SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160102:1–160102:16
https://doi.org/10.1007/s11432-021-3300-2

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

HAPNet: a head-aware pedestrian detection network associated with the affinity field

Jiali DING[1], Tie LIU[1*], Yun ZHAO[2], Zejian YUAN[2*] & Yuanyuan SHANG[1]

[1]*College of Information Engineering, Capital Normal University, Beijing* 100048*, China;*
[2]*College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an* 710049*, China*

Received 25 March 2021/Revised 20 May 2021/Accepted 30 June 2021/Published online 7 April 2022

Abstract Pedestrian detection has made great progress with the rapid development of deep learning, but pedestrian detection under occlusion remains a challenge. To solve the occlusion problem, some research endeavors have been carried out based on visible parts, such as the body, but there is still substantial room for improvement due to the introduction of additional calculation. Aiming to improve the detection performance under occlusion, this paper proposes a head-aware pedestrian detection network (HAPNet) by using the inherent structural relationship between the human body and head. The postprocessing stage is redesigned to include a scoring module and an augmented non-maximum suppression (NMS) algorithm. Specifically, HAPNet detects head and body simultaneously through different layers of a feature map. We then propose a head-side affinity model, which can represent the association between the head and body sides. Detection and affinity prediction tasks are implemented through different branches of HAPNet. In the scoring module, the head and body detection scores are fused to match the head-body pairs to improve the detection performance. On this basis, an enhanced NMS algorithm is proposed, which achieves a good balance between reducing false positives and missing detection. The experimental results verify the effectiveness of this method in pedestrian detection under occlusion.

Keywords pedestrian detection, head detection, head-aware pedestrian network, affinity module, occlusion

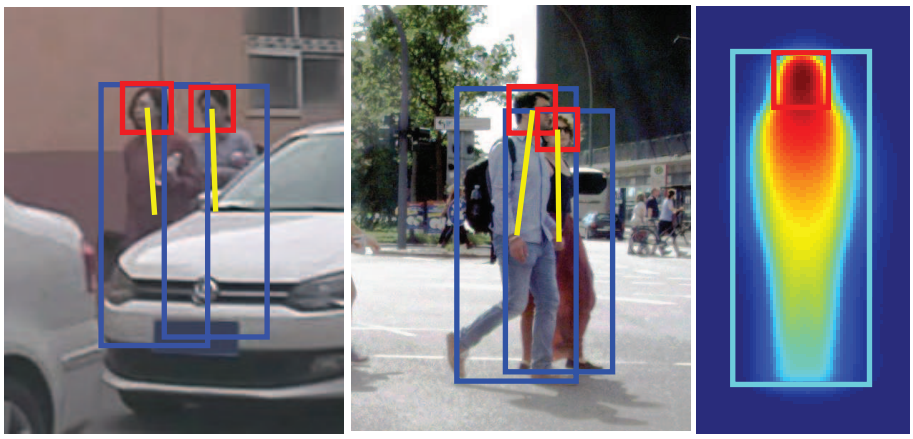
Citation Ding J L, Liu T, Zhao Y, et al. HAPNet: a head-aware pedestrian detection network associated with the affinity field. Sci China Inf Sci, 2022, 65(6): 160102, https://doi.org/10.1007/s11432-021-3300-2

1 Introduction

Pedestrian detection has attracted wide attention due to its applications in video surveillance and automatic driving. With the introduction of the convolutional neural network (CNN) [1–6], the performance of pedestrian detection has substantially improvements. However, occlusion remains a significant challenge. Generally, occlusion can be divided into two types: interclass occlusion and intraclass occlusion [6]. In terms of pedestrian detection, the former occurs when pedestrians are occluded by vehicles, buildings, trees or other non-human categories, as shown in Figure 1(a), and the latter occurs when pedestrians are occluded by each other, as shown in Figure 1(b). Both types of occlusion cause varying degrees of overlap and increase the missing rate of detection. Moreover, occlusion occurs on different parts of the body, complicating the visible region-based approach.

Some efforts have been devoted to part-based methods, including methods based on manually selecting parts [1, 7]. However, these methods are suboptimal because the computational cost grows linearly with increasing number of parts. Additionally, multiple parts may complicate the postprocessing. To simplify the postprocessing, a two-branch structure [2] with the full body and a visible part was introduced. Nevertheless, the visible region part is not reliable and robust enough for its various location configurations. To overcome the above shortcomings, this paper introduces a crucial part, the head, to help detect individuals in occlusion cases. As shown in Figure 1(c), compared with other parts or the whole body, the head part is hardly occluded by nearby pedestrians or objects from other categories.

* Corresponding author (email: liutiel@163.com, yuan.ze.jian@xjtu.edu.cn)



(a) (b) (c)

Figure 1 (Color online) (a) Pedestrians occluded by vehicles (interclass occlusion). (b) Pedestrians that overlap with each other (intraclass occlusion). (c) Heat map produced by averaging the visible masks of occluded pedestrians (red color indicates the highest score). The yellow line represents the assignment of the head and its corresponding full body.

In this paper, we propose a head-aware pedestrian detection network (HAPNet) for occluded pedestrian detection. First, HAPNet is divided into three branches after the backbone, and the top two branches are used for head and pedestrian detection tasks, as shown in Figure 2. These two branches exploit different feature maps with different scales because the shallow layer contains a wealth of detailed information and semantic information is stored in high-level features. Then, a head-side affinity module is introduced to assign the head and full body detections, which represents the association of the head and body sides.

The body side is the general definition for the direction of one side of the upper body, which is broad sensed. In this paper, we select the trisection points instead of the middle points, while taking the center of the head as the center of the circle. This indicates that HAPNet is more discriminative and reliable than the detections of human-designed parts or multiple/single visible regions. In short, the affinity module is a 2D vector field encoding the localizations and orientations from head to body sides, which is built based on the observation that there is a sharp edge along the line segment from the head center to the body side.

In the postprocessing, we propose a scoring module that assigns each head detection to its corresponding pedestrian detection by choosing the maximal integrated association score and fuses the scores of the head and pedestrian detections. The scoring module is well-adapted to the proposed head-aware pedestrian detection method. Inspired by the representative region non-maximum suppression (NMS) [8], we propose an augmented NMS to deal with the heavily occluded cases, which considers the overlap of both the head and full body bounding boxes simultaneously. Specifically, we introduce two thresholds and perform NMS on the head and full body detections with the two thresholds. The experiments indicate that the head part can offer a more accurate guide for distinguishing individuals with the proposed augmented-NMS algorithm.

To further evaluate the performance of the pedestrian detector under occlusion conditions, this paper builds a dataset containing scenes of crowds of pedestrians (CrowdHuman-Ped) and the original data belongs to the CrowdHuman dataset [9]. To select the pedestrian instances from the CrowdHuman dataset, two criteria are introduced, which refer to the width-height ratio of the full body box and the area ratio between the head and full body boxes in the CityPersons dataset.

The major contributions of this paper include three aspects. First, we propose HAPNet to deal with the occluded pedestrian detection task. Specifically, we introduce the head to form a head-body pair and associate it with the proposed head-side affinity module. Second, we propose to modify the postprocessing stage, including designing a re-scoring module and formulating an augmented NMS criterion. The proposed method greatly matches the head and pedestrian detections and achieves an effective trade-off between precision and recall. Third, we collect a crowded pedestrian dataset (CrowdHuman-Ped), established on top of CrowdHuman [9], to evaluate the pedestrian detectors under different occlusion cases. We achieve considerable performance on CrowdHuman-Ped and CityPersons with the proposed detector, especially in occluded cases.

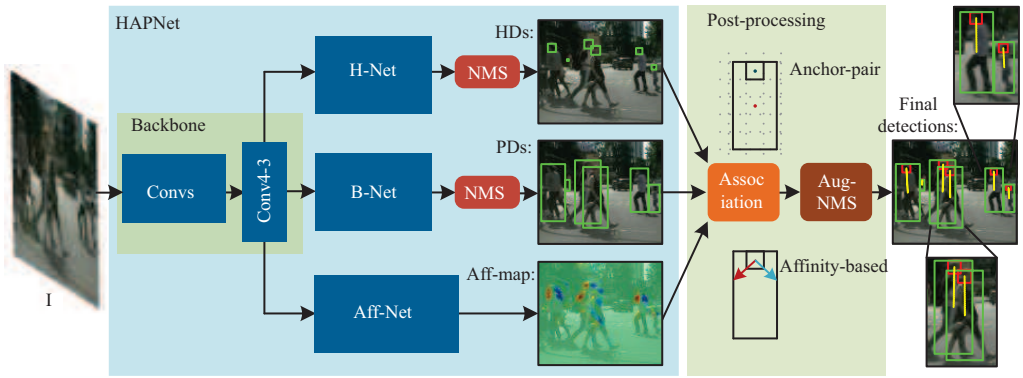


Figure 2 (Color online) Overview of the proposed HAPNet. H-Net denotes the head detection network, and B-Net is the body detection network. The third branch, Aff-Net, is the affinity module. In the final detections, the red box indicates the head detection, the green box denotes the body detection, and the yellow line is the head-body assignment.

2 Related work

2.1 Pedestrian detection

Pedestrian detection has achieved considerable improvement through both traditional methods [10–13] and deep learning-based methods [3–5, 14–17] over the last decade. With the revival of convolutional neural networks, some traditional methods relying on manually selected features [10], classifiers and decision trees [11] are gradually losing their application value, whereas deep learning-based methods [18, 19] yield unusually brilliant results. General object detectors are commonly divided into two types: one-stage or two-stage. They perform well for pedestrian detection. With the success of Faster R-CNN [14], a typical two-stage object detection method, a wave of pedestrian detection approaches [3,4,7,17,20,21] are proposed following the custom architecture. For instance, Zhang et al. [17] properly adapted the model of Faster R-CNN and gain a significant improvement in detection performance. Mao et al. [4] presented a new network architecture that aggregates extra features into the detection framework and jointly learns pedestrian detection as well as the given extra features.

To accelerate the speed of pedestrian detection, a series of one-stage CNN-based pedestrian detectors [22–25] are proposed. Noh et al. [22] implemented pedestrian detection on state-of-the-art one-stage models, e.g., YOLO9000 [26], SSD [27], and DSSD [28]. Liu et al. [23] proposed an effective ALF module that stacks multiple predictors to gradually evolve the detection from SSD to improve the localization performance in pedestrian detection. Song et al. [29] proposed a progressive refinement network that contains three sequential refinement phases; the network uses confidence-aware calibration for adaptive initialization of anchors. Considering the efficiency of the one-stage pedestrian detector, the proposed detector similarly employs the one-stage structure to perform the pedestrian and head detection.

Despite the fact that the single-stage method is more efficient, the results for small-scale pedestrians still trouble many researchers. Small-scale pedestrians are more blurred and noisier, and many attempts have been made to overcome these challenges. For instance, Cao et al. [30] solved the problem of small-scale pedestrian detection and occlusion pedestrian detection at the same time by focusing on the predictive bounding box with low positioning accuracy and extracting more contextual information around the object. Inspired by FPN [31], employing different scales has become popular. For example, Li et al. [32] proposed to detect small-scale pedestrians and large-scale pedestrians simultaneously. In addition, Li et al. [33] proposed a box guided convolution (BGC) that can dynamically adjust the sizes of convolution kernels guided by the predicted bounding boxes. In this way, it can better address the challenge of large variations of scale.

2.2 Occlusion pedestrian detection

Occlusion often accompanies the task of pedestrian detection. In term of handling occlusions for pedestrian detection, many approaches have been proposed recently, which can be mainly divided into three categories: exploiting the attention mechanism [20, 34], feature transformation [35] and part-based detection [2, 5, 7, 12, 22, 36]. Several methods describe the pedestrian using a part-based model to handle occlusion. Zhou et al. [2] addressed the occlusion by regressing two bounding boxes to localize the full

body and the visible part of the pedestrian simultaneously. Xie et al. [5] proposed PSC-Net, which is designed to explicitly capture co-occurrence information of different pedestrian body parts through a graph convolutional network (GCN). However, too many designed parts or visible regions still bother the researchers. Differently, we employ the head as partial information for occluded pedestrian detection. Detecting a pedestrian between the head and the body is reliable because of the inherent structure of the head-body pair.

Several studies [9, 34, 37–39] are focusing on pedestrian head detection. Shao et al. [9] used FPN [31] and RetinaNet [40] to perform head detection. Chi et al. [37] designed a mask-guided module to leverage the head information to enhance the feature representation learning of the backbone network. Chen et al. [39] presented a novel approach that learns a semantic connection between the pedestrian head and other body parts. Instead of simply using the information of the head, we model the association between the head and body detection using a newly defined affinity field, which makes good use of the inherent human body structure.

Additionally, some studies are being conducted on designing overlap-aware loss [6, 7] or modifying the NMS criterion [41, 42] to minimize the effect of occlusion. Wang et al. [6] proposed a repulsion loss for bounding box regression that attaches each proposal to its designated target and pushes it away from other close-by ground truth boxes and proposals. However, the additional loss terms disturb the localization of the pedestrian to some extent. In non-maximum suppression, it is hard to choose a perfect intersection over union (IoU) threshold to delete the close-by detections correctly [43]. To handle this problem, Liu et al. [41] proposed novel adaptive-NMS, which applies a dynamic suppression strategy and designs an auxiliary and learnable subnetwork to predict the adaptive NMS threshold. In addition, to reduce the missing rate in the crowded cases, Soft-NMS [44] would offer a little value instead of the zero in the traditional NMS algorithm. To adapt to our method, we propose a two-thresholds-augmented NMS algorithm, different with the traditional NMS. General methods implicitly model the relationship of close-by detections; differently, we introduce the head detection and build an explicit criterion for NMS.

3 Proposed method

3.1 Overview

The overview of the framework is shown in Figure 2. By observing its structure, it can be found that the framework is mainly divided into two parts, HAPNet and the postprocessing stage. HAPNet consists of three separate branches with a shared backbone, and the first two branches are used for head and full body detections with a single-stage detection scheme. Moreover, the head and pedestrian detections are performed based on the feature maps from different layers (Conv4-3, Conv5-3, Conv6-1, and Pool6; similar structure in SSD [27], MSCNN [3]). As for the third branch, the head-side affinity module is employed to predict the head-side affinity field. Furthermore, we design a new associated module to match the head-body pair and propose an augmented NMS algorithm to adapt HAPNet.

3.2 Head and full-body detection

It is a great challenge to select the visible region part for the part-based occluded pedestrian detection methods. Different from other body parts, the head is more distinct, and its authenticity was verified on CityPersons [17] a long time ago. We consider exploiting the head to assist the body in positioning pedestrians. In this paper, VGG16 [45] is used as the backbone. Specifically, the structure of the first 5 layers remains unchanged, but a $3\times3\times512$ convolution layer (Conv6-1) and a followed max-pooling layer (Pool6) are added after Conv4-3. The overview of the detection framework is depicted in Figure 3(a). For the sake of readability, this paper only shows the branches on the feature maps from one layer (Conv4-3) in Figure 3(a). The head branch and the body branch are built separately using a set of convolutional filters with the same structure pipeline. In the both branches, the number of channels is 512, and different numbers after the channels denote the width and height of different rectangular convolution kernels, which can deal with the situation that pedestrian detecting and head detecting boxes are mostly rectangular boxes. To choose an appropriate anchor box size for head detection, the proposed detector builds an anchor pool and selects the one with the maximum IoU between anchor boxes and ground truths.

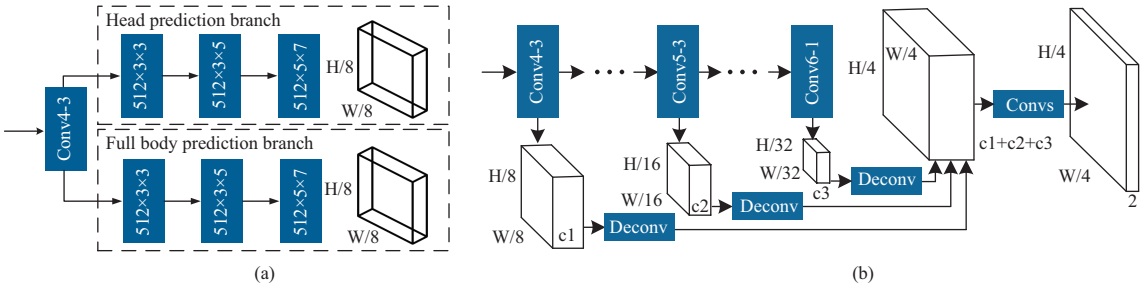


Figure 3 (Color online) (a) Structure of the head and body detections; (b) structure of the head-side affinity module.

3.3 Head-side affinity module

To better assign the head to the corresponding individual, this paper design a head-side affinity module to represent the association of the head and the full body. Inspired by the part affinity field [46], we visualize the orientation of association between the limbs of the upper body. The part affinity field realizes the association between limbs by encoding the position and direction of the limbs on each pixel between the nodes. As shown in Figure 4(a), there is consistency of orientation in the upper body. The blue area shows the right orientation of the upper body, while the left orientation is shown in red. Based on this appearance cue, the proposed detector models the association of the head and corresponding body by the proposed head-side affinity module. Moreover, the head-side affinity module shows robustness against occlusion. As shown in Figure 4(d), when one pedestrian is occluded by the right person, there is sufficient edge information in the region of the left side. This ensures the robust assignment of the head and the corresponding human body.

We first define the module as follows. Assume that the center point of the head and the right body side point are defined as $\boldsymbol{x}_h^k$ and $\boldsymbol{x}_b^k$, where k indicates the k-th pedestrian. The two points are shown in Figure 4(b). Here, we select the trisection points instead of the middle points of the body sides because the mean orientation of the left/right upper body points is closer to the orientation from the head center to the trisection points of the left/right body sides. Then, a unit vector $\boldsymbol{v}_r^k$ is introduced to specify the direction from $\boldsymbol{x}_h^k$ to $\boldsymbol{x}_b^k$ as follows:

$$\boldsymbol{v}_r^k = \frac{\boldsymbol{x}_b^k - \boldsymbol{x}_h^k}{||\boldsymbol{x}_b^k - \boldsymbol{x}_h^k||}. \tag{1}$$

A point $\boldsymbol{p}$ located in the dotted box in Figure 4(b) needs to meet the following requirements:

$$0 \leqslant \boldsymbol{v}_r^k \cdot (\boldsymbol{p} - \boldsymbol{x}_h^k) \leqslant \rho_r^k, \quad |\boldsymbol{v}_{r\perp}^k \cdot (\boldsymbol{p} - \boldsymbol{x}_h^k)| \leqslant \sigma_r^k, \tag{2}$$

where $\boldsymbol{v}_{r\perp}^k$ is a unit vector perpendicular to $\boldsymbol{v}_r^k$. ρ_r^k is the length from the head to the right-side body, and σ_r^k is half the width of the dotted box. An illustration is shown in Figure 4(b). Finally, the head-side affinity field for the right side of the k-th person $\boldsymbol{A}_r^k(\boldsymbol{p})$ can be defined as

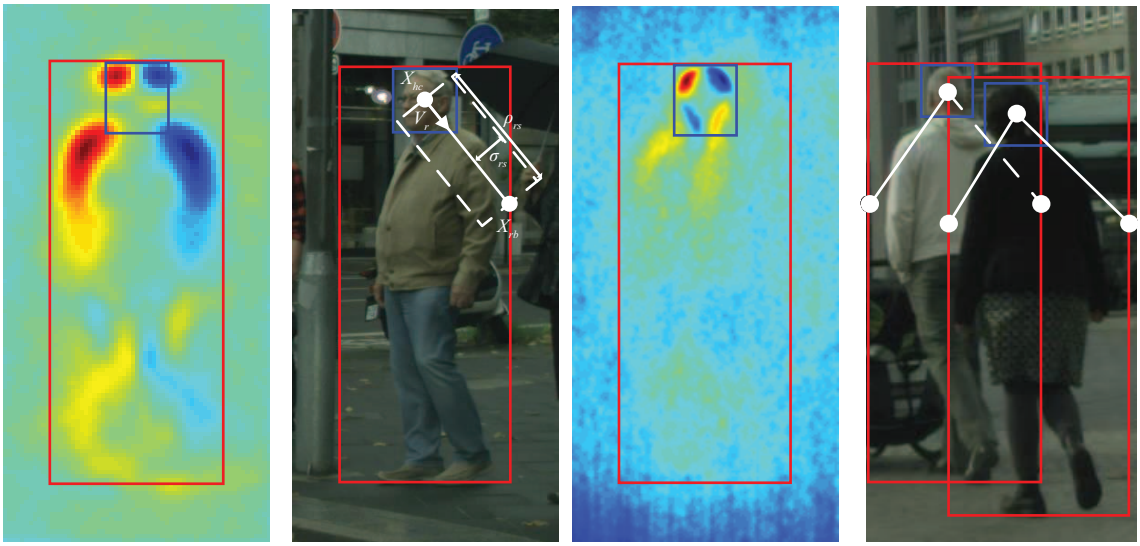
$$\boldsymbol{A}_r^k(\boldsymbol{p}) = \begin{cases} \boldsymbol{v}_r^k, & \text{if } \boldsymbol{p} \text{ on the right head-body side,} \\ \boldsymbol{0}, & \text{otherwise.} \end{cases} \tag{3}$$

During training, the ground truth head-side affinity field at each point $\boldsymbol{A}^*(\boldsymbol{p})$ can be calculated by averaging both orientations of all pedestrians as follows:

$$\boldsymbol{A}^*(\boldsymbol{p}) = \frac{1}{n(\boldsymbol{p})} \sum_{o \in \{r,l\}} \sum_k \boldsymbol{A}_o^k(\boldsymbol{p}), \tag{4}$$

where $n(\boldsymbol{p}) = \max(1, n_r(\boldsymbol{p}) + n_l(\boldsymbol{p}))$ is a non-zero value representing the coverage times of point $\boldsymbol{p}$ by both the right and left head-side of all the pedestrians.

In the overview of the affinity module, as depicted in Figure 3(b), multiple deconvolution layers are attached to different convolution layers to increase the resolution of the feature maps. Moreover, to capture the different visual characteristics of the pedestrian instances from different scales, these multiple feature maps from different layers are concatenated, and a 1×1×512 convolution layer is employed to reduce the channels of the concatenated feature map.



(a) (b) (c) (d)

Figure 4 (Color online) Statistical heat maps of pedestrians. The blue box indicates the bounding box of the head, and the red box denotes the whole body bounding box. (a) The average orientation of the whole body. Red: left orientation. Blue: right orientation. (b) Instance of the definition of head-side affinity. (c) Average orientation of individuals overlapped by pedestrians on the right side. (d) Corresponding connections of overlapped pedestrians.

3.4 Loss function

HAPNet is jointly trained using a multitask loss; the first part is detection, and the definitions are as follows. Assume that a training sample is denoted by (X, Y), and the bounding box coordinates are known. Here, X is the image patch and $Y = (c^*, \boldsymbol{b}^*)$ denotes its class label: 0 for background and 1 for object (head or body). The objective loss function is defined as a weighted sum of the classification loss l_{cls} and the localization loss l_{loc} as in [14, 47]. The objective loss for training includes the head and the body; however, in the following formula, we do not separate them.

$$l = l_{\text{cls}}(\hat{s}, c^*) + \lambda_h [c^* = 1] l_{\text{loc}}(\hat{\boldsymbol{b}}, \boldsymbol{b}^*, \boldsymbol{b}_a), \tag{5}$$

where $\hat{s}$ is the predicted score, and $\hat{\boldsymbol{b}}$, $\boldsymbol{b}^*$, and $\boldsymbol{b}_a$ indicate the bounding boxes of the prediction, ground truth and corresponding anchor, respectively. They are all made up of four parameters, denoted as $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{h})$. The localization loss is defined using a smooth L1 loss [47] between the predicted box $\hat{\boldsymbol{b}}$ and the ground truth box $\boldsymbol{b}^*$. Specifically, l_{loc} is performed on the offset vector $\Delta = (\delta_x, \delta_y, \delta_w, \delta_h)$ based on the anchor box $\boldsymbol{b}_a$. The offset vector between ground truth $\boldsymbol{b}^*$ and anchor $\boldsymbol{b}_a$ is computed as follows:

$$\begin{aligned} \delta_x = (x^* - x_a)/w_a, \quad \delta_y = (y^* - y_a)/h_a, \\ \delta_w = \log(w^*/w_a), \quad \delta_h = \log(h^*/h_a). \end{aligned} \tag{6}$$

In practice, Δ is normalized by its mean and variance to improve the effectiveness of localization [3,14,48].

As for the second part, the affinity module, the loss is calculated as follows. Given an image $I \in R^{W \times H \times 3}$ including multiple pedestrians, the head-side affinity map is defined as $\hat{\boldsymbol{A}} \in \mathbb{R}^{W_a \times H_a \times 2}$, where W_a and H_a are the width and height of the affinity map, with a ratio identical to that of the original image size. During training, the mean squared error (MSE) is employed to measure the loss between the predicted association map $\hat{\boldsymbol{A}}$ and the ground truth $\boldsymbol{A}^*$ over each position $\boldsymbol{p}$ as

$$l_{\text{aff}} = \frac{1}{N(\boldsymbol{p})} \cdot \sum_{\boldsymbol{p}} \boldsymbol{M}(\boldsymbol{p}) \cdot ||\hat{\boldsymbol{A}}(\boldsymbol{p}) - \boldsymbol{A}^*(\boldsymbol{p})||_2^2, \tag{7}$$

where $N(\boldsymbol{p})$ is the number of points in affinity maps. $\boldsymbol{M} \in \mathbb{R}^{W_a \times H_a}$ denotes the mask with the same spatial size as the head-side affinity map, in which a point covered by any ignored pedestrians or pedestrians without annotated heads is set to 0 and a point is otherwise set to 1.

The training loss is described as

$$l = l_b + \alpha_h \cdot l_h + \alpha_{\text{aff}} \cdot l_{\text{aff}}. \tag{8}$$

The first two components l_b and l_h represent the loss of full body and head detections, respectively, as described in (5), and the third term l_{aff} denotes the loss of the head-side affinity module. In the following experiments, α_h is set to 1 and α_{aff} is set to 0.1. During the training process, the proposed detector employs a multi-step training strategy. In the first step, each branch is trained separately and simultaneously. The shared networks in the backbone are updated along with each task. In the second step, all layers are jointly optimized using the final loss in (8).

3.5 Postprocessing

3.5.1 *Anchor-based assignment*

A straight approach to assign the head and pedestrian detections is associating them based on an anchor pair, as shown in Figure 2. Both the head and full body detections can be traced back to an anchor centered at a specified feature pixel. Considering the statistical localization of the head and its corresponding full body, this approach defines an anchor pair whose relative position relation is fixed as illustrated in Figure 2. A head detection and a full body detection are aligned if their original anchors are matched as an anchor pair. This kind of assignment method is simple and efficient. It is performed on the initialized proposals, so it can guarantee that there is an assigned head proposal for each full body detection.

3.5.2 *Head-side affinity module-based assignment*

Despite the efficiency, the anchor-based assignment approach yields numerous incorrect alignments when the pedestrians overlap with each other. To address this shortcoming, the proposed detector refines the assignment of the head box and its corresponding full body detection based on the head-side affinity module. For the j-th pedestrian candidate, this approach first selects a set of overlapped head candidates and builds a head candidate pool $\boldsymbol{P}_h^j$. Then, the association score $\hat{s}^{i,j}$ between j and a head candidate $i \in \boldsymbol{P}_h^j$ is calculated via integration over the predicted head-side field map $\hat{\boldsymbol{A}}$ along the line segment from the head center to the side points:

$$\hat{s}^{i,j} = \sum_{o\in\{r,l\}} \int_{u=0}^{u=1} \hat{\boldsymbol{A}}(\boldsymbol{p}_o(u)) \cdot \frac{\hat{\boldsymbol{x}}_o - \hat{\boldsymbol{x}}_h}{||\hat{\boldsymbol{x}}_o - \hat{\boldsymbol{x}}_h||_2} \mathrm{d}u, \tag{9}$$

where $\hat{\boldsymbol{x}}_h$ and $\hat{\boldsymbol{x}}_o$ represent the prediction localization of head center and body side (right or left), respectively, and $\boldsymbol{p}_o(u)$ indicates the interpolations along the line segments. Then, the $\tilde{i}$-th head with the maximum combined score is selected as the registered head detection for the j-th pedestrian detection as follows:

$$\tilde{i} = \arg\max_{i\in\boldsymbol{P}_h^j} (\hat{s}_h^i + \hat{s}^{i,j}), \tag{10}$$

where $\hat{s}_h^i$ is the detection score of head candidate i. Finally, the prediction boxes of the j-th pedestrian are banded as $(\boldsymbol{b}_b, \boldsymbol{b}_h)_j$.

While the head and full body detections are aligned, we propose to fuse their detection scores to improve the performance of the pedestrian detection. Let $s_b = (s_b^0, s_b^1)$ and $s_h = (s_h^0, s_h^1)$ specify the raw outputs of the detections from PDNets and HDNets without softmax, where the superscript indexes the background (0) and object (1). The final pedestrian detection score $\tilde{s}$ is re-scored by fusing these raw scores with a softmax operation as [49]

$$\tilde{s} = \frac{\exp(s_b^1 + \beta_h \cdot s_h^1)}{\exp(s_b^1 + \beta_h \cdot s_h^1) + \exp(s_b^0 + \beta_h \cdot s_h^0)}, \tag{11}$$

where β_h is a hyperparameter that controls the balance between the body and head detections. In this way, the proposed detector can increase the detection score of occluded pedestrians and improve the detection robustness.

3.5.3 *Augmented NMS algorithm*

To further explore the role of the head for NMS, this subsection investigates the occlusion cases of the head and full body with the ground truths of standing humans (w/h<0.5) in CrowdHuman [9]. First, the approach performs NMS with different IoU thresholds on head and full body ground truths, and the

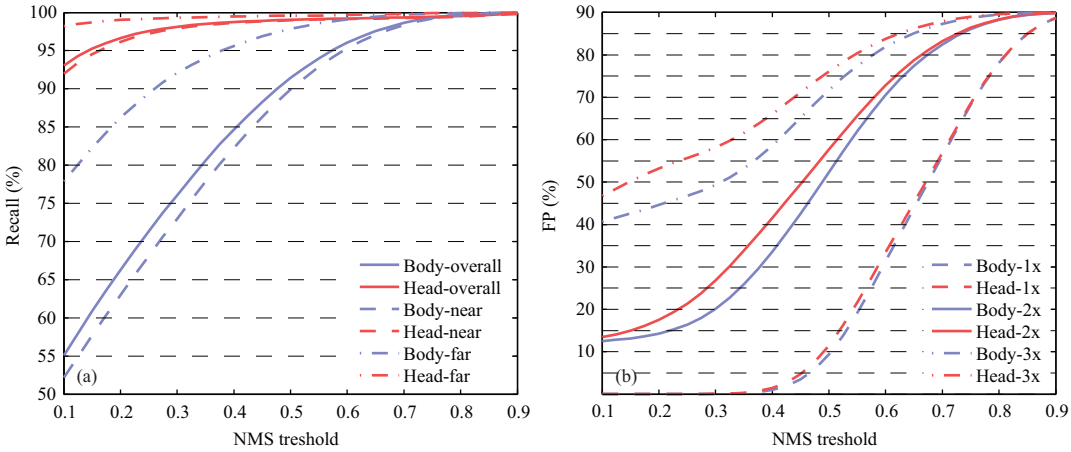


Figure 5 (Color online) (a) Recall of the ground truths using the traditional NMS algorithm with different thresholds. (b) False positive rate of the generated proposals based on the ground truths with different disturbances. 1x: variance of the detections and the corresponding ground truths.

recall rate under different conditions are reported in Figure 5(a). Even though the detection is perfect, numerous missing detections persist when only the body bounding boxes are used during evaluation. Second, to show the abilities of the head-based and body-based NMS algorithms to suppress the false positives around the ground truths during postprocessing, this approach generates 10 proposals around each respective head and body ground truth based on a Gaussian distribution with an identical variance. The variance is computed based on the true body/head detections of MSCNN [3] on the evaluation subset of CityPersons [17]. The final detections are generated through an NMS process with different IoU thresholds and evaluated using an IoU threshold (0.5). The false positive rates are shown in Figure 5(b). The results show that using only the head-based NMS process may introduce more false positives.

To effectively distinguish each individual from nearby persons, this paper proposes an augmented NMS criterion using both the body and head detections. The traditional NMS algorithm greedily selects a highest scoring detection and deletes the nearby neighbors with overlap exceeding a threshold. Similarly, the augmented NMS algorithm firstly selects the highest scoring detection $\boldsymbol{b}^m = (\boldsymbol{b}_b^m, \boldsymbol{b}_h^m, \tilde{s}^m)$ with the fused score. Then, a new measurement is employed to decide whether to delete the close-by detections $\boldsymbol{b}^n$. Let $(\boldsymbol{b}_h^m, \boldsymbol{b}_b^m)$ specify the head and full body boxes of the m-th pedestrian with highest score and $(\boldsymbol{b}_h^n, \boldsymbol{b}_b^n)$ indicate the close-by n-th pedestrian. The proposed algorithm introduces two thresholds θ_h and θ_b and performs NMS on the head and full body detections with these two thresholds, respectively. The close-by n pedestrian detection is deleted if

$$\mathrm{IoU}(\boldsymbol{b}_h^m, \boldsymbol{h}_h^n) \geqslant \theta_h \quad \text{or} \quad \mathrm{IoU}(\boldsymbol{b}_b^m, \boldsymbol{b}_b^n) \geqslant \theta_b. \tag{12}$$

In practice, θ_b is higher than the threshold in traditional NMS. Considering that there are some pedestrian boxes with no matched head detection, the traditional NMS algorithm is employed on the neighbors if any pedestrian among them has no corresponding head detection. These methods can effectively delete the redundant detections and retain the heavily overlapped detections.

4 Experiments

4.1 Datasets

CityPersons. CityPersons [17] is a pedestrian dataset, which is a subset of the semantic segmentation dataset CityScapes [50]. It consists of 5000 images recorded on streets from 27 cities and a total of $\sim$ 35000 annotated persons with an additional $\sim$ 13000 ignored regions. All of the experiments involving this dataset are conducted on the training (2975 images)/validation (500 images) sets for training and testing.

Head annotation on CityPersons. All of the pedestrians in CityPersons [17] are cropped, and the other categories of persons, for instance, sitting persons, riders, and fake humans, are ignored. Then, these cropped patches are resized into a uniform shape and presented to the annotators in a random

Table 1 Evaluation criteria for different conditions

Subset	Height	Occlusion	Occlusion-Peds
Reasonable (Rea.)	[50, inf)	[0, 0.35]	–
Partial	[50, inf)	(0.1, 0.35]	–
Heavy	[50, inf)	(0.35, 0.8]	–
Inter	[50, inf)	(0.1, 0.35]	[0, 0.1)
Crowd	[50, inf)	(0.1, 0.35]	[0.1, 1]

order. The pedestrian head is annotated following the same protocol [51] by drawing a line from the middle of the head top to the middle of the head bottom. Considering that the aspect ratio of all head annotations from CrowdHuman [9] is approximately equal to 1, a head bounding box is automatically generated such that its center coincides with the center point of the drawn line and its width and height are equal to the vertical height of the drawn line. Finally, these annotated head bounding boxes from the cropped patches are mapped back to the original images.

CrowdHuman-Ped. Considering that the human instances in CrowdHuman [9] are in different postures and recorded from different camera poses, to evaluate different pedestrian detectors, we select a subset from CrowdHuman that only consists of pedestrians and define it as CrowdHuman-Ped. For convenience, we select the subset from CrowdHuman using two simple criteria and add all of the remaining humans into the ignore regions. The criteria are as follows:

(1) The aspect ratio w/h of each full body box should be smaller than 0.5.

(2) The area of the head box should be smaller than 1/10 of the area of the full body box.

Finally, we collect 19370 images and 273000 pedestrian instances with an additional 293000 ignore regions. In the CrowdHuman-Ped dataset, more than 62.75% of pedestrian annotations are overlapped by other annotations with IoU above 0.3. In the following experiments, the detectors are trained and evaluated on the training/validation sets with the new annotations.

4.2 Evaluation metric

For evaluation, this paper follows the standard Caltech evaluation metric [52]: log-average miss rate (MR), which is computed by averaging the miss rate over nine false-positive-per-image rates evenly spaced in log-space in the range of $[10^{-2}, 10^{0}]$. To analyze the performance of pedestrian detectors under different conditions, different evaluation protocols [6, 52] are also adopted based on the pedestrian height and occlusion level. Our method mainly focuses on the people with height greater than 50 pixels in an image. Following the usual approach [52], the cases are subdivided between partial occlusion (1%–35% of area occluded) and heavy occlusion (35%–80% occluded). More details are shown in Table 1.

4.3 Implementation details

The original images are resized into multiple scales, and multiple patches (640×640) are randomly cropped from the resized images as the training samples. Multiple scales and multiple patches are commonly used by most of the related methods. It will not bring huge impact to our performance, but can be the icing on the cake. The parameters of the backbone are initialized with VGG16 [45]. HNets and PNets are first trained with randomly sampled negative samples and fine-tuned by performing bootstrapping in negative selection. λ_h and λ_b are both set to 1. During joint optimization, α_h is set to 1, and α_{aff} is set to 0.1. Stochastic gradient descent is used to optimize the model, with 10000 iterations for the first training procedure and 25000 for fine-tuning. Another 25000 iterations are employed for the joint optimization. The learning rates in all processes are initialized as 0.0001 and decay 10× after every 10000 iterations.

While training the three branches, different measurements of ignore regions [52] are adopted. For the body detection branch, the regions of sitting persons, riders, fake humans, etc. are treated as ignored pedestrian regions. During the training of the head detection branch, in addition to the ignored pedestrian regions, the regions of the pedestrians with no annotated heads are also added into the ignored regions. For the affinity prediction branch, an ignored region selection strategy identical to that for the head detection task is adopted.

Table 2 Body detection evaluation on CityPersons

Method	Rea.	Partial	Heavy	Crowd	Inter
Adapted FRCNN [17]	15.14	16.31	54.89	16.49	14.99
B-Net-D	14.72	13.56	55.48	14.29	10.06
Head⇒Ped(Res)	42.44	44.04	66.90	42.94	38.92
Head⇒Ped(Anchor)	43.62	36.62	67.96	33.94	28.19

Table 3 Head detection evaluation on CrowdHuman

Head height	$[0, \text{inf})$	$[0, 20]$	$(20, 40]$	$(40, \text{inf})$
H-Net-D	61.55	74.18	42.44	25.04
RetinaNet [40]	60.64	–	–	–
FPN [31]	52.06	–	–	–

4.4 Ablation study

4.4.1 *Different branches*

We evaluate the effectiveness of different branches on CityPersons [17] and CrowdHuman [9] datasets. In addition, we measure the influence of the head score for pedestrian detection and evaluate different head-body assignment methods. B-Net-D denotes a pedestrian detector that includes only the full body detection branch. It is treated as the baseline for evaluating the proposed detector. H-Net-D represents a head detector with only the head detection branch.

Body detection. As shown in Table 2, compared with the state-of-the-art two-stage detector (Adapted FRCNN [17]), B-Net-D achieves better performance overall in various conditions, which demonstrates the effectiveness of our full body detection branch. Additionally, we attempt to research whether the head detection is adequate for pedestrian detection. We generate the body detections using the head detections in two ways. Head⇒Ped(Res) denotes that the pedestrian bounding box is generated with a preset transformation from a detected head box. Head⇒Ped(Anchor) represents that we select the pedestrian proposals based on the detected head through the anchor-pair as the final pedestrian detections. The detections generated by these two methods are all scored using the corresponding head scores. As shown in Table 2, both of these methods perform worse (42.44 and 43.62 vs. 14.72 MR under reasonable conditions) than the straightforward pedestrian detector. The reason for this is that the head detection module is insufficient to detect relatively small objects and easily generates false positives. In short, the transformation from head to body is too weak to generate a reliable pedestrian detection. Considering that building a fine pedestrian detector is our primary goal, we treat B-Net-D as our baseline and refine the pedestrian detection through the head.

Head detection. Considering that the head appearance of humans in CrowdHuman is similar to that of the pedestrians in CrowdHuman-Ped, the CrowdHuman dataset enables some evaluation of head detection in previous studies. Therefore, we perform the head detection evaluation on the CrowdHuman validation set. The results are shown in Table 3. In the CrowdHuman benchmark [9], FPN [31] and RetinaNet [40] are used as two baseline detectors to represent the two-stage algorithms and single-stage algorithms with ResNet-50 [53]. At first sight, one may be surprised at the difference between the two-stage method and the one-stage method. However, it is normal. The two-stage method mainly functions through region proposal to generate a large number of potential bounding boxes, which may contain objects to be detected, and then uses the classifier to determine whether each bounding box contains objects. The one-stage method treats the object detection task as a regression problem and does not use sliding windows or region proposal. Actually, the classifier can acquire the local information of the image, while the one-stage method obtains information from the whole image during training and testing, which is more prone to error in object location. Thus, the results of small objects would not be good, especially for dense pedestrian heads. Compared with the one-stage method, our head detector achieves similar performance (61.55 MR) as RetinaNet [40] (60.64 MR) under the overall conditions. Additionally, we also show the performance of the head detector under different scale conditions. As shown in Table 3, the performance of the head detector is promoted rapidly when detecting larger-scale heads. An important reason for this is that the smaller heads lack effective features for reliable detection because of the smaller receptive fields. To alleviate the erroneous guidance for pedestrian detection, we ignore the head detections with height smaller than 20.

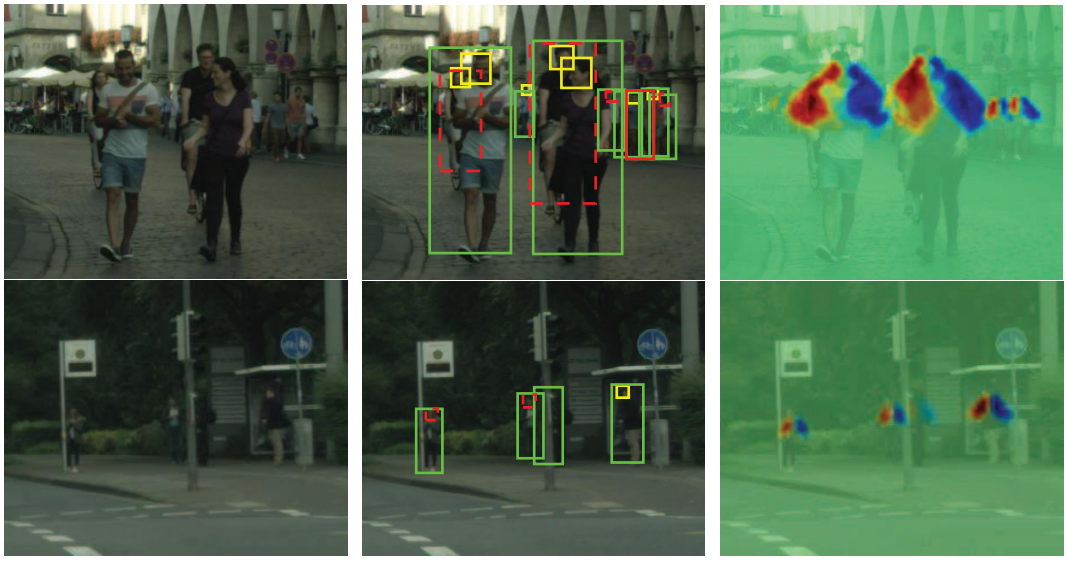


Figure 6 (Color online) Head and full body detections and head-side affinity predictions. The red dashed boxes indicate the missing detections. The red solid box indicates the false detection. The green boxes represent the body detections by B-Net-D. The yellow boxes are the head detections by H-Net-D. The color in (c) represents the orientation of the regressed vector in the head-side affinity field. Red: left orientation. Blue: right orientation.

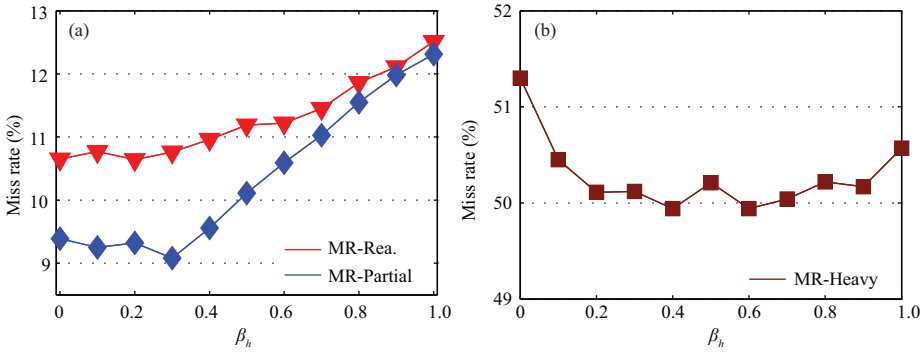


Figure 7 (Color online) Pedestrian detection performance under different validation subsets with different β_hs.

Affinity map prediction. We show the predicted affinity fields in Figure 6(c). The predicted affinity field can effectively represent the location relationship between the head and body detections. It works well under various scale conditions and indifferent occlusion cases. However, it inevitably fails when the pedestrian is too small or under excessively heavy occlusion conditions.

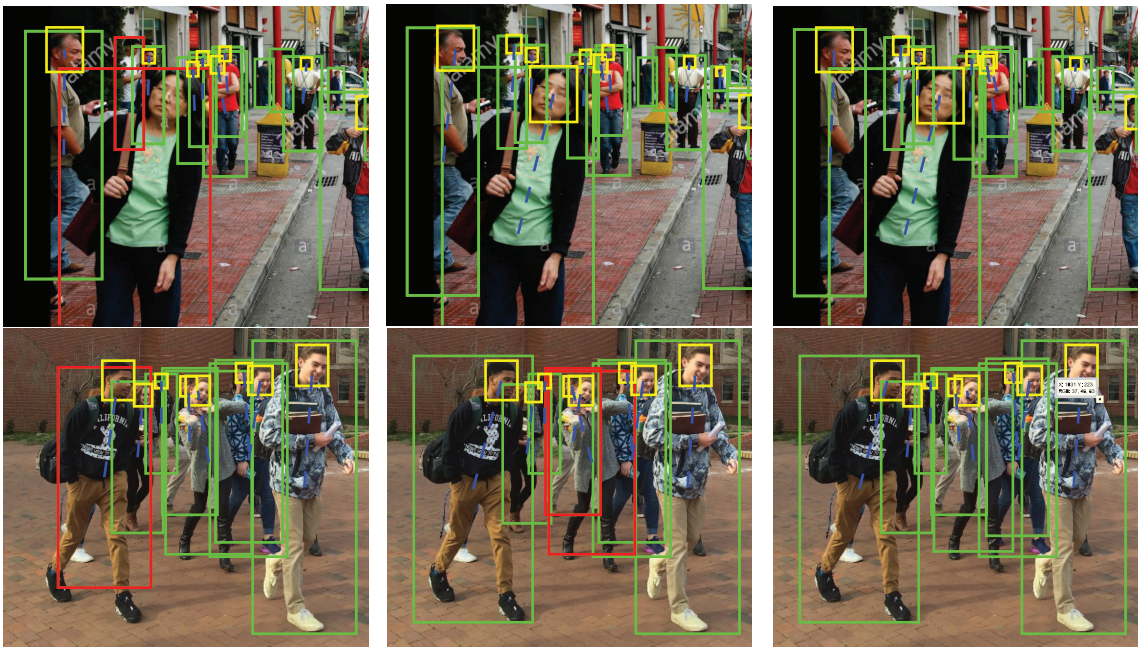
We show the complementarity of these three tasks in Figure 6. In the first row, there are some missing detections due to occlusion or the NMS process. However, the head detections provide a good guide to recognize them, and the head-side affinity map provides a cue to associate the detected head with the corresponding full body box. When the head detection branch fails, as shown in the second row, the body detector can still provide reliable detection.

Fused score. We use an affinity-based method to assign the head and body detections and fuse the detection scores with different β_hs. The detection performance on CityPersons under different conditions is shown in Figure 7. Increasing β_h means that the final detection relies more on the head detection. Larger β_h is therefore good for detecting occluded pedestrians. However, an overly large β_h decreases the performance. Considering the performance under different conditions, we choose $\beta_h = 0.2$ and adopt this value in the following experiments.

Assignment evaluation. We assign the head detections to the body detections from B-Net-D using three different modes and evaluate the effectiveness of these assignment methods on the CrowdHuman-Ped validation dataset. The "Greedy-based" mode represents assignment of the detected head to the body with the minimum distance from the preset head box. The position and size of the preset head

Table 4 Different assignment modes on CrowdHuman-Ped

Method	Accuracy	Recall	Rea.	Heavy	Crowd
B-Net-D	–	–	14.31	29.58	13.14
+Greedy-based	0.70	0.64	14.64	26.70	13.02
+Aug-NMS	–	–	14.32	25.29	11.97
+Anchor-pair-based	0.60	0.55	15.23	28.08	13.39
+Aug-NMS	–	–	15.05	26.78	12.53
+Affinity-based	0.96	0.56	13.65	26.46	12.14
+Aug-NMS	–	–	12.95	24.74	11.32



(a) (b) (c)

Figure 8 (Color online) Instance results of different assignment methods. Red box: missing or poorly assigned associations of head and body. Green box: pedestrian detection. Yellow box: head detection. Blue dashed line: head-body assignment. (a) Greedy-based; (b) Anchor-pair-based; (c) Affinity-based.

box are relative to the corresponding body box and calculated according to the value of the head and full body ground truths in CrowdHuman-Ped. The performance under different conditions is reported in Table 4. Considering that the head detection performance is poor for small-scale heads, we only consider the pedestrian instances in which heads are larger than 20 pixels when measuring the accuracy and recall.

In Table 4, we can observe that the head-body assignment gains more obvious improvements in occlusion cases. Among the three different assignment modes, the "Anchor-pair-based" mode performs the worst. This is because the head detection from the corresponding anchor always results in a low score and easily causes confusion with the heads of nearby pedestrians. This also can be seen from the low accuracy score in Table 4. If the deviation of the body and the corresponding head is too great, then the "Greedy-based" mode encounters difficulty in assigning the detection head box to the right body. Some unsuccessfully assigned instance results are shown in Figure 8. The "Affinity-based" mode can accurately assign the body and the corresponding head (as shown in Figure 8) and achieve better detection performance under different conditions (as shown in Table 4). Moreover, when the augmented NMS algorithm is used in the postprocessing, benefiting from the better alignment results, the "Affinity-based" mode achieves more improvement under the heavy occlusion case.

4.4.2 *Augmented NMS*

We aim to find the best thresholds θ_h and θ_b for the pedestrian detector. We adopt the "Affinity-based" pedestrian detector and calculate the MR on a subset selected from the CrowdHuman-Ped training set. Considering that the annotated visible region in the subset is confused, we re-define the occluded region

Table 5 Augmented NMS with head and body IoU thresholds

θ_h	θ_b				
	0.4	0.5	**0.6**	0.7	0.8
0.05	0.4041	0.3752	0.3666	0.3713	0.3716
0.10	0.3996	0.3658	0.3565	0.3616	0.3646
0.15	0.3984	0.3625	**0.3534**	0.3608	0.3660
0.20	0.3976	0.3611	0.3542	0.3649	0.3728
0.25	0.3975	0.3605	0.3553	0.3727	0.3857

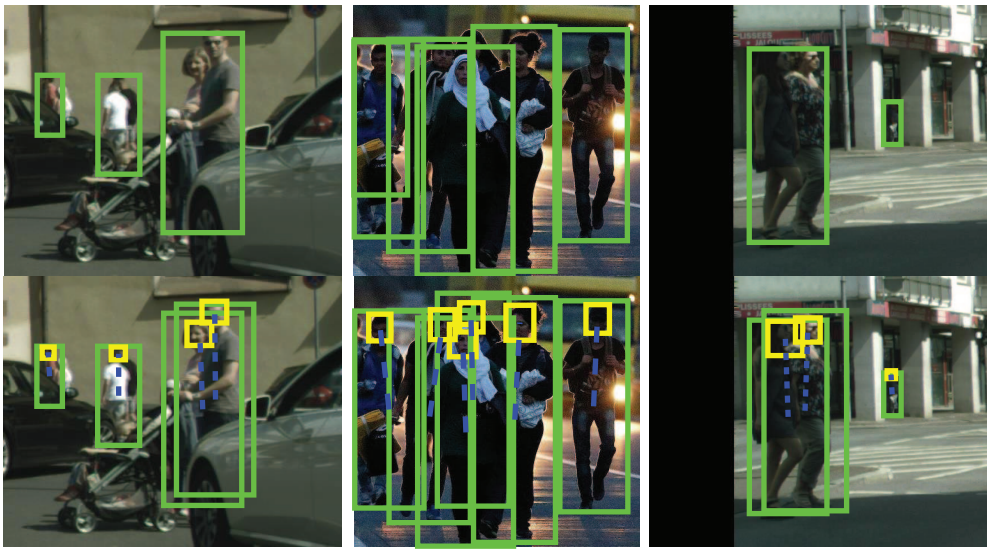


Figure 9 (Color online) Comparison of different NMS algorithms. First row: traditional NMS approach. Second row: augmented-NMS approach. Green box: body detection. Yellow box: head detection. Blue dashed line: head-body assignment.

of a pedestrian according to the region of maximum overlap with nearby pedestrians. The performance is evaluated on the occluded subset wherein the pedestrians are occluded by more than 35%.

Considering that there are some body detections with no corresponding head detections, we adopt two steps to carry out the augmented NMS process. (1) If one or two pedestrians in an instance pair have no head detections, then we use the traditional NMS algorithm with a threshold 0.5 on the body bounding box to suppress the nearby detections. (2) If both the pedestrians have head detections, then we adopt the augmented NMS algorithm to reject the nearby detections. The results are shown in Table 5.

As shown in Table 5, $\theta_b = 0.6$ and $\theta_h = 0.15$ are the best configurations for the pedestrian detector. Increasing or decreasing the thresholds will worsen the detection performance. θ_h is much smaller than θ_b because the head is less occluded than the whole body. In the following experiments, we use the two-step augmented NMS approach to acquire the final detections with the best threshold configuration. Some instance results with different NMS approaches are shown in Figure 9.

4.5 Comparisons with state-of-the-art methods

4.5.1 *CityPersons*

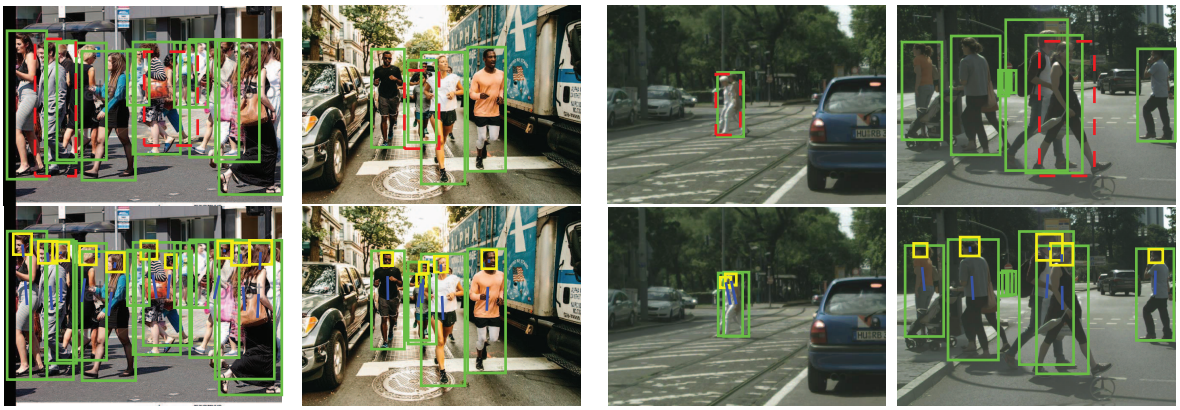
We compare our detector with the state-of-the-art detectors under different occlusion levels. Table 6 summarizes the results on the CityPersons validation dataset. We mark our method as HAPNet†. All the results refer to the corresponding papers or their result files. According to the MR results for subsets with different occlusion degrees, our VGG16-based HAPNet achieves the best performance. Taking a closer look, there is remarkable improvement for the Partial subset, Heavy subset and Crowd subset, especially in the second bar. The algorithms in the second bar show the performance of the state-of-the-art pedestrian detection under occlusion. OR-CNN [7] yields lower MR; however, our method offers a value only 0.1% higher. Our method yields better results with respect to heavy occlusion: 1.4% lower than that of OR-CNN. Moreover, although AdaptiveNMS [41] produces the same MR as our method, the performance of our method on the Heavy subset is better than that of this method.

Table 6 Pedestrian detection results for different conditions on the CityPersons validation set

Method	Backbone	Reasonable	Partial	Heavy	Crowd
ATT-part [20]	VGG16	16.0	–	56.7	–
Adapted FRCNN [17]	VGG16	15.1	16.3	54.9	16.5
RepLoss [6]	ResNet-50	13.2	16.8	56.9	13.5
OR-CNN [7]	VGG16	**12.8**	–	55.7	–
AdaptiveNMS [41]	VGG16	12.9	–	56.4	–
HAPNet†(ours)	VGG16	12.9	**13.6**	**54.3**	**12.6**

Table 7 Results for different conditions on CrowdHuman-Ped

Method	Reasonable	Heavy	Crowd
B-Net-D	14.31	29.58	13.14
HAPNet	12.95	24.74	11.32



(a) (b)

Figure 10 (Color online) Instance results of different detectors. (a) Results of B-Net-D (first row) and HAPNet on the CrowdHuman-Ped dataset; (b) results of Adapted FRCNN [17] (first row) and HAPNet on CityPersons. Red dashed box: missing detection. Green box: pedestrian detection. Yellow box: head detection. Blue dashed line: head-body assignment.

With regard to inference time, we get the result of 120 ms per image. Inference time (ms) is measured on scale $\times 1$ images, and the testing time was performed on GTX 1060. Although we use a three-branch structure and propose corresponding modules, we rarely use per-element multiplication which reduce a lot of expenses. Most of the time may be consumed in the head detection, because the one-stage method is not robust enough to the small-scale object. PRNet [29] performs the experiments on 2 GTX 1080Ti with 220 ms per image, and the BGCNet [33] is tested on the Tesla V100 with 156 ms per image. Our proposed HAPNet performs well in the occlusion pedestrian detection, which is in accordance with our intention that HAPNet is specifically designed to address the occlusion problem.

4.5.2 *CrowdHuman-Ped*

We have also evaluated HAPNet on our proposed CrowdHuman-Ped dataset. The results are shown in Table 7. The fact that CrowdHuman-Ped consists of numerous crowd scenes makes it a challenge for pedestrian detection. Compared with the baseline B-Net-D, our HAPNet achieves improvement with respect to a reasonable subset (+1.36 MR) and achieves more significant improvement for occlusion cases (+4.84 and +1.82 MR for heavy and crowd conditions, respectively).

4.6 Instance results

We show some instance results of different pedestrian detectors on CrowdHuman-Ped and CityPersons. As shown in Figure 10, B-Net-D and Adapted FRCNN [17] easily miss the heavily occluded ground truths. One reason for this is that the occluded pedestrians cause low detection scores, and another reason is that the detections nearby are easily suppressed by the traditional NMS algorithm. Due to the introduction of head detection and the augmented NMS algorithm, our detector exhibits better performance during detection of occluded pedestrians.

5 Conclusion

In this paper, we explore the connection between the head and the body by proposing HAPNet. The network is designed for the pedestrian detection task, especially for occluded cases. Through a three-branch detection structure with a shared backbone, we show the robustness and the validity on CityPersons and CrowdHuman-Ped. Specifically, the proposed head-side affinity module exploits the inherent structure of the head-body pair, effectively expressed as the association between the two parts. Moreover, during the postprocessing, the re-scoring module and the augmented NMS algorithm are both designed for matching the head-body pair in order to achieve an effective trade-off of reducing false positives and retaining the overlapped pedestrians. In summary, the proposed method represents an innovation in the network and postprocessing, and its effectiveness is proven by the experimental results.

Acknowledgements This work was supported by the Beijing Natural Science Foundation (Grant No. L201022) and in part by the National Natural Science Foundation of China (Grant Nos. 61876112, 61976170).

References

1 Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
2 Zhou C, Yuan J. Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of European Conference on Computer Vision, 2018
3 Cai Z, Fan Q, Feris R, et al. A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of European Conference on Computer Vision, 2016
4 Mao J, Xiao T, Jiang Y, et al. What can help pedestrian detection? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
5 Xie J, Pang Y W, Cholakkal H, et al. PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection. Sci China Inf Sci, 2021, 64: 120103
6 Wang X, Xiao T, Jiang Y, et al. Repulsion loss: detecting pedestrians in a crowd. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
7 Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: Proceedings of European Conference on Computer Vision, 2018
8 Huang X, Ge Z, Jie Z, et al. NMS by representative region: towards crowded pedestrian detection by proposal pairing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020
9 Shao S, Zhao Z, Li B, et al. Crowdhuman: a benchmark for detecting human in a crowd. 2018. ArXiv:1805.00123
10 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005
11 Dollár P, Tu Z, Perona P, et al. Integral channel features. In: Proceedings of British Machine Vision Conference, 2009
12 Zhou C, Yuan J. Multi-label learning of part detectors for heavily occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2017
13 Liu T, Duan H B, Shang Y Y, et al. Automatic salient object sequence rebuilding for video segment analysis. Sci China Inf Sci, 2018, 61: 012205
14 Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Conference and Workshop on Neural Information Processing Systems, 2015
15 Zhang L, Liang L, Liang X, et al. Is faster R-CNN doing well for pedestrian detection? In: Proceedings of European Conference on Computer Vision, 2016
16 Wu J, Zhou C, Yang M, et al. Temporal-context enhanced detection of heavily occluded pedestrians. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020
17 Zhang S, Benenson R, Schiele B. CityPersons: a diverse dataset for pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
18 Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63: 120106
19 Sun H Q, Pang Y W. GlanceNets—efficient convolutional neural networks with adaptive hard example mining. Sci China Inf Sci, 2018, 61: 109101
20 Zhang S, Yang J, Schiele B. Occluded pedestrian detection through guided attention in CNNs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
21 Brazil G, Liu X. Pedestrian detection with autoregressive network phases. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
22 Noh J, Lee S, Kim B, et al. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
23 Liu W, Liao S, Hu W, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of European Conference on Computer Vision, 2018
24 Lin C, Lu J, Wang G, et al. Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of European Conference on Computer Vision, 2018
25 Liu W, Liao S, Ren W, et al. High-level semantic feature detection: a new perspective for pedestrian detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
26 Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
27 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of European Conference on Computer Vision, 2016
28 Fu C, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector. 2017. ArXiv:1701.0665
29 Song X, Zhao K, Chu W, et al. Progressive refinement network for occluded pedestrian detection. In: Proceedings of European Conference on Computer Vision, 2020

30 Cao J, Pang Y, Han J, et al. Taking a look at small-scale pedestrians and occluded pedestrians. IEEE Trans Image Process, 2020, 29: 3143–3152
31 Lin T, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
32 Li J, Liang X, Shen S M, et al. Scale-aware fast R-CNN for pedestrian detection. IEEE Trans Multimedia, 2018, 20: 985–996
33 Li J, Liao S, Jiang H, et al. Box guided convolution for pedestrian detection. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020
34 Pang Y, Xie J, Khan M, et al. Mask-guided attention network for occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019
35 Zhou C, Yang M, Yuan J. Discriminative feature transformation for occluded pedestrian detection. In: Proceedings of IEEE International Conference on Computer Vision, 2019
36 Zhao Y, Yuan Z, Zhang H. Joint holistic and partial CNN for pedestrian detection. In: Proceedings of British Machine Vision Conference, 2018
37 Chi C, Zhang S, Xing J, et al. Pedhunter: occlusion robust pedestrian detector in crowded scenes. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020
38 Chi C, Zhang S, Xing J, et al. Relational learning for joint head and human detection. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020
39 Chen G, Cai X, Han H, et al. HeadNet: pedestrian head detection utilizing body in context. In: Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition, 2018
40 Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of IEEE International Conference on Computer Vision, 2017
41 Liu S, Huang D, Wang Y. Adaptive NMS: refining pedestrian detection in a crowd. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
42 Chu X, Zheng A, Zhang X, et al. Detection in crowded scenes: one proposal, multiple predictions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020
43 Zhao Y, Yuan Z, Chen B. Training cascade compact CNN with region-IoU for accurate pedestrian detection. IEEE Trans Intell Transp Syst, 2020, 21: 3777–3787
44 Bodla N, Singh B, Chellappa R, et al. Soft-NMS-improving object detection with one line of code. In: Proceedings of European Conference on Computer Vision, 2017
45 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
46 Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
47 Girshick R. Fast R-CNN. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
48 Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018
49 Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection and segmentation. In: Proceedings of IEEE International Conference on Computer Vision, 2017
50 Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
51 Zhang S, Benenson R, Omran M, et al. How far are we from solving pedestrian detection? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
52 Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 743–761
53 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016