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], [104], [226], [253], person search [65], [135], [211], [219], pose estimation [21], [133], [137], [166], and face detection [100], [130], [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 computer vision. 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% 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% 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 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.

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 [117], [118] do not have the step of proposal generation, while the recent PED
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 [69], [147], objectness methods [28], [69], [184], [266], 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], [116], [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 [113]) 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], [92], [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
TABLE 2
Summary of 21 typical handcrafted features based methods for pedestrian detection. 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.

| Method     | Publication | Family | Proposal | Feature | Classifier | Post-proc. | Scale-avails. | Part-avails. | Context | Description |
|------------|-------------|--------|----------|---------|------------|------------|--------------|-------------|---------|-------------|
| VJ [38]   | ICCV2004    | CF     | SW       | RichHarr| boosting   | NMS        | no           | no          | no      | a robust real-time face detector with Haar features |
| HOG-LBP [38] | CVPR2005    | DPM    | SW       | HOG     | SVM        | NMS        | no           | no          | no      | a novel histogram of gradient feature descriptor |
| CDNet [42] | BMVC2009    | CF     | SW       | CDNet   | boosting   | NMS        | no           | no          | no      | the simple and effective integral channel features |
| DPM [49]  | PAMI2010    | DPM    | SW       | HOG     | SVM        | NMS        | no           | yes         | yes     | a new feature by self-similarity of low-level features |
| HOC-CSS [188] | CVPR2010   | DPM    | SW       | HOG-CSS | SVM        | NMS        | no           | no          | no      | a multisolution model based on DPM & HOG |
| MultiBosC [135] | ECCV2010  | DPM    | SW       | HOG     | SVM        | NMS        | yes          | yes         | ground plane |
| VeryFast [4] | CVPR2012   | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | yes         | ped/car relation |
| MT-DPM [218] | CVPR2013   | DPM    | SW       | HOG     | SVM        | NMS        | no           | no          | no      | mapping ped. of various scales to a common space |
| M2T [151] | CVPR2013    | CF     | SW       | ChnFtrs | SVM        | NMS        | no           | yes         | no      | remove camera motion and object motion |
| SquaresChnFtrs [3] | CVPR2013 | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | use square features to reduce randomness |
| Franken [129] | ICCV2013   | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | yes         | no      | a fast training of many occlusion-specific classifiers |
| ACF [129]  | ICCV2013    | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | aggregate local features by downsmoothing operation |
| InformedHarr [229] | CVPR2014 | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | yes         | no      | local ternary features based on pedestrian shape |
| LDIF [131] | NIPS2014    | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | remove correlations in local neighborhoods |
| 2Pd [132] | PAMI2015    | DPM    | SW       | HOG     | SVM        | NMS        | no           | yes         | no      | spatial configuration patterns of nearby pedestrians |
| FCN [243] | CVPR2015    | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | construct a filtered channel framework |
| SpatialPooling [146] | PAMI2016 | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | extract the features based on spatial pooling |
| SCL [17]  | CVPR2016    | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | add segmentation features as additional channels |
| NNNF [17] | CVPR2016    | CF     | SW       | ChnFtrs | boosting   | NMS        | no           | no          | no      | non-neighbouring features based on inner attributes |

Fig. 5. Two different classes of single-spectral pedestrian detection approaches: handcrafted features based and deep features based methods. We further categorize the handcrafted methods based 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.

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 [5, 42], [48], [244] employ this kind of 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], [92], [227], a variety of handcrafted feature descriptors, including SIFT [122], LBP [136], 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], [131], [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 fea-
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], [131], [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. [131] 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 [131]) 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. [119] designed pixel neighborhood differential features for pedestrian detection. Li et al. [102] 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. [263] proposed additional high-level semantic features by 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
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 Tesema 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. [96] 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 joint different tasks (i.e., feature extraction, deformation handling, occlusion handling, and classification). Luo et al. [126] proposed to automatically learn 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.

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], [92], semantic segmentation [120], [125], [150], and object detection [56], [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 [138], [139], [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. 7(a)), whereas some other methods use handcrafted methods for proposal generation and CNN features for proposal classification (i.e., CNN for proposal classification in Fig. 7(b)). These hybrid approaches share some common
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 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. 7(c) 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. 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 from proposals from different layers (called scale-aware RoI). Yang et al. [223] and Liu et al. [114] 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. [265] flexibly chose the RoI features for regression and combined the RoI features from multiple layers for classification. SAF-RCNN [98] 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. [109], [110] 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. [107] built two-branch networks for head-shoulder and full body predictions and introduced an adaptive fusion mechanism. Lu et al. [124] 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. [267] 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. [105], [106] 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 preprocessing 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. [115] 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. [117] 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. [118] 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. [127] 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. [299] 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. [281] 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. [283] transformed the LR feature space into a new LR classification space using an optimal Mahalanobis 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. [113] 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 R2NMS 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. [128] 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. [90] 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. [88] 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. [97] proposed to use the depthwise 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 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. [112] 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. [91] 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. [95] 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. [262] 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 multimodal re-weighted module was further introduced to generate reliable features. To generate more diversified proposals and learn better features, Li et al. [94] 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 without manual annotations, Cao et al. [38] designed a cross-modality learning 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 image and infrared image. Afterwards, features extracted from the cross-modality network are fused with the features extracted from the traditional detection network.

5 Dataset and evaluation

Pedestrian datasets can be split into two classes: earlier pedestrian datasets and modern pedestrian datasets. Additionally, some multispectral pedestrian datasets have been also built. We first provide an introduction to these datasets. Then, we discuss different evaluation metrics. Finally we give a detailed comparison and analysis on these datasets.

5.1 Earlier pedestrian datasets

The earlier datasets in Table 5 (top) are relatively small and mainly used by handcrafted features based methods.

MIT [152] is one of the first pedestrian datasets. The training set contains 924 positive samples and 11,361 negative samples, where the resolution of samples is of 128×64 pixels. There are 123 images for testing in this dataset.

INRIA [38] is one of the most popular pedestrian datasets for handcrafted features based methods. This dataset mainly contains personal digital images. The training set contains 614 positive images with 1,208 pedestrians and 1,218 negative images. The test set has 288 images.

ETH [46] is taken on different days in busy shopping streets. The dataset contains three subsets for evaluation (i.e., ‘BAHNHOF’, ‘JELMOLI’, and ‘SUNNY DAY’). The ‘BAHNHOF’ set has 999 images, the ‘JELMOLI’ set has 450 images, and the ‘SUNNY DAY’ has 354 images.

TUD-Brussels [203] is recorded from a driving car in the city of Brussels and contains 508 images at a fixed resolution of 650×480 pixels. Additionally, this dataset contains temporal image pairs to learn motion features.

DAIMLER [45] is a large-scale pedestrian dataset recorded at various times of day. There are 6,755 images for training and 21,790 images for testing. Different from other datasets, this dataset consists of gray images instead of color images.

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 [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 testing. Generally, the training images are captured by every 3rd 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.

KITTI2 [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 1,240×376 pixels.

CityPersons [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

| Name | Publication | #Images | #Pedestrians | Resolutions | Annotations | Time | Description |
|------|-------------|---------|--------------|-------------|-------------|------|-------------|
| MIT  | IJCV2000    | -       | 924          | 64×128      | full        | day  | one of earliest pedestrian datasets |
| INRIA| ICCV2005    | 2120    | 1774         | 640×480     | full        | day  | one of earliest popular pedestrian datasets |
| ETH  | ICCV2007    | 1803    | 12k          | 640×480     | full        | day  | a pair of images in busy shopping streets |
| CityPersons | CVPR2009 | 508     | 1326         | 640×480     | full        | day  | pedestrians in the inner city of Brussels |
| Daimler | PAMI2009 | 29k     | 72k          | 640×480     | full        | day  | gray-color images in urban traffic |
| Caltech | PAMI2010 | 250k    | 289k         | 640×480     | full, visible | day  | a standard and complete pedestrian datasets |
| KITTI | CVPR2012   | 15k     | 9k           | 1240×376    | full        | day  | a real-world computer vision benchmarks |
| CityPersons | CVPR2017 | 5k      | 32k          | 2048×1024   | full, visible | day  | extensions on top of the Cityscapes [33] |
| CrowdHuman | axriv2018 | 24k     | 552k         | -           | full, visible, head | day, night | humans in crowded scenes from website |
| EuroCity | PAMI2019 | 47k     | 219k         | 1920×1024   | full        | day, night | images in multiple European Cities |
| NightOwl | ACCV2019 | 28k     | 56k          | 1024×640    | full        | night | pedestrians at night in three countries |
| WIDER Pedestrian | TMM2019 | 13k     | 39k          | -           | full        | day, night | pedestrians in traffic and surveillance scenes |
| TJU-Pedestrian | TIP2020 | 75k     | 373k         | -           | full, visible | day, night | a diverse dataset in traffic and campus |
| KAIST | Sensors2016 | 95k    | 103k         | 640×480     | full        | day, night | color-thermal image pairs in traffic scene |
| CVC-14 | Sensors2016 | 5051   | 7795         | 640×512     | full        | day, night | multimodal (FIR-visible) videosequences |
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 cities. It contains three different person categories (i.e., pedestrians, cyclists, and other riders). There are 47,337 images in total at a resolution of 1,920×1,080 pixels.

NightOwls \(^6\) [132] is a pedestrian dataset recorded at nighttime, which aims to promote progress in pedestrian detection at night. There are about 128 \(k\) training images, 51\(k\) validation images, and 103 \(k\) 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 400\(k\) 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 specifically 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 was recently introduced for crowded pedestrian detection. Before discussing these evaluation metrics in more 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

$$O = \frac{B_d \cup B_g}{B_d \cap B_g}. \quad (1)$$

If the overlap \(O\) is larger than a threshold of \(\alpha\), 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 strategy in Algorithm 1 splits the detection results into true positives and false positives along with the generated false negatives (missed positives). The threshold \(\alpha\) is usually set to be 0.5. Based on true positives, false positives, and false negatives on the whole test set, log-average miss rate and average precision can be calculated to compare the detector performance.

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

$$MR = 1 - \frac{N_{tp}}{N_g}, \quad (2)$$

and false positives per image (FPPI) can be calculated by dividing false positives by the number of images. By varying detection confidence threshold, miss rates against false positives per image (FPPI) can be plotted in log-space. Finally, the log-average miss rate MR is calculated
by averaging miss rates under 11 FPPI equally spaced in [10^{-2}, 10^{1}]. 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% 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% 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% 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

\[ R = \frac{N_{tp}}{N_{g}}. \]

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

\[ P = \frac{N_{tp}}{N_{d}}. \]

By varying the threshold of confidence score, precision against recall can be plotted as a curve. Based on the precision-recall curve, the average 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.

**Jaccard index** The recently proposed Jaccard index (JI) represents the overlap between detection results and ground-truths [171], which is used to evaluate the performance in crowded scenes. The Jaccard index score \( S_{JI} \) under a given confidence threshold can be computed as:

\[ S_{JI} = \frac{\text{IoU}(D, G)}{|D| + |G| - \text{IoU}(D, G)}, \]

where \( D \) represents the set of detection results, and \( G \) is the set of ground-truths. \( M_{D}, M_{G} = \text{IoU}(D, G) \), where \( M_{D} \) and \( M_{G} \) are maximum matching detection sets and ground truth sets, respectively. The IoU match is computed using the Hungarian algorithm. When increasing the confidence threshold, the JI score first increases and then decreases. To generate the best JI score for a detector, the greedy searching algorithm is used to compare the JI scores under different confidence thresholds. If the detector has a larger best JI score, the detector has a better performance.

### 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 [131], Katamari [6], SCCPriors [225], and Checkboarders [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 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 Med 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 pedestrian.
As a result, the Hard set provides more than 10% 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 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 R²NMS [78]) using part information (e.g., visible and head annotations) have a better performance. Table 9 further shows miss-rates of 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, GFDFL [105] and ALFNet [117] are superior in terms of speed on Caltech test set, while PRNet [178] and CSP [118] 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 10 further compares some state-of-the-art methods on multispectral KAIST R test set [79], including hybrid methods and pure CNN based methods. With illumination aware feature alignment and single-stage structure, MBNet [262] achieves the best performance on both speed and accuracy.

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 11. We perform the 1x 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% occlusion are used for training. The original FPN achieves 18.11% MR on R set. By ignoring the proposals located in ignored regions, it provides 3.30% improvement on R set. When further setting the aspect ratio as 0.41 and uniform scale, it provides 0.55% improvement. With these simple modifications 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 12 shows the impact of extra-class training on pedestrian detection. Experiments are performed on two datasets, including TJU-Ped-Traffic [148] and MS COCO [111]. TJU-Ped-Traffic [148] has 5 object classes, while MS COCO [111] 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 13. The first (top) half of Table 13 shows the impact of different intra-dataset data processing strategies. (1) Instead of the pedestrians over 20 pixels in height with less than 80% occlusion for training, we use the pedestrians over 50 pixels in height with less than 65% 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 13 shows the impact of pre-training on different different datasets, including Caltech [43], MS COCO [111], 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% 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 14, 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 might be 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 15. It is difficult to achieve the optimal results on all the sets by using a single threshold \( \theta \). 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 16. 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 [118]. Furthermore, we analyze the 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

Despite this great success, pedestrian detection still faces several challenges discussed as follows.
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 ideas 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 is necessary to treat small-scale pedestrian detection as a standalone problem.

6.2 Occlusion

Pedestrian occlusion is a very common problem. For instance, 70% 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 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 solve this problem, one solution is the improved NMS strategy, for example, dynamic NMS threshold. Another solution is NMS-free strategy by avoiding bounding-box combination operation.

Besides, temporal information is important for occluded pedestrian detection in driving and surveillance scenes.

6.3 Domain adaptation

Most existing methods focus on a specific pedestrian dataset and cannot guarantee a good domain adaptation ability [195], [212]. For instance, the pedestrian 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], [87], [89], [121], feature-level [168], [215], and instance-level [24], [264] methods. However, they mainly focus on generic object detection. In future, the domain adaptation methods about pedestrian detection can be exploitsd by considering pedestrian/scene characteristic.

Besides, we can exploit both the same-domain and cross-domain evaluations in future, which guides the researchers to design the pedestrian detector with good comprehensive performance.

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 have 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 past decade, pedestrian detection has witnessed significant success, which has gone from the handcrafted features 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 (e.g., occlusion, scale variance, and domain adaptation) in pedestrian detection. We hope that this deep survey can help the researchers to develop new methods in the field of pedestrian detection.

### References

[1] Anelia Angelova, Alex Krizhevsky, Vincent VanhouckeI, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. Proc. British Machine Vision Conference, 2015.

[2] Jeonghyun Baek, Junhyuk Hyun, and Euntai Kim. A pedestrian detection system accelerated by kernelized proposals. IEEE Transactions on Intelligent Transportation Systems, 21(3):1216–1226, 2020.

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. Proc. European Conference on Computer Vision, 2006.

[4] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[5] Rodrigo Benenson, Markus Mathias, Tinne Tuytelaars, and Luc Van Gool. Seeking the strongest rigid detector. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[6] Rodrigo Benenson, Mohammad Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? Proc. European Conference on Computer Vision, 2014.

[7] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms—improving object detection with one line of code. Proc. IEEE International Conference on Computer Vision, 2017.

[8] Markus Braun, Sebastian Krebs, Fabian Flohr, and Daria M. Gavrila. Euroc people: A novel benchmark for person detection in traffic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8):1844–1861, 2019.

[9] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[10] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection and segmentation. Proc. IEEE International Conference on Computer Vision, 2017.

[11] Zhaowei Cai, Quanfu Fan, Rogerio S. Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. Proc. European Conference on Computer Vision, 2016.

[12] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. Proc. IEEE International Conference on Computer Vision, 2015.

[13] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[14] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[15] Jiale Cao, Yanwei Pang, Jungong Han, Bolin Gao, and Xuelong Li. Taking a look at small-scale pedestrians and occluded pedes-

[16] Jiale Cao, Yanwei Pang, and Xuelong Li. Learning multilayer channel features for pedestrian detection. IEEE Transactions on Image Processing, 26(7):3210–3220, 2016.

[17] Jiale Cao, Yanwei Pang, and Xuelong Li. Pedestrian detection inspired by appearance constancy and shape symmetry. IEEE Transactions on Image Processing, 25(12):5538–5551, 2016.

[18] Jiale Cao, Yanwei Pang, Shengjie Zhao, and Xuelong Li. High-level semantic networks for multi-scale object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[19] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yin, Yanlong Cao, and Yu Qiao. Pedestrian detection with unsupervised multispectral feature learning using deep neural networks. Information Fusion, 46:206–217, 2019.

[20] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. ISPRS Journal of photogrammetry and remote sensing, 150:70–79, 2019.

[21] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017.

[22] Rui Chen, Haizhou Ai, Chong Shang, Long Chen, and Zijie Zhuang. Learning lightweight pedestrian detector with hierarchical knowledge distillation. Proc. IEEE International Conference on Image Processing, 2019.

[23] Tao Chen, Shijian Lu, and Jiayuan Fan. S-cnn: Subcategory-aware convolutional networks for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(10):2522–2528, 2018.

[24] Yuhua Chen, Wen Li, Christos Sakaridis, Dangxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[25] Zhila Chen and Ximing Huang. Pedestrian detection for autonomous vehicle using multi-spectral cameras. IEEE Transactions on Intelligent Vehicles, 4(2):211–219, 2019.

[26] Zhe Chen, Wanli Ouyang, Tongliang Liu, and Dacheng Tao. A shape transformation-based dataset augmentation framework for pedestrian detection. arXiv:1912.0701, 2019.

[27] Zhichang Chen, Li Zhang, Abdul Mateen Khattak, Wanlin Gao, and Minjuang Wang. Deep feature fusion by competitive attention for object detection. IEEE Access, 7:21981–21989, 2019.

[28] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[29] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. Proc. AAAI Conference on Artificial Intelligence, 2020.

[30] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Li, and Xudong Zou. Relational learning for joint head and human detection. Proc. AAAI Conference on Artificial Intelligence, 2020.

[31] Jie Chu, Zhixian Gao, and Lu Long. Object detection based on multi-layer convolution feature fusion and online hard example mining. IEEE Access, 6:19959–19967, 2018.

[32] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.
d based template matching pedestrian detection. Proc. IEEE International Conference on Robotics and Automation, 2016.

[81] Xiaoheng Jiang, Yanwei Pang, Xuelong Li, and Jing Pan. Speed up deep neural network based pedestrian detection by sharing features across multi-scale models. Neurocomputing, 185:163–170, 2016.

[82] Xiaoheng Jiang, Yanwei Pang, Manli Sun, and Xuelong Li. Cascaded subpatch networks for effective cnns. IEEE Transactions on Neural Networks and Learning Systems, 29(7):2684–2694, 2017.

[83] Sang-II Jung and Ki-Sang Hong. Deep network aided by guiding network for pedestrian detection. Pattern Recognition Letters, 90:43–49, 2017.

[84] Sang-II Jung and Ki-Sang Hong. Direct multi-scale dual-stream network for pedestrian detection. Proc. IEEE International Conference on Image Processing, 2017.

[85] Srinu Kavanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia I. Camps, and Richard J. Radke. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3):523–536, 2020.

[86] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D Bagdanov, Maria Vanrell, and Antonio M Lopez. Color attributes for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[87] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and Wael Scemama. A robust learning approach to domain adaptive object detection. Proc. IEEE International Conference on Computer Vision, 2019.

[88] My Kieu, Andrew D. Bagdanov, Marco Bertini, and Alberto del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. Proc. European Conference on Computer Vision, 2020.

[89] Taeyoung Kim, Minki Jeong, Seunghyeon Kim, Seokoon Choi, and Changik Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[90] Perla Sai Raj Kishore, Sudip Das, Partha Sarathi Mukherjee, and Uijwal Bhattacharya. Cluenet: A deep framework for occluded pedestrian pose estimation. Proc. British Machine Vision Conference, 2019.

[91] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2017.

[92] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. Proc. Advances in Neural Information Processing Systems, 2012.

[93] Jason Ku, Alex O. Pon, and Steven L. Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[94] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. Proc. British Machine Vision Conference, 2017.

[95] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition, 85:161–171, 2019.

[96] Chao Li, Xinggang Wang, and Wenyu Liu. Neural features for pedestrian detection. Neurocomputing, 289:420–432, 2017.

[97] Guofo Li, Yifan Yang, and Xingda Qu. Deep learning approaches for pedestrian detection in hazy weather. IEEE Transactions on Industrial Electronics, 67(10):8889–8999, 2019.

[98] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingta Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. IEEE Transactions on Multimedia, 20(4):985–996, 2018.

[99] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingta Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[100] Jian Li, Xabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[101] Kunquan Li, Shengyin Jin, and Junjie Yan. Mimicking very efficient network for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[102] Qiming Li, Hanzi Wang, Yan Yan, Bo Li, and Chang Wen Chen. Local co-occurrence selection via partial least squares for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems, 18(6):3046–3055, 2017.

[103] Zxin Zhao Li, Yuehu Liu, Zeqi Chen, Jiahuan Zhou, and Ying Wu. Fused discriminative metric learning for low resolution pedestrian detection. Proc. IEEE International Conference on Image Processing, 2018.

[104] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[105] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. Proc. European Conference on Computer Vision, 2018.

[106] Chunze Lin, Jiwen Lu, and Jie Zhou. Multi-grained deep feature learning for robust pedestrian detection. IEEE Transactions on Circuits and Systems for Video Technology, 29(12):3608–3621, 2019.

[107] Chih-Yang Lin, Hong-Xia Xie, and Hua Zheng. Pedjointnet: Joint head-shoulder and full body deep network for pedestrian detection. IEEE Access, 7:47687–47697, 2019.

[108] Matthieiu Lin, Chuming Li, Xingyuan Bu, Ming Sun, Chen Lin, Junjie Yan, Wanli Ouyang, and Zhidong Deng. Detr for pedestrian detection. arXiv:2012.06783, 2020.

[109] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[110] Tsung-Yi Lin, Priya Goyal, Ross B Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2):318–327, 2020.

[111] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. Proc. European Conf. Computer Vision, 2014.

[112] Jilong Liu, Shaojie Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. Proc. British Machine Vision Conference, 2016.

[113] Yang Liu, Shengcai Liao, Weijiang He, Chunmei Li, and Jing Pan. Efficient network for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[114] Tianrui Liu, Mohamed Elmikaty, and Tania Stathaki. Sam-rcnn: Scale-aware multi-resolution multi-channel pedestrian detection. Proc. British Machine Vision Conference, 2018.

[115] Tianrui Liu, Wenhan Luo, Lin Ma, Jun-Jie Huang, Tania Stathaki, and Tianhong Dai. Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling. arXiv:1912.08661, 2019.

[116] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. Proc. European Conference on Computer Vision, 2016.

[117] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. Proc. European Conference on Computer Vision, 2018.

[118] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[119] Xing Liu, Kar-Ann Toh, and Jan P. Allebach. Pedestrian detection using pixel difference matrix projection. IEEE Transactions on Intelligent Transportation Systems, 21(4):1441–1454, 2020.

[120] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[121] Adrian Lopez, Rodriguez, and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. Proc. British Machine Vision Conference, 2019.

[122] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60:91–110, 2004.

[123] Ruiqi Lu and Huimin Ma. Occluded pedestrian detection with visible iou and box sign predictor. Proc. IEEE International Conference on Image Processing, 2019.

[124] Ruiqi Lu and Huimin Ma. Semantic head enhanced pedestrian detection in a crowd. arXiv:1911.11985, 2019.

[125] Xiankai Liu, Wenguang Wang, Chao Ma, Jinjun Qian, Sheng Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2019.

[126] Ping Luo, Yonglong Tian, Xiaogang Wang, and Xiaoou Tang. Switchable deep network for pedestrian detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[127] Yan Luo, Chongyang Zhang, Muming Zhao, Hao Zhou, and Jun Sun. Where, what, whether: Multi-modal learning meets pedestrian detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[128] Xun Liu, Miao Cao, Yujing Jiang, and Zhimin Cao. What can help pedestrian detection? *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[129] Markus Mathias, Rodrigo Benenson, Radu Timofte, and LucVan Gool. Handling occlusions with franken-classifiers. *Proc. IEEE International Conference on Computer Vision*, 2013.

[130] Vipin Naik, Bharat Singh, and Larry S. Davis. Fa-rpn: Floating region proposals for face detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[131] Woohyun Nam, Piotr Döllar, and Joon Hee Han. Local decorrelation for improved pedestrian detection. *Proc. Advances in Neural Information Processing Systems*, 2013.

[132] Lukas Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, and Bernt Schiele. Nightowl: A pedestrians at night dataset. *Proc. Asian Conference on Computer Vision*, 2018.

[133] Alejandro Newell, authorKaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *Proc. European Conference on Computer Vision*, 2016.

[134] Junhyug Noh, Sooehun Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[135] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Person recognition in personal photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):203–216, 2020.

[136] Timo Ojala, Matti Pietikäinen, and Topi Maenpää. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2004.

[137] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[138] Wanli Ouyang and Xiaogang Wang. A discriminative deep model for pedestrian detection with occlusion handling. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[139] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. *Proc. IEEE International Conference on Computer Vision*, 2013.

[140] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[141] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Single-pedestrian detection aided by multi-pedestrian detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[142] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Single-pedestrian detection aided by two-pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1875–1889, 2015.

[143] Wanli Ouyang, Xingyu Zeng, Xiaogang Wang, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Hongyang Li, Kun Wang, Junjie Yan, Chen Change Loy, and Xiaou Tang. Deeponet: Object detection with deformable part based convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1320–1334, 2017.

[144] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1874–1887, 2018.

[145] Sakrapee Paitisriangkrai, Chunhua Shen, and Anton van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. *Proc. European Conference on Computer Vision*, 2014.

[146] Sakrapee Paitisriangkrai, Chunhua Shen, and Anton van den Hengel. Pedestrian detection with spatially pooled features and structured ensemble learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1243–1257, 2016.

[147] Yanwei Pang, Jiale Cao, and Xuelong Li. Learning sampling distributions for efficient object detection. *IEEE Transactions on Cybernetics*, 47(1):117–129, 2017.

[148] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanging Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 2020.

[149] Yanwei Pang, Jiale Cao, Jian Wang, and Jungong Han. Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. *IEEE Transactions on Information Forensics and Security*, 14(12):3322–3331, 2019.

[150] Yanwei Pang, and Ling Shao Jie Sheng. Towards bridging semantic gap to improve semantic segmentation. *Proc. IEEE International Conference on Computer Vision*, 2019.

[151] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. *Proc. IEEE International Conference on Computer Vision*, 2019.

[152] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 30(1):15–33, 2000.

[153] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. *Proc. European Conference on Computer Vision*, 2010.

[154] Dennis Park, Larry Zitnick, Deva Ramanan, and Piotr Döllar. Exploring weak stabilization for motion feature extraction. *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[155] Kihong Park, Seungyong Kim, and Kwanghoon Sohn. unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 80:143–155, 2018.

[156] Xishuai Peng, Yi Murphey, Simon Stent, Yuanni Li, and Zihao Zhao. Spatial focal loss for pedestrian detection in fisheye imagery. *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2019.

[157] Yeqiang Qian, Ming Yang, Xu Zhao, Chunxiang Wang, and Bing Wang. Oriented spatial transformer network for pedestrian detection using fish-eye camera. *IEEE Transactions on Multimedia*, 22(2):431–431, 2020.

[158] Bijia Radosevic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Döllar. Designing network design spaces. arXiv:2003.13678, 2020.

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

[160] Rakesh Nattoji Rajaram, Eshed Ohn-Bar, and Mohan Manubhai Trivedi. Looking at pedestrians at different scales: A multiresolution approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3565–3576, 2016.

[161] Rajeev Ranjan, Vishal M. Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2020.

[162] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[163] Jimmy Ren, Xiaohao Chen, Jianbio Liu, Wenxiu Sun, Jiahao Pang, Qiong Yan, Yu-Wing Tai, and Li Xu. Accurate single stage detector using recurrent rolling convolution. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[164] Shaoqin Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. Advances in Neural Information Processing Systems*, 2015.

[165] David Ribeiro, Jacinto C. Nascimento, and Gustavo Carneiro. Improving the performance of pedestrian detectors using convolutional learning. *Pattern Recognition*, 61:641–649, 2017.

[166] Grégoire Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcrnet++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1146–1161, 2020.

[167] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[168] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[169] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[170] Chong Shang, Haizhou Ai, Zijie Zhuang, Long Chen, and Junliang Xing. Zoomnet: Deep aggregation learning for high-
performance small pedestrian detection. Proc. Asian Conference on Machine Learning, 2018.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. arXiv:1805.00123, 2018.

Jifeng Shen, Xin Zuo, Jun Li, WanKong Yang, and Haibin Ling. A novel pixel neighborhood differential statistic feature for pedestrian and face detection. Pattern Recognition, 63:127–138, 2017.

Jifeng Shen, Xin Zuo, WanKong Yang, Danil Prokhorov, Xue Mei, and Arthur Lin. Differential feature for pedestrian detection: A taylor series perspective. IEEE Transactions on Intelligent Transportation Systems, 20(8):2913–2922, 2018.

Jifeng Shen, Xin Zuo, Lei Zhi, Jun Li, WanKong Yang, Baochang Zhang, and Changxin Sun. Filtered shallow-deep feature channels for pedestrian detection. Neurocomputing, 249:19–27, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

Tao Song, Lei Yu, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. Proc. European Conference on Computer Vision, 2018.

Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. Proc. European Conference on Computer Vision, 2020.

Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

Fiseha B. Tesema, Hong Wu, Mingjian Chen, Junpeng Lin, Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Crowdhuman: A benchmark for detecting human in a crowd. arXiv:1805.00123,

Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. arXiv:1805.00123, 2018.

Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.

Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[263] Chao Zhu and Yuxin Peng. Discriminative latent semantic feature learning for pedestrian detection. *Neurocomputing*, 238:126–136, 2017.

[264] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[265] Yousong Zhu, Jinqiao Wang, Chaoyang Zhao, Haiyun Guo, and Hanqing Lu. Scale-adaptive deconvolutional regression network for pedestrian detection. *Proc. Asian Conference on Computer Vision*, 2016.

[266] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. *Proc. European Conference on Computer Vision*, 2014.

[267] Tengtao Zou, Shangming Yang, Yun Zhang, and Mao Ye. Attention guided neural network models for occluded pedestrian detection. *Pattern Recognition Letters*, 131:91–97, 2020.

Jiale Cao received the Ph.D. degree in information and communication engineering from the Tianjin University, Tianjin, China, in 2018. He is currently an associate professor at the Tianjin University. His research interests include object detection and deep learning, in which he has published 15 paper in CVPR, ICCV, ECCV, IEEE TIP, IEEE TCSVT and IEEE TIFS.

Yanwei Pang received the Ph.D. degree in electronic engineering from the University of Science and Technology of China. Currently, he is a professor at the Tianjin University and is also the founding director of the Tianjin Key Laboratory of Brain Inspired Intelligence Technology (BIIT). He has published 150 scientific papers including 70 papers in top journals/conferences. He is an associate editor of both IEEE T-NNLS and Neural Networks (Elsevier).

Jin Xie received the B.S. degree in electronic engineering from the Tianjin University, Tianjin, China, in 2016. He is currently pursuing the Ph.D. degree in the Tianjin University and his supervisor is Prof. Yanwei Pang. His research interests include machine learning and computer vision, in which he has published 5 paper in CVPR, ICCV, ECCV, IEEE TCSVT, IEEE TCyb.

Fahad Shahbaz Khan received the Ph.D. degree in Computer Vision from Computer Vision Center Barcelona and Autonomous University of Barcelona, Spain. He is a faculty member at 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 is the CEO and the Chief Scientist of the Inception Institute of Artificial Intelligence (IIAI), 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.