IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 7, 2022, accepted January 21, 2022, date of publication February 22, 2022, date of current version February 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150988

Occlusion Handling and Multi-Scale Pedestrian Detection Based on Deep Learning: A Review

FANG LI[1], XUEYUAN LI[1], QI LIU[1], AND ZIRUI LI[1,2]
[1]School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100089, China
[2]Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, The Netherlands

Corresponding author: Xueyuan Li (lixueyuan@bit.edu.cn)

ABSTRACT Pedestrian detection is an important branch of computer vision, and has important applications in the fields of autonomous driving, artificial intelligence and video surveillance. With the rapid development of deep learning and the proposal of large-scale datasets, pedestrian detection has reached a new stage and has achieved better performance. However, the performance of state-of-the-art methods is far behind expectations, especially when occlusion and scale variance exist. Therefore, many works focused on occlusion and scale variance have been proposed in the past few years. The purpose of this article is to make a detailed review of recent progress in pedestrian detection. First, a brief progress of pedestrian detection in the past two decades is summarized. Second, recent deep learning methods focusing on occlusion and scale variance are analyzed. Moreover, the popular datasets and evaluation methods for pedestrian detection are introduced. Finally, the development trends in pedestrian detection are discussed.

INDEX TERMS Deep learning, pedestrian detection, occlusion handling, scale variance.

I. INTRODUCTION

One of the most exciting opportunities at the intersection of robotics and deep learning is autonomous driving, a comprehensive intelligent system that integrates perception, positioning, planning, decision making and motion control [1]–[3]. As the top layer of autonomous driving, the perception system needs to be further improved to achieve a comprehensive understanding of the scene to make the best driving decisions.

As an important part of the real world, pedestrians often occupy the largest number in most datasets, as shown in Figure 1. Therefore, human-centered tasks (e.g., pedestrian re-identification [4], pedestrian detection [5], [6], pedestrian trajectory prediction [7], person search [8] and pedestrian counting [9]) have received considerable attention. Among them, pedestrian detection is a basic task in real-world applications. Pedestrian detection aims to detect all instances and predict their bounding boxes from a given input image or a video, which requires high accuracy and efficiency. Compared to image detection, video detection can utilize temporal context information. Making full use of the temporal context can solve data redundancy in videos and improve the detection speed. It can also improve the detection performance and solve the problems of motion blur, occlusion, and various poses. During the last decade, object detection has made breakthroughs and achieved high performance in popular datasets, such as ImageNet [10], Pascal VOC [11], and MS COCO datasets [12], which is driven by machine learning, especially deep learning techniques. Pedestrian detection has also received considerable attention as a specific category of generic objects, as shown in Figure 2.

Pedestrian detection methods can mainly be divided into two categories: hand-crafted features based [13]–[16] and deep features based [17]–[21]. In first category, hand-crafted features such as Histogram of Oriented Gradients (HOG) [13] and Integral Channel Features (ICF) [14] are extracted to train classifier. These methods are sufficient for some simple cases. However, the efficiency is low, and the performance is not satisfactory. With the rapid development of deep learning, especially the proposal of generic object detection, deep learning-based methods for pedestrian detection achieves significant improvements in terms of speed and accuracy. However, state-of-the-art pedestrian detection performance is still not comparable to that of human perception. Pedestrian detection still faces many challenges, such as following points:

The associate editor coordinating the review of this manuscript and approving it for publication was Gianluigi Ciocca.

(a) PASCAL VOC (20 Classes)

(b) MS COCO (80 Classes)

FIGURE 1. Most frequent object classes in (a) PASCAL VOC and (b) COCO. The size of each word is proportional to the frequency of that class in the training dataset.

1) **Large differences in appearance** Environment condition is various in the real world, such as lighting (i.e., dawn, day, and dusk), weather conditions, backgrounds, illuminations, occlusion, and viewing distances. On the other hand, there are many differences among people, such as clothing, and attachments on the body. All these conditions produce significant variations in pedestrian appearance, such as pose, scale, occlusion, clutter, shading, blur, and motion, as shown in Figure 3.
2) **Occlusion** In many real-time applications, pedestrians are extremely dense. Pedestrians are often occluded by other objects (Figure 3(a)) or dense pedestrians (Figure 3(b)); therefore, only a part of the human body can be seen. Highly overlapped instances are likely to have very similar features, which poses great difficulty in detection.
3) **Scale variance** Pedestrians with different spatial scales may exhibit dramatically different features, as shown in Figure 3(c). Small-scale pedestrians are very common in real scenes, and accurately localizing them is challenging owing to blurred boundaries and obscure appearance.
4) **Complex background** The background is very complex both indoors and outdoors, as shown in Figure 3(d). Some objects resemble human bodies in appearance, shape, color, and texture, making it difficult to accurately distinguish pedestrians.

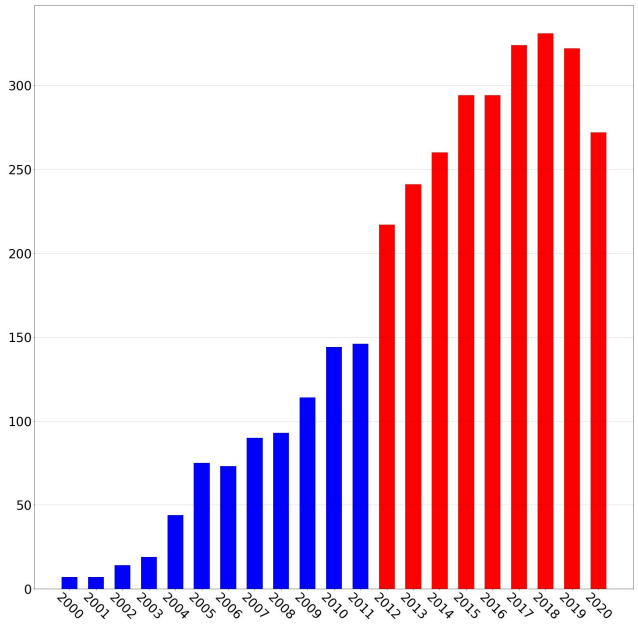


FIGURE 2. The increasing number of publications about pedestrian detection from 2000 to 2020. The red bars represent the publications since DCNN achieves great success. The search results are from Google Scholar using allintitle: "Pedestrian Detection."

5) **Real-time performance** Pedestrian detection is essential in real-time applications; therefore, it must meet real-time requirements. Driven by deep learning, complex models have been applied for pedestrian detection, which requires a large amount of computation and poses challenges to real-time performance.

Faced with these challenges, pedestrians can still be studied as an independent problem, although they are a category of generic object detection. In [22], Zhang *et al.* compare the state-of-the-art methods with human baseline and find that there is a large gap in the performance of occluded and small-scale pedestrian detection. Based on their conclusion, occlusion and scale variance are two key challenges affecting pedestrian detection. This conclusion is also easily obtained from the recently proposed datasets, such as CityPersons [23] where occlusion accounts for 43% and CrowdHuman [24] where occlusion accounts for 70%. The impact on the performance of pedestrian detection is obvious. The state-of-the-art method [25] obtains an 8.3% miss rate (MR) on the reasonable subset in CityPersons while 43.5% on the heavily occluded subset. Figure 4 shows the performance (measured as miss rate) comparison of the several representative works over the years, which is evaluated on the reasonable (**R**) and heavily occluded (**HO**) set of Caltech and CityPersons. The methods evaluated in Caltech perform better overall than those in CityPersons because the intra-class occlusion in CityPersons is relatively serious. In addition, deep learning techniques have significantly improved pedestrian detection. The performance of deep learning methods (e.g. MS-CNN [18]) is better than that of hand-crafted features based methods (e.g. LDCF [26]). It is also clear that the

(a) Inter-class occlusion (b) Intra-class occlusion

(c) Scale variance (d) Complex background

(e) Illumination (f) Rainy day

FIGURE 3. Some challenges of pedestrian detection. (a) Inter-class occlusion. (b) Intra-class occlusion. (c) Pedestrian of different scales, small-scale pedestrian in red box and large-scale pedestrian in blue box. (d) Complex background. The dummy in the green box. (e) Illumination. (f) Rainy day and blur.

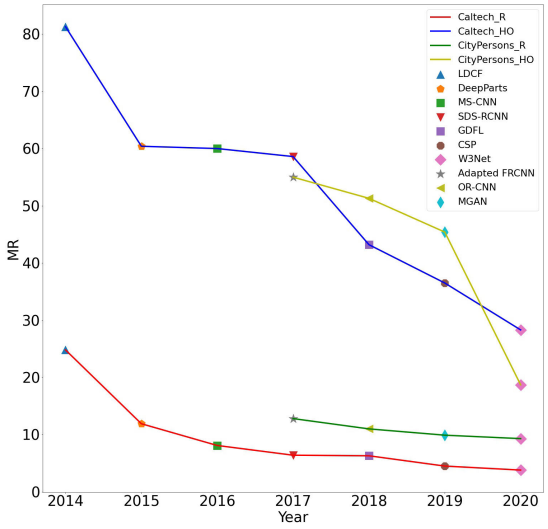


FIGURE 4. Pedestrian detection performance of representative works over the years in Caltech and CityPersons. R and HO represent reasonable subset and heavy occlusion subset, respectively.

performance on the **R** set is approaching saturation, and the gap between performance in two sets is narrowing. However, the detection performance is far behind expectations, especially when heavy occlusion exists.

As a common problem in generic object detection, scale variance has also gradually received widespread attention. Pedestrians with different scales have large intra-category variance in features that may severely hurt the performance of detectors. Statistically, over 60% of the instances in the Caltech training set have a height smaller than 100 pixels. It is very challenging to accurately locate them. To achieve better performance in real-world applications, occlusion handling and multi-scale pedestrian detection have become the mainstream. Besides detecting pedestrian as a simple category with general detectors, many studies have specially addressed occlusion and scale-variation problems.

A. COMPARISON WITH PREVIOUS REVIEWS

A series of reviews on pedestrian detection has been published, as summarized in Table 1. However, there are relatively few recent surveys focusing on the deep learning methods, except for the work by Cao *et al.* [33] who conduct a comprehensive survey on pedestrian detection from hand-crafted to deep features. However, the development of pedestrian detection is very rapid, and we mainly focus on the current research hotspots, that is, occlusion handling and multi-scale detection based on deep learning.

B. SCOPE

As shown in Figure 2, the quantity of publications on pedestrian detection is very large, and we cannot make a detailed summary for all of them. We mainly limit our focus to papers from conferences and some top journals. In addition, as we discuss above, occlusion and scale variance are the two main challenges. In this way, we can pay more attention to pedestrian detection based on deep learning to solve occlusion and scale variance and provide a relatively comprehensive summary. In addition, this survey focuses only on pedestrian detection from images.

This paper aims to review the progress of deep learning-based methods for handling occlusion and scale-variation problems in pedestrian detection over the past few years and propose future research directions. The remainder of this paper is organized as follows. Section II briefly introduces the progress in pedestrian detection over the past two decades. Section III and IV discuss methods for occlusion and scale-variation problems. Section V introduces the popular datasets and evaluation protocol for pedestrian detection. Section VI mentions about the discussion, followed by the research trends. Finally, Section VII provides a summary of the review.

II. A BRIEF REVIEW TO PROGRESS OF PEDESTRIAN DETECTION

Pedestrian detection is a fundamental research topic in computer vision. It can be divided into two main categories: hand-crafted features based and deep learning features based. In recent years, a lot of works have been proposed to improve pedestrian detection. The success relies heavily on large-scale datasets, such as KITTI [34], CityPersons [23], Caltech [35], and CrowdHuman [24]. The milestones of pedestrian detection in recent years are presented in Figure 5. The following is a brief summary of the progress in pedestrian detection.

A. HAND-CRAFTED FEATURES BASED

Before the emergence of deep learning, traditional methods applied the sliding window to obtain patches of

TABLE 1. Summary of related pedestrian detection surveys.

Title	Year	Publication	Description
Monocular Pedestrian Detection: Survey and Experiments [27]	2009	TPAMI	Overview of the current state of the art in person detection from both methodological and experimental perspectives.
Survey of Pedestrian Detection for Advanced Driver Assistance Systems [28]	2010	TPAMI	A survey about technology of pedestrian protection systems, the different methods are analyzed according to each processing stage.
Pedestrian Detection: An Evaluation of the State of The Art [29]	2012	TPAMI	A comprehensive evaluation of detectors in monocular images.
Ten Years of Pedestrian Detection, What Have We Learned [30]	2014	ECCV	A complete evaluation of trends in improving pedestrian detection.
Computer Vision and Deep Learning Techniques for Pedestrian Detection and Tracking: A Survey [31]	2018	Neurocomputing	A survey about pedestrian detection using Machine Learning and Deep Learning.
Pedestrian detection in automotive safety: Understanding state-of-the-art [32]	2019	IEEE Access	A study of different techniques used in pedestrian detection in automotive application
From Handcrafted to Deep Features for Pedestrian Detection: A Survey [33]	2021	TPAMI	A review of pedestrian detection based on handcrafted features and deep features.

different scales. Hand-crafted features such as HOG [13], LBP [47], SIFT [48], and Haar [49] were extracted to train classifiers such as SVM, AdaBoost, and random forest to filter background.

In 2003, Viola and Jones applied their VJ detector [15] to the task of pedestrian detection. In 2005, Dalal and Triggs proposed Histogram of Oriented Gradients (HOG) [13] feature descriptor for representing pedestrians, which is also a milestone of pedestrian detection. The HOG feature describes the shape and appearance of pedestrians and is insensitive to changes in light and spatial translation. However, HOG features only focus on edge and shape information, making it difficult to handle occlusion. Moreover, HOG feature is sensitive to noise owing to the characteristics of the gradient. Although some works have changed SVM to Adaboost to solve the problem of complex computation, the feature extraction is still not improved. Therefore, Dollar *et al.* proposed Integral Channel Features (ICF) [14], which combine channels of LUV, gradient magnitude and gradient histogram. These channels can be computed efficiently and capture different types of information from the input image. Compared with HOG feature, ICF has faster detection speed and better detection performance. Subsequently, it has been improved in various aspects, including ACF [16] and LDCF [26]. In 2010, Felzenswalb *et al.* proposed a deformable part model (DPM) [36] to address object deformation. Humans are divided into different parts, and the features extracted from different parts are fused to detect pedestrians. Owing to the use of HOG features and independent modeling of different pedestrian parts, DPM achieved good performance. However, DPM also has obvious limitations, such as complex feature computation, low computational efficiency, and poor performance for pedestrians with different poses.

Although the combination of hand-crafted features and classifiers was effective for some simple cases, these hand-crafted features presented limited performance. First, the detection performance for pedestrians with different appearances and poses remains poor. Second, feature extraction is inefficient, and the extracted features are too simple and not compact enough. Finally, low computational efficiency cannot meet the real-time requirements.

B. DEEP FEATURES BASED

The detection pipelines of hand-crafted features dominated computer vision until Deep Convolution Neural Network (DCNN) achieved record-breaking results in 2012. Influenced by the success of the DCNN, object detection develops rapidly. The models designed for generic object detection are applied to pedestrian detection after appropriate changes. These methods can be divided into two categories: two-stage methods and single-stage methods. In two-stage frameworks (i.e., RCNN [50], SPPNet [51], Fast RCNN [52], Faster RCNN [37]), the input image is first processed to generate region proposals by sliding window, or selective search. Subsequently, the convolutional features of these regions are extracted by CNNs, and classifiers are utilized to determine the classes of these proposals. For pedestrian detection, many methods are variations of Faster R-CNN [37], as shown in Figure 6. It generates proposals by region proposal network (RPN), and then Fast RCNN [52] leverages feature maps and proposals to detect objects. In RPN+BF [53], researchers find that the classifier in the second stage degrades the results because of insufficient resolution. They replace the classifier with boosted forests and achieve better performance. Adapted FRCNN [23] proposes key adaptations including finer feature stride and ignore region handling to enable FRCNN to obtain state-of-the-art results. MS-CNN [18] extends Faster R-CNN with a multi-scale network to deal with scale variance. Two-stage frameworks have achieved significant breakthroughs in detection performance. Nevertheless, two-stage frameworks are computationally expensive, and their detection speed is relatively slow.

After compromising on speed and accuracy, single-stage frameworks are proposed. They speed up detection by removing the region proposal generation stage. For single-stage framework, they directly predict class probabilities and bounding box offsets from full images simultaneously.

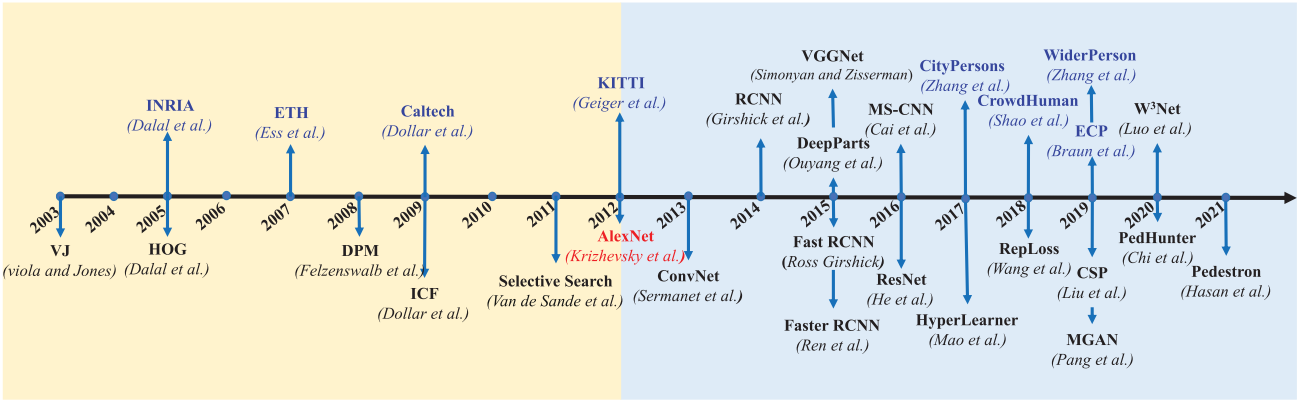


FIGURE 5. Some representative works of pedestrian detection including detection frameworks [13], [15], [17], [20], [21], [35]–[43] and datasets [13], [23], [24], [34], [35], [44]–[46].

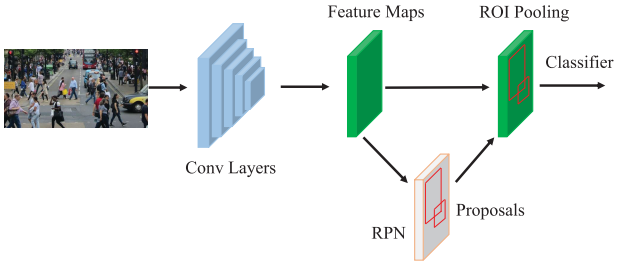


FIGURE 6. Structure of Faster R-CNN [37]. In the first stage, it generates proposals by region proposal network (RPN), and then Fast RCNN [52] leverages feature maps and proposals to detect objects.

The representative works of single-stage frameworks include YOLO series [54], [55]. To improve the accuracy of single-stage pedestrian detection, ALFNet [56] proposes an asymptotic localization fitting module to refine the default anchor boxes of SSD [57] step by step into final detection results.

Over the past two decades, pedestrian detection has evolved from hand-crafted features to deep learning features and the latter can be divided into two-stage and single-stage methods. In general, single-stage methods exhibit fast performance; however, two-stage methods can more easily achieve a more robust performance. However, with the proposal of some large datasets, such as CityPersons [23], and CrowdHuman [24], researchers have found that occlusion and scale variance limit pedestrian detection performance. Therefore, occlusion handling and multi-scale detection have become popular topics in pedestrian detection.

III. OCCLUSION HANDLING FOR PEDESTRIAN DETECTION

Occlusion can usually be categorized into inter-class and intra-class occlusions as one of the main factors affecting the detection performance. Inter-class occlusion occurs when pedestrians are occluded by other objects (i.e., trees, cars, and traffic signs). Intra-class occlusion generally occurs in crowded scenes and seriously affects performance for the following reasons. First, highly overlapped instances have similar features, which is difficult for the detector to generate different predictions for each proposal. Second, some predictions are likely to be incorrectly suppressed by NMS because instances overlap heavily. Many novel works have been proposed to solve occlusion. These methods are summarized in Table 2.

A. PART-BASED METHODS

A common solution for alleviating the occlusion problem is to focus on instance parts. Most methods handle occlusion by exploiting visible parts as additional supervision to improve detection performance. These methods adopt a strategy of learning and integrating a set of part detectors or using more distinctive body parts (e.g., the head or visible region) to learn extra supervision, reweight feature maps, or guide the anchor selection.

Before some large-scale datasets are proposed, most of the methods still solve the inter-class occlusion problem. Some works [17], [85] train ensemble models for different occlusion patterns. In [17], Tian *et al.* propose DeepParts which makes decisions based on an ensemble of extensive part detectors. Nevertheless, the computational cost is extremely high for real-time applications. To solve this problem, Zhou *et al.* [86] propose a joint learning part detector to mine part associations and reduce calculation costs. In contrast to these methods, more recent works ([39], [66], [68], [87]) aim to use visible information as auxiliary supervision to address occlusion. OR-CNN [39] proposes a part-aware RoI pooling unit to integrate the prior structural information of the human body with visibility prediction into the Fast R-CNN module of the detector. Xie *et al.* [64] propose a part spatial co-occurrence module that captures intra-part and inter-part spatial co-occurrence of different body parts using a graph convolutional network.

Several recent pedestrian detection methods utilize visible-part proposals to boost the full-body detection performance. PRNet [62] first performs visible-part estimation. Subsequently, a statistical analysis of occlusion patterns on two popular datasets is derived to bridge the gap

TABLE 2. Summary of typical methods for Occlusion Handling. CA, CH, CP and IN represent Caltech, CrowdHuman, Citypersons and INRIA.

	Methods	Backbone	Datasets	Stage	Anchor	Publication
Part-based	DeepParts [17]	-	-	-	-	ICCV2015
	OHNH [58]	-	CA/CP	Single-stage	Anchor-Based	CVPR2018
	OR-CNN [39]	VGG16	CA/CP/ETH/IN	Two-stage	Anchor-Based	ECCV2018
	PCN [59]	VGG16	CA/IN	Two-stage	Anchor-Based	BMVC2018
	PDOE [19]	VGG16	CA/CP	Two-stage	Anchor-Based	ECCV2018
	DA R-CNN [60]	ResNet-50	CH/COCOPersons	Two-stage	Anchor-Based	arXiv2019
	JointDet [61]	ResNet-50	CP/CH/CA	Two-stage	Anchor-Based	AAAI2020
	PedHunter [25]	ResNet-50	CA/CP/CH	Two-stage	Anchor-Based	AAAI2020
	PRNet [62]	ResNet-50	CA/CP/ETH	Single-stage	Anchor-Based	ECCV2020
	FC-Net [63]	ResNet-50	CA/CP	Two-stage	Anchor-Based	ITS2020
	PSC [64]	VGG16	CA/CP	Two-stage	Anchor-Based	arXiv2020
	V2F-Net [65]	ResNet-50	CH/CP	-	Anchor-Based	arXiv2021
Attention-based	FRCNN+ATT [66]	VGG16	CA/CP/ETH	Two-stage	Anchor-Based	CVPR2018
	GDFL [67]	VGG16	CA/KITTI/IN	Single-stage	Anchor-Based	ECCV2018
	SSA-CNN [6]	VGG16	CA/CP	Two-stage	Anchor-Based	CVPR2019
	MGAN [68]	VGG16	CA/CP	Two-stage	Anchor-Based	ICCV2019
	AGNN [69]	VGG16	CA/CP	Two-stage	Anchor-Based	PR2020
	Beta RCNN [70]	ResNet-50	CP/CH	Two-stage	Anchor-Based	NIPS2020
	PS-RCNN [71]	ResNet-50	CH/WP	Two-stage	Anchor-Based	ICME2020
	PED [72]	-	CP/CH	Single-stage	Anchor-Free	arXiv2021
Loss-based and Post-processing	Reploss [21]	ResNet-50	CP	-	-	CVPR2018
	OR-CNN [39]	VGG16	CA/CP/ETH/IN	Two-stage	Anchor-Based	ECCV2018
	Adaptive NMS [41]	VGG16	CP/CH	-	Anchor-Based	CVPR2019
	SG-Det [73]	ResNet-50/ResNet-101	KITTI/CP	Two-stage	Anchor-Based	arXiv2019
	PBM [74]	Resnet-50/VGG16	CP/CH	Two-stage	Anchor-Based	CVPR2020
	NOH NMS [75]	ResNet-50	CP/CH	-	Anchor-Based	ACM MM2020
	APD [76]	ResNet-50/DLA34	CP/CH	Single-stage	Anchor-Free	TMM 2020
	MAPD [77]	ResNet-50	CP/CH	Single-stage	Anchor-Free	Neurocomputing
	LLA [78]	ResNet-50/ResNet-101	CP/CH	-	Based/Free	arXiv2021
	NMS-ped [79]	Resnet-50	CA/CP	-	Anchor-Based	arXiv2021
Others	EMD-RCNN [80]	ResNet-50	CP/CH/COCO	-	Anchor-Based	CVPR2020
	TFAN [81]	ResNet-101	CA/KAIST	Single-stage	Anchor-Based	CVPR2020
	W^3Net [40]	ResNet-50	CA/CP	Single-stage	Anchor-Free	CVPR2020
	CaSe [82]	ResNet-50	CP/CH	Two-stage	Anchor-Based	ECCV2020
	AEVB [83]	ResNet-50	CP/CH	-	Based/Free	CVPR2021
	IterDet [84]	ResNet-50	CH/WP	-	Anchor-Based	arXiv2021

between the visible and full-body anchors. The new proposed module refines the final full-body localization. Similarly, V2F-Net [65] first detects the visible regions of all pedestrians and then estimates the full-body box from the visible box. To improve the accuracy of full-body estimation from visible region, the feature of detected visible region is utilized to compute its response on each part to determine whether it is visible in the given visible box. In contrast to using visible proposals to guide the detection of full-body, some other methods utilize different branches to generate proposals separately. Bi-Box [19] proposes to perform the full-body estimation and visible-part estimation simultaneously so that the visible part estimation can be fused with the full-body estimation to improve the detection performance. In [88], two different branches generate visible-part proposals and full-body proposals separately. The proposed mutual-supervised feature modulation module calculates the similarity loss between full-body boxes and visible-body boxes to learn more robust feature representations of occluded pedestrians. In [74], the pair RPN generates visible proposals and full-body proposals simultaneously. The aggregate pair of proposal features are utilized to predict pairs of BBoxes.

In other novel methods, additional visibility classifiers are used to incorporate the predicted confidence into the final score. In [58], Noh *et al.* use the confidence of the visible parts to correct the final detection confidence of a pedestrian to address the low confidence of occluded pedestrian. Similarly, PCN [59] also divides the pedestrian box into several part grids and produces score maps, but it uses LSTM to process different permutations of part scores as sequences. Some methods utilize score-level fusion to further improve the final score. Bi-box [19] and MSFMN [88] construct visible-part and full-body branches and then fuse the scores of two branches during inference.

As another intuitive clue in a crowd, the head generally has less overlap. The head features are more stable and robust than the human body, which can be used as auxiliary information to full body prediction to boost pedestrian detection performance. In DA R-CNN [60], double anchor RPN generates proposals in pairs of heads and bodies simultaneously. A proposal crossover strategy is utilized to generate high-quality proposals for both parts. In addition, features of heads and bodies are aggregated efficiently to make the final prediction more reliable. In JointDet [61],

RPN only generates head proposals, then they apply a statistical head-body ratio on these head proposals to obtain full-body proposals. A relationship discriminating module is designed to learn to discriminate the relationships between the head-body pairs and recalls suppressed body detections by head detections. In [89], Lin *et al.* propose PedJointNet, which incorporates the prediction of head-shoulder region and full-body region into a unified architecture. Different from DA-RCNN and JointDet, proposals in PedJointNet are produced independently in two branches. Then an adaptive weighted fusion layer is used to fuse the detection of two branches adaptively. Different from the above methods, Chi *et al.* [25] design a mask guidance module to enhance the feature representation of the backbone by using head information. In HBAN [90], Lu *et al.* propose an extra branch to conduct semantic head detection parallel with traditional body detection to improve the performance and robustness to occlusion.

B. ATTENTION-BASED METHODS

Attention mechanism is originally used in machine translation and has become an important concept in neural networks. It has been widely used in natural language processing and computer vision. In a crowd, the full-body detector would be deceived by the blurred features of occluded pedestrians. Therefore, attention mechanisms are employed to enable the detectors to focus on the features of the visible parts. Some methods use attention mechanisms to enhance the features of pedestrians and suppress background, while others leverage semantic segmentation features with convolutional feature maps to boost pedestrian detection accuracy.

Zhang *et al.* [66] find that many channel features are localizable and often correspond to different body parts. Hence, they propose a channel-wise attention mechanism that can focus more on visible parts to handle occlusion. They add a separate part attention net on Faster R-CNN to generate a channel-wise attention vector to reweight the channel features to handle various occlusion patterns, as shown in Figure 7(a). In [91], Guo *et al.* leverage a semantic segmentation map from the depth images to guide the reweighting of the convolutional features extracted from RGB images, as shown in Figure 7(b). In GDFL [67], Lin *et al.* leverage scale-aware pedestrian attention masks and a zoom-in-zoom-out module to improve the capability of the feature maps to detect small and occluded pedestrians. In AGNN [69], Zou *et al.* propose an attention-guided network that guides LSTM to focus on the important feature sequences of pedestrians. Transformers are introduced as a new attention-based building block for machine translation. Carion *et al.* apply it to object detection and propose the DETR [92] that views object detection as a direct set prediction problem. It replaces hand-designed components such as NMS and anchors using the transformer architecture. However, DETR is unsuitable for pedestrian detection in a crowd. In [72], they find that cross attention is not suitable for crowd detection, so they propose a RF (Rectified attention Field) module to rectify it. In addition, they also propose a new decoder for DETR, which significantly improves DETR for pedestrian detection. In [70], Xu *et al.* adopt an attention mechanism with 2D beta distribution to highlight the features of visible parts and suppress other noise simultaneously, which could induce the network to pay more attention to the discriminative features and achieve better localization accuracy and higher confidence.

Some methods use attention mechanism to enhance pedestrian features of pedestrians and suppress background. In MGAN [68], Pang *et al.* introduce a novel mask-guided attention network, that emphasizes visible pedestrian regions while suppressing the occluded parts by modulating extracted features. Similarly, Zhang *et al.* [63] propose a self-activation module that can reinforce the features in the visible parts while suppressing those in occluded regions. Ge *et al.* [71] propose a PS-RCNN with two parallel RCNN modules. The P-RCNN module is used for the first round of detecting instances with non or slightly occluded instances. Then the features of the heavily occluded pedestrians are highlighted by suppressing the detected pedestrians with human-shaped masks. Then the S-RCNN module is used to detect the rest missed pedestrians. Finally, they ensemble the outputs from these two RCNNs.

In addition, some works leverage semantic segmentation to boost pedestrian detection accuracy. Zhou *et al.* [6] design a multi-task network to co-learn semantic segmentation and pedestrian detection with weak box annotations. The semantic segmentation feature map is connected to the corresponding convolution feature map to provide more discriminating features for pedestrian detection. Brazil *et al.* [5], and Du *et al.* [93] leverage additional semantic segmentation to supervise pedestrian detection. SDS RCNN [5] presents a multi-task infusion framework for joint supervision on pedestrian detection and semantic segmentation while segmentation in [93] is an optional module to improve the performance.

C. LOSS-BASED AND POST-PROCESSING METHODS

Generally, object detectors employ non-maximum suppression (NMS) as a post-processing strategy. Several previous works have investigated improving NMS for generic object detection [94]–[96]. However, it is still very challenging for crowded detection using these NMS. In generic object detection, the traditional pipeline works well because the instance rarely stands with highly overlapped cases. However, an instance is often highly overlapped with multiple instances in crowd scenes, which will be ambiguous for NMS. Usually, it is difficult for the traditional pipeline to choose bounding boxes in a crowd. As shown in Figure 8, it is challenging to distinguish the bounding boxes generated by multiple pedestrians occluded together using a rigid threshold because a lower threshold will increase the miss rate while a higher threshold will keep more false positives. Improving NMS for occluded pedestrian detection is an open problem, as most existing pedestrian detectors still employ traditional post-processing strategies. In [21], [39], the effect of the NMS threshold for crowded detection is explored. To alleviate

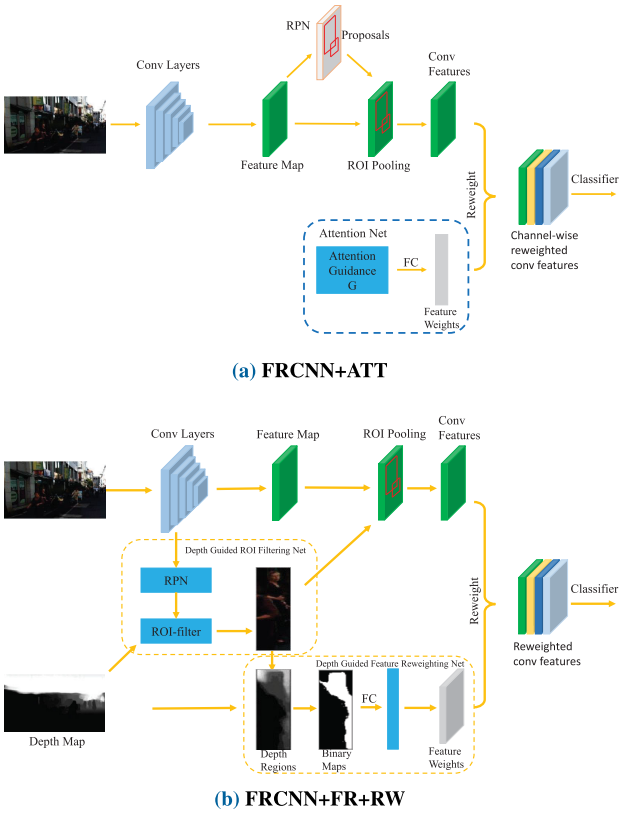


FIGURE 7. The comparison of two attention-based methods. In FRCNN+ATT [66], they propose a channel-wise attention mechanism. In FRCNN+FR+RW [91], they utilize depth information to reweight the convolutional features.

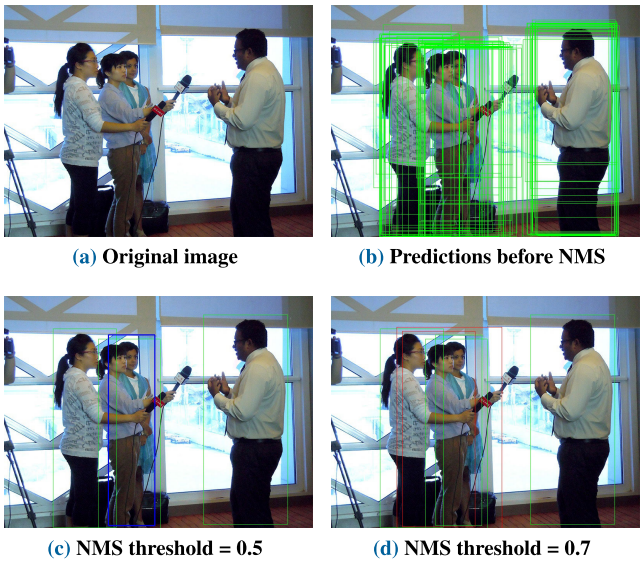
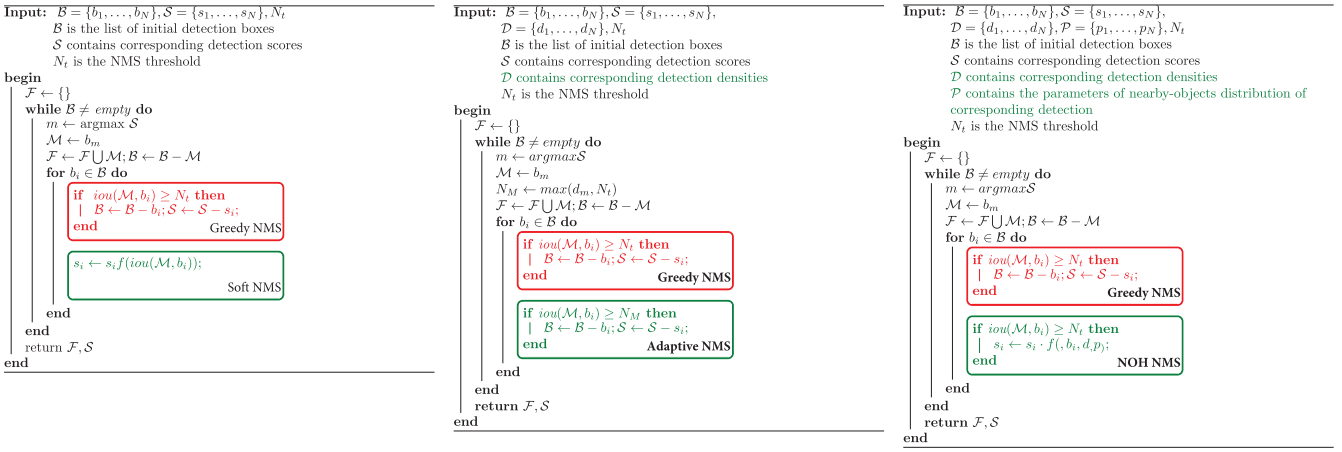


FIGURE 8. Illustration of results using greedy-NMS with different thresholds. The bounding boxes in (b) are generated using Faster R-CNN. The blue boxes show the false negative, while the red ones highlight false positives. In a crowd, a lower NMS threshold may remove true positives (c) while a higher NMS threshold may increase false positives (d).

the influence of rigid threshold to detection, some advanced NMS strategies ([41], [74], [75], [97]) are proposed.

Soft NMS [95] tries to degrade the score of nearby highly overlapped proposals instead of eliminating them, but just like Greedy NMS, it still blindly penalizes the highly overlapped boxes. Some works apply additional information (i.e., density, diversity) beyond location and many object proposals to NMS to solve rigid thresholds. Adaptive NMS [41] uses a larger one of the predicted density around the instance and the initial threshold as the dynamic suppression threshold to refine the bounding boxes, which means the threshold rises as instances occlude each other and decays when instances appear separately. However, novel loss is still required to achieve better performance. Although Adaptive NMS can predict the density of proposals, it is not aware of the locations and spread of the crowded regions, so in [75], Zhou *et al.* propose NOH NMS to pinpoint the objects nearby each proposal with a Gaussian distribution, which is aware of the existence of other nearby objects to address the rigid NMS threshold problem in pedestrian detection. In APD [76], Zhang *et al.* propose an attribute-aware pedestrian detector to explicitly model semantic attributes of the pedestrian in a high-level feature detection manner. Meanwhile, they apply an attribute map that includes density and diversity information to NMS to reject the false-positive results adaptively. In MAPD [77], Wang *et al.* improve the APD and propose a novel multi-attribute NMS algorithm based on density and id information, which can adaptively distinguish predicted boxes of different pedestrians. In [74], Huang *et al.* propose R^2 NMS. They find that the IOU between the boxes of full-body is large while the IOU between boxes of the visible area is relatively small in occlusion. Therefore, the relatively low IOU threshold can effectively remove the redundant boxes and avoid many false positives based on the visible area. Some essentially identical NMS algorithms are shown in Figure 9, which shows the similarities and differences between different NMS algorithms. Furthermore, other NMS strategies have been proposed to adapt to their own methods such as joint NMS [60], set NMS [80], Beta NMS [70], SG NMS [73], pos NMS [98] and CAS NMS [82].

Additionally, some works propose novel loss to address pedestrian detection in a crowd. OR-CNN [39] proposes aggregation loss to enforce proposals to be close to the corresponding objects and to minimize the internal region distances of proposals associated with the same objects. RepLoss [21] introduces a bounding box regression loss to not only push each proposal to reach its designated target but also to keep it away from other surrounding objects. In [79], Luo *et al.* propose NMS-Loss, which pulls predictions with the same objects close to each other and pushes predictions with different objects away from each other so that the false detections caused by NMS can be reflected in the loss function. In [82], Xie *et al.* propose an approach by leveraging pedestrian count and proposal similarity information within a two-stage pedestrian detection framework. Moreover, they introduce a count-weighted detection loss function that assigns higher weights to the detection errors occurring at highly overlapping pedestrians. LLA [78] proposes a loss



(a) Soft NMS **(b)** Adaptive NMS **(c)** NOH NMS

FIGURE 9. Pseudo codes of Greedy NMS, Soft NMS [95], Adaptive NMS [41] and NOH NMS [75].

as a new label assignment strategy to boost the performance in crowd scenarios.

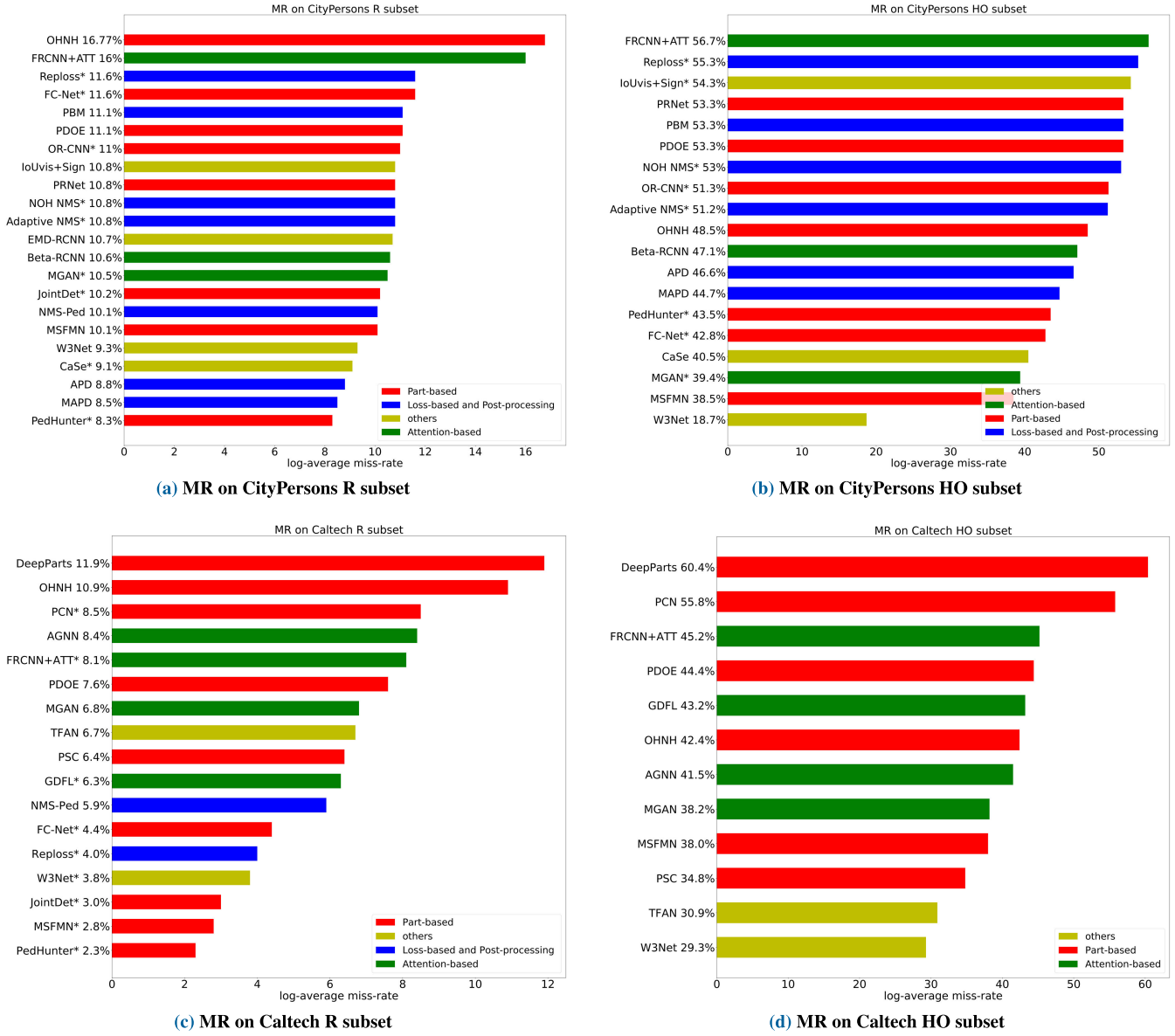
D. OTHERS

In addition to aforementioned methods, some other novel methods are also effective to address occlusion. In IterDet [84], Danila *et al.* propose an iterative detection scheme. In each iteration, a new subset of objects is detected, and all boxes detected in previous iterations are considered in the current iteration to ensure that the same objects will not be detected repeatedly. W^3Net [40] decouples the pedestrian detection task into where, what, and whether problem directing against pedestrian localization, scale prediction, and classification by generating a bird view map to address occlusion. In severe occlusion, it is difficult for single image to provide effective features. Therefore, local temporal context is utilized to enhance the feature representations of heavily occluded pedestrians in TFAN [81]. Chu *et al.* [80] utilize the concept of multiple instance prediction and propose a method which let each proposal predict an instance set. In [83], Zhang *et al.* redefine single-stage pedestrian detection as a variational inference problem and propose a auto-coding variational bayesian algorithm to optimize the problem. In [99], Lu *et al.* propose a visible IoU which can select positive samples correctly to improve the training results. Moreover, a box sign predictor is designed at the final stage to improve localization accuracy.

Summary Occlusion is a critical challenge in pedestrian detection. The performance of the different methods for occlusion handling on CityPersons [23], and Caltech [35] are shown in Figure 10. It is clear that the detection performance is still far from satisfactory when occlusion exists. Therefore, solving the occlusion problem is critical for improving the overall pedestrian detection performance. Occlusion can be categorized into inter-class occlusion and intra-class occlusion. Inter-class occlusion occurs when pedestrians are occluded by other obstacles such as trees, cars, and traffic signs. The background features confuse the model, leading to a high missing rate. The most important information for addressing the inter-class occlusion is the visible information. Part-based methods utilize this information to learn extra supervision, reweight feature maps, guide the anchor selection or generate part proposals to improve the quality of full-body prediction. Other methods utilize attention mechanisms to enhance the features of the visible parts while suppressing the features of other obstacles or background. Intra-class occlusion is also called crowd occlusion and occurs in crowded scenes where pedestrians have large overlaps with each other. Highly overlapped instances have very similar features, which makes it difficult for detector to generate different predictions. As a result, detectors may give a lot of positives in overlapped areas. Therefore, some methods propose additional penalties to remove the redundant BBoxes. On the other hand, the highly overlapped BBoxes may also be suppressed by non-maximum suppression (NMS). To solve this problem, some methods utilize head proposals or visible proposals to recall suppressed body detections. Besides, variants of NMS are proposed to soften the sensitivity of NMS threshold in a crowd, which is helpful for removing redundant BBoxes or recalling suppressed detections. Although many works have been proposed to solve occlusion, there is still a huge gap between detectors and human. As shown in Figure 10, the performance on the reasonable set is approaching saturation, and the gap between different methods is narrowing. However, the detection performance under heavy occlusion is far behind expectations. In term of different methods, part-based and loss-based methods are popular at present. In general, part-based method is more effective than other methods.

IV. MULTI-SCALE PEDESTRIAN DETECTION

Multi-scale object detection is one of the basic challenges in computer vision. Objects have a large variance of scales, which is critical for accurate detection owing to the difference



(a) MR on CityPersons R subset

(b) MR on CityPersons HO subset

(c) MR on Caltech R subset

(d) MR on Caltech HO subset

FIGURE 10. MR of different methods for occlusion handling on CityPersons [23] and Caltech [35]. '*' in Caltech means the methods use new annotations from [22], and in CityPersons means the images used in methods are $1.3_{\times}$ the original image size.

of features between small and large instances. The existing methods are not friendly to small-scale pedestrian detection. Firstly, large downsampled factors lead to the loss of information of small objects. Secondly, large receptive field contains many surrounding features, which may be blurred for detector. Lastly, most detection methods do not achieve the balance between deep and shallow feature maps in terms of semantic and localization information. Therefore, many methods have been developed to solve these problems. Table 3 provides an overview of some methods whose results are published on the Caltech and CityPersons pedestrian detection benchmark and Figure 11 shows the timeline of multi-scale detection methods.

A. LEVERAGE MULTI-SCALE FEATURE FUSION

In generic object detection, the main idea to address scale variance is to use multi-scale feature map for detection. The multi-scale image pyramid [108] is a common strategy to improve the detection performance. It uses images of different scales as input to extract multi-scale feature maps and detect instances independently (Figure 12(a)). These methods are effective but suffer from the problem of long inference time. With hand-crafted features replaced by deep features, most methods extract high-level semantic features for regression and classification [37], [51], [52] (Figure 12(b)). However, detection based on single-scale feature maps is not sufficiently robust to scale variance, which leads to

TABLE 3. Some methods for multi-scale pedestrian detection. Note that the symbol '-' means the results are unavailable.

Methods	Backbone	Stage	Anchor	Citypersons			Caltech			Publication
				Small	Middle	Large	Near	Middle	Far	
SADR [100]	VGG-16	Two-stage	Anchor-based	-	-	-	-	-	-	ACCV2016
MS-CNN [18]	VGG-16	Two-stage	Anchor-based	-	-	-	2.6	49.13	97.23	ECCV2016
FRCNN+seg [23]	VGG-16	Two-stage	Anchor-based	8	6.7	22.6	0	51.8	100	CVPR2017
SAF R-CNN [38]	VGG-16	Two-stage	Anchor-based	-	-	-	0	51.83	100	TMM2017
FDNN+SS [93]	VGG-16	Two-stage	Anchor-based	-	-	-	2.82	33.15	77.47	WACV2017
SAM-RCNN [101]	VGG-16	Two-stage	Anchor-based	-	-	-	-	-	-	arXiv2018
TLL+MRF [102]	ResNet-50	Single-stage	Anchor-free	-	-	-	0.67	25.55	67.69	ECCV2018
GDFL [67]	VGG-16	Single-stage	Anchor-based	-	-	-	-	32.5	-	ECCV2018
ADM [103]	ResNet-50	Two-stage	Anchor-based	-	-	-	0.41	30.82	74.53	TIP2018
CSP [20]	ResNet-50	Single-stage	Anchor-free	16	3.7	6.5	-	-	-	CVPR2019
MagnifierNet [104]	ResNet-101	Two-stage	Anchor-based	12.6	5.5	7.7	-	-	-	ICPR2020
PRF-Ped [105]	ResNet-50	Single-stage	Anchor-free	12.9	3.9	5.8	-	-	-	ICPR2020
DHRNet [106]	DHRNet-W18	Two-stage	Anchor-based	13.43	2.69	6.21	-	-	-	ICPR2020
W^3Net [40]	ResNet-50	Single-stage	Anchor-free	-	-	-	-	-	51.05	CVPR2020
AP^2M [107]	ResNet-50	Single-stage	Anchor-free	15.3	3.5	5.3	-	-	-	AAAI2021

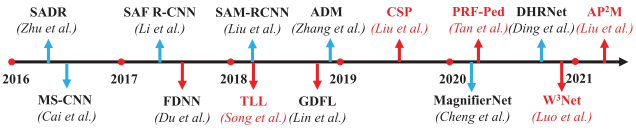


FIGURE 11. Timeline of multi-scale pedestrian detection. The red font represents anchor-free, the black font represents anchor-based, the red arrow represents single-stage methods, and the blue arrow represents two-stage methods.

insufficient and inaccurate information for detecting smaller objects. To solve this problem, some methods e.g., SSD [57] and MS-CNN [18], predict objects at multiple layers of the feature hierarchy independently (Figure 12(c)). However, the feature maps of different depths exhibit significant semantic differences. The shallow feature map has a strong activation effect on small-scale objects but lacks rich semantic information. Deeper ones tend to encode large instances while ignoring small instances and losing more accurate localization information. Therefore, researchers have explored various effective multi-scale feature representations. As the representative model architectures to generate pyramidal feature representations, Feature pyramid network (FPN) [109] (Figure 12(d)) proposes lateral connections and top-down pathway to combine multi-scale features. This structure can combine low-resolution feature maps with strong semantic information and high-resolution feature maps with rich spatial information under the premise of increasing less computation. However, there is a long path from low-level structure to topmost features, increasing difficulty to access accurate localization information. To improve this problem, PANet [110], which is originally used for segmentation, adds an additional bottom-up path augmentation to shorten the information path and further enhance the feature hierarchy with accurate localization signals in low-level layers. Then, YOLOv4 [111] and YOLOv5 use this structure for detection (Figure 12(e)). More recently, some variations, such as Bi-FPN [112] and NAS-FPN [113] also develop more novel network structures. NAS-FPN uses neural architecture search algorithm to design a new pyramidal representation, whereas Bi-FPN improves the connection of PANet, and introduces a simple attention mechanism at the connection point.

The aforementioned feature fusion structures play a great role in generic object detection. Some works like [67], [103], [105], [114], [115] borrow from these ideas and propose some new fusion strategies to adapt to pedestrian detection. Some typical frameworks are shown in Figure 13. Zhang *et al.* [103] propose a method that uses an active detection model based on a set of initial bounding box proposals, executes sequences of coordinate transformation actions across multi-layer features representations to deliver accurate prediction of pedestrian locations. In GDFL [67], they introduce a scale-aware pedestrian attention mask and a zoom-in-zoom-out module to improve the capability of the feature maps to identify small pedestrians. In [115], Xie *et al.* propose a feature enrich unit which involves semantic segmentation feature learning to enrich features to improve detection. In SADR [100], Zhu *et al.* introduce the deconvolutional layers to adaptively upsample the feature map for small pedestrians. In addition, they fuse features from multiple layers to provide both local characteristics and global semantic information, which improves the detection performance. Du *et al.* [93] propose F-DNN, which leverage SSD to generate pedestrian candidates and fuse multiple DNNs in parallel to detect pedestrians by using a soft-reject strategy. In PRF-Ped [105], Tan *et al.* present a bidirectional feature enhancement module (BFEM), which enhances the semantic information of low-level features and enriches the localization information of high-level features. In [116], Zhang *et al.* build a cross-scale feature aggregation module, which merges a top-down path, lateral connections and a bottom-up augmented path by addition to adaptively aggregating multi-scale context information from convolutional layers at adjacent scales to generate more discriminative features. Subsequently, a newly proposed scale-aware hierarchical network uses feature maps of different scales to detect pedestrians of different scales, respectively.

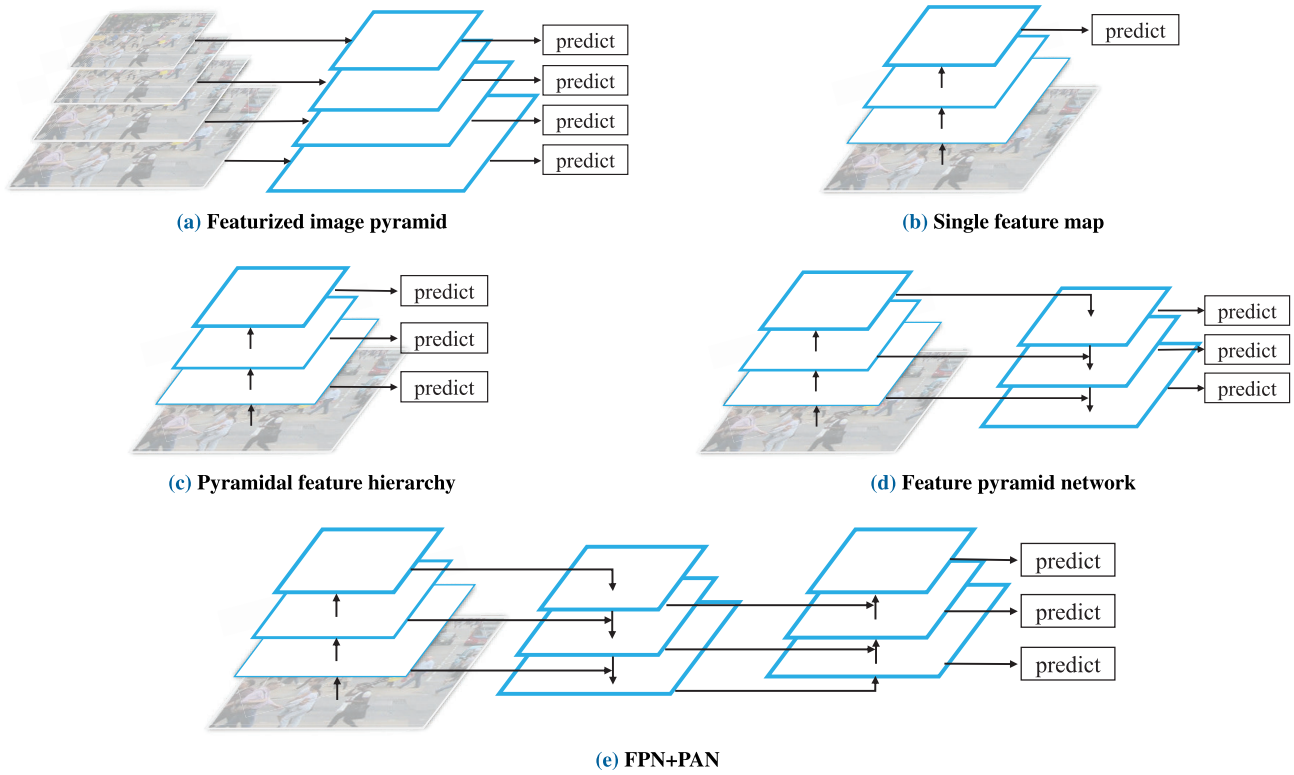


FIGURE 12. (a) Building feature pyramid from image pyramid to detect objects of different scales. (b) Detecting on a single feature map. (c) Using ConvNet to build pyramidal feature hierarchy. (d) Feature pyramid network(FPN). (e) FPN+PAN.

In general, multi-scale feature fusion considers both shallow localization information and deep semantic information, which can effectively improve the performance of small-scale pedestrian detection. However, existing multi-scale detection methods also increase the computation cost, and compromise the real-time performance.

B. ANCHOR-FREE METHODS

Anchors play an important role in object detection. Many state-of-the-art object detection methods have been designed based on the anchor mechanism, which is very unfriendly to small object detection. The existing design of anchor is difficult to balance the contradiction between recall and computation cost of small objects. These methods also lead to an extreme imbalance between positive samples of small objects and large objects, which makes the model pay more attention to the detection performance of large objects, while ignoring the performance of small objects. In addition, the use of anchor introduces extra hyperparameters, such as the number of anchors, aspect ratio and size, which makes it difficult to train the network. Anchor-based methods can achieve satisfactory performance, but it also brings extra computing overhead. In recent years, anchor-free mechanism has become a research hotspot and has achieved good results in small object detection.

In [102], Song *et al.* propose a novel method integrated with somatic topological line localization and temporal feature aggregation for detecting multi-scale pedestrians. In [20], Liu *et al.* propose CSP, which selects and fuses the optimal combination of multi-scale feature maps from each stage and simplify pedestrian detection to a straightforward center and scale prediction task, which breaks the limitation of anchor-based methods and eliminates the complex post-processing of keypoint pairing based detectors. After that, Wang further refines the CSP in [117]. CSP uses the vanilla ResNet50 to extract multi-level feature maps and then simply fuses them into a single one for predicting. Although CSP achieves brilliant accuracy, it ignores the fact that the difference of semantic information of feature map with different depth may harm the effect of feature fusion. Motivated by these observations and analysis of feature fusion, Cai *et al.* [118] propose PP-Net, an anchor-free method for center-based pedestrian detection. They leverage a novel deep guidance module to tackle the dilemma of information sparsity on the top-down pathway of standard FPN architecture and fuse FPN structure and the output of DGM to solve the problem of ignoring the semantic gap between feature maps of different depths when directly fusing them in CSP. In W^3Net [40], Luo *et al.* model the dependency between depth and scale to generate depth-guided scales to address scale-variation problems.

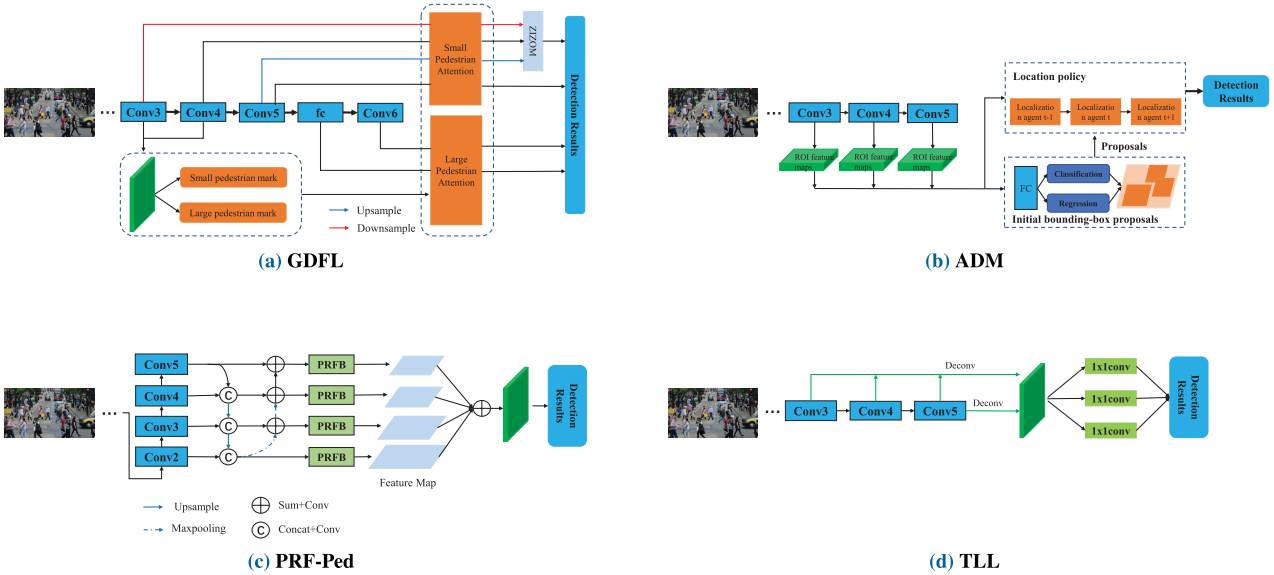


FIGURE 13. Different frameworks for multi-scale pedestrian detection. (a) GDFL [67] (b) ADM [103] (c) TLL [102] (d) PRF-Ped [105].

C. DATA AUGMENTATION

In recent years, deep learning methods, which rely heavily on datasets, have become increasingly popular. Therefore, the quality and quantity of data have a great impact on detection performance. In some datasets, the distribution of objects with different scales is not balanced, which may cause inconsistent detection performance for objects with different scales. Using data augmentation strategies can enrich the diversity of datasets, so as to enhance the robustness and generalization of the frameworks. In early studies, strategies such as elastic distortions, random pruning, and translation have been widely used in object detection. In recent years, some state-of-the-art methods use other data augmentation strategies to improve detection performance, for example, the standard horizontal image flipping used in Fast R?CNN [52] and CSP [20], the random adjustment of exposure and saturation in the HSV color space used in YOLO [54] and YOLOv2 [55]. In addition, more novel data augmentation strategies (Mixup [119], Cutout [120], CutMix [121], Mosaic [111]) are also widely used. In [122], popular data augmentation methods are evaluated in terms of model robustness, and then they propose a data augmentation scheme that uses stylization but only patches of the original image.

Data augmentation is a simple and effective method to improve the performance of small object detection. It can effectively improve the generalization ability of the network. However, it also brings an increase in computation cost. In addition, if the data augmentation causes a large difference in sample distribution, the model performance may be damaged, which also brings challenges.

D. OTHERS

In addition to several categories summarized above, there are many other novel methods in the field of multi-scale pedestrian detection. In recent years, with the increase of computing power, more and more networks are using cascading thinking to improve performance. In WIDER pedestrian challenge, many methods use Cascade RCNN [123] as basic detection framework and add some powerful structures to achieve better performance. Another idea is to leverage parallel branches to detect pedestrians at different scales separately. In [38], Li *et al.* propose SAF R-CNN which incorporated a large-size sub-network and a small-size sub-network into a unified architecture, final results are outputs of the two weighted sub-networks with weights learned from the scale-aware weighting layer. In [106], Ding *et al.* construct multiple branches in DHRNet to generate scale-specific feature maps. Then, different branches are used to detect objects of different scales.

Summary The limitations of small-scale pedestrian detection are obvious. Large-scale instances can provide rich information, while small-scale instances are difficult to recognize. The best solution for scale variance is to fuse multi-scale feature maps in network structure. Various effective multi-scale feature representations are explored to handle scale-variation problems. Besides, some other methods leverage different data augmentation strategies to reduce the impact of unbalanced data distribution on detection performance, while anchor-free methods remove the anchor design to reduce the influence of anchors on small-scale pedestrian detection. In addition, other tips can also be helpful, e.g., replacing RoI Pooling with RoI Align, changing the design of anchors, and multi-scale training. All these methods are effective and have achieved good performance.

V. DATASETS AND PERFORMANCE EVALUATION

A. DATASETS

During the last decades, significant efforts have been made to develop various methods for learning supervised

pedestrian detectors. Therefore, their success depends significantly on large-scale datasets. In contrast to the generic object detection datasets, some datasets specially used for pedestrian detection have been collected over the years, such as MIT [124] INRIA [13], ETH [44], USC [125], [126], TUD-Brussels [127], and Daimler [27]. In addition, some datasets, such as KITTI [34], Caltech [35], CityPersons [23], and ECP [45] are acquired by sensors mounted on actual vehicles, so they are more suitable for solving autonomous driving tasks. In recent years, more diverse datasets, e.g., CrowdHuman [24], WidePedestrian and WiderPerson [46] are proposed. These datasets are more diverse and more dense, which can greatly help improve the robustness and generality of the network. The attributes of these datasets are summarized in Table 4, and the selected example images are shown in Figure 14. Table 2 and Table 3 state that Caltech [35], CityPersons [23], CrowdHuman [24], and KITTI [34] are widely used for validation; therefore, a detailed introduction is provided here.

Caltech [35] The Caltech Pedestrian Dataset consists of approximately 10 hours video taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute-long segments) with a total of 350,000 bounding boxes and 2300 unique pedestrians are annotated. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels.

CityPersons [23] The CityPersons dataset is a subset of Cityscapes which only consists of person annotations. There are 2975 images for training, 500 and 1575 images for validation and testing. The density of pedestrians in the dataset is very high, and the average number of pedestrians in an image is 7. What's more, scenarios of the datasets are rich, and it contains multiple occlusion cases.

KITTI [34] The KITTI dataset is the popular dataset for evaluating computer vision algorithms in autonomous driving. The dataset is used to evaluate the performance of computer vision technologies such as Stereo, Optical Flow, Visual Odometry, 3D Object Detection, and 3D Tracking. KITTI contains real image data from urban, rural and highway scenes, with up to 15 cars and 30 pedestrians per image, with varying degrees of occlusion and truncation.

CrowdHuman [24] CrowdHuman is a benchmark dataset to better evaluate detectors in crowd scenarios. The CrowdHuman dataset is large, rich-annotated and contains high diversity. There is a total of 470K human instances from train and validation subsets and 23 persons per image, with various kinds of occlusions in the dataset. Each human instance is annotated with a head bounding-box, human visible-region bounding-box and human full-body bounding-box.

B. EVALUATION METHODS

There are two main criteria for evaluating the performance of detection model: average precision (AP) and log miss rate (MR).

Average Precision AP is the most commonly used metric in generic object detection and is typically evaluated in a category-specific manner. Before explaining the calculation of AP, we first explain how to choose true positives. A predicted detection is regarded as a True Positive (TP) If

(1) The predicted category equals the ground truth label;

(2) The IOU(Intersection Over Union) between the predicted BBox b_{pre} and the ground truth b_{gt}, as shown in (1) is not smaller than a predefined threshold λ.

$$IOU(b_{pre}, b_{gt}) = \frac{area(b_{pre} \cap b_{gt})}{area(b_{pre} \cup b_{gt})} \tag{1}$$

Otherwise, it is considered as a False Positive (FP). The specific algorithm can be found in [11]. The confidence level is usually compared with a threshold β to determine whether the prediction is accepted.

The precision and recall curve is computed from output of the network. AP is computed separately for each of the object classes, based on the Precision and Recall curve. Recall is defined as the proportion of all ground truths that are from the true positives. Precision is the proportion of all predictions that are from the true positives. Their calculations can be shown in (2) and (3).

$$Recall = \frac{N_{TP}}{(N_{TP} + N_{FN})} \tag{2}$$

$$Precision = \frac{N_{TP}}{(N_{TP} + N_{FP})} \tag{3}$$

For a given task and class, the results returned by a detector are ranked by confidence in decreasing order. Each detection is determined as TP or FP according to the algorithm in [12]. Based on the TP and FP detections, the precision P(β) and recall R(β) can be computed as a function of the confidence threshold β. P-R curve can be obtained by varying the confidence threshold, and then the Average Precision (AP) can be found.

Log-average miss rate The log-average Miss Rate is a bit similar to recall and refers to the objects that are not detected. MR is defined as the ratio of the number of False Negatives (N_{FN}) to the number of ground truth (N_{GT}) in test set as

$$MR = \frac{N_{FN}}{N_{GT}} \tag{4}$$

In addition, false positives per image (FPPI) can be calculated by dividing False Positives (N_{FP}) by the number of images (N) as

$$FPPI = \frac{N_{FP}}{N} \tag{5}$$

Similar to the PR curve, miss rates against false positives per image (FPPI) can be plotted in the log-space by varying the detection confidence threshold. Finally, the log-average miss rate (lower is better) is calculated by averaging miss rates under 11 FPPI equally spaced in $[10^{-2} : 10^{0}]$.

TABLE 4. Summary of pedestrian datasets. 'full' means fully-body bounding-box, 'visible' means visible-body bounding-box, and 'head' means head bounding-box.

Dataset Name	Publish Year	Total Images	Pedestrians	Annotations			Description
				Full	Head	Visible	
MIT [124]	2000	924	924	✓			An earlier published pedestrian dataset.
INRIA [13]	2005	2573	3542	✓			A widely used dataset for pedestrian detection.
ETH [44]	2007	1803	12K	✓			A pair of images in busy shopping streets.
TUD-Brussels [127]	2009	508	1326	✓			Pedestrians in the inner city of Brussels.
Daimler [27]	2009	112K	111K	✓			Including PNG image, disparity map and ground truth shape,rig in an urban environment.
Caltech [35]	2009	250K	289K	✓		✓	A widely used dataset which includes larger amount data and complete annotations.
KITTI [34]	2012	15K	9K	✓			A very popular dataset for computer vision tasks.
CityPersons [23]	2017	5K	32K	✓		✓	A rich and diverse pedestrian detection dataset on top of the Cityscapes dataset.
CrowdHuman [24]	2018	24K	470K	✓	✓	✓	A large, rich-annotated and high-diversity dataset
ECP [45]	2019	47K	219K	✓			Images in multiple European Cities
WiderPerson [46]	2019	13.5K	400K	✓			A large and diverse dataset for dense pedestrian detection in the wild.
WIDER Pedestrian	2019	97K	307K	✓			Covering traffic and surveillance scenarios, with a large number of occlusion instances.

C. COMPARISON

In this subsection, we compare and discuss the performance of some methods mentioned in this article on three popular datasets (Caltech [35], CityPersons [23], CrowdHuman [24]).

Table 5 presents a comparison of the results of several methods in Caltech [35]. On **R** subset, the best performances under original annotations and new annotations are obtained by NMS-Ped [79] which proposes a NMS loss to address the crowd occlusion, and PedHunter [25] which is an anchor-based and two-stage method. In addition, some methods (i.e., JointDet [61], AP^2M [107], MSFMN [88] and W^3Net [40]) also achieve relatively low miss rate on **R** subset. On the **HO** subset, the best performance is obtained by W^3Net [40] which leverages multi-modal information. Moreover, it can be seen that the performance of methods based on new annotations is better than that of methods based on original annotations, which demonstrates that the quality of the dataset has a significant influence on performance.

Table 6 compares some methods on CrowdHuman benchmark [24]. Since each image in the dataset contains dense pedestrians, the MR of all methods is higher than that on Caltech [35], and CityPersons [23]. The MR of most methods ranges between 40% and 50%. The best performance is obtained by MAPD [77] (12%˜24% improvement than other methods) which adapts a better positive settings strategy to mitigate class imbalance problems and proposes a novel piecewise NMS algorithm to reduce false positive. MAPD [77] is an improvement of APD [76]. Similarly, APD [76] also obtains better performance compared with other methods, which proves that the anchor-free method can be effective in crowd detection.

Table 7 shows the results of several advanced methods in CityPersons validation dataset. We separate these methods according to the different image sizes used. It is similar to the other two datasets that APD [76], MAPD [77], and W^3Net [40] achieve almost the best performance. Apart from these methods, part-based methods e.g., MSFMN [88] and attention-based methods e.g., MGAN [68], CaSe [82] also achieve satisfactory performance in HO subset.

Table 8 shows the performance and runtime comparisons on Caltech and CityPersons. Although nearly all works aim to develop a fast and accurate pedestrian detector, they end up compromising on speed and accuracy. Many of them add some modules to the baseline to improve accuracy, but they also increase inference time. Among reported methods, GDFL [67] significantly outperforms the others in terms of both speed and accuracy, while other methods e.g., [62], [63], [76] achieve a favorable trade-off between speed and accuracy.

VI. DISCUSSION AND RESEARCH TRENDS

Pedestrian detection is a challenging problem in computer vision and has received considerable attention. After deep learning achieved great success in generic object detection, pedestrian detection based on deep learning also made great progress. Despite the excellent detection performance, recent results on popular benchmarks show that there is still much room for improvement in occlusion handling and multi-scale detection. In this section, we discuss some open issues and future research trends according to the existing limitations.

A. DISCUSSION

With many dozens of methods discussed throughout this paper, we would now like to make a brief discussion to open issues that have emerged in pedestrian detection focusing on scale variance and occlusion based on deep learning.

1) SINGLE-STAGE VS. TWO-STAGE

Pedestrian detection based on deep learning can be divided into two categories: two-stage and single-stage. As shown in Table 2 and Table 3, most existing methods employ a two-stage strategy as their model architectures, especially

(a) Caltech (b) CrowdHuman (c) Widerperson

(d) KITTI

FIGURE 14. Some example images from CrowdHuman [24], WiderPerson [46], Caltech [35] and KITTI [34].

for occlusion handling, because they are better able to add different modules to meet different challenges. Although the best performance in some benchmarks is achieved by two-stage methods like JointDet [61] and PedHunter [25], these methods have higher computational cost, and the detection speed is relatively lower. Therefore, single-stage methods are becoming more and more popular owing to their faster detection speed. In early works, the detection performance for small-scale objects of single-stage methods like YOLO and SSD is relatively poor. Some recent methods like W^3Net [40] and AP^2M [107] have been modified to improve multi-scale detections. Therefore, more attempts should be made to integrate the advantages of single-stage and two-stage methods to build faster and more accurate detectors.

2) ANCHOR-BASED VS. ANCHOR-FREE

Anchor-based methods achieve state-of-the-art performance in generic object detection and are also very popular in pedestrian detection, as shown in Table 2 and Table 3. However, it remains challenging to accurately distinguish pedestrians in a crowd for anchor-based methods because of highly overlapped instances. Generally, there are more hyperparameters, which makes the network difficult to train. Some researchers have attempted to explore anchor-free methods. They abandon the troublesome anchor setting and use CNN to directly predict the scale and location. Some methods [20], [40] demonstrate the effectiveness of anchor-free methods. However, their performance is still worse than that of anchor-based methods in general. Therefore, effective anchor design or complete removal of anchor needs to be further explored to obtain better performance than the original anchor-based methods.

3) DETECTION ACCURACY AND DETECTION SPEED

In pedestrian detection, accuracy and speed are usually mutually compromised. In real-world applications, a balance between detection accuracy and speed is desirable. However, most of these methods have higher detection accuracy while the detection speed is lower. Therefore, it is very important to design a detector that can meet the requirements of accuracy and detection speed.

4) GENERALIZATION

Although current methods achieve high performance, they are almost always trained and tested on a single dataset. In [43], Hasan *et al.* find that most existing state-of-the-art pedestrian detectors though perform quite well when trained and tested on the same dataset, and generalize poorly in cross dataset evaluation. Consequently, their performance on different datasets is often inconsistent. For example, the detector trained on the Caltech has a good performance, but its performance on KITTI may be poor. The reason why such a problem occurs may be that the diversity of existing datasets is not enough. In addition, the detector obtained by training with a single dataset is more dependent on the dataset and its designs (e.g., anchor settings). Therefore, the generalization

TABLE 5. Miss rates (MR) of selected methods on the Caltech pedestrian dataset. The results of top parts are based on the original annotations from Caltech dataset [35], and the results bottom part are based on new annotations from [22].

Methods	Backbone	R	HO	Publications
DeepParts [17]	-	11.9	60.4	ICCV2015
MS-CNN [18]	VGG-16	10.0	60.0	ECCV2016
PCN [59]	VGG-16	9.4	55.8	BMVC2017
SDS-RCNN [5]	VGG-16	7.4	58.6	ICCV2017
SAF R-CNN [38]	VGG-16	9.7	64.4	TMM2017
FRCNN+ATT [66]	VGG-16	10.3	45.2	CVPR2018
PDOE+RPN [19]	VGG-16	7.6	44.4	ECCV2018
GDFL [67]	VGG-16	7.9	43.2	ECCV2018
FDNN [93]	VGG-16	8.7	55.1	ECCV2018
TLL+TFA [102]	ResNet-50	7.4	28.7	ECCV2018
ADM [103]	ResNet-50	8.6	30.4	TIP2018
UDN+ [85]	VGG-16	11.5	70.3	PAMI2018
AR-Ped [128]	VGG-16	6.5	48.8	CVPR2019
FRCNN+A+DT [87]	VGG-16	8.0	37.9	ICCV2019
MGAN [68]	VGG-16	6.8	38.2	ICCV2019
W^3Net [40]	ResNet-50	6.4	**28.3**	CVPR2020
TFAN [81]	ResNet-101	6.7	30.9	CVPR2020
MSFMN [88]	VGG-16	6.5	38.0	ICPR2020
PSC-Net [64]	VGG-16	6.4	34.8	SCIS2021
NMS-Ped [79]	ResNet-50	5.9	-	ICML2021
MS-CNN [18]	VGG-16	8.1	-	ECCV2016
PCN [59]	VGG-16	8.5	-	BMVC2017
SAF R-CNN [38]	VGG-16	7.5	-	TMM2017
SDS-RCNN [5]	VGG-16	6.4	-	ICCV2017
HyperLearner [129]	VGG-16	5.5	-	CVPR2017
Adapted FRCNN [23]	VGG-16	5.8	-	CVPR2017
FRCNN+ATT [66]	VGG-16	8.1	-	CVPR2018
RepLoss [21]	ResNet-50	4.0	-	CVPR2018
FDNN+SS [93]	VGG-16	6.7	-	ECCV2018
GDFL [67]	VGG-16	6.3	-	ECCV2018
ALFNet [56]	MobileNet	4.5	-	ECCV2018
OR-CNN [39]	VGG-16	4.1	-	ECCV2018
SAM-RCNN [101]	VGG-16	4.9	-	arXiv2018
CSP [20]	ResNet-50	4.5	-	CVPR2019
JointDet [61]	ResNet-50	3.0	-	AAAI2020
PedHunter [25]	ResNet-50	**2.3**	-	AAAI2020
W^3Net [40]	ResNet-50	3.8	-	CVPR2020
FC-Net [63]	ResNet-50	4.4	-	ITS2020
PRF-Ped [105]	ResNet-50	4.2	-	ICPR2020
MSFMN [88]	VGG-16	2.8	-	ICPR2020
AP^2M [107]	ResNet-50	3.3	-	AAAI2021

TABLE 6. Evaluation of full-body detection on CrowdHuman benchmark.

Methods	Backbone	MR	AP	Recall	Publication
Shao *et al.* [24]	ResNet-50	50.4	85	90.2	arXiv2018
RepLoss [21]	VGG-16	45.7	85.6	88.4	CVPR2018
DA-RCNN [60]	ResNet-50	51.8	-	-	arXiv2019
Adaptive NMS [41]	VGG-16	49.7	84.7	91.3	CVPR2019
MGAN [68]	VGG-16	49.3	-	-	ICCV2019
JointDet [61]	ResNet-50	46.5	-	-	AAAI2020
PedHunter [25]	ResNet-50	39.5	-	-	AAAI2020
NOH-NMS [75]	ResNet-50	43.9	89	92.9	ACM MM2020
PBM(R2NMS) [74]	ResNet-50	43.4	89.3	93.3	CVPR2020
EMD-RCNN [80]	ResNet-50	41.4	90.7	83.7	CVPR2020
CaSe [82]	VGG-16	47.9	-	-	ECCV2020
PS-RCNN [71]	ResNet-50	-	87.9	95.1	ICME2020
Feature-NMS [97]	-	75.4	68.7	-	ICPR2020
Beta R-CNN [70]	ResNet-50	40.3	88.2	-	NIPS2020
APD [76]	ResNet-50	35.8	-	-	TMM2020
MAPD [77]	ResNet-50	**27.8**	-	-	Nerocomputing
IterDet [84]	ResNet-50	49.4	88.1	**95.8**	arXiv2020
PED [72]	-	45.6	89.5	94	arXiv2021
LLA [78]	ResNet-50	47.9	88	94	arXiv2021
V2F-Net [65]	ResNet-50	42.3	**91**	84.2	arXiv2021

ability in different scenarios is very important owing to their applications in the real world.

TABLE 7. Miss Rates (MR) comparison of several advanced methods in CityPersons validation dataset.

Method	Input Scale	R	HO	Publication
Adapted FR-CNN [23]	×1	15.4	55	CVPR2017
HPCNN [98]	×1	12.5	-	BMVC2018
RepLoss [21]	×1	13.2	56.9	CVPR2018
FRCNN+ATT [66]	×1	16.0	56.7	CVPR2018
TLL+MRF [102]	×1	14.4	52	ECCV2018
OR-CNN [39]	×1	12.8	55.7	ECCV2018
ALFNet [56]	×1	12.0	51.9	ECCV2018
Adaptive-NMS [41]	×1	11.9	55.2	CVPR2019
CSP [20]	×1	11.0	49.3	CVPR2019
MGAN [68]	×1	11.5	51.7	ICCV2019
R^2NMS [74]	×1	11.1	53.3	CVPR2020
W^3Net [40]	×1	9.3	**18.7**	CVPR2020
PRNet [62]	×1	10.8	53.3	ECCV2020
CaSe [82]	×1	10.5	40.5	ECCV2020
FC-Net [63]	×1	13.5	44.3	ITS2020
PRF-Ped [105]	×1	9.7	47.3	ICPR2020
MagnifierNet [104]	×1	10.8	42.2	ICPR2020
BETA-RCNN [70]	×1	10.6	47.1	NIPS2020
APD [76]	×1	8.8	46.6	TMM2020
MAPD [77]	×1	**8.5**	44.7	Nerocomputing
AP^2M [107]	×1	10.4	48.6	AAAI2021
NMS-Ped [79]	×1	10.1	-	ICMR2021
Adapted FR-CNN [23]	×1.3	12.8	-	CVPR2017
RepLoss [21]	×1.3	11.6	55.3	CVPR2018
OR-CNN [39]	×1.3	11.0	51.3	ECCV2018
PDOE+RPN [19]	×1.3	11.2	44.2	ECCV2018
Adaptive-NMS [41]	×1.3	10.8	54.2	CVPR2019
FRCNN+A+DT [87]	×1.3	11.1	44.3	ICCV2019
MGAN [68]	×1.3	10.5	39.4	ICCV2019
JointDet [61]	×1.3	10.2	-	AAAI2020
PedHunter [25]	×1.3	8.3	43.5	AAAI2020
NOH-NMS [75]	×1.3	10.8	53	ACM MM2020
EMD-RCNN [80]	×1.3	10.7	-	CVPR2020
CaSe [82]	×1.3	9.1	43.6	ECCV2020
FC-Net [63]	×1.3	11.6	42.8	ITS2020
MSFMN [88]	×1.3	10.1	**38.5**	ICPR2020
0.5Stage [130]	×1.3	**8.1**	-	WACV2020

TABLE 8. Comparison with the state-of-the-art methods on the Caltech (upper) and CityPersons (lower) heavy occlusion subset in terms of speed and miss rate. Results are obtained from original paper.

Methods	Miss Rate (%)	Inference Time (s)	GPU
FDNN+SS [93]	53.7	2.48	Titan X
DeepParts [17]	60.4	1.0	-
JL-Tops [86]	49.2	0.6	K5200
SA-FastRCNN [38]	64.3	0.59	Titan X
RPN+BF [53]	74.4	0.50	K40
FDNN [93]	55.1	0.3	Titan X
SDS RCNN [5]	58.6	0.26	Titan X
SSA-CNN [6]	-	0.11	1080Ti
MS-CNN [18]	60.0	0.08	Titan X
GDFL [67]	43.2	0.05	1080Ti
ALF [56]	-	0.05	1080Ti
HBAN [90]	48.1	0.73	Titan X
CSP [20]	49.3	0.33	1080Ti
CaSe [82]	50.3	0.33	Titan X
FC-Net [63]	41.1	0.29	V100
ALF [56]	51.9	0.27	1080Ti
PRNet [62]	42.0	0.22	1080Ti
APD [76]	49.8	0.12	1080Ti

5) HIGH-QUALITY DATASETS

Most current state-of-the-art methods usefully supervised models learned from labeled data with object bounding

boxes, making the performance heavily dependent on the datasets. Hence, the diversity of datasets is important. We should know that data annotation by human is very difficult, so efficient data annotation will make a great contribution to pedestrian detection or generic object detection. In the past, we have been using datasets to evaluate our proposed algorithm. It is worth studying whether we can use a network to assist data annotation. In addition, as mentioned in [22], the quality of data also has a significant impact on the performance of the detector. The training data are usually manually annotated to ensure the quality of the datasets, but this is not completely accurate. Therefore, the detector should have higher robustness for such wrong data.

B. RESEARCH TRENDS

It can be seen from the results on the popular benchmark that state-of-the-art methods in this article have achieved good performance. The performance is basically saturated in **R** subset, but there is still a large gap under heavy occlusion. Based on these open challenges, we propose some works to close the gap with humans in the future.

1) WEAKLY SUPERVISED OR UNSUPERVISED PEDESTRIAN DETECTION

As discussed above, most current methods are fully supervised methods. More attention should be paid to weakly supervised or unsupervised methods to eliminate the problems associated with inefficient data annotation. Furthermore, it is valuable to study the performance of detectors on partially annotated data.

2) PEDESTRIAN DETECTION IN DIFFERENT MODALITIES

Most detectors are based on 2D images. Other modalities (such as depth [91], video [81], and point clouds) will be helpful for pedestrian detection. This conclusion is also proved in W^3Net [40], which achieves the best performance under heavy occlusion. In addition, it is also worth exploring how to combine the information of different modalities to obtain better performance.

3) CROSS-DATASET EVALUATION

Existing state-of-the-art pedestrian detectors perform quite well when trained and tested on the same dataset, generalize poorly in cross-dataset evaluation. However, different datasets have different scenarios, which may negatively impact the model trained on the single dataset. Therefore, more emphasis should be put on cross-dataset evaluation to achieve better generalization performance in real-world applications.

4) GENERIC PEDESTRIAN DETECTION

Most of the current works focus on addressing occlusion or scale-variation problems separately, but these challenges exist simultaneously in the real world. Therefore, methods should be able to address multiple challenges simultaneously.

VII. CONCLUSION

In recent years, tremendous progress has been made towards more accurate pedestrian detection. In this study, we attempt to comprehensively understand the methods for occlusion handling and multi-scale pedestrian detection. Therefore, many dozens of methods are discussed in this paper and we would now like to focus on the key factors which have emerged in pedestrian detection.

Occlusion Handling Occlusion is a critical challenge for pedestrian detection at present. As the intuitive clues of occlusion handling, visible and head information are widely used in many methods. Part-based methods make use of this information to learn extra supervision, reweight feature maps, guide the anchor selection or generate part proposals to improve the quality of full-body prediction. Attention-based methods leverage attention mechanisms to focus on visible information and suppress the occluded parts or background. In addition to using the visible part to make the feature more robust to occlusion, some methods make the proposal more discriminant to occlusion from the perspective of loss. Besides, variants of NMS have been proposed to soften the sensitivity of the NMS threshold in crowded scenarios.

Multi-scale Pedestrian Detection Multi-scale pedestrian detection is still a very challenging problem because real-time applications usually contain pedestrians of various scales. The effective solution for multi-scale pedestrian detection is to fuse multi-scale feature maps to get more information. The key idea behind these methods is that shallow feature maps contain accurate localization information, whereas deeper ones tend to encode rich semantic information. Nevertheless, some other methods leverage different data augmentation strategies to reduce the impact of unbalanced data distribution, while anchor-free methods remove the anchor design to reduce the influence of anchors on small-scale pedestrians.

Most works are working on developing a robust and real-time solution. However, the detection performance as well as the computational cost of available solutions is far behind expectations. Different methods are categorized to understand current research trends and to guide the design of new frameworks in this study. In addition, the results for different benchmarks also show the effectiveness of the different methods. Therefore, we hope our survey can be helpful for developing novel methods for pedestrian detection in the future.

REFERENCES

[1] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 384–406, Apr. 2018.
[2] S. D. Pendleton, H. Andersen, and X. Du, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, p. 6, 2017.
[3] Z. Li, J. Gong, C. Lu, and Y. Yi, "Interactive behavior prediction for heterogeneous traffic participants in the urban road: A graph-neural-network-based multitask learning framework," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 3, pp. 1339–1349, Jun. 2021.
[4] C. Liu, X. Chang, and Y.-D. Shen, "Unity style transfer for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6887–6896.

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

[6] C. Zhou, M. Wu, and S.-K. Lam, "SSA-CNN: Semantic self-attention CNN for pedestrian detection," 2019, *arXiv:1902.09080*.

[7] I. Bae and H.-G. Jeon, "Disentangled multi-relational graph convolutional network for pedestrian trajectory prediction," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2021, vol. 35, no. 2, pp. 911–919.

[8] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 734–750.

[9] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4594–4603.

[10] O. Russakovsky, J. Deng, H. Su, and J. Krause, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[12] T. Lin, M. Maire, S. Belongie, J. Hays, and P. Perona, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

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

[14] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. BMVC*, 2009, pp. 1–11.

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

[16] 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.

[17] 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.

[18] 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. (ECCV)*, Oct. 2016, pp. 354–370.

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

[20] 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. (CVPR)*, Jun. 2019, pp. 5182–5191.

[21] 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.

[22] 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.

[23] 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.

[24] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.

[25] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2020, vol. 34, no. 7, pp. 10639–10646.

[26] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014, pp. 424–432.

[27] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2008.

[28] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[29] C. Wojek, P. Dollar, 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, p. 743, 2012.

[30] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 613–627.

[31] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[32] N. Ragesh and R. Rajesh, "Pedestrian detection in automotive safety: Understanding state-of-the-art," *IEEE Access*, vol. 7, pp. 47864–47890, 2019.

[33] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 30, 2021, doi: 10.1109/TPAMI.2021.3076733.

[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[35] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[36] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[37] 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. 2017.

[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, Apr. 2018, doi: 10.1109/TMM.2017.2759508.

[39] 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)*, Sep. 2018, pp. 637–653.

[40] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun, "Where, what, whether: Multi-modal learning meets pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14065–14073.

[41] 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. 6452–6461.

[42] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3626–3633.

[43] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11328–11337.

[44] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[45] M. Braun, S. Krebs, F. Flohr, and D. Gavrila, "EuroCity persons: A novel benchmark for person detection in traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1844–1861, Aug. 2019.

[46] 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, no. 2, pp. 380–393, Feb. 2020.

[47] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[48] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[49] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 2001, pp. 1–10.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jan. 2014.

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

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

[54] 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. (CVPR)*, Jun. 2016, pp. 779–788.
[55] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[56] W. Liu, S. Liao, and W. Hu, "Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding," *IEEE Trans. Image Process.*, vol. 29, pp. 1413–1425, 2020.
[57] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.
[58] 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. (CVPR)*, Jun. 2018, pp. 966–974.
[59] S. Wang, "PCN: Part and context information for pedestrian detection with CNNs," in *Proc. Procedings Brit. Mach. Vis. Conf.*, 2017, pp. 34.1–34.13.
[60] K. Zhang, F. Xiong, P. Sun, L. Hu, B. Li, and G. Yu, "Double anchor R-CNN for human detection in a crowd," 2019, *arXiv:1909.09998*.
[61] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational learning for joint head and human detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2020, vol. 34, no. 7, pp. 10647–10654.
[62] 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. (ECCV)*, Oct. 2020, pp. 32–48.
[63] T. Zhang, Q. Ye, B. Zhang, J. Liu, X. Zhang, and Q. Tian, "Feature calibration network for occluded pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, early access, Dec. 14, 2021, doi: 10.1109/TITS.2020.3041679.
[64] J. Xie, Y. Pang, H. Cholakkal, R. Anwer, F. Khan, and L. Shao, "PSC-Net: Learning part spatial co-occurrence for occluded pedestrian detection," *Sci. China Inf. Sci.*, vol. 64, no. 2, pp. 1–13, Feb. 2021.
[65] M. Shang, D. Xiang, Z. Wang, and E. Zhou, "V2F-Net: Explicit decomposition of occluded pedestrian detection," 2021, *arXiv:2104.03106*.
[66] 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.
[67] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820–3834, 2020.
[68] 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. (ICCV)*, Oct. 2019, pp. 4966–4974.
[69] T. Zou, S. Yang, Y. Zhang, and M. Ye, "Attention guided neural network models for occluded pedestrian detection," *Pattern Recognit. Lett.*, vol. 131, pp. 91–97, Dec. 2020.
[70] Z. Xu, B. Li, Y. Yuan, and A. Dang, "Beta R-CNN: Looking into pedestrian detection from another perspective," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2020, pp. 19953–19963.
[71] 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. Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[72] M. Lin, C. Li, X. Bu, M. Sun, C. Lin, J. Yan, W. Ouyang, and Z. Deng, "DETR for crowd pedestrian detection," 2020, *arXiv:2012.06785*.
[73] C. Yang, V. Ablavsky, K. Wang, Q. Feng, and M. Betke, "Learning to separate: Detecting heavily-occluded objects in urban scenes," 2019, *arXiv:1912.01674*.
[74] 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.
[75] P. Zhou, C. Zhou, P. Peng, J. Du, X. Sun, X. Guo, and F. Huang, "NOH-NMS: Improving pedestrian detection by nearby objects hallucination," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1967–1975.
[76] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-C. Chen, Y. Hu, and S. C. H. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Trans. Multimedia*, vol. 23, pp. 3085–3097, 2021.
[77] Y. Wang, C. Han, G. Yao, and W. Zhou, "MAPD: An improved multi-attribute pedestrian detection in a crowd," *Neurocomputing*, vol. 432, pp. 101–110, Oct. 2021.
[78] Z. Ge, J. Wang, X. Huang, S. Liu, and O. Yoshie, "LLA: Loss-aware label assignment for dense pedestrian detection," 2021, *arXiv:2101.04307*.
[79] Z. Luo, Z. Fang, S. Zheng, Y. Wang, and Y. Fu, "NMS-loss: Learning with non-maximum suppression for crowded pedestrian detection," 2021, *arXiv:2106.02426*.
[80] 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. (CVPR)*, Jul. 2020, pp. 12211–12220.
[81] 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. (CVPR)*, Jun. 2020, pp. 13430–13439.
[82] X. Jin, H. Cholakkal, M. A. Rao, F. S. Khan, and M. Shah, "Count and similarity-aware R-CNN for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 88–104.
[83] Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, "Variational pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11622–11631.
[84] D. Rukhovich, K. Sofiiuk, D. Galeev, O. Barinova, and A. Konushin, "IterDet: Iterative scheme for object detection in crowded environments," 2020, *arXiv:2005.05708*.
[85] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1874–1887, Aug. 2018.
[86] 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. 3506–3515.
[87] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9556–9565.
[88] Y. He, C. Zhu, and X.-C. Yin, "Mutual-supervised feature modulation network for occluded pedestrian detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8453–8460.
[89] C. Y. Lin, H. X. Xie, and H. Zheng, "PedJointNet: Joint head-shoulder and full body deep network for pedestrian detection," *IEEE Access*, vol. 7, pp. 47687–47697, 2019.
[90] R. Lu, H. Ma, and Y. Wang, "Semantic head enhanced pedestrian detection in a crowd," *Neurocomputing*, vol. 400, pp. 343–351, Oct. 2020.
[91] Z. Guo, W. Liao, Y. Xiao, P. Veelaert, and W. Philips, "Deep learning fusion of RGB and depth images for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2019, pp. 1–13.
[92] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020, pp. 213–229.
[93] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
[94] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness NMS and bounded IoU loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6877–6885.
[95] 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. (ICCV)*, Oct. 2017, pp. 5562–5570.
[96] 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. (AAAI)*, Feb. 2020, vol. 34, no. 7, pp. 12993–13000.
[97] N. O. Salscheider, "FeatureNMS: Non-maximum suppression by learning feature embeddings," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7848–7854.
[98] Y. Zhao, Z. Yuan, H. Zhang, and S. Innovation, "Joint holistic and partial CNN for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, p. 81.
[99] R. Lu and H. Ma, "Occluded pedestrian detection with visible IoU and box sign predictor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1640–1644.
[100] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu, "Scale-adaptive deconvolutional regression network for pedestrian detection," in *Proc. Asian Conf. Comput. Vis.*, Nov. 2016, pp. 416–430.
[101] T. Liu, M. Elmikaty, and T. Stathaki, "SAM-RCNN: Scale-aware multi-resolution multi-channel pedestrian detection," 2018, *arXiv:1808.02246*.
[102] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 1–16.

[103] 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.
[104] Q. Cheng, M. Chen, Y. Wu, F. Chen, and S. Lin, "MagnifierNet: Learning efficient small-scale pedestrian detector towards multiple dense regions," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1483–1490.
[105] Y. Tan, H. Yao, H. Li, X. Lu, and H. Xie, "PRF-PED: Multi-scale pedestrian detector with prior-based receptive field," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6059–6064.
[106] M. Ding, S. Zhang, and J. Yang, "Learning a dynamic high-resolution network for multi-scale pedestrian detection," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9076–9082.
[107] M. Liu, C. Zhu, J. Wang, and X.-C. Yin, "Adaptive pattern-parameter matching for robust pedestrian detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2021, vol. 35, no. 3, pp. 2154–2162.
[108] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Eng.*, vol. 29, no. 6, pp. 33–41, 1984.
[109] T.-Y. Lin, P. Dollar, 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.
[110] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8759–8768.
[111] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[112] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.
[113] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.
[114] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M.-H. Yang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1125–1137, Feb. 2021.
[115] X. Xie and Z. Wang, "Multi-scale semantic segmentation enriched features for pedestrian detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2196–2201.
[116] X. Zhang, S. Cao, and C. Chen, "Scale-aware hierarchical detection network for pedestrian detection," *IEEE Access*, vol. 8, pp. 94429–94439, 2020.
[117] W. Wang, "Adapted center and scale prediction: More stable and more accurate," 2020, *arXiv:2002.09053*.
[118] J. Cai, F. Lee, S. Yang, C. Lin, H. Chen, K. Kotani, and Q. Chen, "Pedestrian as points: An improved anchor-free method for center-based pedestrian detection," *IEEE Access*, vol. 8, pp. 179666–179677, 2020.
[119] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[120] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[121] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6023–6032.
[122] S. Cygert and A. Czyzewski, "Toward robust pedestrian detection with data augmentation," *IEEE Access*, vol. 8, pp. 136674–136683, 2020.
[123] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 6154–6162.
[124] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.
[125] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2005, pp. 90–97.
[126] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multi-view, multi-pose object detection," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[127] 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.
[128] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7224–7233.
[129] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6034–6043.
[130] U. Ujjwal, A. Dziri, B. Leroy, and F. Bremond, "A one-and-half stage pedestrian detector," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 776–785.

FANG LI received the B.S. degree from the Harbin Institute of Technology at Weihai, Weihai, China, in 2020. He is currently pursuing the M.S. degree in mechanical engineering with the Beijing Institute of Technology, Beijing, China. His research interests include autonomous driving and object detection.

XUEYUAN LI received the B.S., M.S., and Ph.D. degrees in vehicle engineering from the Beijing Institute of Technology, Beijing, China, in 1999, 2002, and 2010, respectively. He was the Director of the National Key Laboratory of Vehicular Transmission, from 2008 to 2014. Since 2002, he has been an Associate Professor with the School of Mechanical Engineering, Beijing Institute of Technology, where he is currently the Vice Director of the Department of Vehicle Engineering. His research interests include vehicle transmission theory and technology, unmanned vehicle theory and technology, and machine learning.

QI LIU received the B.S. degree from the Beijing Institute of Technology (BIT), Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include intelligent vehicles, environmental perception, and decision making.

ZIRUI LI received the B.S. degree from the Beijing Institute of Technology (BIT), Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include intelligent vehicles, driver behavior modeling, and transfer learning.