Bi-box Regression for Pedestrian Detection and Occlusion Estimation

Chunluan Zhou[1,2][0000−0003−0284−6256] and Junsong Yuan[2][0000−0002−7324−7034]

[1] Nanyang Technological University, Singapore
[2] The State University of New York at Buffalo, USA
czhou002@e.ntu.edu.sg, jsyuan@buffalo.edu

Abstract. Occlusions present a great challenge for pedestrian detection in practical applications. In this paper, we propose a novel approach to simultaneous pedestrian detection and occlusion estimation by regressing two bounding boxes to localize the full body as well as the visible part of a pedestrian respectively. For this purpose, we learn a deep convolutional neural network (CNN) consisting of two branches, one for full body estimation and the other for visible part estimation. The two branches are treated differently during training such that they are learned to produce complementary outputs which can be further fused to improve detection performance. The full body estimation branch is trained to regress full body regions for positive pedestrian proposals, while the visible part estimation branch is trained to regress visible part regions for both positive and negative pedestrian proposals. The visible part region of a negative pedestrian proposal is forced to shrink to its center. In addition, we introduce a new criterion for selecting positive training examples, which contributes largely to heavily occluded pedestrian detection. We validate the effectiveness of the proposed bi-box regression approach on the Caltech and CityPersons datasets. Experimental results show that our approach achieves promising performance for detecting both non-occluded and occluded pedestrians, especially heavily occluded ones.

Keywords: Pedestrian detection · Occlusion handling · Deep CNN

1 Introduction

Pedestrian detection has a wide range of applications including autonomous driving, robotics and video surveillance. Many efforts have been made to improve its performance in recent years [3, 8, 17, 6, 33, 40, 5, 39, 41, 37, 34, 4]. Although reasonably good performance has been achieved on some benchmark datasets for detecting non-occluded or slightly occluded pedestrians, the performance for detecting heavily occluded pedestrians is still far from being satisfactory. Take the Caltech dataset [9] for example. One of the top-performing approaches, SDS-RCNN [4], achieves a miss rate of about 7.4% at 0.1 false positives per image (FPPI) for non-occluded or slightly occluded pedestrian detection, but its miss rate increases dramatically to about 58.5% at 0.1 FPPI for heavily occluded

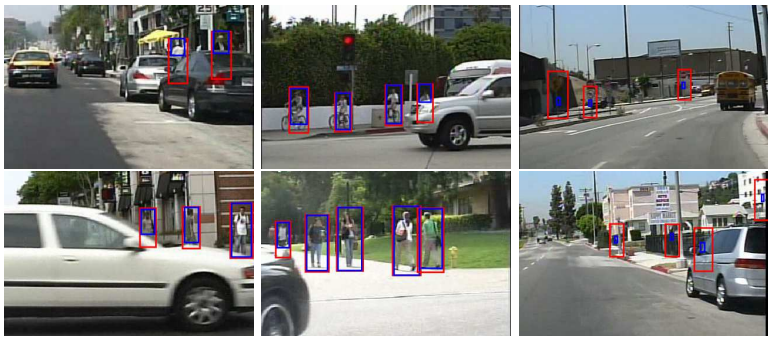


Fig. 1. Detection examples of our approach. The red and blue boxes on each detection represent the estimated full body and visible part respectively. For a pedestrian detection, its visible part is estimated normally as shown in columns 1 and 2. For a non-pedestrian detection, its visible part is estimated to be the center of its corresponding pedestrian proposal as shown in column 3. Since the red box of each detection is obtained by adding estimated offsets to its corresponding pedestrian proposal, the blue box of a non-pedestrian detection is often not exactly at the center of the red box.

pedestrian detection (See Fig. 6). Occlusions occur frequently in real-world applications. For example, pedestrians on a street are often occluded by other objects like cars and they may also occlude each other when walking closely. Therefore, it is important for a pedestrian detection approach to robustly detect partially occluded pedestrians.

Recently, part detectors are commonly used to handle occlusions for pedestrian detection [22, 21, 25, 23, 31, 43, 44]. One drawback of these approaches is that parts are manually designed and therefore may not be optimal. In [22, 21, 25, 31, 43], part detectors are learned separately and then integrated to handle occlusions. For these approaches, the computational cost for testing the part detectors grows linearly with the number of part detectors. A deep convolutional neural network (CNN) is designed to jointly learn and integrate part detectors [23]. However, this approach does not use part annotations for learning the part detectors, which may limit its performance. In [44], a multi-label learning approach is proposed to learn part detectors jointly so as to improve the performance for heavily occluded pedestrian detection and reduce the computational cost of applying the part detectors, but for non-occluded or slightly occluded pedestrian detection, it does not perform as well as state-of-the-art approaches. In addition, for a pedestrian, all these approaches only output one bounding box which specifies the full body region of the pedestrian but does not explicitly estimate which part of the pedestrian is visible or occluded. Occlusion estimation is not well explored in the pedestrian detection literature, but it is critical for applications like robotics which often requires occlusion reasoning to perform interactive tasks.

In this paper, we propose a novel approach to simultaneous pedestrian detection and occlusion estimation by regressing two bounding boxes for full body

and visible part estimation respectively. Deep CNNs [10, 34, 4] have achieved promising performance for non-occluded or slightly occluded pedestrian detection, but their performance for heavily occluded pedestrian detection is far from being satisfactory. This motivates us to explore how to learn a deep CNN for accurately detecting both non-occluded and occluded pedestrians. We thus adapt the Fast R-CNN framework [16, 34, 4] to learn a deep CNN for simultaneous pedestrian classification, full body estimation and visible part estimation. Our deep CNN consists of two branches, one for full body estimation and the other for visible part estimation. Each branch performs classification and bounding box regression for pedestrian proposals. We treat the two branches differently during training such that they produce complementary outputs which can be further fused to boost detection performance. The full body estimation branch is trained to regress full body regions only for positive pedestrian proposals as in the original Fast R-CNN framework, while the visible part estimation branch is trained to regress visible part regions for both positive and negative pedestrian proposals. The visible part region of a negative pedestrian proposal is forced to shrink to its center. Figure 1 shows some detection examples of our approach. For training a deep CNN, positive pedestrian proposals are usually selected based on their overlaps with full body annotations [39, 5, 20, 37, 41, 34, 4], which would include poorly aligned pedestrian proposals for heavily occluded pedestrians (See Fig. 4(b)). To address this issue, we introduce a new criterion which exploits both full body and visible part annotations for selecting positive pedestrian proposals to improve detection performance on heavily occluded pedestrians.

The proposed bi-box regression approach has two advantages: (1) It can provide occlusion estimation by regressing the visible part of a pedestrian; (2) It exploits both full body and visible part regions of pedestrians to improve the performance of pedestrian detection. We demonstrate the effectiveness of our approach on the Caltech [9] and CityPersons [41] datasets. Experimental results show that our approach has comparable performance to the state-of-the-art for detecting non-occluded pedestrians and achieves the best performance for detecting occluded pedestrians, especially heavily occluded ones.

The contributions of this paper are three-fold: (1) A bi-box regression approach is proposed to achieve simultaneous pedestrian detection and occlusion estimation by learning a deep CNN consisting of two branches, one for full body estimation and the other for visible part estimation; (2) A training strategy is proposed to improve the complementarity between the two branches such that their outputs can be fused to improve pedestrian detection performance; (3) A new criterion is introduced to select better positive pedestrian proposals, contributing to a large performance gain for heavily occluded pedestrian detection.

2 Related Work

Recently, deep CNNs have been widely adopted for pedestrian detection [6, 5, 17, 1, 23, 31, 32, 38, 39, 10, 37, 20, 41, 34, 4] and achieved state-of-the-art performance [10, 34, 4]. In [38, 39], a set of decision trees are learned by boosting to form

a pedestrian detector using features from deep CNNs. A complexity-aware cascaded pedestrian detector [6] is learned by taking into account the computational cost and discriminative power of different types of features (including CNN features) to achieve a trade-off between detection accuracy and speed. A cascade of deep CNNs are proposed in [1] to achieve real-time pedestrian detection by first using tiny deep CNNs to reject a large number of negative proposals and then using large deep CNNs to classify remaining proposals. In [31, 23], a set of part detectors are learned and integrated to handle occlusions. A deep CNN is learned to jointly optimize pedestrian detection and other semantic tasks to improve pedestrian detection performance [32]. In [5, 37, 20, 41, 34, 4], Fast R-CNN [16] or Faster R-CNN [27] is adapted for pedestrian detection. In this paper, we explore how to learn a deep CNN to improve performance for detecting partially occluded pedestrians.

Many efforts have been made to handle occlusions for pedestrian detection. A common framework for occlusion handling is learning and integrating a set of part detectors to handle a variety of occlusions [36, 28, 12, 11, 22, 21, 25, 23, 43, 31, 44]. The parts used in these approaches are usually manually designed, which may not be optimal. For approaches (e.g. [21, 31, 43]) which use a large number of part detectors, the computational cost of applying the learned part detector could be a bottleneck for real-time pedestrian detection. In [23], part detectors are learned and integrated with a deep CNN, which can greatly reduce the detection time. However, the part detectors in this approach are learned in a weakly supervised way, which may limit its performance. In [44], a multi-label learning approach is proposed to both improve the reliability of part detectors and reduce the computational cost of applying part detectors. Different part detector integration approaches are explored and compared in [43]. Different from these approaches, we learn a deep CNN without using parts to handle various occlusions. There are also some other approaches to occlusion handling. In [18], an implicit shape model is adopted to generate a set of pedestrian proposals which are further refined by exploiting local and global cues. The approach in [35] models a pedestrian as a rectangular template of blocks and performs occlusion reasoning by estimating the visibility statuses of these blocks. Several approaches [24, 30, 26] are specially designed to handle occlusion situations in which multiple pedestrians occlude each other. A deformable part model [13] and its variants [15, 2, 42] can also be used for handling occlusions.

3 Proposed Approach

Given an image, we want to detect pedestrians in it and at the same time estimate the visible part of each pedestrian. Specifically, our approach produces for each pedestrian two bounding boxes which specify its full body and visible part regions respectively. Considering promising performance achieved by deep CNNs for pedestrian detection [39, 5, 20, 37, 41, 34, 4], we adapt the Fast R-CNN framework [16] for our purpose. Figure 2 shows the overview of the proposed bi-box regression approach. A set of region proposals which possibly contain pedestri-

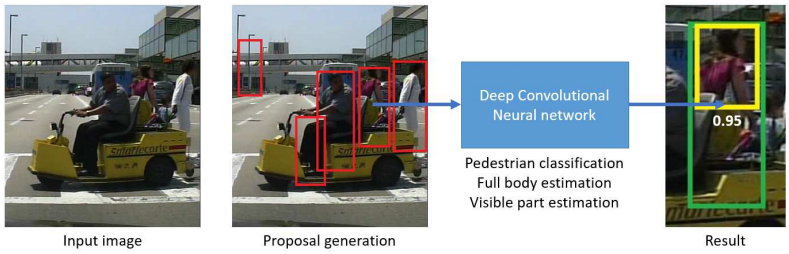


Fig. 2. Overview of our bi-box regression approach.

ans are generated for an input image by a proposal generation approach (e.g. [39, 4]). These pedestrian proposals are then fed to a deep CNN which performs classification, full body estimation and visible part estimation for each proposal.

3.1 Network Structure

We adapt a commonly used deep CNN, VGG-16 [29], to achieve simultaneous pedestrian detection and occlusion estimation. Figure 3 shows the structure of our deep CNN. We keep convolution layers 1 through 4 in VGG-16 unchanged. It is reported in [39, 5] that a feature map with higher resolution generally improves detection performance. As in [39, 5], we remove the last max pooling layer and convolution layer 5 from VGG-16. A deconvolution layer (Deconv5), which is implemented by bilinear interpolation, is added on top of Conv4-3 to increase the resolution of the feature map from Conv4-3. Following Deconv5 is a ROI pooling layer on top of which are two branches, one for full body estimation and the other for visible part estimation. Each branch performs classification and bounding box regression as in Fast R-CNN [16].

3.2 Pedestrian Detection

For detection, an image and a set of pedestrian proposals are fed to the deep CNN for classification, full body estimation and visible part estimation. Let $P = (P^{\mathrm{x}}, P^{\mathrm{y}}, P^{\mathrm{w}}, P^{\mathrm{h}})$ be a pedestrian proposal, where P^{x} and P^{y} specify the coordinates of the center of P in the image, and P^{w} and P^{h} are the width and height of P respectively. For the pedestrian proposal P, the full body estimation branch outputs two probabilities $p_1 = (p_1^0, p_1^1)$ (from the Softmax1 layer) and four offsets $f = (f^{\mathrm{x}}, f^{\mathrm{y}}, f^{\mathrm{w}}, f^{\mathrm{h}})$ (from the FC11 layer). The visible part estimation branch also outputs two probabilities $p_2 = (p_2^0, p_2^1)$ (from the Softmax2 layer) and four offsets $v = (v^{\mathrm{x}}, v^{\mathrm{y}}, v^{\mathrm{w}}, v^{\mathrm{h}})$ (from the FC13 layer). p_1^1 and $p_1^0 = 1 - p_1^1$ represent the probabilities of P containing and not containing a pedestrian, respectively. p_2^0 and p_2^1 are similarly defined. f^{x} and f^{y} specify the scale-invariant translations from the center of P to that of the estimated full body region, while f^{w} and f^{h} specify the log-space translations from the width and height of P to those of the estimated full body region respectively. v^{x}, v^{y}, v^{w} and v^{h} are

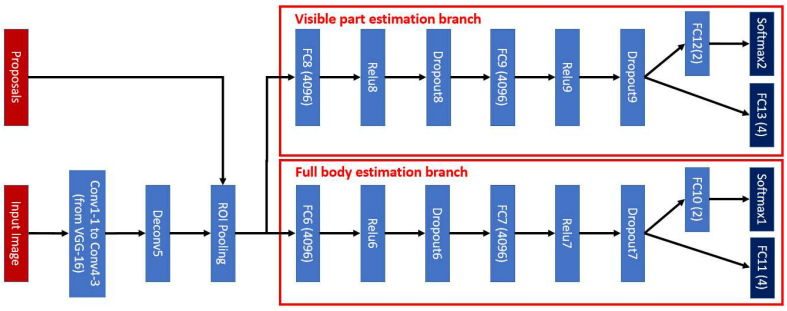


Fig. 3. Network architecture. The number in each fully connected (FC) layer is its output dimensionality. Softmax1 and Softmax2 perform the same task, pedestrian classification. FC11 is for full body estimation and FC13 is for visible part estimation.

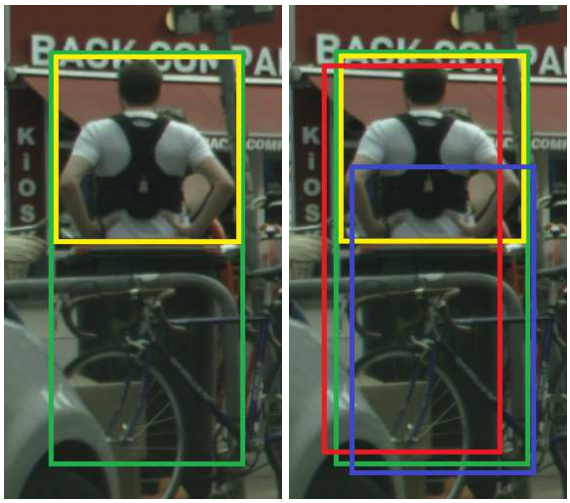
similarly defined for visible part estimation. We define f and v following [16]. With f and v, we can compute the full body and visible part regions for the pedestrian proposal P (See [16] for more details).

We consider three ways to score a pedestrian proposal P. Let $s_1 = (s_1^0, s_1^1)$ and $s_2 = (s_2^0, s_2^1)$ be the raw scores from FC10 and FC12 respectively. The first way scores P with $p_1^1 = \frac{\exp(s_1^1)}{\exp(s_1^1)+\exp(s_1^0)}$ and the second way scores P with $p_2^1 = \frac{\exp(s_2^1)}{\exp(s_2^1)+\exp(s_2^0)}$. The third way fuses the raw scores from the two branches with a softmax operation $\hat{p}^1 = \frac{\exp(s_1^1+s_2^1)}{\exp(s_1^1+s_2^1)+\exp(s_1^0+s_2^0)}$. It can be proved that $\hat{p}^1 - p_1^1 > 0$ if $p_2^1 > 0.5$, i.e. $s_2^1 > s_2^0$. When two branches agree on a positive example, i.e. $p_1^1 > 0.5$ and $p_2^1 > 0.5$, the fused score $\hat{p}^1$ becomes stronger, i.e. $\hat{p}^1 > p_1^1$ and $\hat{p}^1 > p_2^1$. When one branch gives a low score ($p_1^1 < 0.5$) to the positive example, the other branch can increase its detection score if it gives a high score ($p_2^1 > 0.5$). This guides us to increase the complementarity between the two branches so to improve detections robustness as described in next section.

3.3 Network Training

To train our deep CNN, each pedestrian example is annotated with two bounding boxes which specify its full body and visible part regions respectively. Figure 4(a) shows an example of pedestrian annotation. Besides these annotated pedestrian examples, we also collect some pedestrian proposals for training. To achieve this, we match pedestrian proposals in a training image to annotated pedestrian examples in the same image. Let $Q = (\bar{F}, \bar{V})$ be an annotated pedestrian example in an image, where $\bar{F} = (\bar{F}^{\mathrm{x}}, \bar{F}^{\mathrm{y}}, \bar{F}^{\mathrm{w}}, \bar{F}^{\mathrm{h}})$ and $\bar{V} = (\bar{V}^{\mathrm{x}}, \bar{V}^{\mathrm{y}}, \bar{V}^{\mathrm{w}}, \bar{V}^{\mathrm{h}})$ are the full body and visible part regions respectively. A pedestrian proposal P is matched to Q if it aligns well with Q. Specifically, P and Q form a pair if they satisfy

$$\mathrm{IOU}(P, \bar{F}) \geq \alpha \text{ and } \mathrm{C}(P, \bar{V}) \geq \beta, \tag{1}$$



(a) (b)

Fig. 4. Pedestrian annotation and positive pedestrian proposal selection. (a) The green and yellow bounding boxes specify the full body and visible part of a pedestrian example respectively. (b) The red bounding box is a good pedestrian proposal and the blue bounding box is a bad pedestrian proposal.

where $\mathrm{IOU}(P, \bar{F})$ is the intersection over union of the two regions P and $\bar{F}$:

$$\mathrm{IOU}(P, \bar{F}) = \frac{\mathrm{Area}(P \cap \bar{F})}{\mathrm{Area}(P \cup \bar{F})}, \tag{2}$$

and $\mathrm{C}(P, \bar{V})$ is the proportion of the area of $\bar{V}$ covered by P:

$$\mathrm{C}(P, \bar{V}) = \frac{\mathrm{Area}(P \cap \bar{V})}{\mathrm{Area}(\bar{V})}. \tag{3}$$

In Fig. 4(b), the pedestrian proposal (red bounding box) is matched to the annotated pedestrian example (green bounding box) with $\alpha = 0.5$ and $\beta = 0.5$, while the pedestrian proposal (blue bounding box) is not matched due to its poor alignment with the annotated pedestrian example.

Denote by I the image where P is generated. For each matched pair (P, Q), we construct a positive training example $X^+ = (I, P, c, \bar{f}, \bar{v})$, where $c = 1$ indicating P contains a pedestrian, and $\bar{f} = (\bar{f}^{\mathrm{x}}, \bar{f}^{\mathrm{y}}, \bar{f}^{\mathrm{w}}, \bar{f}^{\mathrm{h}})$ and $\bar{v} = (\bar{v}^{\mathrm{x}}, \bar{v}^{\mathrm{y}}, \bar{v}^{\mathrm{w}}, \bar{v}^{\mathrm{h}})$ are regression targets for full body and visible part estimation respectively. As in [14, 16], we define $\bar{f}$ as

$$\begin{aligned} \bar{f}^{\mathrm{x}} &= \frac{\bar{F}^{\mathrm{x}} - P^{\mathrm{x}}}{P^{\mathrm{w}}}, \quad \bar{f}^{\mathrm{y}} = \frac{\bar{F}^{\mathrm{y}} - P^{\mathrm{y}}}{P^{\mathrm{h}}}, \\ \bar{f}^{\mathrm{w}} &= \log(\frac{\bar{F}^{\mathrm{w}}}{P^{\mathrm{w}}}), \quad \bar{f}^{\mathrm{h}} = \log(\frac{\bar{F}^{\mathrm{h}}}{P^{\mathrm{h}}}). \end{aligned} \tag{4}$$

Similarly, $\bar{v}$ is defined as

$$\begin{aligned} \bar{v}^{\mathrm{x}} &= \frac{\bar{V}^{\mathrm{x}} - P^{\mathrm{x}}}{P^{\mathrm{w}}}, \quad \bar{v}^{\mathrm{y}} = \frac{\bar{V}^{\mathrm{y}} - P^{\mathrm{y}}}{P^{\mathrm{h}}}, \\ \bar{v}^{\mathrm{w}} &= \log(\frac{\bar{V}^{\mathrm{w}}}{P^{\mathrm{w}}}), \quad \bar{v}^{\mathrm{h}} = \log(\frac{\bar{V}^{\mathrm{h}}}{P^{\mathrm{h}}}). \end{aligned} \tag{5}$$

We consider P as a negative pedestrian proposal if $\mathrm{IOU}(P, \bar{F}) < 0.5$ for all annotated pedestrian examples Q in the same image. There are two types of negative pedestrian proposals: background proposals which have no visible part region and poorly aligned proposals ($0 < \mathrm{IOU}(P, \bar{F}) < 0.5$). To better distinguish negative pedestrian proposals from positive ones, we choose to shrink the visible part regions of negative pedestrian proposals to their centers. Specifically, for each negative pedestrian proposal P, we construct a negative example $X^- = (I, P, c, \bar{f}, \bar{v})$, where $c = 0$ indicating P does not contain a pedestrian, $\bar{f} = (0, 0, 0, 0)$ and $\bar{v} = (0, 0, a, a)$ with $a < 0$. Since the height and width of the visible part region are both 0, i.e. $\bar{V}^{\mathrm{w}} = 0$ and $\bar{V}^{\mathrm{h}} = 0$, we have $\bar{v}^{\mathrm{w}} = -\infty$ and $\bar{v}^{\mathrm{h}} = -\infty$ according to the definition of $\bar{v}$ in Eq. (5). Ideally, a should be set to $-\infty$. In experiments, we find that if a is too small, it can cause numerical instability. Thus, we set $a = -3$ which is sufficient for the visible part region of a negative pedestrian proposal to shrink to a small region ($\sim \frac{1}{400}$ of the proposal region) at its center.

Let $\mathcal{D} = \{X_i = (I_i, P_i, c_i, \bar{f}_i, \bar{v}_i) | 1 \leq i \leq N\}$ be a set of training examples. Denote by W the model parameters of the deep CNN. Let p_{1i}, p_{2i}, f_i, and v_i be the outputs of the network for the training example X_i. We learn the model parameters W by minimizing the following multi-task training loss:

$$L(W, \mathcal{D}) = L_{\mathrm{C1}}(W, \mathcal{D}) + \lambda_{\mathrm{F}} L_{\mathrm{F}}(W, \mathcal{D}) + \lambda_{\mathrm{C2}} L_{\mathrm{C2}}(W, \mathcal{D}) + \lambda_{\mathrm{V}} L_{\mathrm{V}}(W, \mathcal{D}), \tag{6}$$

where L_{C1} and L_{F} are the classification loss and bounding box regression loss respectively for the full body estimation branch, and L_{C2} and L_{V} are the classification loss and bounding box regression loss respectively for the visible part estimation branch. L_{C1} is a multinomial logistic loss defined by

$$L_{\mathrm{C1}}(W, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} -\log(p^*_{1i}), \tag{7}$$

where $p^*_{1i} = p^0_{1i}$ if $c_i = 0$ and $p^*_{1i} = p^1_{1i}$ otherwise. Similarly, L_{C2} is defined by

$$L_{\mathrm{C2}}(W, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} -\log(p^*_{2i}), \tag{8}$$

where $p^*_{2i} = p^0_{2i}$ if $c_i = 0$ and $p^*_{2i} = p^1_{2i}$ otherwise. For L_{F} and L_{V}, we use the smooth L1 loss proposed for bounding box regression in Fast R-CNN [16]. The bounding box regression loss L_{F} is defined by

$$L_{\mathrm{F}}(W, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} c_i \sum_{*\in\{\mathrm{x,y,w,h}\}} \mathrm{Smooth}_{L1}(\bar{f}^*_i - f^*_i), \tag{9}$$

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

$$\text{Smooth}_{L1}(s) = \begin{cases} 0.5s^2 & \text{if } |s| < 1; \\ |s| - 0.5 & \text{otherwise.} \end{cases} \tag{10}$$

Similarly, L_{V} is defined by

$$L_{\text{V}}(W, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{*\in\{\text{x,y,w,h}\}} \text{Smooth}_{L1}(\bar{v}_i^* - v_i^*). \tag{11}$$

The difference between L_{F} and L_{V} is that negative examples are not considered in L_{F} since $c_i = 0$ for these examples in Eq. (9), while both positive and negative examples are taken into account in L_{V}. During training, the visible part regions of negative examples are forced to shrink to their centers. In this way, the visible part estimation branch and the full body estimation branch are learned to produce complementary outputs which can be fused to improve detection performance. If the visible part estimation branch is trained to only regress visible parts for positive pedestrian proposals, the training of this branch would be dominated by pedestrian examples which are non-occluded or slightly occluded. For these pedestrian proposals, their ground-truth visible part and full body regions overlap largely. As a result, the estimated visible part region of a negative pedestrian proposal is often close to its estimated full body region and the difference between the two branches after training would not be as large as the case in which the visible part regions of negative examples are forced to shrink to their centers. As shown in our experiments, forcing the visible part regions of negative examples to shrink to their centers achieves a larger performance gain than not doing this when the two branches are fused.

We adopt stochastic gradient descent to minimize the multi-task training loss L in Eq. (6). We initialize layers Conv1-1 to Conv4-3 from a VGG-16 model pre-trained on ImageNet [7]. The other layers are randomly initialized by sampling weights from Gaussian distributions. In our experiments, we set $\lambda_{\text{F}} = \lambda_{\text{C2}} = \lambda_{\text{V}} = 1$. Each training mini-batch consists of 120 pedestrian proposals collected from one training image. The ratio of positive examples to negative examples in a training mini-batch is set to $\frac{1}{6}$.

3.4 Discussion

Our bi-box regression approach is closely related to Fast R-CNN [16, 39, 4]. The major difference between our approach and Fast R-CNN is that the deep CNN used in our approach has the additional visible part estimation branch. This branch brings two advantages. First, it can provide occlusion estimation for a pedestrian by regressing its visible part. Second, it can be properly trained to be complementary to the full body estimation branch such that their outputs can be further fused to improve detection performance. This is achieved by training the visible part estimation branch to regress visible part regions for positive pedestrian proposals normally but force the visible part regions of negative pedestrian

proposals to shrink to their centers. To train the visible part estimation branch, we introduce visible part annotations. Also, we exploit both visible part and full body annotations to select better positive pedestrian proposals. Typically, Fast R-CNN selects a pedestrian proposal P as a positive training example if it has large overlap with the full body region of a annotated pedestrian example $Q = (\bar{F}, \bar{V})$, i.e. IOU$(P, \bar{F}) \geq \alpha$. This is a weak criterion for selecting positive pedestrian proposals for partially occluded pedestrian examples as illustrated in Fig. 4(b). For $\alpha = 0.5$, the blue bounding box which poorly aligns with the groud-truth pedestrian example is also selected as a positive training example. With visible part annotations, we can use the stronger criterion defined in Eq. (1). According to this criterion, the blue bounding box would be rejected since it does not cover a large portion of the visible part region.

4 Experiments

We evaluate our approach on two pedestrian detection benchmark datasets: Caltech [9] and CityPersons [41]. Both datasets provide full body and visible part annotations which are required for training our deep CNN.

4.1 Experiments on Caltech

The Caltech dataset [9] contains 11 sets of videos. The first six video sets S0-S5 are used for training and the remaining five video sets S6-S10 are used for testing. In this dataset, around 2,300 unique pedestrians are annotated and over 70% unique pedestrians are occluded in at least one frame. We evaluate our approach on three subsets: Reasonable, Partial and Heavy. The Reasonable subset is widely used for evaluating pedestrian detection approaches. In this subset, only pedestrian examples at least 50 pixels tall and not occluded more than 35% are used for evaluation. In the Partial and Heavy subsets, pedestrians used for evaluation are also at least 50 pixels tall but have different ranges of occlusions. The occlusion range for the Partial subset is 1-35 percent, while the occlusion range for the Heavy subset is 36-80 percent. The Heavy subset is most difficult among the three subsets. For each subset, 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 between 10^{-2} and 10^{0} in log space.

Implementation Details We sample training images at an interval of 3 frames from the training video sets S0-S5 as in [17, 39, 41, 34, 44, 4]. Ground-truth pedestrian examples which are at least 50 pixels tall and are occluded less than 70% are selected for training as in [44]. For pedestrian proposal generation, we train a region proposal network [4] on the training set. ∼1000 pedestrian proposals per image are collected for training and ∼400 pedestrian proposals per image are collected for testing. We train the deep CNN in Fig. 3 with stochastic gradient decent which iterates 120,000 times. The learning rate is set to 0.0005 initially

Table 1. Results of Fast R-CNN with varying β. Numbers are log-average miss rates.

	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 0.9$
Reasonable	10.6	10.5	10.1	**9.9**	10.2
Partial	19.1	18.1	**17.2**	18.5	19.2
Heavy	48.9	48.4	**46.1**	48.1	50.6

Table 2. Results of different approaches on the Caltech dataset. Numbers refer to log-average miss rates.

	FRCN	FRCN+	VPE	FBE	PDOE-	PDOE	PDOE+RPN
Reasonable	10.3	10.1	9.8	10.0	9.7	9.4	**7.6**
Partial	19.1	17.2	17.5	17.7	16.4	14.6	**13.3**
Heavy	49.4	46.1	45.5	45.3	45.1	**43.9**	44.4

and decreases by a factor of 0.1 every 60,000 iterations. Since Fast R-CNN is the most relevant baseline for our approach, we also implement Fast R-CNN using the full body estimation branch of our deep CNN.

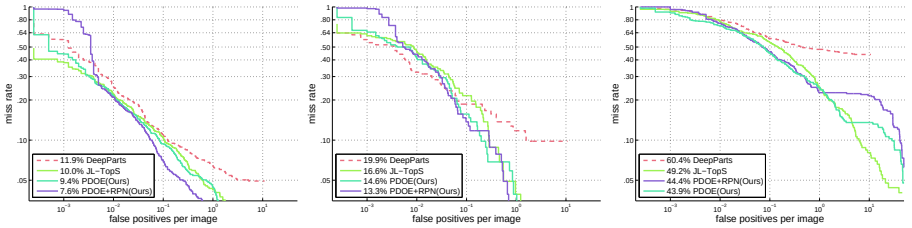
Influence of Positive Pedestrian Proposals We first analyze the influence of positive pedestrian proposals on Fast R-CNN. We conduct a group of experiments in which Fast R-CNN uses the criterion defined in Eq. (1) with α set to 0.5 and β set to 0.1, 0.3, 0.5, 0.7 and 0.9 respectively. The results on the Reasonable, Partial and Heavy subsets are shown in Table 1. We can see that Fast R-CNN works reasonably well with $\beta = 0.5$. When α is fixed, β controls the quality and number of positive pedestrian proposals for training. When β is small, more poorly aligned pedestrian proposals are included. A large β excludes poorly aligned pedestrian proposals but reduces the number of positive training examples. From the results in Table 1, we can see that both the quality and number of positive pedestrian proposals are important for Fast R-CNN. $\beta = 0.5$ achieves a good trade-off between the two factors. In the remaining experiments, we use $\alpha = 0.5$ and $\beta = 0.5$ unless otherwise mentioned.

Ablation Study Table 2 shows the results of different approaches on the Caltech dataset. FRCN is a standard implementation of Fast R-CNN using the full body estimation branch with $\alpha = 0.5$ and $\beta = 0$ for positive pedestrian proposal selection. FRCN+ uses the same network as FRCN but sets $\alpha = 0.5$ and $\beta = 0.5$. We can see that FRCN+ performs better than FRCN on all the three subsets since it uses a sufficient number of better positive pedestrian proposals for training. VPE, FBE and PDOE are three approaches which use the same deep CNN learned by the proposed approach, but score pedestrian proposals in different ways as described in Section 3.2. They score a pedestrian proposal by the visible part estimation branch (VPE), by the full body estimation branch (FBE) and by combining the outputs from both branches (PDOE) respectively. FRCN+, VPE and FBE have similar performances since they uses the same network structure.

PDOE outperforms VPE and FBE on all the three subsets, which shows that the full body and visible part estimation branches complement each other to achieve better pedestrian classification. To demonstrate the effectiveness of forcing the estimated visible parts of negative pedestrian proposals to shrink to their centers, we implement a baseline PDOE- in which negative examples are ignored in the training loss L_{V} in Eq. (11). Although PDOE- also outperforms VPE and FBE, the performance gain achieved by PDOE- is not as significant as that achieved by PDOE. It is pointed out in [4] that the output from a region proposal network can be fused with the output from a detection network to further improve detection performance. As in [4], we further fuse the outputs from the two networks to score a pedestrian proposal P by $\bar{p}^1 = \frac{\exp(s_1^1+s_2^1+s_3^1)}{\exp(s_1^1+s_2^1+s_3^1)+\exp(s_1^0+s_2^0+s_3^0)}$, where $s_1 = (s_1^0, s_1^1)$ and $s_2 = (s_2^0, s_2^1)$ are raw scores from the pedestrian detection network and $s_3 = (s_3^0, s_3^1)$ are raw scores from the region proposal network. We call this approach PDOE+RPN. PDOE+RPN further improves the performance over PDOE on the Reasonable and Partial subsets.

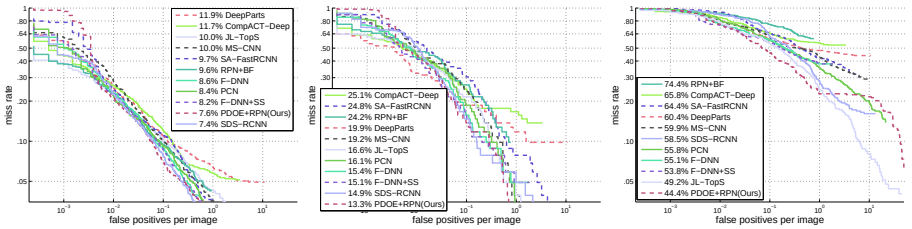
Comparison with Occlusion Handling Approaches To demonstrate the effectiveness of our approach for occlusion handling, we compare it with two most competitive occlusion handling approaches on the Caltech dataset, DeepParts [31] and JL-TopS [44]. Both approaches use part detectors to handle occlusions. Figure 5 shows the results of our approach and the two approaches on the Caltech dataset. Our approach, PDOE, outperforms the two approaches on all the three subsets. Particularly, PDOE outperforms JL-TopS by 0.6%, 2.0% and 5.3% on the Reasonable, Partial and Heavy subsets respectively. The performance improvement on the Heavy subset is significant, which demonstrates that our deep CNN has the potential to handle occlusions reasonably well. PDOE+RPN outperforms JL-TopS on the three subsets with performance improvements of 2.4%, 3.3% and 4.8% respectively. Besides performance improvement over DeepParts and JL-TopS, our approach is able to perform occlusion estimation by regressing visible part regions for pedestrians.

Comparison with State-of-the-art Results In Figure 6, we compare our approach with some state-of-the-art approaches including DeepParts [31], CompACT-Deep [6], SA-FastRCNN [19], MS-CNN [5], RPN+BF [39], F-DNN [10], F-DNN+SS [10], PCN [34], JL-TopS [44], SDS-RCNN [4]. Our approach PDOE+RPN performs slightly worse (0.2%) than SDS-RCNN on the Reasonable subset, but outperforms it by 1.6% and 14.1% on the Partial and Heavy subsets, respectively. The performance gain on the Heavy subset is significant. PCN and F-DNN+SS are two competitive approaches which work fairly well for detecting both non-occluded and occluded pedestrians. Our approach works better than the two approaches on all the three subsets. Note that as our approach, all F-DNN+SS, PCN and SDS-RCNN integrate two or more networks for pedestrian classification. For heavily occluded pedestrian detection, our approach outperforms JL-TopS by 4.8% on the Heavy subset.



(a) Reasonable (b) Partial (c) Heavy

Fig. 5. Comparison of our approach with two competitive occlusion handling approaches on the Caltech dataset. Numbers in legends refer to log-average miss rates.



(a) Reasonable (b) Partial (c) Heavy

Fig. 6. Comparison of our approach and state-of-the-art approaches on the Caltech dataset. Numbers in legends refer to log-average miss rates.

4.2 Experiments on CityPersons

The CityPersons dataset [41] is a recently released pedestrian detection dataset which is diverse in terms of the numbers of countries, cities and seasons it covers. The dataset has a higher pedestrian density than Caltech. 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. 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 Small setup, the height and visibility ranges of pedestrian examples are [50, 75] and [0.65, 1] respectively. In the Heavy setup, the height and visibility ranges of pedestrian examples are [50, ∞] and [0.2, 0.65] respectively. In the All setup, the height and visibility ranges of pedestrian examples are [20, ∞] and [0.2, 1] respectively. As for the Caltech dataset, detection performance is summarized by the log-average miss rate.

Implementation Details We use the Train set for training and the Val set for testing. As in [41], we only use pedestrian examples to collect positive pedestrian proposals and ignore other person examples. Specifically, ground-truth pedestrian examples which are at least 50 pixels tall and are occluded less than 70%

Table 3. Results of different approaches on the CityPersons dataset. Numbers refer to log-average miss rates.

	FasterRCNN	FRCN	FRCN+	VPE	FBE	PDOE-	PDOE	PDOE+RPN
Reasonable	12.81	12.93	12.62	13.01	12.51	12.14	11.53	**11.24**
Small	-	50.12	49.81	50.02	49.65	49.16	47.84	**47.35**
Heavy	-	48.91	47.30	47.54	47.61	46.94	44.91	**44.15**
All	-	46.70	46.20	46.03	45.95	44.90	43.89	**43.41**

are used for training. We also train a region proposal network on the Train set to generate ∼1000 pedestrian proposals per image for training and ∼400 pedestrian proposals per image for testing. Stochastic gradient descent iterates 90,000 times and the learning rate is set to 0.001 initially and decreases by a factor of 0.1 every 45,000 iterations.

Results Table 3 shows the results of different approaches on the CityPersons dataset. Our implementation of Fast R-CNN, FRCN, performs slightly worse than FasterRCNN [41] in the Reasonable setup. With better positive pedestrian proposals for training, FRCN+ outperforms FRCN in all the four setups. FRCN+, VPE and FBE have comparable log-average miss rates due to the same network structure they use. PDOE outperforms both VPE and FBE since the full body estimation branch and visible part estimation branch produce complementary scores which can be further fused to boost detection performance. Compared with PDOE, the performance of the downgraded version of our approach, PDOE-, decreases by 0.61%, 1.32%, 2.03% and 1.01% in the Reasonable, Small, Heavy and All setups respectively. PDOE outperforms the baseline FRCN by 1.4%, 2.28%, 4% and 2.81% in the four setups respectively. Fusing the detection network and region proposal network, PDOE+RPN achieves the best performance.

5 Conclusion

In this paper, we propose an approach to simultaneous pedestrian detection and occlusion estimation by regressing two bounding boxes to localize the full body and visible part of a pedestrian respectively. To achieve this, we learn a deep CNN consisting of two branches, one for full body estimation and the other for visible part estimation. The two branches are properly learned and further fused to improve detection performance. We also introduce a new criterion for positive pedestrian proposal selection, which contributes to a large performance gain for heavily occluded pedestrian detection. The effectiveness of the proposed bi-box regression approach is validated on the Caltech and CityPersons datasets.

Acknowledgement. This work is supported in part by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2015-T2-2-114 and start-up grants of University at Buffalo.

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., Ferguson, D.: Real-time pedestrian deteciton with deep network cascades. In: British Machine and Vision Conference (BMVC) (2015)
2. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: European Conference on Computer Vision (ECCV) (2012)
3. Benenson, R., Mathias, M., Tuytelaars, T., Van Gool, L.: Seeking the strongest rigid detector. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
4. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection and segmentation. In: International Conference on Computer Vision (ICCV) (2017)
5. Cai, Z., Fan, Q., Feris, R., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European Conference on Computer Vision (ECCV) (2016)
6. Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: International Conference on Computer Vision (ICCV) (2015)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
8. Dollar, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI) (2014)
9. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI) (2012)
10. Du, X., El-Khamy, M., Lee, J., Davis, L.S.: Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. CoRR (2016), http://arxiv.org/abs/1610.03466
11. Duan, G., Ai, H., Lao, S.: A structural filter approach to human detection. In: European Conference on Computer Vision (ECCV) (2010)
12. Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.: Multi-cue pedestrian classification with partial occlusion handling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
13. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI) (2010)
14. Girshick, R., Donahue, J., Darrel, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
15. Girshick, R., Felzenszwalb, P., McAllester, D.: Object detection with grammar models. In: Advances in Neural Information Processing Systems (NIPS) (2011)
16. Girshick, R.: Fast r-cnn. In: International Conference on Computer Vision (ICCV) (2015)
17. Hosang, J., Omran, M., Benenson, R., Schiele, B.: Taking a deeper look at pedestrians. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
18. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recogntion (CVPR) (2005)

19. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. CoRR (2015)
20. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Mathias, M., Benenson, R., Timofte, R., Van Gool, L.: Handling occlusions with franken-classfiers. In: International Conference on Computer Vision (ICCV) (2013)
22. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
23. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: International Conference on Computer Vision (ICCV) (2013)
24. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
25. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship in pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
26. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Occlusion patterns for object class detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
28. Shet, V., Neumann, J., Ramesh, V., Davis, L.: Bilattice-based logical reasoning for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR (2014)
30. Tang, S., Andriluka, M., Schiele, B.: Detection and tracking of occluded people. In: British Machine Vision Conference (BMVC) (2012)
31. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: International Conference on Computer Vision (ICCV) (2015)
32. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
33. Tu, Z., Xie, W., Dauwels, J., Li, B., Yuan, J.: Semantic cues enhanced multimodality multi-stream cnn for action recognition. IEEE Transaction on Circuits and Systems for Video Technology (TCSVT) (2018)
34. Wang, S., Cheng, J., Liu, H., Tang, M.: Pcn: Part and context information for pedestrian detection with cnns. In: British Machine Vision Conference (BMVC) (2017)
35. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: International Conference on Computer Vision (ICCV) (2009)
36. Wu, B., Nevatia, R.: 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)
37. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-model deep representations for robust pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
38. Yang, B., Yan, J., Lei, Z., Li, S.: Convolutional channel features. In: International Conference on Computer Vision (ICCV) (2015)

39. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: European Conference on Computer Vision (ECCV) (2016)
40. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
42. Zhou, C., Yuan, J.: Non-rectangular part discovery for object detection. In: British Machine Vision Conference (BMVC) (2014)
43. Zhou, C., Yuan, J.: Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Asian Conference on Computer Vision (ACCV) (2016)
44. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: International Conference on Computer Vision (ICCV) (2017)