From Handcrafted to Deep Features for Pedestrian Detection: A Survey

Jiale Cao, Yanwei Pang, *Senior Member, IEEE*, Jin Xie, Fahad Shahbaz Khan, *Senior Member, IEEE*, and Ling Shao, *Fellow, IEEE*

Abstract—Pedestrian detection is an important but challenging problem in computer vision, especially in human-centric tasks. Over the past decade, significant improvement has been witnessed with the help of handcrafted features and deep features. Here we present a comprehensive survey on recent advances in pedestrian detection. First, we provide a detailed review of single-spectral pedestrian detection that includes handcrafted features based methods and deep features based approaches. For handcrafted features based methods, we present an extensive review of approaches and find that handcrafted features with large freedom degrees in shape and space have better performance. In the case of deep features based approaches, we split them into pure CNN based methods and those employing both handcrafted and CNN based features. We give the statistical analysis and tendency of these methods, where feature enhanced, part-aware, and post-processing methods have attracted main attention. In addition to single-spectral pedestrian detection, we also review multi-spectral pedestrian detection, which provides more robust features for illumination variance. Furthermore, we introduce some related datasets and evaluation metrics, and a deep experimental analysis. We conclude this survey by emphasizing open problems that need to be addressed and highlighting various future directions. Researchers can track an up-to-date list at https://github.com/JialeCao001/PedSurvey.

Index Terms—Pedestrian detection, handcrafted features based methods, deep features based methods, multi-spectral pedestrian detection

1 Introduction

Human-centric computer vision tasks (e.g., pedestrian detection [6], [43], [242], person re-identification [85], [105], [226], [253], person search [65], [136], [211], [219], pose estimation [21], [134], [138], [166], and face detection [101], [131], [161], [224]) have gained significant attention over the past decade. Among these tasks, pedestrian detection is one of the most fundamental tasks with a wide range of real-world-applications. In addition to its standalone value in a variety of applications (e.g., video surveillance and self-driving), pedestrian detection is also a prerequisite that serves as the basis for several other vision tasks (e.g., person re-identification and person search). For instance, both person re-identification and person search need to first accurately detect all the existing pedestrians.

The aim of pedestrian detection is to accurately localize and classify all pedestrian instances in a given image. In the past decade, pedestrian detection has received significant attention with over two thousands research publications (see Fig. 1). The increasing number of publications suggest that it is an active research problem. In recent years, pedestrian detection performance has also obtained a consistent improvement on standard benchmarks. Fig. 2 (top) shows the improvement in pedestrian detection accuracy (in terms of log-average miss rate) on the test set of Caltech [43], which is one of the most popular pedestrian detection benchmarks. The detection performance is evaluated on the reasonable **R**. The reasonable **R** set comprises pedestrians over 50 pixels in height, with less than 35 percent of their body occluded. We compare the performance of 30 methods, including handcrafted, deep learning and hybrid approaches. Note that we split methods into (a) pure deep learning based approaches, comprising end-to-end training where pedestrian proposal generation and classification are learned jointly, and (b) hybrid approaches. In hybrid approaches, some methods use deep features for proposal generation and shallow classifier, such as Support Vector Machines (SVM) [34] or AdaBoost [50], for proposal classification, while some other methods use handcrafted approaches for proposal generation and deep features for proposal classification. In addition, we show recent deep learning based methods (represented by white cross hatch) that utilize more accurate annotations [241]. Despite the consistent progress, we argue that there is still sufficient room for improvement in order to meet real-world application requirements. For instance, Fig. 2 (bottom) shows the detection performance under severe occlusions (heavy occlusion **HO** set of Caltech test dataset). The **HO** set comprises pedestrians over 50 pixels in height, with 35 to 80 percent of their body occluded. These results suggest that the detection performance under real-world challenges, such as occlusion, is still far from satisfactory.

Table 1 compares pedestrian detection with the related human detection and object detection. Compared with human

- *Jiale Cao, Yanwei Pang, and Jin Xie are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. E-mail: {connor, pyw, jinxie}@tju.edu.cn.*
- *Fahad Shahbaz Khan is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi 51133, UAE, and also with the Linköping University, 581 83 Linköping, Sweden. E-mail: fahad.khan@mbzuai.ac.ae.*
- *Ling Shao is with the Inception Institute of Artificial Intelligence, Abu Dhabi 51133, UAE. E-mail: ling.shao@ieee.org.*

Manuscript received 25 September 2020; revised 28 February 2021; accepted 7 April 2021. Date of publication 30 April 2021; date of current version 4 August 2022.
(Corresponding author: Yanwei Pang.)
Recommended for acceptance by V. Morariu.
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2021.3076733, provided by the authors.
Digital Object Identifier no. 10.1109/TPAMI.2021.3076733

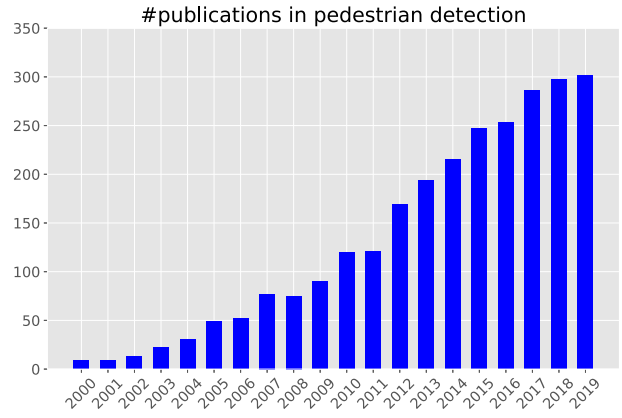

Fig. 1. The increasing number of publications on pedestrian detection from the year 2000 to 2019, obtained through Google scholar search with the key-words: allintitle: "pedestrian detection".

detection, pedestrian detection primarily focuses on driving/surveillance scenes based on visible-light camera and infrared camera. In addition, pedestrian detection detects fully body, including with partial or severe occlusions, and also poses large variations in scale. Compared with generic object detection, pedestrian detection focuses on single category of pedestrian, and faces the challenges, such as frequent occurrence of partial or severe occlusions and large variations in scale. Owing to specific challenges, pedestrian detection has been studied as a standalone problem.

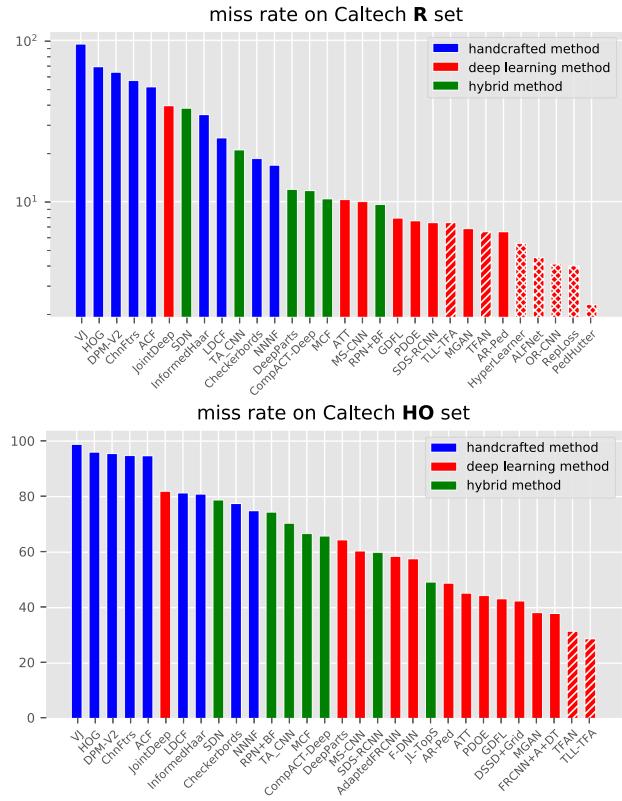

Fig. 2. Detection performance improvements, in terms of log-average miss rate (lower is better), on Caltech test set [43] in past decade. Top: we show the performance comparison on the reasonable (**R**) set. Bottom: We show the comparison on the heavy occluded (**HO**) set. The white cross hatch in bar indicates that more accurate annotations [241] are used for training and test. The white line hatch in bar indicates that motion cue is utilized in addition to appearance information.

TABLE 1
Comparison With Object Detection and Human Detection

Name	Pedestrian detection	Human detection	Object detection
Class	single	single	multiple
Image	RGB/Thermal	RGB	RGB
Scene	driving/ surveillance	unspecific	unspecific
Target	full body	visible part	visible part
Orientation	upright	any	any
Occlusion frequency	large	large	medium
Scale variance	large	medium	medium

Owing to specific challenges, pedestrian detection has been studied as a standalone problem.

In this work, we divide existing pedestrian detection works into single-spectral pedestrian detection and multi-spectral pedestrian detection. Single-spectral pedestrian detection means that only a single sensor is used for detection (e.g., visible-light camera and fisheye camera). Here, we mainly focus on methods based on visible-light camera. Different to single-spectral methods, multi-spectral pedestrian detection adopts multiple sensors of different types. For this class of methods, we mainly focus on those based on visible-light and infrared cameras. Compared with single-spectral pedestrian detection, multi-spectral pedestrian detection is more robust with respect to illumination variation and has attracted considerable attention in the past few years.

The rest of this article is organized as follows. We first introduce the detection pipeline of pedestrian detection in Section 2. After that, we present a detailed review and analysis of single-spectral pedestrian detection approaches in Section 3, including both handcrafted features based methods and deep features based approaches. Then, we introduce multi-spectral pedestrian detection in Section 4, which is a supplement to single-spectral pedestrian detection. An experimental analysis is provided in Section 5. Finally, we discuss several existing challenges in Section 6.

Some surveys about pedestrian detection [6], [54], [159] have been published in past years. Compared with these previous surveys, we focus more attention on the recent deep features based methods, instead of handcrafted features based methods. Further, we present a more detailed analysis. Based on this analysis, we summarize the ongoing challenges in pedestrian detection research. We hope that this survey will not only provide a better understanding of pedestrian detection but also facilitate future research activities and various application developments in the field.

2 Pedestrian Detection Pipeline

Most pedestrian detection methods, including handcrafted [38], [42], deep learning [11], [244] and hybrid [16], [181] approaches, typically comprise three consecutive steps: proposal generation, classification (and regression), and post processing. Fig. 3 shows the overall pipeline depicting these three steps. Note that not all approaches have these three steps. For example, some approaches [118], [119] do not have the step of proposal generation, while the recent PED [109] does not need the NMS post-processing. Without loss of generality, we discuss these three steps in detail.

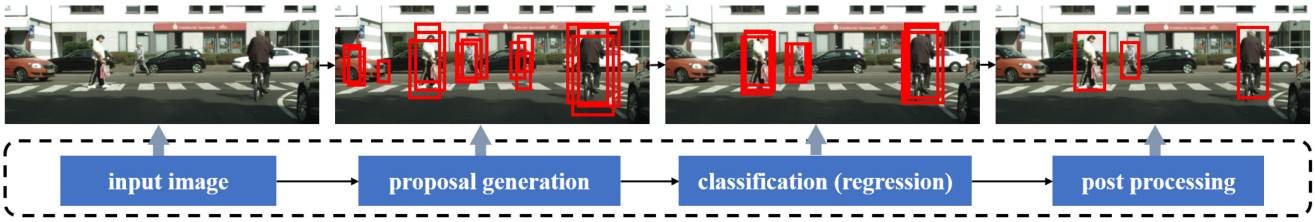

Fig. 3. Most pedestrian detection approaches typically comprise three consecutive steps. The first step, proposal generation, involves generating candidate proposals from an input image. The second step, proposal classification (and regression), involves assigning the proposals to either the positive class (pedestrian) or the negative class (background). Consequently, the post-processing step aims to suppress duplicate bounding-boxes belonging to the same pedestrian. In proposal generation and proposal classification, feature extraction is the key. A variety of feature extraction strategies ranging from handcrafted to deep features have been used in the literature.

(a) *Proposal Generation.* This step aims to extract some candidate proposals of pedestrians from an input image. The proposals indicate a set of bounding-boxes which potentially represent the objects. Common strategies include sliding-window methods [38], [42], [186], particle-window methods [60], [147], objectness methods [28], [69], [184], [265], and region proposal networks [11], [164], [190]. The sliding-window methods (SW) adopt a greedy search strategy with a fixed-sized step to scan the image from the top-left to bottom-right region. The particle-window methods adopt the coarse-to-fine cascaded search where the proposals generated at current stage follow the likelihood distribution of previous stage. The objectness methods typically employ a variety of low-level features (e.g., edge and color features) to extract the proposals in a bottom-up fashion. Recently, a region proposal network (RPN) was introduced for proposal generation, which shares the deep features with the following proposal classification and regression.

(b) *Proposal Classification.* This step assigns these candidate proposals to the positive class (pedestrians) or the negative class (background) based on the extracted features of these proposals. The handcrafted features based methods [6], [38] adopt a shallow classifier (e.g., SVM or boosting) for classification, whereas deep features based methods [37], [117], [164] generally integrate the feature extraction and classification into a unified framework by utilizing a softmax (or sigmoid) layer. Additionally, deep features based methods add *regression* in parallel with *classification* to refine the location quality of the bounding-boxes.

(c) *Post Processing.* As shown in Fig. 3, a single pedestrian may be detected by multiple bounding-boxes after proposal classification, which is the issue of duplicate detections. The technique of non-maximum suppression (NMS) selects the best bounding-box for each object and suppresses other duplicate bounding-boxes. The related methods can be divided into two categories: heuristic-based and learning-based methods. The heuristic-based methods (e.g., greedy NMS, Soft-NMS [7], SGE-NMS [221], and Adaptive NMS [114]) combine the bounding-boxes according to classification scores, where the overlapped bounding-boxes with lower scores are suppressed. The learning-based methods, including Gnet [70] and Relation Network [73], learn a mapping to retain the most accurate bounding-boxes.

Feature extraction is the key component in proposal generation and classification, where the aim is to represent the proposals with discriminative features. A variety of features, ranging from handcrafted [38], [39], [42] to deep features [68], [82], [93], [176], are proposed. Based on the underlying feature extraction scheme, pedestrian detection approaches can be roughly divided into handcrafted features based approaches and deep features based approaches. Most handcrafted features are based on the operations of local difference or sum. One of the most popular handcrafted features is the histogram of oriented gradients (HOG) [38], which captures the changes in local intensity. The fusion of HOG features with other visual cues, such as texture [196] and color [86], has also been investigated. Different to handcrafted features, deep features are typically extracted from the convolutional neural network (CNN). The CNNs learn invariant features through a series of convolution and pooling operations followed by one or more fully-connected (FC) layers. Features from deeper layers are discriminative, whereas the shallow layers contain low-level features with high spatial resolution. Fig. 4 shows visualizations of both handcrafted and deep features on several example images.

3 Single-Spectral Pedestrian Detection

Most vision applications, including pedestrian detection, acquire data using visible-light cameras since they are inexpensive and easily available. As such, most existing pedestrian detection methods [6], [42], [43], [244] employ this kind of

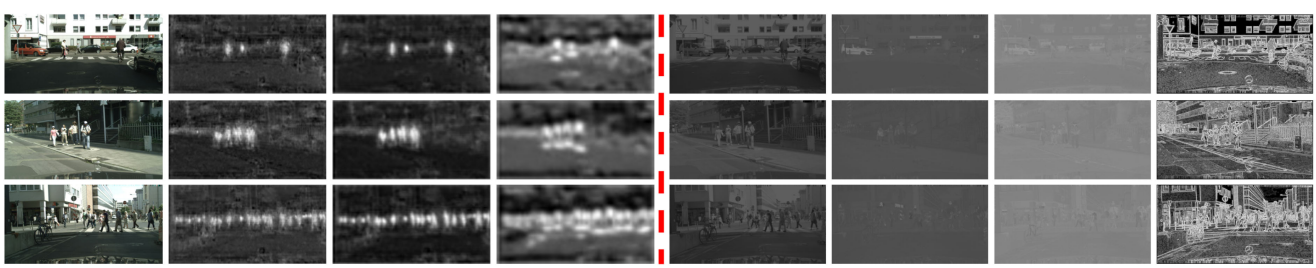

Fig. 4. Visualization of the deep features and handcrafted features used in pedestrian detection. On the left (before the red dotted line): different layers (P2, P3, and P4) of the feature pyramid network [110]. Here, we show feature channels with maximum responses. On the right (after the red dotted line): handcrafted features of three color channels (i.e., LUV) followed by gradient magnitude (last column).

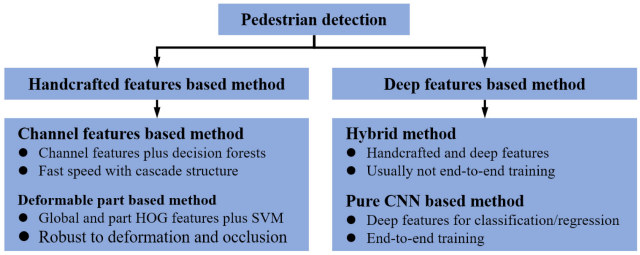

Fig. 5. Two different classes of single-spectral pedestrian detection approaches: handcrafted features based and deep features based methods. We further categorize the handcrafted based methods into channel features based and deformable part model based approaches. Further, deep features based pedestrian detection methods are categorized into hybrid and pure CNN based approaches.

data. We further separate (Fig. 5) these pedestrian detection methods into two main categories: handcrafted features based approaches and deep features based approaches. Moreover, the deep features based approaches are split into pure CNN based methods and hybrid methods. Next, we discuss the handcrafted features methods and then present a summary of deep features based methods.

3.1 Handcrafted Features Based Pedestrian Detection

Before the success of deep convolutional neural networks in computer vision tasks [57], [93], [227], a variety of handcrafted feature descriptors, including SIFT [123], LBP [137], SURF [3], HOG [38], and Haar [186], have been investigated. These handcrafted features usually extract color, texture, or edge information. One of the most widely used handcrafted features for pedestrian detection [38] is histogram of oriented gradients (HOG). Further, most existing handcrafted based approaches either employ channel features [39], [132], [243] or deformable part models [48], [218], [218] for the underlying model learning mechanism. Table 2 summarizes some handcrafted features based methods. Before presenting a detailed introduction of these two kinds of approaches, we first describe their common inference and training steps.

Inference. Given an input image, handcrafted features are first extracted on different proposals (detection windows) generated by sliding the window with a fixed step (e.g., 2). Once the proposals are represented by handcrafted features, they are input to the trained pedestrian detector for prediction (classification). Since real-world pedestrians appear at different scales, input image is first resized at various scales and the detector is then applied on each scale to obtain predictions. Consequently, non-maximum suppression (NMS) is utilized to remove duplicate bounding-boxes (proposals).

Training. The training proposals are generated by sliding-window methods. The proposals that have high overlap with ground-truths are treated as positive samples, otherwise they are treated as negative samples. Given positive and negative samples, the handcrafted features are extracted to represent these samples. Based on the extracted features, shallow classifiers (e.g., boosting or SVM) is used to learn a pedestrian detector to distinguish pedestrians (positive class) and the background (negative class). To improve detection performance, bootstrap technique [186] is commonly adopted to select the hard samples over several training stages, where the hard negative samples at current stage are aggregated to the next one. The detector trained after last stage is used during inference.

3.1.1 Channel Features Based Methods

Most channel features based methods extract a variety of local features from different types of channels (e.g., color and gradient channels) to represent each detection window (proposal) in an image. Then, they employ boosting technique together with decision forests to select a set of most discriminative features, which are used to train a pedestrian detector. One of the earlier detection methods belonging to this category is the popular Viola and Jones (VJ) detector [186]. The VJ method first extracts the candidate Haar features for each proposal and then utilizes cascade AdaBoost [50] to learn the detector. Initially, the VJ method was for face detection. However, compared to face detection, pedestrian detection is more challenging, and the initial VJ framework has been shown to be less effective on this task.

The seminal work of ChnFtrs [42] improves the VJ method for pedestrian detection. It first computes multiple registered image channels and then extracts the local sum features over these image channels. Afterwards, it utilizes the cascade AdaBoost to learn the detector. The ChnFtrs method shows that using ten registered channels (i.e., six gradient histograms, one gradient magnitude, and three LUV color channels) leads to state-of-the-art performance.

Based on the aforementioned registered channels (HOG +LUV) and boosting classifier, several variants of ChnFtrs have been proposed. Some methods [5], [39], [132], [243] focus on extracting better local features from HOG+LUV channels. To reduce computational costs, Dollár *et al.* [39] proposed aggregated channel features (ACF) that aggregate feature values of every block as the candidate features. To avoid a large number of candidate features, Benenson *et al.* [5] selected the local sums of all the squares inside the detection window as the candidate features. To remove the local correlations, Nam *et al.* [132] and Zhou *et al.* [257] proposed to convolve the image channels with a fixed filter bank, learned from the training data, as the candidate features. Zhang *et al.* [243] built a generalized filtered channel framework, where several other detectors (e.g., ChnFtrs [42], SquaresChnFtrs [5], and LDCF [132]) can be seen as the special cases. To avoid the large variety in the types of filter bank, Zhang *et al.* [241] further developed a small set of filter banks inspired by LDCF. Shen *et al.* [172], [173], [174] and Liu *et al.* [120] designed pixel neighborhood differential features for pedestrian detection. Li *et al.* [103] constructed the co-occurrence features in local neighborhoods using a binary pattern. Fu *et al.* [51] exploited the self-similar features based on linear discriminant analysis (LDA). You *et al.* [229] proposed to use several convolutional layers to generate the channel features for pedestrian detection.

Besides the HOG+LUV channel features, other channel features have also been investigated. Costea *et al.* [35], [36] added semantic channel features as additional features. Paisitkriangkrai *et al.* [145], [146] added the low-level visual features (i.e., covariance descriptor and LBP) and spatial pooling to the channel features. Trichet and Bremond [183] introduced LBP-based channels to replace HOG+LUV channels. Zhu *et al.* [262] proposed additional high-level semantic features by

TABLE 2
Summary of 21 Typical Handcrafted Features Based Methods for Pedestrian Detection

Method	Publication	Family	Proposal	Feature	Classifier	Post-proc.	Scale-aware	Part-aware	Context	Description
VJ [187]	IJCV2004	CF	SW	RGB(haar)	boosting	NMS	no	no	no	a robust real-time face detector with Haar features
HOG [38]	CVPR2005	DPM	SW	HOG	SVM	NMS	no	no	no	a novel histogram of gradient feature descriptor
HOG-LBP [197]	ICCV2009	DPM	SW	HOG	SVM	NMS	no	yes	no	an occlusion likelihood map for occlusion handling
ChnFtrs [42]	BMVC2009	CF	SW	Chntrs	boosting	NMS	no	no	no	the simple and effective integral channel features
DPM [49]	PAMI2010	DPM	SW	HOG	SVM	NMS	no	yes	no	deformable part model with six parts and one root
HOF+CSS [189]	CVPR2010	DPM	SW	HOG+CSS	SVM	NMS	no	no	motion info.	a new feature by self-similarity of low-level features
MultiResC [154]	ECCV2010	DPM	SW	HOG	SVM	NMS	yes	yes	ground plane	a multiresolution model based on DPM & HOG
VeryFast [4]	CVPR2012	CF	SW	ChnFtrs	boosting	NMS	yes	no	geometric info.	very fast pedestrian detector running at 135 fps
CrossTalk [40]	ECCV2012	CF	SW	ChnFtrs	boosting	NMS	no	no	no	exploite local correlations for fast cascade design
MT-DPM [219]	CVPR2013	DPM	SW	HOG	SVM	NMS	no	yes	ped./car relation	mapping ped. of various scales to a common space
sDt [155]	CVPR2013	CF	SW	ChnFtrs	SVM	NMS	no	yes	motion info.	remove camera motion and object motion
SquaresChnFtrs [5]	CVPR2013	CF	SW	ChnFtrs	boosting	NMS	no	no	no	use square features to reduce randomness
Franken [130]	ICCV2013	CF	SW	ChnFtrs	boosting	NMS	no	yes	no	a fast training of many occlusion-specific classifiers
ACF [39]	PAMI2014	CF	SW	ChnFtrs	boosting	NMS	no	no	no	aggregate local features by downsampling operation
InformedHaar [240]	CVPR2014	CF	SW	ChnFtrs	boosting	NMS	no	no	no	local ternary features based on pedestrian shape
LDCF [132]	NIPS2014	CF	SW	ChnFtrs	boosting	NMS	no	no	no	remove correlations in local neighborhoods
2Ped [143]	PAMI2015	DPM	SW	HOG	SVM	NMS	no	yes	no	spatial configuration patterns of nearby pedestrians
FCF [244]	CVPR2015	CF	SW	ChnFtrs	boosting	NMS	no	no	no	construct a filtered channel framework
SqatialPooling [147]	PAMI2016	CF	SW	ChnFtrs	boosting	NMS	no	no	no	extract the features based on spatial pooling
SCF [35]	CVPR2016	CF	SW	ChnFtrs	boosting	NMS	no	no	semantic seg.	add segmentation features as additional channels
NNNF [17]	CVPR2016	CF	SW	ChnFtrs	boosting	NMS	no	no	no	non-neighbouring features based on inner attributes

A shallow classifier is used to learn pedestrian detector. 'CF' means channel features based method, 'DPM' means deformable part model based method, and 'SW' means sliding-window strategy.

using a sparse coding algorithm on mid-level image representations. In addition, some methods [154], [188] use motion information to aid pedestrian detection.

Since pedestrian detection only focuses on pedestrian category, some specific pedestrian characteristics can be exploited for feature design. Zhang *et al.* [239] incorporated the prior knowledge that pedestrians usually contain head, upper body, and lower body into Haar-like feature design. Motivated by the human visual system, Zhang *et al.* [240] further developed the center-surround contrast features. Inspired by appearance constancy and shape symmetry of pedestrians, Cao *et al.* [17] designed two non-neighbouring features (i.e., side-inner difference features and shape symmetrical features). These non-neighbouring features have been shown to be complementary to the local features.

Fig. 6 shows some typical handcrafted channel features. The features contains local sum features, local difference features, haar features, non-neighbouring features, etc. From the left to the right, the features have a larger freedom degree in shape and space, and the corresponding methods have better performance in accuracy.

Most methods above strive for improved detection accuracy. In contrast, several other methods focus on improving speed. To reduce computational costs caused by image pyramid, Dollár *et al.* [41] introduced fast feature pyramids, where the channel features at a single scale are used to approximate channel features at nearby scales. Further, Dollár *et al.* [40] designed several fast cascade structures (i.e., soft cascade, excitatory cascade and inhibitory cascade). Pang *et al.* [147] proposed to sample proposals in cascade stages according to the sampling distribution. Benenson *et al.* [4] developed a fast pedestrian detector (called VeryFast) that trains multiple pedestrian detectors and shares the features for different detectors. Rajaram *et al.* [160] trained multiple multi-resolution ACF detectors for fast pedestrian detection.

3.1.2 Deformable Part Based Methods

To better capture the deformation of objects such as pedestrians, the deformable part based model [49] (DPM) was introduced. DPM is one of the most popular handcrafted approaches for detecting both generic objects and pedestrians, which consists of a coarse root model and a set of higher-resolution parts deformation models. The final score is equal to the score of the root model plus the sum over parts of the maximum of the part score minus a deformation cost. In each model, histograms of oriented gradients (HOG) [38] are used to extract the features. By dividing detection window into multiple spatial sub-regions (cells), the gradient histogram features are computed for each cell. Consequently, histograms of each cell are concatenated in a single feature representation to describe detection window. Since HOG features encode the variance in local shape (e.g., edge and gradient structure) very well and the part based model is able to capture the deformations, their combination in the deformable part based models yield promising results in 2006 PASCAL object detection challenge.

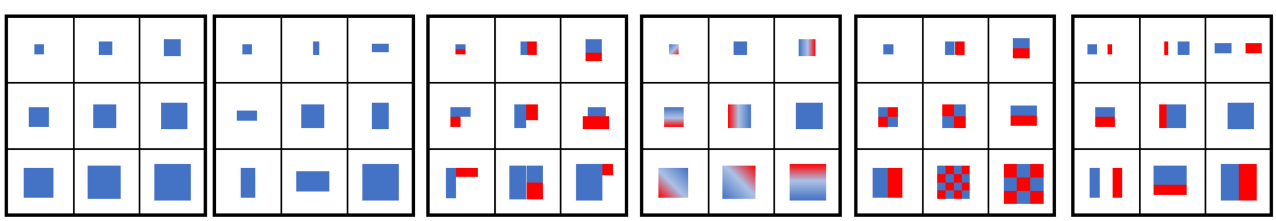

Fig. 6. Some typical features in channel features based methods, including SquaresChntrs [5], Chntrs [42], InformedHaar [239], LDCF [132], Checkererbords [243], and NNNF [17]. The feature freedom degree in shape and space (local or non-local) becomes larger from left to right.

Several variants of the DPM model have been proposed in the literature [48], [80], [153], [218]. Some of these variants focus on further improving pedestrian detection accuracy. Park *et al.* [153] developed multi-resolution feature representations for pedestrians of various scales. Specifically, a deformable and high-resolution part-based model is used for large-sized pedestrian detection, while a rigid and low-resolution template is used for small-sized pedestrian detection. Yan *et al.* [218] proposed to map features from different resolutions to the same subspace using the proposed resolution-aware transformations based on the DPM detector. Ouyang *et al.* [141], [142] improved single-pedestrian detection with the help of multi-pedestrian detection. The multi-pedestrian detector is learned by a mixture of deformable part-based models, where each single pedestrian is also treated as a part. Afterwards, the relationship between single-pedestrian detection and multi-pedestrian detection is modelled to refine pedestrian detection performance. Wan *et al.* [192] incorporated a scale prior and occlusion analysis into deformable part models. Other than striving for improved accuracy, several works focus on improving detection speed. Felzenszwalb *et al.* [48] proposed partial hypotheses to early reject some low scoring samples by using fewer models. As a result, the speed of DPM detector is 20 times faster, without sacrificing detection accuracy. Baek *et al.* [2] developed an additive kernel SVM and BING proposal generation method for fast pedestrian detection.

3.2 Deep Features Based Pedestrian Detection

In recent years, deep convolutional neural networks (CNN) have achieved great success in many computer vision tasks, (e.g., image classification [68], [77], [93], semantic segmentation [121], [126], [150], and object detection [56], [87], [143], [162]). With the success of deep learning in generic object detection, several attempts have been made to apply deep CNN features to pedestrian detection [139], [140], [169]. Table 3 summarizes some deep features based pedestrian detection methods. In this sub-section, we split the related approaches into two categories: hybrid and pure CNN based methods.

3.2.1 Hybrid Pedestrian Detection Methods

As in handcrafted approaches, hybrid methods also have proposal generation and classification steps. According to what CNN features are used for, we divide hybrid approaches into two classes. Some approaches employ CNN features for proposal generation and a shallow classifier for proposal classification (i.e., CNN for proposal generation in Fig. 7a), whereas some other methods use handcrafted methods for proposal generation and CNN features for proposal classification (i.e., CNN for proposal classification in Fig. 7b). These hybrid approaches share some common training and inference protocols, described next. Then, we present a discussion on different hybrid methods.

Inference (1) *CNN for Proposal Generation*. Deep features are first extracted from the entire image. After that, the trained detector slides over the extracted feature map with a fixed step. At each position, the trained detector assigns detection window as either positive class (pedestrian) or negative class based on corresponding features. (2) *CNN for proposal classification*. A handcrafted features based method is used to extract some candidate proposals. Then, the trained CNN classifier classifies these proposals into either the positive or negative class. For both these two kinds of methods, a non-maximum suppression technique is finally used to suppress the duplicate bounding-boxes.

Training (1) *CNN for Proposal Generation*. For positive and negative samples, deep features are extracted from pre-trained CNN. Based on these extracted features, a shallow classifier (e.g., boosting or SVM) along with the bootstrap technique is used to learn the pedestrian detector. The training samples are generated by sliding-window or handcrafted features based methods. (2) *CNN for proposal classification*. First, the handcrafted features based methods are used to generate some candidate proposals. Based on these candidate proposals, a CNN with the softmax layer is trained in an end-to-end fashion (both proposal generation and classification) on the specific pedestrian dataset.

Some pedestrian detection approaches employ CNNs for proposal generation and a shallow classifier for proposal classification. Yang *et al.* [220] proposed to replace the handcrafted filtered channel features (FCF) [243] with convolutional channel features (CCF), where each pixel in the last convolutional layer is used as a single feature. Hu *et al.* [76] trained an ensemble of boosted decision forests based on the features from the different layers of a CNN. Zhang *et al.* [236] and Tesemaa *et al.* [180] utilized the region proposal network (RPN) as an initial pedestrian detector and further trained a shallow classifier with deep features to refine detection results. Li *et al.* [97] proposed to extract multi-resolution deep features from different convolutional networks to learn a pedestrian detector. Sheng *et al.* [175] integrated deep semantic segmentation features and shallow handcrafted channel features into a filtered channel framework. Tesema *et al.* [180] proposed to pool both handcrafted features and deep features to learn pedestrian detector with decision forests. Wang *et al.* [189] developed a multi-scale region proposal network to deal with a variance in scale and integrated a decision forest for classification.

Several other pedestrian detection works treat CNNs as a deep classifier to classify the candidate proposals. Hosang *et al.* [71] provided a deep analysis on the effectiveness of CNNs for pedestrian detection. Based on their careful design, the simple CNNs were shown to achieve promising results for pedestrian detection. Tian *et al.* [181] trained the multiple part detectors and then trained a linear SVM to combine the scores of part detectors. Ribeiro *et al.* [165] trained multiple deep networks with different inputs (e.g., color and segmentation images) to refine the results of ACF detector [39]. Ouyang *et al.* [140], [144] built a unifying deep learning model to join different tasks (i.e., feature extraction, deformation handling, occlusion handling, and classification). Luo *et al.* [127] proposed to automatically learn hierarchical features, salience maps, and mixture representations of different body parts by a Switchable Restricted Boltzmann Machine. Jung *et al.* [83] developed a guiding network to assist the training of pedestrian detector.

Besides the aforementioned approaches that strive for higher accuracy, other methods aim to improve detection speed. Cai *et al.* [12] designed a complexity aware cascade strategy (CompACT) to balance accuracy and computational

TABLE 3
Summary of 45 Typical Deep Features Based Methods for Pedestrian Detection

Method	Publication	Family	Proposal	Feature	Classifier	Post-proc.	Scale-aware	Part-aware	Context	Description
DDN [139]	CVPR2012	Hybrid	SW	HOG	softmax	NMS	no	yes	no	first deep model for pedestrian detection
UMS [170]	CVPR2013	P-CNN	SW	CNN	softmax	NMS	no	no	ms fusion	one of the earliest deep pedestrian detectors
UDN [140]	ICCV2013	P-CNN	SW	CNN	softmax	NMS	no	yes	no	join different components by a deep network
SDN [127]	CVPR2014	Hybrid	HOG	CNN	boosting	NMS	no	yes	no	model mixture of visual variations by networks
ConvNet [71]	CVPR2015	Hybrid	ACF	CNN	softmax	NMS	no	no	no	a state-of-the-art performance using convnets
TA-CNN [183]	CVPR2015	Hybrid	ACF	CNN	boosting	NMS	no	no	attributes	join detection with multiple semantic tasks
DeepCascades [1]	BMVC2015	Hybrid	VeryFast	CNN	softmax	NMS	no	no	no	one of first real-time and very accurate detector
CCF [221]	ICCV2015	Hybrid	ACF	CNN	boosting	NMS	no	no	no	extend FCF [244] to conv. channel features
DeepParts [182]	ICCV2015	Hybrid	ACF	CNN	SVM	NMS	no	yes	no	handle occlusion with deep part pool
CompACT-Deep [12]	ICCV2015	Hybrid	SW	ChnFtrs+CNN	boosting	NMS	no	no	no	a complexity-aware cascade training structure
EEPD [180]	CVPR2016	P-CNN	-	CNN	LSTM	no	no	no	no	end-to-end approach directly predicting objects
MS-CNN [11]	ECCV2016	P-CNN	RPN	FPN	softmax	NMS	yes	no	contextual RoI	multi-scale features for scale-ware detection
RPN+BF [237]	ECCV2016	Hybrid	RPN	CNN	softmax	NMS	no	no	no	analyse limitations of FR-CNN for pedestrians
MCF [16]	TIP2017	Hybrid	SW	ChnFtrs+CNN	boosting	NMS	no	no	no	construct a multi-layer channel framework
SubCNN [211]	WACV2017	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	joint detection and subcategory classification
F-DNN [44]	WACV2017	P-CNN	SSD	CNNs	softmax	NMS	no	no	segmentation	a deep fusion of multiple networks
PGAN [100]	CVPR2017	P-CNN	RPN	R-CNN	softmax	NMS	yes	no	no	narrow feature differences by GAN
HyperLearner [129]	CVPR2017	P-CNN	RPN	R-CNN	softmax	NMS	no	no	segmentation	learning extra features by multi-task learning
Adapted FR-CNN [245]	CVPR2017	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	improved Faster R-CNN for pedestrians
JL-TopS [260]	ICCV2017	Hybrid	RPN	CNN	boosting	NMS	no	yes	no	joint part detectors by multi-label learning
SDS-RCNN [10]	ICCV2017	P-CNN	RPN	R-CNN	softmax	NMS	no	no	segmentation	joint semantic segmentation and detection
PCN [194]	BMVC2017	P-CNN	RPN	R-CNN	softmax	NMS	no	yes	contextual RoI	use body parts semantic and context information
CFM [76]	TCSVT2018	Hybrid	SW	CNN	boosting	NMS	no	no	no	ensemble of boosted models by inner features
SAF-RCNN [99]	TMM2018	Hybrid	ACF	R-CNN	softmax	NMS	yes	no	no	two built-in sub-networks for different scales
SCNN [23]	PAMI2018	Hybrid	ACF	CNN	softmax	NMS	no	no	no	subcategory-aware network for intra-class variance
RepulsionLoss [200]	CVPR2018	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	novel repulsion loss for box regression
OHNH [135]	CVPR2018	P-CNN	-	SSD	softmax	NMS	yes	no	no	part-aware score added in single-shot detector
FR-CNN ATT [249]	CVPR2018	P-CNN	RPN	R-CNN	softmax	NMS	no	yes	no	channel attention mechanism for occlusion
Bi-Box [261]	ECCV2018	P-CNN	RPN	R-CNN	softmax	NMS	no	yes	no	two RoIs for fully/visible-body detections
GDFL [106]	ECCV2018	P-CNN	-	SSD	softmax	NMS	yes	no	ms fusion	encode fine-grained attention masks
OR-CNN [246]	ECCV2018	P-CNN	RPN	R-CNN	softmax	NMS	no	yes	no	part occlusion-aware RoI pooling layer
TTL [178]	ECCV2018	P-CNN	-	CNN	softmax	NMS	no	no	ms fusion	use topological somatic line for detection
ALFNet [118]	ECCV2018	P-CNN	-	SSD	softmax	NMS	yes	no	no	stack a series of predictors on SSD
CSP [119]	CVPR2019	P-CNN	-	CNN	softmax	NMS	no	no	ms fusion	one of first anchor-free pedestrian detector
Adaptive-NMS [114]	CVPR2019	P-CNN	-/RPN	SSD/FPN	softmax	Adapt. NMS	yes	no	no	dynamic NMS threshold based on target density
AR-Ped [9]	CVPR2019	P-CNN	AR-RPN	FPN	softmax	NMS	yes	no	ms fusion	multi-phase autoregressive module for RPN
FRCN+A+DT [259]	ICCV2019	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	narrow the occluded/unoccluded features
MGAN [152]	ICCV2019	P-CNN	RPN	R-CNN	softmax	NMS	no	yes	no	mask-guided attention for RoI regions
PedHunter [29]	AAAI2020	P-CNN	RPN	FPN	softmax	NMS	yes	yes	no	mask-guided module encoding head information
JointDet [30]	AAAI2020	P-CNN	RPN	R-CNN	softmax	RDM	no	yes	no	head-body relationship discriminating module
PRNet [179]	ECCV2020	P-CNN	-	SSD	softmax	NMS	no	yes	ms fusion	a novel progressive refinement network
Case [214]	ECCV2020	P-CNN	RPN	R-CNN	softmax	CaSe-NMS	no	no	no	a count-weighted detection loss
PBM [78]	CVPR2020	P-CNN	RPN	R-CNN	softmax	R^2NMS	no	yes	no	a novel NMS based on a paired-box model
TFAN [207]	CVPR2020	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	a tube feature aggregation network for occlusion
CrowdDetection [32]	CVPR2020	P-CNN	RPN	R-CNN	softmax	Set NMS	no	no	no	predict multiple correlated instances per proposal

These methods are typically built on convolutional neural networks. 'P-CNN' means the pure CNN method, 'Hybrid' means the hybrid method, and 'SW' means the sliding-window strategy. 'R-CNN' means the feature extraction fashion in R-CNN series, including R-CNN, Fast R-CNN, and Faster R-CNN.

complexity. Specifically, CompACT uses the features of lower computational complexity at early stages and the features of higher computational complexity at later stages. Cao *et al.* [16] designed multi-layer channel features (MCF), where the handcrafted channels and each layer of CNNs are integrated together. Based on the multi-layer feature channels, a multi-stage cascade detector is learned. MCF not only makes full use of features of different layers, but also efficiently rejects many samples at a lower computational cost. Angelova *et al.* [1] proposed to cascade the handcrafted detector (i.e., VeryFast [4]) and multiple deep networks for faster pedestrian detection. Jiang *et al.* [81] proposed to share deep features for multi-scale pedestrian detection.

3.2.2 Pure CNN Based Pedestrian Detection Methods

The success and popularity of Faster R-CNN [164] for generic object detection prompted the construction of pure CNN based pedestrian detection approaches, where CNNs are used for both proposal generation and classification. Fig. 7c shows the architecture of a pure CNN based pedestrian detector. Initially, the direct usage of Faster R-CNN for pedestrian detection resulted in below-expected performance. Zhang *et al.* [244] introduced several modifications (e.g., anchor scale and ignored region handling) to Faster R-CNN for improved pedestrian detection. Compared with hybrid methods, the pure CNN based approaches are more effective and simpler. Moreover, they are typically trained in an end-to-end fashion. We first describe some common training and inference protocols. Afterwards, we present a discussion on different pure CNN based methods.

Inference. Given a test image, deep features (of the entire image) are first extracted using a CNN. Some candidate proposals are first generated based on the default anchors and then classified by the corresponding features with a softmax layer and regressed to obtain a more accurate location.

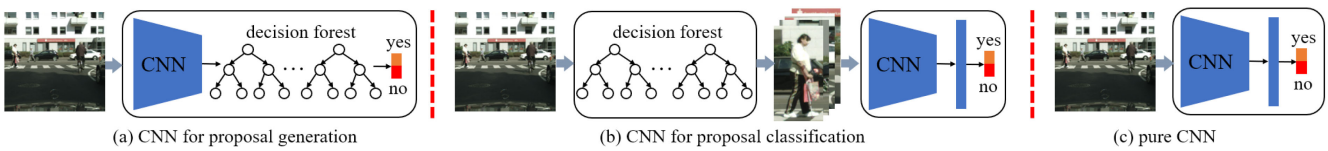

Fig. 7. Architectures of deep features based methods. (a) and (b) show two techniques in hybrid methods, while (c) is pure CNN based method. In (a), CNN extracts deep features for proposal generation and a shallow classifier is used for proposal classification. In (b), a handcrafted features based method is used for proposal generation and CNN is for proposal classification. In (c), pure CNN is used for both proposal generation and classification in an end-to-end fashion.

Consequently, a non-maximum suppression (NMS) technique is used to suppress duplicate bounding-boxes.

Training. Given a training image, deep features (of the entire image) are extracted using a CNN. The anchors are set as each position of feature maps and assigned as positive and negative samples. Based on the gradient back-propagation algorithm, CNNs are updated at each iteration. The learning rate, weight decay, and batch size are set according to the specific pedestrian detection dataset.

In recent years, a variety of pure CNN based pedestrian detection methods have been introduced in the literature. Next, we present a summary of these approaches.

Scale-Aware Methods. Scale-aware methods generally use the different in-network layers (or sub-networks) to detect objects at different scales. Some of these methods extract RoI features of proposals from different layers (*called scale-aware RoI*). Yang *et al.* [223] and Liu *et al.* [115] extracted the RoI features of proposals according to their scales. If the object has a smaller scale, the RoI features from the earlier layer are extracted. If the object has a larger scale, the RoI features from the later layers are extracted. Zhu *et al.* [264] flexibly chose the RoI features for regression and combined the RoI features from multiple layers for classification. SAF-RCNN [99] integrates the scores of large-scale and small-scale sub-networks according to the proposal scales. Some methods generate candidate proposals from different layers (*called scale-aware RPN*). To make the receptive fields match the objects of different scales, Cai *et al.* [11] proposed to extract proposals from multiple in-network layers. Specifically, the lower layers with smaller receptive fields are employed for small objects, whereas the later layers with larger receptive fields are utilized for large objects. To enhance the feature semantics, Lin *et al.* [110], [111] adopted a top-down structure to integrate the features from deep layer with the features from shallow layer (called FPN). Hu *et al.* [74] further modified FPN by reducing the convolutional stride from 2 to 1 at earlier layers to retain more information for small-scale pedestrian detection.

Part-Based Methods. Features from the local part of an object play an important role in capturing occluded or deformable pedestrians. Several methods have investigated the integration of part-based information. Xu *et al.* [217] proposed to first detect the key-points of each proposal and then generate six parts based on these key-points. Afterwards, they combined the features of these parts together. Zhao *et al.* [256] introduced two branches for holistic and part predictions and built a tree-structured module to integrate them. Zhang *et al.* [245] proposed to combine the features of different parts for classification and regression.

Several recent pedestrian detection methods use the visible-body information of a pedestrian. Zhou *et al.* [260] trained a deep network with two output branches to detect full body and visible part. The results of two branches are fused to obtain improved pedestrian detection. Pang *et al.* [151] developed a novel mask-guided attention network to enhance the features of visible regions. Some methods use the head information to aid pedestrian detection. Chi *et al.* [30] designed a joint network for head and pedestrian detection with relationship discriminating module for detection in crowd. Zhang *et al.* [235] proposed double anchor region proposal networks to respectively detect human heads and bodies and a joint NMS to combine the detection results. Lin *et al.* [108] built two-branch networks for head-shoulder and full body predictions and introduced an adaptive fusion mechanism. Lu *et al.* [125] proposed semantic head detection in parallel with a body branch.

Attention-Based Methods. These methods aim to enhance the features of pedestrians while suppressing the features of background. According to the underlying attention mechanism, we divide the related approaches into *self-attention methods* and *semantic-guided attention methods*. Self-attention methods use the attention mechanism to relate different positions of features. Zhang *et al.* [248] observed that different channels represent different parts of an object and utilized a channel-wise attention [75] for occluded pedestrian detection. Zou *et al.* [266] proposed a spatial attention module to up-weight the features of visible part based on class activation mapping technique. Chen *et al.* [27] proposed the competitive attention to fuse the features from different convolutional layers.

Semantic-guided attention approaches aim at joining some high-level semantic task and pedestrian detection and can be also treated as multi-task methods. Brazil *et al.* [10] proposed a multi-task infusion framework for joint pedestrian detection and semantic segmentation for both proposal generation and classification. Lin *et al.* [106], [107] designed a scale-aware attention module to make the detector better focus on the regions of pedestrians. Gajjar *et al.* [52] and Yun *et al.* [231] proposed to use the visual saliency task as a pre-processing step to better focus on the regions of a pedestrian.

Feature-Fused Methods. These methods aim to capture the useful contextual and semantic information by multi-scale feature fusion. Ren *et al.* [163] built a recurrent rolling convolution architecture to gradually aggregate contextual information from the different layers. Based on MS-CNN [11], Jung *et al.* [84] further combined the features of consecutive layers. In contrast, Chu *et al.* [31] combined the features from all different layers together to generate high-level features. Liu *et al.* [116] proposed a gated feature extraction module by adaptively fusing multi-layer features. Shang *et al.* [170] introduced a complementary sub-network to generate the high-resolution feature map for small-scale object detection. Zhang *et al.* [233] concatenated the RoI features from different layers along with the global context. Zhang *et al.* [232] proposed a context feature embedding with a standard convolution and a deformable convolution. Fei *et al.* [47] developed a new pixel-level context embedding module by integrating multi-cue context into a deep CNN feature hierarchy. Wang *et al.* [194] proposed a local competition mechanism (maxout) for adaptive context fusion. Cao *et al.* [15] embedded the large-kernel convolution into feature pyramid structure to exploit contextual information. Wu *et al.* [206] proposed to adaptively fuse the features of current frame and nearby frames.

Cascade-Based Methods. To improve localization quality, cascade structure has been widely used in generic object detection [13], [14], [246]. Recently, some methods have adopted cascade structure for pedestrian detection. Liu *et al.* [118] stacked multiple head predictors for multi-stage regressions. Brazil *et al.* [9] designed a multi-phase auto-regressive module, where each module is trained using increasingly precise labeling policies. Zhang *et al.* [250] proposed to detect the

low resolution and occluded objects again at a finer scale by mimicking the process of humans. Ujjwal *et al.* [185] first utilized semantic segmentation to select a small set of anchors and then re-pooled the features for classification and regression. Du *et al.* [44] proposed to fuse the detection scores of multiple networks by a cascade soft-region rejection strategy. Hasan *et al.* [66] combined Cascade R-CNN [13] and HRNet [191] to achieve improved pedestrian detection performance. Song *et al.* [178] proposed to divide pedestrian detection into three-phase steps: visible part prediction, anchor calibration, and full-body prediction.

Anchor-Free Methods. These methods directly predict the score and pedestrian location or shape at each position. Compared with the anchor-based methods, the anchor-free methods avoid the handcrafted design with respect to the scale and aspect ratio of anchors, thereby having a simpler design and good generalization ability on different datasets. Song *et al.* [177] propose to localize pedestrians by the somatic topological line. Liu *et al.* [119] proposed to predict the center point and the height of the pedestrian based on the high-level semantic feature maps. Zhang *et al.* [234] treated each positive instance as a feature vector to encode both density and diversity information simultaneously.

Instead of focusing on the network design, some methods, discussed next, focus on data augmentation, loss learning, post processing, and multi-task learning.

Data-Augmentation Based Methods. These methods aim to improve detection performance. Some methods focus on generating more pedestrians or images (*data generation*). Based on the prior knowledge of camera parameters, Hattori *et al.* [67] generated a variety of geometrically accurate images of synthetic pedestrians. Vobecky *et al.* [187] used GAN [59] to generate people images in a required pose. Wu *et al.* [205] developed a multi-modal cascaded generative adversarial network with U-net structure to generate pedestrian data. Chen *et al.* [26] transformed real pedestrians from the same dataset into different shapes using the shape-guided deformation. Some methods focus on making full use of current data (*data processing*). To improve occluded pedestrian detection, Chi *et al.* [29] added some occlusions to pedestrians. To generate better positive samples, Lu *et al.* selected samples based on visible intersection-over-union. Zhao *et al.* [255] introduced a strict matching metric for training sample generation by considering the alignments of different parts simultaneously. Wei *et al.* [202] proposed to use soft-NMS [7] to select some occluded samples for training. Luo *et al.* [128] proposed to use multi-modal data, including bird view map, depth and corpus information, for pedestrian localization, scale prediction and classification.

Loss-Driven Methods. These methods either use new functions or add extra loss functions for pedestrian detection. Wang *et al.* [199] proposed two types of repulsion loss (i.e., RepGT loss and RepBox loss) for crowded pedestrian detection. The RepGT loss penalizes the predicted bounding-box near other objects, whereas the RepBox loss makes the predicted bounding-box farther away from other predicted bounding-boxes, in case of belonging to different objects. Wu *et al.* [204] developed a weighted loss that emphasizes challenging samples. Xiang *et al.* [210] explicitly employed the loss of sub-category classification for pedestrian detection. Some methods use the loss function to narrow the feature gap between different samples. Li *et al.* [100] developed an architecture that internally lifts representations of small objects to that of large objects. Zhou *et al.* [258] proposed a discriminative feature transformation to make the pedestrian features approach the feature center of non-occluded pedestrians. Li *et al.* [102] proposed to narrow the feature gap between the small network and the large network for efficient detection. Chen *et al.* [22] performed multi-stage distillation to learn the light-weight network for acceleration. Li *et al.* [104] transformed the LR feature space into a new LR classification space using an optimal Mahanalobis metric. Xie *et al.* [213] proposed to assign a large weight to the proposal in crowded scene.

Post-Processing Methods. Some methods improve NMS to better combine detection results. Liu *et al.* [114] applied a dynamic suppression threshold based on the target density. Yang *et al.* [221] developed bounding-box-level Semantics-Geometry Embedding (SGE) to distinguish two heavily-overlapping boxes. Huang *et al.* [78] proposed R^2NMS that uses the IoU between visible regions to determine whether or not two full-body boxes overlap. Stewart *et al.* [179] built an end-to-end network to directly predict the objects without post-processing. Wang *et al.* [197] proposed to set the confidence threshold by investigating the relationship between the scores and scales of pedestrians. Zhang *et al.* [249] designed an accurate localization-quality estimation module to refine classification scores. Yang *et al.* [222] developed a Kalman filter-based convolutional network to remove some false positives for pedestrian detection in videos.

Multi-Task Methods. Some methods utilize semantic information to aid pedestrian detection. Mao *et al.* [129] investigated the impact of aggregating additional features (e.g., segmentation, heatmap, disparity, and optical flow) for pedestrian detection by using a multi-task learning network. Wang *et al.* [198] proposed joint semantic segmentation and pedestrian detection to better distinguish the background and foreground. Kishore *et al.* [91] and Zhao *et al.* [254] proposed to join occluded pedestrian detection and pose estimation in cascade structure. Han *et al.* [65] proposed to join pedestrian detection and person search.

Others. Most methods above focus on pedestrian detection in color images. Recently, some methods have been proposed for thermal images or fish-eye images. Guo *et al.* [63] designed a domain adaptation component to use the abundant color images associated with pedestrian annotations in thermal domain. Ghose *et al.* [55] proposed to use saliency to augment pedestrian detector in the thermal domain. Kieu *et al.* [89] designed a task-conditioned architecture to adapt pedestrian detector to the thermal domain. Qian *et al.* [157] introduced a projective model to transform normal images into fish-eye images and designed an oriented spatial transformer network to rectify warped pedestrian features for better recognition. Peng *et al.* [156] proposed a new cost function for training object detectors on fish-eye images. Li *et al.* [98] proposed to use the depth-wise separable convolution, linear bottleneck, and multi-scale feature fusion for pedestrian detection in hazy weather. To avoid annotating a large number of pedestrians, Wu *et al.* [208], [209] developed a semi-supervised approach to train deep convolutional networks on partially labeled data.

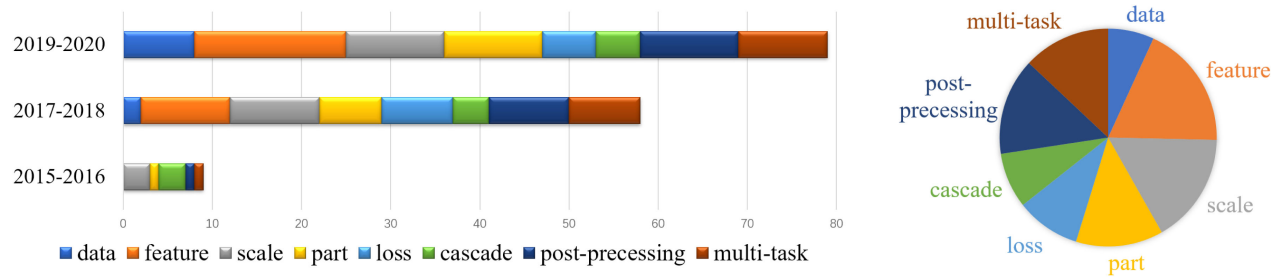

Fig. 8. Statistics of deep features based methods. The left part shows the change in the number of methods belonging different classes. Here, feature in legend indicates the union of feature-fused and attention-based methods which both aim to improve feature description ability. We ignore the anchor-free methods and others due to the limited number of these methods. The right part shows the percentage of different classes.

Fig. 8 provides the statistics of deep features based methods from 2015 to 2020. The left part shows the change in the number of methods belonging to different classes. It can be seen that data-augmentation based methods, feature enhanced methods, and part-based methods have a large increment in past two years. The right part shows the percentage of all the methods over the past six years. The feature-enhanced, post-processing, scale-aware, part-based methods are the dominant approaches. Among these methods, post-processing and part methods usually focus on occluded pedestrian detection, while scale-aware and feature-enhanced methods mainly deal with scale variance problem. Thus, most recent methods still focus on solving the problems of occlusion and scale-variance.

4 Multispectral Pedestrian Detection

In Section 3, most reviewed methods focus on detecting pedestrians in single-spectral images (e.g., color image). However, single-spectral pedestrian detection is not very robust to illumination variations. For instance, the color camera is ineffective at acquiring useful information at night. Therefore, multispectral pedestrian detection [25] becomes important for self-driving and video surveillance, where the color and thermal images provide complementary visual information. Table 4 summarizes some typical methods for multispectral pedestrian detection.

Some methods explore how to deep fusion of multispectral images, including input fusion, feature fusion, and decision fusion. Liu *et al.* [113] exploited four kinds of fusions at different stages (called low-level fusion, middle-level fusion, high-level fusion, and score fusion). It is found that middle-level fusion (Highway Fusion) achieves the best detection performance. To take advantage of the pre-trained model on ImageNet [167], Konig *et al.* [92] added one 1×1 convolutional layer to reduce the number of fused channels to the same number of input channel of VGG. Further, boosted decision trees were used as in [236]. To better combine the features from different modalities, Zhang *et al.* [237] introduced a cross-modality interactive attention module to exploit the complementary nature of different modalities. Guan *et al.* [61] and Li *et al.* [96] explored an illumination-aware mechanism by using the predicted illumination value to re-weight the results of the day and night sub-networks. To solve the modality imbalance problem, Zhou *et al.* [261] proposed a single-stage detector that contains a differential modality aware fusion module and an illumination aware feature alignment module.

Besides, to solve the position mismatch problem between color image and thermal image, Zhang *et al.* [238] proposed to capture the position shift and align the region features. Based on the aligned features, a multi-modal re-weighted module was further introduced to generate reliable features. To generate more diversified proposals and learn better features, Li *et al.* [95] added two head-networks for joint semantic segmentation and pedestrian detection to the color image branch and thermal image branch during training.

For unsupervised domain adaptation, Guan *et al.* [62] proposed to iteratively generate training labels and update the network parameters in the target domain. To automatically transfer a detector from a visible domain to a new multispectral domain, Cao *et al.* [19] presented an auto-annotation framework to label pedestrian instances in visible and thermal channels by leveraging the complementary information of multispectral data. Xu *et al.* [216] designed a cross-modality learning framework to model the relations between color and infrared images. Afterwards, features extracted from the cross-modality network are fused with the features extracted from the traditional detection network.

5 Dataset and Evaluation

5.1 Earlier Pedestrian Datasets

The earlier datasets in Table 5 (top) are relatively small and mainly used by handcrafted features based methods, including MIT [152], INRIA [38], ETH [46],TUD-Brussels [203], Daimler [45]. Appendix A, (available online), provides a detailed description.

5.2 Modern Pedestrian Datasets

In the recent era of deep learning, performance on the early datasets has become saturated. As such, the research community has dedicated significant efforts towards building new modern pedestrian datasets. Modern pedestrian datasets in Table 5 (medium) are significantly larger and aim for a more standard evaluation. Specifically, the number of images and pedestrians is usually over 10 times larger, and a more unifying training and test data are provided.

Caltech[1] [43] is one of most complete benchmarks for pedestrian detection. This dataset contains 11 video sets taken from the urban, where the first 6 video sets are used for training and the remaining 5 video sets are used for

1. www.vision.caltech.edu/Image_Datasets/CaltechPedestrians

testing. Generally, the training images are captured by every 3*rd* frame and the test images are captured by every 30th frame. In addition to the full-body bounding-box annotations, the visible-body bounding-box annotations are also provided.

KITTI[2] [53] is a challenging computer vision benchmark, including multiple different vision tasks. For object detection (car detection, pedestrian detection, and cyclist detection), there are 7,481 training images and 7,518 test images. The resolution of images is about 1,240 $\times$ 376 pixels.

CityPersons[3] [244] is a diverse pedestrian dataset built on the Cityscapes dataset [33]. There are 2,975 training images, 500 validation images, and 1,575 test images. Compared with the early pedestrian dataset, CityPersons dataset is richer in diversity (i.e., different cities, different seasons, various weather conditions, and more persons per image).

CrowdHuman[4] [171] is a recently collected dataset for crowded human detection. It contains 15,000 training images, 4,370 validation images, and 5,000 test images. There are about 23 persons per image. The bounding-box annotations of full-body, visible-body, and head are provided.

EuroCity persons[5] [8] is a large-scale dataset of urban scenes captured at both daytime and nighttime in multiple European cites. It contains three different person categories (i.e., pedestrians, cyclists, and other riders). There are 47,337 images in total at a resolution of 1,920 $\times$ 1,080 pixels.

NightOwls[6] [133] is a pedestrian dataset recorded at nighttime, which aims to promote progress in pedestrian detection at night. There are about 128k training images, 51k validation images, and 103k test images.

WIDER Pedestrians[7] focuses on detecting pedestrians and cyclists for surveillance and car-driving. It contains 96,500 images with 307,183 annotations. There are 8,240 images (58,190 annotations) in surveillance scenes and 88,260 images (248,993 annotations) in car-driving scenes.

WiderPerson[8] [247] focuses on pedestrian detection in the wild under multiple different scenes and is not limited to traffic scenes. The dataset contains 13,382 images with 400k annotations with various kinds of occlusions, which are collected from the website using 50 keywords.

TJU-Pedestrian[9] [148] is a diverse high-resolution dataset, including two sets of TJU-Pedestrian-Traffic and TJU-Pedestrian-Campus. TJU-Pedestrian-Traffic has 20,388 images with 43,618 annotations, while TJU-Pedestrian-Campus has 55,088 images with 329,623 annotations.

5.3 Multispectral Pedestrian Datasets

Multispectral pedestrian datasets in Table 5 (bottom) are built on visible-light and thermal cameras. Thus, more useful information is provided for robust pedestrian detection.

KAIST[10] [79] is a large-scale multispectral pedestrian dataset recorded by a specitically designed imaging hardware device, which can capture the aligned color and thermal image pairs. This dataset has 95,328 image pairs with 103,128 dense annotations and 1,182 unique pedestrians.

CVC-14[11] [58] is a dataset of multimodal (FIR and visible) video sequences recorded during daytime and nighttime. There are 7,085 images for training and 1,433 images for testing. Different from KAIST dataset, the alignment is mainly achieved through post-processing, since the resolution and the field-of-view of two sensors are different.

5.4 Evaluation Metrics

Three evaluation metrics, i.e., log-average miss rate (MR), average precision (AP), and jaccard index (JI), are typically used in pedestrian detection. Among these, MR and AP are widely used, whereas JI (see Appendix B, available in the online supplemental material) was introduced for crowded pedestrian detection. Before discussing these evaluation metrics in detail, we first explain how to determine if a detected bounding-box is a true positive or a false positive. The overlap between a detected bounding-box B_d and a ground-truth B_g can be calculated as $O = (B_d \cup B_g)/(B_d \cap B_g)$. If the overlap O is larger than a threshold of α, the detected bounding-box is a potential matching with the ground-truth. Since a detected bounding-box might match multiple ground-truths, a greedy matching algorithm in Appendix C, available in the online supplemental material, splits the detection results into true positives and false positives along with false negatives (missed positives). Based on true positives, false positives, and false negatives on the whole test set, log-average miss rate and average precision can be calculated for performance comparison.

Log-Average Miss Rate. Given a threshold of detection confidence score, miss rate (MR) can be calculated by the number of true positives (N_{tp}) and the number of ground-truths (N_g) as $MR = 1 - N_{tp}/N_g$, and false positives per image (FPPI) can be calculated by dividing false positives by the number of images. By varying confidence threshold, miss rates against false positives per image (FPPI) can be plotted in log-space. Finally, the log-average MR is calculated by averaging miss rates under 11 FPPI equally spaced in $[10^{-2}{:}10^{0}]$. A lower MR reflects a better performance.

The Caltech [43] and CityPersons [244] datasets evaluate MR under different sets, e.g., **R**, **RS**, **HO**, **R+HO**, and **A**. The **R** set comprises pedestrians over 50 pixels in height with less than 35 percent occlusion. The **RS** set comprises pedestrians over 50 pixels and under 75 pixels with less than 0.35 occlusion. The **HO** set comprises pedestrians over 50 pixels in height with 35-80 percent occlusion. The **R+HO** set is the union set of **R** and **HO**. The **A** set contains the pedestrians over 20 pixels in height with less than 80 percent occlusion.

Average Precision. Given a threshold of confidence score, recall (R) can be calculated by the number of true positives (N_{tp}) and the number of ground-truths (N_g) as $R = N_{tp}/N_g$. The precision (P) can be calculated by the number of true positives (N_{tp}) and the number of all detected bounding-boxes (N_d) as $P = N_{tp}/N_d$. By varying the threshold of confidence score, precision against recall can be plotted as a curve. Based on the precision-recall curve, the average

2. www.cvlibs.net/datasets/kitti
3. https://github.com/cvgroup-njust/CityPersons
4. https://www.crowdhuman.org/
5. https://eurocity-dataset.tudelft.nl
6. https://www.nightowls-dataset.org
7. https://wider-challenge.org/
8. http://www.cbsr.ia.ac.cn/users/sfzhang/WiderPerson
9. https://github.com/tjubiit/TJU-DHD
10. https://soonminhwang.github.io/rgbt-ped-detection
11. http://adas.cvc.uab.es/elektra/datasets

TABLE 4
Summary of 12 Typical Methods for Multispectral Pedestrian Detection

Method	Publication	Family	Proposal	Feature	Classifier	Post-proc.	Scale-aware	Part-aware	Context	Description
ACF-C-T [79]	CVPR2015	CF	SW	ChnFtrs	boosting	NMS	no	no	no	extended ACF with the thermal channel
Halfway [113]	BMVC2016	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	fused channel features at middle-level layers
CMT-CNN [217]	CVPR2017	Hybrid	ACF	R-CNN	softmax	NMS	no	no	no	cross-domain features by cross-modality learning
MRFC [36]	CVPR2017	CF	SW	ChnFtrs	boosting	NMS	no	no	2D/3D	multimodal multiresolution channel features
Fusion RPN [92]	CVPRW2017	Hybrid	RPN	CNN	boosting	NMS	no	no	no	use pre-trained convnet by network in network
APF [156]	PR2018	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	channel weighting & probability fusion
MSDS-RCNN [95]	BMVC2018	P-CNN	RPN	R-CNN	softmax	NMS	no	no	segmentation	joint detection and semantic segmentation tasks
TS-RPN [19]	IF2019	P-CNN	TS-RPN	CNN	softmax	NMS	no	no	no	adapt visible detector to multispectral domain
IAFR-CNN [96]	PR2019	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	adaptively merge results by illumination value
HMFFN [20]	ISPRS2019	P-CNN	RPN	R-CNN	softmax	NMS	no	no	no	box-level segmentation supervised learning
AR-CNN [239]	ICCV2019	P-CNN	RPN	R-CNN	softmax	NMS	no	no	Contextual RoI	first work that tackles position shift problem
MBNet [262]	ECCV2020	P-CNN	-	SSD	softmax	NMS	yes	no	no	designing a modality balance network

'P-CNN' means the pure CNN method, 'Hybrid' means the hybrid method.

precision (AP) is calculated by averaging precisions under 41 recalls equally spaced in [0:1]. A higher AP reflects better performance.

The KITTI [53] dataset evaluates AP under three sets, i.e., `Easy`, `Medium`, and `Hard`. The `Easy` set includes pedestrians over 40 pixels in height with no occlusion. The `Medium` set includes pedestrians over 25 pixels in height with less than part occlusion. The `Hard` set includes pedestrians over 25 pixels in height with less than heavy occlusion.

5.5 State-of-the-Art Comparison and Analysis

Here, we provide a comparison and discussion of several state-of-the-art methods on four widely used datasets (i.e., Caltech [43], KITTI [53], CityPersons [244], and KAIST [79]). Besides these state-of-the-art comparison, we provide more comprehensive analysis about pedestrian detection.

Table 6 compares some methods on Caltech pedestrian dataset [43]. Four subsets of **R**, **HO**, **R+HO**, and **A** are used for performance evaluation. Among these methods, only 5 approaches (i.e., ACF [39], LDCF [132], Katamari [6], SCCPriors [225], and Checkerboards [243]) are handcrafted features based methods. The remaining methods are deep features based approaches. Miss rates witnessed a significant drop due to the introduction of deep methods DeepParts [181] and CompaACT-Deep [12]. On **R** set, the best method is two-stage AR-Ped [9]. On **HO**, **R+HO**, and **A** sets, the best method is TLL-TFA [177], which uses time-sequence information for detection. A likely reason is that time-sequence information plays an important role in occluded and small-scale pedestrian detection.

By using the new and accurate annotations [241] of Caltech pedestrian dataset, some state-of-the-art methods (e.g., JointDet [30] and PedHunter [29]) report a relatively lower miss-rate on **R** set. These two methods use the head information to improve pedestrian detection. Further, the lower miss-rate indicates that the performance on Caltech pedestrian dataset is close to be saturated.

Table 7 compares several state-of-the-art methods on the KITTI test set [53]. The methods only using 2D image annotations are selected for fair comparison. Among these these methods, only four methods (i.e., ACF [39], Checkerboards [243], NNNF [17], and Regionlets [200], [201]) are handcrafted features based approaches and the remaining methods are deep features based approaches. On `Medium` and `Hard` sets, RRC [163] and Aston-EAS [202] are the top two methods. On `Easy` set, FFNet [252] and MHN [18] are the top two methods. Most of these methods adopt feature pyramid structure with multi-scale feature fusion and data augmentation strategy (e.g., multi-scale training). Compared with the `Easy` set, the `Hard` set contains more small-sized pedestrians, occluded pedestrians, and truncated pedestrians. As a result, the `Hard` set provides more than 10 percent lower performance, which indicates that the small-sized and occluded pedestrian detection are the two main bottlenecks.

Table 8 compares several state-of-the-art methods on CityPersons validation set [244]. All these methods are deep

TABLE 5
Summary of Pedestrian Datasets

Name	Publication	#Images	#Pedestrians	Resolutions	Annotations	Time	Description
MIT [153]	IJCV2000	-	924	64×128	full	day	one of earliest pedestrian datasets
INRIA [38]	CVPR2005	2120	1774	640×480	full	day	one of earliest popular pedestrian datasets
ETH [46]	ICCV2007	1803	12k	640×480	full	day	a pair of images in busy shopping streets
TUD-Brussels [204]	CVPR2009	508	1326	640×480	full	day	pedestrians in the inner city of Brussels
Daimler [45]	PAMI2009	29k	72k	640×480	full	day	gray-color images in urban traffic
Caltech [43]	PAMI2010	250k	289k	640×480	full, visible	day	a standard and complete pedestrian datasts
KITTI [53]	CVPR2012	15k	9k	1240×376	full	day	a real-world computer vision benchmarks
CityPersons [245]	CVPR2017	5k	32k	2048×1024	full, visible	day	extensions on top of the Cityscapes [33]
CrowdHuman [172]	arXiv2018	24k	552k	-	full, visible, head	day	humans in crowded scenes from website
EuroCity [8]	PAMI2019	47k	219k	1920×1024	full	day, night	images in multiple European Cities
NightOwls [133]	ACCV2019	281k	56k	1024×640	full	night	pedestrians at night in three countries
WIDER Pedestrian	Challenge	97k	307k	-	full	day	pedestrians in traffic and surveillance scenes
WiderPerson [248]	TMM2019	13k	39k	-	full	day	persons in the wild, not only traffic
TJU-Pedestrian [149]	TIP2020	75k	373k	-	full, visible	day, night	a diverse dataset in traffic and campus
KAIST [79]	CVPR2015	95k	103k	640×480	full	day, night	color-thermal image pairs in traffic scene
CVC-14 [58]	Sensors2016	5051	7795	640×512	full	day, night	multimodal (FIR+visible) videosequences

The top is early pedestrian datasets, the middle is modern pedestrian datasets, and the bottom is multispectral pedestrian datasets.'full' means fully-body bounding-box, 'visible' means visible-body bounding-box, and 'head' means head bounding-box.

TABLE 6
Miss Rates (MR) of 30 State-of-the-Art Methods on Caltech Pedestrian Dataset

Method	Family	Backbone	Time	R↓	HO↓	R+HO↓	A↓
ACF [39]	CF	-	0.11^C	44.2	90.2	54.6	79.6
LDCF [132]	CF	-	0.28^C	24.8	81.3	37.7	71.2
Katamari [6]	DPM	-	-	22.5	84.4	36.2	71.3
DeepCascade [1]	Hybrid	-	0.67^C	31.1	81.7	42.4	74.1
SCCPriors [226]	CF	-	3.88^C	21.9	80.9	35.1	70.3
TA-CNN [183]	Hybrid	AlexNet	-	20.9	70.4	33.3	71.2
CCF [221]	Hybrid	AlexNet	10.0^Z	18.7	72.4	30.6	66.7
Checkerboards [244]	CF	-	2.0^X	18.5	77.5	31.8	68.7
DeepParts [182]	Hybrid	GoogleNet	-	11.9	60.4	22.8	64.8
CompACT-Deep [12]	Hybrid	VGG16	0.5^K	11.7	65.8	24.6	64.4
MS-CNN [11]	P-CNN	VGG16	0.13^X	10.0	59.9	21.5	60.9
RPN+BF [237]	Hybrid	VGG16	0.5^K	9.6	74.3	24.0	64.7
F-DNN [44]	P-CNN	GoogleNet	0.3^X	8.6	55.1	19.3	50.6
PCN [194]	P-CNN	VGG16	-	8.4	55.8	19.2	61.9
PDOE [261]	P-CNN	VGG16	-	7.6	44.4	-	-
UDN+ [145]	P-CNN	VGG16	-	11.5	70.3	24.7	64.8
FRCNN+ATT [249]	P-CNN	VGG16	-	10.3	45.2	18.2	54.5
SAF-RCNN [99]	Hybrid	VGG16	0.59^X	9.7	64.4	21.9	62.6
ADM [252]	P-CNN	ResNet50	-	8.6	30.4	13.7	42.3
GDFL [106]	P-CNN	VGG16	0.05^1	7.8	43.2	15.6	48.1
TLL-TFA [178]	P-CNN	ResNet50	-	7.4	**28.7**	**12.3**	**38.2**
AR-Ped [9]	P-CNN	VGG16	0.09^1	**6.5**	48.8	16.1	58.9
FRCN+A+DT [259]	P-CNN	VGG16	-	8.0	37.9	-	-
MGAN [152]	P-CNN	VGG16	-	6.8	38.1	13.8	-
HyperLearner [129]	P-CNN	VGG16	-	5.5	-	-	-
RepLoss [200]	P-CNN	ResNet50	-	4.0	-	-	-
ALFNet [118]	P-CNN	ResNet50	0.05^1	4.5	-	-	-
OR-CNN [246]	P-CNN	VGG16	-	4.1	-	-	-
JointDet [30]	P-CNN	ResNet50	-	3.0	-	-	-
PedHuter [29]	P-CNN	ResNet50	-	**2.3**	-	-	-

*The sets of **R**, **HO**, **R+HO**, and **A** are used for evaluation. The top part is based on the standard annotations of Caltech [43], and the bottom part is based on the new and accurate annotations [241]. The superscripts C/K/X/Z/1 indicate the CPU, NVIDIA K40 GPU, NVIDIA TitanX GPU, NVIDIA TitanZ GPU, and NVIDIA 1080Ti GPU.*

features based methods. Two subsets of **R** and **HO**, which adopt similar metrics as that of the Caltech dataset [43], are used here for performance evaluation. Note that there are two different settings about **HO**. Most of these state-of-the-art methods are two-stage methods, and are the variants of Faster R-CNN [164]. Additionally, 0.5-stage detector [185] that uses the pseudo-segmentation for anchor generation also achieves the state-of-the-art performance. For occluded pedestrian detection, the methods (i.e., MGAN [151], JointDet[30], PedHunter [29], and R^2NMS [78]) using part information (e.g., visible and head annotations) have a better performance. Appendix D, available in the online supplemental material, further compares several state-of-the-art methods on CityPersons test set [244].

In addition to detection accuracy, we also provide the test time of different methods in the three datasets above. By comprehensive analysis over the three datasets, the single-stage methods usually have faster speed. For example, GDFL [106] and ALFNet [118] are superior in terms of speed on Caltech test set, while PRNet [178] and CSP [119] are the fastest on CityPersons validation set. In addition, these single-stage methods have comparable accuracy to state-of-the-art two-stage methods. This suggests that single-stage methods usually have a better trade-off between accuracy and detection speed.

Table 9 further compares some state-of-the-art methods on multispectral KAIST test set [79], including hybrid methods and pure CNN based methods. With illumination aware feature alignment and single-stage structure, MBNet [261] achieves the best performance on both speed and accuracy.

TABLE 7
Average Precisions (AP) of 21 State-of-the-Art Methods on the KITTI Test Set

Method	Family	Backbone	Time	Medium↑	Easy↑	Hard↑
ACF [39]	CF	-	0.2^C	39.81	44.49	37.21
Checkerboards [244]	CF	-	2.0^C	56.75	67.65	51.12
DeepParts [182]	Hybrid	GoogleNet	1.0^C	58.67	70.49	52.78
CompACT-Deep [12]	Hybrid	VGG16	1.0^K	58.74	70.69	52.71
Regionlets [202]	DPM	-	1.0^C	60.83	73.79	54.72
NNNF [17]	CF	-	0.5^C	58.01	69.16	52.77
RPN+BF [237]	Hybrid	VGG16	0.6^K	61.29	75.45	56.08
SDP+RPN [224]	Hybrid	VGG16	0.4^K	70.42	82.07	65.09
IVA [266]	P-CNN	VGG16	0.4^X	71.37	84.61	64.90
MS-CNN [11]	P-CNN	VGG16	0.4^X	74.89	85.71	68.99
SubCNN [211]	P-CNN	GoogleNet	-	72.27	84.88	66.82
GN [83]	Hybrid	VGG16	1.0^X	72.29	82.93	65.56
RRC [164]	P-CNN	VGG16	3.6^X	**76.61**	85.98	**71.47**
CFM [76]	CF	VGG16	2.0^K	62.84	74.76	56.06
SJTU-HW [250]	P-CNN	VGG16	0.85^X	75.81	87.17	69.86
GDFL [106]	P-CNN	VGG16	0.27^1	68.62	84.61	66.86
MonoPSR [94]	P-CNN	ResNet101	0.2^X	68.56	85.60	63.34
FFNet [253]	P-CNN	VGG16	1.07^1	75.99	**87.21**	69.86
MHN [18]	P-CNN	VGG16	0.39^X	75.99	**87.21**	69.50
Aston-EAS [203]	P-CNN	VGG16	0.24^1	76.07	86.71	70.02
AR-Ped [9]	P-CNN	VGG16	-	73.44	83.66	68.12

The superscripts C/K/X/1 indicate the CPU, NVIDIA K40 GPU, NVIDIA TitanX GPU, and NVIDIA 1080Ti GPU.

To give more insights on pedestrian detection, we further analyze pedestrian detection from different aspects, including comparison with generic object detector, the impact of data processing, the impact of feature extraction, the impact of post-processing, and generalization ability analysis.

Comparison With Generic Object Detector. Here, we provide an analysis regarding using a generic object detector FPN for pedestrian detection in Table 10. We perform the 1× training scheme (12 epochs) on two GPUs with the initial learning rate of 0.005. There are two images per GPU. All the pedestrians over 20 pixels in height with less than 80 percent occlusion are used for training. The original FPN achieves 18.11 percent MR on **R** set. By ignoring the proposals located in ignored regions, it provides 3.30 percent improvement on **R** set. When further setting the aspect ratio as 0.41 and uniform scale, it provides 0.55 percent improvement. With these simple modifications

TABLE 8
Miss Rates (MR) of 17 State-of-the-Art Methods on CityPersons Validation Set

Method	Family	Backbone	Scale	Time	R↓	HO↓
Adapted FR-CNN [245]	P-CNN	VGG16	1×	-	15.4	-
RepLoss [200]	P-CNN	ResNet50	1×	-	13.7	56.9†
FRCNN+ATT [249]	P-CNN	VGG16	1×	-	16.0	56.7
TLL+MRF [178]	P-CNN	ResNet50	1×	-	14.4	52.0†
OR-CNN [246]	P-CNN	VGG16	1×	-	12.8	55.7†
ALFNet [118]	P-CNN	ResNet50	1×	0.27^1	12.0	51.9†
CSP [119]	P-CNN	ResNet50	1×	0.33^1	11.0	49.3†
Adaptive-NMS [114]	P-CNN	VGG16	1×	-	11.9	55.2†
MGAN [152]	P-CNN	VGG16	1×	-	11.3	**42.0**
R^2NMS [78]	P-CNN	VGG16	1×	-	11.1	**53.3**†
PRNet [179]	P-CNN	ResNet50	1×	0.22^1	**10.8**	**42.0**
Adapted FR-CNN [245]	P-CNN	VGG16	1.3×	-	12.8	-
RepLoss [200]	P-CNN	ResNet50	1.3×	-	11.6	55.3†
OR-CNN [246]	P-CNN	VGG16	1.3×	-	11.0	51.3†
PDOE [261]	P-CNN	VGG16	1.3×	-	11.2	44.2
Adaptive-NMS [114]	P-CNN	VGG16	1.3×	-	10.8	54.2†
IoU_{vis}+Sign [124]	P-CNN	VGG16	1.3×	-	10.8	54.3†
FRCN+A+DT [259]	P-CNN	VGG16	1.3×	-	11.1	44.3
MGAN [152]	P-CNN	VGG16	1.3×	-	10.5	**39.4**
JointDet [30]	P-CNN	ResNet50	1.3×	-	10.2	-
0.5-stage [186]	P-CNN	ResNet50	1.3×	-	**8.1**	-
PedHunter [29]	P-CNN	ResNet50	1.3×	-	8.3	**43.5**†

The superscript 1 indicates the NVIDIA 1080Ti GPU. The superscript † indicates the pedestrians over 50 pixels in height with more than 35 percent occlusion, instead of pedestrians over 50 pixels in height with 35-80 percent occlusion. Thus, † suggest higher difficulty.

TABLE 9
Miss Rates of State-of-the Art Detectors on KAIST **R** Test Set Using the Annotations Provided by [113]

Method	Family	Backbone	Time	All↓	Day↓	Night↓
Halfway Fusion [113]	Hybrid	VGG16	0.43^{X}	25.75	24.88	26.59
Fusion RPN [92]	Hybrid	VGG16	0.80^{X}	25.75	24.88	26.59
IAFR-CNN [96]	P-CNN	VGG16	0.21^{X}	15.73	14.55	18.26
IATDNN+IASS [61]	P-CNN	VGG16	0.25^{X}	14.95	14.67	15.72
CIAN [237]	P-CNN	VGG16	0.07^{1}	14.12	14.77	11.13
MSDSR-CNN [95]	P-CNN	VGG16	0.23^{X}	11.34	10.53	12.94
AR-CNN [238]	P-CNN	VGG16	0.12^{1}	9.34	9.94	8.38
MBNet [261]	P-CNN	ResNet50	0.07^{1}	**8.13**	**8.28**	**7.86**

The all test set (All) contains the day subset (Day) and the night subset (Night). The superscripts X/1 indicate NVIDIA Titan X/1080Ti.

specific to pedestrian detection, our improved FPN has a comparable, yet not state-of-the-art, performance in pedestrian detection. Finally, our improved FPN is also used as the baseline detector in the following experiments.

Generally, the generic object detector simultaneously detects multiple object classes. A natural question to ask is: does pedestrian detection benefit from additional object classes. Table 11 shows the impact of extra-class training on pedestrian detection. Experiments are performed on two datasets, including TJU-Ped-Traffic [148] and MS COCO [112]. TJU-Ped-Traffic [148] has 5 object classes, while MS COCO [112] has 80 object classes. We observe that performing multi-class training does not improve over single-class training. We believe how to effectively exploit the contextual information and the relation between person and other objects is an interesting future direction.

Impact of Data Processing. Here, we analyze the impact of different data processing on pedestrian detection in Table 12. The first (top) half of Table 12 shows the impact of different intra-dataset data processing strategies. (1) Instead of the pedestrians over 20 pixels in height with less than 80 percent occlusion for training, we use the pedestrians over 50 pixels in height with less than 65 percent occlusion (see (b)). We observe MR on the **R** set to be similar, but MRs on **HO** and **A** sets have a large drop. This shows that adding occluded and small-scale pedestrians during training is important for improving occluded and small-scale pedestrian detection. (2) Using multi-scale training strategy (see (c)) is effective for improving pedestrian detection. (3) We also randomly erase the part of pedestrians with pixel mean value during the training (see (d)), observing improvement in occluded pedestrian detection likely due to enhancing the diversity of pedestrian occlusion. (4) To enlarge the number of pedestrians, we randomly paste the pedestrians (bounding-box) from other images with geometric prior. We observe no significant improvement in performance likely due to the fact that a simple bounding-box copy-paste also incorporates some background information and brings the style inconsistency (e.g., illumination variation).

The bottom part of Table 12 shows the impact of pre-training on different different datasets, including Caltech [43], MS COCO [112], CrowdHuman [171], and TJU-Pedestrian [148]. When performing pre-training on Caltech, we observe no significant improvement in performance on the **R** set. The large-scale generic object dataset MS COCO provides 2.3 percent improvement on **R** set. However, pre-training with MS COCO is inferior to the person related datasets, namely CrowdHuman [171] and TJU-Pedestrian [148]. Among these person related datasets, pre-training on TJU-Pedestrian leads to more favorable performance on **R** set, likely due to a large number of pedestrians and the related scenes (pedestrian scenarios).

Impact of Feature Extraction. Second, we show the impact of feature extraction using different backbones in Table 13, including ResNet50 [68], ResNet101 [68], VGG16 [176], RegNet [158], and HRNet [191]. Among these backbones, HRNet achieves the best detection performance, especially on **HO** and **A**. The reason maybe explained as that high-resolution and high-semantic representations are important for improving pedestrian detection.

Impact of Post-Processing. Third, we show the impact of two post-processing strategies (i.e., NMS and SoftNMS) in Table 14. It is difficult to achieve the optimal results on all the sets by using a single threshold θ. Some recent methods [29], [78] adopt visible part or head for occluded pedestrian detection, which not only need extra annotations but also face the problem of threshold settings. Thus, it is interesting to pay more attention on NMS-free methods in future.

Generalization Ability. Finally, we analyze generalization ability by performing cross-dataset evaluations in Table 15. Both FPN and Cascade R-CNN have a sub-optimal performance during cross-dataset evaluation (see rows 1-4), which indicates poor generalization ability. Similar findings are also presented in CSP [119]. Furthermore, we analyze the

TABLE 10
Improving Generic Object Detector for Pedestrian Detection on the CityPersons Validation Set

Method	Backbone	R↓	HO↓	A↓
FPN-vanilla [110]	ResNet50	18.11	50.99	43.11
+ignored region handling	ResNet50	14.81	48.09	39.88
+anchor aspect ratios & scales	ResNet50	14.26	44.52	38.17

TABLE 11
Impact of Extra-Class Training on Pedestrian Detection

Dataset	Training strategy	**R**↓	**HO**↓	**A**↓
TJU-Ped-Traffic [149]	Multi-class training	22.59	61.23	38.36
	Single-class training	22.36	60.47	37.78
Dataset	Training strategy	AP↑	AP@0.5↑	AP@0.75↑
MS COCO [112]	Multi-class training	52.2	81.5	56.5
	Single-class training	53.2	81.9	57.2

Simply using additional object classes cannot improve pedestrian detection.

TABLE 12
Impact of Different Data Processing Operations During Training on the CityPersons Validation Set

Method	Backbone	R↓	HO↓	A↓
(a) Baseline (FPN)	ResNet50	14.26	44.52	38.17
(b) Reasonable pedestrians only	ResNet50	14.34	57.71	45.10
(c) Multi-scale training	ResNet50	13.36	44.02	35.99
(d) Random erase augmentation	ResNet50	13.97	43.35	37.31
(e) Copy-paste augmentation	ResNet50	14.10	43.92	38.30
(f) Caltech→CityPersons	ResNet50	13.92	45.56	38.35
(g) MS COCO→CityPersons	ResNet50	11.96	39.43	35.10
(h) CrowdHuman→CityPersons	ResNet50	11.62	38.13	34.23
(i) TJU-Pedestrian→CityPersons	ResNet50	9.80	34.68	34.51

effect of the combined training on CityPersons and Caltech datasets (see rows 5-6). In both cases, we observe that the combined training does not achieve better performance, compared to individual dataset-specific training.

6 Challenges

6.1 Scale Variance

Traffic and video surveillance scenes usually contain pedestrians of various scales. Fig. 9 shows some examples. Large-scale and small-scale pedestrians exhibit high intra-class variations. As a result, it is challenging to use a single detector to accurately detect pedestrians of varying scales.

One of straightforward idea is scale independent strategy, including image pyramid based and feature pyramid based methods. For image pyramid based methods, it is important to improve its efficiency. In contrast, feature pyramid strategy is relatively effective. However, how to extract robust scale-independent features is important for feature pyramid strategy. Another idea is reducing the difference between pedestrians of different scales [149], [207], [228]. We can exploit to reduce intra-class variations in both data and feature levels with generative adversarial networks.

In addition, small-scale pedestrian detection is the bottleneck for solving scale variance problem. However, the existing methods lack of the enough research on small scale object detection [64], [230]. Thus, it it necessary to treat small-scale pedestrian detection as a standalone problem.

TABLE 13
Impact of Different Backbones on the CityPersons Validation Set

Method	Backbone	AP on COCO ↑	R↓	HO↓	A↓
(a) FPN	ResNet50	37.4	14.26	44.52	38.17
(b) FPN	VGG16	-	13.79	44.10	39.51
(c) FPN	ResNet101	39.4	14.55	44.10	39.03
(d) FPN	RegNet	39.9	13.29	45.82	37.31
(e) FPN	HRNet	40.2	12.64	41.02	35.19

TABLE 14
Impact of Different NMS Strategies on the CityPersons Validation Set

Method	Backbone	Threshold	R↓	HO↓	A↓
(a) NMS	ResNet50	$\theta = 0.4$	14.67	45.46	38.69
	ResNet50	$\theta = 0.5$	14.26	44.52	38.17
	ResNet50	$\theta = 0.6$	14.46	44.01	39.05
(b) SoftNMS	ResNet50	$\theta = 0.4$	14.02	44.76	38.44
	ResNet50	$\theta = 0.5$	14.11	44.25	38.12
	ResNet50	$\theta = 0.6$	14.46	43.98	39.05

6.2 Occlusion

Pedestrian occlusion is a very common problem. For instance, 70 percent of pedestrians in CityPersons dataset [244] are occluded. Fig. 10 shows two types of pedestrian occlusions: inter-class occlusion and intra-class occlusion. Inter-class occlusion occurs when pedestrians are occluded by other objects (not pedestrians). In contrast, intra-class occlusion occurs when pedestrians are occluded by other pedestrians.

The main solution to inter-class occlusion is enhancing the features of unoccluded part and suppressing the features of occluded part. There are two strategies: implicit strategy [181], [214] and explicit strategy [29], [151], [260]. The implicit strategy usually learns multiple part detectors to cover different occlusion patterns. In this strategy, how to set/combine the parts is an open problem. The explicit strategy uses extra annotations (e.g., head or visible-part annotation). Because

TABLE 15
Generalization Ability of the Detector on Different Datasets

Method	Train	CityPersons ↓	Caltech ↓
FPN [110]	Caltech	57.1	11.0
Cascade R-CNN [13]	Caltech	57.8	10.7
FPN [110]	CityPersons	14.3	24.9
Cascade R-CNN [13]	CityPersons	13.8	24.2
FPN [110]	CityPersons+Caltech	16.5	12.6
Cascade R-CNN [13]	CityPersons+Caltech	16.5	11.8

Fig. 9. Example pedestrians at various scales. From left to right, pedestrians vary from small scale to large scale. Pedestrians of different scales have large-scale variations and small-scale pedestrians are relatively noisy and blurry.

Fig. 10. Example pedestrians under different types of occlusion (i.e., inter-class and intra-class occlusions). Some examples of inter-class occlusion are shown in the left, where the level of occlusion varies from heavy to bare. Some examples of intra-class occlusion are shown in the right, where intra-class occlusion occurs between different pedestrians.

the extra annotations need large resource consumption, it is useful to exploit using a small number of part annotations to help occluded pedestrian detection. Usually, pedestrian detection needs to remove duplicate bounding-boxes by NMS. As a negative effect, NMS combines the bounding-boxes belonging to different pedestrians in crowd scene (intra-class occlusion). To address this, one way is to improve NMS strategy, for example, dynamic NMS threshold. Another way is NMS-free strategy by avoiding bounding-box combination operation.

6.3 Domain Adaptation

Most existing methods focus on a specific pedestrian dataset and can not guarantee a good domain adaptation ability [195], [212]. For instance, the detector trained under good weather often has a sub-optimal performance in poor weather (e.g., fog and rain). Therefore, it is necessary to address the issue of domain adaptation. Most domain adaptation methods are based on adversarial learning, including data-level [72], [88], [90], [122], feature-level [168], [215], and instance-level [24], [263] methods. However, they mainly focus on generic object detection. In future, the domain adaptation methods about pedestrian detection can be exploited by considering pedestrian/scene characteristic. In addition, we can exploit both the same-domain and cross-domain evaluations in future.

6.4 Multi-Sensor Fusion

In Section 4, multispectral pedestrian detection adopts two kinds of sensors (i.e., visible-light/infrared cameras). It provides more robust detection performance for illumination variance. To ensure safety and generate 3D information, we not only need to fuse information from multiple sensors of homogeneous data (visible-light/infrared cameras) but also multiple sensors of heterogeneous data (cameras and LiDAR). LiDAR can provide accurate depth information while cameras has detailed semantic information. However, how to fuse information in heterogeneous data is a challenging and important task in future.

6.5 Real-Time Detection

Most existing pedestrian detection methods focus on improving detection accuracy, while ignoring the efficiency. However, the application to driving/surveillance scenes has limited computational resources but requires real-time detection speed. For example, the fastest method reported on CityPersons has 0.22s inference speed on a single NVIDIA 1080Ti GPU, which can't meet the needs of real applications. Thus, it is necessary to study the light-weight and real-time pedestrian detection methods for the embedded device.

7 Conclusion

In the past decade, pedestrian detection has witnessed significant success, which has gone from the handcrafted features based methods to deep features based approaches. In this paper, we first summarize these two types of methods in detail. Afterwards, we review multispectral pedestrian detection. We review popular pedestrian datasets and a deep analysis on pedestrian detection methods. Finally, we discuss some challenging problems in pedestrian detection. We hope that this deep survey can help the researchers to develop new methods in the field of pedestrian detection.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102800 and in part by the National Natural Science Foundation of China under Grants 61906131 and 61632018, and in part by VR starting grant (2016-05543).

References

[1] A. Angelova, A. Krizhevsky, V. Vanhoucke 1, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 32.1–32.12.

[2] J. Baek, J. Hyun, and E. Kim, "A pedestrian detection system accelerated by kernelized proposals," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 3, pp. 1216–1228, Mar. 2020.

[3] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[4] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2903–2910.

[5] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3666–3673.

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

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

[8] M. Braun, S. Krebs, F. Flohr, and D. M. 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.

[9] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7224–7233.

[10] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation." in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4960–4969.

[11] 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.*, 2016, pp 354–370.

[12] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3361–3369.

[13] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[14] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.

[15] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, "Taking a look at small-scale pedestrians and occluded pedestrians," *IEEE Trans. Image Process.*, vol. 29, pp. 3143–3152, 2020.

[16] J. Cao, Y. Pang, and X. Li, "Learning multilayer channel features for pedestrian detection," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3210–3220, Jul. 2017.

[17] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5538–5551, Dec. 2016.
[18] J. Cao, Y. Pang, S. Zhao, and X. Li, "High-level semantic networks for multi-scale object detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2019, vol. 30, no. 10, pp. 3372–3386, Oct. 2020.
[19] Y. Cao, D. Guan, W. Huang, J. Yanga, Y. Cao, and Y. Qiao, "Pedestrian detection with unsupervised multispectral feature learning using deep neural networks," *Inf. Fusion*, vol. 46, pp. 206–217, 2019.
[20] Y. Cao, D. Guan, Y. Wu, J. Yang, Y. Cao, and M. Y. Yang, "Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 70–79, 2019.
[21] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.
[22] R. Chen, H. Ai, C. Shang, L. Chen, and Z. Zhuang, "Learning lightweight pedestrian detector with hierarchical knowledge distillation," in *IEEE Int. Conf. Image Process.*, 2019, pp. 1645–1649.
[23] T. Chen, S. Lu, and J. Fan, "S-CNN: Subcategory-aware convolutional networks for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2522–2528, Oct. 2018.
[24] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348 .
[25] Z. Chen and X. Huang, "Pedestrian detection for autonomous vehicle using multi-spectral cameras," *IEEE Trans. Intell. Veh.*, vol. 4, no. 2, pp. 211–219, Jun. 2019.
[26] Z. Chen, W. Ouyang, T. Liu, and D. Tao, "A shape transformation-based dataset augmentation framework for pedestrian detection," 2019, *arXiv: 1912.0701*.
[27] Z. Chen, L. Zhang, A. M. Khattak, W. Gao, and M. Wang, "Deep feature fusion by competitive attention for pedestrian detection," *IEEE Access*, vol. 7, pp. 21981–21989, 2019.
[28] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.
[29] 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.*, 2020, pp. 10639–10646.
[30] 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.*, 2020, pp. 10647–10654.
[31] J. Chu, Z. Guo, and L. Leng, "Object detection based on multilayer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018.
[32] 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.*, 2020, pp. 12214–12223 .
[33] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[35] A. D. Costea and S. Nedevschi, "Semantic channels for fast pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2360–2368.
[36] A. D. Costea, R. Varga, and S. Nedevschi, "Fast boosting based detection using scale invariant multimodal multiresolution filtered features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 993–1002.
[37] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[39] 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.
[40] P. Dollar, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 645–659.
[41] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 68.1–68.11.
[42] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.
[43] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[44] X. Du, M. El-Khamy , J. Lee, and L. S. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 953–961.
[45] 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. 2009.
[46] A. Ess, B. Leibe, and L. van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[47] C. Fei, B. Liu, Z. Chen, and N. Yu, "Learning pixel-level and instance-level context-aware features for pedestrian detection in crowds," *IEEE Access*, vol. 7, pp. 94944–94953, 2019.
[48] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester , "Cascade object detection with deformable part models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2241–2248.
[49] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[50] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–137, 1997.
[51] X. Fu, R. Yu, W. Zhang, L. Feng, and S. Shao, "Pedestrian detection by feature selected self-similarity features," *IEEE Access*, vol. 6, pp. 14223–14237, 2018.
[52] V. Gajjar, Y. Khandhediya, A. Gurnani, and V. Mavani, "ViS-HuD: Using visual saliency to improve human detection with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1989–19898.
[53] 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.*, 2012, pp. 3354–3361.
[54] D. Gerónimo, A. M. López, 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. 2009.
[55] D. Ghose, S. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection from thermal images using saliency maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 988–997.
[56] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[57] 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.*, 2014, pp. 580–587.
[58] A. Gonzalez *et al.* "Pedestrian detection at day/night time with visible and fir cameras: A comparison," *Sensors*, vol. 16, no. 6, 2016, Art. no. 820.
[59] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[60] G. Gualdi, A. Prati, and R. Cucchiara, "Multistage particle windows for fast and accurate object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1589–1604, Aug. 2012.
[61] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Inf. Fusion*, vol. 50, pp. 148–157, 2019.
[62] D. Guan *et al.*, "Unsupervised domain adaptation for multispectral pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 434–443.
[63] T. Guo, C. P. Huynh, and M. Solh, "Domain-adaptive pedestrian detection in thermal images," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1660–1664.
[64] B. Han, Y. Wang, Z. Yang, and X. Gao, "Small-scale pedestrian detection based on deep neural network," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 7, pp. 3046–3055, Jul. 2020.

[65] C. Han *et al.*, "Re-ID driven localization refinement for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9813–9822.

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

[67] H. Hattori, N. Lee, V. N. Boddeti, F. Beainy, K. M. Kitani, and T. Kanade, "Synthesizing a scene-specific pedestrian detector and pose estimator for static video surveillance," *Int. J. Comput. Vis.*, vol. 126, pp. 1027–1044, 2018.

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

[69] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Apr. 2016.

[70] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6469–6477.

[71] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4073–4082.

[72] H.-K. Hsu *et al.*, "Progressive domain adaptation for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 738–746.

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

[74] J. Hu, L. Jin, and S. Gao, "Fpn++: A simple baseline for pedestrian detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1138–1143.

[75] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7132–7141.

[76] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli, "Pushing the limits of deep CNNs for pedestrian detection," *IEEE TCSVT*, vol. 28, no. 6, pp. 1358–1368, Jun. 2018.

[77] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[78] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10747–10756.

[79] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[80] O. H. Jafari and M. Y. Yang, "Real-time RGB-D based template matching pedestrian detection," in *IEEE Int. Conf. Robot. Automat.*, 2016, pp. 5520–5527.

[81] X. Jiang, Y. Pang, X. Li, and J. Pan, "Speed up deep neural network based pedestrian detection by sharing features across multi-scale models," *Neurocomputing*, vol. 185, pp. 163–170, 2016.

[82] X. Jiang, Y. Pang, M. Sun, and X. Li, "Cascaded subpatch networks for effective CNNs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no, 7, pp. 2684–2694, Jul. 2017.

[83] S.-I. Jung and K.-S. Hong, "Deep network aided by guiding network for pedestrian detection," *Pattern Recognit. Lett.*, vol. 90, pp. 43–49, 2017.

[84] S.-I. Jung and K.-S. Hong, "Direct multi-scale dual-stream network for pedestrian detection," in *Proc. IEEE Int. Conf. Image Processing*, 2017, pp. 156–160.

[85] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras , O. I. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[86] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3306–3313.

[87] F. S. Khan, J. Xu, J. van de Weijer, A. Bagdanov, R. M. Anwer, and A. Lopez, "Recognizing actions through action-specific person detection," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4422–4432, Nov. 2015.

[88] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 480–490.

[89] M. Kieu, A. D. Bagdanov, M. Bertini, and A. del Bimbo, "Task-conditioned domain adaptation for pedestrian detection in thermal imagery," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp 546–562.

[90] T. Kim, M. Jeong, S. Kim, S. Choi, and C. Kim, "Diversify and match: A domain adaptive representation learning paradigm for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12448–1245.

[91] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "ClueNet: A deep framework for occluded pedestrian pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 94.1–94.15.

[92] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 243–250.

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 84–90.

[94] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3D object detection leveraging accurate proposals and shape reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11859–11868.

[95] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 225.1–225.12.

[96] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, 2019.

[97] C. Li, X. Wang, and W. Liu, "Neural features for pedestrian detection," *Neurocomputing*, vol. 238, pp. 420–432, 2017.

[98] G. Li, Y. Yang, and X. Qu, "Deep learning approaches on pedestrian detection in hazy weather," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8889–8899, Oct. 2019.

[99] 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.

[100] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1951–1959.

[101] J. Li *et al.*, "DSFD: Dual shot face detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5055–5064.

[102] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7341–7349.

[103] Q. Li, H. Wang, Y. Yan, B. Li, and C. W. Chen, "Local co-occurrence selection via partial least squares for pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 18, no. 6, pp. 3046–3055, Jun. 2017.

[104] X. Li, Y. Liu, Z. Chen, J. Zhou, and Y. Wu, "Fused discriminative metric learning for low resolution pedestrian detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 958–962.

[105] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2197–2206.

[106] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 745–761.

[107] C. Lin, J. Lu, and J. Zhou, "Multi-grained deep feature learning for robust pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3608–3621, Dec. 2019.

[108] 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.

[109] M. Lin *et al.*, "Detr for pedestrian detection," 2020, *arXiv: 2012.06785*.

[110] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

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

[112] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp 740–755.

[113] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, 73.1–73.13.
[114] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6452–6461.
[115] T. Liu, M. Elmikaty, and T. Stathaki, "SAM-RCNN: Scale-aware multi-resolution multi-channel pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 236.1–236.13.
[116] T. Liu, W. Luo, L. Ma, J.-J. Huang, T. Stathaki, and T. Dai, "Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling," 2019, *arXiv: 1912.08661*.
[117] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp 21–37.
[118] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 643–659.
[119] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5182–5191.
[120] X. Liu, K.-A. Toh, and J. P. Allebach, "Pedestrian detection using pixel difference matrix projection," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1441–1454, Apr. 2020.
[121] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 640–651.
[122] A. Lopez, Rodriguez, and K. Mikolajczyk, "Domain adaptation for object detection via style consistency," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 204.1–204.14.
[123] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
[124] R. Lu and H. Ma, "Occluded pedestrian detection with visible iou and box sign predictor," in *Proc. IEEEInt. Conf. Image Process.*, 2019, pp. 1640–1644.
[125] R. Lu and H. Ma, "Semantic head enhanced pedestrian detection in a crowd," 2019, *arXiv: 1911.11985*.
[126] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3618–3627.
[127] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 899–906.
[128] 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.*, 2020, pp. 14065–14073.
[129] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6034–6043.
[130] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1505–1512.
[131] M. Najibi, B. Singh, and L. S. Davis, "FA-RPN: Floating region proposals for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7715–7724.
[132] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 424–432.
[133] L. Neumann *et al.*, "NightOwls: A pedestrians at night dataset," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp 691–705.
[134] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp 483–499.
[135] 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.*, 2018, pp. 966–974.
[136] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Person recognition in personal photo collections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 203–220, Jan. 2020.
[137] T. Ojala, M. PietikaEinen, and T. MaEenpaEa , "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.
[138] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2337–2344.
[139] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3258–3265.
[140] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2056–2063.
[141] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3198–3205.
[142] W. Ouyang, X. Zeng, and X. Wang, "Single-pedestrian detection aided by two-pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1875–1889, Sep. 2015.
[143] W. Ouyang *et al.*, "DeepID-Net: Object detection with deformable part based convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1320–1334, Jul. 2017.
[144] 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.
[145] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp 546–561.
[146] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1243–1257, Jun. 2016.
[147] Y. Pang, J. Cao, and X. Li, "Learning sampling distributions for efficient object detection," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 117–129, Jan. 2017.
[148] Y. Pang, J. Cao, Y. Li, J. Xie, H. Sun, and J. Gong, "TJU-DHD: A diverse high-resolution dataset for object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 207–219, 2021.
[149] Y. Pang, J. Cao, J. Wang, and J. Han, "JCS-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3322–3331, Dec. 2019.
[150] Y. Pang, Y. Li, and L. S. J. Shen, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4229–4238.
[151] 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.*, 2019, pp. 4966–4974.
[152] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 30, no. 1, pp. 15–33, 2000.
[153] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp 241–25.
[154] D. Park, L. Zitnick, D. Ramanan, and P. Dollár, "Exploring weak stabilization for motion feature extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2882–2889.
[155] K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognit.*, vol. 80, pp. 143–155, 2018.
[156] X. Peng, Y. Murphey, S. Stent, Y. Li, and Z. Zhao, "Spatial focal loss for pedestrian detection in fisheye imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 561–569.
[157] Y. Qian, M. Yang, X. Zhao, C. Wang, and B. Wang, "Oriented spatial transformer network for pedestrian detection using fisheye camera," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 421–431, Feb. 2020.
[158] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing network design spaces," Iin *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10425–10433.
[159] N. K. Ragesh and R. Rajesh, "Pedestrian detection in automotive safety: Understanding state-of-the-art," *IEEE Access*, vol. 7, pp. 47864–47890, 2019.
[160] R. N. Rajaram, E. Ohn-Bar , and M. M. Trivedi, "Looking at pedestrians at different scales: A multiresolution approach and evaluations," *IEEE Trans. Intell. Transport. Syst.*, vol. 17, no. 12, pp. 3565–3576, 2016.
[161] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2020.
[162] 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.*, 2016, pp. 779–788.

[163] J. Ren *et al.*, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 752–760.

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

[165] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, "Improving the performance of pedestrian detectors using convolutional learning," *Pattern Recognit.*, vol. 61, pp. 641–649, 2017.

[166] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.

[167] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[168] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6949–6958.

[169] 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.*, 2013, pp. 3626–3633.

[170] C. Shang, H. Ai, Z. Zhuang, L. Chen, and J. Xing, "ZoomNet: Deep aggregation learning for high-performance small pedestrian detection," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 486–501.

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

[172] J. Shen, X. Zuo, J. Li, W. Yang, and H. Ling, "A novel pixel neighborhood differential statistic feature for pedestrian and face detection," *Pattern Recognit.*, vol. 63, pp. 127–138, 2017.

[173] J. Shen, X. Zuo, W. Yang, D. Prokhorov, X. Mei, and H. Ling, "Differential features for pedestrian detection: A taylor series perspective," *IEEE Trans. Intell. Transport. Syst.*, vol. 20, no. 8, pp. 2913–2922, Aug. 2019.

[174] J. Shen, X. Zuo, L. Zhu, J. Li, W. Yang, and H. Ling, "Pedestrian proposal and refining based on the shared pixel differential feature," *IEEE Trans. Intell. Transport. Syst.*, vol. 20, no. 6, pp. 2085–2095, Jun. 2019.

[175] B. Sheng, Q. Hu, J. Li, W. Yang, B. Zhang, and C. Sun, "Filtered shallow-deep feature channels for pedestrian detection," *Neurocomputing*, vol. 249, pp. 19–27, 2017.

[176] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[177] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 554–569.

[178] 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.*, 2020, pp. 32–48.

[179] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2325–2333.

[180] F. B. Tesema, H. Wu, M. Chen, J. Lin, W. Zhu, and K. Huang, "Hybrid channel based pedestrian detection," *Neurocomputing*, vol. 389, pp. 1–8, 2020.

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

[182] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5079–5087.

[183] R. Trichet and F. Bremond, "LBP channels for pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1066–1074.

[184] J. R. Uijlings, K. E. V. D. Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2016.

[185] U. Ujjwal, A. Dziri, B. Leroy, and F. Bremond, "A one-and-half stage pedestrian detector," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 765–774.

[186] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.

[187] A. Vobecky, M. Uricar, D. Hurych, and R. Skoviera, "Advanced pedestrian dataset augmentation for autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2367–2372.

[188] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1030–1037.

[189] H. Wang, Y. Li, and S. Wang, "Fast pedestrian detection with attention-enhanced multi-scale RPN and soft-cascaded decision trees," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 12, pp. 5086–5093, Dec. 2020.

[190] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2960–2969.

[191] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2020, doi: 10.1109/TPAMI.2020.2983686.

[192] L. Wang, L. Xu, and M.-H. Yang, "Pedestrian detection in crowded scenes via scale and occlusion analysis," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1210–1214.

[193] S. Wang, J. Cheng, H. Liu, and M. Tang, "PCN: Part and context information for pedestrian detection with CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 34.1–34.13.

[194] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body part semantic and contextual information with DNN," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3148–3159, Nov. 2018.

[195] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7166–7175.

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

[197] X. Wang *et al.*, "S3D: Scalable pedestrian detection via score scale surface discrimination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3332–3344, Oct. 2020.

[198] X. Wang, C. Shen, H. Li, and S. Xu, "Human detection aided by deeply learned semantic masks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2663–2673, Aug. 2020.

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

[200] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 17–24.

[201] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2071–2084, Oct. 2015.

[202] J. Wei, J. He, Y. Zhou, K. Chen, Z. Tang, and Z. Xiong, "Enhanced object detection with deep convolutional neural networks for advanced driving assistance," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1572–1583, Apr. 2020.

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

[204] C.-H. Wu, W. Gan, D. Lan, and C.-C. J. Kuo, "Boosted convolutional neural networks (BCNN) for pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 540–549.

[205] J. Wu, Y. Peng, C. Zheng, Z. Hao, and J. Zhang, "PMC-GANs: Generating multi-scale high-quality pedestrian with multimodal cascaded GANs," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 157.1–157.14.

[206] 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.*, 2020, pp. 13427–13436.

[207] J. Wu, C. Zhou, Q. Zhang, M. Yang, and J. Yuan, "Self-mimic learning for small-scale pedestrian detection," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2012–2020.

[208] S. Wu, S. Wang, R. Laganiere, C. Liu, H.-S. Wong, and Y. Xu, "Exploiting target data to learn deep convolutional networks for scene-adapted human detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1418–1432, Mar. 2018.

[209] S. Wu *et al.*, "Semi-supervised human detection via region proposal networks aided by verification," *IEEE Trans. Image. Process.*, vol. 29, pp. 1562–1574, Oct. 2020.

[210] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 924–933.

[211] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3376–3385.

[212] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 361–371, Feb. 2014.

[213] J. Xie *et al*., "Count- and similarity-aware R-CNN for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp 88–104.
[214] J. Xie, Y. Pang, H. Cholakkal, R. M. Anwer, F. S. Khan, and L. Shao, "PSC-Net: Learning part spatial co-occurence for occluded pedestrian detection," 2020, *arXiv: 2001.09252*.
[215] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang, "Multi-level domain adaptive learning for cross-domain detection," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 3213–3219.
[216] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4236–4244.
[217] M. Xu, Y. Bai, S. S. Qu, and B. Ghanem, "Semantic part RCNN for real-world pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 45–54 .
[218] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3033–3040.
[219] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, "Learning context graph for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2153–2162.
[220] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 82–90.
[221] 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*.
[222] F. Yang, H. Chen, J. Li, F. Li, L. Wang, and X. Yan, "Single shot multibox detector with kalman filter for online pedestrian detection in video," *IEEE Access*, vol. 7, pp. 15478–15488, 2019.
[223] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2129–2137.
[224] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.
[225] Y. Yang, Z. Wang, and F. Wu, "Exploring prior knowledge for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 176.1–176.12.
[226] M. Ye, X. Lan, and P. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 176–193.
[227] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2655–2666, Jan. 2020.
[228] R. Yin, "Multi-resolution generative adversarial networks for tiny-scale pedestrian detection," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1665–1669.
[229] M. You, Y. Zhang, C. Shen, and X. Zhang, "An extended filtered channel framework for pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 19, no. 5, pp. 1640–1651, May 2018.
[230] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, "Scale match for tiny person detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1246–1254.
[231] I. Yun, C. Jung, X. Wang, A. O. Hero, and J. K. Kim, "Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment," *IEEE Access*, vol. 7, pp. 23027–23037, 2019.
[232] C. Zhang and J. Kim, "Object detection with location-aware deformable convolution and backward attention filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9444–9453.
[233] H. Zhang, K. Wang, Y. Tian, C. Gou, and F.-Y. Wang, "MFR-CNN: Incorporating multi-scale features and global information for traffic object detection," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8019–8030, Sep. 2018.
[234] J. Zhang *et al*., "Attribute-aware pedestrian detection in a crowd," 2019, *arXiv: 1912.08661*.
[235] 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*.
[236] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.
[237] L. Zhang *et al*., "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, 2019.
[238] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5126–5136.
[239] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 947–954.
[240] S. Zhang, C. Bauckhage, D. A. Klein, and A. B. Cremers, "Exploring human vision driven features for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1709–1720, Oct. 2015.
[241] 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.*, 2016, pp. 1259–1267.
[242] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.
[243] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1751–1760.
[244] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4457–4465.
[245] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 657–674.
[246] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 657–674.
[247] 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.
[248] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6995–7003.
[249] S. Zhang, X. Zhao, L. Fang, H. Fei, and H. Song, "LED: Localization-quality estimation embedded detector," in *ICIP*, 2018, pp. 584–588.
[250] T. Zhang, Z. Han, H. Xu, B. Zhang, and Q. Ye, "CircleNet: Reciprocating feature adaptation for robust pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, 2019, pp. 4593–4604.
[251] 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.
[252] C. Zhao, Y. Qian, and M. Yang, "Monocular pedestrian orientation estimation based on deep 2D-3D feedforward," *Pattern Recognit.*, vol. 100, 2019, Art. no. 107182.
[253] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3586–3593.
[254] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, Sep. 2019.
[255] Y. Zhao, Z. Yuan, and B. Chen, "Training cascade compact CNN with region-IoU for accurate pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 9, pp. 3777–3787, Sep. 2020.
[256] Y. Zhao, Z. Yuan, and H. Zhang, "Joint holistic and partial CNN for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 81.1–81.12.
[257] C. Zhou, M. Wu, and S.-K. Lam, "Group cost-sensitive boostLR with vector form decorrelated filters for pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 12, pp. 5022–5035, Dec. 2020.
[258] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9556–9565.
[259] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3506–3515.
[260] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–151.
[261] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 787–803.

[262] C. Zhu and Y. Peng, "Discriminative latent semantic feature learning for pedestrian detection," *Neurocomputing*, vol. 238, pp. 126–136, 2017.
[263] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 687–696.
[264] 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.*, 2016, pp. 416–430.
[265] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.
[266] 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, 2020.

Jiale Cao received the PhD degree in information and communication engineering from Tianjin University, Tianjin, China, in 2018. He is currently an associate professor with Tianjin University. He has authored or coauthored 15 papers in the CVPR, ICCV, ECCV, *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, and *IEEE Transactions on Information Forensics and Security* in the area of his reaearch field which include object detection and deep learning.

Yanwei Pang (Senior Member, IEEE) received the PhD degree in electronic engineering from the University of Science and Technology of China. He is currently a professor with Tianjin University and also the founding director of the Tianjin Key Laboratory of Brain Inspired Intelligence Technology. He has authored or coauthored 150 scientific papers including 70 papers in top journals or conferences. He is an associate editor for *IEEE Transactions on Neural Networks and Learning Systems* and *Neural Networks* (Elsevier).

Jin Xie received the BS degree in electronic engineering from Tianjin University, Tianjin, China, in 2016. He is currently working toward the PhD degree with Tianjin University and his supervisor is Prof. Yanwei Pang. He has authored or coauthored five papers in CVPR, ICCV, ECCV, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Cybernetics* in the area of his reaearch field which include machine learning and computer vision.

Fahad Shahbaz Khan (Senior Member, IEEE) received the PhD degree in computer vision from Computer Vision Center Barcelona and Autonomous University of Barcelona, Spain. He is currently a faculty member with the Mohamed bin Zayed University of Artificial Intelligence, UAE, and Linköping University, Sweden. His research interests include a wide range of topics within computer vision.

Ling Shao (Fellow, IEEE) is currently the CEO and the chief scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of the IEEE, the IAPR, the IET, and the BCS.