Discriminative Feature Transformation for Occluded Pedestrian Detection

Chunluan Zhou[1,2,4*] Ming Yang[3] Junsong Yuan[4]
[1]Baidu Research [2]Wormpex AI Research [3]Horizon Robotics [4]State University of New York at Buffalo
chunluan.zhou@bianlifeng.com m-yang4@u.northwestern.edu jsyuan@buffalo.edu

Abstract

Despite promising performance achieved by deep convolutional neural networks for non-occluded pedestrian detection, it remains a great challenge to detect partially occluded pedestrians. Compared with non-occluded pedestrian examples, it is generally more difficult to distinguish occluded pedestrian examples from backgrounds in featue space due to the missing of occluded parts. In this paper, we propose a discriminative feature transformation which enforces feature separability of pedestrian and non-pedestrian examples to handle occlusions for pedestrian detection. Specifically, in feature space it makes pedestrian examples approach the centroid of easily classified non-occluded pedestrian examples and pushes non-pedestrian examples close to the centroid of easily classified non-pedestrian examples. Such a feature transformation partially compensates the missing contribution of occluded parts in feature space, therefore improving the performance for occluded pedestrian detection. We implement our approach in the Fast R-CNN framework by adding one transformation network branch. We validate the proposed approach on two widely used pedestrian detection datasets: Caltech and CityPersons. Experimental results show that our approach achieves promising performance for both non-occluded and occluded pedestrian detection.

1. Introduction

Pedestrian is a core module for a wide range of applications such as video surveillance, robotics and autonomous driving. With the development of deep convolutional neural networks (CNNs), the performance of pedestrian detection has been significantly improved in recent years [4, 38, 3, 14, 39, 31, 34, 12, 40, 27]. As pointed out in [41, 44], although reasonably good performance has been achieved for detecting non-occluded pedestrians, existing approaches still have difficulty in detecting partially occluded pedestrians. It is generally more challenging to detect a pedestrian when some of its parts are occluded, as illustrated in Fig. 1(a-b). Occlusions occur frequently in practical applications. For example, on a street pedestrians are often occluded by other objects like poles or cars and may also be occluded by each other when walking closely. Therefore, it is essential for a pedestrian detector to handle

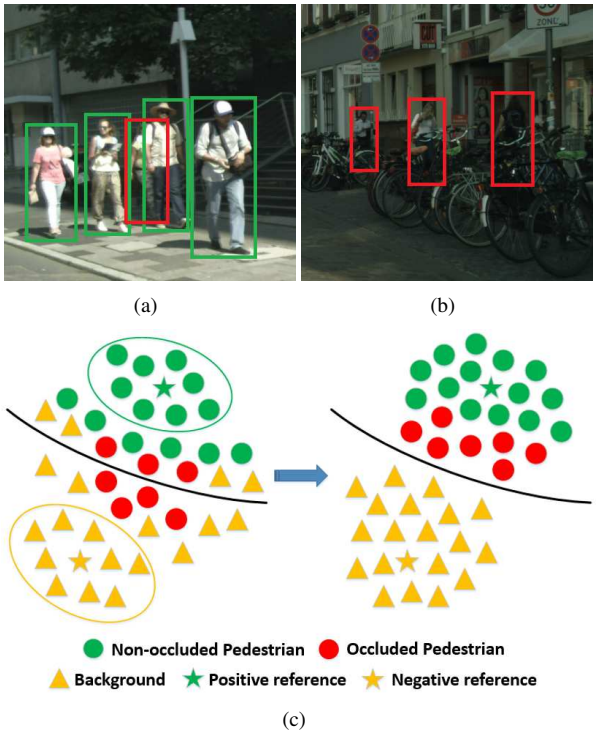


Figure 1. Motivation. (a-b) Occluded and non-occluded pedestrian examples. Green and red boxes represent non-occluded and occluded pedestrian examples respectively. (c) Discriminative feature transformation. (Left) Before transformation, occluded pedestrian examples are difficult to be distinguished from backgrounds. The black line represents the classification boundary. Inside the green ellipse are easy non-occluded pedestrian examples and inside the yellow ellipse are easy background examples. (Right) After transformation, occluded pedestrian examples are moved close to the positive reference and background examples are pushed towards the negative reference.

*The work was done during Chunluan's visit at Baidu and UB

occlusions robustly.

For full-body pedestrian detectors [11, 4, 37, 38, 3, 39, 2, 34], clutters introduced by occlusions within full-body region proposals could degrade detection performance on occluded pedestrians, especially heavily occluded ones. To handle this issue, most occlusion handling approaches [17, 15, 20, 18, 29, 42, 43, 16, 40] adopt a strategy of learning and integrating a set of part detectors. They assume that when a pedestrian is occluded, some part detectors corresponding to visible regions of the pedestrian can still work well. This strategy exploits part correlations and/or complementarity to improve detection performance on occluded pedestrians. Alternatively, a channel-wise attention model [41] is exploited to enhance feature channels activated by visible parts and suppress the other feature channels. In [12], pixel-wise attention is learned to suppress features from background regions. These two approaches adaptively suppress background noise without using part detectors. A bi-box regression framework [44] handles occlusions by estimating the full body and visible part of a pedestrian simultaneously to exploit the complementarity of the two estimation tasks. The above occlusion handling methods improve the robustness to occlusions by exploiting visible parts of pedestrians, but do not make up for the occluded parts. In contrast, we argue that besides the visible parts, enhancing pedestrian representations to compensate missing parts in feature space is a feasible way to further improve occluded pedestrian detection.

In this paper, we propose a discriminative feature transformation to handle occlusions for pedestrian detection. Compared with non-occluded pedestrians, it is usually more difficult to distinguish occluded pedestrians from backgrounds in feature space, since the representations of occluded pedestrians lack the information from their occluded parts, as illustrated in the left part of Fig. 1(c). The proposed feature transformation operates on the representations of pedestrian and non-pedestrian examples to better separate them. Specifically, in feature space it makes pedestrian examples move close to the centroid of easy *non-occluded* pedestrian examples (i.e. ones with high classification scores) and pushes non-pedestrian examples towards the centroid of easy *non-pedestrian* examples (i.e. ones lying far from the classification boundary). We refer to these two centroids as positive and negative reference points in our approach. Figure 1(c) illustrates the idea of the proposed discriminative feature transformation.

Specifically, we adopt the Fast R-CNN framework [9] to implement our approach. First, we learn a Fast R-CNN detector which consists of a feature extractor and a detection branch (See Fig. 3(a) for the structure of the Fast R-CNN detector). As in [12], we incorporate an attention module in the feature extractor to suppress background regions. The detection branch is placed on top of the feature extractor for proposal classification and bounding-box regression. Then, we add a transformation branch on top of the RoI pooling layer in the detection branch. The transformation branch transforms the proposal features from the RoI pooling layer and classifies the pedestrian proposals using the transformed features. The proposed discriminative feature transformation implicitly compensates the missing contribution of occluded parts by pushing occluded pedestrian examples close to the centroid of non-occluded pedestrian examples in feature space, as illustrated in Fig. 2. To our best knowledge, this is the first work that handles occlusions by compensating occluded parts in feature space using a deep CNN. To demonstrate the effectiveness of the proposed approach, we conduct experiments on the Caltech [5] and CityPersons [39] datasets. Experimental results show that our approach achieves promising performance for detecting both non-occluded and occluded pedestrians.

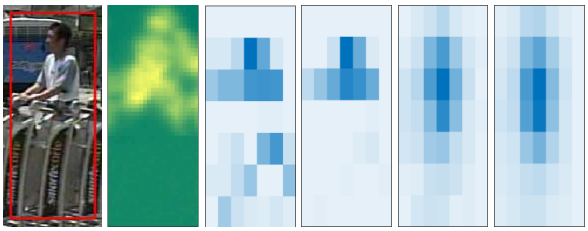


Figure 2. Feature visualization. (a) Pedestrian proposal. (b) Attention map. (c) Features of the pedestrian proposal before applying the attention map. (d) Features of the pedestrian proposal after applying the attention map. The attention map suppresses features corresponding to the background in the pedestrian proposal. (e) Transformed features of the pedestrian proposal. (f) Centroid of easy non-occluded pedestrian examples in feature space. The transformed features are similar to the centroid. In our implementation, the features of a pedestrian proposal from the RoI pooling layer have $7\times7\times512$ dimensions. Here, we only show one typical feature channel from the 512 feature channels.

2. Related work

2.1. Pedestrian detection with CNNs

In recent years, deep CNNs have been widely adopted for pedestrian detection and achieved state-of-the-art performance [34, 13, 27, 40, 12]. In [37, 38, 43], boosting is applied to learn and combine a set of decision trees to form a pedestrian detector using features from a deep CNN. To achieve a trade-off between detection accuracy and speed, a boosting algorithm [4] is proposed to learn complexity-aware cascades by taking into the computational cost and discriminative power of different types of features. In [1], a cascade of deep CNNs of different model sizes is proposed to achieve real-time pedestrian detection by first filtering a large number of negative proposals using tiny CNNs and

then passing the remaining proposals to large CNNs for accurate classification. For fast and accurate detection of multi-scale pedestrians, multi-scale CNNs [3, 13] are designed by adapting the single-stage detector YOLO [22]. In [30], a task-assistant CNN is proposed to exploit both pedestrian attributes and scene attributes to improve pedestrian detection performance. To cope with small pedestrians, a fully convolutional neural newtork is proposed to localize topological lines (lines connecting the head and the middle point between two feet of a pedestrian) instead of bounding-boxes [27]. In [14, 36, 39, 32, 2], Fast R-CNN [9] or Faster R-CNN [23] is adapted for pedestrian detection. In this paper, we adopt the Fast R-CNN framework for occlusion handling.

2.2. Pedestrian detection aided by segmentation

In some works [14, 6, 2, 12], semantic segmentation is exploited to improve performance for pedestrian detection. It is demonstrated in [14] that integrating CNN features with segmentation maps can improve pedestrian detection accuracy. In [6], a segmentation mask is exploited in a post-processing manner to calibrate classification scores output by a deep CNN so as to achieve robust pedestrian detection. A segmentation infusion network [2] is proposed to exploit a segmentation loss to implicitly enhance CNN features from foreground regions and suppress CNN features from background regions. In [12], multi-scale attention maps via supervised segmentation to suppress background regions in the feature maps. Both [2] and [12] use box-level annotations to generate weak ground-truth masks for training. In our approach, we also use box-level annotations to learn an attention map to suppress backgrounds as in [12]. The attention map can better separate the positive and negative reference points used in our approach.

2.3. Occlusion handling for pedestrian detection

Occlusion handling for pedestrian detection has drawn a great deal of attention from researchers due to its importance in practical applications. Learning and integrating a set of part detectors [35, 25, 8, 7, 17, 15, 20, 18, 42, 29, 43, 16, 40] is a widely adopted solution to handle a variety of occlusions. The parts used in these approaches are usually manually designed, which may not be optimal. For approaches [15, 29, 42] which use a large number of independently learned part detectors, the computational cost of applying the learned part detectors could be a bottleneck for real-time pedestrian detection. A multi-label learning approach is proposed in [43] to learn part detectors jointly so as to exploit part correlations as well as reduce the computational cost. In [18, 16, 40], part detectors are learned and integrated in a single deep CNN with the back-end shared by all the part detectors, which can greatly reduce the detection time. Several part detector integration approaches are explored and compared in [42]. In [33], a pedestrian is modeled as a rectangular template of blocks and occlusion reasoning is performed by estimating the visibility statuses of these blocks. Several approaches [19, 28, 21, 34] are specially designed to handle occlusion situations in which multiple pedestrians occlude each other. Particularly, the recent work [34] adopts a repulsion loss to train a deep CNN to improve pedestrian localization accuracy in crowds and achieves promising performance. A bi-box regression framework [44] handles occlusions by estimating the full body and visible part of a pedestrian simultaneously to exploit the complementarity of the two estimation tasks. In [41, 12], attention mechanisms are adopted to suppress background regions and/or enhance foreground regions in feature space for occlusion handling. Deformable part models [10, 45] can also be applied to handle occlusions for pedestrian detection. Considering the importance of occlusion handling, a large-scale dataset, CrowdHuman [24], is proposed for human detection in crowds.

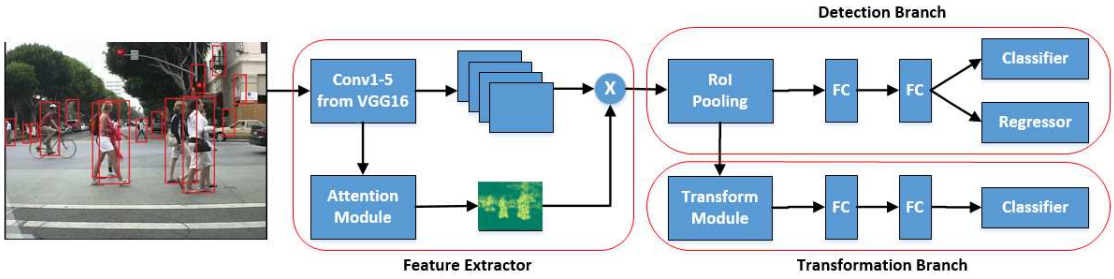
3. Proposed approach

3.1. Overview

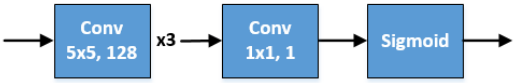
Deep CNNs have achieved promising performance for non-occluded pedestrian detection [34, 12, 13, 40, 27]. However, their performance for occluded pedestrian detection is still far from being satisfactory. To improve the performance of a deep CNN for occluded pedestrian detection, we propose to learn a discriminative transformation in the deep CNN which transforms the features of occluded pedestrians and background regions properly such that they can be better distinguished. We adopt the Fast R-CNN [9] framework to implement our approach. The overview of our approach is shown in Fig. 3. The network used in our approach consists of three components: feature extractor, detection branch and transformation branch. The feature extractor and detection branch form a conventional Fast R-CNN detector. The Fast R-CNN detector takes an image and a set of pedestrian proposals as input and performs classification and bounding-box regression for the pedestrian proposals. This branch transforms the features of pedestrian proposals from the ROI pooling layer to improve classification. At inference stage, we use the detection branch for localization and the transformation branch for classification.

3.2. Fast R-CNN detector

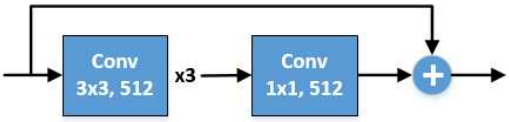
We learn a Fast R-CNN detector for pedestrian detection as well as feature extraction. We use the convolution layers from the VGG-16 network [26] and an attention module [12] to form the feature extractor in our Fast R-CNN detector. The detection branch is placed on top of the feature extractor for classifying pedestrian proposals and refining



(a) Network structure



(b) Attention module



(c) Transform module

Figure 3. Overview of our approach.

their locations.

The attention module takes the feature maps from the last convolution layer as input and outputs an attention map which has the same size as the feature maps. The value at a location of the attention map represents the probability of the image region corresponding to the location belonging to a pedestrian. The attention map is multiplied element-wisely to the feature maps so as to suppress features from background regions as illustrated in Fig. 2(a-d). The structure of the attention module is shown in Fig. 3(b). It consists of three 5x5 convolution layers each with 128 channels, one 1x1 convolution layer with 1 channel and one sigmoid layer. As pixel-level annotations are usually not available in pedestrian detection datasets, we use bounding-box annotations to generate coarse ground-truth segmentation maps for learning the attention module. For each training image, pixels inside ground-truth pedestrian bounding-boxes are labeled as 1 and the others are labeled as 0. The ground-truth segmentation map is scaled to have the same size as the feature maps from the layer Conv5. Let $\bar{S}_i$ and S_i be the ground-truth segmentation map and the predicted attention map (the output from the Sigmoid layer) for the i-th training image, respectively. We use the following Euclidean loss to learn the attention module

$$L_{\text{attn}} = \frac{1}{N}\sum_{i=1}^{N} ||\bar{S}_i - S_i||_2^2, \tag{1}$$

where N is the number of training images.

As in [9], we use cross-entropy and smooth L1 losses to learn the pedestrian proposal classifier and bounding-box regressor in the detection branch, respectively. Let $P_i = (P_i^{\text{x}}, P_i^{\text{y}}, P_i^{\text{w}}, P_i^{\text{h}})$ be a pedestrian proposal, where P_i^{x} and P_i^{y} specify the coordinates of the center of P_i in the image, and P_i^{w} and P_i^{h} are the width and height of P_i, respectively. The pedestrian proposal P_i is associated with a label $c_i \in \{0, 1\}$. P_i is considered as a positive proposal ($c_i = 1$) if there exists at least one ground-truth pedestrian example whose intersection over union (IOU) with P_i is not less than 0.5. Otherwise, it is considered as a negative proposal ($c_i = 0$). Let $p_i = (p_i^0, p_i^1)$ be the output of the pedestrian proposal classifier, where p_i^1 and $p_i^0 = 1 - p_i^1$ represent the probabilities of the pedestrian proposal P_i containing and not containing a pedestrian respectively. We learn the pedestrian proposal classifier with the following loss

$$L_{\text{cls1}} = \frac{1}{M}\sum_{i=1}^{M} -\log(p_i^*), \tag{2}$$

where M is the number of pedestrian proposals, and $p_i^* = p_i^0$ if $c_i = 0$ and $p_i^* = p_i^1$ otherwise. The bounding-box regressor outputs offsets $f_i = (f_i^{\text{x}}, f_i^{\text{y}}, f_i^{\text{w}}, f_i^{\text{h}})$ to refine the location of P_i by

$$\begin{aligned} F^{\text{x}} &= P^{\text{x}} + P^{\text{w}} f^{\text{x}}, \quad F^{\text{y}} = P^{\text{y}} + P^{\text{h}} f^{\text{y}}, \\ F^{\text{w}} &= P^{\text{w}} \exp(f^{\text{w}}), \quad F^{\text{h}} = P^{\text{h}} \exp(f^{\text{h}}). \end{aligned} \tag{3}$$

Let $\bar{f}_i = (\bar{f}_i^{\text{x}}, \bar{f}_i^{\text{y}}, \bar{f}_i^{\text{w}}, \bar{f}_i^{\text{h}})$ be the ground-truth regression targets. We learn the bounding-box regressor with the following loss

$$L_{\text{reg}} = \frac{1}{M}\sum_{i=1}^{M} c_i \sum_{*\in\{\text{x,y,w,h}\}} \text{Smooth}_{L1}(\bar{f}_i^* - f_i^*), \tag{4}$$

where for $s \in \mathbb{R}$

$$\text{Smooth}_{L1}(s) = \begin{cases} 0.5s^2 & \text{if } |s| < 1; \\ |s| - 0.5 & \text{otherwise.} \end{cases} \tag{5}$$

We learn the Fast R-CNN detector by minimizing the following weighted loss

$$L_1 = L_{\text{cls1}} + L_{\text{reg}} + \lambda_1 L_{\text{attn}}, \tag{6}$$

where λ_1 is set to 0.000005 empirically. We refer readers to [9] for more details on Fast R-CNN.

3.3. Discriminative feature transformation

After learning the Fast R-CNN detector, we use features from the RoI pooling layer to represent pedestrian proposals. Generally, it is more difficult to distinguish occluded pedestrians than non-occluded ones from backgrounds in feature space since visual details of occluded parts are missing. To better classify occluded pedestrians and backgrounds, we add one transformation branch to the Fast R-CNN detector as shown in Fig. 3(a). This branch is comprised by a transform module whose structure is shown in Fig. 3(c) and a classifier which classifies pedestrian proposals using their transformed features.

Specifically, we want to learn a transformation which makes occluded pedestrians approach easy non-occluded pedestrians and hard negative proposals approach easy negative proposals in feature space. To achieve this, we first generate two reference points R^+ and R^- in feature space, one for positive proposals and the other for negative proposals. Let H_i be the features of the pedestrian proposal P_i from the RoI pooling layer. In our implementation, H_i is a $K = 7 \times 7 \times 512$ dimensional feature vector. Let o_i be the maximum IOU of P_i with ground-truth pedestrian examples in the same image and v_i be the visibility ratio of the pedestrian example with which P_i has the highest IOU. We collect a set of positive proposals which have high classification scores from the Fast R-CNN detector and overlap largely with at least one non-occluded pedestrian example. Let $\mathcal{H}^+$ be a set consisting of the features of these positive proposals, $\mathcal{H}^+ = \{H_i | p_i^1 \geq s_1, v_i = 1.0 \text{ and } o_i \geq \tau_1\}$, where the thresholds s_1 and τ_1 are set to 0.9 and 0.7, respectively. We define R^+ by

$$R^+ = \frac{1}{|\mathcal{H}^+|} \sum_{H \in \mathcal{H}^+} H, \tag{7}$$

which is the centroid of the feature points in $\mathcal{H}^+$. Similarly, we collect a set of features from easy negative proposals which do not have a large IOU with any ground-truth pedestrian example, $\mathcal{H}^- = \{\mathcal{H}_i | p_i^1 < s_2 \text{ and } o_i < \tau_2\}$, where the thresholds s_2 and τ_2 are set to 0.01 and 0.1, respectively. The reference point R^- is defined by

$$R^- = \frac{1}{|\mathcal{H}^-|} \sum_{H \in \mathcal{H}^-} H, \tag{8}$$

which is the centroid of the feature points in $\mathcal{H}^-$. Let H_i^T be the transformed features of H_i. We learn the transform module with the following loss

$$L_{\text{trans}} = \frac{1}{M} \sum_{i=1}^{M} c_i d_i^+ + (1 - c_i) d_i^-, \tag{9}$$

with

$$d_i^+ = \sum_{k=1}^{K} \text{Smooth}_{L1}(H_i^T(k) - R^+(k)) \tag{10}$$

and

$$d_i^- = \sum_{k=1}^{K} \text{Smooth}_{L1}(H_i^T(k) - R^-(k)), \tag{11}$$

where $H_i^T(k)$ is the k-th feature in H_i^T. The transformation loss in Eq. (9) enables the features of occluded pedestrians to approach the reference point R^+ generated from non-occluded pedestrians during training. It implicitly compensates the missing information of occluded parts in feature space as illustrated in Fig. 2. The reference point R^- attracts the negative proposals to move away from the positive ones to better separate them.

We learn the classifier in the transformation branch via a cross-entropy loss

$$L_{\text{cls2}} = \frac{1}{M} \sum_{i=1}^{M} -\log(q_i^*), \tag{12}$$

where $q_i = (q_i^0, q_i^1)$ is the probabilities output from the classifier, and $q_i^* = q_i^0$ if $c_i = 0$ and $q_i^* = q_i^1$ otherwise. The loss function for the transformation branch is defined by

$$L_2 = L_{\text{cls2}} + \lambda_2 L_{\text{trans}}, \tag{13}$$

where λ_2 is set to 0.1 empirically.

3.4. Training

We train the network in two steps. In the first step, we train the Fast R-CNN detector. The network weights of the Fast R-CNN detector are initialized with the pre-trained VGG-16 network [26] and then updated by minimizing the loss function in Eq. (6). In the second step, we first use the feature extractor in the Fast R-CNN detector to generate postive and negative centroids and then learn the weights of the transformation branch by minimizing the loss function in Eq. (13). The weights of the Fast R-CNN detector are fixed in this stage.

4. Experiments

To demonstrate the effectiveness of our approach, we conduct experiments on two commonly used pedestrian detection datasets: Caltech [5] and CityPersons [39]. Besides the proposed discriminative transformation (DT), we also implement two variants in which only positive examples and negative examples are respectively involved in the transformation loss in Eq. (9) for training the transformation branch. We refer to these two variants as PT and NT in the following sections.

4.1. Experiments on Caltech

The Caltech dataset [5] contains 11 sets of videos collected by a camera mounted on a vehicle driving on urban streets. These videos are divided into two groups: video sets S0-S5 are used for training and video sets S6-S10 are used for testing. In this dataset, there are around 2,300 unique pedestrians and over 70% unique pedestrians are occluded in at least one frame. Some evaluation settings are used in this dataset for evaluating different aspects of pedestrian detection approaches. As our approach is for occlusion handling, we evaluate it in three settings: Reasonable, Partial and Heavy. In the Reasonable setting, only pedestrian examples which have a height of at least 50 pixels and are not occluded more than 35% are used for evaluation. This setting is most widely used for evaluating pedestrian detection approaches. In the Partial and Heavy settings, pedestrians used for evaluation also have a height of at least 50 pixels but are occluded with different ranges. The occlusion range in the Partial setting is 1-35 percent, while the occlusion range in the Heavy setting is 36-80 percent. The Heavy setting is most difficult among the three settings. In each evaluation setting, the detection performance is summarized by a log-average miss rate which is calculated by averaging miss rates at 9 false positives per image (FPPI) points evenly spaced in $[10^{-2}, 10^{0}]$ in log space.

4.1.1 Implementation

We sample training images at an interval of 3 frames from the training video sets S0-S5, resulting in a $10\times$ training set, as commonly done in [38, 39, 32, 43, 2, 41, 44, 12, 27]. Following [43, 44], we select ground-truth pedestrian examples which are at least 50 pixels tall and are occluded less than 70% as positive examples. For pedestrian proposal generation, we train a RPN [38] on the training set. $\sim$1000 pedestrian proposals per image are collected for training and $\sim$400 pedestrian proposals per image are collected for testing. We train our network with SGD for 90,000 iterates. The learning rate is set to 0.0005 initially and decreases by a factor of 10 after 45,000 iterations. We set the batch size to be 160 with foreground-background ratio of $1:3$.

4.1.2 Results

Table 1 shows the results of our approach and some baseline methods. FRCN and FRCN+A are two Fast R-CNN detectors without and with the attention module, respectively. FRCN+A outperforms FRCN by 0.7%, 1.1% and 4.6% in the Reasonable, Partial and Heavy settings, respectively. The improvement in the Heavy setting is significant, which demonstrates the effectiveness of the attention module for suppressing background clutters within heavily occluded pedestrians. From the comparison between FRCN+A and FRCN+A+NT, we can see that NT contributes little to FRCN+A, indicating that learning the transformation branch to only transform features of negative examples does not help much. FRCN+A+PT improves the performance over FRCN+A by 0.4%, 2.0% and 1.0% in the three settings, respectively. The improvements on Partial and Heavy are more significant, showing that the missing information of occluded parts compensated by the transformation branch is helpful for better distinguishing occluded pedestrians from background clutters. DT achieves the most significant improvements among PT, NT and DT. FRCN+A+DT outperforms FRCN+A by 0.8%, 2.9% and 1.8% in the three settings, respectively. DT compensates the missing information of occluded parts for occluded pedestrian examples and forces negative examples to move away from positive examples in feature space, therefore achieving the best performance. We also implement a baseline detector, FRCN+A+TB, which adds the transform branch (TB) to FRCN+A without using the feature transformation loss, i.e. λ_2 is set to 0 in Eq. (13). FRCN+A+TB has the same network structure as FRCN+A+DT but only improves the performance marginally over FRCN+A, indicating that the proposed discriminative transformation is mainly responsible for the performance improvement rather than a classification head with more layers. The transformation loss (Eq. 9) serves as regularization to reduce over-fitting and guides the model training to converge to a better solution.

Method (%)	Reasonable	Partial	Heavy
FRCN	9.5	16.2	44.3
FRCN+A	8.8	15.1	39.7
FRCN+A+NT	8.5	14.9	39.2
FRCN+A+PT	8.4	13.1	38.7
FRCN+A+TB	9.1	14.4	39.1
FRCN+A+DT	**8.0**	**12.2**	**37.9**

Table 1. Results of different approaches on Caltech. Numbers in the table refer to log-average miss rates (lower is better).

τ_1/τ_2	0	0.1	0.2
0.7	8.3/**11.7**/38.4	7.9/12.2/**37.9**	**7.9**/12.3/38.1
0.85	8.4/12.8/38.3	8.4/12.7/38.4	8.2/11.8/38.1
1.0	8.2/12.6/38.4	8.2/12.0/38.0	8.3/12.0/38.0

Table 2. Results with different τ_1 and τ_2 on Caltech. $s_1 = 0.9$ and $s_2 = 0.01$ are used in these experiments.

s_1/s_2	0.01	0.1	0.2
0.7	8.3/12.9/38.3	8.5/12.1/37.6	**7.8**/12.5/38.4
0.8	8.2/12.6/38.4	8.3/12.2/**37.5**	8.4/12.9/38.3
0.9	8.0/12.2/37.9	8.3/12.8/37.8	8.2/**11.8**/37.6

Table 3. Results with different s_1 and s_2 on Caltech. $\tau_1 = 0.7$ and $\tau_2 = 0.1$ are used in these experiments.

Method (%)	Occ	Reason	Partial	Heavy
CompACT-Deep [4]		11.7	25.1	65.8
SA-FastRCNN [11]		9.7	24.8	64.4
MS-CNN [3]		10.0	19.2	59.9
RPN+BF [38]		9.6	24.2	74.4
F-DNN [6]		8.6	15.4	55.1
PCN [32]		8.4	16.1	55.8
F-DNN+SS [6]		8.2	15.1	53.8
TLL(MRF) [27]		8.0	–	–
SDS-RCNN [2]		**7.4**	14.9	58.5
DeepParts [29]	✓	11.9	19.9	60.4
JL-TopS [43]	✓	10	16.6	49.2
FRCN+ATT-vbb [41]	✓	10.3	–	45.2
PDOE+RPN [44]	✓	7.6	13.3	44.4
GDFL [12]	✓	7.8	–	43.2
FRCN+A+DT (Ours)	✓	8.0	**12.2**	**37.9**

Table 4. Comparison with the state-of-the-art approaches on Caltech. Numbers in the table refer to log-average miss rates (lower is better). The Occ column indicates whether an approach is designed for handling occlusions.

Next, we analyze the effect of different positive and negative centroids on the proposed approach, FRCN+A+DT. We conduct experiments with different settings of the overlap thresholds τ_1/τ_2 and score thresholds s_1/s_2 respectively for determining the positive and negative centroids. Table 2 shows the results with different τ_1/τ_2. The miss rates in the Reasonable/Partial/Heavy settings are in the ranges $8.0 \pm 0.4/12.2 \pm 0.6/37.9 \pm 0.5$. Table 3 shows the results with different s_1/s_2. The miss rates are $8.0 \pm 0.5/12.2 \pm 0.7/37.9 \pm 0.5$ in the three settings. Overall, the performance of our approach does not fluctuate much with different choices of τ_1/τ_2 and s_1/s_2 as shown in Tables 2 and 3. We also conduct an experiment in which the positive centroid is determined by all positive examples (i.e. $\mathcal{H}^+ = \{H_i | o_i \geq 0.5\}$) and the negative centroid is determined by all negative examples (i.e. $\mathcal{H}^- = \{\mathcal{H}_i | o_i < 0.5\}$). The performance drops by 0.5%/1.3%/0.9% in the Reasonable/Partial/Heavy settings, indicating that the centroid of easy non-occluded pedestrian examples and the centroid of easy negative examples are a better choice.

We compare our approach with the state-of-the-art approaches using deep CNNs including DeepParts [29], CompACT-Deep [4], SA-FastRCNN [11], MS-CNN [3], RPN+BF [38], F-DNN+SS [6], PCN [32], JL-TopS [43], SDS-RCNN [2], FRCN+ATT-vbb [41], PDOE+RPN [44], TLL(MRF) [27], and GDFL [12]. The results are shown in Table 4. In the Reasonable setting, our approach, FRCN+A+DT, achieves a miss rate of 8.0% in the Reasonable setting, which is comparable to the state-of-the-art performance of 7.4%. In the Partial and Heavy settings,

Figure 4. Detection examples without and with feature transformation.

FRCN+A+DT achieves the best performance of 12.2% and 37.9% respectively. In the Partial setting, FRCN+A+DT outperforms the most competitive approach, PDOE+RPN, by 1.1%. In the Heavy setting, FRCN+A+DT outperforms the most competitive approach, GDFL, by 5.3%. These results validate the effectiveness of our approach for occlusion handling. Figure 4 shows two detection examples of the baseline FRCN+A and our approach. The proposed feature transform help improve the detection scores of the partially occluded pedestrian examples.

4.2. Experiments on CityPersons

CityPersons [39] is a relatively new pedestrian detection dataset. This dataset is more diverse and difficult than Caltech since it covers more countries, cities and seasons and has a higher pedestrian density. This dataset is split into three sets, Train, Val and Test which contain 2975, 500 and 1575 images respectively. Persons in this dataset are classified into six categories: ignored region, pedestrian, rider, group of people, sitting person and other. Results are reported for four setups: Reasonable, Small, Heavy and All. We evaluate the proposed approach in the Reasonable and Heavy setups which are defined according to occlusion ranges. In the Reasonable setup, pedestrian examples which are at least 50 pixels tall and are not occluded more than 35% are used for evaluation. In the Heavy setup, the height and visibility ranges of pedestrian examples are $[50, \infty]$ and $[0.2, 0.65]$ respectively. As on the Caltech dataset, detection performance is summarized by the log-average miss rate.

4.2.1 Implementation

We learn our network on the Train set and evaluate it on the Val set as commonly done in [39, 27, 34, 13, 41, 40, 44]. As in [39], we only use pedestrian examples as positive examples and ignore other person examples. Specifically, ground-truth pedestrian examples which are at least 50 pixels tall and are occluded less than 70% are used for training, as in [43, 44]. We enlarge input images by a factor of 1.3 for training and testing. We also train a RPN on the Train set to generate $\sim 1,500$ pedestrian proposals per image for

Method (%)	Occlusion	Scale	Backbone	Reasonable	Heavy	Partial	Bare
Adapted FasterRCNN [39]		×1	VGG-16	15.4	–	–	–
		×1.3	VGG-16	12.8	–	–	–
TLL(MRF) [27]		×1	ResNet-50	14.4	52.0	15.9	9.2
ALFNet [13]		×1	ResNet-50	12.0	51.9	11.4	8.4
FasterRCNN+ATT-part [41]	✓	×1	VGG-16	15.9	56.7	–	–
RepLoss [34]	✓	×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 [40]	✓	×1	VGG-16	12.8	55.7	15.3	6.7
		×1.3	VGG-16	**11.0**	51.3	13.7	**5.9**
PDOE+RPN [44]	✓	×1.3	VGG-16	11.2	**44.2**	–	–
FRCN+A+DT (Ours)	✓	×1.3	VGG-16	11.1	44.3	**11.2**	6.9

Table 5. Comparison with the state-of-the-art approaches on CityPersons. Numbers in the table refer to log-average miss rates (lower is better). The Occlusion column indicates whether the approach is designed for occlusion handling. The Scale column shows the scale factor the approach uses to enlarge input images. The Backbone column shows the network structure used in the approach.

Method (%)	Reasonable	Heavy
FRCN	12.8	49.2
FRCN+A	12.2	47.4
FRCN+A+NT	11.9	47.2
FRCN+A+PT	11.6	45.8
FRCN+A+TB	12.0	46.6
FRCN+A+DT	**11.1**	**44.3**

Table 6. Results of different approaches on CityPersons. Numbers in the table refer to log-average miss rates (lower is better).

training and $\sim$750 pedestrian proposals per image for testing. Stochastic gradient descent iterates 90,000 times. The intial learning rate is set to 0.001 and decreases by a factor of 0.1 after 45,000 iterations. We set the batch size to be 256 with foreground-background ratio of $1:3$.

4.2.2 Results

Table 6 shows the results of our approach and some baseline methods on the CityPersons dataset. FRCN+A outperforms FRCN by 0.6% and 1.8% in the Reaonable and heavy setups respectively, showing the effectiveness of the attention model for background suppression. Similar to the results on the Caltech dataset, the proposed discriminative transformation, DT, achieves the best performance among the three implementations (NT, PT and DT). FRCN+A+DT improves over FRCN+A by 1.1% and 3.1% respectively in the Reasonable and Heavy setups. FRCN+A+DT also outperforms FRCN+A+TB by 0.9% and 2.3% respectively in the two setups. These results demonstrate the effectiveness of the proposed approach for both non-occluded and occluded pedestrian detection.

We compare our approach with the state-of-the-art approaches including Adapted FasterRCNN [39], TLL(MRF) [27], ALFNet [13], FasterRCNN+ATT-part [41], RepLoss [34], OR-CNN [40] and PDOE+RPN [44] in Table 5. Among these approaches, FasterRCNN+ATT-part, RepLoss, OR-CNN and PDOE+RPN are designed for occlusion handling. As in [34], we also report the results in Partial and Bare setups. In the Partial setup, the height and visibility ranges of pedestrian examples are $[50, \infty]$ and $(0.65, 0.9]$ respectively. In the Partial setup, the height and visibility ranges of pedestrian examples are $[50, \infty]$ and $(0.9, 1]$ respectively. In the Reasonable setup, our approach achieves a miss rate of 11.1% which is comparable to the state-of-the-art performance of 11.0%. In the Heavy setup, our approach has comparable performance to the most competitive occlusion handling approach, PDOE+RPN. Our approach and PDOE+RPN adopt different strategies for occlusion handling. PDOE+RPN focus on how to exploit visible parts for occlusion handling, while our approach learns a feature transformation to better separate pedestrian and non-pedestrian proposals.

5. Conclusion

In this paper, we present a discriminative feature transformation to handle occlusions for pedestrian detection. It forces pedestrian examples to approach the centroid of easy non-occluded pedestrian examples and non-occluded pedestrian examples to approach the centroid of easy non-pedestrian examples in feature space. For occluded pedestrian examples, this transformation compensates the missing information of occluded parts in feature space to some extent, which is a novel way to cope with occluded pedestrian detection. We implement the proposed approach in a Fast R-CNN framework and validate its effectiveness on the Caltech and CityPersons datasets.

Acknowledgement This work is supported in part by the gift grant from Horizon Robotics and start-up funds from University at Buffalo.

References

[1] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian deteciton with deep network cascades. In *British Machine and Vision Conference (BMVC)*, 2015.

[2] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *International Conference on Computer Vision (ICCV)*, 2017.

[3] Zhaowei Cai, Quanfu Fan, Rogerio S. Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision (ECCV)*, 2016.

[4] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *International Conference on Computer Vision (ICCV)*, 2015.

[5] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[6] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry S. Davis. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. *CoRR*, 2016.

[7] Genquan Duan, Haizhou Ai, and Shihong Lao. A structural filter approach to human detection. In *European Conference on Computer Vision (ECCV)*, 2010.

[8] Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[9] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[10] Ross Girshick, Pedro F. Felzenszwalb, and David McAllester. Object detection with grammar models. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[11] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, and Shuicheng Yan. Scale-aware fast R-CNN for pedestrian detection. *CoRR*, 2015.

[12] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *European Conference on Computer Vision (ECCV)*, 2018.

[13] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *European Conference on Computer Vision (ECCV)*, 2018.

[14] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classfiers. In *International Conference on Computer Vision (ICCV)*, 2013.

[16] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Wanli Ouyang and Xiaogang Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[18] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *International Conference on Computer Vision (ICCV)*, 2013.

[19] Wanli Ouyang and Xiaogang Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[20] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Modeling mutual visibility relationship in pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Occlusion patterns for object class detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[24] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *CoRR*, 2018.

[25] Vinay D. Shet, Jan Neumann, Visvanathan Ramesh, and Larry S. Davis. Bilattice-based logical reasoning for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[27] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *European Conference on Computer Vision (ECCV)*, 2018.

[28] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. Detection and tracking of occluded people. In *British Machine Vision Conference (BMVC)*, 2012.

[29] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *International Conference on Computer Vision (ICCV)*, 2015.

[30] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Pedestrian detection aided by deep learning semantic tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] Zhigang Tu, Wei Xie, Justin Dauwels, Baoxin Li, and Junsong Yuan. Semantic cues enhanced multimodality multistream cnn for action recognition. *IEEE Transaction on Circuits and Systems for Video Technology (CSVT)*, 2019.

[32] Shiguang Wang, Jian Cheng, Haijun Liu, and Ming Tang. Pcn: Part and context information for pedestrian detection

with cnns. In *British Machine Vision Conference (BMVC)*, 2017.

[33] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *International Conference on Computer Vision (ICCV)*, 2009.

[34] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[35] Bo Wu and Ram Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *International Conference on Computer Vision (ICCV)*, 2005.

[36] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-model deep representations for robust pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Convolutional channel features. In *International Conference on Computer Vision (ICCV)*, 2015.

[38] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision (ECCV)*, 2016.

[39] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *European Conference on Computer Vision (ECCV)*, 2018.

[41] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] Chunluan Zhou and Junsong Yuan. Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[43] Chunluan Zhou and Junsong Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *International Conference on Computer Vision (ICCV)*, 2017.

[44] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *European Conference on Computer Vision (ECCV)*, 2018.

[45] Chunluan Zhou and Junsong Yuan. Occlusion pattern discovery for object detection and occlusion reasoning. *IEEE Transaction on Circuits and Systems for Video Technology (CSVT)*, 2019.