool, L.J.V.: Pedestrian detection at 100 frames per second. In: CVPR. pp. 2903–2910 (2012)
3. Benenson, R., Mathias, M., Tuytelaars, T., Gool, L.J.V.: Seeking the strongest rigid detector. In: CVPR. pp. 3666–3673 (2013)
4. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection and segmentation. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 4960–4969 (2017)
5. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: ECCV. pp. 354–370 (2016)
6. Cai, Z., Saberian, M.J., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: ICCV. pp. 3361–3369 (2015)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
8. Costea, A.D., Nedevschi, S.: Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In: CVPR (2014)
9. Costea, A.D., Nedevschi, S.: Semantic channels for fast pedestrian detection. In: CVPR. pp. 2360–2368 (2016)
10. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. pp. 886–893 (2005)
12. Dollár, P., Appel, R., Belongie, S.J., Perona, P.: Fast feature pyramids for object detection. TPAMI **36**(8), 1532–1545 (2014)
13. Dollár, P., Tu, Z., Perona, P., Belongie, S.J.: Integral channel features. In: BMVC. pp. 1–11 (2009)
14. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI **34**(4), 743–761 (2012)
15. Du, X., El-Khamy, M., Lee, J., Davis, L.S.: Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In: WACV (2017)
16. Duan, G., Ai, H., Lao, S.: A structural filter approach to human detection. In: ECCV. pp. 238–251 (2010)
17. Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M.: Multi-cue pedestrian classification with partial occlusion handling. In: CVPR. pp. 990–997 (2010)
18. Ess, A., Leibe, B., Gool, L.J.V.: Depth and appearance for mobile scene analysis. In: ICCV. pp. 1–8 (2007)
19. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9) (2010)
20. Girshick, R.B.: Fast R-CNN. In: ICCV. pp. 1440–1448 (2015)
21. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. pp. 249–256 (2010)
22. Hosang, J.H., Omran, M., Benenson, R., Schiele, B.: Taking a deeper look at pedestrians. In: CVPR. pp. 4073–4082 (2015)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)

24. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR. pp. 878–885 (2005)
25. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia (2017)
26. Lim, J.J., Zitnick, C.L., Dollár, P.: Sketch tokens: A learned mid-level representation for contour and object detection. In: CVPR. pp. 3158–3165 (2013)
27. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: CVPR (2017)
28. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. In: ECCV. pp. 21–37 (2016)
29. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: CVPR. pp. 899–906 (2014)
30. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: CVPR. pp. 6034–6043 (2017)
31. Marín, J., Vázquez, D., López, A.M., Amores, J., Leibe, B.: Random forests of local experts for pedestrian detection. In: ICCV. pp. 2592–2599 (2013)
32. Mathias, M., Benenson, R., Timofte, R., Gool, L.J.V.: Handling occlusions with franken-classifiers. In: ICCV. pp. 1505–1512 (2013)
33. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: NIPS. pp. 424–432 (2014)
34. Ohn-Bar, E., Trivedi, M.M.: To boost or not to boost? on the limits of boosted trees for object detection. In: ICPR. pp. 3350–3355 (2016)
35. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR. pp. 3258–3265 (2012)
36. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV. pp. 2056–2063 (2013)
37. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR. pp. 3198–3205 (2013)
38. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: CVPR. pp. 3222–3229 (2013)
39. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: ECCV (2014)
40. Papageorgiou, C., Poggio, T.A.: A trainable system for object detection. IJCV **38**(1), 15–33 (2000)
41. Pepik, B., Stark, M., Gehler, P.V., Schiele, B.: Occlusion patterns for object class detection. In: CVPR. pp. 3286–3293 (2013)
42. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. CoRR **abs/1612.08242** (2016)
43. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. TPAMI **39**(6), 1137–1149 (2017)
44. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: CVPR. pp. 3626–3633 (2013)
45. Shen, C., Wang, P., Paisitkriangkrai, S., van den Hengel, A.: Training effective node classifiers for cascade classification. IJCV **103**(3), 326–347 (2013)
46. Shet, V.D., Neumann, J., Ramesh, V., Davis, L.S.: Bilattice-based logical reasoning for human detection. In: CVPR (2007)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
48. Tang, S., Andriluka, M., Schiele, B.: Detection and tracking of occluded people. In: BMVC. pp. 1–11 (2012)

49. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: ICCV. pp. 1904–1912 (2015)
50. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: CVPR. pp. 5079–5087 (2015)
51. Toca, C., Ciuc, M., Patrascu, C.: Normalized autobinomial markov channels for pedestrian detection. In: BMVC. pp. 175.1–175.13 (2015)
52. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. IJCV **104**(2), 154–171 (2013)
53. Viola, P.A., Jones, M.J.: Robust real-time face detection. IJCV **57**(2) (2004)
54. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: ICCV. pp. 32–39 (2009)
55. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. CoRR **abs/1711.07752** (2017)
56. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV (2005)
57. Xu, H., Lv, X., Wang, X., Ren, Z., Bodla, N., Chellappa, R.: Deep regionlets for object detection. CoRR **abs/1712.02408** (2017)
58. Yan, J., Lei, Z., Yi, D., Li, S.Z.: Multi-pedestrian detection in crowded scenes: A global view. In: CVPR. pp. 3124–3129 (2012)
59. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: CVPR. pp. 3033–3040 (2013)
60. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: ICCV (2015)
61. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: CVPR (2016)
62. Yang, Y., Wang, Z., Wu, F.: Exploring prior knowledge for pedestrian detection. In: BMVC. pp. 176.1–176.12 (2015)
63. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: ECCV. pp. 443–457 (2016)
64. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed haar-like features improve pedestrian detection. In: CVPR. pp. 947–954 (2014)
65. Zhang, S., Benenson, R., Omran, M., Hosang, J.H., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR. pp. 1259–1267 (2016)
66. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR. pp. 1751–1760 (2015)
67. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR. pp. 4457–4465 (2017)
68. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: CVPR (2018)
69. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Detecting face with densely connected face proposal network. In: CCBR. pp. 3–12 (2017)
70. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Faceboxes: A CPU real-time face detector with high accuracy. In: IJCB (2017)
71. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S^3FD: Single shot scale-invariant face detector. In: ICCV (2017)
72. Zhou, C., Yuan, J.: Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: ACCV. pp. 305–320 (2016)
73. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: ICCV. pp. 3506–3515 (2017)
74. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405 (2014)