Joint Holistic and Partial CNN for Pedestrian Detection

Yun Zhao[1]
jiwuyiyi@stu.xjtu.edu.cn
Zejian Yuan*[1]
yuan.ze.jian@xjtu.edu.cn
Hui Zhang[2]
jet.zhang@forward-innovation.com

[1] Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University
Xi'an, China

[2] Shenzhen Forward Innovation Digital Technology Co. Ltd. China

Abstract

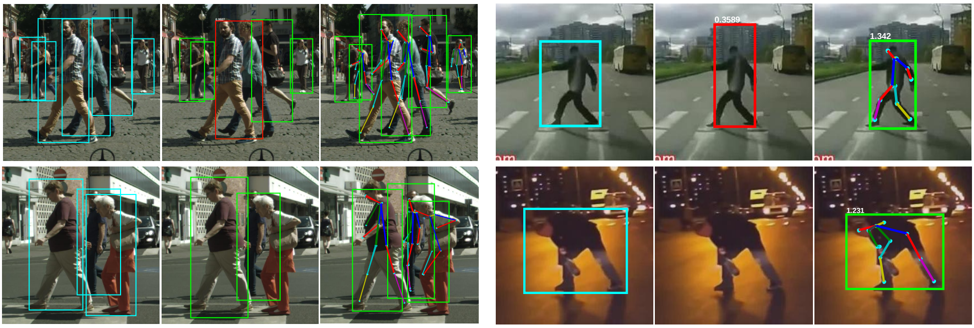
In this paper, we propose a new network combining holistic and partial information for detecting crowd occluded pedestrians and irregularly posed pedestrians. It consists of two parallel sub-networks and a fusion sub-network. We perform holistic bounding box detection and parts prediction respectively in two parallel sub-networks. And then, we embed parts scores into the holistic detected box using a convolutional network. The convolutional network is constructed from a tree-structured model by mapping each step of inference algorithm into an equivalent CNN layer. Finally, we perform non-maximum suppression (NMS) using both detections and their pose to keep the close-by true positive detections. Our detector outperforms the state-of-the-art methods on both CityPersons dataset and Precarious Pedestrian dataset, which demonstrates the effectiveness of our detector, especially for crowed occluded and irregular-posed pedestrians.

1 Introduction

Pedestrian detection is increasingly important for many applications, such as driver assistance system, video surveillance and so on. Though tremendous strides have been made, there are remaining open challenges such as detecting the crowd occluded and irregular-posed pedestrians in near scale [18](80 or more pixels in height).

Crowd occluded pedestrians refer to the pedestrians occluded by each other. As shown in Figure 1(a), the pedestrians gather together and occlude each other. This causes a decrease in feature discrimination, and the appearance pattern formed by close pedestrians is easily confused with that of a single pedestrian. Even worse, a true positive detection is prone to be suppressed by a nearby primary detection of another pedestrian, turning into a missed detection.

Another challenge for pedestrian detection is the large pose variance. Compared with typical walking ones, irregularly posed pedestrians often appear in dangerous scenarios, such as pedestrians tumbled or run across the road. These pedestrians scarcely appear in the dataset and are of diverse pose configurations, which lead to low scoring even missed detections, as shown in Figure 1(b).



(a) (b)

Figure 1: Diverse pedestrians. (a) Crowed occluded pedestrians. (Erroneous detection and promiscuous overlaps) (b) Precarious pedestrians. (Missed detection and low score) Cyan represents ground truth. Green is corrected detection. Red indicates false positives. (Best in color.)

Most competitive pedestrian detectors model the pedestrian as a holistic template. They utilize a bounding box tightly around a person to localize a pedestrian. Various features (ACF [6], LDCF [15], Checkerboards [26]) are extracted according to this bounding box. Then, a holistic template is learned based on these features and detect pedestrians by sliding windows on the feature map. Recently, numerous CNN-based network [2, 21, 27] have been proposed for pedestrian detection. Most of them follow the pipeline of Faster R-CNN [17] which consists of a fully convolutional Region Proposal Network (RPN) for proposing candidate regions and a downstream Fast R-CNN [8] classifier. Although they achieve prevalent successes on the general pedestrian dataset, their performances on detecting crowd occluded pedestrians and irregularly posed pedestrians are still less than satisfactory. An important reason is that a single holistic model is often not expressive enough to represent the rich appearance patterns of the pedestrian.

Modeling a pedestrian with body parts is an effective way to alleviate these difficulties. DPM [7] and its variants [16, 23] are employed for pedestrian detection. They use a star structure to represent an object by a collection of parts arranged in a deformable configuration. Despite the good results on pedestrian detection, the simple star models are incapable of huge pose variance and the hand-crafted HOG [5] features restrict their competitiveness. Numerous human pose estimation methods [3, 20, 22] have been proposed to deal with pose variance. They directly model the anatomical parts and the relationships between parts. They have shown good performances on parts localization. However, the anatomical body parts are not optimal for pedestrian detection, as they might be non-discriminative with respect to the background. Moreover, some of them rely on early committed single-person detections.

We propose a unified CNN to joint holistic and partial information for detecting crowd occluded and irregularly posed pedestrians. It is motivated by the complementary properties of the aforementioned two kinds of models. We impose a shared convolution network to extract low-level features efficiently. Then, the pedestrian detection and parts prediction are done through two branches respectively. After that, we impose a tree-structured model to embed the evidence of each part into the detected bounding box. Specifically, the joint unit implements the inference algorithm of the fusion model as an equivalent convolutional network. To prevent occluded pedestrians from suppression, we define a new non-maximum suppression (NMS) strategy on the basis of detected human pose, named as

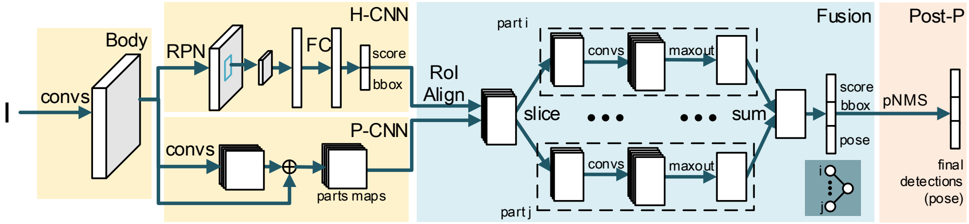


Figure 2: The architecture of our method. It consists of 4 components: body network, partial CNN branch (P-CNN), holistic CNN branch (H-CNN) and fusion module. The fusion module herein shows an example of score propagation on a tree of depth 2 . pNMS is the short for pose-based non-maximum suppression.

poseNMS, which measures the overlap of two bounding boxes by considering both overlapped area and occluded joints in two detections.

The major contributions of this paper are as follows:

1) We propose a unified framework to combine holistic and partial information for pedestrian detection. We design a CNN-based fusion module to integrate evidence of human parts to the detected bounding box, and demonstrate the joint model improves the performance of pedestrian detection.

2) We propose a new NMS strategy (poseNMS) to better select final individual pedestrian from a crowd by using both the information of bounding box and parts.

3) We achieve a state-of-the-art performance on CityPersons dataset, especially on the heavy occluded pedestrian subset. Moreover, our detector gets a significant promotion on the Precarious Pedestrian dataset, demonstrating that our detector is an effective method to detect pedestrians in irregular pose configuration.

2 Proposed method

We integrate holistic detection and parts prediction into one framework, as illustrated in Figure 2. We embed the evidence of each part to the whole bounding box by inferring on a tree-structured model. Specifically, we implement the inference of the fusion model via an equivalent CNN. A diagram of one step in CNN based inference algorithm is shown in Figure 3(c).

2.1 Network Architecture

Our framework is illustrated in Fig. 2. It mainly consists of five components: the body network for feature map generation, a holistic pedestrian detection CNN (H-CNN), a part detection CNN (P-CNN), a fusion model for integrating partial and holistic scores, and a post-process for selecting final detections.

From the very left, an image is forwarded through multiple convolution layers to generate shared feature maps. Then, the network is split into two parallel branches: holistic CNN branch (H-CNN) and partial CNN branch (P-CNN). H-CNN predicts holistic pedestrian detections (boxes and scores) and P-CNN generates parts score maps and parts affinity fields for parts association. Based on the detected proposals from H-CNN, we use a RoIAlign layer [12] to convert the parts maps inside any valid region of interest into small

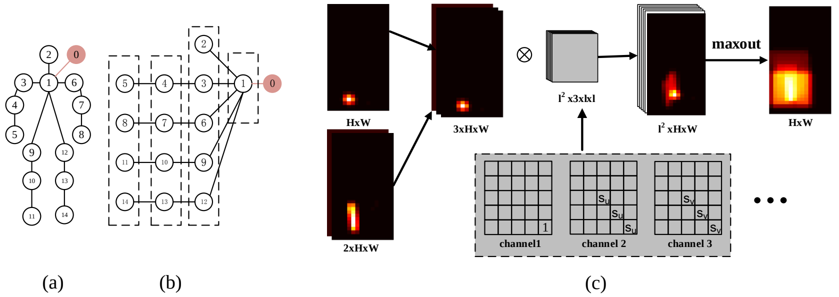


Figure 3: The architecture of our fusion model. (a) The graph of human pose. Each node represent a joint (part) of pedestrian. The 0 node represents the bounding box of detected pedestrian. (b) The parts score is propagate from the left leaf nodes to the root node. (c) CNN-formed computation for maximization in Equation (6). H and W represent the width and height of the unary score map. l is the filter size. The elements in empty cells of filters are set to zeros. $s_u(0.236)$ and $s_v(0.236)$ are computed according Equation (8) under $w_r = 1/3$ and $p'_k = (2,2)$.

maps with a fixed spatial extent of $H \times W$. After that, we embed the score of parts into pedestrian detection through a pictorial structure model, where this process is implemented in a CNN based structure.

Holistic CNN branch: We perform pedestrian detection using the same pipeline proposed in MS-CNN [2], a modification of Faster-RCNN [17] for pedestrian detection. It consists of two sequential stages: region proposal network (RPN) and Fast-RCNN (FRCNN). The RPN takes feature maps from the shared networks and generates candidate object bounding boxes. The FRCNN extracts features using ROIPool from each candidate box and uses these features to perform classification and bounding-box regression. We denote the each detection hypothesis as $b = (x,y,w,h)$ and its score $s(b)$ respectively, where (x,y) is the coordinate of left-top corner and (w,h) is the width and height of bounding box.

Partial CNN branch: We use a pictorial structure to model human parts. Here, we refer human anatomical joints [24] as parts, and denote the configuration of human pose as $\mathbf{p} = (p_1,\ldots,p_K)$, where $p_i = (x_i,y_i)$ and K is the number of the joints. The human pose is organized in a tree structure together with geometric constraints on pairs of parts, as shown in Figure 3(a). The unary score of part i is denoted as $s_i(p_i)$, and the pairwise relational score of each edge (i,j) is $r_{i,j}(p_i,p_j)$.

The part detection branch herein follows the approach proposed in [3]. Given an image, it simultaneously predicts a set of 2D confidence maps $\mathbf{S}$ of body part locations and a set of 2D vector fields $\mathbf{R}$ of part affinities which encode the association between pair-wise parts. This branch is the iterative convolutional networks, following the design in CPM [22].

2.2 Fusion Model

We use the bounding boxes $\{b\}$ from holistic CNN branch as proposal candidate boxes. We convert the parts maps $(\mathbf{S},\mathbf{R})$ inside any valid region of interest b into small maps with a fixed spatial extent of $H \times W$. The inference is performed on the cropped maps.

A pedestrian hypothesis $(b, p_1, \cdots, p_K)$ specifies the proposal bounding box of holistic pedestrian and the location of its K parts. We represent the pedestrian hypothesis by a tree-structured model as shown in Figure 3(a), where node 0 specifies the bounding box and the others indicate the positions of each part (node 1 represents neck). Then, the score of each hypothesis is given as follow:

$$s(b,p_1,\cdots,p_K) = w_0 \cdot s(b) + w_{0,1} \cdot r(b,p_1) + \sum_{i=1} w_i \cdot s_i(p_i) + \sum_{i,j} w_{i,j} \cdot r_{i,j}(p_i,p_j). \quad (1)$$

w_i and $w_{i,j}$ is the weight parameters for each node and connection, which can be trained using SVM. For simplification, they are integrated into the value of corresponding terms and omitted in the following formulation. The unary and pairwise terms are defined as follows:

Unary Terms: The unary terms give evidences for bounding box b and part i to lie at position p_i. The bounding box score $s(b)$ is given by the holistic CNN branch for every bounding box b. The unary score of part $s_i(p_i)$ is got from the score map $\mathbf{S}_i$ at position p_i, where $\mathbf{S}_i$ is generated by partial CNN branch.

Pairwise Relational Terms: These pairwise terms capture the relations between bounding box and parts. $r(b,p_1)$ represents the spatial relationship between bounding box and neck, defined as follows:

$$r(b,p_1) = -||(x_1,y_1) - ((x_0,y_0)+v_1)||^2, \quad (2)$$

where (x_0,y_0) is the position of box left-corner and $v_1 = (x_v,y_v)$ is the anchor position of neck. Instead of using only relative spatial information, we employ an pairwise relational term $r_{i,j}(p_i,p_j)$ as proposed in [3]. It is computed by integrating over the part affinity field maps $\mathbf{R}_{i,j}$ along the line segment as follow:

$$r_{i,j}(p_i,p_j) = \int_{u=0}^{u=1} \mathbf{R}_{i,j}(p_u) \cdot \frac{p_j - p_i}{||p_j - p_i||} du, \quad (3)$$

where $p_u = (1-u)p_i + up_k$ is the interpolation between two joints.

2.2.1 Inference with CNN

To detect pedestrians in an image, we compute an overall score for each proposal bounding box according to the best possible configuration of the parts:

$$\hat{s}(b) = \max_{p_1,\cdots,p_K} s(b,p_1,\cdots,p_K). \quad (4)$$

The computation of the max score can be computed as:

$$\hat{s}(b) = s(b) + \max_{p_1}(r(b,p_1) + \hat{s}_1(p_1)). \quad (5)$$

Here, the upstream score from pose $\hat{s}_1(p_1)$ are recursively computed from the leaf nodes as described in [1]:

$$\hat{s}_i(p_i) = s_i(p_i) + \sum_{k \in \mathbb{K}(i)} \max_{p_k}(r_{i,k}(p_i,p_k) + \hat{s}_i(p_k)), \quad (6)$$

where $\mathbb{K}(i)$ is the set of children of node i in the graph ($\mathbb{K}(i) = \emptyset$), if node i is a leaf.

To perform the whole framework in an elegant end-to-end CNN, we rephrase the inference algorithm of the fusion model into a convolutional manner. Considering that p_k is located in a l width regular grid $\mathcal{M}$ centered at p_i, we reformulated the unary and pairwise terms in Equation 6 into:

$$\begin{aligned} \hat{s}_k(p_k) &= \sum_{q\in\mathcal{M}} \mathbf{1}_{q=p_k} \cdot \hat{s}_k(q), & (7) \\ r_{i,k}(p_i,p_k) &= w_r \sum_{q\in\mathcal{M}} \mathbf{1}_{q\in\{p_u\}} \cdot R_{i,k}(q), & (8) \end{aligned}$$

where w_r is short for $(1/N_u)\cdot(p_k-p_i)/||p_k-p_i||$, N_u is the number of interpolations.

Obviously, these both terms can be computed by filtering with designed filters. We design l^2 filters. Each filter represents a p_k in $\mathcal{M}$. We transform p_k into a coordinate centered at p_i and denote it as p'_k. The filters are constructed as follows:

1) Channel 1: only the one element at p'_k is set to 1 and the others are set to 0.

2) Channel $2-3$: the points on the line from center to p'_k is set to $1/N_u \cdot p'_k/||p'_k||$, and the other points are set to 0. N_u herein represents the number of no-zero elements.
A designed filter is shown in Figure 3(c). The p'_k is $(2,2)$ relative to filter center.

For pair-wise parts i and k, we concatenate $\mathbf{S}_k$ and $\mathbf{R}_{i,k}$ in channel-wise as a slice of 3-channel feature map. We filter the designed filters on the feature map, and then we get a l^2-channel score map. The maximization term in Equation 6 is finally computed by a maximization over l^2 channels, as "maxout" unit described by Goodfellow *et al.* [10].

Given the upstream score map of neck $\hat{s}_1(p_1)$, we compute the final score $\hat{s}(b)$ following Equation (5). This step is implemented using a CNN as proposed in [9].

Our inference algorithm of the fusion model can be seen as an extension of DPM-CNN [9]. There are mainly two modifications. First, we extend the star model to tree-structured model. Second, we design specific filters that rephrase the computation of pairwise cost as convolutional filtering. This CNN form brings conveniences for the implement of the whole framework on GPU.

2.3 Learning

Partial CNN branch: We initialize the parameters of body network and P-CNN using the pre-trained model on COCO [14] dataset. Then we fine-tune the network on CityPersons dataset using the anatomical landmarks annotated identically as COCO key-points. In experiments, we find that, even without fine-tuning, the pose estimation network works well.

Holistic CNN branch: We fix the body network and train only pedestrian detection network. During the training of H-CNN, we use the same loss function as proposed in MS-CNN [2]. Online Hard Example Mining (OHEM) [19] is used to choose negative samples and accelerate convergence. The H-CNN is trained on pedestrian dataset.

Fusion model: Instead of being trained as typical CNNs, the parameters of filters are fixed into designed values as described in Section 2.2.1. We denote the proposals containing pedestrian as positive examples and the proposals without pedestrian as negative examples. We use SVM to learn the weight parameters $(w_0,\cdots,w_i,w_{0,1},\cdots,w_{i,j})$. We compute the anchor position of the neck as [24] using the joints of pedestrian annotated on CityPersons dataset.

2.4 PoseNMS

Given detection hypotheses set $\{(b, p_1, \cdots, p_K)\}$, we try to get the final detections through non-maximum suppression (NMS). We define a new NMS scheme based on estimated pose called poseNMS. This scheme greedily chooses the detections with the highest scores and abandons their neighbors. Different from GreedyNMS which uses the intersection over union (IoU) on bounding boxes to find these rejected neighbors, our method also takes the similarity of poses into consideration.

We firstly select a highest scored detection hypotheses with pose $\mathbf{p}_i$. Then, the m-th part in a close-by j-th detection $\mathbf{p}_j^m$ is labeled overlapped if it falls with $\alpha \cdot \max(h_i, w_i)$ pixels of the corresponding joint $\mathbf{p}_i^m$ from the anchor i-th detection . h_i and w_i are the height and width of the bounding box of i-th detection, and α controls the relative threshold for considering overlap of joints ($\alpha = 0.2$). We denote the subset of occluded joints in j-th proposal as $\mathbf{o}_j$.

In practice, high scoring detection may include some low scoring joints on account of occlusion. So, we select the joints whose scores are higher than β (set to 0.3) and denote them as $\hat{\mathbf{p}}_j$. Meanwhile, we shrink the occluded joint set into $\hat{\mathbf{o}}_j$ by discarding the joints whose scores are higher than those in anchor detection. Then a new metric is defined as:

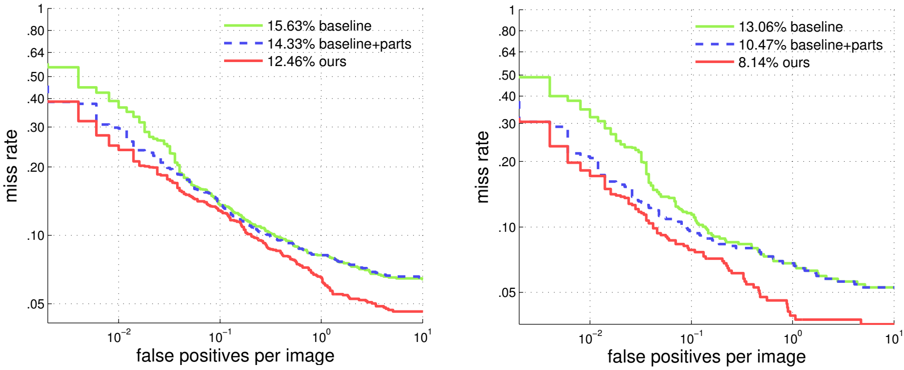
$$IoU_{pose} = \sum_{m \in \hat{\mathbf{o}}_j} s_j^m / \sum_{m \in \hat{\mathbf{p}}_j} s_j^m \tag{9}$$

where s_j^m is the score of m-th joint in j-th detection hypothesis.

The final IoU used for NMS is computed as: $IoU = IoU_{overlap} + w_{pose} IoU_{pose}$, where $IoU_{overlap}$ represents the typical IoU on bounding box. In experiments, we set $w_{pose} = 1$ and the rejection IoU threshold to 1. For these bounding box with no joints or too fewer joints, we use original GreedyNMS. As shown in Figure 1(a), the pose-NMS makes better localizations for crowd occluded pedestrians.

3 Experiments

Dataset and Metric: Our method is evaluated on both CityPersons [27] and Precarious pedestrian dataset [13]. CityPersons [27] is a new pedestrian detection dataset on top of the semantic segmentation dataset CityScapes [4]. It provides a total of ∼ 35k person and ∼ 13k



(a) Reasonable (b) Near scale + Partial occlusion

Figure 4: Ablation experiments on two CityPersons validation subsets.

ignored region annotations in 5000 frames. Both bounding box annotation of all persons and visible parts are provided. The dataset is divided into train/validation/test subsets with a consistent density of persons. All of our experiments involved CityPersons are conducted on the reasonable subsets [18] (height > 50 pixels and overlap < 35%) for training and testing.

Precarious pedestrian dataset [13] is a novel annotated pedestrian dataset of dangerous scenarios with a total of 951 reasonable images. It contains various kinds of scenes, such as children running on the road, people tripping and so on. A pedestrian is annotated using a bounding box totally surrounding a person. The Precarious dataset is split equally for training and testing.

For evaluation on CityPersons, the log miss rate averaged over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ (MR^{-2}) [18] is used. On Precarious dataset, we compare the miss rate at 10^{-1} false positive per image as [13].

Baseline: Our baseline detector follows MS-CNN [2], a modification of Faster-RCNN [17] for pedestrian detection. We use VGG-16 to initialize the backbone network. The difference between our implementation and theirs is that we employ Online Hard Example Mining (OHEM) [19] to choose hard negative samples and accelerate convergence

3.1 Ablation Study

The effect of partial information is verified by comparing our detector with the baseline detector. The *baseline+parts* represents a detector using only part scores for final detections. It means that the holistic CNN only provides candidate bounding boxes for the fusion model. The final detector (*ours*) combines the score of the holistic bounding box and parts and utilizes the poseNms for selecting the final detections. We compare the detectors on two CityPersons validation subsets, Reasonable (height > 50 pixels, occlusion < 35%) and Near scale+Partial occlusion (height>80 pixels, 10% $<$occlusion $< 35\%$). The results are shown in Figure 4.

It can be seen from Figure 4(a), the partial information is more helpful for performance under lower FPPI. With fusing holistic and partial scores and employing poseNms, our method gains an overall improvement. It is due to the facts that the part scores raise the confidence of relatively high scored detections and the poseNms keeps the nearby occluded pedestrians. The improvements are more obvious on the near occluded pedestrian subset, as shown in Figure 4 (b). It demonstrates the effectiveness of our detector on detecting near occluded pedestrians.

3.2 Comparisons with State-of-the-art Methods

3.2.1 Citypersons dataset:

We compare our method with state-of-the-art pedestrian detectors on CityPersons. Checkerboards [26] is the state-of-the-art method without using CNN. Zhang *et al.* [27] adapt FasterRCNN [17] and obtain competitive performance. Wang *et al.* [21] replace the VGG-16 backbone with a faster and liter ResNet-50 [11] network. To demonstrate our effectiveness on occlusion, we divide the reasonable subset into the partial subset (10% $<$occlusion $< 35\%$) and the bare subset (occlusion $< 10\%$). The results are shown in Table 1. Our proposed method achieves 12.5%MR^{-2}, which is an absolute 3.1-point improvement over our baseline. Moreover, it outperforms the state-of-the-art methods. In terms of different occlusion levels, our improvements on the Partial subset is more obvious.

Method	Scale	Reasonable	Partial	Bare
Checkerboards [26]	×1	25	-	-
Zhang *et al.* [27]	×1	15.1	-	-
	×1.3	12.8	-	-
RepLoss [21]	×1	13.2	16.8	7.6
Baseline	×1	15.6	17.4	9.6
Ours	×1	**12.5**	**14.2**	**7.1**

Table 1: Pedestrian detection results ($\mathrm{MR}^{-2}(\%)$) on the CityPersons [27]. Models are trained on train set and tested on validation set respectively.

Method	50% overlap	70% overlap
LDCF [15]	71.64%	87.37%
RPN/BF [25]	54.52%	82.8%
RPN+ [13]	42.47%	73.7%
MSCNN [2]-RPN	27.02%	74.6%
MSCNN [2] (Baseline)	13.72%	54.5%
Ours	**5.82%**	**35.97%**

Table 2: Miss rate of different detectors under different overlap ratio criteria on Precarious [13]. We denote the miss rate at 10^{-1} FPPI. Ours is extremely better.

3.2.2 Precarious dataset

To demonstrate our effectiveness on pose variance, we compare our approach with other state-of-the-art methods. Note that the state-of-the-art RPN+ [13] uses only region proposal network (RPN). We train MS-CNN [2] on the precarious train dataset as our baseline detector. The results on reasonable [18] test dataset are presented in Table 2. Interestingly, our baseline detector (MS-CNN [2]) have already gained huge improvements on performance comparing with RPN+ [13]. The huge improvements attribute to the fine modification [2] for pedestrian detection and the additional detection sub-network [17].

Our final result is shown in Table 2. During testing, we refine the final detected bounding box using the estimated pedestrian pose. This is implemented using four linear functions that map a feature vector containing the positions of the original bounding box and each part to the upper-left corner and width-height of the final bounding box. Our detector significantly outperforms all other methods. By combining the holistic and partial information, our detector gains a 7.9% improvement at 0.5 overlap ratio criteria and a higher 18.53% improvement at stricter 0.7 criteria.

3.2.3 Instance analysis

We show some instance results on both Citypersons and Precarious datasets in Figure 5. The red boxes indicate the wrong detections and the red dash boxes represent the missed detections. The first row shows some wrong results of our baseline detector using only holistic CNN branch. They are caused by crowd occlusion or irregular human pose. The results of our final detector are shown in the second row where our detector can effectively detect the crowd occluded and irregular-posed pedestrians. Meanwhile, we show some detections using only partial CNN branch in the third row of Figure 5. The first two

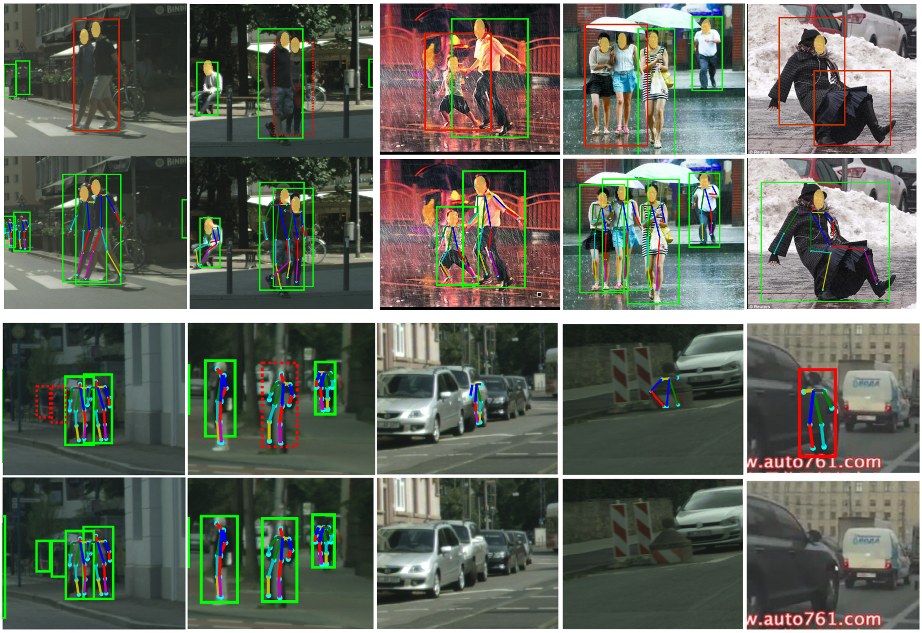


Figure 5: Instance results on Citypersons and Precarious datasets. The first row shows the results of our baseline detector (only holistic CNN branch). The third row shows the results of our detector for parts (only partial CNN branch). The second and fourth rows show the results of our final detector. The red boxes represent false positives. The green ones are true positives. The red dash box represents the missed detection.

columns show that the partial CNN branch is incapable of handling small scaled pedestrians and intermediately truncated pedestrians. The last third columns indicate the partial branch is easy to generate false positives on some vehicles and traffic signs The fourth row shows the results of jointing holistic and partial information. It can effectively detect the small scaled pedestrian and eliminate the false positives.

4 Conclusion

In this paper, we propose a unified CNN to joint holistic and partial information for detecting crowd occluded and irregularly posed pedestrians. We impose a tree-structured model to embed the evidence of each parts into the detected bounding box and perform the joint unit as a convolutional network. Benefit from complementary properties of holistic and partial models, our detector achieves the best performance on Citypersons [27] and Precarious [13] datasets. The superiority of dealing with the unusual pedestrian on Precarious dataset demonstrates that our detector has great potential to be integrated into real-world applications.

Acknowledgement: This work was supported by the National Key R&D Program of China (No.2016YFB1001001), the National Natural Science Foundation of China (No.91648121, No.61573280), and Guangdong Provincial Science and Technology Plan Project (No.2016A010101039).

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[2] 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)*, pages 354–370. Springer, 2016.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEEE Conference on*, volume 1, page 7. IEEE, 2017.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 3213–3223. IEEE, 2016.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition(CVPR), 2005 IEEE Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 304–311. IEEE, 2009.

[7] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[8] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448. IEEE, 2015.

[9] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 437–446, 2015.

[10] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1319–1327, 2013.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 770–778. IEEE, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017.

[13] Shiyu Huang and Deva Ramanan. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4664–4673. IEEE, 2017.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[15] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, 2014.

[16] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In *European conference on computer vision (ECCV)*, pages 241–254. Springer, 2010.

[17] 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*, pages 91–99, 2015.

[18] Bernt Schiele, Bernt Schiele, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[19] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 761–769. IEEE, 2016.

[20] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014.

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

[22] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 4724–4732. IEEE, 2016.

[23] Junjie Yan, Xucong Zhang, Zhen Lei, Shengcai Liao, and Stan Z Li. Robust multi-resolution pedestrian detection in traffic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3033–3040. IEEE, 2013.

[24] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2013.

[25] 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)*, pages 443–457. Springer, 2016.

[26] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Filtered channel features for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, volume 1, page 4. IEEE, 2015.

[27] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, volume 1, page 3. IEEE, 2017.