JCS-Net: Joint Classification and Super-Resolution Network for Small-scale Pedestrian Detection in Surveillance Images

Yanwei Pang, *Senior Member, IEEE,* Jiale Cao, Jian Wang, and Jungong Han

***Abstract*—While Convolutional Neural Network (CNN)-based pedestrian detection methods have proven to be successful in various applications, detecting small-scale pedestrian from surveillance images is still challenging.The major reason is that the small-scale pedestrians lack much detailed information compared to the large-scale pedestrians. To solve this problem, we propose to utilize the relationship between the large-scale pedestrians and the corresponding small-scale pedestrians to help recover the detailed information of the small-scale pedestrians, thus improving the performance of detecting small-scale pedestrians. Specifically, a unified network (called JCS-Net) is proposed for small-scale pedestrian detection, which integrates the classification task and the super-resolution task in a unified framework. As a result, the super-resolution and classification are fully engaged and the super-resolution sub-network can recover some useful detailed information for the subsequent classification. Based on HOG+LUV and JCS-Net, multi-layer channel features (MCF) are constructed to train the detector. Experimental results on the Caltech pedestrian dataset and the KITTI benchmark demonstrate the effectiveness of the proposed method. To further enhance the detection, multi-scale MCF based on JCS-Net for pedestrian detection is also proposed, which achieves the state-of-the-art performance.**

***Index Terms*—Pedestrian Detection, Small-Scale, Large-Scale, Classification, Super-Resolution, MCF.**

I. Introduction

Pedestrian detection based on Convolutional Neural network (i.e., CNN) has achieved great success [54], [49], [8], [48] in various applications, including traffic monitoring [1], crowd event analysis [2], suspicious behavior detection [3] and human (re-) identification [4]. Because the distances of pedestrians from the camera are very different, the scales of pedestrians can arbitrarily be varied and thus the appearances of pedestrians differ across different scales. To carry out pedestrian detection in image, tremendous efforts have been paid, which can be divided into two main streams depending on whether the scale variation is addressed: (1) scale-agnostic methods [11], [53], [28] and (2) scale-aware methods [36], [10], [59].

This work was supported by the National Natural Science Foundation of China (No. 61632018 and No. 61773301), Postdoctoral Program for Innovative Talents (No. BX20180214), China Postdoctoral Science Foundation (No. 2018M641647), and the Nokia. (Corresponding author: Jungong Han.)

Y. Pang, J. Cao, and J. Wang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China. (E-mail: {pyw,connor,jianwang}@tju.edu.cn).

J. Han is with WMG Data Science Group at University of Warwick, UK, E-mail: jungonghan77@gmail.com.

In scale-agnostic methods, the pedestrians of different scales are seen as the same category. In the training process, pedestrians of all the different scales are rescaled to a fixed size (e.g., 128×64) to train one detector. To detect the pedestrians of different scales in the test image, the image pyramid technique is used to rescale the image at multiple sizes. After that, the trained detector respectively scans the images of different sizes. In fact, the pedestrians of different scales are dissimilar in appearance patterns [36], which has been ignored by scale-agnostic methods. For example, the large-scale pedestrians have the rich texture while the small-scale pedestrians are often blurry.

To solve the above problem existed in scale-agnostic methods, scale-aware methods have been proposed for pedestrian detection. They treat pedestrians across scales as the different sub-categories, based on which multiple scale-specific detectors are respectively trained. For example, Yang *et al.* [59] proposed the Scale-Dependent Pooling (SDP) to handle the scale variation problem for object detection. According to the heights of pedestrians, SDP trains multiple ROI pooling models based on different convolutional layers. Specifically, if the height of a pedestrian is small, the ROI pooling will be from the feature maps of early convolutional layer. To solve the inconsistency between the scales of objects and the sizes of filter receptive fields, Cai *et al.* [10] proposed multi-scale CNN (called MSCNN). It extracts the object proposals from multiple different layers, where each layer focuses on the certain scales of pedestrians. Despite their success, the key idea of these methods above is to train multiple scale-specific detectors, and the relationship between the pedestrians of different scales is not fully utilized. Moreover, small-scale pedestrian detection still does not perform well.

In this paper, we propose to utilize the relationship between the pedestrians of different scales to help improve the performance of small-scale pedestrian detection. Firstly, a super-resolution sub-network is trained, given the paired pedestrians (i.e., the large-scale pedestrians and their corresponding small-scale pedestrians). Secondly, a classification sub-network (i.e., VGG16 [50]) is fine-tuned on the large-scale pedestrians. Thirdly, the two sub-networks are integrated into one unified network by Joint the Classification loss and the Super-resolution loss, which is called JCS-Net. Based on HOG+LUV [19] and JCS-Net, multi-layer channel features (MCF) [12] are constructed for small-scale pedestrian detection. To further improve the detection performance, multi-scale MCF is proposed. The contributions and characteristics of this paper

are summarized as follows:

(1) JCS-Net is proposed for small-scale pedestrian detection. By jointly optimizing classification and super-resolution, JCS-Net makes full use of the relationship between the large-scale pedestrians and the small-scale pedestrians to help small-scale pedestrian detection. Based on HOG+LUV and JCS-Net, multi-layer channel features (MCF) are constructed to train the detector for small-scale pedestrian detection.

(2) To solve the scale-variable problem, multi-scale MCF is proposed. Specifically, the detector for the small-scale pedestrians is trained based on HOG+LUV and JCS-Net, and the detector for the large-scale pedestrians is trained based on HOG+LUV and fine-tuned VGG16. The detection results of different detectors are further combined together. Compared to other scale-aware methods, multi-scale MCF not only treats the pedestrians of different scales as the different sub-categories, but also uses the relationship between the pedestrians of different scales for small-scale pedestrian detection.

(3) Experiments on the Caltech pedestrian dataset [20], [21] and the KITTI benchmark [24] show the effectiveness of our proposed method. On the Caltech pedestrian dataset, in particular, our method achieves the state-of-the-art performance (i.e., 8.81% miss rate).

The rest of the paper is organized as follows. Firstly, we review pedestrian detection and the related super-resolution in Sec. II. Then, our proposed method is presented in Sec. III. After that, the experimental results are reported in Sec. IV. Finally, we conclude this paper in Sec. V.

II. Related works

In this section, we begin with a review of pedestrian detection, which is followed by a brief introduction about the related super-resolution.

A. Pedestrian detection

Pedestrian detection can mainly be divided into two paradigms: the handcrafted features based methods and the CNN based methods. Cascade AdaBoost detector based on Haar features is one of the most classical handcrafted features based methods [56]. Thanks to the cascade structure [45], it achieves the real-time detection speed with no loss of detection performance. Dollár *et al.* [19] proposed Integral Channel Features (ICF) for pedestrian detection. It generates the local sum features from multiple registered image channels (i.e., HOG [16] and LUV) along with the integral image trick. Following ICF [19], many methods based on channel features (e.g., ACF [18], InformedHaar [61], LDCF [42], Checkerboards [62], and NNNF [13]) have been proposed. By using fast feature pyramids and aggregate channel features, ACF [18] dramatically accelerates the detection speed. LDCF [62] and Checkerboards [62] convolve original image channels (i.e., HOG+LUV) with PCA-like filters and handcrafted filters to generate new image channels, respectively. InformedHaar [61] and NNNF [13] incorporate the statistical characteristics of pedestrians into the design of pedestrian features.

After the success of CNN on image classification [33], the CNN has been applied to many other visual tasks [6], [35], [39], [31], [44], where object detection is probably the most successful example [26], [43], [47], [51]. The most famous object recognition method is R-CNN [26], which carries out three steps - it firstly extracts region proposals by selective search [55], then computes the CNN features of these proposals, and finally classifies these proposals by class-specific linear SVMs. Faster R-CNN [47] integrates the above three steps of R-CNN into a unified and end-to-end network. Based on the famous CNN models [33], [50], many CNN based methods for pedestrian detection have been also proposed. Hosang *et al.* [28] made extensive experiments for deep pedestrian detection, where the handcrafted features based methods generate candidate proposals while CNN is employed to classify these proposals. By combing some local handcrafted features and the CNN features, Cai *et al.* [11] proposed the CompACT boosting algorithm for pedestrian detection that reaches a good trade-off of accuracy and computational complexity. Based on the CNN features, CCF [58] and RPN+BF [60] both used the decision forest model to learn the pedestrian detector. MCF [12] integrated the handcrafted image channels (i.e., HOG+LUV) and each layer of CNN into multi-layer image channels. Mao *et al.* [41] exploited semantic features to help pedestrian detection. Zhang *et al.* [63] proposed an attention mechanism for occluded pedestrian detection. Because these methods treat the pedestrians of different scales as the same category and do not consider the scale-variable problem of pedestrians, these methods are called scale-agnostic methods.

In fact, the scale-variable problem is one of the most important problems in pedestrian detection. On the one hand, the pedestrians of different scales have very different characteristics. For example, the large-scale pedestrians have very rich texture, while the small-scale pedestrians are very blurry. On the other hand, there exists the inconsistency between the receptive fields of the last CNN layer and the scales of pedestrians [10]. To solve the scale-variable problem, scale-aware methods have been proposed. Li *et al.* [36] proposed to use two sub-networks to respectively capture the characteristics of the large-scale and the small-scale pedestrians. To reduce computation cost, the two sub-networks share the first few convolutional layers. Yang *et al.* [59] trained multiple ROI pooling models from the outputs of different convolutional layers for pedestrians of different scales. Recently, Cai *et al.* [10] proposed multi-scale proposal network to generate the candidate proposals from multiple output layers. To further improve the detection performance, a detection sub-network is added after the multi-scale proposal network.

Compared to scale-agnostic methods, scale-aware methods generally perform better, especially on small-scale pedestrian detection. Even though scale-aware methods have been successful, we argue that there is still room for further improvement: (1) As most scale-aware methods only treat the pedestrians of different scales as the different sub-categories, the relationship between the pedestrians of different scales is ignored; (2) Because the small-scale pedestrians lose much useful information, detecting small-scale pedestrians is much harder than detecting large-scale pedestrians in image. In this

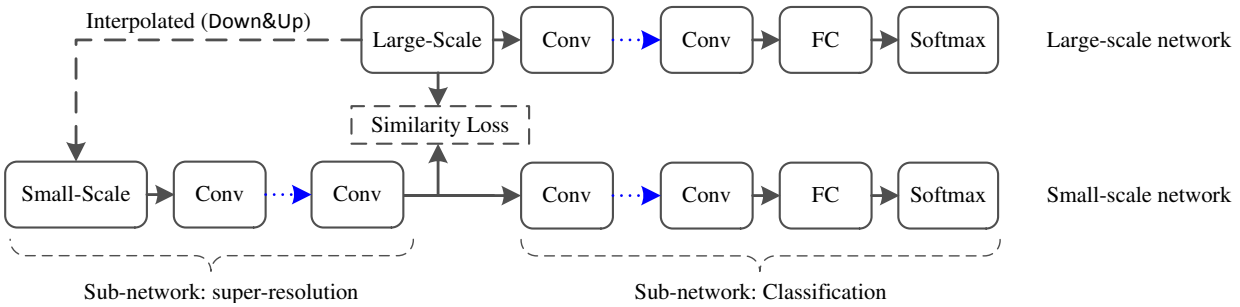

Fig. 1. The proposed JCS-Net for small-scale pedestrian detection by joint classification and super-resolution. Specifically, it consists of two sub-networks: a sub-network of super-resolution and a sub-network of classification. The classification sub-network is firstly initialized by the large-scale network which is trained on the large-scale pedestrians and then jointly trained with the super-resolution sub-network on the small-scale pedestrians. The blue arrows mean some undisplayed convolutional layers or pooling layers.

paper, we propose to use the large-scale pedestrian detection to aid small-scale pedestrian detection. Though the proposed JCS-Net is related to some following methods [46], [27], [5], [37], it is different from these methods in various aspects. Compared to [46], our proposed JCS-Net aims to recover some detailed information of small-scale pedestrians to help small-scale pedestrian detection. Though TDSR [27] indeed considers the super-resolution for object detection, the way of using super-resolution by TDSR is different from that by our JCS-Net. TDSR firstly trains the detection sub-network and then trains the super-resolution sub-network by freezing the network parameters of detection sub-network. We argue that the training process of freezing detection sub-network in TDSR is not optimal, because these two sub-networks are not fully engaged. In addition, the super-resolution in TDSR works on the whole image with no distinction of foreground and background. However, in practice, the area of background is usually much larger than that of foreground (pedestrian), hence we argue that the important foreground (pedestrian) cannot be reconstructed very well. Compared to TDSR, our JCS-Net only focuses on the important foreground (pedestrian) reconstruction. In [5], SOD-MTGAN uses two separate networks (generator network and discriminator network) to respectively generate the fine-scale image and recognize the specific object category. Compared to SOD-MTGAN, our proposed JCS-Net uses a single network for the fine-scale image generation and object classification. As a result, our proposed method is more efficient and less complicated in the training process. In [37], PGAN uses an extra sub-network to narrow down the ROI feature difference between the small-scale objects and the large-scale objects. Compared to PGAN, our proposed JCS-Net aims to narrow down the image difference between the small-scale objects and the large-scale objects. Namely, PGAN focuses on the feature domain, while our proposed JCS-Net focuses on the image domain. More importantly, the results show that our JCS-Net is superior to PGAN.

B. Image super-resolution

Image super-resolution aims to recover the high-resolution image from the low-resolution image. Recently, CNN based methods have achieved the great success on image super-resolution. Dong *et al.* [22] are the first to propose an end-to-end and fully convolutional neural network for image super-resolution. After that, many variants have been proposed [32], [64], [30], [52]. For example, Kim *et al.* [32] proposed a deep super-resolution network with the residual-learning and gradient clipping. Zhang *et al.* [64] proposed a residual dense network to utilize the hierarchical features for image super-resolution. Hui *et al.* [30] proposed a distillation block to gradually extract the abundant features to reconstruct the high-resolution image. In this paper, we use the image super-resolution technique to recover more detailed information of the small-scale pedestrians with the aid of the corresponding large-scale pedestrians.

III. Our Proposed Method

In this section, we start by presenting our proposed JCS-Net for small-scale pedestrian detection, and then show how to construct multi-layer channel features (MCF) from HOG+LUV and JCS-Net, and finally propose the multi-scale MCF.

A. JCS-Net

In this subsection, we propose to integrate the super-resolution sub-network and the classification sub-network together for small-scale pedestrian detection. Generally, the large-scale pedestrians have more abundant detailed information. Thus, the super-resolution sub-network aims to recover the detailed information of small-scale pedestrians from their large-scale counterparts. Moreover, by engaging the super-resolution and the classification, the reconstructed small-scale pedestrians are much more suitable for small-scale pedestrian detection.

First of all, we give an overall review of the proposed network for small-scale pedestrian detection in Fig. 1. The top row shows the network for large-scale pedestrian detection, and the bottom row shows the proposed network for the small-scale pedestrian detection by joint classification and super-resolution. For simplification, the proposed network for small-scale pedestrian detection is called JCS-Net. JCS-Net consists of two sub-networks: one sub-network of super-resolution and another sub-network of classification. The specific process of training JCS-Net is explained as follows: (1) Firstly, the

network (e.g., VGG16 [50]) for the large-scale pedestrians is fine-tuned on the large-scale pedestrians and negatives. (2) Secondly, the sub-network of the super-resolution in JCS-Net is pre-trained following the existing technique of image super-resolution [22], [32]. The small-scale pedestrians are generated by firstly downsampling the large-scale pedestrians and secondly upscaling them. (3) Thirdly, the sub-network of the classification in JCS-Net is initialized by the weights of network for large-scale pedestrian detection. (4) Finally, JCS-Net for small-scale pedestrian detection is trained by joining the sub-network of classification and the sub-network of super-resolution together. The loss of JCS-Net is the joint loss of the two sub-networks.

Assuming that $\mathbf{y}_i$ refers to the large-scale pedestrian, $\mathbf{x}_i$ refers to the corresponding small-scale pedestrian, and $F(\mathbf{x}_i)$ refers to the reconstructed pedestrian by the sub-network of super-resolution. Then, the loss of the super-resolution sub-network is expressed by mean squared error as follows:

$$L_{similarity} = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - F(\mathbf{x}_i)||^2, \tag{1}$$

where n is the number of the training positive samples. And the loss of the classification sub-network is expressed as

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} -\log p_c(\mathbf{x}_i), \tag{2}$$

where c is the ground-truth label of the sample $\mathbf{x}_i$, and $p_c(\mathbf{x}_i)$ is the output of the softmax layer which means the probability that the sample $\mathbf{x}_i$ belongs to class c (i.e., pedestrian or non-pedestrian). The loss of JCS-Net that joins super-resolution and classification can be finally expressed as

$$L_{JCS} = L_{cls} + \lambda L_{similarity}, \tag{3}$$

where λ is used to balance the two terms which is set to be 0.1 by cross-validation.

Most multi-scale methods only treat the pedestrians of different scales as the different sub-categories. Thus, these methods do not make full use of the relationship between the pedestrians of different scales. Compared to these methods, our proposed JCS-Net uses the relationship between the large-scale pedestrians and the small-scale pedestrians to help improve small-scale pedestrian detection.

B. MCF by JCS-Net for small-scale pedestrian detection

First of all, a review of multi-layer channel features (i.e., MCF) [12] is given. It integrates HOG+LUV (i.e., L1) and each layer of CNN (i.e., C1 to C5) into a unified framework. Fig. 2 gives the illustration of the original MCF: (1) Firstly, multi-layer image channels (i.e., L1 to L6) are constructed; (2) Secondly, the candidate features (i.e., F1 to F6) are extracted from each layer, respectively. (3) Finally, multi-stage cascade AdaBoost (i.e., S1 to S6) is learned from the candidate features of the corresponding layer. The original MCF based on HOG+LUV and the fine-tuned VGG16 are used for large-scale pedestrian detection in this paper.

For small-scale pedestrian detection, MCF is constructed by HOG+LUV and JCS-Net. Fig. 3 gives the illustration about

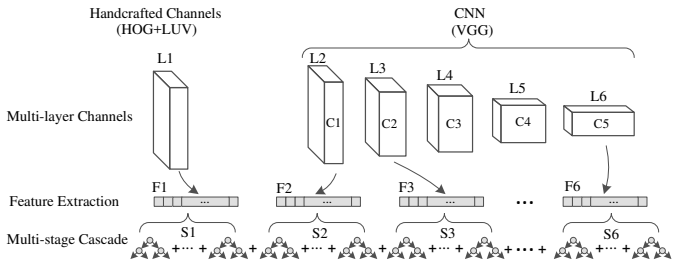

Fig. 2. Multi-layer Channel Features (MCF) based on HOG+LUV and each layer of CNN (i.e., VGG16). HOG+LUV and each layer of VGG16 (i.e., C1-C5) are used. It is used for large-scale pedestrian detection.

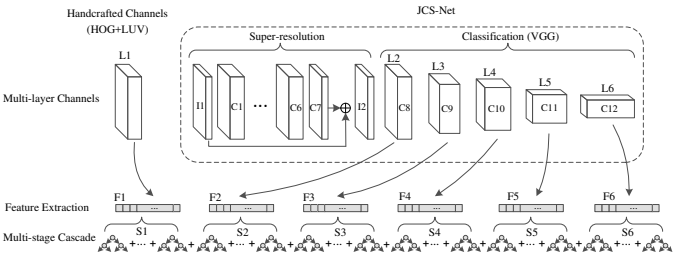

Fig. 3. Multi-layer Channel Features (MCF) based on HOG+LUV and JCS-Net. HOG+LUV and each layer of the classification sub-network in JCS-Net (i.e., C8-C12) are used. "I1" refers to the original small-scale pedestrian, and "I2" refers to the reconstructed pedestrian by the super-resolution sub-network. It is used for small-scale pedestrian detection.

how to construct MCF based on HOG+LUV and the trained JCS-Net. In JCS-Net, the super-resolution sub-network adopts the similar residual structure of VDSR [32]. It contains 7 convolutional layers. The first layer (C1) has 64 filters with the size of $3 \times 3 \times 3$, the middle five layers (C2-C6) have 64 filters with the size of $3 \times 3 \times 64$, and the last layer (C7) has 3 filters with the size of $3 \times 3 \times 3$. Because it learns the residual between the large-scale pedestrian and corresponding small-scale pedestrian, the reconstructed small-scale pedestrian (I2) is the addition of the original small-scale pedestrian (I1) and the output of the super-resolution sub-network (C7). Note that some other networks (e.g., [22]) can be also used for super-resolution. The classification sub-network in JCS-Net is based on the original structure of VGG16 [50]. JCS-Net is trained based on the joint loss of two sub-networks. Finally, the traditional HOG+LUV and each layer of classification sub-network (i.e., C8-C12) in the trained JCS-Net are used to construct multi-layer image channels (i.e., L1 to L6). Multi-stage cascade AdaBoost (i.e., S1 to S6) is learned based on multi-layer image channels. Note that the channels in the sub-network of the super-resolution are not used for constructing the multi-layer image channels.

At the test stage, the image channels in L1 (i.e., HOG+LUV) are firstly computed given the input image. Detection windows are generated by sliding the input image. For the detection windows accepted by S1, they are then put into the sub-network of super-resolution. The image channels in the sub-network of classification are computed based on the output of the super-resolution sub-network. The rest of the process is the same as the original MCF [12].

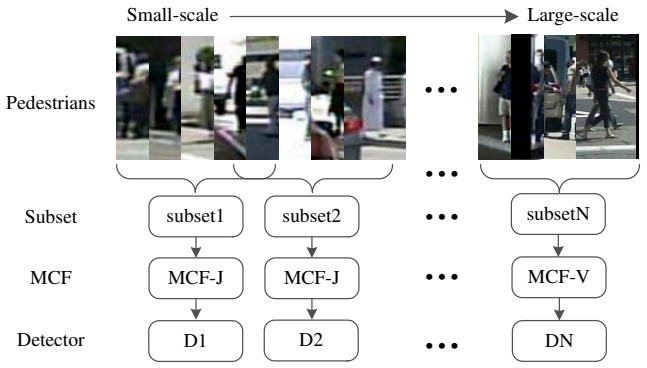

Fig. 4. The illustration of multi-scale MCF. It consists of multiple detectors. MCF-V means that MCF is constructed by HOG+LUV and the fine-tuned VGG16, which is used for large-scale pedestrian detection. MCF-J means that MCF is constructed by HOG+LUV and JCS-Net, which is used for small-scale pedestrian detection.

C. Multi-scale MCF

Usually, the scales of pedestrians are arbitrarily variable. For example, the reasonable height of pedestrians in the Caltech pedestrian dataset [20], [21] ranges from 50 pixels tall to 480 pixels tall, and the moderate height of pedestrians in the KITTI benchmark [24] ranges from 25 pixels tall to 374 pixels tall. Scale-agnostic methods treat the pedestrians of different scales as the same category and use the image pyramid technique to detect the pedestrians of different scales. To achieve much better detection performance, scale-aware methods were proposed recently [36], [59], [10].

In this subsection, the multi-scale MCF is proposed based on JCS-Net in Fig. 4. It consists of multiple detectors which are trained on different subsets. The pedestrians of different scales are split into several different subsets (e.g., subset 1, subset 2, ..., subset N) according to the height of pedestrians. To enlarge the number of samples in each subset, the different subsets can overlap. For each subset, MCF is used to train a multi-stage cascade AdaBoost. If MCF is constructed by HOG+LUV and JCS-Net, it is called MCF-J. If MCF is constructed by HOG+LUV and fine-tuned VGG16, it is called MCF-V. Generally speaking, MCF-J is used for small-scale pedestrian detection and MCF-V is used for large-scale pedestrian detection. In Fig. 4, the two detectors (D1 and D2) on the first two subsets (subset 1 and subset 2) are trained based on MCF-J, and the remaining detectors (D3-DN) on the remaining subsets are trained based on MCF-V. At the test stage, multiple detectors detect the input image, respectively. The scores of the same detection window are added together before NMS.

IV. Experiments

In this section, experiments on the Caltech pedestrian dataset [20], [21] and the KITTI benchmark [24] are shown to demonstrate the effectiveness of the proposed method and compare with some state-of-the-art methods.

The Caltech pedestrian dataset [20], [21] consists of 6 training sets and 5 test sets. The original training images and the test images are generated by every $30th$ frame. Thus, there are 4250 training images and 4024 test images. To enlarge the training data, we sample one image from every $3rd$ frame on the training set. The enlarged training data is called the Caltech 10x training set [62]. Thus, there are 42782 training images on the Caltech 10x set. The KITTI benchmark [24] is a very challenging computer vision benchmark, which consists of several different vision tasks, such as stereo, visual odometry, and object detection. Pedestrian detection is one sub-task in object detection, which consists of 7481 training images and 7518 test images.

The network of VGG16 [50] is used for pedestrian detection in this paper. Some original network parameters should be changed as follows: the input size of 227×227 is replaced by that of 128×64 and then the filter size in the first fully-connected layer is set as 4×2. The weights of VGG16 pre-trained on the ImageNet [17] is used for weight initialization. It is then fined-trained on the pedestrian dataset.

JCS-Net consists of two sub-networks: the super-resolution sub-network and the classification sub-network. The pedestrians over 50 pixels tall are used as the ground-truth of the super-resolution sub-network, and the interpolated versions of them are the input (i.e., "I1" of Fig. 3). Based on these pedestrians and their interpolated versions, the super-resolution sub-network is trained firstly. The classification sub-network (VGG16) is also fine-tuned on the pedestrians over 50 pixels tall. After that, the two sub-networks are jointly trained by Eq. (3). Finally, the trained JCS-Net is used to construct MCF-J for small-scale pedestrian detection.

For large-scale pedestrian detection, VGG16 is fine-tuned on the pedestrians over 50 pixels tall and MCF-V is constructed by HOG+LUV and each layer of the fine-tuned VGG16. For small-scale pedestrian detection, MCF-J is constructed by HOG+LUV and each layer of the classification sub-network in JCS-Net. Feature extraction in the first layer is based on the NNNF [13], and feature extraction in the CNN layers is single pixel of each channel. Soft Cascade AdaBoost [9] is used for learning the detector and the rejected threshold of each weak classifier.

A. Experiments on the Caltech pedestrian dataset

In this subsection, some experiments on the Caltech pedestrian dataset are conducted to show the effectiveness of the proposed method. Miss rates log-averaged over the range of FPPI=$[10^{-2},10^{0}]$ are used for evaluating detection performance, where FPPI means false positives per image.

To demonstrate the effectiveness of JCS-Net and train multi-scale MCF, the original pedestrians in the training data are split into three different subsets, which are called "train-all", "train-small", and "train-large", respectively. "train-all" subset contains all the pedestrians, "train-small" subset contains the pedestrians under 100 pixels tall and the interpolated pedestrians over 100 pixels tall, and "train-large" subset contains the pedestrians over 80 pixels tall.

*1) **Effectiveness of JCS-Net for small-scale pedestrian detection**:* To show the effectiveness of JCS-Net for small-scale pedestrian detection, MCF-J and MCF-V are both trained

(a) (b) (c)

Fig. 5. The reconstructed pedestrians by super-resolution sub-network of JCS-Net are compared to the interpolated pedestrians. The left of each sub-figure is the interpolated pedestrian, and the right is the reconstructed pedestrian.

TABLE I
MISS RATES (MR) OF MCF-V AND MCF-J ARE SHOWN ON CALTECH TEST SET. MCF-V IS LEARNED BASED ON HOG+LUV AND THE FINE-TUNED VGG16. MCF-J IS LEARNED BASED ON HOG+LUV AND THE PROPOSED JCS-NET.

method	training set	reasonable	small
MCF-V	"train-small"	13.20%	14.28%
MCF-J	"train-small"	11.07%	11.72%
ΔMR	-	2.13%	2.56%
ablation experiments:			
MCF-C	"train-small"	12.23%	13.02%
MCF-S	"train-small"	12.65%	13.50%

TABLE II
MISS RATES (MR) OF MS-V AND MS-J ARE SHOWN ON CALTECH TEST SET. MS-V MEANS MULTI-SCALE MCF BASED ON FINE-TUNED VGG16. MS-J MEANS MULTI-SCALE MCF BASED ON JCS-NET.

method	detectors	training set	reasonable	small
MS-V	MCF-V	"train-small"	9.67%	10.48%
	MCF-V	"train-large"		
	MCF-V	"train-all"		
MS-J	MCF-J	"train-small"	8.81%	9.57%
	MCF-V	"train-large"		
	MCF-V	"train-all"		
ΔMR	-	-	0.86%	0.91%

based on "train-small" subset. The training processes of MCF-V and MCF-J are similar. The positives come from "train-small" subset. The negatives are generated by the bootstrap technique with five rounds of the original NNNF [13], where the number of decision trees in each round is 32, 128, 512, 2048, and 4096, respectively. Based on the positives and negatives, multi-stage cascade AdaBoost is learned, which consists of 4096 depth-4 decision trees.

Table I compares Miss Rates (MR) of MCF-V and MCF-J on the Caltech test set. The two subsets of the Caltech test set (i.e., reasonable and small) are used for evaluation. The reasonable test set means that the pedestrians are over 50 pixels tall under no or partial occlusion, and the small test set means that the pedestrians are under 100 pixels tall and over 50 pixels tall. Namely, the small test set belongs to the reasonable test set. On the reasonable test set, MR of MCF-V is 13.20% and that of MCF-J is 11.07%. Thus, MCF-J outperforms MCF-V by 2.13%. On the small test set, MCF-J outperforms MCF-V by 2.56%. This means that the proposed JCS-Net is useful for small-scale pedestrian detection.

To further demonstrate the effectiveness of the proposed JCS-Net, we conduct two additional ablation experiments in Table I. (1) The first one is setting λ of Eq. (3) to zero during the fine-tuning to turn off the term of the similarity loss. This aims to show whether a higher capacity network leads to a good performance. Based on it, MCF-C is learned. Thus, MCF-C has a higher capacity compared to MCF-V and has the same capacity compared to MCF-J. In Table I, MCF-C has the better performance than MCF-V and has the worse performance than MCF-J. It means that the higher capacity is not the main reason that MCF-J can improve the performance of small-scale pedestrian detection; (2) The second one is using the vanilla super-resolution for small-scale pedestrian detection. Namely, the super-resolution sub-network and the classification sub-network are independently trained. It aims to demonstrate the importance of the joint training of the super-resolution and classification. Based on it, MCF-S is learned. MCF-J also outperforms MCF-S. Even though using the vanilla super-resolution can also improve the detection performance, it cannot create the best detection performance improvement compared with the joint training of JCS-Net.

Fig. 5 gives some examples about the interpolated pedestrians and correspondingly reconstructed pedestrians by the super-resolution sub-network on the Caltech test set. The left of each sub-figure is the interpolated pedestrian, and the right is the reconstructed pedestrian by the super-resolution sub-network. It can be seen that the reconstructed pedestrians are relatively clear and their contours are more obvious. For example, the stairs in the right of Fig. 5(b) are much clearer. Thus, JCS-Net can help improve the following classification sub-network by providing more detailed information.

*2) **Multi-scale MCF**:* For multi-scale MCF-V (MS-V), the three detectors on the three different training subsets (i.e., "train-all", "train-small", and "train-large") are all trained by

TABLE III
MULTI-SCALE MCFS (I.E., MS-V AND MS-J) ARE COMPARED TO SINGLE-SCALE MCF ON CALTECH TEST SET (REASONABLE).

	MCF	MS-V	**MS-J**
MR	10.40%	9.67%	8.81%
ΔMR	-	0.73%	1.59%

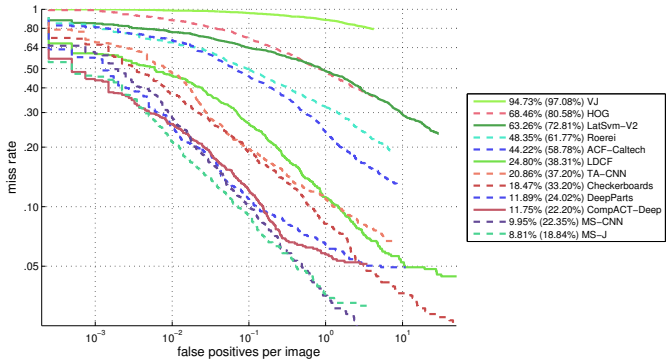

Fig. 6. ROC of some state-of-the-art methods on Caltech test set (reasonable). Miss rates log-averaged over the range of FPPI=$[10^{-2},10^{0}]$ and FPPI=$[10^{-4},10^{0}]$ are both shown in the legend.

MCF-V. For multi-scale MCF-J (MS-J), two detectors on the "train-all" and "train-large" subsets are trained by MCF-V and one detector on the "train-small" subset is trained by MCF-J. The scores of the same detection windows predicted by different detectors are added together before NMS. Table II compares the Miss Rates (MR) of multi-scale MCF-V (MS-V) and multi-scale MCF-J (MS-J) on the Caltech test set. On the reasonable test set, MS-J outperforms MS-V by 0.86%. On the small test set, MS-J outperforms MS-V by 0.91%. It demonstrates that incorporating the relationship between large-scale pedestrians and small-scale pedestrians into the network design can improve pedestrian detection performance.

Multi-scale MCFs (i.e., MS-V and MS-J) are also compared to single-scale MCF on the Caltech test set (reasonable) in Table III. Single-scale MCF treats all the pedestrians as the same category and uses them (i.e., "train-all") to train one detector. MS-V and MS-J both have the lower miss rates than MCF. For example, MS-J outperforms MCF by 1.59%. Moreover, MS-J has the best performance.

*3) **Comparison with some state-of-the-art methods**:* Finally, we compare our proposed method (MS-J) with some state-of-the-art methods. Fig. 6 shows the ROC curves of these methods on the Caltech test set (reasonable). Miss rates log-averaged over the range of FPPI=$[10^{-2},10^{0}]$ and FPPI=$[10^{-4},10^{0}]$ are both shown in the legend. VJ [56], HOG [16], LatSvm-V2 [23], Roerei [7], ACF-Caltech [18], LDCF [42], and Checkerboards [62] are the traditional hand-crafted features based methods. TA-CNN [54], DeepParts [53], CompACT-Deep [11], MS-CNN [10], and our proposed MS-J are the CNN based methods. MS-J achieves the best detection performance (i.e., 8.81% MR). For examples, MRs of CompACT-Deep [11] and MS-CNN [10] are 11.75% and 9.95%, respectively. MS-J outperforms CompACT-Deep [11] and MS-CNN [10] by 2.94% and 1.14%. Meanwhile, The proposed method outperforms PGAN [37] by 0.67%. If FPPI=$[10^{-4},10^{0}]$ is used for evaluation, MS-J outperforms CompACT-Deep [11] and MS-CNN [10] by 3.36% and 3.51%. Because many state-of-the-art methods only provide the detection results of pedestrians over 50 pixels tall, Fig. 7 further evaluates the detection performance of small-scale pedestrian detection on the small test mentioned above instead of the medium and far sets mentioned in [21]. It can be seen that MS-J also achieves the state-of-the-art performance.

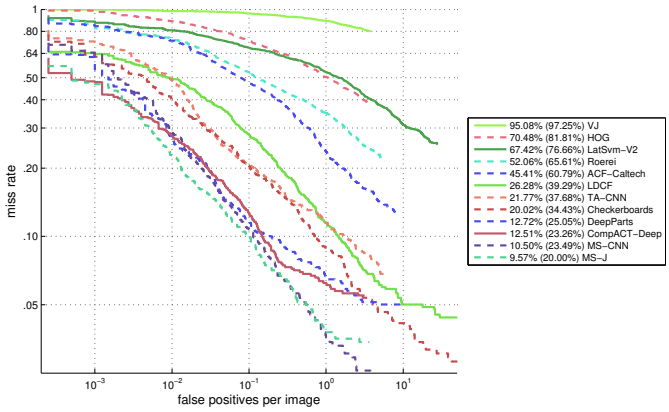

Fig. 7. ROC of some state-of-the-art methods on Caltech test set (small). Small set means the pedestrians over 50 pixels tall and under 100 pixels tall.

TABLE IV
AVERAGE PRECISION (AP) OF MCF-V AND MCF-J ARE SHOWN ON KITTI VALIDATION SET. MCF-V IS LEARNED BASED ON HOG+LUV AND THE FINE-TUNED VGG16. MCF-J IS LEARNED BASED ON HOG+LUV AND THE PROPOSED JCS-NET.

method	training set	moderate	small
MCF-V	"train-small"	61.95%	47.21%
MCF-J	"train-small"	65.12%	50.64%
ΔAP	-	3.17%	3.43%

B. Experiments on the KITTI benchmark

In this subsection, some experiments on the KITTI benchmark are further conducted to show the effectiveness of the proposed method. Instead of miss rate (MR) used on the Caltech dataset, precision-recall (PR) is used for the evaluation on the KITTI benchmark. Average precision (AP) is averaged by summing in 10% recall steps.

To demonstrate the effectiveness of the proposed method, the training data is split into two parts: the training set and the validation set. Following [15], each part contains about half of the training data. The pedestrians in the training set are split into three subsets according to the height of pedestrians, including "train-small", "train-medium", and "train-large". Because the pedestrians over 50 pixels tall and 25 pixels tall are used for the evaluation on the Caltech and KITTI datasets respectively, the partition of pedestrians on the KITTI training set is different from that on the Caltech training set. Specifically, "train-small" subset consists of the pedestrians

TABLE V
AVERAGE PRECISION (AP) OF MS-V AND MS-J ARE SHOWN ON KITTI VALIDATION SET. THREE SUBSETS WITH DIFFERENT DIFFICULTIES (I.E., EASY, MODERATE, AND HARD) ARE USED FOR EVALUATION. MS-V MEANS MULTI-SCALE MCF BASED ON FINE-TUNED VGG16. MS-J MEANS MULTI-SCALE MCF BASED ON JCS-NET.

method	detectors	training set	Easy	Moderate	Hard
	MCF-V	"train-small"			
MS-V	MCF-V	"train-medium"	77.65%	71.80%	61.96%
	MCF-V	"train-large"			
	MCF-J	"train-small"			
MS-J	MCF-V	"train-medium"	78.19%	72.47%	63.26%
	MCF-V	"train-large"			
ΔAP	-	-	0.54%	0.67%	1.30%

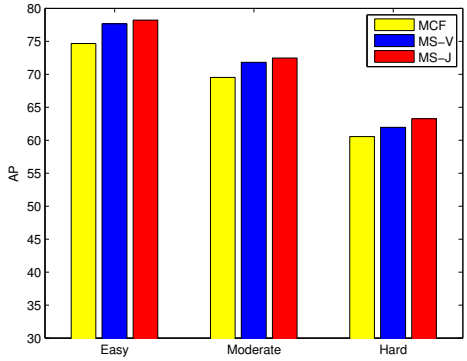

Fig. 8. Multi-scale MCFs (i.e., MS-V and MS-J) are compared to single-scale MCF on KITTI validation set.

under 100 pixels tall, "train-medium" subset contains the pedestrians under 150 pixels tall and over 40 pixels tall, and "train-large" subset means the pedestrians over 50 pixels tall.

*1) **Effectiveness of JCS-Net for small-scale pedestrian detection**:* In the training processes of MCF-V and MCF-J, the positives come from "train-small" subset, and the negatives are generated by the bootstrap technique with five rounds of the original NNNF [13]. Because the number of training data is relatively limited compared to the Caltech 10x training images, depth-2 decision tree instead of depth-4 decision tree is used. The final detectors of MCF-V and MCF-J for small-scale pedestrian detection both contain 4096 depth-2 decision trees. The number of decision trees in each stage is the same as that of Sec. IV-A.

Table IV compares Average Precisions (AP) of MCF-V and MCF-J on the validation set. The moderate set on the validation set is used for evaluation. AP of MCF-V is 61.95% and that of MCF-J is 65.12%. MCF-J outperforms MCF-V by 3.17%. To show the improvement on small-scale pedestrian detection, the small subset is further split from the moderate set, which refers to the pedestrians under 100 pixels tall in the moderate set. On the small subset, MCF-J outperforms MCF-V by 3.43%. Thus, MCF-J has the better performance than MCF-V, especially on small-scale pedestrian detection. The reason is that MCF-J makes full use of the relationship between the large-scale pedestrians and the small-scale pedestrians.

TABLE VI
AVERAGE PRECISION (AP) OF SOME STATE-OF-THE-ART METHODS ON KITTI TEST SET. THREE DIFFERENT DIFFICULTIES (I.E., EASY, MODERATE, AND HARD) ARE USED FOR EVALUATION. THE CNN WHETHER USED FOR PROPOSAL EXTRACTION IS ALSO GIVEN.

method	Proposal Extraction	Easy	Moderate	Hard
ACF [18]	without CNN	44.49%	39.81%	37.21%
Checkerboards [62]	without CNN	67.65%	56.75%	51.12%
NNNF [14]	without CNN	69.16%	58.01%	52.77%
DeepParts [53]	without CNN	70.49%	58.67%	52.78%
CompACT-Deep [11]	without CNN	70.69%	58.74%	52.71%
MCF [12]	without CNN	70.87%	59.45%	54.28%
CFM [29]	without CNN	74.21%	63.26%	56.44%
MS-J	without CNN	75.94%	63.41%	59.03%
RPN+BF [60]	with CNN	75.45%	61.29%	56.08%
SubCNN [57]	with CNN	83.17%	71.34%	66.36%
MSCNN [10]	with CNN	83.70%	73.62%	68.28%
MS-J	with CNN	85.62%	74.99%	69.65%

*2) **Multi-scale MCF**:* Table IV demonstrates the effectiveness of JCS-Net for small-scale pedestrian detection. To achieve much better performance, multi-scale MCF is proposed (i.e., MS-V and MS-J). In MS-V, the three detectors on the three different subsets (i.e., "train-small", "train-medium", and "train-large") are all trained based on MCF-V. In MS-J, the detector on the "train-small" subset is trained based on MCF-J and two other detectors on the "train-medium" and "train-large" subsets are trained based on MCF-V. The scores of the same detection windows predicted by different detectors are added together before NMS. Table V compares the average precision (AP) of MS-V and MS-J on the three different difficulties (i.e., "Easy", "Moderate", and "Hard") of the validation set. MS-J outperforms MS-V on all the three different difficulties. For example, MS-J outperforms MS-V by 1.30% on the hard set. MS-J joins super-resolution and classification for small-scale pedestrian detection and treats the large-scale pedestrians and the small-scale pedestrians as the different sub-categories, while MS-V only treats the pedestrians of different scales as the different sub-categories. Thus, MS-J has the better detection performance than MS-V.

Fig. 8 further compares multi-scale MCFs (i.e, MS-V and MS-J) with single-scale MCF on the validation set. Single-scale MCF only trains one detector based on all the pedestrians of the training set. No matter which difficulty is used (i.e., Easy, Moderate, or Hard), MS-V and MS-J both have the better performance than single-scale MCF.

*3) **Comparison with some state-of-the-art methods**:* Finally, the proposed MS-J is compared with some state-of-the-art methods on the KITTI test set in Table VI. (1) Firstly, some methods which do not use the CNN to extract the candidate proposals are compared. Among thse methods, ACF [18], Checkerboards [62], and NNNF [13] use the traditional handcrafted features to learn the pedestrian detector. DeepParts [53], CompACT-Deep [11], MCF [12], CFM [29], and our proposed MS-J use the handcrafted features to extract the candidate proposals and use the CNN features to further classify these proposals. Among these methods, MS-J achieves

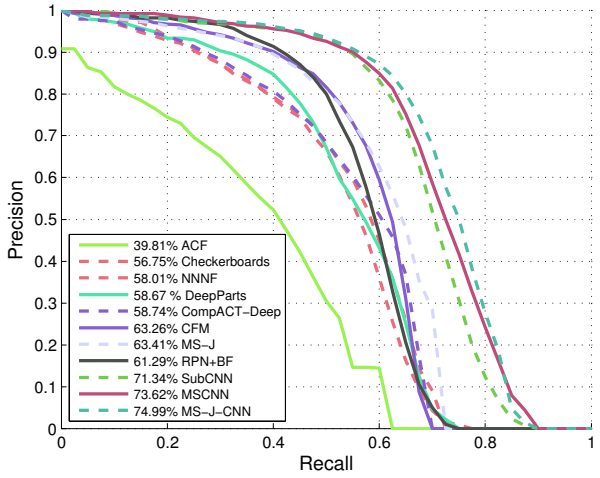

Fig. 9. PR curves of some state-of-the-art methods on the KITTI test set (Moderate).

the best detection performance. On the moderate set, AP of CompACT-Deep [11] is 58.74% and that of MCF [12] is 59.45%. MS-J outperforms CompACT-Deep [11] and MCF [12] by 4.67% and 3.96%. (2) Secondly, some methods which use the CNN to extract the candidate proposals (i.e., [60], [57], [10]) are further compared. Instead of using NNNF, MS-J uses FPN [38] to extract the candidate proposals. It can be seen that the proposed method also outperforms the other methods. For example, MS-J outperforms MSCNN and SubCNN by 1.37% and 3.65% on the moderate subset. Fig. 9 further shows the PR curves of these methods on the moderate subset. It can be seen that MS-J outperforms the other methods.

V. Conclusion and Future Work

In this paper, we proposed a unified framework (called JCS-Net) for small-scale pedestrian detection. It consists of two sub-networks: one for super-resolution and another for classification. The loss of JCS-Net is the joint loss of the super-resolution sub-network and the classification sub-network. Due to the incorporation of the relationship between the large-scale pedestrians and the small-scale pedestrians, MCF based on JCS-Net (MCF-J) provided better detection performance for small-scale pedestrian detection. Experiments on two public datasets (the Caltech [20], [21] and KITTI [24] datasets) showed the effectiveness of the proposed JCS-Net. To have a better performance for pedestrian detection, multi-scale MCF based on JCS-Net (MS-J) was also proposed. It achieved a state-of-the-art performance on the pedestrian datasets.

However, we have observed that the detection performance drops when the pedestrian and its background are similar in appearance. The reason might be that we only consider the similarity between the reconstructed pedestrian and the large-scale pedestrian, but ignore the dissimilarity between the reconstructed background and the large-scale pedestrian. In the future work, we will consider this dissimilarity to further improve pedestrian detection.

References

[1] Z. Zhang, Y. Zhao, Y. Wang, J. Liu, Z. Yao and J. Tang, "Transferring training instances for convenient cross-view object classification in surveillance," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 10, pp. 1632-1641, 2013.

[2] H. Su, H. Yang, S. Zheng, Y. Fan, and S. Wei, "The large-scale crowd behavior perception based on spatio-temporal viscous fluid field," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 10, pp. 1575-1589, 2013.

[3] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 10, pp. 1590-1599, 2013.

[4] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 10, pp. 1642-1653, 2013.

[5] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. European Conf. Computer Vision*, 2018.

[6] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," *CoRR, abs/1512.04143*, 2015.

[7] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[8] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. European Conf. Computer Vision*, 2014.

[9] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[10] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. European Conf. Computer Vision*, 2016.

[11] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

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

[13] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[14] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," *IEEE Trans. on Image Processing*, vol. 25, no. 12, pp. 5538–5551, 2016.

[15] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[18] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fastest feature pyramids for object detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vo. 36, no. 8, pp. 1532-1545, 2014.

[19] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. British Machine Vision Conference*, 2009.

[20] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[21] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vo. 34, no. 4, pp. 743-761, 2012.

[22] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 2016.

[23] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[25] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018.
[26] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014.
[27] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-Driven Super Resolution: Object detection in low-resolution images," *CoRR, abs/1803.11316*, 2018.
[28] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.
[29] Q. Hu, P. Wang, C. Shen, A. Hengel and F. Porikli, "Pushing the limits of deep CNNs for pedestrian detection," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
[30] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018.
[31] Z. Jie, X. Liang, J. Feng, W. F. Lu, E. H. F. Tay, and S. Yan, "Scale-aware pixelwise object proposal networks," *IEEE Trans. Image Processing*, vol. 25, no. 10, pp. 4525-4539, 2016.
[32] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016.
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012.
[34] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate superresolution," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[35] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Trans. Image Processing*, vol. 25, no. 11, pp. 5012-5024, 2016.
[36] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *CoRR, abs/1510.08160*, 2015.
[37] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[38] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[39] H. Liu, J. Lu, J. Feng, and J. Zhou, "Learning deep sharable and structural detectors for face alignment," *IEEE Trans. Image Processing*, vol. 26, no. 4, pp. 1666-1678, 2017.
[40] X. Mao, C. Shen, and Y. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Advance in Neural Information Process Systems*, 2016.
[41] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[42] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Advances in Neural Information Processing Systems*, 2014.
[43] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Computer Vision*, 2015.
[44] Y. Pang, M. Sun, X. Jiang, and X. Li, "Convolution in Convolution for Network in Network," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1587-1597, 2018.
[45] Y. Pang, J. Cao, and X. Li, "Cascade Learning by Optimally Partitioning," *IEEE Trans. Cybernetics*, vol. 47, no. 12, pp. 4148-4161, 2017.
[46] A. Rozantsev, M. Salzmann, and P, Fua, "Beyond sharing weights for deep domain adaptation," *CoRR, abs/1603.06432*, 2016.
[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Advance in Neural Information Process Systems*, 2015.
[48] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, and Y. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[49] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR, abs/1409.1556*, 2014.
[51] H. Sun and Y. Pang, "GlanceNets Efficient Convolutional Neural Networks with Adaptive Hard Example Mining," *SCIENCE CHINA Information Sciences*, vol. 61, no. 10, pp.109-101, 2018.
[52] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.
[53] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int'l Conf. Computer Vision*, 2015.
[54] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015.
[55] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Computer Vision*, vol. 104, no. 2, pp. 154-171, 2013.
[56] P. Viola and M. Jones, "Robust real-time face detection," *Int'l J. Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.
[57] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, 2017.
[58] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int'l Conf. Computer Vision*, 2015.
[59] F. Yang, W. Choi, and Y. Lin, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016.
[60] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection," in *Proc. European Conf. Computer Vision*, 2016.
[61] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014.
[62] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2015.
[63] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2018.
[64] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-Resolution," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2018.
[65] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu, "Scale-adaptive deconvolutional regression network for pedestrian detection," in *Proc. Asian Conf. Computer Vision*, 2016.

Yanwei Pang (M'07-SM'09) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China in 2004. Currently, he is a professor of the Tianjin University, China. His research interests include object detection and image recognition, in which he has published 145 scientific papers including 35 IEEE Transactions papers. He was a member of the editorial boards of the International Journal of Computer Mathematics (Taylor & Francis), the Neurocomputing (Elsevier), the International Journal of Image and Graphics (World Scientific), and the International Journal of Creative Computing.

Jiale Cao received the Ph.D. in information and communication engineering from the Tianjin University, Tianjin, China, in 2018. He is currently a postdoctoral in the Tianjin University. His research interests include object detection and deep learning, in which he has published five IEEE Trans. papers and two CVPR paper.

Jian Wang received the Ph.D. degree in electronic engineering from the Shanghai Jiao Tong University in 2004. Currently his is a lecture of the Tianjin University. His research interests include image processing and image recognition.

Jungong Han is currently a tenured Data Science Associate Professor at University of Warwick, UK. Dr. Han's research interests include Computer Vision, Artificial Intelligence and Machine Learning. He has published over 180 papers, including 35+ IEEE Trans and 30+ A* conference papers.