

REVIEW

Deep learning for occluded and multi-scale pedestrian detection: A review

Yanqiu Xiao¹  | Kun Zhou² | Guangzhen Cui¹ | Lianhui Jia² | Zhanpeng Fang¹ | Xianchao Yang¹ | Qiongpei Xia¹

¹ College of Mechanical and Electronic Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

² China Railway Engineering Equipment Group Co. Ltd., Zhengzhou, China

Correspondence

College of Mechanical and Electronic Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China.

Email: xianchaoyang95@gmail.com

Abstract

Pedestrian detection, as a research hotspot in the field of computer vision, is widely used in many fields, such as automatic driving, video surveillance, robots and so on. In recent years, with the rapid development of deep learning, pedestrian detection technology has made unprecedented breakthroughs. However, it fails to saturate pedestrian detection research, and there are still many problems to be solved. This study reviews the current research status of pedestrian detection methods based on deep learning. In the first place, we summarised the research results of two stage and one stage pedestrian detection based on deep learning, also summarised and analysed the improvement methods. Meanwhile, we focused on the occlusion and multi-scale problems of pedestrian detection and discussed the corresponding solutions. At last, we induced the pedestrian detection datasets and evaluation methods and prospected the development trend of deep learning in pedestrian detection.

1 | INTRODUCTION

Pedestrian detection, as an important research topic in the field of computer vision for a long time, has many applications such as autonomous driving, video surveillance, robotics and so on. In addition, pedestrian detection as a special case of object detection, its research achievements play an important role in promoting the development of other object detection methods. The process of pedestrian detection is to predict, locate and mark the position of pedestrian to obtain information such as the position and action of the pedestrian [1], as shown in Figure 1. However, due to the random distribution and dynamic characteristics of pedestrian, many detection algorithms cannot detect pedestrians accurately in real time. There will be false positives and false negatives in the process of pedestrian detection since the influence by weather, similar objects, occlusion and other factors results in poor robustness of current pedestrian detection algorithms in more complex scenes. Therefore, a lot of research is still devoted every year to establish a state-of-the-art method. Figure 2 shows the amount of researches from 2000 to 2018 in pedestrian detection, the data from Google scholar advanced search the pedestrian detection as allintitle.

Pedestrian detection is a special task of object detection. Its technological progress is closely related to the development of general object detection. This connection can be described as follows: The general object detection algorithm can be used for pedestrian detection after appropriate improvement. Pedestrian detection is a kind of object detection, and the problems it studies can promote the development of general object detection from another view.

In 2003, Viola and Jones [2] first used image intensity information and motion information combined with Adaboost classifier to realise pedestrian detection and tracking, which attracted the attention of researchers to the issue. Then, Dalal and Triggs [3] proposed pedestrian detection methods based on histogram of oriented gradient (HOG) and support vector machine (SVM) classifiers, which achieved nearly 100% detection effect on MIT pedestrian dataset [4] and greatly promoted the development of pedestrian detection technology due to its ability to accurately represent objects. Later, pedestrian detection methods based on artificial feature extraction combined with machine learning classifier have become the mainstream [5–9], and most of the researchers have improved or innovated on this paradigm.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

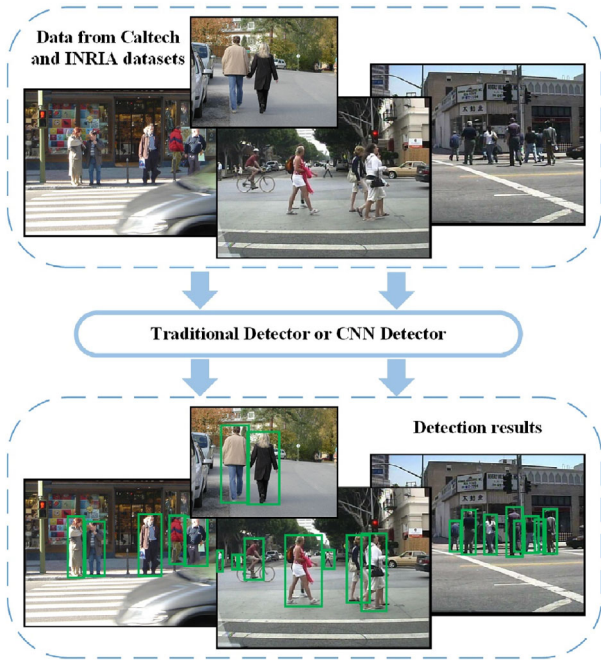


FIGURE 1 The process of pedestrian detection

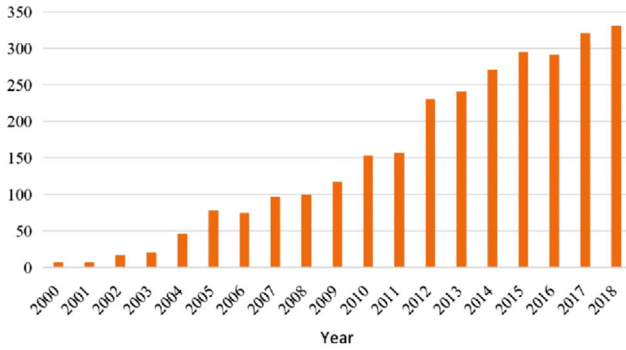


FIGURE 2 The increasing number of publications in pedestrian detection from 2000 to 2018

The design of an artificially extracted feature is critical to the performance of the detector. The development of this kind of method is restricted by the artificial feature extraction since the artificial feature cannot adapt to the change of image background, which may lead to the poor effect of the proposed detection method. Krizhevsky [10] proposed AlexNet model based on deep convolution neural network in 2012, which won the first prize in ImageNet [11] large-scale image recognition contest with absolute superiority, and its performance is far superior to traditional machine learning algorithm. The excellent classification performance of AlexNet model has led researchers to focus on deep learning. Based on the research of artificial features, Girshick [12] used sliding window to extract pedestrian proposals in 2014, and convolutional neural network (CNN) was used to extract pedestrian deep feature in the proposals, which was trained by SVM. Thereafter, the detection performance has improved by the new generation of pedestrian detection methods based on CNN significantly.

Pedestrian detection, as a hotspot and difficult problem in the field of computer vision, has attracted widespread attention due to its great application prospect in many directions. This study presents a comprehensive survey of the existing achievements in pedestrian detection. The main work of this study is as follows: Section 1 briefly summarises the current research results from traditional machine learning detection methods to deep learning detection methods. Section 2 focuses on the principles, performances, advantages of several deep learning pedestrian detection models, and discusses its corresponding improvement methods. Section 3 analyses and explores the occlusion and multi-scale problems and solutions of pedestrian detection. Section 5 summarises the pedestrian detection datasets and evaluation methods, and prospects the development trend of deep learning methods in pedestrian detection.

2 | PEDESTRIAN DETECTION METHOD BASED ON DEEP LEARNING

With the continuous improvement of computer performance, the detection method based on deep learning has attracted the attention of scholars. Pedestrian detection based on CNN can be roughly divided into two categories. One is a two-stage framework which first generates approximate valid regional proposals and then refines them through another sub-network. The other detection method is called single-stage framework which can accelerate the detection speed by removing the generation stage of regional proposals and regress the predefined area directly so that the computational efficiency can be improved. These frameworks are usually built on different network models. The following is a summary of the application of these two detection frameworks and their improved methods in pedestrian detection.

2.1 | Two-stage detection framework

Two-stage network framework detection method is usually called region-based detection method. First, it obtains the proposals of the object by sliding window or selective search, then extracts the convolution feature in the region by using CNN, and finally classifies and recognises the feature by using classifier. Girshick et al. [13] combined traditional machine learning methods with CNN and proposed a detection framework based on *RCNN* as shown in Figure 3, of which the selective search is used to obtain as many object proposals as possible; CNN is used to extract the features of the proposals instead of manual extraction and SVM is used to classify the feature vectors. The results showed that the *RCNN* method owns the powerful processing ability of *CNN* in the field of computer vision. Later, spatial pyramid pooling (*SPP*) Net [14] and *Fast RCNN* [15] have been improved by introducing SPP layer and region of interest (ROI) pooling layer, respectively. However, the number of the proposals is too large, which is accompanied by a large amount of computational consumption in the proposals generation process, and limits its application scenarios. In response

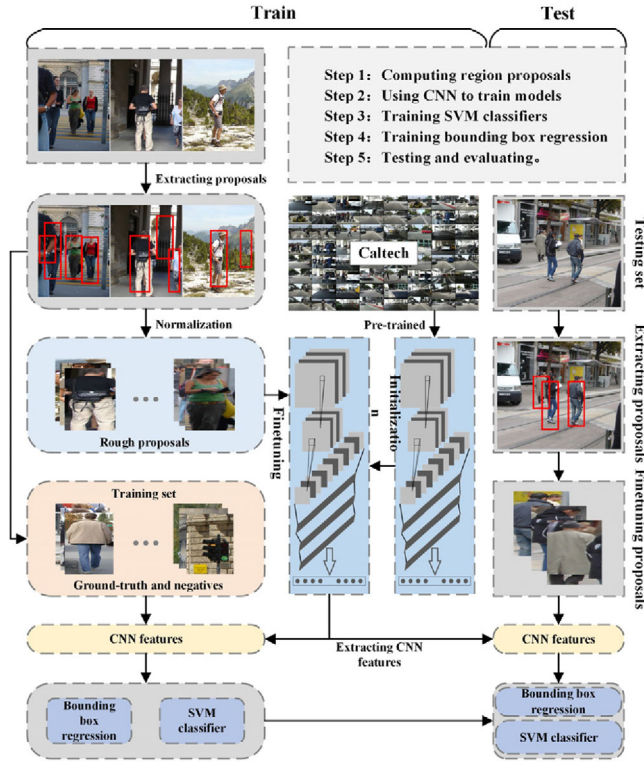


FIGURE 3 The process of RCNN detector

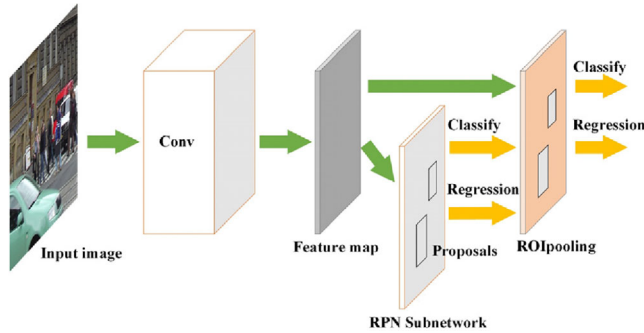


FIGURE 4 Faster RCNN network structure

to this issue, as shown in Figure 4, *Faster RCNN* [16] proposed a region proposal network (RPN), which is more accurate than selective search used by *RCNN*. In addition, the proposals are generated under the unified network framework by means of network sharing, and the training and learning process is completed by using the Softmax classifier. As in Table 1, compared with the initial CNN using staged training and SVM classification, the detection performance is greatly improved.

Because of the repeatability of deep convolution feature extraction and training process, its detection speed is limited. The detection framework based on *RCNN* consists of two parts: Feature extraction and classification training. In order to solve these problems, many methods try to improve the detection speed and accuracy of detectors by improving feature extraction methods, classification strategies and other auxiliary information. In fact, the network of the algorithm will take these three factors into consideration.

2.1.1 | Features extraction

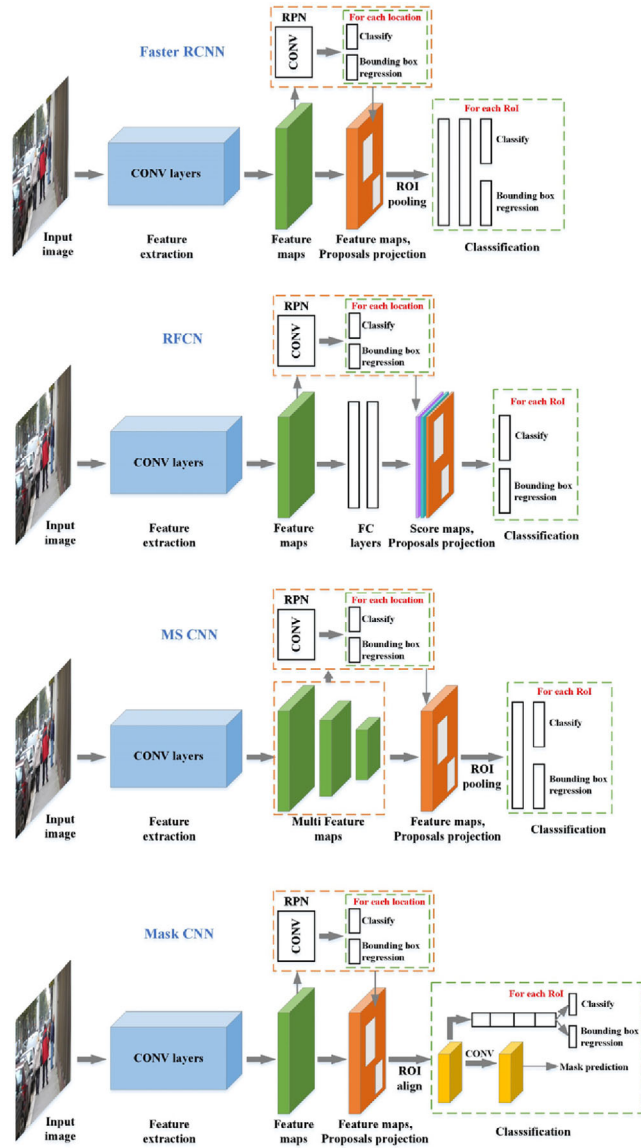
The accuracy of pedestrian detection depends on the accuracy of its feature description and the classification ability of the classifier. Pedestrian detection method based on deep learning is to extract pedestrian features through CNN. However, the convolution neural network has a strong expansibility, so most researchers change the structure of convolution neural network to extract more accurate pedestrian features. Network architecture plays an important role in feature extraction. It is the most basic and effective process to improve the performance of detectors by improving the network architecture in the feature extraction process. The continuous evolution of *RCNN* algorithm series is the continuous improvement of its network architecture. By changing the number, distribution structure and function of each convolution layer in the network, the convolution network can achieve higher detection performance [17–18].

In order to reduce the repeated calculation and improve the translation invariance and sensitivity in the detection process, on the basis of *Faster RCNN*, region-based fully convolutional network (*RFCN*) [19] used fully convolution layer to complete the learning of object features and location, and maximised the shared network, which greatly reduced the network computing consumption. Pedestrian detection requires an accurate response to the target's translation. *Faster RCNN* is convoluted before ROI pooling and has translation invariance. However, once ROI pooling is inserted, the later network structure will no longer have translation invariance. Therefore, *RFCN* used position sensitive score map to integrate pedestrian position information into ROI pooling to improve detection accuracy. *Faster RCNN* has a good achievement in dealing with object proposals, but its performance cannot be guaranteed for small-scale objects with low pixels. Based on this problem, *MS CNN* [20] proposed a fast multi-scale objects detection framework. Similarly with *Faster R-CNN*, there are also proposals sub-networks and detection sub-networks. *MS CNN* detected pedestrian at different scales on different deep convolution layers, low-level is used to detect small-scale objects, and high-level for the large-scale objects. By introducing deconvolution feature up-sampling as an alternative method of input up-sampling, the detection ability of small-scale objects is enhanced. *Mask RCNN* [21] expanded *RCNN* by adding a branch to predict objects parallel of existing branches. It can accomplish many tasks such as object classification, object detection, semantic segmentation, instance segmentation and human pose recognition. It forecasted the region of interest, generated category labels and rectangular box coordinates. Each binary mask generated by the mask prediction branch depends on the classification prediction results, and separated the pedestrians and the background based on the moment. Figure 5 shows the structure comparison of *Faster RCNN*, *MS CNN*, *RFCN* and *Mask RCNN*. It can be found that the differences shown mainly include feature map generation, proposals generation and classifier structure.

Therefore, for the improvement of two-stage detection method, researchers separate feature extraction and classifier to improve the performance of pedestrian detection. In the feature extraction stage, different proposed region generation methods

TABLE 1 Several different training methods and pedestrian detectors of classifiers

Detector name	Classifier	Pipeline	Dataset	MR	Year	Highlights
<i>RCNN</i> [13]	Support vector machine (SVM)	—	Caltech	23.3%	2014	Combining deep convolutional neural network (CNN) features and SVM, better for traditional method
<i>Faster RCNN</i> [16]	Softmax	—	Caltech	12.7%	2015	Region proposal network (RPN) and Softmax for accurate proposals and classification
<i>RPN+BF</i> [24]	Boosted forest (BF)	Faster RCNN	Caltech	9.6%	2016	BF solved the problem of that faster RCNN is bad for small scale pedestrian
<i>Comp ACT</i> [25]	Cascade boosting	RCNN	Caltech	11.7%	2015	By optimising classification risk under a complexity constraint to improve classification accuracy
<i>SAF RCNN</i> [26]	Softmax	Fast RCNN	Caltech	9.3%	2017	Incorporating different size sub-network into a unified architecture for training, to solve various sizes of pedestrian instances in the image

**FIGURE 5** The structure comparison of faster RCNN, MS convolutional neural network (CNN), RFCN and mask RCNN


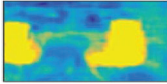
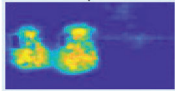

are used to quickly filter the background, and different branch networks are used to refine the extraction of multiple features, so that the detector has better robustness to pedestrians in variable scenes. In addition, a lot of researches are focused on the fusion of multiple visual tasks under the same feature map, which is also helpful to improve the detection performance.

2.1.2 | Training and classification

Detector performance is greatly affected by network training process and classifier. The classifier plays a decisive role in the performance of pedestrian detection. Different training strategies will affect the classification ability of the classifier. For the algorithm based on two-stage detection framework, feature extraction and classifier training are independent. In this process, some researchers focus on feature extraction, while some researchers try to explore more effective training strategies to improve the performance of classifier. Because of the complexity of pedestrian samples, it is necessary to design appropriate classification strategies according to pedestrian characteristics in the process of network training. Table 1 lists several different training methods and pedestrian detectors of classifiers. Among them, miss rate (MR) [22] is the evaluation result in the reasonable subset of Caltech [23], which is the meaning of all MR in this study. The content of dataset and pedestrian detector performance evaluation will be detailed in Section 4.

Zhang et al. [24] used the *Faster RCNN* detection framework to detect pedestrians, they found that the Softmax classifier used by *Faster RCNN* cannot effectively used the features provided by the fully connect layer, resulting in the classifier unable to adapt to low pixel pedestrians. They combined RPN network with Boosted Forests (BFs) on the basis of *Faster RCNN*. Meanwhile, BF classification strategy was introduced on the basis of RPN detector, the ability of classifier to mine difficult cases was strengthened, and the problem of weak generalisation ability of *Faster R-CNN* in pedestrian detection was improved. According to the different features extracted, the corresponding classification strategy was set up to avoid the miscalculation of multiple features by the classifier. The detection accuracy is improved by balancing the ability of feature extraction and classification

TABLE 2 Comparison of methods to enhance detector performance with auxiliary information

Detector name	Auxiliary information	Pipeline	Dataset	MR	Year	Highlights
<i>Deep parts</i> [28]	Part models 	RCNN	Caltech	11.8%	2015	By handling occlusion with an extensive part pool.
<i>SDS RCNN</i> [29]	Semantic features 	Faster RCNN	Caltech	7.3%	2017	Jointing supervision on pedestrian detection and semantic segmentation
<i>GDFL</i> [33]	Attention module 	—	Caltech	7.8%	2018	Incorporating pixel-wise information into deep convolutional feature maps for pedestrian detection
<i>HyperLearner</i> [34]	Channel feature 	Faster RCNN	Caltech	5.5%	2017	Integrated semantic features and edge features into CNN-based pedestrian detectors

training. Cai et al. [25] deduced a complexity-aware cascade training algorithm (*Comp ACT*) to optimise the classification risk under the constraint of feature complexity, so that the high-complexity features can be trained in the later stage which can better combine feature extraction and classifier functions. It is very effective to train the classification sub network for detecting pedestrians of different scales to improve the ability of the detector to deal with low pixel pedestrians, and achieves a high-precision pedestrian detection. Scale-aware fast region-convolutional neural networks (*SAF RCNN*) [26] proposed a weighting mechanism based on scale perception for pedestrian features at different scales. It used sub-networks to train pedestrian images at different scales separately, which improved the performance of private sub-networks at different input scales and ensured the detection performance in a certain scale range.

To sum up, in the improved algorithm based on the two-stage detection framework, researchers balance the ability of classifier and feature extraction to improve the detection accuracy. This process needs to refer to the detailed feature types for effective classification strategy design. At the same time, the designed classification strategy should meet the adaptability of the detector to pedestrians in various scenes.

2.1.3 | Auxiliary information

At present, pedestrian detection mainly uses supervised learning to complete the training of detector. The quality of annotation information in datasets has an important influence on the learning process. For Caltech and Citypersons [27] pedestrian datasets, relevant auxiliary information such as location, occlusion and other hints are marked on their images. In the process of network training, other auxiliary information can be added to enhance the detection performance of the detector. Although current researchers focus on deep learning methods,

the achievements of traditional machine learning methods cannot be ignored. Traditional artificial features have better representation ability, so many studies combine artificial features as additional feature representation with convolution network to improve detector performance. Meanwhile, some advanced semantic information is also used to increase the ability of detectors in the training process.

Table 2 shows a comparison of some ways to enhance detector performance with auxiliary information. Tian et al. [28] integrated pedestrian detection with pre-trained convolution network by building a pedestrian body parts pool (*Deep parts*) in different scenarios, and solved pedestrian detection problems in most complex scenarios through component pools. *Deep parts* can complete the training in the data of only marking the pedestrian body part, that is, it can assist the detection of the whole pedestrian by detecting the part of the pedestrian body. This method can deal with the data that deviate from the actual annotation, and has a better detection effect for the occluded pedestrian. Brazil et al. [29] proposed a segmentation injection network (*SDS RCNN*), which combined semantic segmentation as pedestrian detection auxiliary information with regional detection; additional monitoring information which can help guide the functions in the network sharing layer and still has better detection performance under weak annotation information, and is twice as fast in Caltech and KITTI [30] datasets as before. The research [31], [32] combined semantic segmentation with pedestrian detection in different ways which can promote the generalisation ability of the detector to a certain extent. Lin et al. [33] proposed a texture perception-based depth feature learning pedestrian detector (graininess-aware deep feature learning, *GDFL*), used fine-grained details to construct a pedestrian attention module to guide the detector to focus on the pedestrian areas. Meanwhile, an amplification and reduction modules were introduced, which combined local features with upper and lower frame information by

convolution to enhance the detection ability of the detector for small-scale pedestrians. Mao et al. [34] proposed a CNN architecture that integrates multi-channel features. Integral channel features (ICF), edge, segmentation, thermal map, optical and parallax channels are integrated into CNN for pedestrian detection, and the ability of the detector to learn additional features was improved by real, semantic, dynamic and deep visual features.

In the two-stage detection framework, some researchers use additional feature information assisted detector to better identify pedestrians. These additional features include manually extracted features, heat map, semantic segmentation features and so on. The purpose of this method is to ensure that the detector can better focus on the high dimension feature descriptors of pedestrians. Other researcher's studies are to refine the focus position of the detector such as using part model and attention mechanism. This method is to promote the detector to expand to the whole detection location according to the small parts.

2.2 | Single-stage detection framework

Although the two-stage network framework has made great breakthroughs in accuracy, the performance of end-to-end learning cannot be reflected due to the hierarchical method of region extraction combined with training. To solve this issue, a single-stage network framework is proposed to speed up detection by removing the regional proposals generation stage. By setting anchors in advance, the input image is convoluted directly, and then the anchors in the convolution map are regressed and classified. In practical testing, it has more efficient detection speed and is easily transplanted to embedded system. However, its direct detection on the original image means that the training process is very complex and the trained model is difficult to guarantee better robustness, so the accuracy cannot replace the two-stage framework.

You only look once (YOLO) [35], Single Shot MultiBox Detector (SSD) [36] are representative single-stage network frameworks. YOLO divides the input image into $S \times S$ units, each unit is responsible for the centre of the unit's object detection, using a one-time prediction of the object boundaries, positioning confidence and all kinds of probability vectors. At present, several versions have been updated according to their performance, such as YOLOv2 [37], YOLOv3 [38]. Different from YOLO, SSD detects multi-scale objects directly in the convolution layer by setting anchors of different scales on the image, calculating and regressing all the anchors and confidence in the detection process, and detecting multi-scale objects by setting convolution maps of different scales. SSD has advantages over YOLO in solving small-scale and location problems. The network structure of SSD and YOLO is shown in Figure 6.

The proposer of SSD algorithm used the original SSD in pedestrian detection and found that the results were worse than those of RCNN framework pedestrian detection algorithm. The reason is that SSD has poor ability in reducing false positives when dealing with pedestrians in complex scenarios. Inspired

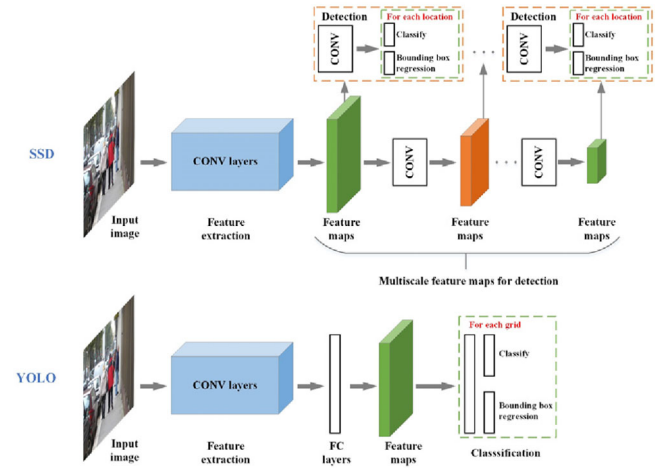


FIGURE 6 Comparison of SSD and YOLO structures

by Cascade RCNN [39], Liu et al. [40] proposed an improved method based on asymptotic localisation fitting (ALF). By setting different IOU thresholds on the feature map for multiple regression, the regression boxes of the upper layer is used as the anchor boxes of the next layer. The results show that the detection accuracy reaches the most advanced level under the condition of guaranteeing high detection speed. Based on the SSD framework, Du et al. [41] proposed a deep neural network fusion structure FPGA-Deep Neural Networks (F-DNN) for fast and robust pedestrian detection. It used a single convolution network to generate pedestrian proposals, and used several deep neural networks to optimise the results in parallel. At the same time, it integrated the pixel-level semantic segmentation network into the detection architecture to enhance the pedestrian detector.

The outstanding advantage of single-stage network is its excellent detection speed, but the accuracy is slightly inadequate. In order to increase the detection accuracy of YOLOv3 in automatic driving applications, by constructing the boundary frame model with Gauss parameters, [42] proposed a new predictive localisation algorithm to improve the reliability of the border. This method guaranteed the excellent detection speed of YOLOv3 and greatly improves the accuracy. The authors of [43] proposed a hybrid attention mechanism HARNet for single-stage object detection. First, spatial, channel and focused attentions are used for single-stage object detection. Then, the consistent attention mechanism was constructed into a deformable filter, and the hybrid attention mechanism is embedded in Retina-Net to complete single-stage object detection. Through the combination of multiple attention mechanisms and single-stage network, HARNet improves the single-stage network to locate the pedestrian area quickly and accurately, and solves the problem of missing detection caused by too many anchors in the single-stage network.

YOLO and SSD play an important role in promoting the real-time application of pedestrian detection algorithm, but so far pedestrian detection methods still focus on the improvement of two-stage network framework, and there are still few pedestrian detection algorithms based on single-stage network framework.

TABLE 3 Summary of the state-of-art method from 2015 to 2019, C-Caltech, E-ETH [8], K-KITTI, I-INRIA [3], CP-Citypersons, T-TUD [44]

Methods	MR in Caltech	Pipeline used	Backbone	Dataset used	Year
<i>ConvNet</i> [45]	23.32%	—	AlexNet	C	2015
<i>TA CNN</i> [46]	20.86%	—	—	C&E	2015
<i>RotatedFilters</i> [47]	10.00%	RCNN	VGG-16	C	2016
<i>MS CNN</i> [20]	10.00%	Faster RCNN	VGG-16	C&K	2016
<i>CMFs</i> [48]	8.93%	RCNN	VGG-16	C&K	2016
<i>JL-TopS</i> [49]	16.60%	Faster RCNN	VGG-16	C	2017
<i>UDN+</i> [50]	13.36%	—	VGG-16	C&E	2017
<i>F-DNN</i> [41]	8.65%	SSD	VGG-16	C&E&T	2017
<i>PCN</i> [51]	8.45%	Faster RCNN	VGG-16	C&I	2017
<i>ADM</i> [52]	8.64%	Faster RCNN	ResNet-50	C&E&T	2018
<i>TLL</i> [53]	8.45%	FCN	ResNet-50	C&CP&K	2018
<i>Bi-box</i> [54]	7.60%	Fast RCNN	VGG-16	C&CP	2018
<i>RepLoss</i> [55]	5.00%	Faster RCNN	ResNet-50	C&CP	2018
<i>ALFNet</i> [40]	4.50%	SSD	ResNet-50	C&CP	2018
<i>OR CNN</i> [56]	4.10%	Faster RCNN	VGG-16	C&CP&E&I	2018
<i>SSA CNN</i> [31]	6.27%	Faster RCNN	VGG-16	C&CP	2019
<i>CSP</i> [57]	3.80%	FCN	ResNet-50	C&CP	2019

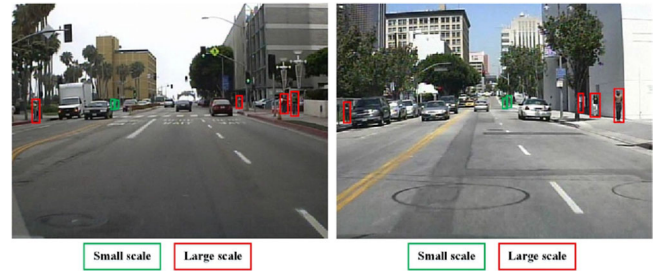
The reason is that the most suitable method is still not found to balance the high speed and high precision of single-stage network framework in dealing with complex scenarios.

3 | OCCLUDED AND MULTI-SCALE PEDESTRIAN DETECTION

According to the results of some datasets, such as Caltech, KITTI, etc., most of the detectors have a good ability for pedestrians with high visibility. However, for pedestrians with low visibility, this performance cannot be guaranteed the same. Especially when pedestrians are in occlusion or long distance, the detector will appear false positives and false negatives. Therefore, there are a lot of researches focused on how to deal with pedestrian occlusion and multi-scale pedestrian detection. The aim of this section is to review progress over the five years of pedestrian detection in deep learning (20+ methods), identify the main ideas explored for occlusion and multi-scale pedestrian detection, and try to quantify which ideas had the most impact on final detection quality.

Table 3 provide a quantitative and qualitative overview of some methods whose results are published on the Caltech pedestrian detection benchmark. Since some methods are shown in Tables 1 and 2, they are not listed in Table 3. These data are from the respective research reports and Caltech data disclosure website.

We refer to the above table, classify each method according to its key performance, and summarise the current methods of dealing with occlusion and multi-scale problems in pedestrian detection. These methods are listed because they have optimal performance at that time. Other methods with similar characteristics will be included in the summary below.

**FIGURE 7** Multi-scale pedestrian sample

3.1 | Multi-scale pedestrians

Although the production of CNN has made outstanding achievements and solved some problems in pedestrian detection, the research on multi-scale pedestrian detection method is still in progress. The main difference between large-scale and small-scale pedestrians is that large-scale pedestrians can provide more abundant information for pedestrian detection, while small-scale pedestrians, because of their low pixels and large noise, have blurred boundaries and appearance which makes it difficult to distinguish them from complex backgrounds and other objects. The comparison of large-scale and small-scale pedestrians is shown in Figure 7.

3.1.1 | Multi-scale proposals or feature maps

The initial *RCNN* focused on the sampling of multi-scale object in the process of generating proposals, but the excessive number of proposals led to the inefficiency of its calculation. Although

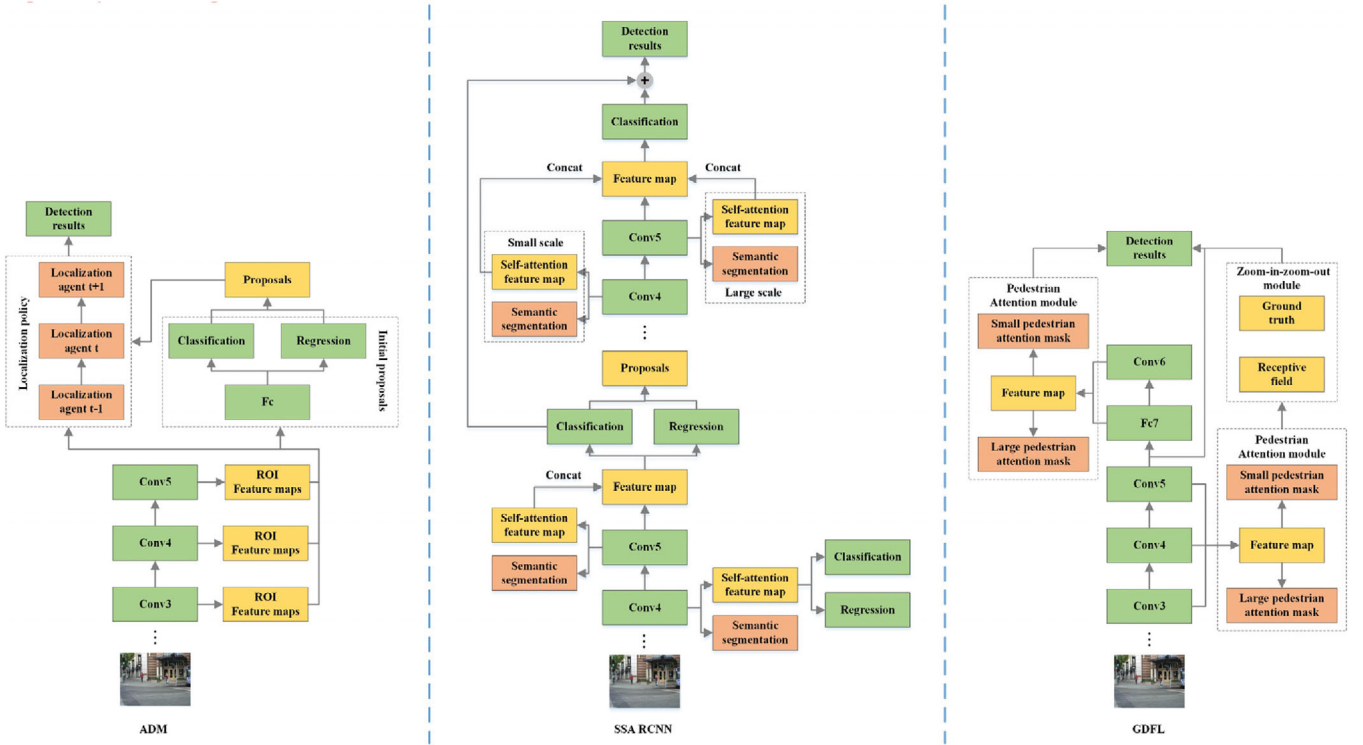


FIGURE 8 The network structure of *ADM*, *SSA RCNN* and *GDFL*

the proposals of RPN solve this problem to some extent, the proposals for small-scale pedestrians are not fully covered. *SAF RCNN* [26] and *MS-CNN* [20] extended *Fast RCNN* and *Faster RCNN* to deal with scale change, respectively. *F-DNN* [41] used multiple deep classifiers combined with soft filters to further validate each proposal. *SSD* [36] divided the output feature map into a set of template boxes by using boundary boxes with different aspect ratios and proportions, and then constructed multi-scale target detector using complementary detection method in different output layers. Recent studies have different views on multi-scale pedestrian detection. But they have similarities, that is, they all consider the impact of the region proposals generation on multi-scale pedestrians.

In Figure 8, we briefly describe the details of the relevant methods to deal with multi-scale detection problems. Zhang et al. [52] proposed a new multi-scale pedestrian detection method (Active Detection method, *ADM*). Based on the characteristics of ResNet and *RCNN*, the multi-layer convolution feature of the input image and the initial pedestrian proposals were taken as input, and the coordinate transformation action sequence was carried out to realise the accurate prediction of boundary frames of different scales. *GDFL* [33] proposed a scale-aware pedestrian attention module to guide the detector to focus on pedestrian regions. It calculated the probability of pedestrian presence at each pixel by generating pedestrian attention masks and integrated the masks with the convolution feature map after coding, which not only highlights the pedestrian but also significantly eliminated the background interference, and improved the recognition ability of small-scale pedestrian and occluded pedestrian. *SSA CNN* [31] proposed a multi-

scale and multi task learning framework. By learning pedestrian detection and semantic segmentation from the multi-scale network layer, the semantic information with different granularity is integrated with the shared feature maps. It connects two semantic segmentation branches to different scale network layer to obtain multi-scale semantic feature map. Then, the multi-scale semantic feature map is used as the semantic clue and connected with the corresponding convolution feature map to provide the pixel level classification information, improve the classification ability of pedestrians, and reduce the difficulty of pedestrian bounding box regression.

By comparing the above literature, the key point of dealing with multi-scale pedestrian detection is whether the low pixel pedestrian features can be accurately extracted in the feature extraction stage. In this process, it plays a decisive role in the generation of the proposed area and the operation of the feature map. Therefore, in order to solve the problem of multi-scale pedestrian detection, it is necessary to eliminate the background interference accurately without increasing the calculation cost, and at the same time to ensure that the deep convolution feature map does not lose the low-pixel pedestrian information.

3.1.2 | Different training and classification strategy

Another way to deal with multi-scale detection problem is to use different stages of classification strategy. In *Comp ACT* [25] and *RPN+BF* [24], cascade boosting and BF classifiers are used to classify images with different resolution under deep feature

maps, the characteristics of small-scale image are fully mined, and are not limited by the structure of pre-training network. Similar methods are used in [48] to classify multi-scale deep convolution feature maps by using the boosted decision forests. It trained a group of enhanced boosted decision forests through multi-layer convolution feature map, and effectively improved the detection ability of the detector for multi-scale pedestrians by using the enhanced boosted decision forests of different scales trained. *SAF RCNN* [26] classifies multi-scale proposals by training sub networks of different sizes. And *ALF* [40] refines the classification results by cascading regression on multi-scale feature maps. In addition, [58] adopted an unsupervised training deep network, which combines multi-step global feature and local feature classification. It used multi-stage features and connections that skip layers to integrate global shape information with local distinctive motif information, especially the unsupervised method based on convolutional sparse coding to pre-train the filters at each stage. The method of unsupervised training and fusion of various feature maps ensures that the detector can adapt to the changes of pedestrians with different pixel sizes, so as to enhance the detection ability of pedestrians with small pixels.

It is proved to be very effective to improve the detector's ability to detect small-scale pedestrians based on improved training and classification methods. The core of these methods is to train classifiers for different scale feature maps and enhance the sensitivity of classifiers to low pixel features. In addition, combining the feature map of CNN with classifier training at different depths also has a better performance.

3.1.3 | Annotation method

The pedestrian detection method based on deep learning needs to input a certain number of labelled images to train the CNN. The quality of the input image determines the detection ability of the trained detector. Among them, the size, resolution and label position of the image affect the accuracy of the detector after training. Therefore, some researchers explore how to label pedestrian images with different scales to guide the feature extraction ability of CNN for small-scale pedestrians. In order to better realise the ability of the detector to learn small-scale pedestrians, Song et al. [53] analysed the bias of image boundary frames in the training stage, and a multi-scale pedestrian detection method (*TLL*) was proposed based on the topological line localisation and temporal feature aggregation. By establishing the topological information of human body model in different scales, the topological information is used as the annotated training detector in the training stage as shown in Figure 9. Pedestrians over different scales could be modelled as a group of 2D Gaussian kernels. And a post-processing scheme based on Markov random field is proposed to improve the positioning accuracy under crowd occlusion. Zhang et al. [47] verified that the initial annotation information plays a decisive role in detector training. Through more detailed post annotation training in Caltech dataset, MR is 3% lower than before. In addition, *CMFs* [48] used additional pixel annotation to improve the perfor-



FIGURE 9 The annotation method of *TLL* [53]

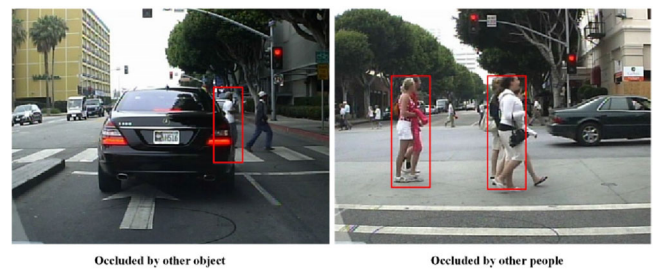


FIGURE 10 The examples of pedestrian occlusion

mance of the detector. They proposed a combination of sliding window detectors and semantic pixel labelling, and used the weighted sum of pixel labelling scores within a proposal region to represent the score of the proposal. This method can ensure that the pixel information can be retained in the training process and achieve the purpose of identifying low pixel pedestrians. On the other hand, the annotation method can also improve the detector's ability to deal with occlusion problems, and the typical ones are *JL-TopS* [49]. The human body parts are labelled in different levels during training, which promotes the detector's perception of occlusion.

The multi-scale detection capability is only part of the performance of pedestrian detector. Its development is from the initial single-scale detection method to the feature pyramid, then to the multi-detection model and advanced semantics assistant detection. Its inherent purpose is to enable the detector to deeply mine the feature differences of different scale images in the training process, so as to obtain more general purpose detection performance.

3.2 | Pedestrian occlusion

Although good performance has been achieved on some benchmark datasets for detecting non-occluded or lightly occluded pedestrians, the performance of detecting heavily occluded pedestrians cannot be guaranteed. Generally speaking, occlusion can be divided into two categories: Inter-class and intra-class occlusions, as shown in Figure 10. The former occurs

when objects are occluded by the other categories of objects, and the latter is also called group occlusion which occurs when objects are occluded by the same category of objects. The processing of occlusion problem has always been a difficult and key point in pedestrian detection. Its research direction is divided into two parts: Recognition and location. Recognition-based occlusion processing mainly focuses on how to identify the person under occlusion by constructing the detector of the corresponding component model, but it cannot locate a specific occlusion effectively. Location-based occlusion processing methods focus on pedestrian positioning frame regression and non-maximum suppression algorithm.

3.2.1 | Part classifier integration

Due to the lack of information when the pedestrian is occluded, the detector cannot accurately identify whether the information is from a pedestrian, and cannot effectively locate the pedestrian's part. The deformable part model was used to deal with occlusion at the beginning of the proposed method, but the problem of large consumption in the calculation of the model cannot be solved well. According to the characteristics of different parts of human body, local detectors of different parts were constructed in research [59] and *Deep parts* [28] to deal with occlusion. By designing feature descriptions of different parts, classifiers are used to train this part. Because most of these local detectors are designed manually, their performance may not be the most ideal, and multiple local detectors will also result in higher computational costs.

Zhou et al. [49] proposed a multi-label learning method which has better detection ability for pedestrian under severe occlusion, and can improve the accuracy of the detector while reducing the calculation cost, but has lower detection performance for normal pedestrian than other methods. In order to break the limitations brought by multi-local detectors, Ouyang et al. [60] integrated the detectors with different occlusion degrees to reduce the detection time greatly. On this basis, *UDN* [50] improved it by combining learning deep feature extraction, deformation processing, occlusion processing and pedestrian detection model, as in Figure 11, and merged deformation layer into convolution neural network, so that component model can play a greater role. In addition, [61] proposed a detection method to improve the processing performance of the single-stage detector for occlusion. By dividing the prediction confidence into parts to correct the overall detection confidence, the effect is significant in *SqueezeDet+* [62], *YOLOv2*, *SSD* and *DSSD* [63] detectors.

The most widely used method to improve the detection ability of occluded pedestrian is based on part model-assisted global detection. In this process, a series of human part models are designed and combined with convolution features to improve the detection rate. However, when using the part model, there will be additional calculation consumption. Therefore, the core of the detection method based on part model is to improve the recognition rate of the detector to the occluded pedestrian without reducing the detection speed.

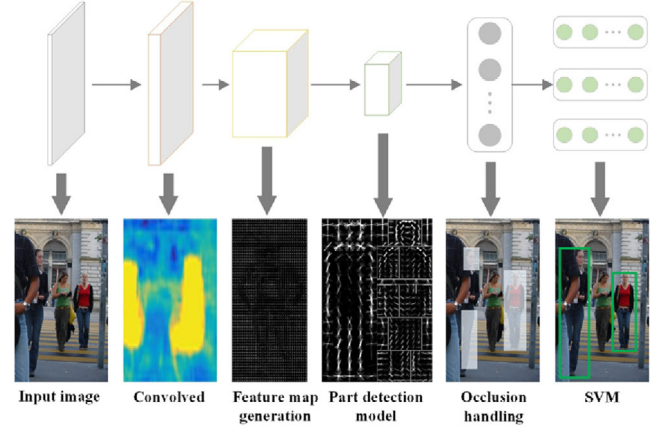


FIGURE 11 *UDN* occlusion processing strategy

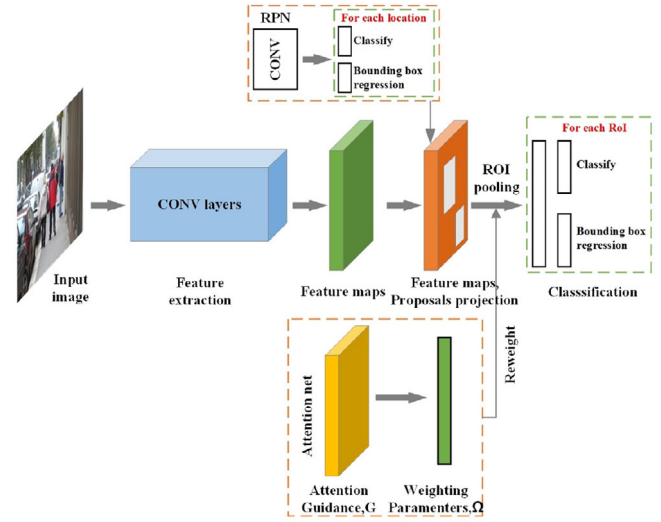


FIGURE 12 *Faster RCNN* and attention mechanism for handling occlusion

3.2.2 | Attention mechanism

The basic paradigm of the component model method is to design and learn a set of pedestrian part detectors, and to process partially occluded pedestrians by adjusting the fusion part detection using the corresponding mechanism. Similar to this method from local detection to overall detection, some researchers use the attention mechanism to focus on the key parts of pedestrians to achieve the purpose of accurate detection of the whole pedestrian. Research [64] founded that the channel characteristics of specific parts of the human body can be positioned. By adding an additional attention network to the *Faster RCNN* architecture, the channel attention mechanism is applied to deal with different degrees of occlusion as shown in Figure 12. During the training process, self-attention, visual frame attention and partial attention are added to increase the ability of detectors in occlusion situation. *SSA RCNN* [31] and *GDFL* [33] also used attention mechanism to improve the accuracy of the detector, and combine semantic segmentation to ensure the robustness of the detector. As a high-level feature, semantic segmentation is gradually used in pedestrian

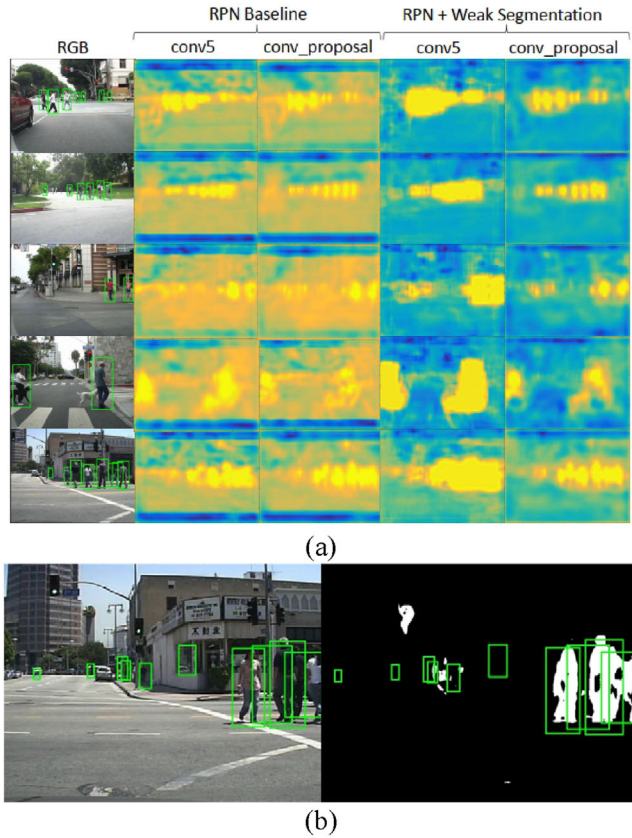


FIGURE 13 Semantic segmentation for pedestrian detection in *SDS RCNN* [29] and *F-DNN* [41]

detection. Costea [32] used semantic segmentation as visual cues to help the detector better understand the environment by dividing pixels into pedestrian, road, vehicle and other semantic classes to quickly distinguish foreground and background; using semantic segmentation as the context information of pedestrian detection, semantic channel integration detection process, optimisation feature extraction, multi-scale sliding window and classification scheme to reduce the calculation cost. Furthermore, as in Figure 13, *SDS RCNN* [29] and *F-DNN* [41] injected semantic segmentation model to promote the mining of deep features at the stage of network impassability, respectively. In this way, the possible pedestrian regions are generated quickly, and the feature map generated by the end of convolution network is adjusted according to the attention mechanism of semantic segmentation, that is, to guide the CNN to focus on the blocked pedestrian parts. *PCN* [51] combined component model and semantic features to deal with occlusion and uses LSTM [65] to refine the semantic features of different parts, and obtains the final confidence score for recognition. At present, center and scale prediction base detection (*CSP*) [57], the best performance detection algorithm, also uses high-level semantic features to simplify the pedestrian detection process into the prediction of the centre point and scale, reducing the complexity of the post-processing stage.

The key to improve pedestrian detection performance by attention mechanism is to combine semantic segmentation fea-

tures with convolutional feature maps. Pixel segmentation of pedestrians and other objects, on the one hand, can quickly locate the possible areas of pedestrians and improve the detection efficiency. On the other hand, it can mine the pedestrian pixels that are not easy to be found due to occlusion, so as to solve the problem of occlusion pedestrian detection.

3.2.3 | Location regression and network post-processing

The above two methods are to construct detectors from the perspective of how to recognise the occluded pedestrians. Simultaneously, in order to improve the detection rate of the detector under occlusion in crowd, researchers analysed the occlusion problem from other angles. Hosang et al. [71] were committed to improving the robustness of NMS [66], but an additional network is needed to post-process occlusion ultimately. The accuracy of the post-processing phase is improved by combining NMS with detection network and defining double penalty to complete joint training. It solves the problem of false detection caused by too many positioning frames in the post-processing stage of occluded pedestrian detection. Liu et al. [67] present a new *Adaptive-NMS* method to better refine the bounding boxes in crowded scenarios and applied a dynamic suppression strategy, where an additionally learned sub-network is designed to predict the threshold according to the density for each instance. This threshold increases when pedestrians gather or occluded each other and decreases when pedestrians appear alone. In addition, an additional subnetwork is used to predict the adaptive NMS threshold of each pedestrian so as to avoid positioning errors caused by mutual occlusion between pedestrians. *Bi-box* [54] constructed a convolution network consisting of two branches, the pedestrian's whole body and visible part were located by regressing two boundary boxes. In the training process, the pedestrian proposals with different confidence levels are optimised to ensure the detection accuracy under severe occlusion.

To solve the problem of high false negative rate in crowd scenarios, occlusion-aware convolutional neural networks (*OR CNN*) [56] designed a new aggregation loss to force the proposals to approach and locate the corresponding object compactly. And it constructed a new local occlusion ROI pool unit, which integrated the prior structure information and visibility prediction of human body into the network to process occlusion. *RepLoss* [55] designed a new boundary frame repulsion loss function for crowd scenes specially, by analysing the positioning accuracy of various advanced detectors in dealing with occluded crowds. The robustness of pedestrian localisation in the group is improved by preventing the proposals from transferring to the nearby objects.

Regardless of the two-stage or single-stage framework, positioning strategy is the most important in the process of network design. Aiming at the operation mechanism of positioning processing occlusion gradually focusing on the deep convolution network itself, the performance of positioning processing occlusion is changing from improving the overall performance

TABLE 4 Comparison of several mainstream pedestrian detection datasets

Dataset name	Started year	Number of labels	Bounding box range and pixel ratio		Size	Characteristic
Caltech	2009	350,000	Reasonable	$0.65 < [50, \infty) < \infty$	640×480	Large amount of data and rich annotation information
			Small	$[30, 80]$		
			Occluded	$0.2 < [50, \infty) < 0.65$		
KITTI	2012	80,000	Visible	Visibility above 20%	1392×512	Including 3D annotation, mostly for autopilot
			Occluded			
			Truncation			
Cityperson	2016	35,016	Reasonable	$0.65 < [50, \infty) < \infty$	2048×1024	Add city background annotations for training

of the network to deepening the performance of each processing stage.

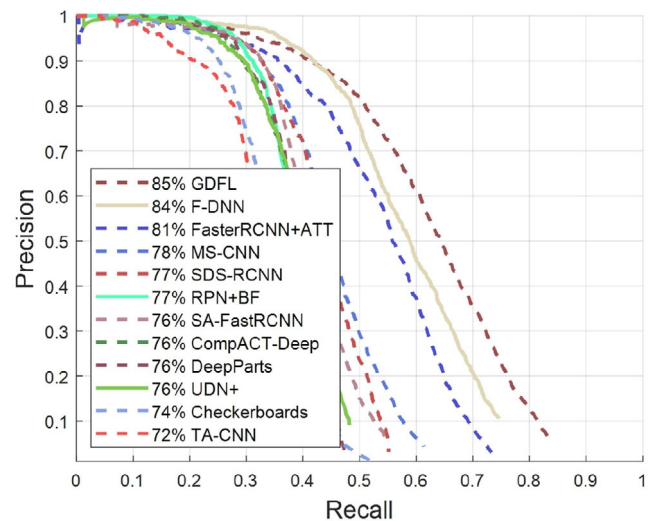
4 | DATASET AND EVALUATION METHOD

In the past decades, more research results have been achieved and a large-scale research system has been formed. The application of pedestrian detector has promoted its rapid development in recent years. In order to test the performance of pedestrian detector better, unlike the general object detection datasets, the field of pedestrian detection has formed some special datasets and evaluation methods for the performance evaluation of pedestrian detector.

4.1 | Pedestrian detection dataset

Dataset is one of the foundations and decisive factors in the process of pedestrian detection research. It is not only the common basis for measuring and comparing the performance of competitive algorithms, but also a powerful assistant to promote the development and progress of research in this field. The number of datasets and the quality of annotated information are critical for training detectors. Detectors need more data to enrich their ability to adapt to multiple scenarios, and accurate annotation information can better guide the detector to learn what it needs. So far, the published pedestrian datasets include MIT [4], INRIA [3], Daimler [7], Caltech [23], KITTI [30], TUD [44], NICTA [68], ETH [8], CVC [69], USC [70], and Citypersons [27] pedestrian datasets. According to the different content of each dataset, each dataset has its own characteristics. Among them, Caltech, KITTI and Citypersons pedestrian datasets have more complete annotation information and better annotation for occluded and multi-scale scenes, hence, they are most widely used. We summarise these three common pedestrian detection datasets in detail as shown in Table 4.

Caltech dataset is the largest pedestrian dataset at present, which is photographed by car camera with about 250,000 frames (about 137 min), 350,000 bounding boxes and 2300 pedestrian annotations. In addition, the time correspondence between rectangular frames and their occlusion are also labelled.

**FIGURE 14** Precision–recall curve in Caltech dataset

KITTI dataset is the largest dataset under automatic driving scenario in the world including 180G data. It can be used to evaluate the performance of 3D image, optical flow, visual ranging, 3D object detection and 3D tracking in vehicle environment.

Citypersons dataset is a pedestrian detection dataset obtained by tagging on cityscapes dataset which is mainly urban road pedestrians, tagging occlusion information and small-scale pedestrian information between objects and pedestrians.

4.2 | Evaluation method

The detection ability of pedestrian detector is reflected by the corresponding evaluation index and the correct evaluation method plays a decisive role in the process of detection performance evaluation. At present, the evaluation of detector performance is based on the test set of dataset. In this section, taking the results of various detection algorithms in Caltech dataset as examples, the measurement standards of different evaluation methods for detector performance are described in detail.

Figure 14 shows the comparison of the results of several deep learning based pedestrian detection algorithms in Caltech dataset. The precision-recall curve [22] is used to evaluate the performance of each algorithm. The more convex the curve is

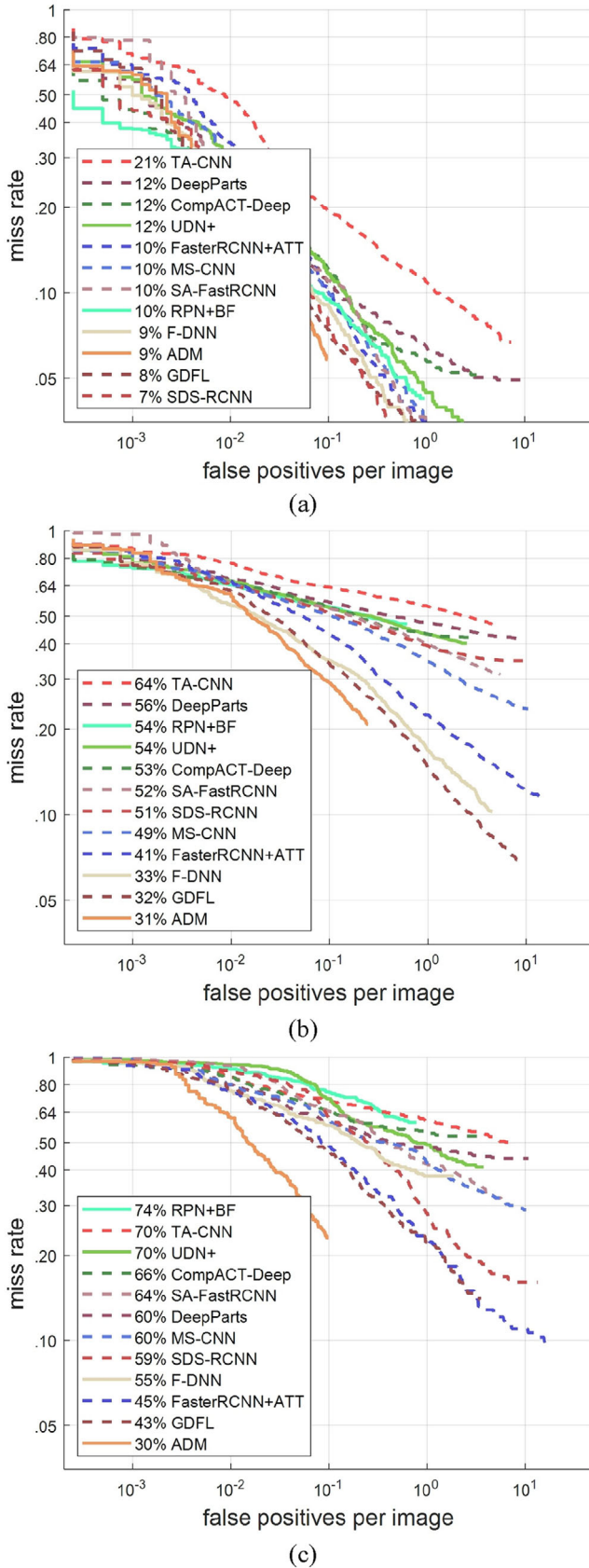


FIGURE 15 MR-FPPI curve comparison of several advanced algorithms in Caltech dataset

on the right, the better the performance is. The optimal algorithm *GDFL* obtains 85% detection accuracy. Precision and recall are common evaluation indexes in general object detection. Precision refers to the correct proportion of the object predicted by the detector, and recall rate is the proportion of the object correctly positioned to the total number of objects.

Scholars in the analysis of detector error cases found that only use precision-recall curve cannot be refined to prove the effectiveness of the detector for a variety of scenarios, that is, false positives and false negatives cannot be better displayed. Based on the general object evaluation method, Piotr Dollar puts forward the curve of the number of false positive per image (or false positive per window) of each image with the log average miss rate referred to as MR-FPPI/FPPW curve [22]. Figure 15 shows the evaluation results of several advanced algorithms in Caltech dataset. Take the average miss rate of FPPI under 10^{-2} , the lower the curve, the better the performance.

Figures 15 (a), (b) and (c), respectively, show the detection results of three different types of pedestrians: Reasonable, small-scale and occlusion, in Caltech dataset. Among them, for reasonable pedestrians, *SDS-RCNN* achieves an average miss rate of 7% of the optimal performance energy. In the detection of small-scale pedestrians, *ADM* achieves the best 31% average miss rate. At the same time, *ADM* has a better detection ability for pedestrians under occlusion.

5 | CONCLUSION

Pedestrian detection is an important and challenging problem in computer vision which has attracted wide attention of researchers. It has made tremendous progress from the initial manual design features combined with traditional machine learning methods to the widely used deep learning methods, but at the same time accompanied by new problems. In this study, we have not only extensively reviewed some milestone detectors (e.g. *RCNN*, *Faster-RCNN*, *YOLO*, *SSD* etc.), but also analysed the two-stage and single-stage detection frameworks based on deep learning and the state-of-the-art improved methods emphatically. In addition, we have discussed the challenges of the occlusion and multi-scale, which currently met by all of the detectors, and how these detectors solve these problems further extended and improved.

In the past few years, pedestrian detection has achieved tremendous success. The latest *CSP* has achieved 3.8% of MR results on ordinary Caltech datasets, but it still has a huge gap with human capabilities. Especially when the detector is embedded in the embedded system, its actual effect is often not as good as the test results. Therefore, in order to promote the pedestrian detection to be better applied to the fields of automatic driving, robots and so on, there is still much work to be done:

Pedestrian dataset annotation method: Although the most representative Caltech and Cityperson datasets currently include pedestrians in various scenes, annotation information also includes occlusion, small scale, etc. But for CNN, there is still room for expansion. For the expansion of datasets and

high-quality annotation, it can be of great help for training detectors. For image annotation in special scene, it can help the detector to recognise occlusion and multi-scale pedestrian.

Additional information to enhance pedestrian characteristics: Deep convolution network can extract higher dimension pedestrian features, but with the increase of the deep of the network some useful shallow information will also be missing. The fusion of convolution features of different dimensions is an effective method to enhance the feature description. In addition, some of the features extracted artificially can also enhance the deep features of pedestrians. Therefore, enhancing occlusion and small-scale pedestrian characteristics is an important way to improve the performance of the detector.

Pedestrian detection based on multi-task fusion: The reason why visual information is widely praised is that it is rich in information, while object detection is only one of the tasks of machine vision, and other tasks include semantic segmentation, instance segmentation, human pose recognition, etc. Using special methods can make one visual task assist another visual task and is an effective way to improve the effect. Therefore, it is an important direction to integrate multiple visual tasks.

Multi-form and 3D pedestrian detection: At present, most of the pedestrian detection objects are for upright pedestrians, but considering its application scenarios, the constructed detector may not be able to recognise some special pedestrian states such as sitting, squatting, riding and so on. Therefore, it is necessary to deeply mine the common features of multi-modal pedestrians to enhance the generalisation ability of pedestrian detector. In addition, the current research on pedestrian detection is basically carried out on 2D images, and the research on pedestrian detection on 3D images is still less.

Detection with information fusion: Pedestrian detection with multiple sources/modalities of data, etc, RGB-D image, 3D point cloud, LIDAR, etc, is of great importance for autonomous driving and drone applications. Some open questions include: How to immigrate well-trained detectors to different modalities of data; how to make information fusion to improve detection, etc.

ACKNOWLEDGEMENTS

This research is supported by National Natural Science Foundation of China (51805490) and (51805491). In addition, this research is supported by National Key Research and Development Program of China (2017YFD07012042), and Major Scientific and Technological Projects in Henan Province (191110210100).

ORCID

Yanqiu Xiao  <https://orcid.org/0000-0001-5152-1736>

REFERENCES

- Benenson, R., et al.: Ten years of pedestrian detection, what have we learned? In: The 2014 European Conference on Computer Vision, pp. 613–627. Springer, Berlin (2014)
- Paul, V., et al.: Detecting pedestrians using patterns of motion and appearance. In: The 9th IEEE International Conference on Computer Vision, pp. 734–741. IEEE Computer Society, Washington DC (2003)
- Navneet, D., Bill, T.: Histograms of oriented gradients for human detection. In: The 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–93. IEEE Computer Society, Washington DC (2005)
- Constantine, P., Tomaso, P.: A trainable system for object detection. *Int. J. Comput. Vision* 38(1), 15–33 (2000)
- Dollar, P., et al.: Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(8), 1532–1545 (2014)
- Lahouli, I., et al.: Hot spot method for pedestrian detection using saliency maps, discrete Chebyshev moments and support vector machine. *IET Image Proc.* 12(7), 1284–1291 (2018)
- Enzweiler, M., Gavrilu, D.: Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(12), 2179–2195 (2009)
- Ess, A., et al.: A mobile vision system for robust multi-person tracking. In: The 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2008)
- Felzenszwalb, P., et al.: A discriminatively trained, multiscale, deformable part model. In: The 26th IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2008)
- Alex, K., et al.: ImageNet classification with deep convolutional neural networks. In: The 26th Annual Conference on Neural Information Processing Systems, pp. 1097–1105. Curran Associates Inc., Red Hook, New York (2012)
- Deng, J., et al.: ImageNet: A large-scale hierarchical image database. In: The 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE Computer Society, Washington DC (2009)
- Ross, G., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: The 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE Computer Society, Washington DC (2014)
- Girshick, R., et al.: Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(1), 142–158 (2015)
- He, K., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: The 13th European Conference on Computer Vision, pp. 346–361. Springer Verlag, Berlin (2015)
- Girshick, R.: Fast R-CNN. In: The 15th IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE Computer Society, Washington DC (2016)
- Ren, S., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6), 1137–1149 (2015)
- Xiaofei, W., et al.: Similarity-preserving hashing based on deep neural networks for large-scale image retrieval. *J. Visual Commun. Image Represent.* 61, 260–271 (2019)
- Yao, L., Wang, B.: Pedestrian detection framework based on magnetic regional regression. *IET Image Proc.* 13(9), 1431–1436 (2019)
- Dai, J., et al.: RFCN: Object detection via region based fully convolutional networks. In: The 30th Annual Conference on Neural Information Processing Systems, pp. 379–387. Neural Information Processing Systems, La Jolla (2016)
- Cai, Z., et al.: A unified multi-scale deep convolutional neural network for fast object detection. In: European Conference on Computer Vision, pp. 354–370. Springer Verlag, Berlin (2016)
- Kaiming, H., et al.: Mask R-CNN. In: The 16th IEEE International Conference on Computer Vision, pp. 298–2988. IEEE Computer Society, Washington DC (2018)
- Wojek, C., et al.: Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* 34(4), 743–761 (2012)
- Dollar, P., et al.: Pedestrian detection: A benchmark. In: The 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 304–311. IEEE Computer Society, Washington DC (2009)

24. Zhang, L., et al.: Is Faster R-CNN doing well for pedestrian detection? In: European Conference on Computer Vision, pp. 443–457. Springer Verlag, Berlin (2016)
25. Cai, Z., et al.: Learning complexity-aware cascades for deep pedestrian detection. In: The 2015 IEEE International Conference on Computer Vision, pp. 3361–3369. IEEE Computer Society, Washington DC (2015)
26. Li, J., et al.: Scale-aware Fast R-CNN for pedestrian detection. *IEEE Trans. Multimedia* 20(4), 985–996 (2018)
27. Shanshan, Z., et al.: Citypersons: A diverse dataset for pedestrian detection. In: The 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 4457–4465. IEEE Computer Society, Washington DC (2017)
28. Tian, Y., et al.: Deep learning strong parts for pedestrian detection. In: The 15th IEEE International Conference on Computer Vision, pp. 1904–1912. IEEE Computer Society, Washington DC (2015)
29. Brazil, G., et al.: Illuminating pedestrians via simultaneous detection and segmentation. In: The 16th IEEE International Conference on Computer Vision, pp. 4960–4969. IEEE Computer Society, Washington DC (2017)
30. Geiger, A., et al.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE Computer Society, Washington DC (2012)
31. Chengju, Z., et al.: SSA-CNN: Semantic self-attention CNN for pedestrian detection. In: IEEE International Conference on Computer Vision. pp. 4321–4330. IEEE Computer Society, Washington DC (2019)
32. Costea, A.D., Nedevschi, S.: Semantic channels for fast pedestrian detection. In: Computer Vision and Pattern Recognition, pp. 2360–2368. IEEE Computer Society, Washington DC (2016)
33. Chunze, L., et al.: Graininess-aware deep feature learning for pedestrian detection. In: European Conference on Computer Vision, pp. 745–761. Springer Verlag, Berlin (2018)
34. Jiayuan, M., et al.: What can help pedestrian detection? In: Computer Vision and Pattern Recognition, pp. 6034–6043. IEEE Computer Society, Washington DC (2017)
35. Redmon, J., et al.: You only look once: Unified, real-time object detection. In: Computer Vision and Pattern Recognition, pp. 779–788. IEEE Computer Society, Washington DC (2015)
36. Wei, L., et al.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer Verlag, Berlin (2016)
37. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Computer Vision and Pattern Recognition, pp. 6517–6525. IEEE Computer Society, Washington DC (2017)
38. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. In: Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2018)
39. Zhaowei, C., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: Computer Vision and Pattern Recognition, pp. 6154–6162. IEEE Computer Society, Washington DC (2017)
40. Wei, L., et al.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: European Conference on Computer Vision, pp. 643–659. Springer Verlag, Berlin (2018)
41. Xianzhi, D., et al.: Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In: 17th IEEE Winter Conference on Applications of Computer Vision, pp. 953–961. Institute of Electrical and Electronics Engineers Inc, New Jersey (2017)
42. Jiwoong, C., et al.: Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2019)
43. Ya-Li, L., Shengjin, W.: HAR-Net: Joint learning of hybrid attention for single-stage object detection. In: Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2019)
44. Wojek, C., et al.: Multi-cue onboard pedestrian detection. In: Computer Vision and Pattern Recognition, pp. 794–801. IEEE Computer Society, Washington DC (2009)
45. Hosang, J., et al.: Taking a deeper look at pedestrians. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4073–4082. IEEE Computer Society, Washington DC (2015)
46. Tian, Y., et al.: Pedestrian detection aided by deep learning semantic tasks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5079–5087. IEEE Computer Society, Washington DC (2015)
47. Zhang, S., et al.: How far are we from solving pedestrian detection? In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2016)
48. Hu, Q., et al.: Pushing the limits of deep CNNs for pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* 28(6), 1358–1368 (2016)
49. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: IEEE International Conference on Computer Vision, pp. 3506–3515. Institute of Electrical and Electronics Engineers Inc., New Jersey (2017)
50. Ouyang, W., et al.: Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1874–1887. Institute of Electrical and Electronics Engineers Inc., Washington DC (2017)
51. Wang, S., et al.: PCN: Part and context information for pedestrian detection with CNNs. In: British Machine Vision Conference. British Machine Vision Association, Durham (2017)
52. Zhang, X., et al.: Too far to see? Not really! Pedestrian detection with scale-aware localization policy. *IEEE Trans. Image Process.* 27(8), 3703–3715 (2018)
53. Song, T., et al.: Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation. In: 15th European Conference on Computer Vision, pp. 554–569. Springer Verlag, Berlin (2018)
54. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: European Conference on Computer Vision, pp. 138–154. Springer Verlag, Berlin (2018)
55. Wang, X., et al.: Repulsion loss: Detecting pedestrians in a crowd. In: The IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7774–7783. IEEE Press, Washington DC (2018)
56. Zhang, S., et al.: Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: European Conference on Computer Vision, pp. 657–674. Springer Verlag, Berlin (2018)
57. Liu, W., et al.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Computer Vision and Pattern Recognition. IEEE Computer Society, Washington DC (2019)
58. Sermanet, P., et al.: Pedestrian detection with unsupervised multi-stage feature learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3626–3633. IEEE Computer Society, Washington DC (2013)
59. Zhou, C., Yuan, J.: Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Asian Conference on Computer Vision, pp. 305–320. Springer Verlag, Berlin (2016)
60. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: IEEE International Conference on Computer Vision, pp. 2056–2063. Institute of Electrical and Electronics Engineers Inc, Washington DC (2013)
61. Noh, J., et al.: Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 966–974. IEEE Computer Society, Washington DC (2018)
62. Wu, B., et al.: SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. *arXiv:1612.01051* (2016)
63. Fu, C.-Y., et al.: DSSD: Deconvolutional single shot detector. *arXiv:1701.06659* (2017)
64. Zhang, S., et al.: Occluded pedestrian detection through guided attention in CNNs. In: Conference on Computer Vision and Pattern Recognition, pp. 6995–7003. IEEE Computer Society, Washington DC (2018)
65. Karpathy, A., et al.: Visualizing and understanding recurrent networks. <http://vision.stanford.edu/pdf/KarpathyICLR2016.pdf> (2015)
66. Bodla, N., et al.: Soft-nms: Improving object detection with one line of code. In: IEEE International Conference on Computer Vision, pp. 5562–5570. Institute of Electrical and Electronics Engineers Inc, Washington DC (2017)

67. Liu, S., et al.: Adaptive NMS: Refining pedestrian detection in a crowd. arXiv:1904.03629 (2019)
68. Overett, G., et al.: A new pedestrian dataset for supervised learning. In: IEEE Intelligent Vehicles Symposium, pp. 373–378. Institute of Electrical and Electronics Engineers Inc., Washington DC (2008)
69. Marin, J., et al.: Learning appearance in virtual scenarios for pedestrian detection. In: The IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 137–144. IEEE Computer Society, Washington DC (2010)
70. Mi, Z., Sawchuk, A.A.: USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In: The 2012 ACM Conference on Ubiquitous Computing, Association for Computing Machinery, pp. 1036–1043. ACM, New York (2012)
71. Hosang, J., et al.: Learning non-maximum suppression. In: IEEE Conference on Computer Vision and Pattern Recognition 2017. pp. 1063–6919. IEEE Computer Society, Washington DC (2017)

How to cite this article: Xiao Y, Zhou K, Cui G, Jia L, Fang Z, Yang X, Xia Q. Deep learning for occluded and multi-scale pedestrian detection: A review. *IET Image Process* 2021;15:286–301.
<https://doi.org/10.1049/ipr2.12042>