Occluded Pedestrian Detection via Distribution-Based Mutual-Supervised Feature Learning

Ye He, Chao Zhu, and Xu-Cheng Yin, *Senior Member, IEEE*

***Abstract*— Pedestrian detection is a very important task in intelligent transportation system. State-of-the-art detectors work well on non-occluded pedestrians, but they are still far from satisfactory for heavily occluded ones. Recently, to deal with occlusion problems, the popular two-stage approaches are to build a two-branch architecture with the help of additional visible body annotations. However, these methods still have disadvantages. Either the two branches only use score-level fusion, which cannot guarantee the detectors to learn more robust pedestrian features. Or they only focus on the features of visible part via the attention mechanisms. However, the visible body features of heavily occluded pedestrians are only concentrated in a relatively small area, which may easily lead to missed detections. To alleviate the above issues, we propose a novel Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN), to better deal with occluded pedestrian detection. The key DMSFL module in our network is to learn more discriminative feature representations of pedestrians by minimizing the similarity loss between feature distributions of full body and visible body, which has two advantages: enhancing the feature representations of occluded pedestrians and reducing the intra-class variance in pedestrians. To facilitate the DMSFL module, we also propose a novel two-branch network architecture, which is trained in a mutual-supervised way with both full body and visible body annotations respectively. Extensive experiments are conducted on four challenging pedestrian datasets: Caltech, CityPersons, CrowdHuman and CUHK occlusion. Our approach achieves superior performance compared to other state-of-the-art methods, especially on heavy occlusion subsets.**

***Index Terms*— Pedestrian detection, occlusion handling, mutual-supervised feature learning, visible body information.**

I. Introduction

With the rapid development of urbanization, traffic pressure increases rapidly, and traffic accidents occur frequently. Therefore, the development of intelligent transportation system has attracted more and more attention. Automatic pedestrian detection based on machine learning and computer vision techniques plays an important role in the intelligent transportation system and has a broader application prospect. For example, it can be applied in the auxiliary driving system of intelligent vehicles, which is conducive to reducing the occurrence of traffic accidents and casualties. It can also be applied to the statistics of pedestrian flow and analysis of abnormal behavior in intelligent traffic monitoring systems. In addition, it can also be applied to the scene where a large number of pedestrians gather in the front lane of the airport terminal, the waiting hall of the railway station, the front of the bus station and the parking lot. It will be very important to carry out effective pedestrian detection and pedestrian behavior analysis in these scenes. Therefore, pedestrian detection possesses very important practical significances in intelligent transportation system.

In recent years, pedestrian detection gains a significant improvement on detection accuracy with the advances in convolutional neural networks (CNNs) and object detection. A large number of pure CNN based pedestrian detection methods are proposed in the literature, which can be divided into several categories, such as feature-based method [1], cascade based method [2], scale aware method [3], multi task method [4], attention based method [5], loss driven method [6] and so on. State-of-the-art pedestrian detectors [1]–[6] have achieved significant progress on non-occluded and slightly occluded pedestrians, yet they are still far from being satisfactory for detecting heavily occluded pedestrians. Handling occlusions is a critical issue. In fact, occluded pedestrians occur quite frequently in many scenarios of pedestrian detection. For example, walking on a street, pedestrians are likely to be occluded by other moving persons, such as drivers, cyclists, motorcyclists, skateboarders, etc. Pedestrians may also be occluded by some fixed obstacles, such as billboards, traffic signs, buildings, etc. Moreover, CityPersons dataset [7] has about 70% of pedestrians affected by various degrees of occlusions. Therefore, how to robustly detect partially or heavily occluded pedestrians has always been one of the most challenging problems for pedestrian detection task.

Many existing approaches [1]–[4] employ a simple detection strategy that assumes entirely visible pedestrians when trained with full body annotations. Despite achieving progressive results for non-occluded pedestrians, such a strategy is still unsatisfied under partial or heavy occlusions since the features of the occluding part are vastly different from the visible part.

Manuscript received 10 November 2020; revised 1 April 2021 and 16 June 2021; accepted 29 June 2021. Date of publication 4 August 2021; date of current version 9 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072032 and Grant 61703039 and in part by the Beijing Natural Science Foundation under Grant 4174095. The Associate Editor for this article was Z. Duric. *(Corresponding author: Chao Zhu.)*

The authors are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: s20190676@xs.ustb.edu.cn; chaozhu@ustb.edu.cn; xuchengyin@ustb.edu.cn).

Digital Object Identifier 10.1109/TITS.2021.3094800

Compared with full bounding boxes of pedestrians, visible parts of pedestrians normally suffer much less from occlusion, which can pinpoint the area for guiding the detector where should be focused on. Recently, several competitive pedestrian detectors [8]–[10] tackle occlusions by building a two-branch architecture with extra visible-region information, available with standard pedestrian detection benchmarks, like Caltech [11] and CityPersons [7]. Either the two branches are trained independently with only score-level fusion, such as Bi-box [8] or the attention mechanisms are exploited to emphasize on the visible regions while suppressing the occluded regions, like MGAN [10]. However, these occlusion handling approaches still have some weaknesses. Firstly, the visible body features of the heavily occluded pedestrians are concentrated in a relatively small area, which will easily result in missing detections. Secondly, some two-branch methods collect positive training samples with full body annotations and visible body annotations simultaneously, which may sacrifice some useful visible features, such as Bi-box [8]. As illustrated in Fig. 1, (b) and (d) depict the feature maps learned by Bi-box [8] and MGAN [10], respectively. It can be noticed that only the visible part has a higher response, while the occluded part almost has no response, which easily leads to inaccurate detection and even missing detection under heavy occlusions. Therefore, we argue that in the case of heavy occlusion, it is insufficient to focus only on the features within the visible part of pedestrians. The assistance of features from the occluding part is also very helpful as a context cue to enhance pedestrian detector against heavy occlusion to get more accurate full body detection box, and this has not been studied thoroughly in previous works.

Therefore in this paper, we aim to design a simple and effective approach to enhance feature representations for heavily occluded pedestrians. Specifically, we propose a novel Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN). A key part of the proposed DMSFLN is the Distribution-based Mutual-Supervised Feature Learning module, which aims at learning more discriminative feature representations of pedestrians by forcing feature distributions between full body and visible body as close to each other as possible. To obtain the visible boxes, we also construct a novel two-branch architecture consisting of a standard full body detection branch and an extra visible body classification branch. Moreover, these two branches sample their training samples supervised by full body annotations and visible body annotations, respectively (as displayed in Eq.3 and Eq.4) to obtain more focused training samples as shown in Fig. 2. Note that our proposed method can be easily applied to any existing region-based detection framework.

By applying the proposed mutual-supervised feature learning approach, there are two main advantages for pedestrian detection in heavy occlusion cases. Firstly, the pedestrian detector could learn more discriminative and robust feature representations with the assist of both reliable visible features and helpful contextual features from the occluding parts. Secondly, the intra-class variances of pedestrian features are reduced, making it easier for the classifier to distinguish the heavily occluded pedestrians from the background.

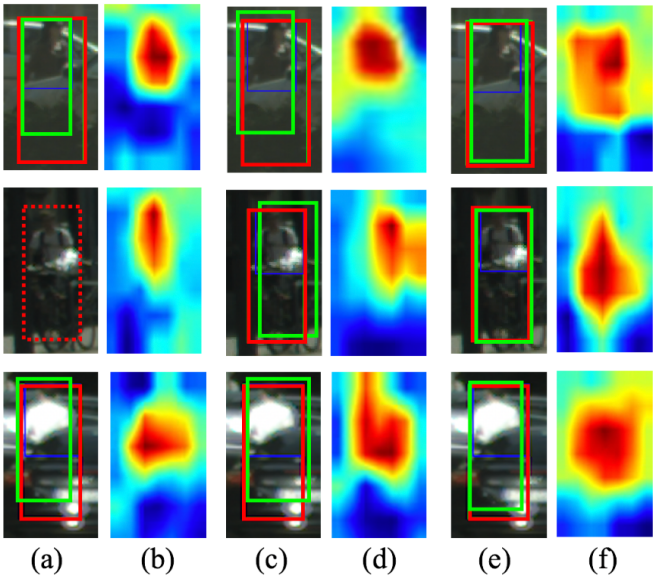


Fig. 1. Visual comparison between Bi-box, MGAN and our DMSFLN. (a), (c), and (e) represent the detection results of different methods. (b), (d), and (f) represent the feature visualization. Solid red boxes represent full body annotations, blue boxes are visible body annotations, green boxes denote detection results, and dashed red boxes represent the missed detections. The detected regions are cropped from the corresponding images in CityPersons val. set. Compared with Bi-box and MGAN, our DMSFLN displays a high response not only on the visible part but also on the occluding part.

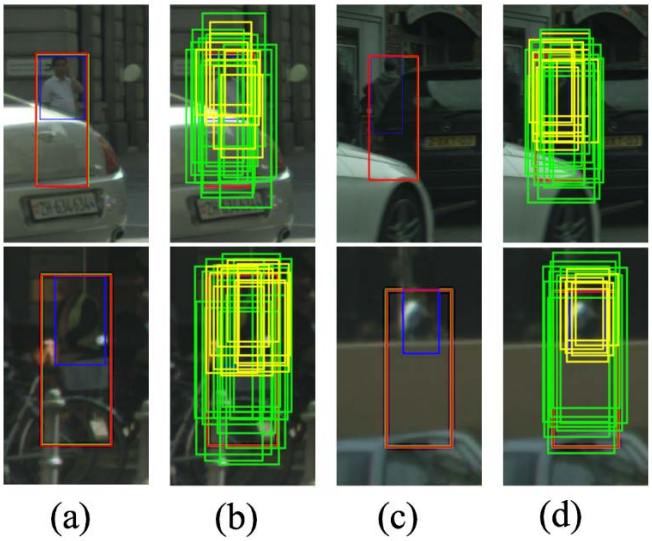


Fig. 2. Visualization of training samples obtained by our proposed sampling method. (a) and (c) are training images in CityPersons train. set. The Red boxes denote full body ground truth box and the blue boxes represent the visible body ground truth box of a pedestrian. (b) and (d) depict the positive training samples. Green boxes are positive training samples collected from full body branch and yellow boxes denote positive training samples collected from visible body branch.

Finally, as shown in Fig. 1 (f), it is obvious that the feature responses of our proposed method are concentrated in a relatively wider region, including not only the visible part but also the occluding part. Therefore, the contextual information from the occluding regions can essentially enhance the discriminability of heavily occluded pedestrian features.

To summarize, the contributions of this work are three-fold:

(1) We propose a novel Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN), to deal with the problem of occluded pedestrian detection. The key Distribution-based Mutual-Supervised Feature Learning module minimizes the similarity loss between feature distributions of full body and visible body to learn more robust feature representations of heavily occluded pedestrians;

(2) Our proposed DMSFLN comprises two branches: a standard full body detection branch and an extra visible body classification branch. The two branches are trained in a mutual-supervised way and collect their more focused training samples supervised by full body annotations and visible body annotations, respectively;

(3) Extensive experiments are carried out on four challenging pedestrian detection benchmarks: CityPersons [7], Caltech [11], CrowdHuman [12] and CUHK occlusion [13] dataset. Our approach has achieved a significant performance compared to other state-of-the-art approaches on all of these datasets, strongly validating its effectiveness for heavily occluded pedestrian detection.

A preliminary version of this work appeared in [14]. This paper includes that work but significantly extends it in the following ways. Firstly, we add a detailed analysis about the influence of occluded pedestrians on the region-based pedestrian detector (taking Faster RCNN [15] as an example), and confirm that most missed detections are caused by heavy occlusions due to insufficient feature representations within detection bounding boxes (See details in Section III). Secondly, we propose a more effective distribution-based mutual-supervised feature learning method to minimize the similarity loss between pedestrian features of full body and visible body. Specifically, we calculate the distance between two feature distributions in the feature space instead of calculating the similarity of features corresponding to the same ground-truth individual sample. By this way, further improved state-of-the-art results are achieved (See details in Section V-C). Thirdly, besides the CityPersons [7] and Caltech [11] datasets used in [14], we add more experimental evaluations on the other two challenging datasets (CrowdHuman [12] and CUHK occlusion [13]) and more ablation studies to extensively validate the effectiveness of the proposed approach (See details in Section V).

II. Related Work

A. Deep Pedestrian Detection

With the rapid development of deep convolutional neural networks (CNNs) [16]–[18], great progress has been made in the pedestrian detection field. Most existing CNN-based pedestrian detectors employ either a one-stage or two-stage strategy as their backbone architecture. One-stage approaches [1], [19], [20] predict locations directly in an unified one-shot structure which is fast and efficient. In contrast to one-stage approaches, two-stage detectors aim to pursue the state-of-the-art performance by generating the region proposals first and then refining these proposals after a scale-invariant feature aggregating operation named RoI Pooling [21] or RoI Align [22]. In recent years, two-stage pedestrian detection approaches [3]–[5], [7], [9], [23], [24] have shown significant performance on standard pedestrian datasets. For example, in [23], RPN [15] is employed to generate proposals and provide CNN features followed by a boosted decision forest. Zhang *et al.* [7] propose five key strategies to adapt the plain Faster R-CNN for pedestrian detection. Because of their promising performance on some pedestrian datasets [7], we also deploy a two-stage detection method as a backbone pipeline in this work.

B. Occlusion Handling in Pedestrian Detection

Many efforts have been made to deal with occlusions for pedestrian detection. A typical strategy [25]–[29] is the part-based approach where a set of part detectors are learned with each part designed to handle a specific occlusion pattern. Franken [25] quantifies the performance of Integral Channel Features detector in various occlusion situations, and proposes an effective method to train a series of detectors for various occlusion situations. Ouyang *et al.* [26] built a unifying deep learning model to join different tasks (i.e., feature extraction, deformation handling, occlusion handling, and classification). Tian *et al.* [27] developed an extensive part pool to train the multiple part detectors and then trained a linear SVM to combine the scores of part detectors. JL-TopS [28] proposed a multi-label learning method to jointly learn part detectors to capture partial occlusion patterns. The part detectors share a set of decision trees to exploit part correlations. OR-CNN [29] uses a new part occlusion-aware region of interest (PORoI) pooling unit to replace the RoI pooling layer in order to integrate the prior structure information of human body with visibility prediction into the network to handle occlusion. The parts used in these approaches are usually manually designed, which may not be optimal.

Different from the above approaches, some other approaches [6], [29]–[31] handle occlusions without using parts information. In [30], an implicit shape model is proposed to generate a set of pedestrian proposals which are further refined by exploiting local and global cues. Repulsion Loss [6] and AggLoss [29] design two unique regression losses to generate more compact proposals to make them less sensitive to the NMS threshold. Besides, in [32], an adaptive NMS strategy is proposed that applies a dynamic suppression threshold to an instance in crowded scenes.

In contrast to the aforementioned methods, recent approaches focus on utilizing annotations of the visible body as additional supervisions together with the standard full body annotations to handle the problem of occluded pedestrian detection. Zhang *et al.* [9] utilize visible body information along with a pre-trained body part prediction model to learn specific occlusion patterns (full, upper-body, left-body, and right-body visible). MGAN [10], a one-way supervision network, incorporates attention mechanisms into pedestrian detection with visible region supervision to emphasize the visible regions while suppressing the occluded regions. The work of Bi-box [8] regresses the full and visible body of a

pedestrian at the same time. However, the two branches of Bi-box [8] are trained separately with only score-level fusion, which cannot guarantee the detectors to learn robust enough pedestrian features.

In this work, we follow the idea of utilizing extra visible annotations to tackle the problem of occluded pedestrian detection. Different to [8], our proposed method effectively integrates the two branches in the feature level to obtain more discriminative features. Different to [10], our proposed method adopts a distribution-based mutual-supervised way to learn more robust pedestrian features from both visible part and occluding part, aiming at enhancing the feature representations against heavy occlusions.

III. Influence Analysis of Heavy Occlusions on Region-Based Pedestrian Detectors

This section mainly investigates the impact of heavy occlusions on region-based pedestrian detectors, taking Faster RCNN [15] as an example.

A. Preliminaries

1) Dataset and Metric: CityPersons [7] dataset is one of the most popular benchmarks for pedestrian detection, which is built upon the semantic segmentation dataset Cityscapes [33]. The dataset contains 5000 images (2975 for training, 500 for validation, and 1525 for testing) with 35k manually annotated persons and 13k ignore region annotations. Following the common practice in previous works, the detectors in this section are trained on the training subset and tested on the validation subset with enlarged resolution by 1.3×, compared to the original one. We report performance using standard average-log miss rate (**MR**) in experiments. Here, **MR** is computed over the false positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ [11]. Lower value represents better detection performance.

2) Baseline Detector: We adopt Adaptive Faster R-CNN [7] detector as our baseline detector which is modified for pedestrian detection. It refines on the standard Faster RCNN [15] by proposing the following 3 modifications: (1) The ratio of all anchors is set to 2.44. (2) Remove the fourth max-pooling layer from VGG-16 [16] and reduce the stride to 8 pixels, helping the detector to handle small-scale pedestrians. (3) During the training process, proposals and RoIs avoid sampling the ignore regions which the annotator cannot tell if a person is present or absent.

B. Failure Analysis

We obtain the evaluation results on Citypersons [7] validation subset and notice that there are less false positive detections while many false negative detections. Therefore, we mainly focus on false negative detections in the following analysis.

1) False Negative: The traditional region-based pedestrian detectors usually have two steps to obtain final detection boxes. These detectors first generate a pool of anchor boxes by the RPN [15] network, and then predict the classification scores and regress the final bounding boxes through the head network. As we all know, the region-based detectors have achieved state-of-the-art performances because a large number of anchor boxes are generated in the RPN network to capture every possible target in an image. However, these detectors still cannot accurately detect heavily occluded pedestrians, resulting in the fact that the recall of heavily occluded pedestrians is far lower than that of the non-occluded and partially occluded pedestrians.

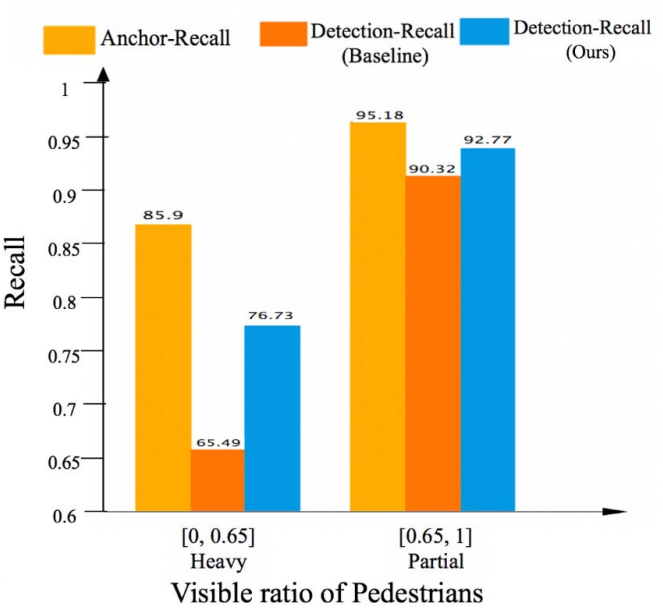


Fig. 3. Recall histogram of pedestrians with different degrees of occlusions. We only consider the detections with height greater than 50 pixels. Pedestrians with visible ratios in [0, 0.65] are regarded as heavily occluded pedestrians.

To further explore the influence of heavily occluded pedestrians on the region-based detectors, we compute the recall of pedestrians with different degrees of occlusion. Specifically, we define two indicators: Anchor-Recall and Detection-Recall. As for the former, we calculate the IoU between all generated anchor boxes and ground truth examples through the RPN [15] network. The anchor box is considered as a positive sample if IoU is greater than 0.5, otherwise it is a negative sample. As for the latter, we calculate the IoU between all detected boxes and ground truth samples. The detection box is regarded as a positive sample if IoU is greater than 0.5, otherwise it is a negative sample. Therefore, this process can be expressed as the following formula:

$$\text{Anchor-Recall} = \frac{|a_p|}{|GT|} \tag{1}$$

$$\text{Detection-Recall} = \frac{|d_p|}{|GT|} \tag{2}$$

where a_p represents the set of positive anchor boxes, d_p is the set of positive detection boxes, GT is the set of ground-truth samples. The symbol $|a_p|$, $|d_p|$, $|GT|$ denotes the number of elements in the set a_p, d_p and GT. In our experiments, 2000 proposals are obtained from RPN [15] network. As illustrated in Fig. 3, the anchor-recall of pedestrians with different degrees of occlusions is 95.18% for partial occlusion (visible

ratio in [0.65, 1]) and 85.9% for heavy occlusion (visible ratio in [0, 0.65]) respectively, showing that heavily occluded pedestrians are truly harder to be covered in the first RPN step. Then, for the baseline detector, the detection-recall of heavily occluded pedestrians decreases sharply from 85.9% to 65.49%, while the detection-recall of partially occluded pedestrians only declines from 95.18% to 90.32%. These results indicate that even most of the heavily occluded pedestrians can still be covered by anchor boxes after the RPN step, however, these pedestrians cannot finally be detected after the following detection head network. We believe that these failures are mainly caused by the insufficient feature representation of heavily occluded pedestrians within the full body bounding boxes, which makes it more difficult for the classifier to output high classification scores.

To better understand the missed detection problem of heavily occluded pedestrians, some detection examples are visualized in Fig. 4. Note that the heavily occluded pedestrians are located accurately in the RPN step. However, they are finally considered as negative samples in the detection head network due to low classification scores which are lower than 0.5.

2) Conclusion: The above analyses indicate that heavy occlusion remains a difficult issue in pedestrian detection. The main reason is that for the heavily occluded pedestrians, the useful features extracted from the full body boxes are limited and insufficient, which makes it difficult for the classifier to predict a higher classification score, resulting in missed detections. Therefore, in this paper, we propose a novel distribution-based mutual-supervised feature learning method to enhance the feature representations of heavily occluded pedestrians, and Fig. 3 shows that it truly helps the detector to achieve much higher detection recall and performance compared to the baseline, especially in heavy occlusion case.

IV. Proposed Method

In this section, we propose a novel Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN) for occluded pedestrian detection. We describe the overall architecture of the proposed network in Section A. To obtain the most focused positive training samples, we propose a novel proposal sampling method in Section B. Next, we detail the design of the novel Distribution-based Mutual-Supervised Feature Learning Module in Section C. Finally, we present the total loss function of multi-task prediction for end-to-end training along with a fusion method of two branches during inference in Section D.

A. Overall Architecture

The overall architecture of our proposed method is illustrated in Fig. 5. Note that our proposed method can be easily applied to any existing region-based detection frameworks. For a fair comparison, we select Faster R-CNN framework [15] and adopt VGG-16 [16] as the backbone which is the most commonly used backbone in pedestrian detection networks. The architecture takes a raw image as input, first deploys a pre-trained ImageNet [34] model. Then extracted feature maps are sent to the region proposal network (RPN [15]) to generate two different sets of candidate proposals, including full body proposals and visible body proposals. For each proposal, a fixed-sized feature representation is obtained through RoI Align [22] layer. Finally, these features go through Full Body (FB) branch and Visible Body (VB) branch separately to generate final predictions. Specifically, FB branch is a standard pedestrian detection branch to generate classification scores and regressed bounding box coordinates. For the VB branch, we need to consider the choice of employing classification task only or both classification and regression tasks. Ref. [35] discussed that classification requires translation invariant features, while regression requires translation covariant features. If both classification and regression tasks are designed in VB branch, the regression task will force the detector to gradually learn the translation covariance features during the training process, which may reduce the performance of the classifier. Therefore, we only implement a classification task in the VB branch to accurately classify the visible proposals. See Section V-C for more comparative results.

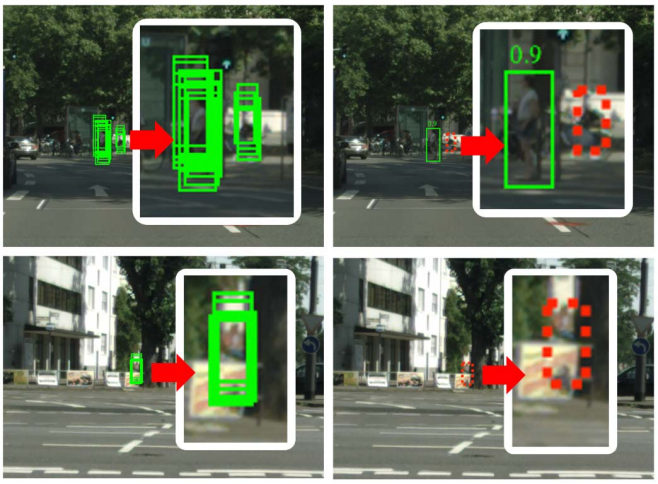


Fig. 4. Visualization of failure detection examples of baseline detector. The first and second columns represent the detection results generated through RPN and head network, respectively. Green boxes represent the detected bounding boxes and dashed red boxes are the missed detections.

B. Proposals Generation of Two Branches

In our architecture, the FB branch collects training samples using full body annotations and then passes them through a classification network (FC11, FC12) to generate the classification scores and the regressed bounding box coordinates. VB branch collects training samples employing visible region annotations and then feeds them into a simple classification network (FC21, FC22) to obtain the classification scores, indicating the probability that this visible proposal contains a pedestrian. The proposals generation process can be expressed with the following equations:

$$P_{FB} = \{x | IoU(x, GT_{FB}) > \alpha\} \cup \{y | IoU(y, GT_{FB}) <= \alpha\} \quad (3)$$

$$P_{VB} = \{x' | IoU(x', GT_{VB}) > \beta\} \cup \{y' | IoU(y', GT_{VB}) <= \beta\} \quad (4)$$

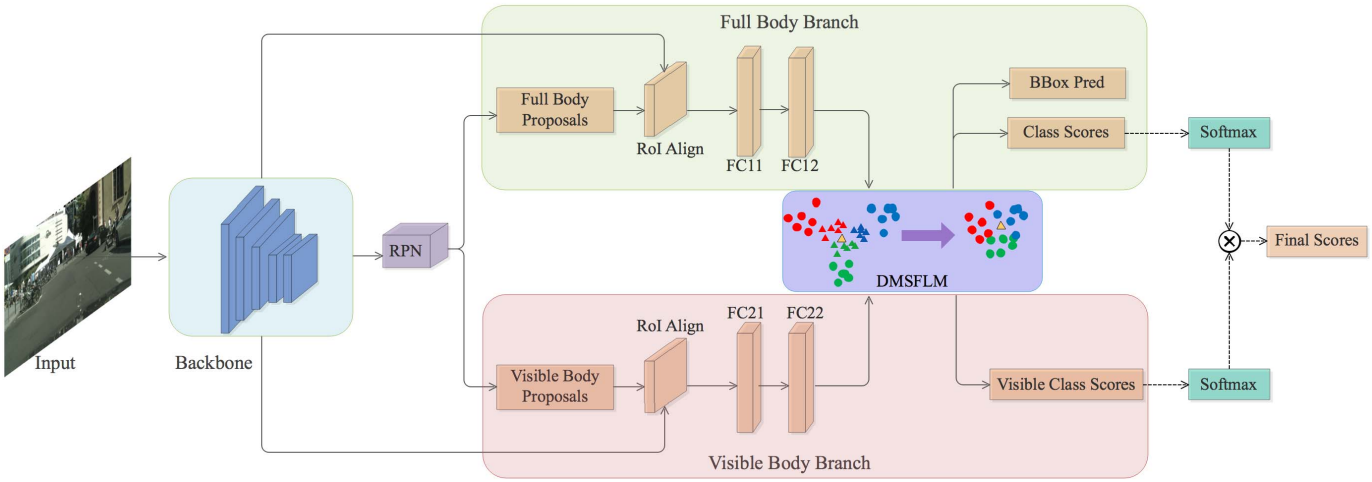


Fig. 5. Overall network architecture of our Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN). It consists of a full body (FB) branch enclosed in the green box and a visible body (VB) branch in the red box. A novel Distribution-based Mutual-Supervised Feature Learning Module (DMSFLM) is enclosed in the purple box. Two feature vectors are obtained from fully connected layer FC12 and FC22 respectively and then sent to the DMSFLM. In our architecture, the FB branch is a standard pedestrian detector branch and the VB branch is proposed to generate classification scores for visible proposals. FC_{ij} denotes the j-th FC layer in the i-th branch. The dotted lines depict the inference process. $\otimes$ represents element-wise product operation.

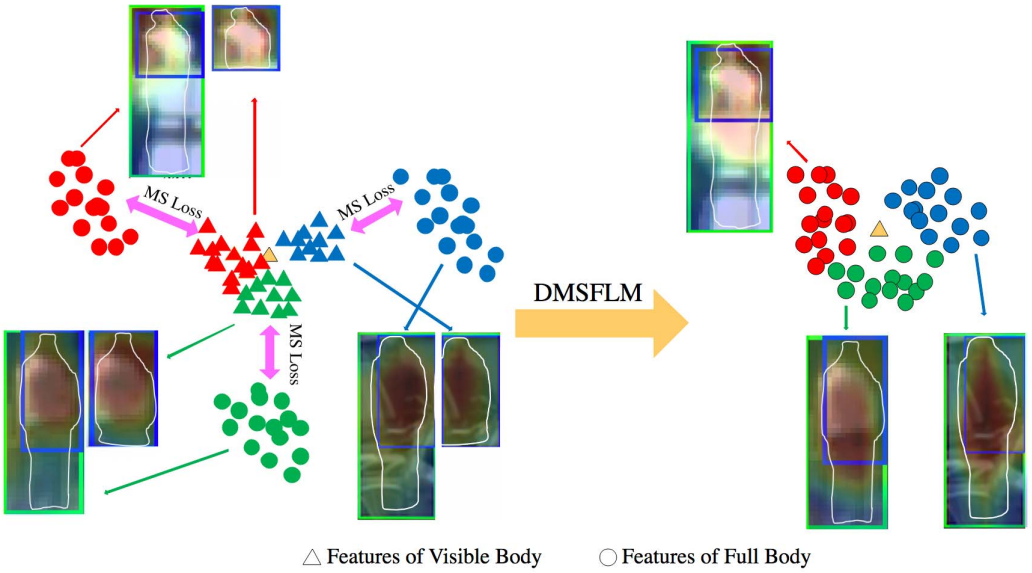


Fig. 6. Illustration of our proposed Distribution-based Mutual-Supervised Feature Learning Module. $MSLoss$ represents the mutual-supervised loss. The triangle represents the features of the visible part of the pedestrians and the circle represents the features of full body of the pedestrian. The yellow triangle represents the average feature of visible body. Different colors represent the training samples corresponding to different ground-truth examples. Note that the silhouettes of the pedestrian examples are enhanced manually for better visualization.

where P_{FB} and P_{VB} represent the training samples collected from FB branch and VB branch respectively, x and x' denote positive training samples collected from FB branch and VB branch respectively, and y and y' are negative training samples collected from FB branch and VB branch respectively. The positive samples generated by the two branches are shown in Fig. 2. For each branch, we sample 512 region proposals, the positive and negative samples are randomly sampled at a ratio of 1:3, following the same parameter settings as in [15]. α and β are the IoU thresholds for the FB branch and the VB branch, respectively. The influence of α and β is experimentally studied and shown in Fig. 8.

C. Distribution-Based Mutual-Supervised Feature Learning Module

From the analyses of Section III, we can see that most of the heavily occluded pedestrians can still be covered by anchor boxes after the RPN step, however these pedestrians are finally missed in the detection head network due to their insufficient feature representations. This observation inspires us that in order to improve the detection accuracy of heavily occluded pedestrians, both features from visible body and full body are required. The former is to provide discriminative information of pedestrian appearance to guide the detectors to

locate the occluded pedestrians more precisely, and the latter is to provide contextual information of pedestrian surroundings to assist the detectors to predict more accurate detection box for the occluding part. Thus, the goal of our proposed DMSFL module is to make better use of the features from both visible body and full body to learn a more complete and robust pedestrian feature, as shown in Fig. 1 (f). In order to achieve the above goal, we propose a mutual-supervised way to learn on the features from both visible body and full body, i.e. to force these two kinds of features as close to each other as possible, by minimizing their similarity loss. In our previous work [14], we measure their similarity by finding the features of visible body and full body corresponding to the same ground-truth individual sample and then calculating their distance. In this paper, we propose a more effective method to do so. As we know, the problem of pedestrian detection can be regarded as a two-class classification problem, and the main objective is to distinguish pedestrians from other objects or background, which means that all pedestrians belong to the same class and we do not need to distinguish one person from other person. Therefore, a better way to measure their similarity would be calculating the distance between two feature distributions of full body and visible body in the feature space.

In terms of formula, x and x' are positive samples obtained through FB and VB branches, respectively. We calculate the average features of x' positive samples for simplicity and then apply mutual-supervised loss for each positive feature of FB branch and the average feature of VB branch.

To calculate the similarity loss between features of full body and visible body, we compare several different methods including Manhattan distance, Euclidean distance, Kullback-Leibler divergence, and Cosine Similarity. Among these measurements, Cosine Similarity achieves the best results. See Section V-C for more comparative results.

The mutual-supervised loss is computed as:

$$L_{DMSFLM} = \frac{1}{P}\sum_{i=1}^{P}[1 - cos(x_i, \bar{x'})] \tag{5}$$

where P represents the number of positive samples from FB branch. $\bar{x'}$ is the average feature of all positive samples from VB branch. x_i represents the full body feature of ith positive sample.

Fig. 6 illustrates the effects of our proposed module. On the left side we visualize the feature maps of three example full body bounding boxes and visible body bounding boxes before our proposed DMSFLM, and on the right side we visualize three corresponding feature maps learned by DMSFLM. It can be seen that the feature responses after our proposed method are concentrated in a relatively wider region, including not only the visible part but also the occluding part, therefore facilitating pedestrian detection in occlusion cases. Note that the silhouettes of the pedestrian examples in Fig.6 are enhanced manually for better visualization.

1) Discussion: The proposed Distribution-based Mutual-Supervised Feature Learning Module has the following two main advantages:

1. By integrating the feature representations of visible body and full body through a mutual-supervised way, the detector can not only be guided by the features of the visible part but also utilize the contextual features of the occluding part to learn more discriminative and enhanced feature representations for heavily occluded pedestrians.

2. By minimizing the mutual-supervised loss, the intra-class variance of pedestrian features becomes smaller in the feature space, as illustrated in Fig. 6. Therefore, it is easier for detectors to distinguish pedestrians from the background, especially in heavy occlusion cases.

D. Multi-Task Optimization & Inference

Here, we present the final loss function for the proposed architecture DMSFLN. The overall loss formulation L is as follows:

$$L = L_{RPN_{cls}} + L_{RPN_{reg}} + L_{FB_{cls}} + L_{FB_{reg}} + L_{VB_{cls}} + L_{DMSFLM} \tag{6}$$

where $L_{RPN_{cls}}$ and $L_{RPN_{reg}}$ refer to the classification and regression loss of RPN, $L_{FB_{cls}}$ and $L_{VB_{cls}}$ refer to the classification loss of FB and VB, $L_{FB_{reg}}$ is the bounding box regression loss of FB and L_{DMSFLM} is the loss of Distribution-based Mutual-Supervised Feature Learning Module. Here, classification loss is Cross-Entropy loss and the bounding box regression loss is Smooth-L1 loss.

In the inference stage, we propose a simple yet effective method to fuse the information from two branches. Since the visible box contains more discriminative information of pedestrians, taking the classification score of the VB branch as a part of the final score of detection box would be helpful to improve the accuracy of pedestrian detector. We compare several different fusion methods to fuse the confidence scores of FB branch and VB branch, including minimum, maximum, average and multiplication. The results of multiplication performances best. See Section V-C for more comparative results. Specifically, the classification scores of the VB branch are multiplied by those of the FB branch as the scores of the final detection box. Formally, the final scores of pedestrians are defined as:

$$\begin{aligned} &Final\ Scores \\ &\quad = Softmax(Classification\ Scores) \\ &\qquad \otimes Softmax(Visible\ Classification\ Scores) \end{aligned} \tag{7}$$

where *Classification Scores* and *Visible Classification Scores* represent the raw scores output from FC12 and FC22, respectively, $\otimes$ denotes element-wise product operation.

V. Experimental Evaluation

In this section, we evaluate our proposed approach on CityPersons [7], Caltech [11], CrowdHuman [12] and CUHK occlusion dataset [13], which are four challenging pedestrian detection datasets with different occlusion settings.

Fig. 7. Illustrative examples from different human dataset benchmarks. The images inside the red, yellow, green and blue boxes are from the CUHK occlusion [13], Caltech [11], CityPersons [7] and CrowdHuman [12] datasets, respectively.

A. Datasets and Evaluation Metrics

1) Datasets: CityPersons [7] has been introduced in Section III. Caltech is a popular dataset [11] for pedestrian detection research containing 11 sets of videos. The first six video sets are for training and the remaining five video sets are for testing. To enlarge the size of the training set, we train our model on Caltech10×. Finally, the training and test sets have 42782 and 4024 images, respectively. In [7], Zhang *et al.* provided high-quality annotations by correcting several types of errors in the original annotations. In our experiments, we utilize the new and old annotations simultaneously for training and testing to compare with different approaches on Caltech dataset. CrowdHuman dataset [12] is recently released for better evaluating pedestrian detectors in crowded scenarios. It is a large benchmark dataset containing 15000, 4370, and 5000 images for training, validation, and testing subsets, respectively. CrowdHuman [12] provides three kinds of bounding boxes annotations which are head bounding-box, visible-region bounding-box, and full-body bounding-box. CUHK occlusion dataset [13] contains 1063 images with occluded pedestrians from the datasets of Caltech [11], ETHZ [36], TUD-Brussels [37], INRIA [38] and recorded images from surveillance cameras collected from [13]. Table I lists the statistical information of training subsets for different pedestrian detection datasets. Since the CUHK [13] occlusion dataset does not publish division rules for training and testing sets, we display the information of all images in Table I. All of the four datasets provide full body and visible body annotations. Fig. 7 illustrates some pedestrian examples in each of the four different datasets.

TABLE I

STATISTICAL COMPARISON OF DIFFERENT PEDESTRIAN DETECTION DATASETS

	Caltech	*CityPersons*	*CrowdHuman*	*CUHK*
#images	42782	2975	15000	1063
#persons	13674	19238	339565	7523
#ignore regions	50363	6768	99227	0
#person/image	0.32	6.47	22.64	7.08
#unique persons	1273	19238	339565	7523

2) Evaluation Metrics: Following the most common evaluation metrics in the related works, we report the performance using a log-average miss rate (**MR**) throughout our experiments. On Cityperons, we follow [7] and report the results across four different subsets: Reasonable (**R**), Bare, Partial, Heavy Occlusion (**HO**). For the Caltech dataset, we report results on Reasonable (**R**), **None**, **Partial**, Heavy Occlusion (**HO**), and the combined Reasonable + Heavy Occlusion (**R+HO**). The visibility ratio in **R** set is larger than 65%, none/partial/heavy occlusion within the original annotation space as defined in [11]. Thus, the visibility ratio in **R+HO** set is larger than 20%. According to [9], the pedestrians with height greater than 50 pixels are taken into consideration. Notice that **HO** set is designed to evaluate the performance under severe occlusions. Specifically, to facilitate the comparisons on the CrowdHuman dataset [12], the results in terms of Average Precision (**AP**) are also provided.

B. Implementation Details

For these four datasets, the network is trained on two GPUs with a total of 2 images per mini-batch. We adopt RoI

TABLE II
COMPARISON(IN LOG-AVERAGE MISS RATES) OF OUR DMSFLN WITH THE BASELINE ON THE CITYPERSONS

Method	*VB Branch*	*DMSFLM*	*R*	*HO*
Baseline	×	×	11.92	47.88
Our DMSFLN	cls+reg	×	11.45	45.35
	cls	×	10.78	44.81
	cls	pos+neg	10.29	40.18
	cls	pos	**9.88**	**38.12**

Align [22] instead of RoI Pooling [21] for feature extraction to get more precise features. Now, we detail settings specific to the four datasets.

1) Citypersons: We fine-tune pre-trained ImageNet VGG model [16] on the trainset of the CityPersons. We follow the same experimental protocol as in [7] and employ two fully connected layers with 1024 instead of 4096 output dimensions. We choose SGD with momentum of 0.9 as our optimizer and set the initial learning rate as 0.0025. We train 15 epochs in total and decrease the learning rate by 0.1 at the 8-*th* and 11-*th* epochs. Following [8], we only consider ground-truth pedestrian examples whose height is greater than 50 pixels and the visible ratio is more than 30%.

2) Caltech: We start with a model pre-trained on CityPersons dataset. The initial learning rate is 0.0025 for the first 3 epochs and is reduced by 10 and 100 times for another 2 and 1 epochs.

3) CrowdHuman: For fair comparisons with state-of-the-art approaches [6], [12], [32], [39] on CrowdHuman dataset, we also adopt Feature Pyramid Network (FPN) [40] with ResNet-50 [17] as our baseline. The anchor scales are set the same as [40], while the aspect ratios are set to $H : W = \{1 : 1, 2 : 1, 3 : 1\}$. We resize training images so that the short edge is 800 pixels while the long edge is smaller than 1400 pixels. Each training runs for 35 epochs and decreases the learning rate by 0.1 at 30-*th* and 33-*th* epochs.

4) CUHK: We first train the model on Caltech [11] and CityPersons dataset [7], respectively and then evaluate on CUHK occlusion dataset [13] to explore the robustness of the model for different degrees of occlusion.

Multi-scale training and testing are not applied to ensure fair comparisons with the other state-of-the-art methods.

C. Ablation Study

To adequately verify the effectiveness of the key components in our proposed method, we conduct detailed ablation studies on CityPersons dataset [7].

1) Baseline Comparison: Table II shows the performance comparisons between the baseline and our proposed method on CityPersons validation subsets. For a fair comparison, we keep the same training data, input scale (×1.3), and network backbone (VGG-16). The best results are boldfaced. Our reproduced baseline detector obtains a log-average miss rate of 11.92% on **R** set, outperforming the adapted Faster-RCNN baseline in the original paper [7] by 0.89%. Thus, our baseline model is strong enough to verify the effectiveness of the proposed components. To analyze the contributions of each key component in the proposed method individually, we gradually apply the VB branch and Distribution-based

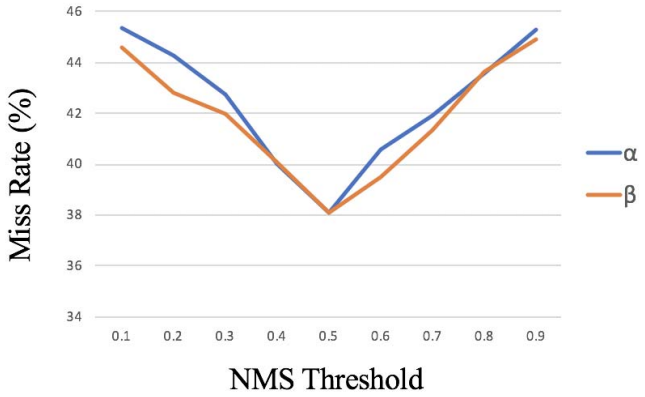


Fig. 8. Log-Average Missing Rates on **HO** subsets across various NMS thresholds at FPPI $= 10^{-2}$. The blue line represents the results when $\beta = 0.5$ and α is varied from small to large. The orange line represents the results when $\alpha = 0.5$ and β is varied from small to large.

Mutual-Supervised Feature Learning Module (DMSFLM) to the baseline model. As shown in Table II, our final DMSFLN significantly improves the performances on both **R** and **HO** subsets. Under heavy occlusions (**HO**), the DMSFLN achieves a significant improvement of 9.76% **MR** compared to the baseline, demonstrating the effectiveness of the DMSFLN towards handling heavy occlusions. Also, as shown in Fig.3 in section III, the proposed DMSFLN outperforms the baseline detector with a remarkable improvement of 11.24% in Detection-Recall of heavily occluded pedestrians, which is very important and helpful for improving the final detection performance in heavy occlusion cases. Besides, our DMSFLN also improves the recall of partially occluded pedestrians compared with the baseline. We measure the detection speed using Titan X (Pascal) and cuDNN v6.0.21. For the VGA-resolution image with batch size 1 using a single GPU, baseline runs at 3.8 FPS, Bi-box runs at 2.5 FPS and our DMSFLN runs at 3.0 FPS. We note that detection efficiency is also an important aspect, especially in practical applications, therefore, how to improve the efficiency of our method is also one of our future work.

2) Influence of VB Branch: Because the visible body boxes contain more discriminative features of pedestrians, we propose to add an additional visible branch supervised by visible annotations. To evaluate the effectiveness of the proposed VB branch, we first add the VB branch based on the baseline model. We not only explore the effect of using classification task only in the VB branch but also explore the effect of using classification and regression tasks simultaneously. The results in Table II validate our analyses in Section IV-A. Specifically, our two branches network with only a classification task in the VB branch outperforms the baseline model on both **R** and **HO** sets by 1.14% **MR** and 3.07% **MR**, respectively.

3) Influence of Different Proposals Generation Policy: We then apply the Distribution-based Mutual-Supervised Feature Learning Module for two-branch model consisting of FB branch and VB branch (with classification task only) to demonstrate its effectiveness. We not only consider applying similarity loss only for positive samples but also consider applying similarity loss for both positive and negative samples.

TABLE III

COMPARISON (IN LOG-AVERAGE MISS RATES) OF DIFFERENT VISIBILITIES OF VISIBLE BODY BOXES

visibility	*R*	*HO*
[0,0.5]	10.5	40.27
[0.2,0.65]	10.23	39.19
[0.5,1.0]	10.04	38.95
[0.2,1.0]	9.92	38.83
[0,1.0]	**9.88**	**38.12**

However, negative proposals have low similarity since they usually contain vastly different features. As shown in Fig. 2, the positive training samples of FB and VB branches have strong similarities. Therefore, it is more reasonable to calculate similarity loss for positive samples. The comparisons in Table II confirm that only calculating the similarity between positive samples of two branches achieves the best results, and outperforms the baseline model by 2.04% **MR** and 9.76% **MR** on **R** and **HO** subsets, respectively.

4) Influence of Different Visibility of Visible Body Boxes: In the proposed network, the VB branch can generate a number of visible body boxes with different visibility according to the ground-truth of pedestrians. We need to choose the most suitable visibility range to be used for mutual-supervised loss computation. As listed in Table III, we conducted a series of experiments to compare the results of different visibility ranges of visible body boxes. The experimental results indicate that the best performances are achieved on both **R** and **HO** subsets when visible ratios are in the range of [0,1], which means that all the visible body boxes should be considered.

5) Influence of Thresholds for Collecting Training Samples: According to Eq.3 and Eq.4, α and β are two parameters controlling the thresholds for collecting the positive and negative training samples for two branches respectively, so we also conduct a series of experiments to explore their influence on the detection performance. Fig.8 depicts the log-average miss rates on **HO** subsets. The blue and orange lines represent the results of varying one threshold from small to large when fixed another threshold to 0.5. It is easy to see that as the threshold increases from 0.1 to 0.5, the generated positive samples contain more effective information of pedestrian, so the detector performs better on the Heavy Occlusion (**HO**) subset. As the threshold increases from 0.5, the generated positive samples will be more closely around the ground truth examples, but at the same time, the number of positive samples will be sharply reduced, which will cause the detector unable to make full use of discriminative features of pedestrians, resulting in the decline of detection results. Therefore, we choose $\alpha = 0.5$ and $\beta = 0.5$ in our experiments to get the best results.

6) Influence of Different Similarity Loss Functions: Next, to calculate the distance between two feature distributions, we compare several different methods, including Manhattan distance, Euclidean distance, Kullback-Leibler divergence, and Cosine Similarity. The Manhattan distance is the distance between two points measured along axes at right angles. Euclidean distance is the "ordinary" straight-line distance between two points in Euclidean space. The Kullback–Leibler divergence (also called relative entropy) is a measure of how one probability distribution is different from a second, reference probability distribution. And Cosine Similarity is to measure the difference between two individuals by cosine value of the angle between two vectors in vector space. As shown in Table IV, the Cosine Similarity method outperforms the other methods on both **R** and **HO** subsets.

TABLE IV

COMPARISON (IN LOG-AVERAGE MISS RATES) OF DIFFERENT SIMILARITY COMPUTATION METHODS

Method	*R*	*HO*
Manhattan distance	11.19	46.38
Euclidean distance	10.55	40.23
KL divergence	10.94	40.14
Cosine Similarity	**9.88**	**38.12**

TABLE V

COMPARISON (IN LOG-AVERAGE MISS RATES) OF DIFFERENT FUSION METHODS FOR INFERENCE

Method	*R*	*HO*
minimum	10.51	40.62
maximum	10.89	40.94
average	10.26	39.84
multiplication	**9.88**	**38.12**

7) Influence of Different Fusion Methods for Inference: In the inference phase, we also compared the results of four different fusion methods. Specifically, we maximize, minimize, add or multiply the confidence scores of FB branch and VB branch as the score of the final detection box. The results shown in Table V show that the multiplication of the two branches can get better detection results on both **R** and **HO** subsets.

D. State-of-the-Art Comparison on CityPersons

We compare our method with other recent state-of-the-art methods including Adapted FasterRCNN [7], Rep. Loss [6], OR-CNN [29], Bi-box [8], Adaptive NMS [32], FRCN+A+DT [41] and MGAN [10] on CityPersons dataset. We report the performance of DMSFLN and other methods on the validation set using the same ground-truth pedestrian examples and input scale during training. As shown in Table VI, the proposed DMSFLN outperforms all the other methods on both **R**, **Bare**, **Partial** and **HO** subsets. Some methods have not reported results on **HO** (visibility ranges is [20%, 65%]) subset, they only show results on **HO**(visibility ranges is [0, 65%]) subset. Therefore * represents the results with visibility in the range of [0, 65%]. Notably, our method reduces the **MR** of state-of-the-art results from 39.40% to 38.12% on **HO**, demonstrating the superiority of our proposed method in handling heavy occlusion cases. Fig.9 displays some example detections from Bi-box [8], MGAN [10], and our proposed DMSFLN on CityPersons val. set. The occlusion degrees of examples vary widely from partial to heavy occlusions. Our DMSFLN detects pedestrians with varying levels of occlusions more accurately.

E. State-of-the-Art Comparison on Caltech

We also evaluate our DMSFLN on Caltech [11] and compare it with state-of-the-art approaches. Table VII shows the comparison on Caltech test set under different occlusion

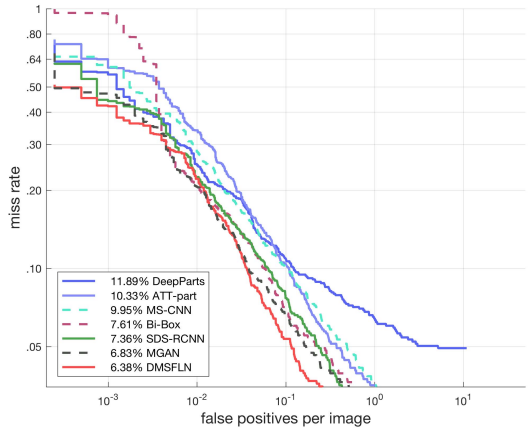
TABLE VI
COMPARISON (IN LOG-AVERAGE MISS RATES) WITH STATE-OF-THE-ART ON THE CITYPERSONS VAL. SET

Method	*Backbone*	*R*	*Bare*	*Partial*	*HO*	*FPS*
Adaptive Faster RCNN [7]	VGG-16	12.81	-	-	-	3.8
Rep.Loss [6]	ResNet-50	11.60	7.00	14.80	55.30*	-
Bi-box [8]	VGG-16	11.24	-	-	44.15	2.5
FRCN+A+DT [41]	VGG-16	11.10	6.90	11.20	44.30	-
OR-CNN [29]	VGG-16	11.00	5.90	13.70	51.30*	-
Adaptive NMS [32]	VGG-16	10.80	6.20	11.40	54.00*	-
MGAN [10]	VGG-16	10.50	-	-	39.40	-
Our DMSFLN	VGG-16	**9.88**	**5.23**	**10.37**	**38.12**	3.0

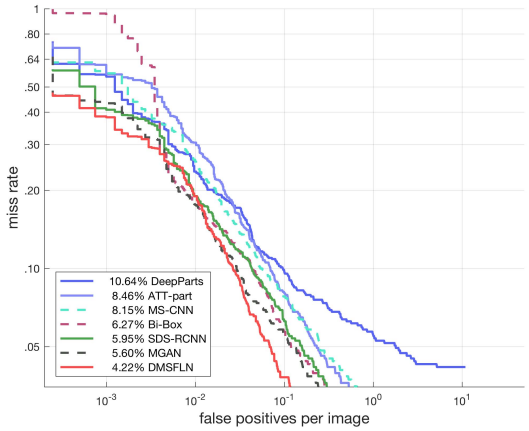
Fig. 9. Comparison of different detection methods on CityPersons dataset. The first row is the detection results of Bi-box [8], the second row is the detection results of MGAN [10] and last row is the final detections of our DMSFLN. The detected regions are cropped from the corresponding images for improved visualization. Note that all detection results are obtained using the same false positive per image (FPPI) criterion. Our DMSFLN accurately detects pedestrians with varying levels of occlusions.

subsets: **R**, **None**, **Partial**, **HO**, and **R** + **HO**. $*^O$ means the result is under the standard (old) test annotations, and $*^N$ means the result is under the new annotations provided by [42]. Compared to the existing methods, our DMSFLN achieves superior detection performance on all these subsets with a log-average miss rate of 6.38%, 4.22%, 12.98%, 37.45%, 13.12%, and 2.68%, respectively. Fig.10 presents the comparison over the wide range of false positives per image metric (FPPI). Numbers in legends refer to log-average miss rates on different subsets on Caltech. Fig. 11 depicts some detection examples of our proposed DMSFLN and Bi-box [8] and MGAN [10]. Our proposed DMSFLN provides more accurate detections under different occlusion scenarios. We also found on the Caltech's official website[1] that the results of ADM [43] and TLL-TFA [44] on the **HO** subset are 30% and 29%, respectively, which are truly better results. However, to achieve such good results, both approaches utilize additional multi-frame or temporal information to help with heavy occlusion cases, while other approaches (including ours) focus on static detection in single frame. More specifically, ADM [43] consists of two main stages: multi-layer feature representations and initial pedestrian proposals are obtained through the first stage, in the second stage, these results are sent to an active detection stage where a localization policy is designed to produce the final detections by executing sequences of coordinate transformation actions. TLL-TFA [44] try to improve the detection quality by exploiting multi-frame temporal feature aggregation method. These two approaches all employ a recurrent neural network (RNN) with LSTM units. Thus, their results are not included in Table VII for fair comparisons.

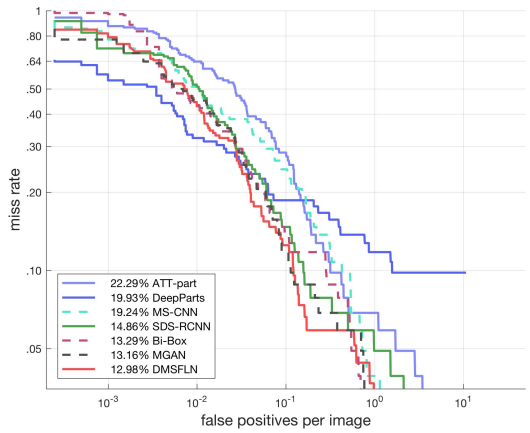
[1] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/rocs/UsaTestRocs.pdf



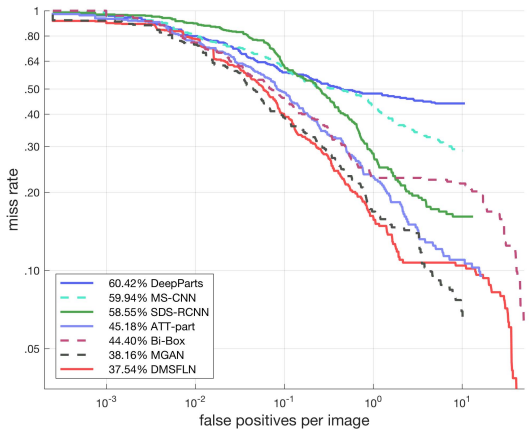
(a) Reasonable



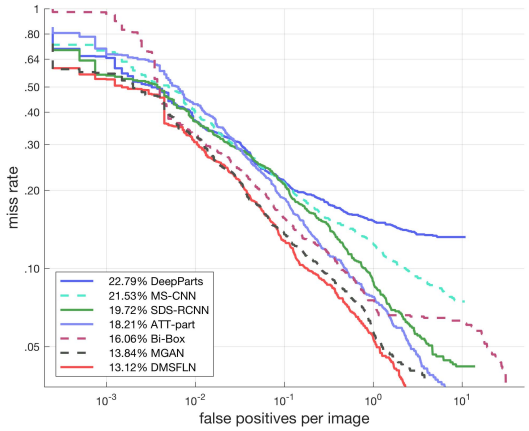
(b) None



(c) Partial



(d) Heavy



(e) R+HO

Fig. 10. Comparison of our approach with other competitive methods on different subsets on Caltech. Numbers in legends refer to log-average miss rates.

F. State-of-the-Art Comparison on CrowdHuman

Table VIII shows the performance of the baseline, AdaptiveNMS [32], Rep.Loss [6], R^2NMS [39] and the proposed DMSFLN on CrowdHuman [12] validation subsets. The result with * stands for our re-implemented baseline, which outperforms the original baseline in CrowdHuman [12] by 3.88% **MR** and 0.54% **AP**, respectively. Therefore, our re-implemented baseline is strong enough to validate the effectiveness of our proposed methods. Compared to our baseline, the proposed DMSFLN achieves significant improvements of 2.95% **MR** and 3.69% **AP** respectively, demonstrating its effectiveness. As for the comparison with other state-of-the-art methods, our DMSFLN outperforms Adaptive NMS [32]

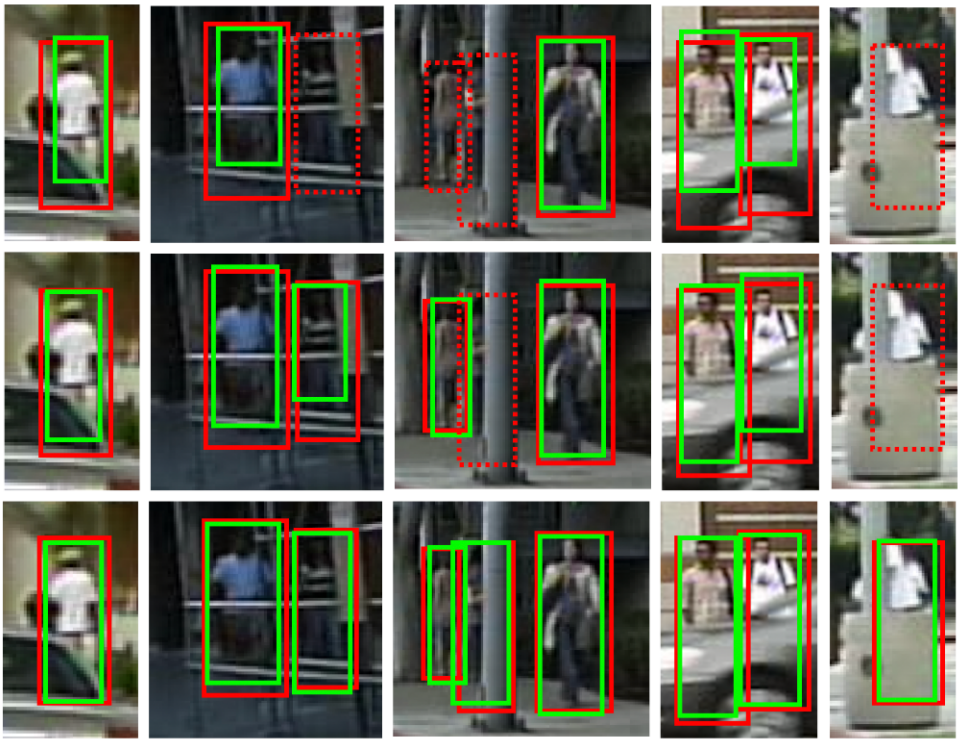


Fig. 11. Comparison of different detection methods on Caltech dataset. The first row is the detection results of Bi-box [8], the second row is the detection results of MGAN [10] and the last row is the final detections of our DMSFLN. All detection results are obtained using the same false positive per image criterion. The solid red boxes denote the ground-truth, dashed red boxes represent the missed detections and the green boxes present detection results. For better visualization, the detected regions are cropped from the corresponding images.

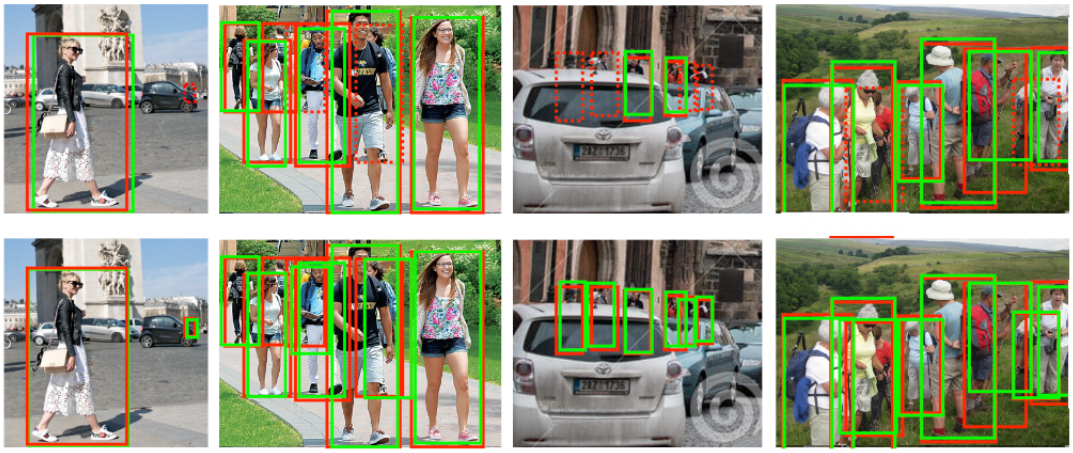


Fig. 12. Comparison of different detection methods on CrowdHuman dataset. The top row is the detection results of the baseline, the bottom row is the detection results of our DMSFLN. All detection results are obtained using the same false positive per image criterion. The solid red boxes denote the ground-truth, dashed red boxes represent the missed detections and the green boxes present detection results. For better visualization, the detected regions are cropped from the corresponding images.

and Rep.Loss [6] and achieves very close performance to R^2NMS [39]. It is worth mentioning that CrowdHuman [12] dataset contains many crowded scenes with around 23 persons per-image, thus it is more suitable to evaluate a special type of occlusion that pedestrians are occluded by each other. And R^2NMS [39] specially designed a novel NMS method for pedestrian detection in such crowded scenes, while our DMSFLN still adopts the standard NMS method, as most other approaches did. Cooperating with new NMS method in our proposed approach would be an interesting future work to further improve the performance under various occlusion conditions. Qualitative detection performance on this dataset is shown in Fig. 12.

G. State-of-the-Art Comparison on CUHK

We also compare with the state-of-the-art methods on the CUHK occlusion [13] dataset. The models are pretrained on CityPersons [7] and Caltech [11], respectively and then evaluated on CUHK occlusion [13] dataset. As shown in Table IX,

Fig. 13. Comparison of different detection methods on CUHK occlusion dataset. The first row is the detection results of Bi-box [8], the second row is the detection results of MGAN [10] and the last row is the final detections of our DMSFLN. All detection results are obtained using the same false positive per image criterion. The solid red boxes denote the ground-truth, dashed red boxes represent the missed detections and the green boxes present detection results. For better visualization, the detected regions are cropped from the corresponding images.

TABLE VII

COMPARISON (IN LOG-AVERAGE MISS RATES) WITH STATE-OF-THE-ART ON CALTECH TEST SET

Method	R^O	$None^O$	$Partial^O$	HO^O	$R+HO^O$	R^N
DeepParts [27]	11.89	10.64	19.93	60.42	22.79	12.90
MS-CNN [3]	9.95	8.15	19.24	59.94	21.53	8.08
ATT-part [9]	10.33	8.46	22.29	45.18	18.21	8.11
SDS-RCNN [5]	7.36	5.95	14.86	58.55	19.72	6.44
OR-CNN [29]	-	-	-	-	-	4.10
Rep.Loss [6]	-	-	-	-	-	4.00
Bi-box [8]	7.61	6.27	13.3	44.40	16.06	-
MGAN [10]	6.83	5.6	13.16	38.16	13.84	-
Our DMSFLN	**6.38**	**4.22**	**12.98**	**37.54**	**13.12**	**2.68**

TABLE VIII

COMPARISON (IN LOG-AVERAGE MISS RATES) WITH STATE-OF-THE-ART COMPARISON ON CROWDHUMAN VAL. SET

Method	*MR*	*AP*
Baseline [12]	50.42	84.95
Baseline*	46.54	85.49
Adaptive NMS [32]	49.73	84.71
Rep.Loss [6]	45.69	85.64
R^2NMS [39]	**43.35**	**89.29**
Our DMSFLN	43.59	89.18

TABLE IX

PERFORMANCE COMPARISON WITH THE BASELINE DETECTOR ON THE CUHK OCCLUSION DATASET

Method	*MR (CityPersons)*	*MR (Caltech)*
Baseline	8.29	7.29
Bi-box [8]	7.94	6.18
MGAN [10]	7.23	5.97
Our DMSFLN	**6.12**	**4.73**

the proposed method outperforms the second best result by 1.11% and 1.24% on **MR**, validating that our approach is effective for occlusions in various scenes. In addition, some visual comparisons of Bi-box [8], MGAN [10] and our DMSFLN on CUHK [13] occlusion dataset are displayed in Fig 13.

VI. CONCLUSION

This paper has presented a novel Distribution-based Mutual-Supervised Feature Learning Network (DMSFLN) for heavily occluded pedestrian detection. A new Distribution-based Mutual-Supervised Feature Learning Module is designed to minimize the similarity loss of feature distributions between full body and visible body to enhance the feature representations of heavily occluded pedestrians. Our DMSFLN consists of two branches, one is the standard full body detection branch, the other is the additional visible body classification branch. Besides, our two branches are supervised by full body annotations and visible body annotations, respectively. Extensive experiments demonstrate that our proposed network outperforms other state-of-the-art approaches, confirming the effectiveness of the proposed method.

1) Future Work: Our method performs better in the case of partial occlusion, but the performance on heavy occlusion, especially in crowded scenes (pedestrians occluded by other pedestrians), still could be further improved. We found that in crowded scenes, the candidate bounding boxes of nearby (highly overlapped) pedestrians are easily removed by mistake due to the standard post-processing (non-maximum suppression) step. Therefore, for the future extension, we plan to focus on the perspective of post-processing and improve the pedestrian detection performance in crowded scenes. Moreover, we plan to improve the inference speed of our approach to make it more suitable for the practical applications.

REFERENCES

[1] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5420–5428.

[2] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 618–634.

[3] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 354–370.

[4] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3127–3136.

[5] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4950–4959.

[6] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[7] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

[8] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–151.

[9] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[10] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[12] S. Shao *et al.*, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*. [Online]. Available: http://arxiv.org/abs/1805.00123

[13] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3258–3265.

[14] Y. He, C. Zhu, and X.-C. Yin, "Mutual-supervised feature modulation network for occluded pedestrian detection," in *Proc. 25th Int. Conf. Pattern Recognit.*, Jan. 2020, pp. 8453–8460. [Online]. Available: https://arxiv.org/abs/2010.10744

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[19] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 732–747.

[20] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 966–974.

[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[23] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 443–457.

[24] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7231–7240.

[25] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1505–1512.

[26] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.

[27] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1904–1912.

[28] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3486–3495.

[29] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 637–653.

[30] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 878–885.

[31] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.

[32] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6459–6468.

[33] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468.

[36] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. ICCV*, 2007, pp. 1–8.

[37] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 794–801.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[39] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10750–10759.

[40] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[41] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9557–9566.

[42] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

[43] X. Zhang, L. Cheng, B. Li, and H.-M. Hu, "Too far to see? Not really!—Pedestrian detection with scale-aware localization policy," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3703–3715, Aug. 2018.

[44] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 536–551.

Ye He received the B.Sc. degree in computer science from the University of Science and Technology Beijing, China, in 2019, where she is currently pursuing the master's degree with the Department of Computer Science and Technology. Her research interests include object detection and pedestrian detection.

Chao Zhu received the bachelor's degree in automation from Xidian University, Xi'an, China, in 2005, the master's degree in system engineering from Xi'an Jiaotong University, Xi'an, in 2008, and the Ph.D. degree in computer science from the Ecole Centrale de Lyon, France, in 2012. He was a Post-Doctoral Fellow with the Multimedia Information Processing Laboratory (MIPL), Institute of Computer Science and Technology (ICST), Peking University, Beijing, from 2013 to 2015. He is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. He has published more than 30 articles in refereed international journals and conference proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, *Pattern Recognition*, AAAI, ICME, ICMR, and ICPR. His research interests include object detection and recognition, pedestrian detection, person re-identification, and image/video classification.

Xu-Cheng Yin (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the University of Science and Technology Beijing, China, in 1999 and 2002, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2006.

He was a Visiting Professor with the College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA, for three times (from January 2013 to January 2014, from July 2014 to August 2014, and from July 2016 to September 2016). He is currently a Full Professor and the Director of the Pattern Recognition and Information Retrieval Laboratory, Department of Computer Science and Technology, University of Science and Technology Beijing. He has published more than 80 research articles, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Pattern Recognition*, CVPR, ACM MM, ACM SIGIR, and IJCAI. His research interests include pattern recognition, computer vision, and document analysis and recognition. From 2013 to 2019, his team had won the first place of a series of text detection and recognition competition tasks for 15 times in ICDAR Robust Reading Competition.