2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT) | 978-1-6654-6910-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/CNIOT55862.2022.00045

Human-Cascaded network for Robust Detection of Occluded Pedestrian

Zhewei Xu, Xiufeng Fu, Dacheng Feng, Wei Li
Beijing Institute of
Computer Technology and Applications
Beijing, China
{xuzhewei,fuxiufeng2022,fengdacheng2022,liweii2022}@163.com

Yang Liu
National Digital Switching System
Engineering Technology Research Center
Henan Province, China
liuyang198610@163.com

Abstract—Occlusion is a key challenge in real-world on-road pedestrian detection task. Due to constrained viewpoint geometry, a pedestrian is very likely to be obstructed by other pedestrians and/or other objects such as cars and bicycles. For Advanced Driver Assistance System (ADAS), heavily occluded pedestrians are as important as reasonable pedestrians because they may burst out from crowds or roadside obstacles. In this paper, a human-cascaded network is proposed for robust detection of heavily occluded pedestrians. Specifically, a sharp-response proposal network (SRPN) is designed to refine the feature responses in a narrow area to handle crowded pedestrian detection, followed by outputting the head and full body proposals for diverse occlusion situations. After RoI-pooling, a visible-guided attention (VGA) module is developed to leverage the head and visible area information. The VGA module also suppresses the feature noise of occluded area to enhance the feature representation learning of the backbone network. Finally, a head-cascade RCNN (HRCNN) network is proposed to predict the pedestrian bounding box from the head proposal. The proposed approach is validated through a widely used pedestrian detection dataset: CityPersons. Experimental results show that our approach achieves promising detection performance (log-average miss rate, MR) improvement of 11.4% on heavy occlusion subset, compared to the baseline detector.

Keywords-Human cascade; Pedestrian Detection; Occluded Pedestrians;

I. Introduction

Pedestrian detection is a key component for a wide range of real-world applications, such as Advanced Driver Assistant Systems (ADAS), video surveillance or crowd analysis [1]. Recently, the performance of pedestrian detector has been rapidly improved with the development of deep convolutional neural networks (CNN) [2], [3], [4], [5], [6], [7]. However, while there is great progress on reasonable pedestrians, real-world pedestrian detection in ADAS is still struggling under frequent and severe occlusions [8], [9], [10], [11], [12], [13]. Recently, many top conferences have reported a lot of work on occlusion pedestrian detection task [2], [14], [15], [16]. For ADAS, the heavily occluded pedestrians are as important as reasonable pedestrians, because they may burst out from crowds or roadside obstacles so that the ADAS cannot detect in time, resulting to accidents or even death [17]. Heavily occluded pedestrians not only are more difficult for the detectors, but also cause great distress to the driver. Since there are large variations in scales, ratios, and poses in occlusion scenes [7], [18], detection robustness becomes a challenging issue. Furthermore, when pedestrians overlap largely with each other, semantic features of different instances also interweave so that the target bounding box always mistakenly shifts to another pedestrian [2], [8]. Lastly, when pedestrians are occluded by roadside obstacles, the detector's discriminative abilities are degraded due to lacking body part information and also including noise background features [15], [16]. Therefore, recent benchmarks specifically focus on heavily occluded pedestrian detection. For instance, the CityPersons [13] dataset has about 70% of pedestrians depicting various degrees of occlusions, and more than half of occluded pedestrians are under heavy occlusion, as shown in Fig. 1.

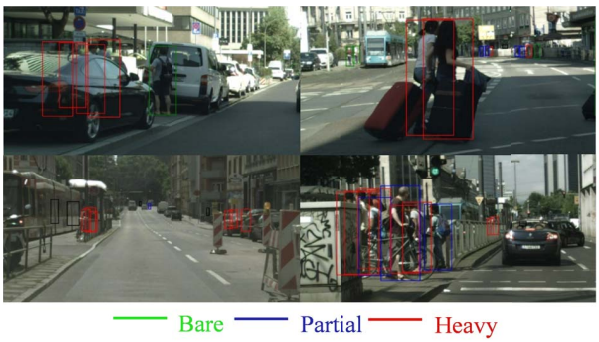

Figure 1. Pedestrian examples in the Citypersons Dataset. The bare occlusion (visible region $\geq$ 90%) pedestrians are green line. The partial occlusion (65% $\leq$ visible region $<$ 90%) pedestrians are blue. The heavy occlusion (visible region $<$ 65%) pedestrians are red. Occluded pedestrians are important and challenging for ADAS.

Generally, pedestrian detectors employ a holistic detection strategy [7], [6], [12], [19] that assumes the CNN can handle obstacle noise when trained using full body bounding box annotations. It is because the detectors only pay attention to reasonable pedestrians whose body

978-1-6654-6910-4/22/$31.00 ©2022 IEEE
DOI 10.1109/CNIOT55862.2022.00045

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

is most visible, e.g. the occlusion area less 35%. However, since clutters are introduced by occlusions within full body annotation, the holistic detection strategy could degrade detection performance on occluded pedestrians, especially the heavily occluded ones. Lately, a common solution to tackle occlusion problem is the learning of a mixture model which contains a series of instance patterns or part detectors. Under this strategy, when the holistic detector cannot recognize a occluded pedestrian, the visible part of the pedestrians still can provide high level of confidence. Then the mixture model summarizes all part detectors to discriminate the occluded pedestrian. Therefore, the robustness of each part detector is very significant. However, most pedestrian detection datasets lack part annotations. Most previous works [1], [20], [21], [22], [23] generate part annotations by analyzing the differences between visible and full-body bounding boxes for each person but it is not accurate because of the large diversity of poses and occlusion patterns. In addition, training of such mixture model is very difficult and complicated fusion strategy is required.

From these perspectives, a human-cascaded network is proposed in this work to improve pedestrian detection in diverse occlusion scenes by utilizing the body, visible part and head information. In real-world scenarios, the pedestrian head is usually less overlapped with a better visibility. Compared with the pedestrian body, it is thus more robust for the detection of diverse obstacle occlusion and crowd scenes. Subsequently, a sharp-response proposal network (SRPN) is proposed to output full-body and head proposals from more acute feature response. However, head is with small size and hence insufficient information, the detection result based on head underperforms the center point [3]. It is because the visible region contains more information without noise feature but the scale of visible region is diverse. Therefore, our proposed method keeps the original full-body detector while adding a visible-guided attention (VGA) module to enhance the feature on visible part and simultaneously suppress the noise area. Apart from this, a head-cascade R-CNN (HRCNN) module is proposed to detect heavily occluded pedestrians from head proposals. To summarize, the main contributions of this work are:

1) A novel human-cascaded network is proposed for robust detection of occluded pedestrians. Compared with the Faster R-CNN, the proposed network has three modifications for occlusion handling: sharp-response proposal network (SRPN), visible-guided attention module (VGA), and head-cascade R-CNN (HRCNN) detection.

2) The SRPN is designed to refine the feature responses in a narrow area to handle crowded pedestrian detection and output the head and full body proposals for diverse occlusion situations. Inspired from human eye system, the SRPN directly predicts the center and scale of the target without pre-defined bounding box.

3) The VGA module is proposed to leverage the visible area information and suppress the feature noise on occluded area so that the feature representation learning of the detection network is enhanced for full-body proposal. It produces a visible attention mask using visible bounding box annotation. Then the mask is applied on the RoI-pooling feature to filter the noise feature.

4) HRCNN uses a resampling procedure to extend the receptive field by different cascade stages. The first stage refines the head location and generates a rough full-body bounding box. The second stage estimates the visible region from the rough full-body bounding box. Finally, the third stage uses the visible region feature to predict the refined full-body bounding box. HRCNN uses head cues and combines the visible features to enhance the reliability of detection results.

In our work, the proposed approach is validated on widely used pedestrian detection datasets: CityPersons. Experimental results show that our approach achieves promising performance (log-average miss rate) of 45.3%, 10.1%, 10.8% on heavy occlusion, partial occlusion and reasonable subsets, respectively.

II. Related Work

A. Generic Pedestrian detection

Early generic pedestrian detectors extended from Viola and Jones paradigm [24] predominated the field of pedestrian detection for years, after that a boosted decision forest, such as ACF [25], LDCF [26], and Checkerboards [18], filtering various Integral Channels Features (ICF) [27]. In recent years, a new generation of more effective pedestrian detectors based on deep convolutional neural network (CNN) significantly improves the state-of-the-art performance. Most of the top accuracy detectors are two-stage approach [19], [28], which first generates a pool of object proposals by a separated proposal generator (e.g., RPN [29]), and then predicts the class label with accurate location and size of each proposal. In [19], the RoI-pooling features were fed into a boosted forest to mining hard examples. In [7], fine-grained attention masks are encoded into convolutional feature maps, which significantly suppress background interference and highlight pedestrians. While in Cai et al. [30], they proposed a Cascaded R-CNN network for object detection by improving localization accuracy based on resampling RoI feature, which outperformance than Faster R-CNN.

B. Occluded Pedestrian Detection

Handling occlusion in pedestrian detection is challenge problem. Most of early works [20], [22], [23] use part-based model to represent the pedestrian, which learn a

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

series of specific part detectors. In [20], [23], they train an ensemble model for most occlusion patterns, but the computation is expensive. Zhou et al. [31] propose a jointly learn part detectors to exploit part correlations and reduce the computation cost. Further, Zhang et al. [32] propose a channel-wise attention mechanism to learn discriminative features for different occlusion patterns in one unit model. Recently, several works have exploited visible and head information to handle occlusion. Zhou et al. [33] design a method to detection full body and visible part simultaneously. Different from [33], Pang et al. [16] propose mask-attention model to enhance visible part area importance. Chi et al. [2], [14] proposed two method handle occlusion pedestrian in crowd scenes using head information. Although numerous pedestrian detection methods are presented in recent years, how to robustly detect occluded pedestrian is still one of the most critical issues for pedestrian detection. Unlike above method, we propose a novel human-cascade network that is not restricted to only certain types of occlusion patterns but pays more attention to important visible area and head area. Furthermore, we designed a new cascade mechanism for detecting pedestrians using head information.

III. Method

The overall framework is shown in Fig. 2. Feature Pyramid Network (FPN) [34] with ResNet-50 [35] is adopted as the backbone network. We first designed an anchor-free method (in our work, sharp-response proposal network (SRPN)) to detect the head and full-body bounding boxes. Afterwards, the full-body proposals are sent into a Fast RCNN with a visible-guided attention (VGA) module. The VGA predicts a visible area probability map and then performs a pixel-wise dot product with the original feature map. This way enhances the visible part importance and suppresses the response of inter-class occluded parts. In addition, the head proposals are sent into a head-cascade RCNN (HRCNN), which is newly designed to regress the pedestrian bounding box from head to visible area, and then to the full-body. The specific details of each module are described below.

A. Sharp-response proposal network

In this work, a sharp-response proposal network (SRPN) is designed to refine the feature responses in a narrow area, which handles crowded pedestrian detection and then outputs the head and full body proposals for diverse occlusion situations. In fact, human vision system recognizes the locations of instance in space and predicts the boundary given the visual cortex map without any pre-defined shape template. From this fact, SRPN is motivated from the human eye mechanism: directly regresses the center and scale of the target without pre-defined bounding box. The architecture of SRPN is illustrated in Fig. 3.

1) Proposal Module. The FPN outputs five scale feature maps, namely P3 to P7, where Pl has a resolution of 2^l lower than the input, and l = 3 to 7 indicates pyramid level as the human eye with different focal lengths. Since the convolution kernels size are same, the shallower feature maps can provide more precise localization information, while the coarser ones contain more semantic information with increasing sizes of receptive fields. Therefore, we perform SRPN on each pyramid levels. Upon the pyramid feature maps P3 to P7, two parallel proposal networks are appended to respectively output the head and body proposals. In this work, each proposal network has the same structure: a single 3x3 convolution layer and three sibling 1x1 convolution layers, which produces the center map, scale map, and shift map, respectively.

2) Target ground truth and loss function. For the center map, the average binary cross-entropy loss is applied to train this branch. The target center map is applied with a 2D Gaussian heatmap G at the location (i, j) of each positive in the image. Formally, it is formulated as:

$$G(i,j,x,y,\sigma) = exp(-(i-x)^2/(2\sigma^2) - (j-y)^2/(2\sigma^2)) \tag{1}$$

$$GT_{center}(i,j) = \max_{k=1,2,\ldots,K} G(i,j,x_k,y_k,\sigma) \tag{2}$$

where K is the number of positives in an image, (x_k, y_k) is the center location of k-th person body, the variance σ of the Gaussian heatmap is related to the stride(s) of pyramid feature, (i, j) is the coordinate of the feature map. To combat the extreme positive-negative imbalance problem, the focal weights [36] on hard examples are also adopted. Thus, the center loss can be formulated as:

$$L^l_{center} = -1/K \sum_{i=1}^{W/2^l} \sum_{j=1}^{H/2^l} a_{ij}(1-p_{ij})^\gamma log(\hat{p}_{ij}) \tag{3}$$

$$\hat{p}_{ij} = \begin{cases} p_{ij} & if\ GT_{center}(i,j) = 1 \\ 1 - p_{ij} & otherwise \end{cases} \tag{4}$$

$$a_{ij} = \begin{cases} 1 & if\ GT_{center}(i,j) = 1 \\ (1 - GT_{center})^\beta & otherwise \end{cases} \tag{5}$$

In the above, $p_{ij} \in [0, 1]$ is the probability of object center, l is the level of feature map, γ and β are hyper-parameters controlling the focal weight distribution. As suggested in [3], [37] $\gamma = 2$ and $\beta = 4$ are set. For scale map, the Smooth L1 loss [38] is applied to train

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

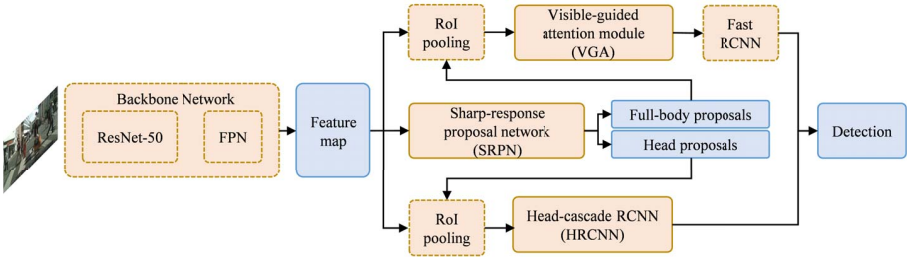

Figure 2. The Human-cascade network overall framework. It contains Backbone network with ResNet-50 and FPN, SRPN, VGA module and HRCNN. The proposed components are shown with orange solid rectangle. Based on pyramid feature map, SRPN generates full-body and head proposal. The full-body proposals input Fast R-CNN branch with VGA module. The head proposals input Head-cascade R-CNN branch.

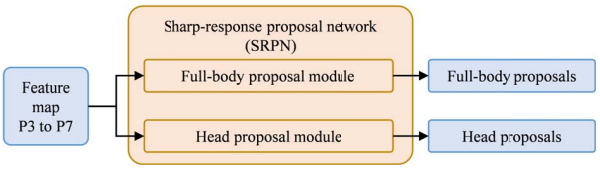

Figure 3. The architecture of SRPN, which mainly comprises two same proposal modules to produce full-body and head proposal respectively. The proposal module contains a 3×3 convolutional layer, followed by three prediction layers, one for center map, one for scale map and one for shift map. Then convert to RoIs results, after NMS processing output proposal.

this branch. In pedestrian detection task, line annotation is proposed by [13] and bounding box is automatically generated with a fixed aspect ratio. Therefore, we only predict the height of each object and generate the bounding box with fixed aspect ratio. The scale map is formulated as:

$$GT_{scale}(i,j) = \begin{cases} \log(h_k) & |i - x_k| < r, |j - y_k| < r, k \in K \\ 0 & otherwise \end{cases} \tag{6}$$

where h_k is the height of k-th object, r is the radius of positive sample area to reduce the ambiguity and empirically $r = 2$. The scale loss can be formulated as:

$$L^l_{scale} = 1/N \sum_{i=1}^{W/2^l} \sum_{j=1}^{H/2^l} [GT_{scale} > 0] SmoothL1(\hat{h}_{ij}, GT_{scale}(i,j)) \tag{7}$$

where $\hat{h}_{ij}$ represents the network's scale prediction, and N is the normalization term. The Iverson bracket indicator function $[GT_{scale} > 0]$ evaluates to 1 when $GT_{scale} > 0$ and 0 otherwise. For shift map, the smooth L1 loss is also applied to refine the center position. The target shift is defined as:

$$t_{ij} = \begin{cases} t_i = x_k/2^l - \lfloor x_k/2^l \rfloor \\ t_j = y_k/2^l - \lfloor y_k/2^l \rfloor \end{cases} \tag{8}$$

Then, the shift loss can be formulated as:

$$L^l_{shift} = 1/K \sum_{i=1}^{W/2^l} \sum_{j=1}^{H/2^l} SmoothL1(\hat{t}_{ij}, t_{ij}) \tag{9}$$

where $\hat{t}_{ij}$ is the predicted shift of the feature center. Therefore, at a pyramid level l, the full optimization objective function is:

$$L_l = \lambda_c L^l_{center} + \lambda_s L^l_{scale} + \lambda_t L^l_{shift} \tag{10}$$

where λ_c, λ_s, and λ_t are weights, which are experimentally set as 0.01, 1, and 0.1, respectively. The proposal branch losses of the SRPN consist of the pyramid level losses for P3 to P7. Then, the whole losses are:

$$L_{SRPN} = \sum_{l=3}^{7} L^l_{head} + L^l_{body} \tag{11}$$

where L^l_{head} and L^l_{body} are loss pyramid level l of head and full body proposal branch. For head proposal, the position and scale of the center are strongly related to those of full body in general pedestrian detection dataset. Based on statistical observations, the position of the scale of pedestrian head is about 20% of the full body height. Therefore, the head annotation is computed as:

$$(x_h, y_h, s_h) = \begin{cases} x_h = x_f \\ y_h = y_f - 0.4h_f \\ s_h = 0.2h_f \end{cases} \tag{12}$$

where (x_f, y_f) and h_f are full body box center and scale. The aspect ratio of head box is simply set to 1.0. For full body proposal, the aspect ratio is set to 0.41 as suggested in [13].

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

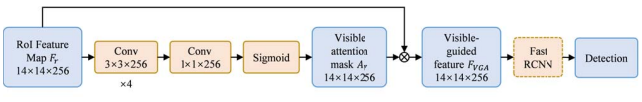

Figure 4. The architecture of VGA in Fast R-CNN network. Visible attention mask is generated from RoI feature with four 3×3 convolutional layer and one 1×1 convolutional layer. The sigmoid layer normalizes the response. The original RoI feature is modulated by the visible attention mask, before input Fast R-CNN.

B. Visible-guided attention module

The proposed VGA module is depicted in Fig. 4. After RoI-pooling, a visible-guided attention (VGA) module is newly designed to leverage the head and visible area information and suppress the feature noise of occluded area in order to enhance the feature representation learning of the backbone network.

1) VGA Architecture. The full body feature extracted by RoI pooling layer is denoted as $F_r \in [C \times H_r \times W_r]$, where C is channel and H_r and W_r are the resolution. The F_r passes through four $C \times 3 \times 3$ convolution layers, one 1×1 convolution layer, and a pre-pixel sigmoid function, and then the visible attention map $A_V \in [1 \times H_r \times W_r]$ is obtained. The visible-guided feature F_{VGA} is achieved by taking the element-wise dot product of every feature channel in F_r with A_V as:

$$F_{VGA}(i,:,:) = F_r(i,:,:) \cdot A_V \tag{13}$$

where i is the channel index. This refined feature F_{VGA} can reduce confusion on the detection network, leading to a relatively high confidence for occluded proposals.

2) Ground truth and Loss. Training VGA requires segmentation annotation on the visible area. However, it requires very heavy labeling work and most datasets for pedestrian detection task do not provide such annotations. Therefore, we adapt visible-area bounding box annotation as an approximate segmentation, which are available for the popular pedestrian detection benchmarks. The ground truth for visible-guided attention is formulated as:

$$GT_{VGA}(x,y) = \begin{cases} 1 & (x,y) \in BB_{v_k} \\ 0 & otherwise \end{cases} \tag{14}$$

where BB_{v_k} is k-th visible-area bounding box. If a pixel lies in the visible region, the label is 1 or 0 otherwise. For VGA, the binary cross-entropy loss (BCE loss) is applied to train this branch, which is formulated as:

$$L_{VGA} = 1/H_r W_r \sum_{i=1}^{H_r} \sum_{j=1}^{W_r} BCELoss(GT(i,j), A_V(i,j)) \tag{15}$$

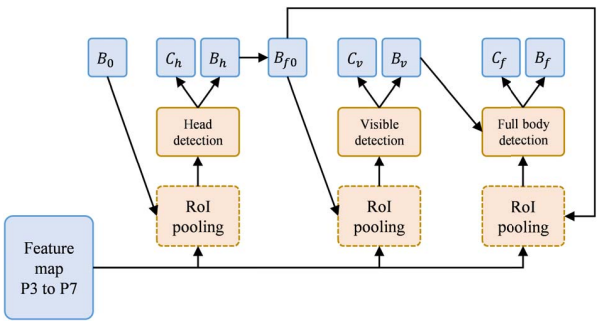

Figure 5. Head-cascade R-CNN progressively regresses the head proposal to refined full-body detection. Its advantage lies in the use of head cues while increasing credibility based on visible information.

C. Head-cascade RCNN network

Finally, a head-cascade RCNN (HRCNN) is proposed to regress the pedestrian bounding box from head proposal. The architecture of the original Cascade R-CNN [30] is a multi-stage extension of the Faster R-CNN [29], which is framed as cascaded regression problem. Cascaded regression is a resampling procedure that changes the distribution of hypotheses to be processed by different stages. Hence, Cascade R-CNN achieves more precise localization than Faster R-CNN. In human visual system, we often resort the head information to locate each pedestrian in occluded scenes. However, due to the small head size and insufficient information, the results of detection based on head underperform the center point [3]. Based on the above observations, HRCNN uses a method of progressively expanding the sampling area for regression of visible-region from head proposal, and regression of full-body from visible area.

1) Architecture. The architecture of the HRCNN is shown in Fig. 5. First, the bounding box for the proposal head (B_0) is applied on first stage to extract the head RoI feature, which is subsequently input to head detection (D_h) module. The D_h outputs the head classification score (C_h) and head bounding box (B_h). The second stage is the visible area detection (D_v) module, but obviously it is hard for regression of visible bounding box (B_v) on head information. Therefore, as in (12), B_h is converted to the full-body bounding box ($B_f 0$) as the proposal for second stage. The D_v predicts visible classification score (C_v) and visible bounding box (B_v). For third stage, the $B_f 0$ is input to the full body detection (D_f) module as the proposal, and the feature is filtered by B_v . Like VGA, D_f only use the feature response in the visible region to suppress noise feature from obstacle objects. Finally, the D_f outputs full body classification score (C_f) and bounding box (B_f). Each detection module has the same structure as original Cascade R-CNN [30].

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

2) Ground truth and loss. In most popular pedestrian detection datasets, the visible and full body annotations (G_v and G_f) are provided but no head annotations. Therefore, the head annotations (G_h) are automatically generated by (12). Each detector module includes a classifier and a regressor. The HRCNN are learned with the loss as follow:

$$L_{HRCNN} = L_h + L_v + L_f \tag{16}$$

$$L_i = L_{cls}(C_i, Y_i) + [Y_i > 0]L_{reg}(B_i, G_i),\ i \in h, v, f \tag{17}$$

where L_i is the loss of i-th stage, L_{cls} is classification cross entropy loss, L_{reg} is regression Smooth L1 loss, and Y_i is the label of proposal under IoU threshold $u_i = 0.7$, $[\cdot]$ is indicator function.

D. Implementation details

Multi-task loss function. The whole network is optimized by $L = L_{SRPN}+L_{VGA}+L_{FRCNN}+L_{HRCNN}$, the Fast RCNN detection loss (L_{FRCNN}) is identical to those defined in [38].

Optimization. The backbone network is initialized by the ImageNet [39] pre-trained ResNet-50 model. The parameter of newly added layers in the SRPN are initialized by the Xavier normal distribution method, and the parameters in the VGA and HRCNN are initialized by the MSRA normal distribution method. We fine-tune the module using Adam optimizer. The proposed network is trained on 4 RTX2080TI GPUs with a mini-batch 2 images per GPUs, the learning rate is 2×10^{-4} and training is stopped after 150 epochs. We also apply the mean-teacher strategy [40] of moving average weights to achieve more stable training.

IV. Experiments

A. Datasets and Evaluation Metric

The bulk of the experiments was performed on CityPersons [13], which is built upon the semantic segmentation dataset Cityscapes [41] to provide a new dataset of interest for pedestrian detection. It is recorded across 18 different cities in Germany with 3 different seasons and various weather conditions. The dataset includes 5,000 images (2,975 for training, 500 for validation, and 1,525 for testing) with about 35k manually annotated persons plus about 13K annotations of ignore region. Both the bounding boxes and visible parts of pedestrians are provided. The full body is annotated by drawing a line from the top of the head to the middle of two feet, the head center is at 10% of the height from the top. The visible bounding box for each instance is the tightest one fully covering the segment mask. There are approximately 7 pedestrians in average per image. The numbers of bare, partial and heavy occlusion are 929:1057:1171 in validation subset.

We follow the standard evaluation metric: log-average miss rate (MR), which is averaged over the false positives per image (FPPI) in [0.01, 1], and its lower value reflects better detection performance. For proposal network, the general metric is recall rate (Recall). The results are reported across three different occlusion degrees: Reasonable, Partial Occlusion, Heavy Occlusion. The visibility ratio in Reasonable set is larger than 65%, the visibility ratio in Partial Occlusion set ranges from 65% to 90%, and the visibility ratio in Heavy Occlusion set ranges from 20% to 65%. In all subsets the height of pedestrians over 50 pixels is taken for training and evaluation, as in [13].

Table I
Comparison with the state of the arts on CityPersons.

Method	Backbone	Scale	MR (%)		
			Reasonable	Partial	Heavy
FRCNN	VGG-16	1	15.4	-	-
FRCNN (reimplement)	ResNet-50	1.3	13.6	15.4	56.7
RepLoss	ResNet-50	1	13.2	16.8	56.9
AFRCNN	VGG-16	1	12.9	-	-
OR-CNN	VGG-16	1	12.8	15.3	55.7
TTL	ResNet-50	-	15.5	17.2	53.6
TTL+MRF	ResNet-50	-	14.4	15.9	52.0
ALFNet	ResNet-50	1	12.0	11.4	51.9
RepLoss	ResNet-50	1.3	11.6	-	55.3
MGAN	VGG-16	1	11.5	-	51.7
OR-CNN	VGG-16	1.3	11.0	13.7	51.3
CSP	ResNet-50	1	11.0	10.4	49.3
Human-cascade	ResNet-50	1.3	10.8	10.1	45.3

B. Benchmark

Table I shows the comparisons with previous state of the arts methods, namely FRCNN [29], RepLoss [12], AFRCNN [13], TLL+MRF [6], ALFNet [42], MGAN [16], OR-CNN [4], CSP [3] on CityPersons validation set. Following [13], we evaluate on the reasonable, partial occlusion and heavy occlusion subsets. We re-implement the FRCNN with ResNet-50 and FPN as baseline detector. Our approach achieves promising detection performance improvement of 11.4% in MR on heavy occlusion subset. Furthermore, compared the sate-of-the-art detector, our proposed method also achieves significant improvements on heavy occlusion subset: MR is reduced from 49.3% to 45.3%. On the reasonable and partial occlusion subsets, the proposed method also outperforms other detectors by achieving a MR of 10.8% and 10.1%, respectively. The experimental results demonstrate that the HRCNN can effectively handle both heavily occluded and normal pedestrians under diverse scenes.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

Table II
Comparison of two proposal networks.

Method	Recall (%)		
	Reasonable	Partial	Heavy
RPN	95.31	96.81	77.37
SRPN-body	97.97	98.40	82.30
SRPN-head	96.83	97.17	89.20
SRPN-both	98.29	98.40	90.53

Table III
Results of VGA on two proposal networks.

Proposal	VGA	MR (%)		
		Reasonable	Partial	Heavy
RPN		13.6	15.4	56.7
RPN	✓	11.9	12.1	52.7
SRPN		12.9	12.7	52.6
SRPN	✓	11.0	10.4	49.3

C. Ablation Studies

In this section, an ablative analysis of the proposed SRPN, VGA, and Head-cascade RCNN modules are conducted on the Citypersons dataset.

SRPN. To verify the effectiveness of the proposed SRPN, we evaluate different proposal methods for Faster R-CNN framework on Citypersons dataset. The results are illustrated in Table II. The baseline proposal network is RPN. The proposal results are filtered by NMS post-process at IoU threshold 0.5. The proposals are sorted by score, then the recall of the first 100 results are counted. As shown in Table II, due to the lack of human information, RPN has a low recall rate (77.37%) for heavily occluded pedestrians. The SRPN refines the feature responses in a narrow area to handle occluded pedestrians. The SRPN-body and SRPN-head only outputs the full-body and head bounding boxes, respectively. The SRPN-both consists of the SRPN-body and the full-body bounding boxes converted from the head proposals as (12). Comparing the results in Table II, we find that the SRPN-body and SRPN-head achieves a recall rate of 4.93% and 11.83% on the Heavy Occlusion set, respectively. The SRPN-both achieves better performance; it means that body and head proposals are complementary. Specifically, the SRPN-body is better than the SRPN-head on low occlusion set (Reasonable and Partial), and the SRPN-head is better than the SRPN-body on Heavy Occlusion set. The SRPN-body perceives the full body information which has more distinguishable feature but is easily occluded. The SRPN-head only gets the head information which is more visible but lacks information and too small to detect. Therefore, the SRPN-both is a better way for proposal of all occlusion level scenes.

Table IV
Comparison of three head-cascade strategies.

Cascade strategy	MR (%)		
	Reasonable	Partial	Heavy
Baseline (SRPN-body + VGA)	11.0	10.4	49.3
A $B_h \rightarrow B_{f0}$	12.4	11.6	47.8
B $B_h \rightarrow B_{f0} \rightarrow B_v \rightarrow B_f$	11.5	10.9	46.8
C $B_h \rightarrow B_{f0} \rightarrow B_v \rightarrow (B_{f0}) \rightarrow B_f$	10.8	10.1	45.3

VGA. The full-body proposals are input into Fast RCNN but the heavily occluded samples would confuse the network during training. The VGA module predicts a visible area to leverage visible area information and suppresses the feature noise of occluded area. Two baselines are selected to verify the effectiveness of our VGA module. Table III shows the baselines comparison. One baseline detector is Faster RCNN with RPN, respectively obtaining MR of 13.6%, 15.4%, and 56.7% on Reasonable, Partial, and Heavy occlusion sets of Cityperosns dataset, respectively. The VGA module combined with baseline detector effectively reduces the MR with 1.7%, 3.3%, and 4.0% on the Reasonable, Partial, and Heavy Occlusion sets, respectively. Another baseline detector is Faster RCNN with the SRPN-body proposal network. The results show that the MR improvement of VGA module is consistent on three subsets (1.9%, 2.3%, and 3.3%). Noteworthy, the reduction in MR on the Heavy Occlusion set demonstrates the effectiveness of the VGA for full-body proposals.

Head-cascade RCNN. The head-cascade module uses a resampling procedure to extend the receptive field under different cascade stages. To demonstrate the effectiveness of the proposed head-cascade strategy, three head cascade procedures are evaluated. The first strategy (A) directly predicts the head bounding box (B_h) in head stage, and B_h is converted to the full-body bounding box (B_{f0}) as the final results. The second strategy (B) adds a visible stage and full-body stage, which regresses the visible bounding box (B_v) and the full bounding box (B_f), respectively. Compared with the second strategy, the final strategy (C) does not regress directly from B_v to the full-body bounding box (B_f), but from B_{f0} with the feature filtered from the occluded area. As shown in Table IV, when the strategy A reduces the MR by 1.5% on Heavy Occlusion set but increases the MR on Reasonable and Partial Occlusion sets due to lacking information on the head proposal. The strategy B improves the MR by 2.5% on Heavy Occlusion set but is still worse than the baseline on other two sets. Since the difference between B_v and B_f is usually greater than B_{f0}, regression from B_{f0} to B_f has smaller deviation. The strategy C achieves the best performance which reduces the MR by 0.2%, 0.3%, and 4.0% on

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

three subsets, respectively. The results verify that the proposed head-cascade module can effectively reduce the MR of pedestrian detection on Heavy Occlusion without increasing the false alarm of pedestrians on low occlusion scenes.

V. Conclusion

Effective detection in heavily occluded pedestrians is challenging and demanding in many applications such as ADAS because misdetection of such pedestrians usually leads to severe accidents or even death. In this paper, a new Human-cascaded R-CNN was proposed (SRPN + VGA + head-cascade) for occluded pedestrian detection by paying more attention on visible body part. The HRCNN is robust in diverse occlusion scenes including crowd and obstacle occlusions. It uses head cues while increasing credibility based on visible information. Experimental results on Citypersons show better performance of the proposed method in reasonable, partial and heavy occlusion subsets. Compared with the baseline detector FRCNN, our human-cascade obtains an absolute gain of 11.4% and 2.8% in MR on heavy occlusion and reasonable subsets. In ablation studies, results also show that the proposed SRPN can improve pedestrian recall rate by detecting full body and head proposal at the same time. Moreover, the proposed VGA can significantly reduce the MR of heavily occluded pedestrians on Faster R-CNN framework, no matter the proposal network is RPN or SRPN. Finally, the head-cascade module can effectively reduce the MR of pedestrians on heavy occlusion without increasing the false alarm of pedestrians on low occlusion scenes.

References

[1] G. Duan, H. Ai, and S. Lao, "A structural filter approach to human detection," in ECCV. Springer, 2010, pp. 238–251.

[2] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "PedHunter: Occlusion Robust Pedestrian Detector in Crowded Scenes," in AAAI, sep 2020. [Online]. Available: http://arxiv.org/abs/1909.06826 https://github.com/ChiCheng123/PedHunter

[3] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level Semantic Feature Detection:A New Perspective for Pedestrian Detection," in CVPR, apr 2019. [Online]. Available: http://arxiv.org/abs/1904.02948

[4] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd," in ECCV, jul 2018.

[5] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," Pattern Recognition, vol. 85, pp. 161–171, jan 2019.

[6] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-Scale Pedestrian Detection Based on Topological Line Localization and Temporal Feature Aggregation," in ECCV, 2018, pp. 554–569.

[7] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-Aware Deep Feature Learning for Pedestrian Detection," in ECCV, 2018, pp. 745–761.

[8] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining Pedestrian Detection in a Crowd," in CVPR, apr 2019.

[9] F. Wen, Z. Lin, Z. Yang, and W. Liu, "Single-Stage Detector with Semantic Attention for Occluded Pedestrian Detection," in International Conference on Multimedia Modeling. Springer, Cham, jan 2019, pp. 414–425.

[10] J. Zhang, L. Lin, Y.-c. Chen, Y. Hu, S. C. H. Hoi, and J. Zhu, "CSID: Center, Scale, Identity and Density-aware Pedestrian Detection in a Crowd," arXiv, 2019.

[11] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and Context Information for Pedestrian Detection with CNNs," arXiv, apr 2018.

[12] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion Loss: Detecting Pedestrians in a Crowd," in CVPR, nov 2017, pp. 7774–7783.

[13] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," in CVPR, vol. 2017-Janua, 2017, pp. 4457–4465.

[14] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational Learning for Joint Head and Human Detection," in AAAI, sep 2020.

[15] C. Zhou, M. Yang, and J. Yuan, "Discriminative Feature Transformation for Occluded Pedestrian Detection," in ICCV, vol. 1, 2019, pp. 9557–9566.

[16] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-Guided Attention Network for Occluded Pedestrian Detection," in ICCV, 2019.

[17] M. Jeong, B. C. Ko, and J.-Y. Nam, "Early Detection of Sudden Pedestrian Crossing for Safe Driving During Summer Nights," IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 6, pp. 1368–1380, jun 2017.

[18] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How Far are We from Solving Pedestrian Detection?" in CVPR. IEEE, jun 2016, pp. 1259–1267.

[19] L. Zhang, L. Lin, X. Liang, and K. He, "Is Faster R-CNN Doing Well for Pedestrian Detection?" in ECCV. Springer, 2016, pp. 443–457.

[20] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in ICCV, 2013, pp. 1505–1512.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.

[21] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in CVPR. IEEE, 2012, pp. 3258–3265.

[22] ——, "Joint Deep Learning for Pedestrian Detection," in ICCV. IEEE, dec 2013, pp. 2056–2063.

[23] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," in ICCV, 2016, pp. 1904–1912.

[24] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in ICCV. IEEE, 2003, p. 734.

[25] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE TPAMI, vol. 36, no. 8, pp. 1532–1545, 2014.

[26] W. Nam, P. Dollar, and J. H. Han, "Local decorrelation for improved pedestrian detection," in NISP, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 424–432.

[27] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in BMVC, 2009, pp. 91.1–91.11.

[28] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in ECCV. Springer, 2016, pp. 354–370.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in NIPS, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99.

[30] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in CVPR, 2018.

[31] C. Zhou and J. Yuan, "Multi-Label Learning of Part Detectors for Heavily Occluded Pedestrian Detection," in ICCV, 2017.

[32] S. Zhang, J. Yang, and B. Schiele, "Occluded Pedestrian Detection Through Guided Attention in CNNs," in CVPR. IEEE, jun 2018, pp. 6995–7003.

[33] C. Zhou and J. Yuan, "Bi-box Regression for Pedestrian Detection and Occlusion Estimation," in ECCV, vol. 11205 LNCS, 2018, pp. 138–154.

[34] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in CVPR, 2017.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[36] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," IEEE TPAMI, vol. 32, no. 9, pp. 1–20, 2009.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, P. Dollar, P. Dollár, P. Dollar, P. Dollár, P. Dollar, and P. Dollár, "Focal Loss for Dense Object Detection," in ICCV, aug 2017.

[38] R. Girshick, "Fast R-CNN," in ICCV, 2015, pp. 1440–1448.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, sep 2014.

[40] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in NIPS, vol. 2017-Decem, mar 2017, pp. 1196–1205.

[41] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in CVPR, 2016, pp. 3213–3223.

[42] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in ECCV, 2018.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on March 16,2023 at 10:40:55 UTC from IEEE Xplore. Restrictions apply.