OAF-Net: An Occlusion-Aware Anchor-Free Network for Pedestrian Detection in a Crowd

Qiming Li, Yijing Su, Yin Gao, Feng Xie, and Jun Li, *Member, IEEE*

Abstract— **Although pedestrian detection has achieved promising performance with the development of deep learning techniques, it remains a great challenge to detect heavily occluded pedestrians in crowd scenes. Therefore, to make the anchor-free network pay more attention to learning the hard examples of occluded pedestrians, we propose a simple but effective Occlusion-aware Anchor-Free Network (namely OAF-Net) for pedestrian detection in crowd scenes. Specifically, we first design a novel occlusion-aware detection head, which includes three separate center prediction branches combining with the scale and offset prediction branches. In the detection head of OAF-Net, occluded pedestrian instances are assigned to the most suitable center prediction branch according to the occlusion level of human body. To optimize the center prediction, we accordingly propose a novel weighted Focal Loss where pedestrian instances are assigned with different weights according to their visibility ratios, so that the occluded pedestrians are up-weighted during the training process. Our OAF-Net is able to model different occlusion levels of pedestrian instances effectively, and can be optimized towards catching a high-level understanding of the hard training samples of occluded pedestrians. Experiments on the challenging CityPersons, Caltech, and CrowdHuman benchmarks sufficiently validate the efficacy of our OAF-Net for pedestrian detection in crowd scenes.**

Index Terms— **Pedestrian detection, occlusion-aware, anchor-free, crowd scenes.**

I. Introduction

PEDESTRIAN detection plays an important role in intelligent transportation systems with various applications, such as automotive systems [1], [2], vehicle safety [3], and video surveillance [4], [5]. Because crowd scenes happen frequently in transportation systems, pedestrian detection in the real-world scenes, where the density of people is high (*e.g.*, signalized intersections [6], [7], airports [8], train stations [9], and shared spaces [10]), has attracted increasing attentions. Over the past few years, coupled with the success of deep learning techniques, the performance of pedestrian detection has been greatly improved with the development of Convolutional Neural Networks (CNNs). However, due to the heavy occlusion issue, detecting pedestrians in crowd scenes still remains a very challenging task. Take the popular Caltech [11] benchmark for example, some state-of-the-art pedestrian detectors (*e.g.*, CSP [12], RepLoss [13], and OR-CNN [14]) could achieve a miss rate of around 4% MR^{-2} at 0.1 false positives per image (FPPI) for partially occluded pedestrians, but their performances drop dramatically (about 40% MR^{-2} at 0.1 FPPI) while the heavy occlusions are present (refer to Fig. 5 for more details).

In recent years, many attempts based on CNNs [15]–[23] have been made to handle the occlusion issue for pedestrian detection in crowd scenes. The common technique of these methods is to design an additional branch in anchor-based networks to generate poor-quality proposals for occluded pedestrians. However, there are some limitations to these methods. First, the design of the hyper-parameters (*e.g.*, size, aspect ratio, and number) for the anchor-based detectors have a great impact on the detection performance. Second, the added branch in the network would greatly increase the memory and time requirements for detection. Third, most of the generated candidate proposals from the anchor-based network are labelled as negative samples, it would results in the class imbalance problem. On the other hand, some methods develop the post-processing Non-Maximum Suppression (NMS) algorithms [24]–[27] to tackle the crowd occlusion issue. Although some promising results can be achieved, these pedestrian detectors using the improved NMS strategies are still extremely restricted by the classification results of bounding boxes which are obtained from the backbone network.

Compared to the anchor-based network, the anchor-free network directly detects pedestrians from the input images without enumerating a large number of candidate proposals, and can be trained in an end-to-end fashion. In addition, the training loss function used in the anchor-free network (*e.g.*, Focal Loss [28]) is effective to combat the class imbalance issue. However, the state-of-the-art anchor-free detectors (*e.g.*, CSP [12] and APD [29]) also still struggle under severe occlusions, and limited anchor-free pedestrian detectors have been specially designed to deal with the occlusion issue. Overall, the anchor-free network has a better potential than the anchor-based network to be applied to the study of occlusion handling for pedestrian detection.

Manuscript received 15 September 2021; revised 20 January 2022 and 18 March 2022; accepted 26 April 2022. Date of publication 6 May 2022; date of current version 7 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62102394, Grant 61906178, and Grant 62001452; in part by the Natural Science Foundation of Fujian Province of China under Grant 2020J05083; and in part by the Science and Technology Program of Quanzhou under Grant 2020C071. The Associate Editor for this article was H. G. Jung. *(Corresponding author: Jun Li.)*

Qiming Li, Yijing Su, Yin Gao, and Jun Li are with the Laboratory of Robotics and Intelligent Systems, Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou, Fujian 362216, China, and also with Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou, Fujian 350108, China (e-mail: qimingli@fjirsm.ac.cn; suyj@fjirsm.ac.cn; yingao@fjirsm.ac.cn; junli@fjirsm.ac.cn).

Feng Xie is with the Department of Traffic and Assistance, Institute of Automation and Communication, 39106 Magdeburg, Germany (e-mail: feng.xie@ifak.eu).

Digital Object Identifier 10.1109/TITS.2022.3171250

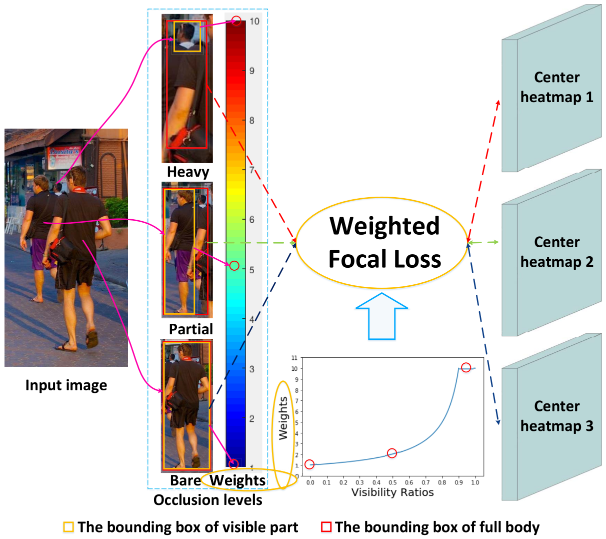


Fig. 1. Illustration of the motivation of our OAF-Net.

However, if following the traditional occlusion handling strategies to design the anchor-free network, the learned part detectors are still not able to cover the large diversity of poses and occlusions of pedestrians in real-world scenarios. Therefore, this paper proposes to optimize the network by selecting the most suitable center prediction branch for each pedestrian instance in the designed detection head according to the Intersection-over-Union (IoU) overlap between the visible part and the full body (*i.e.*, the visibility ratio). Specifically, we propose a simple but effective Occlusion-aware Anchor-Free Network (namely OAF-Net), where only two additional center prediction branches are introduced for the occlusion-aware mechanism and a novel weighted Focal Loss is proposed accordingly to optimize the center prediction. Unlike most traditional occlusion handling detectors, our detector does not need to learn part detectors of human bodies by designing complex frameworks of deep networks, thus too many memory and time requirements can be avoided. And the weighted novel Focal Loss can also be easily extended into existing anchor-free frameworks for occlusion handling. The general concept of our ideal is presented in Fig. 1. As in the typical anchor-free detector, pedestrian detection is formulated as a center and scale prediction task in the detection head by conducting convolution on the concatenation of the feature pyramids. But differently, to better handle the occluded pedestrians, we introduce three separate center prediction branches so that the occluded pedestrian instances can be assigned to the most suitable center heatmap according to the visibility ratio during the training process. Accordingly, to optimize the center prediction in the occlusion-aware detection head, we further propose a novel weighted Focal Loss where pedestrian instances are weighted according to their occlusion levels. By combining the occlusion-aware head and weighted Focal Loss, our OAF-Net can be optimized towards catching a high-level understanding of the hard training samples of occluded pedestrians, and is able to model different occlusion levels of pedestrian instances effectively. At the inference stage, we select the coordinates with maximum scores among the three center heatmaps as the final center locations of detected pedestrians, and the bounding boxes are generated with the obtained centers and their corresponding scales.

In summary, the main contributions of this paper are three-fold. First, according to the occlusion level of the human body, a novel occlusion-aware detection head is designed for the anchor-free detector, where the occluded pedestrian instances are assigned to the most suitable center prediction branch, and the proposed network can be simply and effectively trained in an end-to-end fashion. Second, a novel weighted Focal Loss is proposed to optimize the center prediction branch in the detection head, where the occluded pedestrian instances are up-weighted, so that the network can pay more attention to learning the hard examples of occluded pedestrians. This paper also further investigates the influence of weighted Focal loss with different settings of hyper-parameter on the detection performance of occlusion handling. Third, experiments on three challenging benchmarks in crowd scenes (*i.e.*, CityPersons [30], Caltech [11], and CrowdHuman [31]) show that our OAF-Net outperforms state-of-the-art pedestrian detectors with a large performance gap for heavily occluded pedestrians.

The remainder of our paper is organized as follows. Sec. II provides an overview of the related work. Sec. III presents the details of our OAF-Net. Sec. IV presents extensive experiments to validate the efficacy of OAF-Net on three pedestrian detection benchmarks. Sec. V gives the conclusion of this paper.

II. Related Work

A. Generic Object Detection

Early generic object detectors [32], [33] classify/detect object instances based on the sliding window scheme using handcrafted features. In recent years, object detection has been dominated by deep learning techniques, which can be broadly divided into two categories: anchor-based detectors [16], [34]–[40] and anchor-free detectors [12], [41]–[51]. Anchor-based detectors first pre-define a set of anchors at different scales and then classify these candidate anchor boxes to find the object instances. Most existing anchor-based detectors either adopt the one-stage [16], [34], [40] or two-stage [35]–[39] strategy in the network. Faster R-CNN [39] is one of the most representative two-stage frameworks, where CNNs are utilized for both proposal generation and classification. On the contrary, the one-stage anchor-based detectors (*e.g.*, SSD [34] and ALFNet [40]) discard the separate anchor boxes generation step and directly predict the default anchors into object bounding boxes.

Anchor-free detectors directly find and locate objects from the input images without enumerating a large number of candidate proposals. As one of the most popular anchor-free detectors, YOLO [41] utilizes points around the center of object to predict the bounding boxes. Afterwards, many key-points based anchor-free detectors [42]–[47] are proposed to detect objects by predicting the key points of bounding boxes. Recently, some pedestrian detectors [12], [29], [47], [51] based on the center point prediction have also emerged. For example, CSP [12] transforms the task of pedestrian detection into the predictions of centers and their corresponding scales. To address the scale variation issue, MPAF-Net [51] proposes

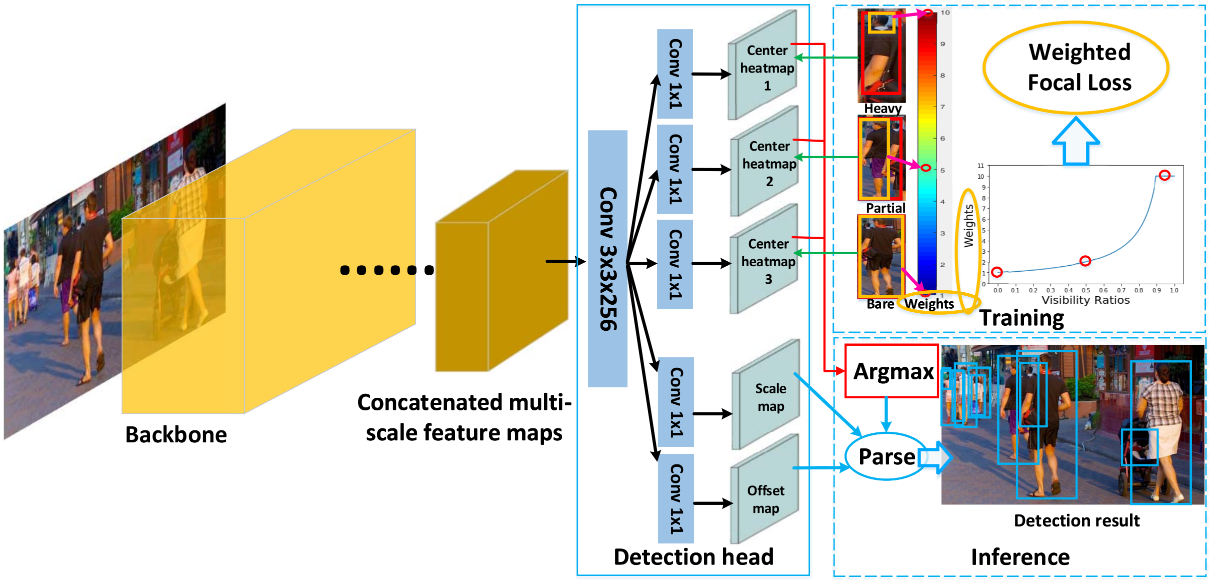


Fig. 2. The overall architecture of our OAF-Net.

a Conditional Random Fields (CRFs) based message passing mechanism to refine the multi-scale features obtained from the backbone network for the center prediction. Apart from the center, scale, and offset prediction in the detection head, APD [29] further integrates an attribute-aware feature prediction branch which encodes the semantic differences among pedestrians, and a novel attribute-aware NMS is introduced to adaptively reject the false-positive proposals and distinguish the individuals based on the obtained semantic information. Note that MPAF-Net, APD, and our OAF-Net all belong to the center point based anchor-free detectors. However, there are some significant differences between them: (1) The motivation of this paper is to handle the occlusion issue by designing an occlusion-aware detection head where the most suitable center prediction branch is selected for each occluded pedestrian instance. However, the main focus of MPAF-Net is to alleviate the scale variation issue by capturing complementary information between multi-scale features. APD focuses on differentiating the bounding boxes of individuals from each other by introducing a pedestrian-oriented attribute feature prediction branch in the detection head, based on which an attribute-aware NMS is proposed as a post-processing step to generate final detection results. (2) Our OAF-Net introduces three separate center prediction branches combining with the scale and offset prediction branches in the occlusion-aware detection head, whereas MPAF-Net and APD only use one center prediction branch. (3) A novel weighted Focal Loss is designed accordingly for each center prediction branch to enforce our network to catch a high-level understanding of different occlusion levels of pedestrians during training, while MPAF-Net and APD directly apply the traditional Focus Loss to optimize the center prediction.

B. Occlusion Handling in Pedestrian Detection

Occlusion is one of the most challenging issues for pedestrian detection in crowd scenes. The widely used strategy for occlusion handling is to learn and integrate a set of part-based detectors [15]–[23], [52]. For example, Franken [52] trains an exhaustive set of occlusion-specific detectors based on the integral channel features classifier. DeepParts [15] constructs a part pool whose features are learned by the fine-tuning CNNs, and explores the ensemble of part detectors during testing to handle occlusions. Noh *et al.* [16] integrate the average grid post-refinement classifiers and part confidence scores into the single-stage pedestrian detectors. Ouyang *et al.* [19]–[21] propose a series of deep models by designing specific convolutional layers for different part models and learning their mutual visibility relationships. Zhou and Yuan [22], [23] propose two kinds of part detectors to handle occlusions with the multi-label learning and bi-box regression strategies. Moreover, another alternative strategy for occlusion handling in deep CNNs is to design some special losses like RepLoss [13] and OR-CNN [14] to discriminate the occlusion levels. Besides that, many variants of NMS (*e.g.*, Adaptive-NMS [26], Soft-NMS [27], R^{2}NMS [24], and NOH-NMS [25]) are proposed to better refine the overlapped bounding boxes for the occluded pedestrians. To validate the robustness of pedestrian detectors for occlusion handling, two large-scale pedestrian datasets (*i.e.*, CityPersons [30] and CrowdHuman [31]) are collected in recent years. Though many efforts have been made, these detectors mentioned above are still far from being practical to deal with heavily occluded pedestrians in crowd scenes.

III. OAF-Net

A. Architecture of OAF-Net

Crowd occlusions occur frequently in real-world scenarios, limited existing anchor-free pedestrian detectors are reported to handle the issue, and their performances for the heavy occlusion are still far from being satisfactory. Therefore, we propose a simple but effective OAF-Net to handle the occlusion issue for pedestrian detection in a crowd. The overall architecture of our OAF-Net is illustrated in Fig. 2. Under the anchor-free detection framework [12], we introduce three separate branches in the detection head to predict the center heatmaps of pedestrians with different occlusion levels. During the training process, we divide the pedestrian instances in input images into three occlusion levels (*i.e.*, Bare occlusion, Partial occlusion, and Heavy occlusion) according to their visibility ratios, and each instance is assigned to the corresponding center prediction branch according to its occlusion level (*i.e.*, visibility ratio). Furthermore, to optimize the center prediction branches in the detection head, we further propose a novel weighted Focal Loss, where pedestrian instances are assigned with different weights according to their visibility ratios as well, so that the occluded pedestrian instances are up-weighted during training. At the inference stage, the maximum confidence scores between the three obtained center heatmaps are selected as the final scores of centers of detected pedestrians, and the final detection results can be parsed with the obtained centers and their corresponding scales.

B. Feature Extraction

The backbone of our OAF-Net can be truncated from the ImageNet [53] pre-trained standard networks like ResNet-50 [54], DLA-34 [55], and HRNet-W32 [56]. This paper adopts HRNet-W32 as the backbone network for the subsequent pedestrian detection task. There is total of four stages in HRNet-W32 with the down-sampling rate of 4, 8, 16, and 32 respectively. Note that the combination of deeper feature maps is of great importance for superior performance [12], thus we concatenate all the multi-scale feature maps in the final stage of HRNet-W32 for the feature extraction. We utilize the deconvolution layer to resize these multi-scale feature maps at the last stage to the same resolution (*i.e.* $W/r \times H/r$, where input image $I \in \Re^{W \times H \times 3}$ and $r = 4$ is the down-sampling rate). It has been proved in [12] and [57] that $r = 4$ gives the best detection performance, because a smaller r introduces more computational burdens, while a larger r means coarser feature maps that struggle on accurate localization. Afterwards, these resized feature maps are concatenated and then fed into the pedestrian detection head. It is worthy to point out that more complicated feature fusion strategies like [58]–[60] may further improve the detection performance, but it is not in the scope of our paper.

C. Occlusion-Aware Pedestrian Detection Head

Unlike traditional anchor-free detectors, we design five branches to predict the feature maps (*i.e.*, three center heatmaps, one scale map, and one offset map) in the detection head. Specifically, we first attach a 3×3 convolution layer with 256 channels to reduce the channel dimension of the concatenated multi-scale feature maps Φ_{det}, then append five sibling 1×1 convolution layers to respectively generate the desired feature maps which have the same size as Φ_{det} (*i.e.*, $W/r \times H/r$). During the training process, pedestrian instances are assigned to the corresponding center prediction branch according to the visibility ratio $\mathbb{R}$ which can be defined as

$$\mathbb{R}_n = IOU(V_n, P_n) = \frac{Area(V_n \cap P_n)}{Area(V_n \cup P_n)}, \qquad (1)$$

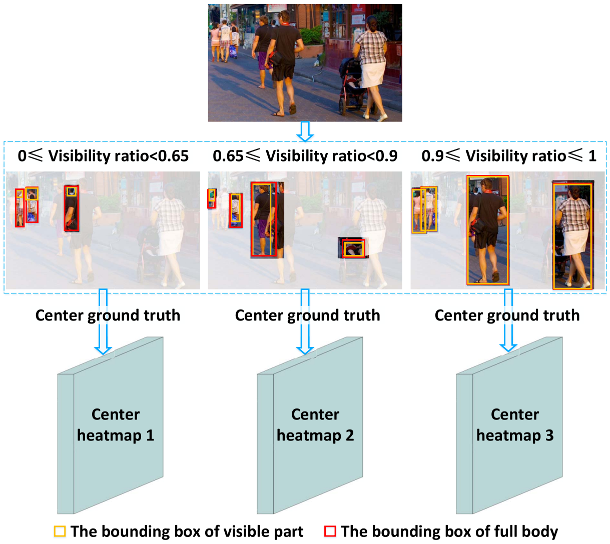


Fig. 3. Illustration of center prediction branches in the occlusion-aware pedestrian detection head.

where $V_n = (V_n^x, V_n^y, V_n^w, V_n^h)$ and $P_n = (P_n^x, P_n^y, P_n^w, P_n^h)$ denote the visible part and full body regions of the n−th pedestrian instance in the image; $IOU(V, P)$ denotes the intersection over the union of visible part V and full body P, which is usually utilized as the occlusion handling information in many pedestrian detectors [15], [17], [24], [26], [29], [50], [61], [62], and some pedestrian datasets [11], [30], [31], [63] which provide the annotations of visible parts of pedestrians have been proposed in recent years.

As shown in Fig. 3, following the strategies in [13] and [14], we divide the pedestrian instances in training images into Bare (occlusion $<$ 10%, *i.e.* visibility ratio $\geqslant$ 90%), Partial (10% $\leqslant$ occlusion $<$ 35%), and Heavy (35% $\leqslant$occlusion) occlusion levels. Furthermore, we design a novel occlusion-aware detection head so that the obtained divided training samples can be fed into the corresponding center prediction branch. The feature map of each branch can be updated at the most suitable occlusion level.

D. Optimization Target

1) Center Loss: Focal Loss [28] and its variants are the widely used strategies to handle the imbalances between positives and negatives for general object detection. However, to the best of our knowledge, there has been no published work to study the impact of weighted Focus Loss for occluded pedestrian instances in pedestrian detection. Therefore, to fully optimize the designed center prediction branches in the detection head, we propose a novel weighted Focal Loss to train the target L_c of center location prediction by taking the visibility ratios of pedestrians into account:

$$L_c = -\frac{1}{N}\sum_{i=1}^{W/r}\sum_{j=1}^{H/r} \omega_{ij}^{\eta}\alpha_{ij}\left(1-\hat{p}_{ij}\right)^{\gamma}\log\left(\hat{p}_{ij}\right), \qquad (2)$$

where N denotes the number of pedestrian instances in an image. The newly defined hyper-parameter η is set to 1, which will be discussed in ablation studies. ω is used to up-weight

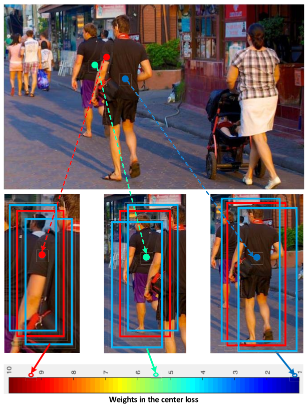


Fig. 4. Illustration of the re-weighted occluded pedestrian instances in the center loss during training. The red box indicates the ground truth annotation, and the blue box indicates the hard training samples around the ground truth.

occluded pedestrian instances during the training process, it is defined as

$$\omega_{ij} = \begin{cases} 10 & \text{if } \mathbb{R}_n \leqslant 0.1 \textit{ and } (i,j) \in P_n, \\ \dfrac{1}{\mathbb{R}_n} & \text{if } \mathbb{R}_n > 0.1 \textit{ and } (i,j) \in P_n, \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

$\hat{p}_{ij}$ and α_{ij} can be defined as

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij} = 1 \\ 1 - p_{ij} & \text{otherwise,} \end{cases}$$
$$\alpha_{ij} = \begin{cases} 1 & \text{if } y_{ij} = 1 \\ \left(1 - M_{ij}\right)^{\beta} & \text{otherwise,} \end{cases} \tag{4}$$

where $y_{ij} = 1$ denotes that a pedestrian's center falls the location (i, j). Let $p_{ij} \in [0, 1]$ denote the network's predicted distribution indicating whether the location (i, j) is the center of a pedestrian or not. Like in [12], to combat the positive-negative imbalance issue, the Gaussian mask M_{ij} is applied to reduce the contribution of negative samples to the full training loss. The hyper-parameters γ and β are set to 2 and 4 as suggested in [28] and [43].

As seen in Fig. 4, the occluded pedestrian instances are re-weighted in the designed center loss of our OAF-Net during training. More specifically, the hard training samples around the center of the heavily occluded pedestrian are assigned with high weights according to the occlusion level, and the weights of the easy training samples around the center of the non-occluded pedestrian stay the same.

2) Scale Loss: We formulate the prediction of scale map as a regression task via the Smooth L1 loss [64]:

$$L_{\mathrm{s}} = -\frac{1}{N}\sum_{n=1}^{N} \text{Smooth } L1\,(s_n, \bar{s}_n), \tag{5}$$

where s_n and $\bar{s}_n$ denote the prediction and the ground truth of $n-$th pedestrian's scale. For the CrowdHuman dataset, we predict both the height and width of each pedestrian instance, so the scale ground truth $\bar{s}_n$ can be defined as $(log(h_n), log(w_n))$, where h and w denote the height and width of a pedestrian. To reduce the ambiguity [12], the pedestrian's center and negatives around it with a radius of 2 are all assigned with $(log(h_n), log(w_n))$. But for the other datasets, only the height prediction is considered, and the final bounding box of a pedestrian can be generated with the height and the pre-defined aspect ratio.

3) Offset Loss: The prediction of offset map is also formulated as a regression task via the Smooth L1 loss:

$$L_{\mathrm{o}} = -\frac{1}{N}\sum_{n=1}^{N} \text{Smooth } L1\,(o_n, \bar{o}_n), \tag{6}$$

where o_n and $\bar{o}_n$ denote the prediction and the ground truth of $n-$th pedestrian's offset. The predicted offset map is used to slightly adjust the center locations of pedestrians. Because the real location of $n-$th pedestrian's center (x_n, y_n) is mapped to the location $(\lfloor \frac{x_n}{r} \rfloor, \lfloor \frac{y_n}{r} \rfloor)$ in outputs, the offset ground truth can be defined as

$$\bar{o}_n = (\frac{x_n}{r} - \left\lfloor \frac{x_n}{r} \right\rfloor, \frac{y_n}{r} - \left\lfloor \frac{y_n}{r} \right\rfloor). \tag{7}$$

4) Full Training Loss: To sum up, the full training loss function of our network is

$$L_{ped} = \lambda_{\mathrm{c}} \sum_{l=1}^{3} L_{\mathrm{c}}^{l} + \lambda_{\mathrm{s}} L_{\mathrm{s}} + \lambda_{\mathrm{o}} L_{\mathrm{o}}, \tag{8}$$

where λ_{c}, λ_{s}, and λ_{o} are experimentally set to 0.01, 1, and 0.1, respectively. L_{c}^{l} denotes the training loss of $l-$th center prediction branch. By combining the three center losses, we exploit the proposed occlusion-aware mechanism to enforce our network to catch a high-level understanding of different occlusion levels of pedestrians in crowd scenes.

E. Inference

During testing, the inference is as simple as forwarding an image through the network to generate the five desired feature maps in the detection head. But before parsing the bounding boxes of pedestrians, we need to take the element-wise maximum of the three center heatmaps, and then a confidence threshold of 0.01 is used to filter out the center predictions with the low confidence in the obtained maximum results. These top center predictions with their corresponding scales are combined to generate the bounding boxes of detected pedestrians, and the predicted offset map is used to adjust the center locations before remapping. Finally, NMS with a threshold of 0.5 is adopted to yield the final detection results.

IV. Experiments

A. Experimental Settings

1) Datasets and Evaluation Metrics: To fully validate the robustness of the proposed OAF-Net on occluded pedestrians, we perform extensive experiments on three challenging pedestrian datasets: CityPersons [30], Caltech [11], and CrowdHuman [31]. CityPersons is a recently released challenging pedestrian detection dataset, which contains 5,050 images (2,975 for training, 500 for validation, and 1,575 for testing), and provides both bounding box annotations of full bodies and visible parts of pedestrians in the training images. CityPersons is very suitable to validate the efficacy of detectors for occluded pedestrians, because it includes about 70% of the pedestrian instances depicting various occlusion levels. As one of the predominant and representative pedestrian datasets, Caltech comprises approximately 2.5 hours of auto-driving video which is divided into 42,782 training images and 4,024 testing images. We perform all the experiments on Caltech using the refined annotations provided in [65]. CrowdHuman is recently presented to specifically target crowd scenes, which collects 15,000, 4,370, and 5,000 images from the Internet for training, validation, and testing subsets respectively. There are totally 470k pedestrian instances in the training and validation subsets, and approximately 22.6 persons per image, which is of much higher crowdedness compared with the previous datasets. The bounding box annotations of heads, visible parts, and full bodies are provided for each pedestrian instance, therefore CrowdHuman is also very suitable to validate the robustness of pedestrian detectors for occlusions handling.

Following the common practice, the log Miss Rate (MR) averaged over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ is exploited as the evaluation metric for all three datasets. In addition, Recall and Average Precision (AP) are also included to evaluate the detection performance for CrowdHuman.

2) Training Details: The proposed OAF-Net is implemented in PyTorch and trained on 4 GTX TITAN X GPUs. HRNet-W32 [56] pre-trained on ImageNet [53] is selected as our backbone network, and Adam [66] solver is applied to optimize the network. The input sizes of training images are set to 640×1280, 336×448, and 800×1200 for CityPersons, Caltech, and CrowdHuman, respectively. We use brightness variation, horizontal flip, cropping, and random scaling (between 0.5 to 1.5) as the data augmentation. For CityPersons and Caltech, our OAF-Net is trained for $300k$ iterations with a mini-batch of 4 images per GPU, and the initial learning rate starts at 2×10^{-4} and decays to 1×10^{-4} after $200k$ iterations. For CrowdHuman, our detector is trained for 200k iterations with an initial learning rate of 2×10^{-4}, which then decreases to 1×10^{-4} after $100k$ iterations.

B. Ablation Studies

In this section, we first conduct the ablative analysis of the hyper-parameter η defined in the weighted Focal Loss on CityPersons. Furthermore, to figure out which component of OAF-Net is working, we also study the impact of the proposed weighted Focal Loss and occlusion-aware head on the performance improvement of our detector respectively, where η is fixed to its best-fit value. Because we mainly focus on the performance of our detector for occluded pedestrians, the results on the Reasonable, Heavy (Occlusion), Partial (Occlusion), and Bare (Occlusion) subsets of CityPersons are reported below.

1) Parameter Analysis of Weighted Focal Loss: To evaluate the influence of η in the weighted Focal Loss of the center prediction branch on the detection performance, we test our detector trained with different settings of η on CityPersons. Table I gives the performance comparison of our OAF-Net with different settings of η. As can be seen, the best performance is achieved on all subsets when η is set to 1. The performance drops when the value of η is greater than 1, it is because that OAF-Net may be over-fitted when the detector becomes too attuned to the hard training samples of occluded pedestrians on which it was trained. Therefore, in the following experiments, we fix the value of η to 1.

TABLE I

PERFORMANCE OF OUR OAF-NET WITH DIFFERENT SETTINGS OF η IN THE WEIGHTED FOCAL LOSS ON CITYPERSONS (RED INDICATES THE BEST PERFORMANCE)

η	$MR^{-2}(\%)$			
	Reasonable	Heavy	Partial	Bare
0.5	9.7	44.9	9.5	6.2
1	9.4	43.1	8.3	5.6
1.5	10.7	48.8	9.6	7.4
2	10.8	47.2	9.7	7.6

2) Impact of Weighted Focal Loss: To demonstrate the effectiveness of the proposed weighted Focal Loss, we first re-implement the baseline (CSP) using a HRNet-W32 backbone, which includes only one center prediction branch where the original Focal Loss is applied. In addition, we further modify the Focal Loss with the weighting factor ω based on CSP+HRNet-W32. The performance comparison of the above detectors is given in Table II. As can be seen, while applying the weighted Focal Loss to the baseline using a HRNet-W32 backbone can bring the performance gain (especially for partial and heavy occlusion subsets), it proves the advantage of the designed weighted Focal Loss.

3) Impact of Occlusion-Aware Head: In this subsection, to explore the effect of the proposed occlusion-aware head, we implement the occlusion-aware head by introducing different numbers of separate center prediction branches according to the occlusion levels. As can be seen in Table II, our OAF-Net with two or three center prediction branches in the detection head outperforms the baseline (CSP+HRNet-W32+weighted Focal Loss) with one center prediction branch consistently on Heavy, Partial, and Bare subsets. While using four center prediction branches in the detection head, OAF-Net still achieves better performance on the Heavy subset than the baselines. These results demonstrate the benefit of the designed occlusion-aware head. Based on the best performance of OAF-Net in Table II, we fix the number of center prediction branches in the detection head to be 3 in our experiments.

C. Benchmark Comparison

1) CityPersons: The proposed OAF-Net is extensively compared with some state-of-the-art pedestrian detectors [12]–[14], [23], [25], [29], [30], [40], [57], [67]–[69], [73]. The results on the Reasonable, Heavy (Occlusion), Partial (Occlusion), and Bare (Occlusion) subsets are listed in Table III. On the three occlusion subsets, our OAF-Net delivers the best performance among all the competing detectors. Even compared to some methods using occlusion-handling strategies (*i.e.*, RepLoss [13], FRCN+A+DT [23], OR-CNN [14], NOH-NMS [25], and APD [29]), our method also provides superior detection performance. Especially on the Heavy subset, our detector achieves a large performance gain of

TABLE II

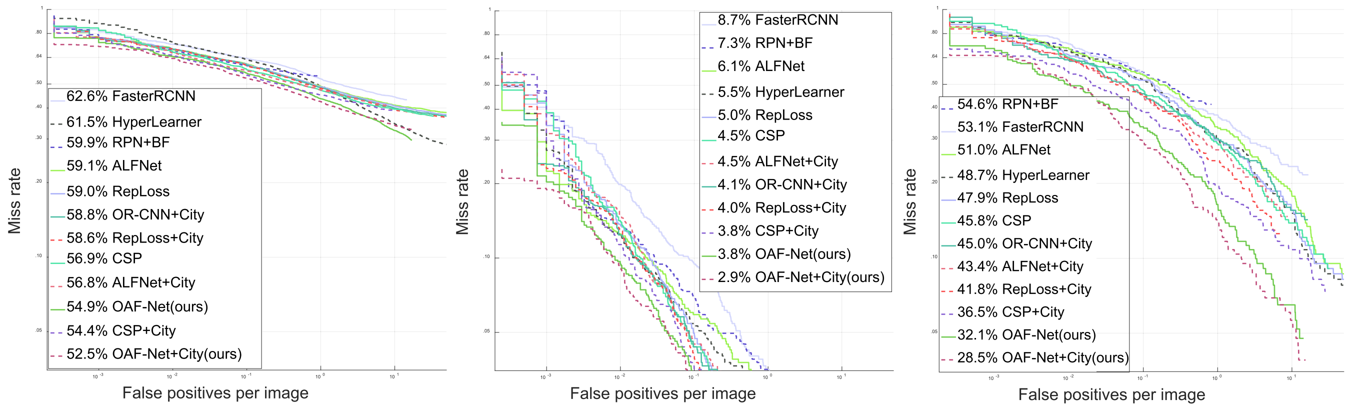
PERFORMANCE OF OAF-NET WITH DIFFERENT IMPLEMENTATIONS OF OCCLUSION-AWARE HEAD IN COMPARISON WITH THE BASELINE ON CITYPERSONS. FOR OAF-NET WITH TWO CENTER PREDICTION BRANCHES, THE PEDESTRIAN INSTANCES ARE DIVIDED INTO TWO OCCLUSION LEVELS (OCCLUSION <50% AND 50% ⩽OCCLUSION). FOR OAF-NET WITH FOUR CENTER PREDICTION BRANCHES, THE PEDESTRIAN INSTANCES ARE DIVIDED INTO FOUR OCCLUSION LEVELS (EVERY 25% VISIBILITY RATIO INTERVALS)

Method	$MR^{-2}(\%)$			
	Reasonable	Heavy	Partial	Bare
CSP+HRNet-W32	10.4	48.1	9.3	6.8
CSP+HRNet-W32 + weighted Focal Loss	10.2	46.5	8.8	6.8
OAF-Net (two center prediction branches)	9.9	45.8	8.6	6.6
OAF-Net (three center prediction branches)	9.4	43.1	8.3	5.6
OAF-Net (four center prediction branches)	12.6	45.4	11.4	8.8

TABLE III

PERFORMANCE OF OAF-NET IN COMPARISON WITH SOME STATE-OF-THE-ART METHODS ON CITYPERSONS

Method	Backbone	$MR^{-2}(\%)$				
		Reasonable	Heavy	Partial	Bare	Test time
RepLoss [13]	ResNet-50	13.2	56.9	16.8	7.6	N.A.
TLL [57]	ResNet-50	15.5	53.6	17.2	10.0	N.A.
TLL+MRF [57]	ResNet-50	14.4	52.0	15.9	9.2	N.A.
FRCN+A+DT [23]	VGG-16	11.1	44.3	11.2	6.9	N.A.
OR-CNN [14]	VGG-16	12.8	55.7	15.3	6.7	N.A.
ALFNet [40]	ResNet-50	12.0	51.9	11.4	8.4	0.27s/img
FRCNN [30]	VGG-16	15.4	N.A.	N.A.	N.A.	N.A.
FRCNN+Seg [30]	VGG-16	14.8	N.A.	N.A.	N.A.	N.A.
WIDEERPERSON [67]	VGG-16	11.1	N.A.	N.A.	N.A.	N.A.
PedHunter [73]	ResNet-50	8.3	N.A.	N.A.	N.A.	N.A.
CSP [12]	ResNet-50	11.0	49.3	10.4	7.3	0.33s/img
CSP [12]	HRNet-W32	10.4	48.1	9.3	6.8	0.28s/img
NOH-NMS [25]	ResNet-50	10.8	53.0	11.2	6.6	0.28s/img
APD [29]	DLA-34	8.8	46.6	8.3	5.8	0.16s/img
BGCNet [68]	HRNet-W32	8.8	43.9	8.0	6.1	0.16s/img
PRNet [69]	ResNet-50	10.8	53.3	10.0	6.8	0.22s/img
DAGN [50]	ResNet-50	11.9	43.9	12.1	7.6	0.22s/img
OAF-Net (ours)	HRNet-W32	9.4	43.1	8.3	5.6	0.25s/img



(a) All (b) Partial Occlusion (c) Heavy Occlusion

Fig. 5. Performance of OAF-Net in comparison with some state-of-the-art methods on Caltech using new annotations.

3.5% MR^{-2} upon the closest competitor APD using an additional NMS strategy). Moreover, for fair comparison with the baseline (*i.e.*, CSP), we also re-implement CSP in HRNet-W32. Compared to CSP and APD which only

TABLE IV

MR^{-2} PERFORMANCE OF OAF-NET IN COMPARISON WITH SOME STATE-OF-THE-ART METHODS ON CALTECH (RED AND BLUE INDICATE THE BEST AND SECOND-BEST RESULTS, PO-PARTIAL OCCLUSION SUBSET, HO-HEAVY OCCLUSION SUBSET)

Method	$MR^{-2}(\%)$		
	All	PO	HO
FasterRCNN [39]	62.6	8.7	53.1
HyperLearner [70]	61.5	5.5	48.7
RPN+BF [37]	59.9	7.3	54.6
ALFNet [40]	59.1	6.1	51.0
RepLoss [13]	59.0	5.0	47.9
OR-CNN+City [14]	58.8	4.1	45.0
RepLoss+City [13]	58.6	4.0	41.8
CSP [12]	56.9	4.5	45.8
ALFNet+City [40]	56.8	4.5	43.4
SDS-RCNN [23]	56.8	6.4	38.7
MS-CNN [35]	55.8	9.5	48.6
BGCNet [68]	N.A.	4.1	42.0
TFAN+TDEM+PRM [72]	N.A.	6.7	30.9
DAGN [50]	46.8	6.0	33.2
PedHunter [73]	39.5	N.A.	N.A.
CSP+City [12]	54.4	3.8	36.5
OAF-Net (ours)	54.9	3.8	32.1
OAF-Net+City (ours)	52.5	2.9	28.5

TABLE V

PERFORMANCE OF OAF-NET IN COMPARISON WITH SOME STATE-OF-THE-ART METHODS ON CROWDHUMAN

Method	MR^{-2}	Recall	AP
FPN [58]	52.4	90.6	83.1
FPN+Soft-NMS [27]	52.0	91.7	83.9
FPN+AdaptiveNMS [26]	49.7	91.3	84.7
RFB-Net [74]	65.2	94.1	78.3
RFB-Net+Soft-NMS [27]	66.3	95.4	78.1
RFB-Net+AdaptiveNMS [26]	63.0	94.8	79.7
GossipNet [75]	49.4	N.A.	80.4
RelationNet [76]	48.2	N.A.	81.6
Repulsion Loss [13]	45.7	88.4	85.6
PBM+R^{2}NMS[24]	43.4	93.3	89.3
NOH-NMS [25]	43.9	92.9	89.0
FCOS+AEVB [77]	47.7	N.A.	N.A.
Faster R-CNN+AEVB [77]	40.7	N.A.	N.A.
OAF-Net (ours)	45.0	96.5	89.8

utilize one center prediction branch and traditional Focal Loss, OAF-Net achieves the best results on three occlusion subsets. More specifically, our detector outperforms the baseline (CSP+HRNet-W32) consistently on all subsets and significantly improves Heavy, Partial, and Bare subsets by 5.0%, 1.0%, 1.2% MR^{-2} respectively. The large performance gain over the traditional center point based pedestrian detectors on the occlusion subsets comes from the designed occlusion-aware mechanism, which up-weights occluded pedestrian instances in the loss function and optimizes the network by selecting the most suitable center prediction branch for each pedestrian instance. The speed of OAF-Net is 0.25s/img with the original 640×1280 images, which is comparable to the other state-of-the-art pedestrian detectors.

2) Caltech: Fig. 5 summarizes the performance of our OAF-Net in comparison with some state-of-the-art methods [12], [13], [30], [37], [40], [70] on All, Partial Occlusion, and Heavy Occlusion subsets of Caltech. We also list MR^{-2} performance of OAF-Net and some state-of-the-art methods [12], [13], [23], [30], [35], [37], [40], [68], [70], [72], [73] on the three subsets in Table IV. Like some other methods (*e.g.*, ALFNet [40], OR-CNN [14], RepLoss [13], and CSP [12]), we also validate our detector with the model pre-trained on CityPersons [30]. As we can see, our OAF-Net+City achieves the best results and shows significant improvement over various existing methods on three subsets. On the Heavy Occlusion subset, OAF-Net+City greatly pushes the performance to 28.5% MR^{-2}, which is over 8.0% better than the closest competitor (36.5% of CSP+City). Even our OAF-Net is trained only with the training set of Caltech, the proposed detector still outperforms all the competing methods on Partial Occlusion subset, and also presents competitive performance compared to the detectors pre-trained on CityPersons. These results firmly demonstrate that our detector is effective to handle occlusions reasonably well in crowd scenes.

3) CrowdHuman: We compare our OAF-Net with some latest state-of-the-art methods [13], [24]–[27], [58], [74]–[77] on the CrowdHuman validation set in Table V. We can see that OAF-Net delivers the best performance in terms of Recall (96.5%) and AP (89.8%). Our detector using only the simple greedy-NMS algorithm still achieves comparable MR^{-2} (45.0%) compared to PBM+R^{2}NMS [24] and NOH-NMS [25] using some additional NMS strategies. The relatively promising detection results for crowd scenes in CrowdHuman also confirm the efficacy of our detector.

V. CONCLUSION

In this paper, we propose a simple but effective OAF-Net for pedestrian detection in crowd scenes. Based on the anchor-free detection framework, we design an occlusion-aware detection head to handle different occlusion levels of pedestrians in crowd scenes, where the occluded pedestrians are assigned to the most suitable center prediction branch. Accordingly, a novel weighted Focal Loss is proposed for these three center prediction branches, so that the occluded pedestrians can be up-weighted according to their visibility ratios. Our OAF-Net can be simply and effectively trained in an end-to-end fashion. Extensive experiments on three challenging pedestrian detection benchmarks (*i.e.*, CityPersons, Caltech, and CrowdHuman) demonstrate that our detector achieves the state-of-the-art results, and strongly validate the superiority of the designed occlusion-aware mechanism for heavily occluded pedestrians in crowd scenes. In the future, we expect to extend the proposed occlusion-aware mechanism into more anchor-free frameworks for many other object detection tasks.

REFERENCES

[1] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, and C. Fox, "Pedestrian models for autonomous driving—Part I: Low-level models, from sensing to tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6131–6151, Oct. 2021.

[2] L. Chen, S. Lin, X. Lu, D. Cao, and F. Y. Wang, "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021.

[3] Z. Chen and X. Huang, "Pedestrian detection for autonomous vehicle using multi-spectral cameras," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 211–219, Mar. 2019.

[4] K. Chen and Z. Zhang, "Pedestrian counting with back-propagated information and target drift remedy," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 639–647, Nov. 2016.

[5] S. H. Semnani and O. A. Basir, "Semi-flocking algorithm for motion control of mobile sensors in large-scale surveillance systems," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 129–137, Jul. 2014.

[6] S. Koehler *et al.*, "Stationary detection of the pedestrians intention at intersections," *IEEE Trans. Intell. Transp. Syst. Mag.*, vol. 5, no. 4, pp. 87–99, Oct. 2013.

[7] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. G. Jung, "A new approach to urban pedestrian detection for automatic braking," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 594–605, Dec. 2009.

[8] Z. Chen *et al.*, "A novel sparse representation model for pedestrian abnormal trajectory understanding," *Expert Syst. Appl.*, vol. 138, no. 30, pp. 1–11, Dec. 2009.

[9] C. Conde, D. Moctezuma, I. M. D. Diego, and E. Cabello, "HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments," *Neurocomputing*, vol. 100, no. 16, pp. 19–30, Dec. 2013.

[10] T. Sayed and R. Alsaleh, "Modeling pedestrian-cyclist interactions in shared space using inverse reinforcement learning," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 70, pp. 37–57, Apr. 2020.

[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] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5187–5196.

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

[14] 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.*, Sep. 2018, pp. 637–653.

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

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

[17] 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 Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4967–4975.

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

[19] 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.

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

[21] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3222–3229.

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

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

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

[25] P. Zhou *et al.*, "NOH-NMS: Improving pedestrian detection by nearby objects hallucination," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1967–1975.

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

[27] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5562–5570.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[29] J. Zhang, L. Lin, J. Zhu, Y. Li, and C. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Trans. Multimedia*, vol. 23, pp. 3085–3097, 2020.

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

[31] S. Shao *et al.*, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.

[32] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[33] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[34] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[35] 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.*, Oct. 2016, pp. 354–370.

[36] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. Davis, "Fused deep neural networks for efficient pedestrian detection," 2018, *arXiv:1805.08688*.

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

[38] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Oct. 2017.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2012.

[40] 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.*, Sep. 2018, pp. 618–634.

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2016, pp. 779–788.

[42] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 850–859.

[43] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 734–750.

[44] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*.

[45] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9657–9666.

[46] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6569–6578.

[47] X. Zhou, D. Wang, and P. Krähenbuhl, "Objects as points," 2019, *arXiv:1904.07850*.

[48] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 840–849.

[49] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detector," 2019, *arXiv:1904.03797*.

[50] H. Xie, W. Zheng, and H. Shin, "Occluded pedestrian detection techniques by deformable attention-guided network," *Appl. Sci.*, vol. 11, no. 13, pp. 1–19, Jun. 2021.

[51] Q. Li, H. Qiang, and J. Li, "Conditional random fields as message passing mechanism in anchor-free network for multi-scale pedestrian detection," *Inf. Sci.*, vol. 550, pp. 1–12, Oct. 2021.

[52] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2014, pp. 1505–1512.

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

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

[55] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[56] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.

[57] 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.*, Sep. 2018, pp. 536–551.

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

[59] S. W. Kim, H. K. Kook, J. Y. Sun, M. C. Kang, and S. J. Ko, "Parallel feature pyramid network for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 234–250.

[60] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, "Deep feature pyramid reconfiguration for object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 169–185.

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

[62] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12214–12223.

[63] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, "TJU-DHD: A diverse high-resolution dataset for object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 207–219, 2020.

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

[65] 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.*, Jun. 2016, pp. 1259–1267.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[67] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[68] J. Li, S. Liao, H. Jiang, and L. Shao, "Box guided convolution, pedestrian detection, receptive fields, scale variation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1615–1624.

[69] X. Song, K. Zhao, W. S. Chu, H. Zhang, and J. Guo, "Progressive refinement network for occluded pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 32–48.

[70] 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.

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

[72] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, and J. Yuan, "Temporal-context enhanced detection of heavily occluded pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13427–13436.

[73] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "PedHunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. IEEE Conf. Comput. Artif. Intell.*, Apr. 2020, pp. 10639–10646.

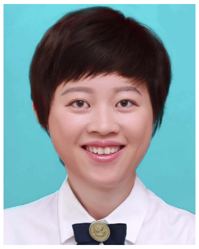
[74] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 404–419.

[75] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4507–4515.

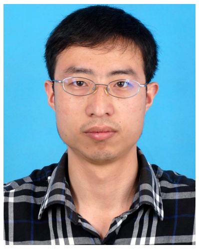
[76] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[77] Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, "Variational pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11617–11626.

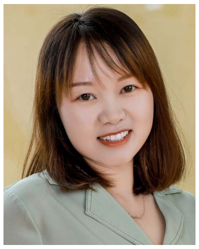
Qiming Li received the Ph.D. degree from the School of Information Science and Technology, Xiamen University, in 2016. He is currently an Associate Research Fellow with the Haixi Institute, Chinese Academy of Sciences. He has authored over 20 SCI indexed journal papers and EI indexed refereed conference papers, including IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and *Signal Processing*. His research interests include computer vision, object detection and tracking, and machine learning. He has presided over several projects at various levels, such as the National Natural Science Foundation of China, China Post-doctoral Science Foundation, and many other provincial-level projects.



Yijing Su received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree from Tsinghua University in 2017. She is currently an Assistant Research Fellow with the Haixi Institutes, Chinese Academy of Sciences. Her research interests include computer vision and fingerprint recognition. She has presided over or participated in projects at various levels, such as the National Natural Science Foundation of China, provincial and municipal science and technology plan.



Yin Gao received the M.S. degree in optical engineering from Yunnan Normal University, China, in 2012. He is currently a Senior Engineer with the Haixi Institute, Chinese Academy of Sciences. He has published more than 20 innovative academic papers in Elsevier and other SCI journals. His research interests include image dehazing, low-light image enhancement, multi-exposure image fusion, fundamental study of image optimization, and product identification and defect detection in industrial applications.



Feng Xie received the bachelor's degree in information engineering from the East China University of Science and Technology in 2015 and the master's degree in electrical engineering and information technology from Magdeburg University, Germany, in 2018. In 2019, she received the VDI-Funded Award from the Association of German Engineers. She is currently a Research Engineer with the Institute of Automation and Communication, Magdeburg, Germany. Her recent work focuses on electric vehicles simulation, battery modeling, and software development for German and European international projects. Her research interests include transportation electrification and intelligent transport systems.

Jun Li (Member, IEEE) received the Ph.D. degree from the University of Munich, Germany. He is currently the Director of the Fujian Robotics Intelligent System Engineering Technology Research Center. He was a Post-Doctoral Research Associate with University of Marburg, Germany. In 2015, he joined the faculty with the Haixi Institute, Chinese Academy of Sciences. His research focuses on the image processing and recognition, robot adaptive control, and human–computer interaction. In the field of images and robots, he has published two personal English and German monograph, published more than 40 innovative academic papers in IEEE and other SCI journals, applied more than 30 invention patents, and obtained three national invention patents. In 2016, he was selected as a member of the Fujian Provincial Double-Hundred-Talent Program. Since 2015, he has been the Leader of the Laboratory of Robotics and Intelligent Systems hosted and participated different scientific research programs, including the national 13th Five-Year Key Research Program and Fujian Major Science and Technology projects.