Multi-Grained Deep Feature Learning for Robust Pedestrian Detection

Chunze Lin, Jiwen Lu, *Senior Member, IEEE,* and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we address the challenging problem of detecting pedestrians which are heavily occluded and/or far from cameras. Unlike most existing pedestrian detection methods which only use coarse-resolution feature maps with fixed receptive fields, our approach exploits multi-grained deep features to make the detector robust to visible parts of occluded pedestrians and small-size targets. Specifically, we jointly train a multi-scale network and a human parsing network in a weakly-supervised manner with only bounding box annotations. We carefully design the multi-scale network to predict pedestrians of particular scales with the most appropriate feature maps, by matching their receptive fields with the target sizes. The human parsing network generates a fine-grained attention map which helps guide the detector to focus on the visible parts of occluded pedestrians and small-size instances. Both networks are computed in parallel and form a unified single stage pedestrian detector, which assures a suitable trade-off between accuracy and speed. Moreover, we introduce an adversarial hiding network to make our detector more robust to occlusion situations, which generates occlusions on pedestrians in the goal to fool the detector that in turn adapts itself to learn to localize these adversarial instances. Experiments on three challenging pedestrian detection benchmarks show that our proposed method achieves state-of-the-art performance and executes 2× faster than competitive methods.**

***Index Terms*—Pedestrian detection, human parsing, attention, deep feature learning**

I. Introduction

Pedestrian detection is one of the most important topics in computer vision and has attracted great attention over the past few years [1]–[13]. It is a key technology in many practical applications such as automotive safety, intelligent video surveillance and human behavior analysis. Despite the recent progress, it is still a challenging problem to detect occluded pedestrians due to the noisy representation and small-size targets because of the low resolution. Fig. 1 illustrates some examples of small and occluded pedestrian images.

Existing pedestrian detection methods can mainly be mainly classified into two categories: hand-crafted features based [2], [4], [5], [15], [16] and deep learning features based [6], [7], [17]–[19]. For the first category, prior knowledge such as edges and human shapes are considered to generate features and decision trees are usually learned by applying boosting to these features to form a pedestrian detector. For the second category, features are learned via a series of convolutional and pooling layers according to the training data. With its deep structure, the convolutional neural network (CNN) generates more abstract and high-level semantic features which significantly improve the pedestrian detection performance.

The authors are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Beijing National Research Center for Information Science and Technology, Beijing 100084, China. E-mail: lcz16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn. (Corresponding author: Jiwen Lu)

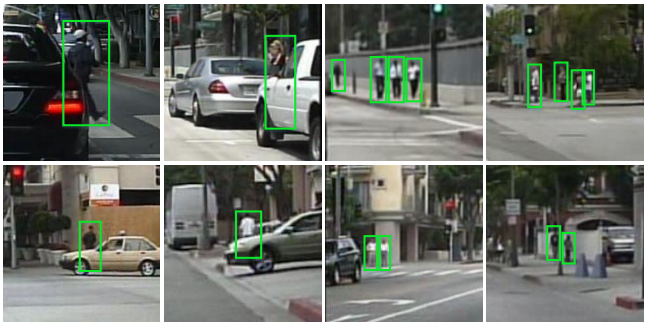


Fig. 1. Several examples of small and occluded pedestrians. (a) Pedestrians are often occluded by cars from Caltech dataset [14]. (b) In addition to occlusion, pedestrians are usually of small sizes, which makes the detection even more challenging.

While many CNN-based pedestrian detection methods have been proposed in recent years, there are still two main shortcomings: 1) most of them usually use feature maps with a single receptive field to deal with multi-scale pedestrians. The mismatch between the sizes of targets and receptive fields limits the performance. The small-size instances especially suffer from this inconsistency, which are often ignored when the receptive field is too large; 2) most of them are full-body detectors which are not efficient when dealing with occlusion situation. Even if some methods learn a set of human part detectors to handle the occlusion issue, the feature maps of coarse resolution are often employed for detection. However, the body parts are of small sizes and their information on the coarse-resolution feature maps are limited, making them indistinguishable from the background. The representations of high resolution are therefore necessary for effective detection.

In this paper, we propose a multi-grained deep feature learning (MDFL) method to simultaneously handle the occlusion and small-size problems in pedestrian detection. Fig. 2 illustrates an overview of the proposed framework. Instead of using feature maps with single resolution and fixed receptive field, we introduce a multi-scale network which exploits multiple feature maps for detection and a human parsing network which incorporates pixel-wise information to guide the detector. The multi-grained feature maps make the detector

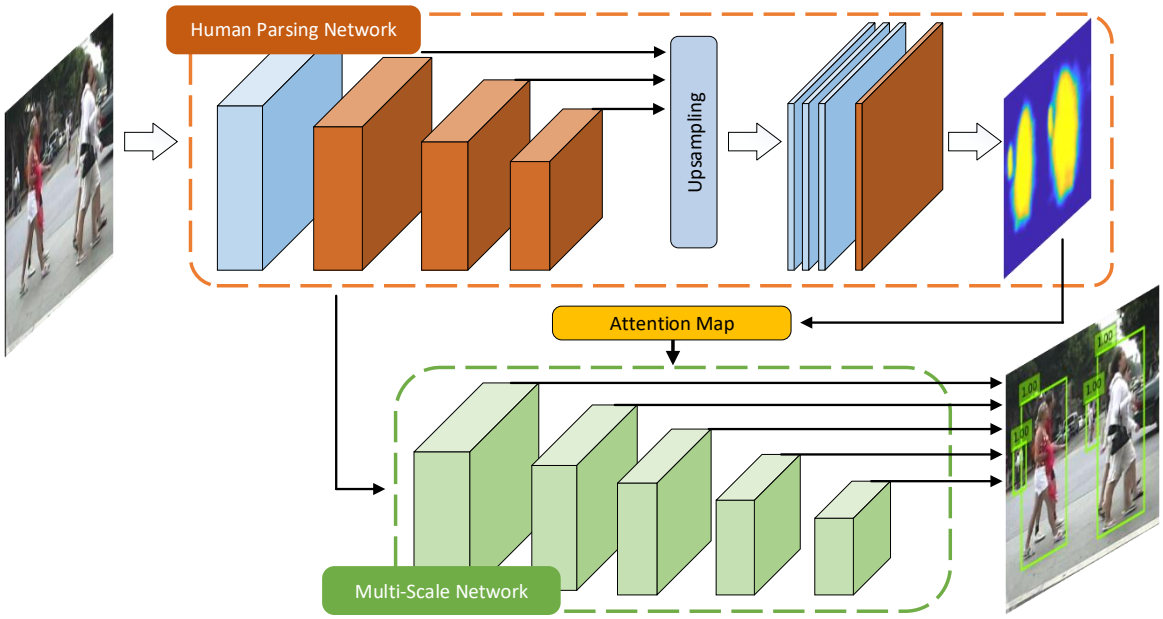


Fig. 2. Overview of the proposed framework. Given an input image, the human parsing network generates fine-grained feature maps which are encoded into the feature maps of the multi-scale network as a segmentation-aware attention map to help focus on visible parts of occluded targets and small-size pedestrians. The multi-scale network employs multiple feature maps with different receptive field sizes and resolutions to detect pedestrians of specific scale.

more robust to large variation of scales and especially to small-size targets. We collaboratively learn the multi-scale network and the human parsing network, which form a single-shot pedestrian detector and can directly output predicted bounding boxes without any post-processing, except a simple NMS. The multi-scale network is carefully designed to form a feature pyramid and detects pedestrian of specific size with most appropriate feature maps. Shallower feature maps with small receptive field are employed for detecting small-size targets while deeper layers are used for large instances. The human parsing network generates a fine-grained human parsing mask, which is then converted into an attention map to guide the multi-scale network focus on pedestrians. In addition, we also propose an adversarial hiding network (AHN) which artificially generates occlusion on pedestrians to make our detector be more robust to occlusion issue. The AHN aims to hide most discriminative parts of pedestrians so that the well trained detector becomes unable to recognize the targets. In turn, the detector will adapt itself to learn to recognize these adversarial instances. Experiments results on challenging pedestrian detection datasets including Caltech [14], KITTI [20] and INRIA [15] demonstrate the effectiveness of the proposed method, which achieves state-of-the-art performance and executes at least $2\times$ faster than competitive approaches.

This paper is an extended version of our conference paper [21]. There are several new contributions:

1) We exploit the adversary to generate hard occlusion cases to make our pedestrian detector be more robust when dealing with occlusion situations.
2) We perform extensive ablation analysis and examine the impacts of different main components of our model on the detection performance.
3) We conduct more additional experiments on the widely used pedestrian dataset INRIA [15] and achieve the state-of-the-art performance.

II. Related Work

In this section, we briefly review two topics: 1) pedestrian detection and 2) simultaneous detection and segmentation.

A. Pedestrian Detection

Visual human analysis [22]–[29] is one of the most important topics in computer vision since human is the central component in real world applications. Pedestrian detection, as a key technology in visual human analysis, has attracted great attention over the past decade and many efforts have been made to improve the performance of pedestrian detection. Existing methods can be mainly categorized into two classes: hand-crafted features based [30]–[34] and deep convolutional features based [6], [7], [17], [35], [36]. The Integrate Channel Features (ICF) [2] is among the most popular pedestrian detectors without using deep features. It exploited channel feature pyramids and boosted classifiers. The feature representations of ICF have been widely studied and many variants have been proposed [4], [5], [37], [38]. With the prevalence of deep convolutional neural network [39]–[41], most recent pedestrian detection approaches are deep CNN-based. Since the region proposal based detectors [42], [43] have achieved great results in general object detection, many pedestrian detection methods were variations of Faster R-CNN [43]. MS-CNN [7] integrated a feature pyramid property into Faster-RCNN [43] to address the scale problem. While SA-FastRCNN [44] proposed multiple built-in sub-networks to detect pedestrians with scales from disjoint ranges. Instead of using RoIPooling [43] followed by some fully connected

layers, some methods used stronger classifiers to boost the performance. In RPN+BF [6], given pedestrian candidates generated by the Region Proposal Network (RPN) [43], convolutional features were extracted and fed into a boosted forest (BF) to perform hard negative samples mining and make the detector more robust. SDS-RCNN [17] replaced the boosted forest by a VGG16 network [45] for classification and exploited an additional semantic segmentation loss to implicitly supervise and guide the detector. While F-DNN [46] used SSD [47] for region proposals and a series of deep classifiers in parallel to post verify each candidate. Besides, apart the above full-body detectors, some methods [11], [19], [48]–[50] learned occlusion-specific detectors, where each one was responsive to detect a human part to handle occlusion issue. These detectors would give a high confidence score based on the parts which are still visible when the full-body detector is confused by the presence of background. Different from most of the above pedestrian detection methods which adopted two-stage pipeline [43], [51] needing considerable computing time, we propose a single stage framework. Instead of part-level detection we exploit pixel-wise classification to deal with occlusion and small-size issues.

B. Simultaneous Detection and Segmentation

There are two main lines of research on simultaneous detection and segmentation. The first aims to improve the performance of both detection and segmentation tasks [52]–[54]. For example, Dai *et al.* [56] constructed a cascaded network on top of the region proposal network [43] to predict instance-aware segmentation mask. Qi [55] proposed hierarchically gated deep networks to customize a suitable scale for targets of different sizes. He *et al.* [57] extended the Faster-RCNN framework [43] by adding a branch for segmentation in parallel with existing branch for bounding box prediction. The second tends to use segmentation as a strong cue to improve detection. For example, Mao *et al.* [58] considered features of different semantic levels and fused with deep convolutional features. They demonstrated that fusing semantic segmentation features with deep convolutional features improves significantly the detection performance than other features. Tian *et al.* [8] incorporated the scene attributes to improve the detection accuracy. This method discarded hard negative samples due to the complex background with the scene attributes. Du *et al.* [46] utilized the semantic segmentation (SS) information as an additional deep classifier in their F-DNN+SS framework. The segmentation mask was used in a post-processing manner to eliminate the predicted bounding boxes that contain only the background. While most of the above approaches require external pixel-wise annotations to train the segmentation module, we explore a weakly supervised strategy and only need bounding box annotations for training. Instead of using the segmentation mask as a post-processing or hard mining strategy, our attention mechanism is computed in parallel with the multi-scale network and explicitly activates the pedestrian regions.

C. Generative Adversarial Networks

Generative adversarial networks (GANs) [59] have shown very exciting results for numerous generative tasks such as image generation [60]. Recently, the adversarial training and its ability of generating hard data have been used to improve many practical tasks. For example, Zhao *et al.* [61] introduced a generative network for hard triplet generation to optimize the network ability in distinguishing similar examples of different categories while grouping instances of the same categories. Li *et al.* [62] proposed a perceptual GAN that generate super-resolution representation of small objects to boost the detection performance of these small samples. More similar with our work, Wang *et al.* [63] extended the Faster-RCNN framework [43] with an adversarial network for generating examples with occlusions and deformations to challenge original object detectors. However, in this work the occlusion shape is handcrafted defined and the generator has limited interaction with the detector. In contrast, our adversarial hiding network learns to generate deformable occlusion according to the targets and receive information directly from the detector, which makes it generate more challenging occlusion examples.

III. Approach

In this section, we first present our proposed MDFL detector which predicts a series of bounding boxes and scores, followed by non-maximum suppression (NMS) to produce the final detection results. Then we introduce the adversary which has the goal to improve the detection performance in occlusion conditions. Lastly, we present how to optimize these deep convolutional networks and give the details of our implementation.

Our MDFL framework is composed of two key parts: a multi-scale network which detects pedestrian using multi-grained features and a human parsing network which generates a fine-grained attention map to help the multi-scale network focus on regions that contain pedestrians. These two networks are computed in parallel and form a single stage detection framework [47], which offers a great trade-off between accuracy and speed. Fig. 2 illustrates an overview of the proposed architecture of our approach.

A. Multi-scale Network

In natural scenes, pedestrian images usually have large variations and appear at multiple scales. Instead of conventional image pyramids, we design the multi-scale network to detect the pedestrians of specific size using feature maps with appropriate resolution and receptive field. Specifically, high-resolution feature maps are used for smaller targets detection while feature maps with larger receptive field are extracted for large-size pedestrians detection.

1) Architecture of the network: The multi-scale network, as shown in Fig. 3, is composed of the following structures:

- Trunk Network: The multi-scale network is based on a truncated VGG16 network [45] where the fully connected layers are converted into convolutional layers and the classification layers are removed. Extra convolutional layers are added to the end of the base network. These

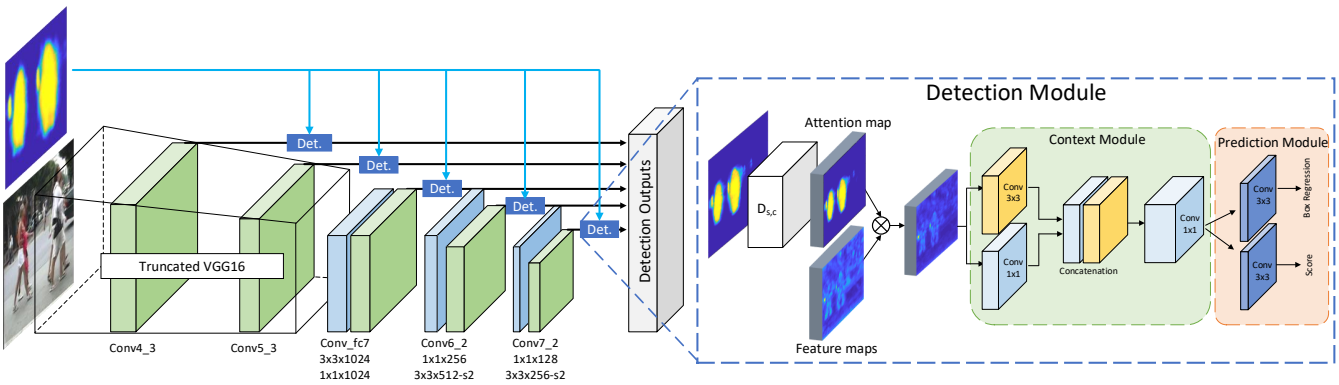


Fig. 3. The architecture of the multi-scale network. It mainly consists of a truncated VGG16 net and detection modules. The detection layers presented in green, are used for multi-scale pedestrian detection. Each detection layer is followed by a detection module for the final prediction. In the detection module, the operator $D_{s,c}$ firstly converts the segmentation mask into the attention map. We then encode the attention map into the feature maps of the multi-scale network via element-wise multiplication operation, activating most relevant parts of the feature maps. The concatenation of convolutional layers with different receptive fields allows to incorporate context information. The 1×1 convolution filter then selects the best features before predictions.

TABLE I
THE HEIGHT OF REFERENCE BOX IN PIXELS ASSOCIATED TO EACH PREDICTION LAYER. THE CONTEXT MODULE PROVIDES TWO DIFFERENT TRFS WHICH MIMICS THE CONTEXT INFORMATION INCORPORATION.

Detection Layer	Box Height	TRF
conv4_3	30, 60	108, 124
conv5_3	90, 120	228, 260
conv_fc7	150, 180	292, 324
conv6_2	240, 270	356, 420
conv7_2	320, 350	484, 612

layers decrease in size and increase in receptive field progressively in order to cover multi-scale pedestrians.

- Detection Layers: We select conv4_3, conv5_3, conv_fc7, conv6_2 and conv7_2 as detection layers according to their increasingly large receptive field sizes.
- Context Module: In two-stage detector, it is common to incorporate context information by enlarging the region proposal. We simulate this effect in a simple convolutional manner. Concretely, a feature map with larger receptive field size is fused with an initial feature map to mimic the context incorporation. Fig. 3 illustrates the details of the context module, where an 1×1 and an 3×3 filters are computed in parallel. The convolutional layer with the kernel size 3×3 has larger receptive field which permits to introduce additional context information.
- Prediction Layer: The context module is followed by two 3×3 convolutional layers to produce classification scores and bounding box offsets respectively.

2) Design of Reference Boxes: We evaluate a series of reference boxes of different aspect ratios at each location in prediction layers. For each reference box, we predict both the shape offsets and the confidence for pedestrian. More the reference boxes concord with the ground truth bounding boxes, easier will be the bounding box regression. Since the reference boxes have an important effect on the regression performance, they are carefully designed based on the receptive field of the prediction layers. According to [64], in the theoretical receptive field (TRF) of a convolutional layer, center pixels have much more impact comparing to the rest, and as a result, the effective area is in general of Gaussian form. Based on this observation, the height of our reference boxes is designed significantly smaller than the TRF in order to match the effective area (see Table I). Once the height of the box is determined, the width is computed according to the aspect ratios of pedestrians: $\{0.25, 0.41, 0.52\}$. As the result, six reference boxes of different scales are considered at each location in the feature maps of prediction layers.

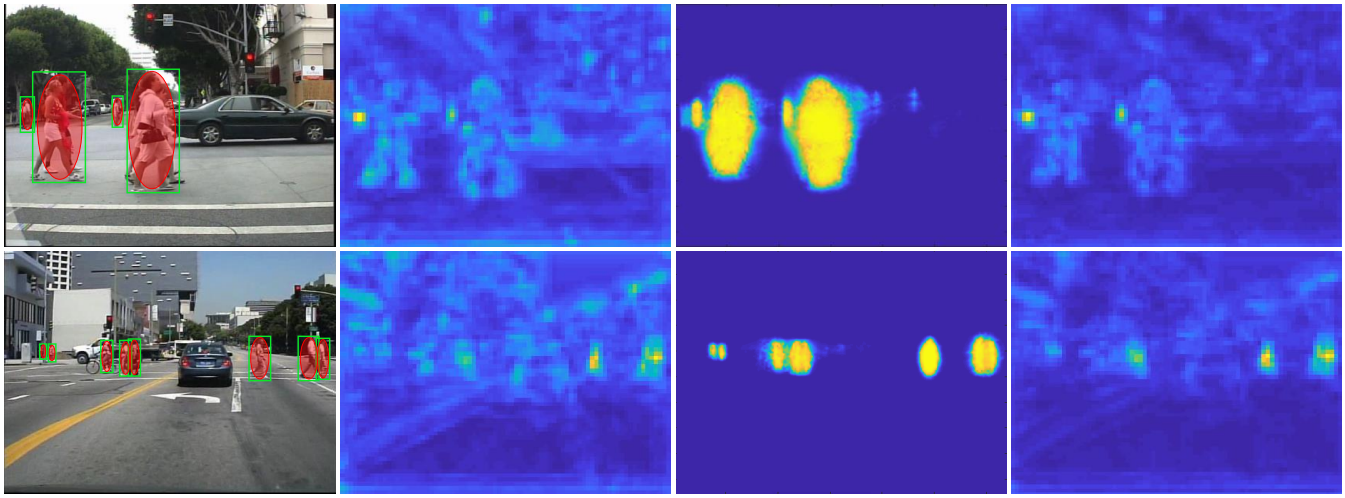
3) Multi-scale Network Learning: Our framework has two sibling output layers. The first outputs bounding-box regression offsets, $\mathbf{d} = (d^x, d^y, d^w, d^h)$. The parameterization for $\mathbf{d}$ is as in [42], in which it specifies a scale-invariant translation and log-space height/width shift relative to a reference box. The second branch outputs the detection confidence score, computed by a softmax over two classes (pedestrian v.s. background). We use the following loss function to supervise the multi-scale network:

$$L_{\text{SAN}} = L_{\text{conf}} + \lambda_b L_{\text{box}} \tag{1}$$

The box regression loss L_{box} targets at minimizing the Smooth L1 loss $R(\mathbf{d}, \hat{\mathbf{g}})$ [42], between the estimated parameters ($\mathbf{d}$) and the ground truth box regression targets ($\hat{\mathbf{g}}$), where $\hat{\mathbf{g}}$ has the same parametrization as $\mathbf{d}$.

$$L_{\text{box}} = \frac{1}{N} \sum_{i \in Pos}^{N} \sum_{k \in \{x,y,w,h\}} x_{ij} R(d_i^k - \hat{g}_j^k) \tag{2}$$

where $x_{ij} = \{1, 0\}$ is an indicator for matching the i-th reference box to the j-th ground truth box and N is the number of matched reference boxes. If $N = 0$, we set the loss of the detector to 0. In our implementation, we begin by matching each ground truth box to the reference box with the best intersection over union (IoU) and we then match reference boxes to any ground truth with IoU higher than 0.5. This training strategy and the definition of reference boxes constraint different layers to detect pedestrians of specific size. The confidence score loss L_{conf} is the softmax loss. In our experiments, we regularize our multi-task loss by setting the weight terms $\lambda_b = 1$.



Input image with ground-truth | Feature maps | Segmentation mask | Feature maps with attention

Fig. 4. Illustration of the segmentation results and the attention map effects on the convolutional feature maps. First column: images with the ground truth bounding boxes drawn in green and the artificial foreground areas presented in red which are used for training the human parsing network. Second column to fourth column: conv4_3 convolutional feature maps of the given images (left), the segmentation mask (middle), and the feature maps with attention map (right). The resulting feature maps highlight pedestrians while ignoring most background regions. Best viewed in color.

B. Human Parsing Network

In parallel with the multi-scale network, the human parsing network generates a semantic segmentation mask which classes the regions that contain pedestrians as foreground and the rest as background. We then convert the mask into an attention map and encode into the feature maps to guide the detection. The pixel-wise classification permits to highlight small-size pedestrian and especially the body parts of occluded instances which are often ignored by full-body detector.

1) Architecture of the Network: The human parsing network is based on the VGG16 network truncated at conv5_3. We change the layer pool4 from 2×2-s2 to 3×3-s1 and adopt the atrous algorithm [65] to compute more dense feature maps. Each convolutional stage (conv2_2, conv3_3 and conv5_3) is up-sampled to generate feature maps at the size of the input image. The concatenation of these hierarchical maps forms discriminative feature maps which are then followed by an 1×1 convolutional layer and a sigmoid layer to output pedestrian segmentation mask. Note that the stem parts (conv1-conv2) are computationally expensive, we share these layers with the multi-scale network. The architecture of the human parsing network is depicted in the top part of Fig. 2.

2) Weakly Supervised Learning: In general, only bounding box annotations are provided in pedestrian detection tasks. Therefore, to train our human parsing network, we follow a weakly supervised strategy by creating artificial foreground segmentation using bounding box information. In practice, we consider the center area of the bounding box (65% of pixels within the box) as foreground, as shown the first column in Fig. 4. This process considerably eliminates background inside the bounding box while keeping the main parts of pedestrian. We use the cross-entropy loss L_{seg} to supervise our human parsing network, which aims to minimize the difference between the prediction and the ground truth mask generated as discussed above. Some segmentation results are illustrated in the third column of Fig. 4. We can verify that, despite the weak annotations, the pedestrians are effectively highlighted.

3) Segmentation-Aware Attention Map: In order to make our detector more robust to small-size targets and occluded pedestrians, we exploit the fine-grained features generated by the human parsing network to help guide the detection. Specifically, by encoding the segmentation-aware attention map into the feature maps of the detection layers, we substantially reduce the background interference and enhance the features representing pedestrians and visible body parts. The occluded targets can be then inferred based on these visible parts. Fig. 3 illustrates the architecture of the detection module with the attention map inserted. Specifically, given the segmentation mask M, we convert it into an attention map by down-sampling the size and increasing the channel number, in order to match with the feature maps of the detection layer $F_{s,c}$. The resulting activated feature maps $A_{s,c}$ can be formulated as:

$$A_{s,c} = D_{s,c}(M) \odot F_{s,c} \tag{3}$$

where $D_{s,c}(M)$ down-samples the segmentation mask M by s times and outputs with c channels. In practice, we down-sample the spatial size of the segmentation mask with average pooling layer. We then harmonize the channel number with 1×1 convolutional layers. $\odot$ represents the Hadamard operator. Some results are depicted in Fig. 4, which show that the conv4_3 feature maps with the attention mechanism become more focused on pedestrians and the background is significantly smoothed.

C. Adversarial Hiding Network

Occlusion cases occur occasionally in the real world applications, *e.g.* pedestrian trying to across the street may be

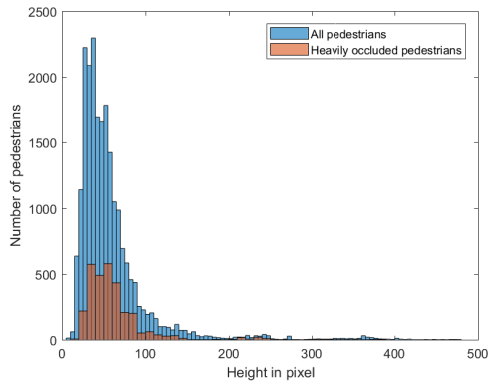


Fig. 5. Distribution of pedestrians with respect to their heights on the Caltech training set. The distribution of all pedestrians is plotted in blue while the distribution of heavily occluded pedestrians are presented in brown.

occluded by cars. Failed detection in the such situation will result in dramatical accident. It is therefore important to improve the robustness of our detector to deal with occlusion issues. Besides the difficulty of the task, the lack of occluded instances consists also a main reason that the current pedestrian detectors struggle to recognize occluded targets. By considering the Caltech training set [14], we observe that there are only 2,973 heavily occluded instances (36-80% occluded) among the total 21,666 pedestrians. Fig. 5 illustrates the distribution of all and occluded pedestrians on Caltech dataset.

1) AHN Overview: In order to increase the number of occluded instances, we develop an adversarial hiding network (AHN) to generate occlusion on pedestrians. Given the feature maps of the input image, the AHN aims to hide the parts of pedestrians making the detector unable to recognize effectively the target. More specifically, our AHN predicts a mask M_{AHN} which is either 0 or 1 after thresholding on pedestrian regions. The AHN hides the body parts by assigning zeros while keeps intact the rest of the feature maps. Let $M_{AHN}^{i,j}$ be the value for the ith row and j^{th} column of the mask. We perform an element-wise multiplication with the feature maps X across all channels, so that if $M_{AHN}^{ij} = 0$, $X_{ijk} = 0$. Where X_{ijk} denotes the value in channel k at location i, j of the feature maps. However, without any constraint, our adversarial hiding network would mask out totally the pedestrians, so that the detector cannot localize any target. In order to avoid this situation, for each bounding box that contains a pedestrian, we only mask out the $\tau\%$ areas within the bounding box, which correspond to the most relevant parts for our adversarial hiding network. In contrast to our human parsing network which aims to highlight pedestrian and helps the detector, our adversarial hiding network tends to hide most important parts of body and force the detector to focus on less discriminative details.

2) Network Architecture: The adversarial hiding network needs firstly to find out the pedestrians, before effectively masking out the body parts. We therefore build the AHN on top of our human parsing network to avoid redundant computations. The adversarial hiding network is composed of 5 convolutional layers and takes the last feature maps from the human parsing network as input. Fig. 6 gives an overview of our framework with the adversarial hiding network.

3) Adversarial Learning: The goal of our adversarial hiding network is to make the detector least capable to distinguish pedestrian by masking out body parts. Mathematically, let $\mathcal{F}(X)$ represent the original pedestrian detector, where X corresponds to the feature maps. The detector outputs confidence scores $\mathcal{F}_c$ and bounding box locations $\mathcal{F}_l$ at multiple positions. When training the pedestrian detector, we aim to minimize the difference between the prediction and the ground truth bounding boxes. In contrast, the adversary tends to maximize this difference. We supervise therefore our adversarial hiding network via the following loss function:

$$L_{AHN} = -L_{conf}(\mathcal{F}_c(\mathcal{A}(X)), G_c) \quad (4)$$

where $\mathcal{A}(X)$ denotes the feature maps with artificial occlusion generated by the adversarial hiding network and G_c is the ground truth label. The loss will be high when the generated feature maps with occlusion are easy for the detector. If the relevant body parts are effectively hided, the detector becomes unable to find out the pedestrian. We will then get a high loss for the detector while a low loss for the adversarial hiding network. It is worth to note the loss supervising our human parsing network can also be used for training our adversarial hiding network. However, the segmentation loss is pixel-wise based and is less helpful as we constantly mask out $\tau\%$ pedestrian parts. Without this constraint, our AHN will inevitably hide the entire body of the targets, which makes the detector unable to see anything, leading thus to a high detection loss.

The output of the adversarial hiding network (AHN) is a mask with values between 0 and 1, and we aim this network to return a binary mask with values as 0 or 1. However, the back-propagation of gradients would be impossible with a sampling operation. In order to avoid this problem, instead of a hard sampling operation, we perform a soft sampling process with an intermediate supervision signal during the training. To be specific, during the training of our adversarial hiding network, we generate a pseudo-ground-truth binary mask based on the output of the AHN and use it as a supervision signal. To generate this pseudo-ground-truth binary mask, we set the 40% smallest values inside the bounding boxes containing pedestrians to 0 while set the rests and the points which are outside of the bounding boxes to 1. With this artificial intermediate supervision signal, the adversarial hiding network tends to yield a binary mask and is constrained to hide 40% parts of pedestrians. In practice, we use a cross-entropy loss between the pseudo-ground-truth binary mask and the output of the AHN for learning. The hiding mask generated by the AHN is then encoded into the feature maps for detection and can receive information from the detector by back propagation as we do not directly employ any sampling operation. The adversarial loss function becomes as follows:

$$L_{AHN} = -L_{conf}(\mathcal{F}_c(\mathcal{A}(X)), G_c) - L_{CE}(\mathcal{A}, \mathcal{M}_b) \quad (5)$$

where $L_{CE}(\mathcal{A}, \mathcal{M}_b)$ corresponds to the cross-entropy loss between the output of the adversarial hiding network $\mathcal{A}$ and the pseudo-ground-truth binary mask $\mathcal{M}_b$. Note that during the

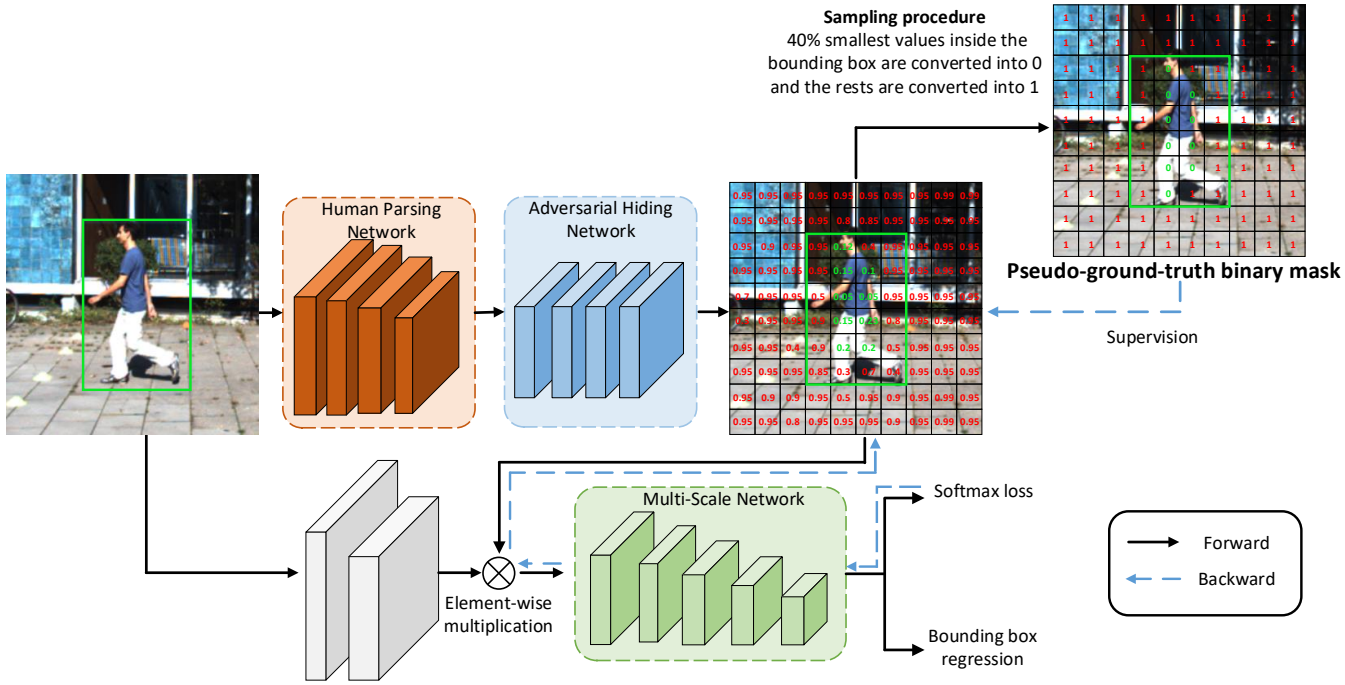


Fig. 6. Training pipeline of the adversarial hiding network (AHN). Given the output of the AHN, we generate a pseudo-ground-truth binary mask to supervise the adversarial network. This supervision makes the adversarial hiding network compute a binary mask and only hide 40% of pedestrian. In addition, with this strategy the information from the multi-scale network can be directly back-propagated.

Algorithm 1 MDFL+Adv Training

Input: Training images, annotations
Output: Pedestrian detector
Step 1: Freeze the well trained MDFL
Step 2: Train the adversarial hiding network with the loss function L_{AHN}
Step 3: Optimize the multi-scale network with occluded feature maps

testing, we can use a hard sampling process to have a strict binary mask.

D. Implementation Details

1) MDFL Training: Our multi-scale network and human parsing network were partially initialized with the detection model of [47] and the DeepLab segmentation model [65], respectively. All new additional layers were randomly initialized with the Xavier [66]. In order to facilitate the convergence, we first trained the two networks separately and then the both networks were jointly optimized. Specifically, the multi-scale network was fine-tuned for $50k$ iterations where we used 10^{-4} learning rate for the first $40k$ iterations then continued with 10^{-5} for the rest iterations. The human parsing network was fine-tuned for $80k$ iterations with a learning rate of 10^{-8}. The both networks are then jointly optimized for $10k$ iterations supervised with the following multi-task loss:

$$L = L_{\text{conf}} + \lambda_b L_{\text{box}} + \lambda_s L_{\text{seg}} \quad (6)$$

where $\lambda_b = 2$ and $\lambda_s = 1$ are the weight terms to balance the loss. All our implementations were based on Caffe framework [67].

2) Adversarial Training: The adversarial hiding network and the detector were trained alternatively. We first trained our multi-scale network and human parsing network which form an effective pedestrian detector. Given the well trained pedestrian detector, we froze the parameters of the model and only optimized the adversarial hiding network. In turn, once the adversarial hiding network was able to hide the relevant body parts, we fixed its parameters and trained the multi-scale network with the occluded feature maps. Note that since the AHN is constructed on top of the human parsing network, the latter remains frozen and will not be trained again. In our experiments, the adversarial hiding network was initialized with the Xavier [66] and was trained for $50k$ using a learning rate of 10^{-6}. The multi-scale network was then trained for $8k$ iterations with occluded feature maps using a learning rate of 10^{-6}. Algorithm 1 summarizes the adversarial learning steps.

3) Hard Negative Mining: Our detector has to evaluate a considerable number of reference boxes, yet only a few locations contain pedestrians, which causes a significant class imbalance during the training. For more stable training, instead of using all negative samples, we sorted them by the highest loss values and kept the top ones so that the ratio between the negatives and positives is at most 5:1. We filtered out most simple samples and made the detector focus on the foreground and the most confusing negative instances.

4) Data Augmentation: To make our model more robust to sizes and illumination variations, we adopted following

data augmentation strategies: color distortion, expansion and horizontal flip [47]. We randomly expanded the training image with a factor $\alpha \in [1, 4]$ to create more small training examples.

IV. Experiments

A. Datasets and Evaluation Protocols

We conducted experiments on three challenging pedestrian detection datasets, Caltech [14], KITTI [20] and INRIA [15], to evaluate our proposed method and compared it with state-of-the-art pedestrian detection approaches. Here we give a brief description of these datasets.

1) Caltech [14]: The Caltech dataset is one of most challenging pedestrian detection benchmark due to the presence of large number of small targets. We can see in Fig. 5 that about 60% of the pedestrians from the Caltech training set have a height smaller than 100 pixels. The dataset consists of about 10 hours videos collected from a vehicle driving through regular urban traffic. The log-average miss rate is used to evaluate the detection performance and is calculated by averaging miss rates at 9 false positive per-image (FPPI) points sampled within the range of $[10^{-2}, 10^{0}]$. In our experiments, three subsets were considered to demonstrate the performance on occlusion and small-size issues: *Heavy Occluded*, *Medium* and *Overall*. In the *Heavy Occluded* subset, pedestrians are 36-80% occluded; in the *Medium* subset, pedestrians are of 30-80 pixels height without occlusion; and the *Overall* subset consists of all pedestrians taller than 20 pixels with or non occlusion.

2) KITTI [20]: The KITTI dataset contains 7,481 training images and 7,518 test images, comprising about 80 thousands annotations of cars, pedestrians and cyclists. KITTI evaluates the PASCAL-style mean Average Precision (mAP) under three difficulty levels: easy, moderate and hard. Under moderate setting, which is used to rank the competing methods in the benchmark, the pedestrians taller than 25 pixels with or non occlusion are considered.

3) INRIA [15]: The INRIA dataset includes 614 positive and 1,218 negative training images. There are 288 test images available for evaluating pedestrian detection methods. The log-average miss rate on FPPI is employed as the evaluation metric. In the INRIA dataset, pedestrians taller than 50 pixels with partial or non occlusion are considered for evaluation.

B. Comparison with State-of-the-art Methods

1) Caltech: We used the Caltech training set, which contains 42,782 training images, to train our detection system and evaluated it on the Caltech testing set. Among the ground truth annotations, we considered 'person' and 'people' as targets to be detected and ignored the rests. We compared our proposed MDFL and MDFL+Adv models with the methods that have achieved great performance on Caltech [6], [7], [11], [17], [19], [44], [46], [49], [68]. As shown in Fig. 7 and Fig. 8, our MDFL method achieves 31.46% and 46.85% miss rate on the *Medium* and *Overall* subsets respectively, which outperforms the current state-of-the-art methods. The performance on the *Medium* subset shows the capability of our approach to deal with small-size pedestrians. Our MDFL method achieves great results on the *Overall* subset, surpassing the competitive approaches by 3%. The performance under this more general setting points out the robustness of our method, which can perform well under different conditions. Our MDLF method achieves an impressive 38.53% miss rate on the *Heavy Occluded* subset, which outperforms considerably the current pedestrian detectors. When we further introduce the adversarial hiding network to generate the occlusion situations, we make the detector even more robust and get 37.45% miss rate on the *Heavy Occluded* subset. The second column of Table II shows the comparison with the state-of-the-art methods using the *Heavy Occluded* setting. The comparison with the recent part-detector based approach [11] which has achieved 49.20% miss rate, demonstrates the effectiveness of our MDFL and MDFL+Adv models to handle occlusion issues.

However the performances on the *Medium* and *Overall* subsets are slightly impacted by the adversary training. To understand this degradation, we qualitatively analyzed the

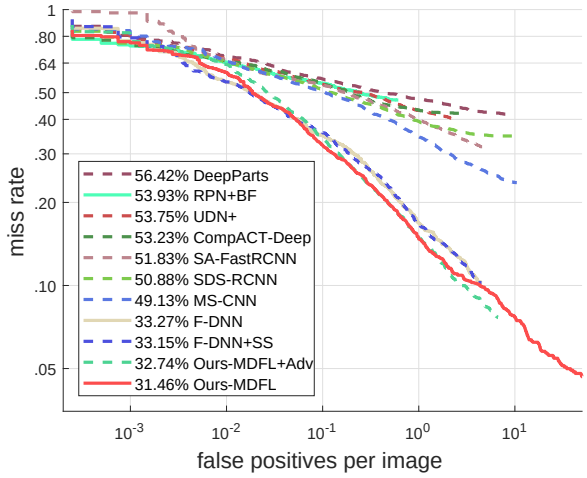


Fig. 7. Comparison with the state-of-the-art methods on the Caltech dataset using the *medium* setting.

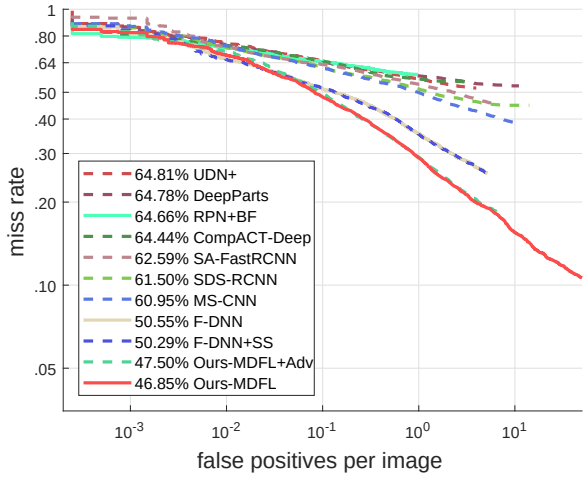


Fig. 8. Comparison with the state-of-the-art methods on the Caltech dataset using the *overall* setting.

Fig. 9. Some confusing examples. The detector is alarmed by the objects that have similar features with occluded pedestrians.

TABLE II
COMPARISON OF OUR METHOD WITH THE STATE-OF-THE-ART APPROACHES IN TERMS OF TRADE-OFF BETWEEN ACCURACY AND SPEED. CALTECH HEAVY OCCLUDED MISS RATE (%), KITTI MAP SCORE (%) AND RUNTIME ARE TABULATED.

Method	Caltech	KITTI	Runtime
RPN+BF [6]	74.36	61.29	0.5s
SA-FastRCNN [44]	64.35	65.01	0.5s
DeepParts [49]	60.42	58.67	1s
MS-CNN [7]	59.94	73.70	0.14s
SDS-RCNN [17]	58.55	63.05	0.21s
F-DNN [46]	55.13	-	0.3s
F-DNN+SS [46]	53.76	-	2.48s
JL-Tops [11]	49.20	-	0.6s
Ours-MDFL	38.53	66.32	0.07s
Ours-MDFL+Adv	37.45	67.29	0.07s

results and found out the following reason. Our model is more robust to occlusion issues thanks to the supplement artificial occlusion instances during the training. However, with this strategy, our model becomes more likely to be alarmed by objects that share similar features with occluded pedestrians. Fig. 9 shows some failed examples where our model is confused by the objects resembling to pedestrians occluded by cars.

2) KITTI: We used the KITTI training set to train our pedestrian detector and evaluated on the designated test set. As our main task is pedestrian detection, we only considered the pedestrian class. Our MDFL and MDFL+Adv methods achieve competitive 66.32%mAP and 67.29%mAP, respectively on the moderate setting for pedestrian class, which outperform most approaches [6], [17], [44], [49]. The comparison results are tabulated in the third column of Table II. The adversary brings an improvement of 1%mAP to our MDFL model. This result proves the capability of our adversary to make the pedestrian detector be more robust to the occlusion issues.

Note that in the KITTI evaluation, cyclists are counted as false positives and persons-sitting are ignored, while in Caltech these two classes are labeled as pedestrians. Since the semantic segmentation information are useful for detecting human body, this advantage is less helpful on the KITTI dataset. Fig. 10 illustrates some examples of images containing cyclists and pedestrians in unusual pose from the Caltech test set. Cyclists

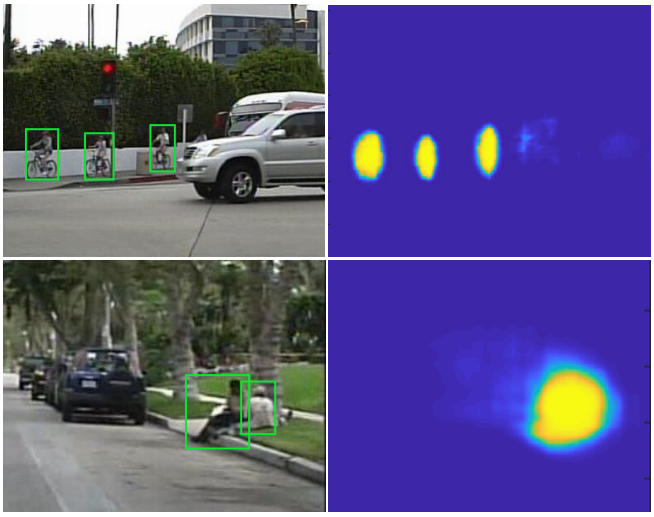


Fig. 10. Illustration of the effects of the attention mask on cyclists and pedestrians in unusual pose. As they share similar features with pedestrians, our human parsing network will consider them as pedestrians.

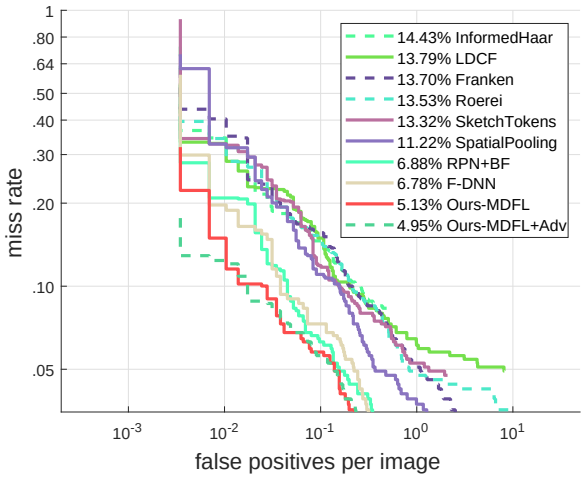


Fig. 11. Comparison with the state-of-the-art methods on the INRIA dataset.

and sitting people have very similar features with pedestrian, especially the top parts of the body. The attention mask tends to activate these instances even if we ignored these two classes during the training on KITTI. This capability is helpful on the Caltech dataset, but can be harmful on the KITTI dataset which considers cyclists as an independent class and ignores persons-sitting.

3) INRIA: We trained our model with 614 positive images by excluding negative images as there is non pedestrians in these images. We evaluated our method on the available test set. Fig. 11 illustrates the results of our approaches and the methods that perform best on the INRIA dataset [37], [69]–[73]. Our MDFL and MDFL+Adv detectors achieve state-of-the-art performances with 5.13% and 4.95% miss rate, outperforming the previous best competitors by about 2 points. The comparison demonstrates the superiority of our method. Since the INRIA dataset is relatively small, this performance

TABLE III
MISS RATE (%) ANALYSIS ON THE CALTECH TEST SET WHEN THE KEY COMPONENTS OF OUR MDFL FRAMEWORK ARE SUCCESSIVELY DISABLED.

Component disabled	Medium	Heavy Occ.	Overall
Multi-scale	42.85	48.23	55.53
Context module	35.31	44.37	49.99
Attention map	33.27	40.27	47.83
Our MDFL	31.46	38.53	46.85

TABLE IV
COMPARISON OF MISS RATE (%) WITH DIFFERENT SUPERVISION STRATEGIES FOR THE HUMAN PARSING NETWORK.

Method	Medium	Heavy Occ.	Overall
SAN	33.27	40.27	47.83
SAN+Att_{center}	**31.46**	**38.53**	**46.85**
SAN+Att_{all}	32.39	40.51	47.45

points out that our model can converge easily and achieves great results even if the training set is limited.

4) Efficiency Comparison: Most of the real world applications require that the pedestrian detector be accurate and execute at real-time. It is therefore necessary to evaluate the computation time. Our method executes 0.07 second per image with an input image of size 640×480 on a single Nvidia 1080Ti GPU. Compared to the most methods, our approach executes $2\times$ faster (fourth column of Table II). Specifically, compared to JL-Tops [11], which was proposed to handle occlusion issue, our method is $8\times$ faster. The previous best approach on the Caltech medium subset F-DNN+SS [46] needs 2.48 seconds to process one image, which is $35\times$ slower than our method. The comparison shows the efficiency of our proposed MDFL detector which achieves great detection performance while executes much faster than competitive methods. Since the adversarial hiding network is only introduced for the training, the computing time of our MDFL+Adv model remains the same as the MDFL detector.

C. Ablation Studies

1) MDFL Analysis: In order to analyze the contribution of the key components of our MDFL framework on the detection performance, we successively removed each component and evaluated on Caltech. We mainly studied the effects of the attention map generated by our human parsing network, the context module and the multi-scale property. Table III summarizes the results of the ablative experiments.

We first disabled the segmentation-aware attention map and observed a degradation of the performance by 2% on the *Medium* and *Heavy Occluded* subsets and 1% on the *Overall* subset. These results point out the importance of the fine-grained information for recognizing small and occluded pedestrians. The attention map effectively guides our detector to focus on pedestrian regions and in particular visible human body parts. The human parsing network alleviates the issue that heavy occluded instances are often ignored by the multi-scale network which has coarser resolution and tends to be confused by the background.

When we further removed the context module, the performance for detecting small-size targets and occluded pedestrians dropped by 2% and 4%, respectively. The context information have great impacts for our detector to distinguish pedestrians of small size and occluded instances from the complex background. These experiments confirm our intuition that the context information are helpful for inferring correctly small and/or occluded targets.

We then analyzed the effects of our multi-scale network which allows the detection of instances of difference sizes with appropriate feature maps. Instead of multiple detection layers, we only used the conv_fc7 layer for prediction. We placed all the reference boxes on top of this detection layer. The results degrade by more than 4% on all subsets, which demonstrates the importance to use the multi-scale strategy to effectively detect pedestrians of different sizes.

2) Weakly Supervised Attention Map: As we only used the bounding box annotations to weakly supervise our human parsing network, it is important to analysis how to explore the pixels within the box. We investigated two possibilities: (a) consider 65% of pixels at the center area of the bounding box as foreground and the rest as background, which we denote Att_{center}. Some examples are illustrated in the first column of Fig. 4; (b) consider all pixels within the bounding box as foreground, which we denote Att_{all}. The experiments results are reported in Table IV. We can see that SAN+Att_{center} achieves better miss rate than SAN+Att_{all}. As the SAN+Att_{all} method takes into the account considerable number of background for training, the model is more sensitive and more likely to be confused by background, which leads to an increase of false alarms. While in the SAN+Att_{center} approach, we minimize the noises caused by the background and make the human parsing network focus on the supervision signal from the main body of pedestrians.

3) Occlusion Strategies Analysis: We compared our adversary with different occlusion strategies to examine its effect. Some examples of these strategies are illustrated in Fig. 12. We performed the experiments on the Caltech test set using the heavy occluded setting. We masked out the same $\tau = 40\%$ areas of pedestrian in the following analysis for fair comparison. Table V summarizes the impacts of the different strategies on the detection performance.

The first simple strategy consists to randomly occlude the human body parts on the convolutional feature maps (Fig. 12(a)). As Table V shows, the miss rate of random occlusion is 39.35% which is slightly worse than our MDFL. This drop of performance can be explained by the following reason. The random occlusion disperses on the entire space within the pedestrian region. Instead of generating effective occlusion, this strategy introduces noises.

We then tried a bottom-occlusion strategy (Fig. 12(b)) since this type of occlusion occurs most frequently in real world scene (Fig. 1). Specifically, given the bounding box containing a pedestrian, we masked out $\tau\%$ bottom part of the region within the box. With this strategy, we obtain 37.68% miss rate, which is slightly better than our MDFL model. This performance demonstrates the effectiveness of the data augmentation to make the detector more robust to occlusion situation.

Instead of randomly masking out or occluding all the

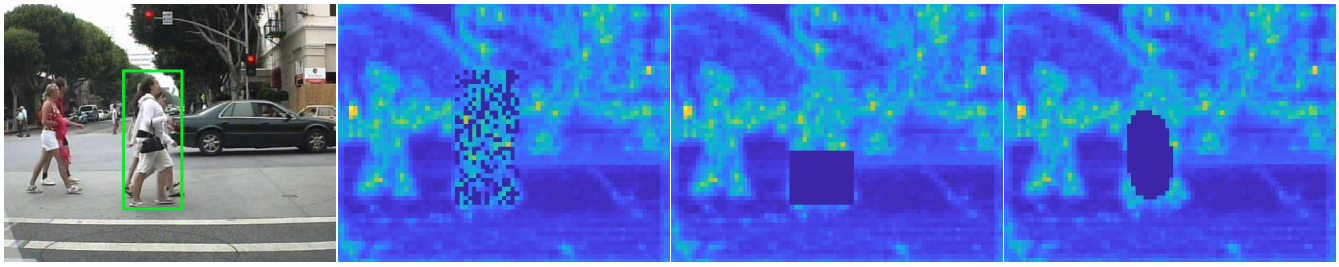


Fig. 12. Some examples of various occlusion strategies. Given an input image and a bounding box containing pedestrian, these strategies generate different types of occlusion and hide 40% areas within the bounding box. From left to right, we (a) randomly mask out areas, (b) hide the bottom parts, and (c) learn an adversarial hiding network which aims to mask out most relevant parts.

Fig. 13. Qualitative comparison of pedestrian detection results with other state-of-the-art methods. The first column shows the input images with the ground truth bounding boxes drawn in red. The rest columns display the detection results (green bounding boxes) of F-DNN+SS [46], SDS-RCNN [17] and our MDLF+Adv respectively. We illustrate the predicted bounding boxes with confidence score higher than 0.2. Our proposed method successfully detects heavily occluded pedestrians and is more robust compared to other approaches. Best viewed in color.

TABLE V
THE EFFECT OF VARIOUS OCCLUSION STRATEGIES ON THE DETECTION PERFORMANCE UNDER OCCLUSION CONDITIONS. ABLATION EXPERIMENTS EVALUATED ON THE CALTECH HEAVY OCCLUDED SUBSET.

Method	Miss rate (%)
MDFL	38.53
MDFL+random occlusion	39.35
MDFL+bottom occlusion	37.68
MDFL+Adv	**37.45**

bottom part of the bounding box containing pedestrian, we then explored the adversary (Fig. 12(c)). Our AHN learns automatically to hide the most relevant body parts that make the detector less capable to recognize the targets. As Fig. 12 shows, our AHN tends to mask out the main parts of human body, letting only the head and feet visible. With this strategy, we get 37.45% miss rate, which represents 1% improvement compared to our MDFL approach without artificial occlusion. The result demonstrates the superiority of our adversary occlusion strategy compared to the random and handcrafted occlusion manners.

4) Dropout Percentage: How much parts should we hide is a crucial question and has non negligible effect on the detection performance. In order to answer this question, we investigated the impact of different values of the parameter τ. Table VI tabulates the detection performance according to this parameter. The experiments point out that with a large value of τ, we force the detector to focus on the least important

TABLE VI
DETECTION PERFORMANCE WITH DIFFERENT LEVELS OF ARTIFICIAL HIDING, REPRESENTED BY THE PARAMETER τ. ABLATION EXPERIMENTS EVALUATED ON THE CALTECH HEAVY OCCLUDED TEST SUBSET.

MDFL+Adv	Miss rate (%)
$\tau = 20\%$	39.06
$\tau = 30\%$	37.88
$\tau = 40\%$	**37.45**
$\tau = 50\%$	38.26
$\tau = 60\%$	39.19

body parts, which increases the false detection rate and results in the drop of performance. The resulting features are not relevant enough for recognizing pedestrians and make the detector confused. For example, when we mask out 60% areas within the pedestrian region, the performance drops to 39.19%. While with small value of the parameter, the occlusion effect is limited and can be considered as noises. With $\tau = 30\%$, the detection performance is slightly better than our MDFL. However, when we mask out only 20% parts, the result deteriorates to 39.06%. According to the above experiments, we demonstrate that hiding more parts of pedestrian does not necessarily lead to a better performance. A trade-off should be made and it seems masking out 40% parts of human body leads to best results, with a miss rate of 37.45%.

5) Visualization of Detection Results: Qualitative detection results of our method and state-of-the-art approaches [17], [46] are illustrated in Fig. 13. The first column shows the input images with the ground truth and the rest three columns sequentially show the detection results by F-DNN+SS [46], SDS-RCNN [17] and our method. Our proposed detector successfully locates heavily occluded pedestrians which the other two methods have missed. The miss detection in the situation of the second image endangers the pedestrian who may across the road. Due to occlusion issue, these detectors are unable to warn the users to anticipate, which may lead to dramatical accident. In addition, our proposed approach gets better localization and has much less redundant detections compared to the two other methods. The qualitative detection results further demonstrate the superiority of our MDFL+Adv method in detecting occluded and small-size instances.

V. CONCLUSION

In this paper, we have proposed a multi-grained deep feature learning based method for pedestrian detection. By jointly training a multi-scale network and a human parsing generator, our approach exploits pixel-wise segmentation information, background context and multi-scale property to simultaneously handle the occlusion and small-size issues. The whole detection system is a single stage framework, assuring a great accuracy/speed trade-off. We have further proposed an adversarial hiding network, which artificially generates occluded instances, to make our detector more robust to occlusion issues. The proposed method has achieved impressive performance on challenging pedestrian detection datasets such as Caltech, KITTI and INRIA, outperforming most existing approaches while executing $2\times$ faster.

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant 61672306, Grant U1713214, Grant 61572271, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564.

REFERENCES

[1] S. Zhang, C. Bauckhage, D. A. Klein, and A. B. Cremers, "Exploring human vision driven features for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1709–1720, 2015.
[2] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.
[3] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1259–1267.
[4] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, 2014.
[5] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, no. 2, 2015, p. 4.
[6] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–457.
[7] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.
[8] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5079–5087.
[9] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5363–5371.
[10] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body-part semantic and contextual information with dnn," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3148–3159, 2018.
[11] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3486–3495.
[12] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, 2018.
[13] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 732–747.
[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 304–311.
[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[16] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2903–2910.
[17] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4950–4959.
[18] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli, "Pushing the limits of deep cnns for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1358–1368, 2018.
[19] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1874–1887, 2018.
[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[21] C. Lin, J. Lu, and J. Zhou, "Multi-grained deep feature learning for pedestrian detection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, accepted, 2018.

[22] J. Lu, V. E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, 2018.

[23] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-aware local binary feature learning for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1139–1153, 2018.

[24] H. Liu, J. Lu, J. Feng, and J. Zhou, "Two-stream transformer networks for video-based face alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2546–2554, 2018.

[25] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, 2015.

[26] J. Lu, J. Hu, and Y.-P. Tan, "Discriminative deep metric learning for face and kinship verification," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4269–4282, 2017.

[27] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, 2015.

[28] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Learning rotation-invariant local binary descriptor," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3636–3651, 2017.

[29] J. Lu, J. Hu, and J. Zhou, "Deep metric learning for visual understanding: An overview of recent advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76–84, 2017.

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

[31] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1243–1257, 2016.

[32] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1140–1151, 2008.

[33] P. Dollár, S. J. Belongie, and P. Perona, "The fastest pedestrian detector in the west." in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, no. 3, 2010, p. 7.

[34] M. Jeong, B. C. Ko, and J.-Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1368–1380, 2017.

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

[36] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. S. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades." in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, 2015, p. 4.

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

[38] M. You, Y. Zhang, C. Shen, and X. Zhang, "An extended filtered channel framework for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1640–1651, 2018.

[39] J. Lu, V. E. Liong, and J. Zhou, "Deep hashing for scalable image search," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2352–2367, 2017.

[40] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 281–292, 2018.

[41] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[42] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

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

[44] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

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

[46] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Wint. Conf. Appli. Comp. Vis.*, 2017, pp. 953–961.

[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[48] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3258–3265.

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

[50] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2123–2137, 2016.

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

[52] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[53] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, 2018.

[54] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, accepted, 2018.

[55] G.-J. Qi, "Hierarchically gated deep networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2267–2275.

[56] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3150–3158.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[58] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3127–3136.

[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[60] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[61] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, and X.-s. Hua, "An adversarial approach to hard triplet generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 501–517.

[62] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1222–1230.

[63] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3039–3048.

[64] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4898–4906.

[65] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[66] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 249–256.

[67] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. Int. Conf. Multimedia*, 2014, pp. 675–678.

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

[69] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 947–954.

[70] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool, "Handling occlusions with franken-classifiers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1505–1512.

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

[72] J. J. Lim, C. L. Zitnick, and P. Dollár, "Sketch tokens: A learned mid-level representation for contour and object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3158–3165.
[73] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 546–561.

Chunze Lin received the B.S. degree in engineering from Ecole Centrale de Nantes, France. He is currently pursuing the M.Eng degree at the department of Automation, Tsinghua University, China. His research interests include computer vision, pattern recognition and deep learning.

Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored/co-authored over 200 scientific papers in these areas, where 60 of them are IEEE Transactions papers (including 11 T-PAMI papers) and 40 of them are CVPR/ICCV/ECCV/NIPS papers. He serves an Associate Editor for several international journals including the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Biometrics, Behavior, and Identity Science, and Pattern Recognition. He is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He also served as Workshop Chair/Special Session Chair/Area Chair for more than 20 international conferences such as ICIP, ICPR, ICME, ACCV and WACV. He was a recipient of the National 1000 Young Talents Program of China in 2015, and the National Science Fund of China for Excellent Young Scholars in 2018, respectively. He is a senior member of the IEEE.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.