MSAGNet: Multi-Stream Attribute-Guided Network for Occluded Pedestrian Detection

Hong Zhang, Chaoqi Yan, Xuliang Li, Yifan Yang, and Ding Yuan, *Member, IEEE*

***Abstract*—Pedestrian detection plays an indispensable role in human-centric applications. Although having enjoyed the merits of generic object detectors based on deep learning frameworks, pedestrian detection is still a persistent crucial task since the pedestrians often gather together and occlude each other. In this study, we propose a simple yet effective Multi-Stream Attribute-Guided Network (MSAGNet) to regard occluded pedestrian detection as a standard central point and height estimation problem. Specifically, we focus on searching for the central points of the pedestrians and predicting the scales and offsets of the corresponding pedestrians. Meanwhile, an adaptive weighting parameter, i.e., Intersection over the Visible part region of ground truth (IoV), is utilized to conduct accurate bounding box regression. Furthermore, a novel nonlinear Non-Maximum Suppression (NMS) is proposed to flexibly prune false positives and decrease the miss rate of adjacent overlapping pedestrians. Experimental results on Caltech-USA, CityPersons, CrowdHuman and WiderPerson pedestrian datasets show that the proposed MSAGNet can obtain significant performance boosts, while maintaining a reasonable run-time speed.**

***Index Terms*—Pedestrian detection, deep learning, bounding box regression, non-maximum suppression, false positives.**

I. Introduction

PEDESTRIAN detection has drawn considerable attention in the last decade. It is one of the key problems in many human-centric applications, and has been widely applied to autonomous vehicles, person identification, and video surveillance. Although impressive success has been achieved over the years [1], [2], [3], [4], [5], [6], the pedestrian detection performance still considerably deteriorates when dealing with crowded scenes, as occluded pedestrians are extremely difficult to detect and locate precisely.

According to statistics, there are approximately 48.8% adjacent overlapping pedestrians in the CityPersons dataset [7], which presents significant challenges in tackling occlusion cases. As shown in Fig. 1(b), pedestrian (A) is heavily overlapped by pedestrian (B). It is difficult for the detectors to conduct discriminable predictions for such proposals, generating false positive predictions, such as (F). Also, because the detectors are sensitive to the threshold of NMS [8], the predictions of some overlapped pedestrians are likely to be mistakenly suppressed by NMS, such as the missed detection of pedestrian (M). Given this situation, the exploration of occluded pedestrian detection raises some open questions: *How to design a simple yet effective anchor-free method to tackle the occlusion problem for the pedestrian detection task? How to design a novel bounding box regression loss designed explicitly for the occluded pedestrians by incorporating the occlusion information? And how to design an effective NMS mechanism to refine the redundant detection boxes for adjacent overlapping pedestrians?*

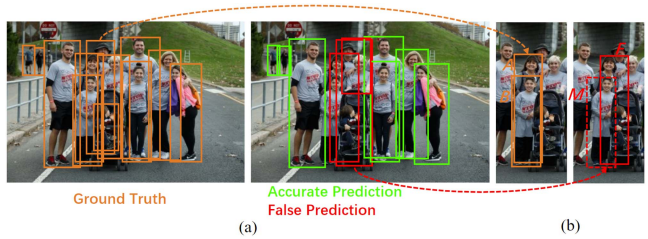


Fig. 1. Illustration of challenge in crowded scenes. The orange bounding boxes in (a) are the pedestrian ground-truths. The green and red bounding boxes denote the true and false positives, respectively. The dashed red box in (b) shows the common miss detection result (suppressed by NMS mistakenly).

To accurately localize a visual pedestrian, several studies [9], [10], [11], [12], [13] have been conducted to address occlusion problems. Early pedestrian detectors extended from the Viola and Jones paradigm [14] rely primarily on low-level handcrafted features to detect pedestrians, such as ACF [15], MultiFtr [16] and Checkerboards [17]. Recently, models based on Convolutional Neural Networks (CNNs) have pushed pedestrian detection to an attractive success since CNNs are verified to have a stronger capability to extract high-level image features, and they are usually divided into anchor-based and anchor-free methods [18], [19], [20], [21], [22]. In anchor-based methods, Faster R-CNN [19] and SSD [18] are used as predominant baselines for pedestrian detection. However, without any specific occlusion handling, it remains challenging for these methods to robustly detect pedestrians in a crowd. Most recently, anchor-free methods have achieved rapid development due to their flexibility in network model design. For example, CornerNet [20] detects bounding boxes as pairs of corners. CenterNet [21] identifies

Manuscript received 14 July 2022; revised 13 October 2022; accepted 14 October 2022. Date of publication 20 October 2022; date of current version 2 November 2022. This work was supported by the National Natural Science Foundation of China under Grants 61872019, 61972015, 62002005, and 62002008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Victor Sanchez. *(Hong Zhang and Chaoqi Yan contributed equally to this work.) (Corresponding author: Ding Yuan.)*

Hong Zhang, Chaoqi Yan, Xuliang Li, and Ding Yuan are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: dmrzhang@buaa.edu.cn; cqyan92@gmail.com; xulli8997@buaa.edu.cn; dyuan@buaa.edu.cn).

Yifan Yang is with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China (e-mail: stephenyoung@buaa.edu.cn).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LSP.2022.3215920, provided by the authors.

Digital Object Identifier 10.1109/LSP.2022.3215920

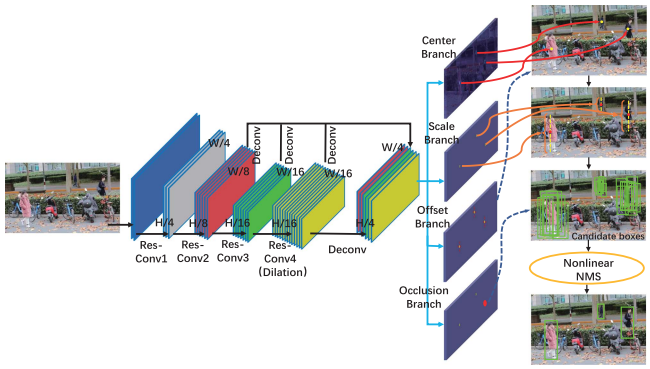


Fig. 2. An overview of the proposed single-shot MSAGNet pedestrian detection architecture. Our detector utilizes truncated ResNet-50 to extract representative features. The detection head consists of center, scale, offset and occlusion attribute branches and a Nonl-NMS is introduced to refine the predicted boxes at the post-processing stage.

that a single point at the center of the bounding box is feasible for object localization. However, heavy occlusion between pedestrians poses a significant challenge to the standard NMS [23], [24]. Soft-NMS [25] lowers the detection scores of neighbors instead of setting them to zero to reduce the number of false positives. Adaptive-NMS [8] develops a dynamic suppression strategy by setting an adaptive threshold for a high localization accuracy. Based on the above considerations, we devise a unified end-to-end anchor-free detection network for occluded pedestrian instances in crowd scenarios. The primary contributions of this study are summarized as follows:

1) We propose a simple yet effective MSAGNet by solving the crowd occlusion issue from the basic semantic features, including center position, target's scale, offset, and occlusion attributes.

2) We propose a novel optimized bounding box regression loss fusion designed explicitly for detecting occluded pedestrians by introducing an adaptive weighting parameter.

3) We propose a nonlinear NMS (denoted as Nonl-NMS) to reduce false positives and decrease the miss-rate by performing a greedy fashion for crowded scenes.

4) Extensive experiments conducted on challenging occluded pedestrian detection benchmarks demonstrate the proposed MSAGNet has better overall performance than the other state-of-the-art methods.

II. Method

The overall architecture of the proposed MSAGNet pedestrian detection model is illustrated in Fig. 2. The backbone of the proposed MSAGNet is truncated from the ResNet-50 network. Specifically, the corresponding output feature maps are down-sampled by 2, 4, 8, 16, and 32 w.r.t. the input image. Moreover, dilated convolutions are adopted to maintain the spatial size of the final feature map at 1/16 of the input image. Based on representation theory, we extract features from the last layer of each res-block, that is, Resnet50-Conv2, Conv3, and Conv4, and utilize deconvolution to unify spatial resolutions to 1/4 of the input size. Finally, multi-scale feature maps with the same resolution are skip-connected for the final detection, which is performed on the aggregated feature map Φ_{det} with a size of $H/r \times W/r$. Based on the experiments, the down-sampling factor $r = 4$ yields the best performance. The detection head consists of four branches: the center, scale, offset, and occlusion attribute branches. All of these multi-stream attribute branches have the same size as the aggregated feature map Φ_{det} and are subsequently appended into the network. Finally, a novel Nonl-NMS is used to produce more compact bounding boxes.

A. Center Branch

The center branch is designed to predict the center points of pedestrians. However, it is difficult to detect an 'accurate' center point of a pedestrian. Thus, a 2D Gaussian mask $G(.)$ is exploited at the center location of each positive to alleviate the training difficulties. Formally, it can be defined as:

$$G\left(i,j;x,y,\sigma_w,\sigma_h\right) = e^{-\left(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(j-y)^2}{2\sigma_h^2}\right)}, \tag{1}$$

$$P_{ij} = \max_{k=1,2,\ldots,K} G\left(i,j;x_k,y_k,\sigma_{w_k},\sigma_{h_k}\right), \tag{2}$$

where P_{ij} denotes the overall mask map about the center points of pedestrians, K is the number of positive instances in an image, (x_k, y_k, w_k, h_k) represent the center coordinates, width, and height of the k-th pedestrian, respectively. In addition, the variances (σ_w^k, σ_h^k) are proportional to the height and width of each individual pedestrian. We also utilize the focal weights on hard examples to alleviate the positive-negative imbalance problem. Based on this, the final classification loss can be expressed as:

$$L_{center} = -\frac{1}{K}\sum_{i=1}^{W/r}\sum_{j=1}^{H/r}\alpha_{ij}\left(1-\hat{p}_{ij}\right)^{\gamma}\log\left(\hat{p}_{ij}\right), \tag{3}$$

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij}=1 \\ 1-p_{ij} & \text{otherwise} \end{cases}, \tag{4}$$

$$\alpha_{ij} = \begin{cases} 1 & \text{if } y_{ij}=1 \\ \left(1-P_{ij}\right)^{\beta} & \text{otherwise} \end{cases}, \tag{5}$$

where $p_{ij} \in [0, 1]$ represents the network's estimated probability, illustrating whether there is an instance's center in location (i, j). $y_{ij} = 1$ denotes a positive location. To handle the ambiguity from the negatives surrounding the positives, α_{ij} is exploited to weaken their weights to the total loss according to the Gaussian mask in (2). We empirically set the hyper-parameters $\gamma = 2$ and $\beta = 4$, as suggested in [26] for a fair comparison.

B. Scale Branch

The scale branch is designed to predict the height and/or width of pedestrians. The height of the pedestrian is merely predicted for simplicity in this study, and the bounding boxes can be generated using a fixed aspect ratio (0.41) [7]. The scale ground truth can be formulated as:

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

where h_k denotes the height of k-th object. We assign $log(h_k)$ to the negatives within a radius $r = 2$ of the positives, and all other regions are set to zero to alleviate ambiguity.

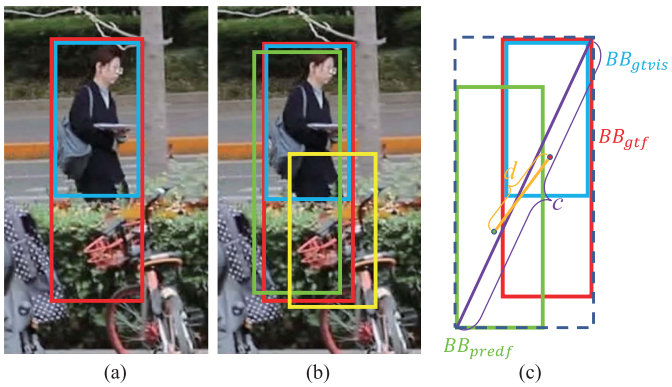


Fig. 3. (a) Two forms of pedestrian annotation, i.e., full body annotation and visible part region annotation. (b) Illustration of good pedestrian proposal (green box) and bad pedestrian proposal (yellow box). (c) Illustration of different bounding boxes. c is the diagonal distance of the small enclosing box containing BB_{gtf} and BB_{predf}. d represents the Euclidean distance of the central points of the two boxes.

An optimized Smooth L1 loss function [19] is proposed to improve the robustness of our model by introducing occlusion information. As shown in Fig. 3(c), BB_{predf} represents the full-body region of the predicted bounding box. Meanwhile, BB_{gtf} and BB_{gtvis} are the full-body region and visible part region of the pedestrian annotation (i.e., ground truth), respectively. IoV is the proportion of the area of BB_{gtvis} covered by BB_{predf}, and is calculated as:

$$\text{IoV} = \frac{Area\left(BB_{predf} \cap BB_{gtvis}\right)}{Area\left(BB_{gtvis}\right)}. \tag{7}$$

IoV is incorporated as an adaptive weighting parameter into the Smooth L1 loss (denoted as IoV-S loss), and the final regression loss for the scale branch can be formulated as follows:

$$L_{scale} = \frac{1}{K}\sum_{k\in A}\text{IoV}_k\sum_{k=1}^{K}\text{Smooth L1}\left(\hat{s}_k, s_k\right), \tag{8}$$

where A represents all bounding boxes. $\hat{s}_k$ and s_k denote the predicted height and ground truth of each pedestrian, respectively. Note that if BB_{predf} has a high overlap with BB_{gtvis}, a larger penalty will be added to the regressor. In this manner, the IoV assigns more weight to the loss function, and thus the regressor is trained to narrow the gap between the proposals and ground-truth pedestrians more efficiently.

C. Offset Branch

The offset branch is designed to fine-tune the center position of pedestrians in the case of the center-shift problem [26]. For the offset target $\mathcal{O} \in \mathcal{R}^{\frac{W}{r}\times\frac{H}{r}\times 1}$, the corresponding ground truth can be formulated as:

$$t_{ij} = \begin{cases} t_i = x_k/r - \lfloor x_k/r \rfloor \\ t_j = y_k/r - \lfloor y_k/r \rfloor \end{cases}. \tag{9}$$

It is also considered a regression task via Smooth L1 loss, as follows:

$$L_{offset} = \frac{1}{K}\sum_{k=1}^{K}\Phi\left(\hat{o}_k, o_k\right), \tag{10}$$

where $\hat{o}_k$ is the predicted offset, and o_k represents the ground truth of each positive instance. In summary, the overall multi-task loss function is defined as follows:

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset}, \tag{11}$$

where λ_c, λ_s, and λ_o are the weights for the corresponding branches and are experimentally set to 0.01, 1, and 0.1, respectively.

D. Occlusion Branch

The occlusion branch is designed to propose a robust NMS algorithm to suppress the detection boxes in the post-processing stage. The widely used Soft-NMS [25] penalizes the detection score of neighbors using a continuous function as follows:

$$s_i = \begin{cases} s_i, & IoU\left(\mathcal{M}, b_i\right) < N_t \\ s_i f\left(IoU\left(\mathcal{M}, b_i\right)\right), & IoU\left(\mathcal{M}, b_i\right) \geq N_t \end{cases}, \tag{12}$$

where s_i is the classification score of box b_i, N_t is the NMS threshold, $\mathcal{M}$ represents the detection box with the maximum score, and $f(IoU(\mathcal{M}, b_i))$ is the overlap-based weighting function. However, Soft-NMS inevitably detects one object with multiple adjacent regression boxes and generates more false positives. Based on this work, a novel Nonl-NMS that introduces a nonlinear Gaussian penalty is proposed to suppress the redundant detection boxes $\mathcal{B}$. The pruning step can be obtained by:

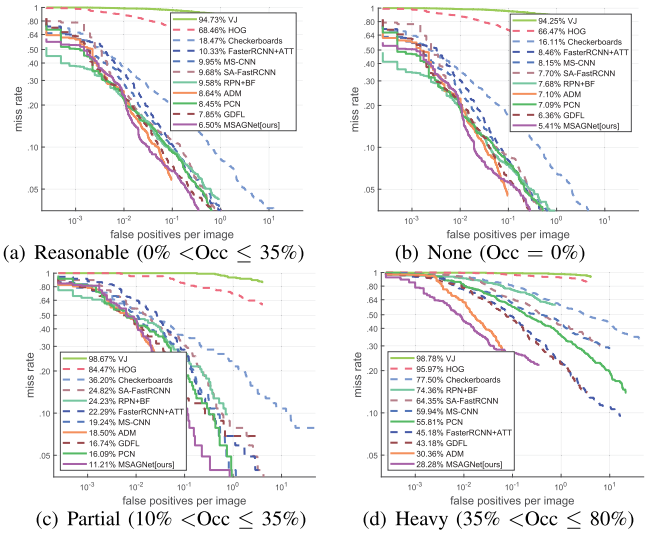
$$s_i = \begin{cases} s_i, & DIoU\left(\mathcal{M}, b_i\right) < N_t \\ s_i\left(e^{-DIoU(\mathcal{M}, b_i)}\right), & DIoU\left(\mathcal{M}, b_i\right) \geq N_t \end{cases}, \tag{13}$$

where $DIoU(\mathcal{M}, b_i)$ is calculated using [27] for a better bounding box regression instead of the previous IoU metric. In this study, N_t is the threshold and is empirically set to 0.5. We choose DIoU as our evaluation metric because it incorporates the normalized distance between the predicted box BB_{predf} and target box BB_{gtf} as shown in Fig. 3(c). It can largely minimize the distance between two central points. Other evaluation metrics, GIoU [28] and CIoU [27], are also encouraged to be explored for potential performance boost in future work.

III. Experiments

A. Experimental Settings

We evaluate the proposed approach on widely used benchmarks including Caltech-USA [29], CityPersons [7], CrowdHuman [30], and WiderPerson [31] for qualitative and quantitative comparisons. For the Caltech benchmark, we use the set00∼set05 training data augmented by 10 folds (42,782 frames) and test on 4,024 frames in the standard test set. For the CityPersons benchmark, we use all 2,975 images in the training set for training and 500 images for validation. The log-average Miss Rate over False Positive Per Image (FPPI) [29] within the range of 10^{-2} to 10^{0} (denoted as MR^{-2}) is employed as the evaluation metric. *The qualitative and quantitative results on CrowdHuman and WiderPerson datasets, and an overview of this study along with more details are given in the supplementary material I and II.*



(a) Reasonable (0% <Occ ≤ 35%) (b) None (Occ = 0%)

(c) Partial (10% <Occ ≤ 35%) (d) Heavy (35% <Occ ≤ 80%)

Fig. 4. L-AMR versus FPPI plot of different pedestrian detection algorithms on the new Caltech-USA test subset. Numbers in legend are the log-average miss rates. "Occ" is the abbreviation of "Occlusion".

TABLE I
QUANTITATIVE COMPARISONS TO OTHER STATE-OF-THE-ART PEDESTRIAN DETECTORS ON THE CITYPERSONS VALIDATION SET

Method	Backbone	MR^{-2}				Test time
		R	HO	Partial	Bare	
OR-CNN (ECCV'18) [2]	VGG-16	12.8	55.7	15.3	6.7	-
ALFNet (ECCV'18) [42]	ResNet-50	12.0	51.9	11.4	8.4	0.27s/img
TLL+MRF (ECCV'18) [43]	ResNet-50	14.4	52.0	15.9	9.2	-
CSPNet (CVPR'19) [26]	ResNet-50	11.0	49.3	10.4	7.3	0.33s/img
Adaptive-NMS (CVPR'19) [8]	ResNet-50	10.8	54.0	11.4	6.2	-
NOH-NMS (ACM'20) [44]	ResNet-50	10.8	53.0	11.2	6.6	0.28s/img
PRNet (ECCV'20) [45]	ResNet-50	10.8	53.3	10.0	6.8	0.22s/img
AP^2M (AAAI'21) [41]	ResNet-50	10.4	**48.6**	9.7	6.2	-
PRNet++ (Neuro'22) [46]	ResNet-50	10.7	51.2	9.9	6.9	0.20s/img
MSAGNet (Ours)	ResNet-50	**9.2**	49.3	**9.1**	**5.6**	**0.19s/img**

TABLE II
ABLATION STUDY FOR GREEDY-NMS, SOFT-NMS AND NONLINEAR NMS

Method	Greedy	Soft	Nonl-NMS	Backbone	MR^{-2}
Adapted FRCNN [7]	✓			VGG-16	15.4
		✓		VGG-16	14.2
			✓	VGG-16	**12.9**
CSPNet [26]	✓			ResNet-50	11.0
		✓		ResNet-50	10.2
			✓	ResNet-50	**9.8**

TABLE III
ABLATION STUDY ABOUT TWO PROPOSED COMPONENTS ON THE CITYPERSONS VALIDATION SET

Component		R	HO	Partial	Bare
IoV-S Loss	Nonl-NMS				
		10.6	50.4	10.6	7.2
✓		10.2	49.9	10.2	6.8
	✓	9.8	49.6	9.7	6.4
✓	✓	**9.2**	**49.3**	**9.1**	**5.6**

B. Experimental Results on Caltech-USA

We evaluate the performance on four setups of the Caltech-USA dataset: Reasonable (**R**), **None**, **Partial** and Heavy Occlusion (**HO**). We compare the proposed MSAGNet with the following state-of-the-art methods: Checkerboards [17], FasterRCNN+ATT [32], MS-CNN [33], SA-FastRCNN [34], RPN+BF [35], ADM [36], PCN [37], GDFL [38]. VJ [39] and HOG [40] are baselines. Fig. 4 shows the quantitative results, which show that our method consistently outperforms the previous methods by a significant margin.

C. Experimental Results on CityPersons

We also compare the proposed MSAGNet with the existing best-performing detectors on the CityPersons dataset, and the quantitative results are shown in Table I. It can be observed that the proposed MSAGNet achieves MR^{-2} of 9.2% on the **R** setting, which clearly exceeds the best existing result, 10.4% of AP^2 M [41]. In addition, our detector achieves a leading performance consistently in **HO**, **Patrial**, and **Bare** settings, 49.3%, 9.1%, and 5.6%, respectively.

To evaluate the effectiveness of the proposed Nonl-NMS, we conduct an ablation study on the CityPersons validation set [7] on two types of baseline detectors: two-stage adapted Faster R-CNN [7] detector and one-stage CSPNet [26] detector. The best performance of the traditional greedy-NMS and soft-NMS [25] with the "linear" method is reported at $N_t = 0.5$ for a fair comparison. As shown in Table II, with the proposed Nonl-NMS method, both the adapted Faster R-CNN and CSPNet can achieve the best results on the **R** setting.

An ablative analysis is also conducted on the CityPersons validation set to evaluate the effectiveness of the two proposed components (IoV-S Loss and Nonl-NMS). As shown in Table III, our design performs better than solely utilizing IoV-S Loss or Nonl-NMS on four subsets with different occlusion levels [7] of CityPersons.

D. Runtime Efficiency

Given the anchor-free architecture, our method can achieve a conspicuous speed with 0.19 s per image of 1024×2048 pixels as shown in Table I. It has a slight advantage over the recent state-of-the-art pedestrian detector, 0.20 s/img of PRNet++ [46]. The experiment is conducted on a PC with an Intel(R) Core(TM) i7-10850K@3.60 GHz processor and one NVIDIA RTX3070 GPU.

IV. CONCLUSION

In this study, we take the pedestrian detection task as a standard central point and height estimation problem and propose a novel Multi-Stream Attribute-Guided Network (MSAGNet) for occluded pedestrian detection. By introducing occlusion information as an adaptive weighting parameter to balance the regression calculation of the Smooth L1 loss, our detector can be more flexibly robust to occlusion scenes. Meanwhile, the proposed Nonl-NMS can adaptively avoid the highly overlapped instances being falsely suppressed in the post-processing stage. Our proposed method has been demonstrated to be competitive over state-of-the-art pedestrian detectors while achieving a reasonable speed-accuracy trade-off.

REFERENCES

[1] G. Lee, S. Hong, and D. Cho, "Self-supervised feature enhancement networks for small object detection in noisy images," *IEEE Signal Process. Lett.*, vol. 28, pp. 1026–1030, 2021.

[2] 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.*, 2018, pp. 637–653.

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

[4] 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.*, 2018, pp. 7774–7783.

[5] 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.*, 2020, pp. 10747–10756.

[6] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12211–12220.

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

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

[9] C. Zhuang, Z. Li, X. Zhu, Z. Lei, and S. Z. Li, "SADet: Learning an efficient and accurate pedestrian detector," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2021, pp. 1–8.

[10] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 303–312.

[11] J. Wang, C. Zhao, Z. Huo, Y. Qiao, and H. Sima, "High quality proposal feature generation for crowded pedestrian detection," *Pattern Recognit.*, vol. 128, 2022, Art. no. 108605.

[12] Y. He, N. He, R. Zhang, K. Yan, and H. Yu, "Multi-scale feature balance enhancement network for pedestrian detection," *Multimedia Syst.*, vol. 28, no. 3, pp. 1135–1145, 2022.

[13] H. Li, J. Li, H. Wei, Z. Liu, Z. Zhan, and Q. Ren, "Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles," 2022, *arXiv:2206.02424*.

[14] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[15] 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.

[16] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *Proc. Joint Pattern Recognit. Symp.*, Berlin, Heidelberg: Springer, 2008, pp. 82–91.

[17] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection.," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. 2, 2015, pp. 1751–1760.

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

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 91–99, 2015.

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

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

[22] J. Zhang et al., "Attribute-aware pedestrian detection in a crowd," *IEEE Trans. Multimedia*, vol. 23, pp. 3085–3097, 2021.

[23] Z. Ge, Z. Jie, X. Huang, R. Xu, and O. Yoshie, "PS-RCNN: Detecting secondary human instances in a crowd via primary object suppression," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2020, pp. 1–6.

[24] D. Rukhovich, K. Sofiiuk, D. Galeev, O. Barinova, and A. Konushin, "Iterdet: Iterative scheme for object detection in crowded environments," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. Struct. Syntactic Pattern Recognit.*, Cham, Switzerland: Springer, 2021, pp. 344–354.

[25] 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.*, 2017, pp. 5561–5569.

[26] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5182–5191.

[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 12993–13000.

[28] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.

[29] P. Dollar, 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.

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

[31] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "Widerperson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Trans. Multimedia*, vol. 22, pp. 380–393, 2020.

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

[33] 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.

[34] 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, pp. 985–996, 2018.

[35] 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.

[36] 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.

[37] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and context information for pedestrian detection with CNNs," 2018, *arXiv:1804.04483*.

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

[39] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

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

[41] M. Liu, C. Zhu, J. Wang, and X.-C. Yin, "Adaptive pattern-parameter matching for robust pedestrian detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2154–2162.

[42] 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.*, 2018, pp. 618–634.

[43] 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.*, 2018, pp. 536–551.

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

[45] 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.*, Cham, Switzerland: Springer, 2020, pp. 32–48.

[46] X. Song, B. Chen, P. Li, B. Wang, and H. Zhang, "PRNet++ : Learning towards generalized occluded pedestrian detection via progressive refinement network," *Neurocomputing*, vol. 482, pp. 98–115, 2022.