# Graininess-aware Deep Feature Learning for Robust Pedestrian Detection

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

***Abstract*—In this paper, we propose a graininess-aware deep feature learning method for pedestrian detection. Unlike most existing methods which utilize the convolutional features without explicit distinction, we appropriately exploit multiple convolutional layers and dynamically select most informative features. Specifically, we train a multi-scale pedestrian attention via pixel-wise segmentation supervision to efficiently identify the pedestrian of particular scales. We encodes the fine-grained attention map into the feature maps of the detection layers to guide them to highlight the pedestrians of specific scale and avoid the background interference. The graininess-aware feature maps generated with our attention mechanism are more focused on pedestrians, and in particular on the small-scale and occluded targets. We further introduce a zoom-in-zoom-out module to enhances the features by incorporating local details and context information. Extensive experimental results on five challenging pedestrian detection benchmarks show that our method achieves very competitive or even better performance with the state-of-the-arts and is faster than most existing approaches.**

***Index Terms*—Pedestrian detection, attention, deep learning, graininess**

## I. Introduction

Pedestrian detection is an important research topic in computer vision and has attracted a considerable attention over past few years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. It plays a key role in several applications such as autonomous driving, robotics and intelligent video surveillance. Despite the recent progress, pedestrian detection task still remains a challenging problem because of large variation of scales, low resolution of small-size targets and occlusion issues.

Existing methods for pedestrian detection can mainly be grouped into two categories: hand-crafted features based [2], [3], [12], [13] and deep learning features based [6], [7], [8], [9]. In the first category, human shape based features such as Haar [1] and HOG [2] are extracted to train SVM [2] or boosting classifiers [3]. While these methods are sufficient for simple applications, these hand-crafted feature representations are not robust enough for detecting pedestrian in complex scenes. In the second category, deep convolutional neural network (CNN) learns high-level semantic features from raw pixels and assimilates useful information from large amount of data, which shows more discriminative capability to recognize pedestrian with complex poses from noisy background [14], [15], [16]. Deep learning features have significantly improved pedestrian detection performance and many CNN based methods have recently been proposed [6], [17], [8], [18], [19]. However, there are still some shortcomings for most of the methods in this category. 1) They usually employ the convolutional features without distinction, which leads to sub-optimal results due to the distraction from complex background, such as occlusion and hard negative samples. The detectors tend to be confused by background interference, which leads to large number of miss detections of occluded pedestrians and false alarms due to cluttered background. 2) Deepest convolutional layer with coarse resolution and large receptive field is often used for prediction, which is inefficient for localizing small-scale targets, despite its high-level semantic features. 3) Most methods employ heavy deep network and need refinement stage to boost the detection results. The inference time is scarified to ensure accuracy, making these methods unsuitable for real-time application.

In this paper, we propose a graininess-aware deep feature learning (GDFL) method for pedestrian detection to address the above issues. Instead of only using deepest layer and treating all features equally, we utilize appropriate convolutional layers for multi-scale detection, and enhance the features about human body parts while discard the interference features due to background. Specifically, we employ multiple convolutional layers with different resolutions and receptive field sizes for detection and introduce a multi-scale pedestrian attention mechanism to guide the detector to focus on pedestrian regions. We train the attention via the pixel-wise segmentation supervision signals to assimilate fine-grained information and acquire high capability to recognize small-scale targets and human body parts. By encoding the attention into the convolutional feature maps of detection branches, they significantly eliminate background interference while highlight pedestrians of specific scale for each branch. The resulting graininess-aware deep features have much more discriminative capability to distinguish pedestrians, especially the small-scale and occluded ones from cluttered background. Fig. 1 illustrates the effect of the multi-scale attention on the feature maps, where shallower layers become more focus on small-scale pedestrians while deeper layers on large-scale targets. In addition, we further propose a zoom-in-zoom-out module (ZIZOM) to improve the detection performance of small-scale targets. It mimics the intuitive zoom in and zoom out processes of the human-annotators, when they aim to locate an object in an image. The module incorporates local details

Chunze Lin, Jiwen Lu and Jie Zhou 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. Email: lcz16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn.

Gang Wang is with AI Laboratories, Alibaba Group, Hanzhou, 310052, China. Email: gangwang6@gmail.com.

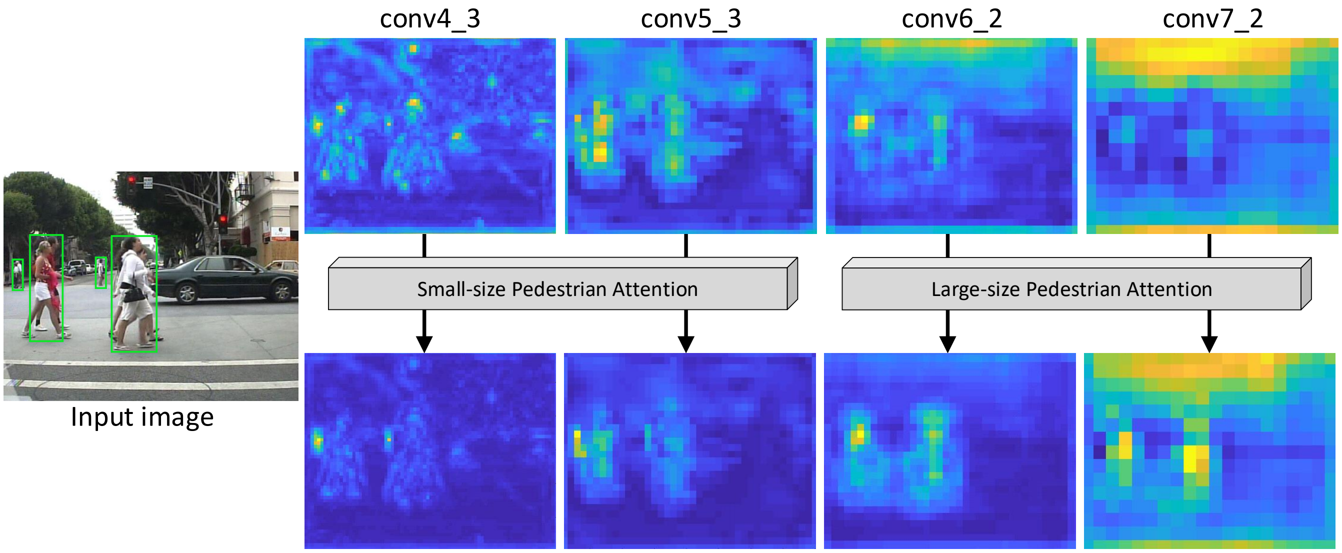

conv4_3
conv5_3
conv6_2
conv7_2
Small-size Pedestrian Attention
Large-size Pedestrian Attention
Input image

Fig. 1. Visualization of the original feature maps from different detection layers of the backbone network (top), and graininess-aware feature maps obtained with our pedestrian attention (bottom). With our attention mechanism, the background interference is significantly attenuated and each detection layer is more focused on pedestrians of specific size. Best viewed in color.

and context information in a convolutional manner to enhance the feature maps of shallower layers which are responsive of small-scale targets detection. Extensive experimental results on five widely used pedestrian detection benchmarks demonstrate the effectiveness of the proposed method. Without any extra refinement steps, our single stage detector achieves competitive performance on Caltech [20], INRIA [2], KITTI [21], MOT17Det [22] and CityPersons [23] datasets, and can execute about 4 times faster than competitive methods.

This paper is an extension of our conference paper [24]. In summary, the main contributions of this paper are as follows:

1) We propose to learn a multi-scale spatial attention via pixel-level supervision which has high capability to identify pedestrian body parts. The attention guides the detector to focus on pedestrians and avoids the background interference.
2) We introduce a zoom-in-zoom-out module to enhance the feature maps of shallow layers by incorporating details and context information.
3) Based on the convoultional layer dissection results, we propose a more natural multi-scale attention framework, which simultaneously guides the detector at spatial and channel dimensions.
4) We conduct extensive experiments on five challenging pedestrian detection benchmarks and achieve very competitive and state-of-the-art performance. Furthermore, we perform comprehensive ablation study to analyze the contribution of main components.

## II. Related Work

In this section, we briefly review three related topics: pedestrian detection, segmentation in detection, and attention mechanism.

### A. Pedestrian Detection

Existing pedestrian detection methods can mainly be grouped into two categories: hand-crafted feature based and deep learning feature based. The Integratal Channel Features (ICF) [3] is among the most efficient pedestrian detectors without deep feature learning, which applies integral channel features with respect to oriented gradient (HOG), color features (LUV) and gradient magnitude, and employed boosted decision forests as classifier. Due to the success of the ICF framework, the feature representations of ICF have been widely studied and many variants have been proposed [25], [26], [27], [13], [28], [29], [30].

With the prevalence of deep convolutional neural network, which has achieved impressive results in various domains, most recent pedestrian detection methods are CNN-based. Many methods were variations of Faster R-CNN [31] which has shown great accuracy in general object detection. RPN+BF [5] replaced the downstream classifier of Faster R-CNN with a boosted forest and used aggregated features with a hard mining strategy to boost the small size pedestrian detection performance. SA-FastRCNN [32] and MS-CNN [33] extended Fast and Faster R-CNN [34], [31] with a multi-scale network to deal with the scale variations problem, respectively. Instead of a single downstream classifier, F-DNN [7] employed multiple deep classifiers in parallel to post verify each region proposal using a soft-reject strategy. Different from these two stages methods, our proposed approach directly outputs detection results without post-processing [35], [36]. Apart the above full-body detectors, several human part based methods [37], [38], [39], [10], [4], [40], [8] have been introduced to handle occlusion issues. These occlusion-specific methods learned a set of part-detector, where each one was responsive to detect a human part. The results from these part detections were then fused properly for locating partially occluded pedestrians.

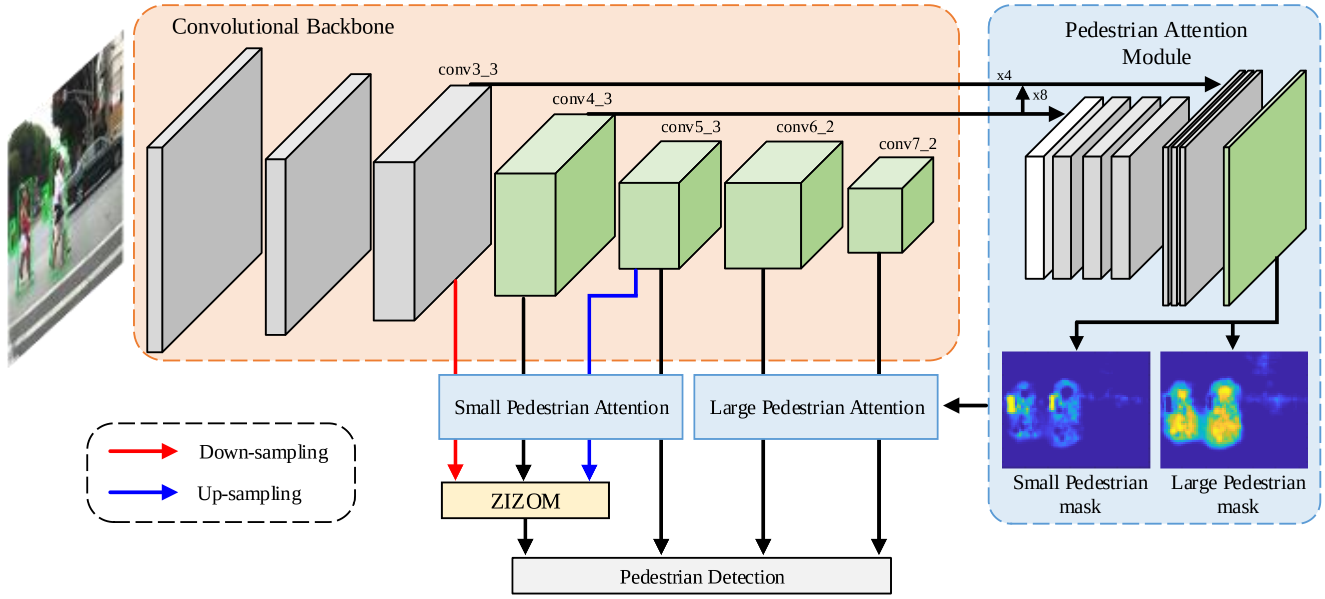


Fig. 2. Overview of the proposed GDFL framework. The model includes three key parts: convolutional backbone, pedestrian attention module and zoom-in-zoom-out module (ZIZOM). Given an image, the backbone generates multiple features representing pedestrians of different scales. The attention maps are encoded into backbone feature maps to highlight pedestrians and suppress background interference. ZIZOM incorporates local details and context information to further enhance the feature maps.

The occlusion-specific detectors were able to give a high confidence score based on the visible parts when the full-body detector was confused by the presence of background. Tian *et al.* [4] employed a large number of part detectors to cover various occlusion patterns, which however dramatically augmented the computational time. Zhou *et al.*[8] proposed a multi-label learning strategy to jointly learn part detectors which takes into the account the relations between different parts. In order to save the computation time, several works integrated the parts information into deep convolutional network [10], [16]. Instead of part-level classification, we explore pixel-level masks which guide the detector to pay more attention to human body parts.

### *B. Segmentation in Detection*

Since our pedestrian attention maps are generated in a segmentation manner [41], [42], [43], we present here some methods that have also exploited semantic segmentation information. Zhao *et al.* [44] enhanced object detection network with top-down segmentation feedbacks, where two connections are established between segmentation and detection branch for segmentation message propagation. Tian *et al.* [45] optimized pedestrian detection with semantic tasks, including pedestrian attributes and scene attributes. Instead of simple binary detection, this method considered multiple classes according to the attributes to handle pedestrian variations and discarded hard negative samples with scene attributes. Mao *et al.* [46] have demonstrated that fusing semantic segmentation features with detection features improves the performance. Du *et al.* [7] exploited segmentation as a strong cue in their F-DNN+SS framework. The segmentation mask was used in a post-processing manner to suppress prediction bounding boxes without any pedestrian. Brazil *et al.* [9] extended Faster R-CNN [31] by replacing the downstream classifier with an independent deep CNN and added a segmentation loss to implicitly supervise the detection, which made the features be more semantically meaningful. In DES [47], a segmentation branch is designed to augment the low level detection feature map with strong semantic information. Instead of exploiting segmentation mask for post-processing or implicit supervision, our attention mechanism directly encodes into feature maps and explicitly highlights pedestrians.

### *C. Attention Mechanism*

In general, attention can be viewed as a tool to reconfigure the allocation of available processing resources towards the most informative parts of an input data. Recently, Attention mechanism has gained great success across a range of tasks [48], [49], [50], [51], such image classification [52], person re-identification [53] and image captioning [54]. Shen *et al.* [55] presented a sharp attention network using differentiable Gumbel-Softmax sampler to produce sharper attention maps that can more assertively distinguish relevant visual structures from irrelevant ones. While most of existing attention models investigated spatial relation of features, in SENet [56], a ”Squeeze-and-Excitation” block was proposed to adaptively recalibrate channel-wise feature responses. Inspired by SENet, Zhang *et al.* [16] introduced body part attention via a human body keypoints detector to alleviate the occlusion problem. Besides, the aforementioned human body part detectors [4], [40], [8] can also be seen as a type of the attention mechanism, which tends to focus on different body parts. Different from the previous works, in this paper, we propose a multi-scale attention by simultaneously considering

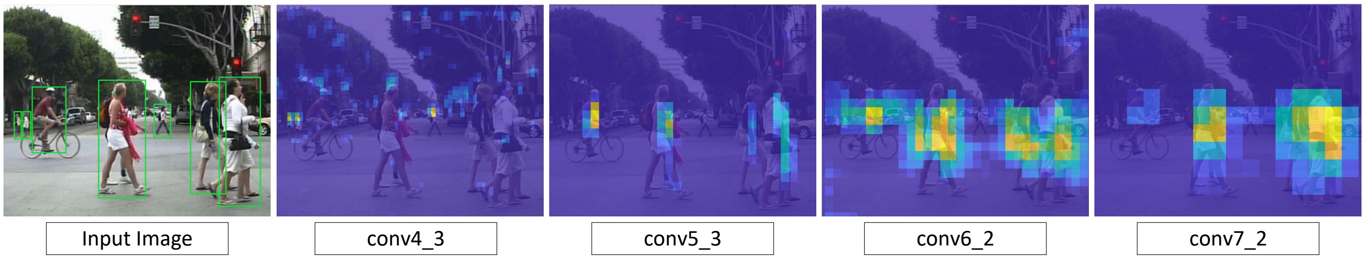


Fig. 3. Visualization of feature maps from different convolutional layers. Shallow layers have strong activation for small size targets but are unable to recognize large size instances. While deep layers tend to encode pedestrians of large size and ignore small ones. For clarity, only one channel of feature maps is shown here. Best viewed in color.

TABLE I
THE HEIGHT AND ASPECT RATIO OF DEFAULT BOX ASSOCIATED TO EACH DETECTION LAYER. TRF REFERS TO THE THEORETICAL RECEPTIVE FIELD OF CORRESPONDING LAYER.

| Detection Layer | Box Height | TRF |
|---|---|---|
| conv4_3 | 40, 60, 80 | 108 |
| conv5_3 | 110, 130, 150 | 228 |
| conv6_2 | 180, 210, 240 | 292 |
| conv7_2 | 270, 300, 330 | 356 |
| Aspect ratio: 0.41 | | |

both spatial and channel-wise relation of features to guide pedestrian detection. Our attention is learned in a weakly-supervised manner, without external model, information or hand-crafted designed occlusion patterns.

## III. Graininess-Aware Deep Feature Learning

In this section, we present the proposed GDFL method for pedestrian detection in detail. Our framework is composed of three key parts: a convolutional backbone, a scale-aware pedestrian attention module and a zoom-in-zoom-out module. The convolutional backbone generates multiple feature maps for representing pedestrian at different scales. The scale-aware pedestrian attention module generates several attention maps which are encoded into these convolutional feature maps. It forms graininess-aware feature maps which have more capability to distinguish pedestrians and body parts from background. The zoom-in-zoom-out module incorporates extra local details and context information to further enhance the features. We then slide two sibling $3 \times 3$ convolutional layers over the resulting feature maps to output a detection score and a shape offset relative to the default box at each location [35]. An overview of the proposed GDFL framework is illustrated in Fig. 2.

### A. *Multi-layer Pedestrian Representation*

Pedestrians have a large variance of scales, which is a critical problem for an accurate detection due to the difference of features between small and large instances [57], [35]. We exploit the hierarchical architecture of the deep convolutional network to address this multi-scale issue. The network computes feature maps of different spatial resolutions with successive sub-sampling layers, which forms naturally a feature pyramid. We use multiple feature maps to detect pedestrians at different scales. Specifically, we tailor the VGG16 network [58] for detection, by removing all classification layers and converting the fully connected layers into convolutional layers. Two extra convolutional layers are added on the end of the converted-VGG16 in order to cover large scale targets. The architecture of the network is presented on the top of Fig. 2. Given an input image, the network generates multiple convolutional feature layers with increasing sizes of receptive field. We select four intermediate convolutional layers {conv4_3, conv5_3, conv6_2, conv7_2} as detection layers for multi-scale detection. As illustrated in Fig. 3, shallower convolutional layers with high resolution feature maps have strong activation for small size targets, while large-size pedestrians emerge at deeper layers. We regularly place a series of default boxes [35] with different scales on top of the detection layers according to their representation capability. The detection bounding boxes are predicted based on the offsets with respect to these default boxes, as well as the pedestrian probability in each of those boxes. The high resolution feature maps from layers conv4_3 and conv5_3 are associated with default boxes of small scales for detecting small target, while those from layers conv6_2 and conv7_2 are designed for large pedestrian detection. The sizes of default boxes for each detection layer are tabulated in Table I, which are designed smaller than the theoretical receptive field (TRF). Indeed, according to [59] the effective impact area of convolutional layers is much smaller than theoretical receptive field.

### B. *Pedestrian Attention Module*

Despite the multi-layer representation, the feature maps from the backbone are still too coarse, *e.g.*, stride 8 on conv4_3, to effectively locate small size pedestrians and recognize human body parts. In addition, even if each detection layer tends to represent pedestrian of particular size, it would also consider target of other scales, which is undesirable and may lead to box-in-box detection. We propose a scale-aware pedestrian attention module to make our detector pay more attention to pedestrians, especially small size ones, and guide feature maps to focus on target of specific scale via pixel-wise attention maps. By encoding the fine-grained attention maps into the convolutional feature maps, the features representing pedestrian are enhanced, while the background interference is significantly reduced. The resulting graininess-aware features have more powerful capability to recognize human body parts

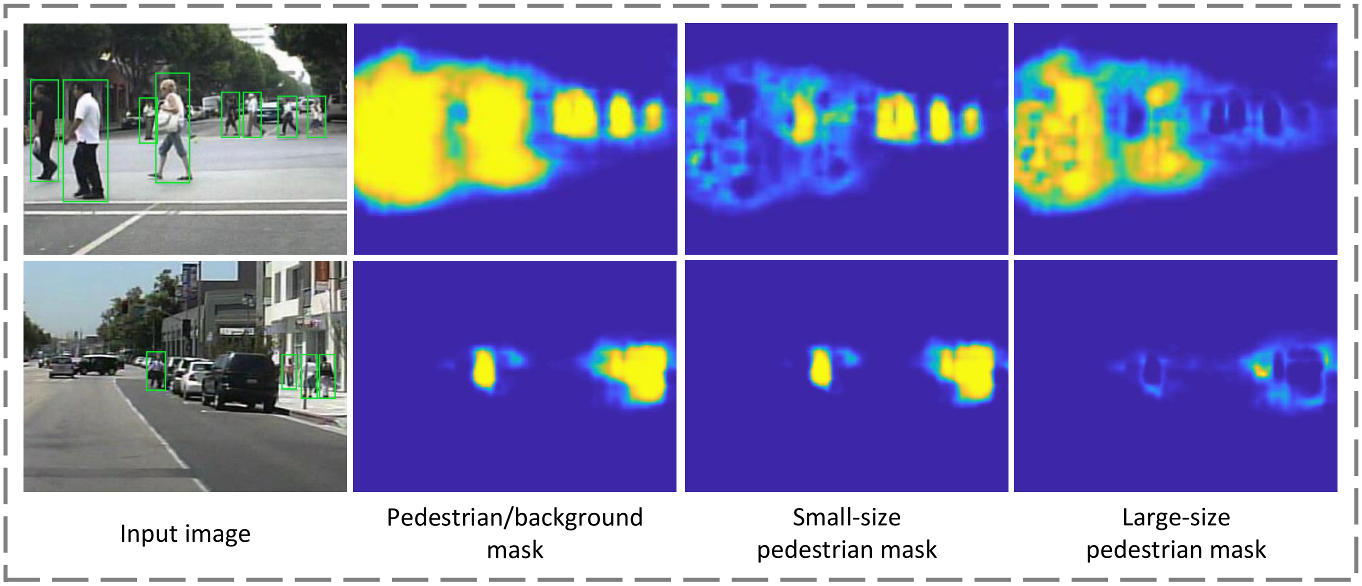

Input image
Pedestrian/background mask
Small-size pedestrian mask
Large-size pedestrian mask

Fig. 4. Visualization of pedestrian attention maps generated from Caltech test images. From left to right are illustrated: images with the ground truth bounding boxes, pedestrian v.s. background mask, small-size pedestrian mask, and large-size pedestrian mask. The pedestrian/background mask corresponds to the sum of the last two masks and can be seen as a single scale pedestrian mask. Best viewed in color.

and are able to infer occluded pedestrian based on the visible body parts.

The attention module is built based on the layers conv3_3 and conv4_3 of the backbone network. It generates multiple masks that indicate the probability of pedestrian of specific size at each pixel location. The architecture of the attention module is illustrated on the right part of Fig. 2. We construct a max-pooling layer and three atrous convolutional layers [60] on top of conv4_3 to get a conv_mask layer which has high resolution and large receptive field. Each of the conv3_3, conv4_3 and conv_mask layers is first reduced into $(S_c+1)$-channel maps and spatially up-sampled into the image size. They are then concatenated together and followed by a $1\times1$ convolutional and softmax layer to output the attention maps. Where $S_c$ denotes the number of scale-class and the remaining dimension corresponds to background. In our GDFL framework, we distinguish small and large pedestrians according to a height threshold of 120 pixels and set $S_c = 2$. Fig. 4 illustrates some examples of pedestrian masks, which effectively highlight pedestrian regions.

Once the attention maps $M \in \mathcal{R}^{W\times H\times 3}$ are generated, we encode them into the feature maps of the convolutional backbone to obtain our graininess-aware feature maps by resizing the spatial size and element-wise multiplication:

$$\tilde{F_i} = F_i \odot R(M_S, i), \quad i \in \{\text{conv4}, \text{conv5}\} \tag{1}$$

$$\tilde{F_j} = F_j \odot R(M_L, j), \quad j \in \{\text{conv6}, \text{conv7}\} \tag{2}$$

where $M_S \in \mathcal{R}^{W\times H\times 1}$ and $M_L \in \mathcal{R}^{W\times H\times 1}$ correspond to the attention maps highlighting small and large pedestrians, respectively. $W$ and $H$ are the size of input image. $R(\cdot, i)$ is the function that resizes the input into the size of $i^{th}$ layer. $\odot$ is the element-wise multiplication operator. $F_i$ represents the feature maps from backbone network while $\tilde{F_i}$ is the graininess-aware feature maps with pedestrian attention. The mask $R(M_S, i)$ is encoded into the feature maps from layers conv4_3 and conv5_3, which are responsive for small pedestrian detection. While the mask $R(M_L, i)$ is encoded into the feature maps from conv6_2 and conv7_2, which are used for large pedestrian detection. The feature maps with and without attention maps are shown in Fig. 1, where we can see that pedestrian information are highlighted while the background is smoothed with the attention.

### *C. Zoom-in-zoom-out Module*

When the human annotators try to find and recognize a small object in an image, they often zoom in and zoom out several times to correctly locate the target. The zoom-in process allows to get details information and improve the location precision. While the zoom-out process permits to import context information, which is a key factor when reasoning the probability of a target in the region, *e.g.*, pedestrians tend to appear on the ground or next to cars than on sky [61]. Inspired by these intuitive operations, we introduce a zoom-in-zoom-out module (ZIZOM) to further enhance the features. It explores rich context information and local details to facilitate pedestrian detection.

We implement the zoom-in-zoom-out module in a convolutional manner by exploiting the feature maps of different receptive fields and resolutions. Feature maps with smaller receptive fields provide rich local details, while feature maps with larger receptive fields import context information. Fig. 5(b) depicts the architecture of the zoom-in-zoom-out module. Specifically, given the graininess-aware feature maps $\tilde{F_i}$, we incorporate the features from directly adjacent layers $F_{i-1}$ and $F_{i+1}$ to mimic zoom-in and zoom-out processes. Each adjacent layer is followed by a convolutional layer of kernel size $1\times1$ to select features and an up- and down-sampling operation to harmonize the spatial size of feature

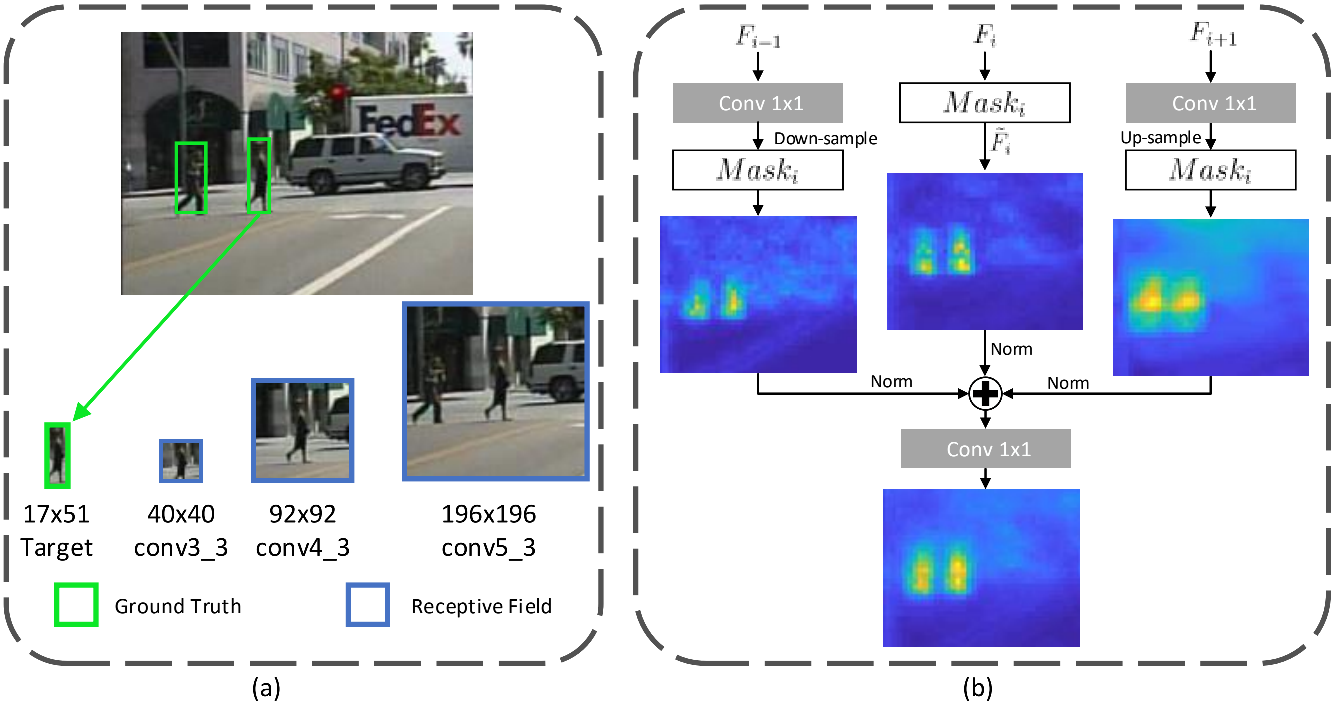


Fig. 5. Zoom-in-zoom-out module. (a) According to their receptive fields, the layer conv5_3 has more capability to get context information while the layer conv3_3 is able to get more local details. (b) Architecture of the module. Features from adjacent detection layers are re-sampled and encoded with the corresponding attention map before to be fused with current detection features.

maps. The sampling operations consist of max-pooling and bilinear interpolation without learning parameters for simplicity. The attention map of the current layer, $Mask_i$, is encoded into these sampled feature maps, making them focus on targets of the corresponding size. We then fuse these feature maps along their channel axis and generate the feature maps for final prediction with an $1 \times 1$ convolutional layer for dimension reduction as well as features recombination. Since the feature maps from different layers have different scales, we use L2-normalization [62] to rescale their norm to 10 and learn the scale during the back propagation.

Fig. 5(a) analyzes the effects of the ZIZOM in terms of receptive field with some convolutional layers. The features from conv5_3 enhance the context information with the presence of a car and another pedestrian. Since the receptive field of conv3_3 matches with size of target, its features are able to introduce more local details about the pedestrian. The concatenation of these two adjacent features with conv4_3 results in more powerful feature maps as illustrated in Fig. 5(b).

### D. Objective Function

All the three components form a unified framework which is trained end-to-end. We formulate the following multi-task loss function $L$ to supervise our model:

$$L = L_{\text{conf}} + \lambda_l L_{\text{loc}} + \lambda_m L_{\text{mask}} \tag{3}$$

where $L_{\text{conf}}$ is the confidence loss, $L_{\text{loc}}$ corresponds to the localization loss and $L_{\text{mask}}$ is the loss function for pedestrian attention maps. $\lambda_l$ and $\lambda_m$ are two parameters to balance the importance of different tasks. In our experiments we empirically set $\lambda_l$ to 2 and $\lambda_m$ to 1.

The confidence score branch is supervised by a Softmax loss over two classes (pedestrian vs. background). The box regression loss $L_{\text{loc}}$ targets at minimizing the Smooth L1 loss [34], between the predicted bounding-box regression offsets and the ground truth box regression targets. We develop a weighted Softmax loss to supervise our pedestrian attention module. There are two main motivations for this weighting policy: 1) Most regions are background, but only few pixels correspond to pedestrians. This imbalance makes the training inefficient; 2) The large size instance occupies naturally larger area compared to the small ones. This size inequality pushes the classifier to ignore small pedestrians. To address the above imbalances, we introduce an instance-sensitive weight

$$\omega_i = \alpha + \beta \frac{1}{h_i}, \tag{4}$$

and define the attention map loss $L_{\text{mask}}$ as a weighted Softmax loss:

$$L_{\text{mask}} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{l_s=0}^{S_c} \omega_i^{\mathbb{1}\{l_s \neq 0\}} \hat{c}_i^{l_s} \log(c_i^{l_s}) \tag{5}$$

where $N_s$ is the number of pixels in mask, $S_c$ is the number of scale-class, and $h_i$ is the height of the target representing by the $i^{th}$ pixel. $\mathbb{1}\{\cdot\}$ is the indicator function. $\hat{c}_i^{l_s}$ is the ground truth label of the pixel $i$, $l_s = 0$ corresponds to the background label and $c_i^{l_s}$ is the predicted score of $i^{th}$ pixel for $l_s$ class. The constants $\alpha$ and $\beta$ are set to 3 and 10 by cross validation.

## IV. Channel Adaptive Graininess-Aware Deep Feature Learning

In this section, we present the proposed CA-GDFL method for pedestrian detection in details. We first analyze and interpret the different channels of a convolutional layer to better understand the internal representations of the network. Based on the channel-wise observations, we introduce the channel adaptive pedestrian attention module to dynamically highlight the channels that are responsive to represent the pedestrians.

### A. Convolutional Layer Dissection

Our fully-convolutional network is trained for the specific pedestrian detection task, which makes the network have high activation for the human body. Generally, by taking the mean or max operation across the channels of convolutional layer, we observe that the pedestrian regions have relatively high activation comparing to the rest. However, do all channels of the feature maps have high activation for pedestrians? To answer this question, we perform the convolutional layer dissection to visualize the internal representation of the feature maps and analyze the reactions of each channel.

Qualitatively, we observe that in a detection layer, a series of feature maps have high activation for pedestrians while another sets react to different background regions but not pedestrians. Some examples are depicted in Fig. 6. Meanwhile, we perform the quantitative analysis to further interpret this observation. Given the feature maps $F_i$ of the $i^{th}$ detection layer and the pedestrian mask $M_i$ spatially reduced at the same size of the feature maps, we compute the intersection over union ($IoU_i$) metric between the mask and the $c^{th}$ channel of the feature maps $F_{i,c}$ to measure their spatial agreement:

$$IoU_{i,c} = \frac{\mathbb{1}(F_{i,c} > \tau_i) \cap M_i}{\mathbb{1}(F_{i,c} > \tau_i) \cup M_i} \tag{6}$$

where $\cap$ and $\cup$ correspond to the intersection and union operations. $\mathbb{1}(F_{i,c} > \tau_i)$ produces a binary feature map by thresholding $F_{i,c}$ with $\tau_i$. In our experiments, we set the value $\tau_i$ to retain the $5\%$ strongest activation in the feature maps. We utilize the $IoU_{i,c}$ to rank the spatial correlation between the feature maps and the mask. We observe that about $28\%$ channels show strong activation for pedestrians.

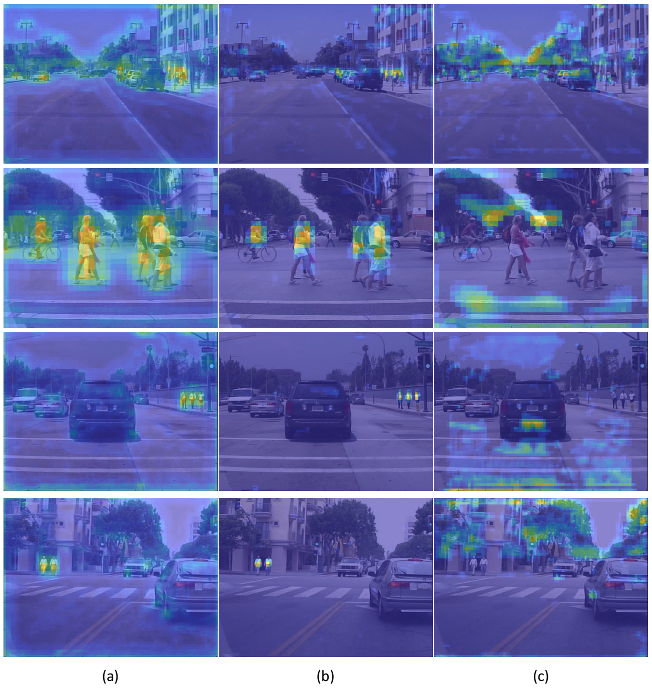


Fig. 6. Qualitative analysis of the convolutional layer dissection. (a) We compute a mean feature map across the channel dimension and observe that pedestrians are highlighted. By looking into details we observe that (b) some feature maps have strong activation for pedestrians, (c) while some other ones activate for background objects. Best viewed in color.

### B. Channel Adaptive Spatial Attention

According to the convolutional layer dissection, the convolutional layer is composed of a portion of feature maps that encode for pedestrians and the remaining ones that encode for other objects. Based on this observation, instead of simple spatial attention, we propose a channel adaptive graininess-aware feature learning framework (CA-GDFL) to dynamically highlight the channels of convolutional layer that activate for pedestrians. An overview of the CA-GDFL framework is illustrated in Fig. 7.

Specifically, we simultaneously train a spatial and channel-wise attention to enhance the features of pedestrians while diminish the importance of features representing the background objects. Instead of using the final pedestrian mask as a single-channel attention map that provides only spatial information, we generate the attention map from the intermediate result of segmentation branch to both explore spatial and channel-wise information. To achieve that, on top of each detection layer, we construct an attention branch and a segmentation branch in parallel. The attention branch is composed of 4 convolutional layers and the segmentation branch consists of 6 layers. The first three layers of these two branches share the same parameters. By sharing the parameters, the attention maps can benefit from the fine-grained segmentation supervision, which is of key importance for generating high quality attention. The last layer of the attention branch computes the channel adaptive spatial attention maps that have the same channel and spatial dimension as the feature maps of the detection layer. We then encode the attention into the feature maps via the element-wise multiplication to obtain the channel adaptive graininess-aware deep features for pedestrian detection. The bottom part of Fig. 7 depicts the architecture of the attention branch and segmentation branch.

### C. Scale Specific Pedestrian Segmentation

The goal of the segmentation branch is to provide fine-grained information for attention learning and highlight pedestrians. This branch consists of three $3\times3\times256$ and one $1\times1\times2$ convolutional layers followed by an bilinear up-sampling operation and a final $1 \times 1 \times 2$ convolutional prediction layer. In order to guide each detection layer to focus on the pedestrians of the specific scale, we reformulate the loss function for each

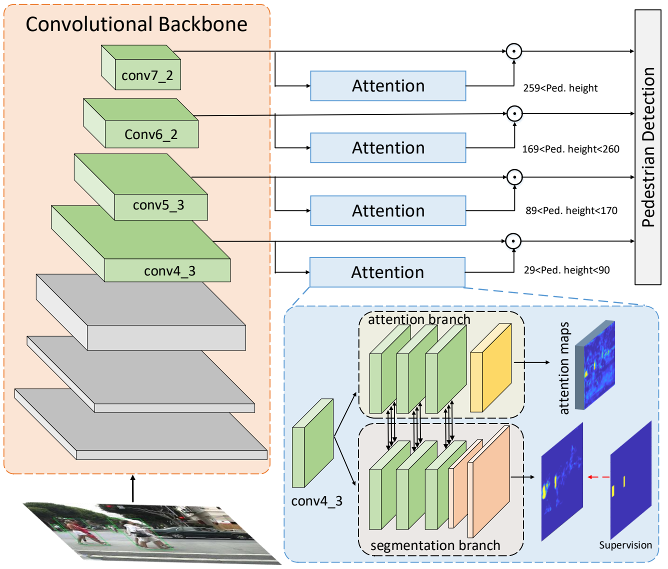


Fig. 7. Pipeline of our CA-GDFL method. Each detection layer is followed by a channel adaptive spatial attention. The attention is trained by sharing parameters with the segmentation branch to capture fine-grained information. Each detection layer is focused on the pedestrian of particular scale (Ped. height) guided by the scale specific segmentation supervision. The double arrows mean the parameters sharing.

segmentation branch as

$$L_{\text{mask}_j} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{l_s=0}^{1}\omega_{i,j}^{\mathbb{1}\{l_s\neq 0\}}\hat{c}_i^{l_s}\log(c_i^{l_s}), \tag{7}$$

where $N_s$ is the number of pixels in mask, $\hat{c}_i^{l_s}$ is the ground truth label of the pixel $i$, $l_s = 0$ corresponds to the background label and $c_i^{l_s}$ is the estimated score of the $i^{th}$ pixel for $l_s$ class. The instance-sensitive weight is redefined as follows

$$\omega_{i,j} = \begin{cases} \alpha + \beta\frac{1}{h_i} & \text{if } h_i \in [h_{\min,j}, h_{\max,j}] \\ 0 & \text{else} \end{cases}, \tag{8}$$

where $[h_{\min,j}, h_{\max,j}]$ corresponds to the range of pedestrian height that the $j^{th}$ detection layer is responsive. The value of $[h_{\min,j}, h_{\max,j}]$ for different detection layers are tabulated in Table II. With this weight, we make the segmentation branch to focus on the pedestrians of specific scales by ignoring the supervision signal from non-interested targets. Indeed, pedestrians of different scales have similar features and by naively considering those with scale out of the range as background will confuse the network and leads to sub-optimal results. Note that, as the segmentation branch is introduced for helping the attention learning, we can omit this branch during the inference to save the computation time.

In summary, we supervise the CA-GDFL framework with the following loss

$$L \quad = \quad L_{\text{conf}} + \lambda_l L_{\text{loc}} + \lambda_m \sum_{j}^{N_d} L_{\text{mask}_j}, \tag{9}$$

where $N_d$ denotes the number of detection branches, which is set to 4 in our experiments. Comparing to the previous relatively coarse small/large scale distinction, the CA-GDFL framework allows a finer distinction and select more appropriate pedestrian regions for each detection branch.

TABLE II
THE HEIGHT OF PEDESTRIAN FOR EACH DETECTION LAYER WHEN GENERATING ATTENTION MASKS IN CA-GDFL. TRF REFERS TO THE THEORETICAL RECEPTIVE FIELD OF CORRESPONDING LAYER.

| Detection Layer | TRF | $\{h_{min}, h_{max}\}$ |
|---|---|---|
| conv4_3 | 108 | {30, 89} |
| conv5_3 | 228 | {90, 169} |
| conv6_2 | 292 | {170, 259} |
| conv7_2 | 356 | {260, -} |

## V. EXPERIMENTS AND ANALYSIS

In this section, we first introduce the pedestrian detection benchmarks and evaluation metrics that we utilized for experiments. We then provide the implementation details and compare the proposed GDFL and CA-GDFL methods with the state-of-the-art approaches. Finally, we present the comprehensive ablation study which shows the contribution of different key components in our framework.

### A. Datasets and Evaluation Protocols

We comprehensively evaluated our proposed method on five benchmarks: Caltech [20], INRIA [2], KITTI [21], MOT17Det [22] and CityPersons [23]. Here we give a brief description of these benchmarks.

The Caltech dataset [20] consists of ∼10 hours of urban driving video with $350K$ labeled bounding boxes. It results in 42,782 training images and 4,024 test images. The log-average miss rate is used to evaluate the detection performance and is calculated by averaging miss rate on false positive per-image (FPPI) points sampled within the range of $[10^{-2}, 10^{0}]$, denoted as $MR_{-2}$. As the main purpose of our approaches is to address occlusion issues, we evaluated on the two subsets: *Reasonable* and *Heavy Occlusion*. The *Reasonable* subset consists of pedestrians taller than 50 pixels with partial occlusion (occlusion $<$ 35%). In the *Heavy Occlusion* subset, pedestrians are taller than 50 pixels and 36 to 80% occluded.

The INRIA dataset [2] includes 614 positive and 1,218 negative training images. There are 288 test images available for evaluating pedestrian detection methods. In this dataset, pedestrians are presented at various scenes and with different postures. The evaluation metric is the log-average miss rate on FPPI with the *Reasonable* setting.

The KITTI dataset [21] consists of 7,481 training images and 7,518 test images, comprising about 80K annotations of cars, pedestrians and cyclists. KITTI evaluates the PASCAL-style mean Average Precision (mAP) with three metrics: *easy*, *moderate* and *hard*. The difficulties are defined according to the minimum pedestrian height, occlusion and truncation level.

The MOT17Det dataset [22] consists of 14 video sequences in unconstrained environments, which results in 11,235 images. The dataset was collected from multiple different scenes with the camera installed at different position, resulting in various point of views. The dataset is split into two parts for

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE CALTECH DATASET IN TERMS OF MISS-RATE (%) AND RUNNING TIME ($s$) PER IMAGE. THE ADDITIONAL ANNOTATIONS INDICATE WHETHER EXTRA INFORMATION ARE EXPLOITED FOR TRAINING.

| | Reasonable | | Heavy Occlusion | | Additional | Running |
|---|---|---|---|---|---|---|
| Method | $MR^O_{-2}$ | $MR^N_{-2}$ | $MR^O_{-2}$ | $MR^O_{-4}$ | annotations | time (s) |
| DeepCascade+ [63] | 26.21 | 26.43 | 82.19 | – | – | 0.06 |
| MS-CNN [33] | 9.95 | 8.08 | 59.94 | 74.58 | – | 0.14 |
| SA-FastRCNN [32] | 9.68 | 7.47 | 64.35 | – | – | 0.59 |
| RPN+BF[5] | 9.58 | 7.28 | 74.36 | 83.58 | – | 0.50 |
| F-DNN+SS[7] | 8.18 | 6.89 | 53.76 | 69.73 | – | 2.40 |
| SDS-RCNN[9] | 7.36 | 6.44 | 58.55 | 74.19 | – | 0.20 |
| DeepParts [4] | 11.89 | 12.90 | 60.42 | 74.45 | Visible-part bbox | 1.00 |
| PCN [64] | 8.45 | 8.47 | 56.70 | 72.15 | – | – |
| JL-Max [8] | 10.3 | – | 48.40 | – | Visible-part bbox | 0.60 |
| FasterRCNN+ATT-vbb [16] | 10.33 | 8.11 | 45.18 | 63.52 | Visible-part bbox | – |
| PDOE+RPN [65] | 7.6 | – | 44.40 | – | Visible-part bbox | – |
| Noh *et al.* [66] | 10.85 | – | 42.42 | – | Visible-part bbox | – |
| GDFL (ours) | 7.84 | 6.32 | 43.18 | 62.02 | – | 0.05 |
| CA-GDFL (ours) | 7.84 | 6.04 | 39.35 | 59.02 | – | 0.04 |

training and testing, which are composed of 7 video sequences respectively. The Average Precision (AP) is used for evaluating different methods.

The CityPersons [23] is a recently released pedestrian detection dataset built on top of the semantic segmentation benchmark CityScapes [67]. It was recorded in multiple cities across Europe and consisted of 5,000 images. CityPersons is a challenging large-scale pedestrian detection benchmark including a large number of occlusion situations. Similar to Caltech dataset, we evaluate the log-average miss rate on FPPI on *Reasonable* and *Heavy Occlusion* subsets.

### B. Implementation Details

**Weakly supervised training for attention module:** To train the pedestrian attention module, we only use the bounding box annotations in order to be independent of any pixel-wise annotation. To achieve this, we explore a weakly supervised strategy by creating artificial foreground segmentation using bounding box information. In practice, we consider pixels within the bounding box as foreground while the rest are labeled as background. We assign the pixels that belong to multiple bounding boxes to the one that has the smallest area. As illustrated in Fig. 4, despite the weak supervised training, our generated pedestrian masks carry significant semantic segmentation information. Note that the generated attention is not exactly a rectangle mask for each pedestrian. We can see that for large scale pedestrians, the generated attention is highlighting particular part of human such as head and legs. This is because the attention module is not only supervised by the pseudo rectangular ground-truth but also by the final pedestrian detection loss. The weakly supervised training with the rectangular ground-truth guides the attention to better distinguish pedestrian regions from the background, while the detection loss makes it highlighting most relevant human parts for more effective localization.

**Training:** We optimize our detector using the stochastic gradient descent (SGD) algorithm with $0.9$ momentum and $0.0005$ weight decay unless otherwise stated. We partially initialize our model with the pre-trained model in [35], and all new additional layers are randomly initialized with the "xavier" method [68]. We adopt the data augmentation strategies as in [35] to make our model more robust to scale and illumination variations. Besides, during the training phase, negative samples largely over-dominate positive samples, and most are easy samples. For more stable training, instead of using all negative samples, we sort them by the highest loss values and keep the top ones so that the ratio between the negatives and positives is at most 3:1.

**Inference:** We use the initial size of input image to avoid loss of information and save inference time: $480 \times 640$ for Caltech and INRIA, $384 \times 1280$ for KITTI, $1080 \times 1920$ for MOT17Det and $1024 \times 2048$ for CityPersons. In inference stage, a large number of bounding boxes are generated by our detector. We perform non-maximum suppression (NMS) with a Intersection over Union (IoU) threshold of 0.45 to filter redundant detection. We use a single GeForce GTX 1080 Ti GPU for computation and our detector executes about 20 frames per second with inputs of size $480 \times 640$ pixels.

### C. Comparison with State-of-the-Art Methods

We evaluated the proposed GDFL and CA-GDFL models on five challenging pedestrian detection benchmarks, Caltech [20], INRIA [2], KITTI [21], MOT17Det [22] and CityPersons [23], and compared with existing best performing methods.

**Caltech:** We trained our model on the Caltech training set with the original annotations and evaluated on the Caltech testing set. Table III lists the comparison with state-of-the-art methods [25], [63], [4], [5], [69], [32], [9], [33], [7] on Caltech dataset in terms of miss rate and execution time. The top raws correspond to methods that perform well on the reasonable subset, while the bottom raws indicate occlusion-specific approaches. On the reasonable subset, we evaluate on both the original and new annotations denoted as $MR^O_{-2}$ and $MR^N_{-2}$, respectively. The new annotations are recently released [11] to correct the official annotations in terms of box alignment and consistency. Our approaches achieve comparable performance with the state-of-the-art method [9] using the

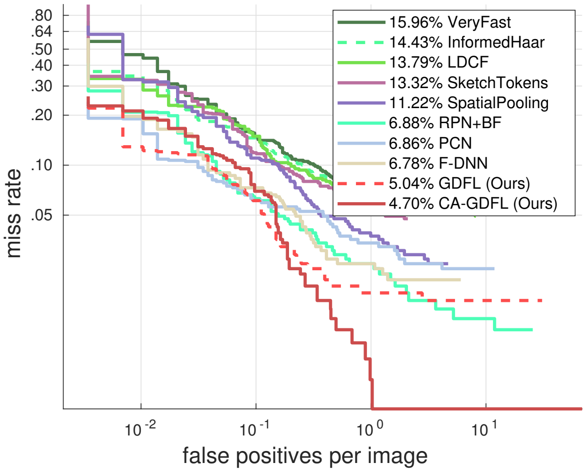


Fig. 8. Comparison with state-of-the-art methods on the INRIA dataset using the *reasonable* setting.

TABLE IV
COMPARISON WITH PUBLISHED PEDESTRIAN DETECTION METHODS ON THE KITTI DATASET. THE MAP (%) AND RUNNING TIME ($s$) ARE COLLECTED FROM THE KITTI LEADERBOARD.

| Method | Easy | Moderate | Hard | Time |
|---|---|---|---|---|
| FilteredICF [29] | 69.05 | 57.12 | 51.46 | 2 |
| DeepParts [4] | 70.49 | 58.68 | 52.73 | 1 |
| CompACT-Deep [69] | 69.70 | 58.73 | 52.73 | 1 |
| RPN+BF [5] | 77.12 | 61.15 | 55.12 | 0.6 |
| SDS-RCNN [9] | - | 63.05 | - | 0.21 |
| CFM [6] | 74.21 | 63.26 | 56.44 | 2 |
| SA-FastRCNN [32] | 77.93 | 65.01 | 60.42 | - |
| MS-CNN [33] | 83.70 | 73.62 | 68.28 | 0.4 |
| GDFL (384 × 1280) | 83.78 | 67.73 | 60.07 | 0.15 |
| GDFL (576 × 1920) | 84.61 | 68.62 | 66.86 | 0.27 |
| CA-GDFL (576 × 1920) | 85.05 | 68.87 | 66.93 | 0.25 |

original annotations and outperform this method with the new annotations. These results mean that our methods predict relatively more precise localization. In the heavy occlusion cases, we also examine a more challenging metric with the FPPI over the range $[10^{-4}, 10^{0}]$, referred as $MR^{O}_{-4}$. Our GDFL and CA-GDFL respectively achieve 43.18% and 39.35 $MR^{O}_{-2}$, and 62.02% and 59.02 $MR^{O}_{-4}$. These results are significantly better than most existing occlusion-specific detectors. Our CA-GDFL outperforms the most competitive method [66] by 3%, which suggests that our detector, guided by fine-grained information at both spatial and channel dimensions, has better capability to identify human body parts and thus to locate occluded pedestrians. Note that most occlusion-specific methods explore additional annotations, such as visible-part bounding box to supervise the models. In contrast, we only require the bounding box annotations for training our detector.

**INRIA:** We trained our model with 614 positive images by excluding the negative images and evaluated on the test set. Fig. 8 illustrates the results of our approach and the methods that perform well on the INRIA dataset [72], [13], [28], [73], [74], [75], [76]. Our GDFL and CA-GDFL detectors achieve the state-of-the-art performance with 5.04% and 4.70% miss rate respectively, outperforming the competitive methods by approximatively 2%. It proves that our methods can achieve great results even if the training set is of small scale.

TABLE V
COMPARISON WITH PUBLISHED STATE-OF-THE-ART METHODS ON MOT17DET BENCHMARK.

| Method | Average Precision | External data |
|---|---|---|
| ACF [27] | 0.32 | × |
| DPM [37] | 0.61 | × |
| FRCNN [31] | 0.72 | × |
| SDP [70] | 0.81 | × |
| KDNT [71] | 0.89 | ✓ |
| GDFL (Ours) | 0.81 | × |
| CA-GDFL (Ours) | 0.82 | × |

TABLE VI
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CITYPERSONS VALIDATION DATASET. MISS RATE ON THE ORIGINAL IMAGE RESOLUTION (1024 × 2048) ARE REPORTED.

| Method | Reasonable | Heavy occ. |
|---|---|---|
| FRCNN+ATT-part [16] | 15.9 | 56.6 |
| FRCNN [23] | 15.4 | 55.0 |
| TLL [77] | 15.5 | 53.6 |
| TLL+MRF [77] | 14.4 | 52.0 |
| RepLoss [78] | 13.2 | 56.9 |
| OR-CNN [79] | 12.8 | 55.7 |
| ALFNet [80] | 12.0 | 51.9 |
| Cascade MS-CNN [81] | 12.0 | 49.4 |
| GDFL (Ours) | 14.84 | 44.17 |
| CA-GDFL (Ours) | 13.6 | 43.2 |

**KITTI:** We trained our model on the KITTI training set and evaluated on the designated test set. We compared our GDFL and CA-GDFL approaches with the current pedestrian detection methods on KITTI [29], [4], [69], [5], [9], [6], [33]. The results are listed in Table IV. Our detector achieves competitive performance with MS-CNN [33] yet executes about 3× faster with the original input size. Apart its scale-specific property, MS-CNN [33] has explored input and feature upsampling strategies which are crucial for improving the small objects detection performance. Following this process, we upsampled the inputs by 1.5 times and we observed a significant improvement on the hard subset but with more execution time. Note that in the KITTI evaluation protocol, cyclists are regarded as false detections while people-sitting are ignored. With this setting, our pedestrian attention mechanism is less helpful since it tends to highlight all human-shape targets including person riding a bicycle. This explains the reason our model does not perform as well as on KITTI than that on Caltech or INRIA.

**MOT17Det:** We trained and evaluated our detector on the designated training and testing sets respectively, where the model trained on Caltech was used as initialization. Table V tabulates the detection results of our method and the state-of-the-art approaches. Our GDFL and CA-GDFL detectors achieve competitive 0.81 and 0.82 average precision (AP) without using external datasets for training. This performance demonstrates the generalization capability of our models. The state-of-the-art method KDNT [71] achieves impressive 0.89 AP, but a large number of additional data such as ETH [82], Caltech pedestrian [20] and the self-collected surveillance dataset are included for training.

TABLE VII
ABLATION EXPERIMENTS EVALUATED ON THE CALTECH TEST SET. ANALYSIS SHOW THE EFFECTS OF VARIOUS COMPONENTS AND DESIGN CHOICES ON THE DETECTION PERFORMANCE IN TERMS OF $MR^{O}_{-2}$.

| Component | Choice | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single-layer detection | ✓ | | | | | | | | | | |
| Multi-layer detection | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Instance-sensitive weight | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| Single scale attention | | | ✓ | | | | | | | | |
| Two-scale attention | | | | ✓ | | | ✓ | ✓ | | ✓ | |
| Multi-scale spatial-channel att. | | | | | ✓ | | | | ✓ | | |
| Self multi-scale attention | | | | | | ✓ | | | | | |
| ZIZOM on $\tilde{F}_{conv4\_3}$ | | | | | | | | ✓ | ✓ | ✓ | |
| ZIZOM on $\tilde{F}_{conv5\_3}$ | | | | | | | | | | ✓ | |
| ZIZOM on $F_{conv4\_3}$ | | | | | | | | | | | ✓ |
| Miss rate on Reasonable | 16.86 | 9.44 | 9.16 | 8.44 | 8.33 | 9.35 | 9.59 | 7.84 | **7.84** | 8.01 | 8.86 |
| Miss rate on Heavy Occ. | 53.44 | 50.21 | 47.60 | 44.68 | 41.34 | 46.24 | 47.69 | 43.18 | **39.35** | 42.86 | 45.73 |

**CityPersons:** We trained our model on the training set using only the bounding box annotations and evaluated on the validation set. As CityPersons dataset has larger resolution, we added two extra convolutional layers to cover largest pedestrians. Following [23], we optimize our model with the Adam solver.Table VI shows the comparison with state-of-the-art methods on CityPersons. It can be observed that our CA-GDFL achieves competitive performance under the reasonable situations and outperforms the existing approaches by a large margin on heavy occlusion cases. We obtain $43.2\%$ miss rate on the heavy occlusion subset, which corresponds to a gain of $9\%$ upon the closest pedestrian detection competitor (ALFNet [80]). The multiple refinement stages in AFLNet help to improve the results on the reasonable subset with more precise localization, but are less helpful for recognizing heavily occluded pedestrians. Recently, Cao *et al.* applied the cascade training strategy of Cascade RCNN [83] on the MS-CNN framework [33], namely Cascade MS-CNN [81]. Although ALFNet and Cascade MS-CNN utilize similar multiple refinement steps, the two-stage Cascade MS-CNN detector outperforms single-stage detector AFLNet. However, both methods have much more failures than the proposed CA-GDFL on the hard cases such as occlusion situations. These results demonstrate the robustness of our method in the challenging scenarios such as heavy occlusion cases.

**Efficiency Analysis:** Since our goal is to propose a fast and accurate pedestrian detector, we also examined the efficiency of our method. Tables III and IV compare the running time on Caltech dataset and KITTI benchmark, respectively. Our methods are much faster than F-DNN+SS [7] and is about $10\times$ faster than the previous best method on Caltech heavy occlusion subset, JL-Max [8]. On the Caltech reasonable subset, SDS-RCNN [9] performs similar results with our methods, but it requires $4\times$ more inference times than our approaches. On the KITTI dataset, when we utilize the original scale of input data, our model can execute at $0.15$ second per image which is faster than all tabulated approaches. The efficiency is sacrified to enhance the detection performance by upsampling the input image, but it is still faster than most of pedestrian detectors. In summary, the comparison shows that the proposed approaches achieve a favorable trade-off between speed and accuracy.

From the tables, we can see that CA-GDFL is slightly faster than GDLF. The main reason is that in GDFL, the attention maps are produced at the resolution of input image and then down-sampled at resolutions of different features. The up-sampling process to obtain high-resolution attention is relatively time consuming. While in CA-GDFL, the attention branch is supervised by the pixel-level segmentation masks only during the training. At inference, the attention map is directly computed at features resolution without any up-sampling computations, which makes it slightly faster.

### D. Ablation Study

To better understand our model, we conducted ablation experiments using the Caltech dataset and CityPersons dataset. We considered our convolutional backbone as baseline and successively added different key components to examine their contributions on performance. Table VII summarizes our comprehensive ablation experiments.

**Multi-layer detection:** We first analyzed the advantages of using multiple detection layers. To this end, we trained a model with multiple detection branches and another one which only used the layer conv6_2 for predicting pedestrians of all scales. The experimental results of these two architectures demonstrate the superiority of multi-layer detection with a notable gain of $7\%$ and $3\%$ on the Caltech reasonable and heavy occlusion subsets, respectively. The mismatch between the receptive field size and the scale of targets has a significant impact on the detection performance, which leads to a deterioration of results. For the following experiments, we employed the multi-layer detection framework.

**Attention mechanism:** We analyzed the effects of our attention mechanism, in particular the contribution of the single-scale attention and the multi-scale attention. To control this, we compared three models with single-scale, two-scale and multi-scale attention designs. From Table VII, we can see that these three models improve the results with $0.28\%$, $1\%$ and $1.11\%$ gain on the Caltech reasonable subset and $2.61\%$, $5.53\%$ and $8.87\%$ gain on the Caltech heavy occlusion subset, respectively. The models with the two-scale and multi-scale attention perform clearly better than single-scale atten-

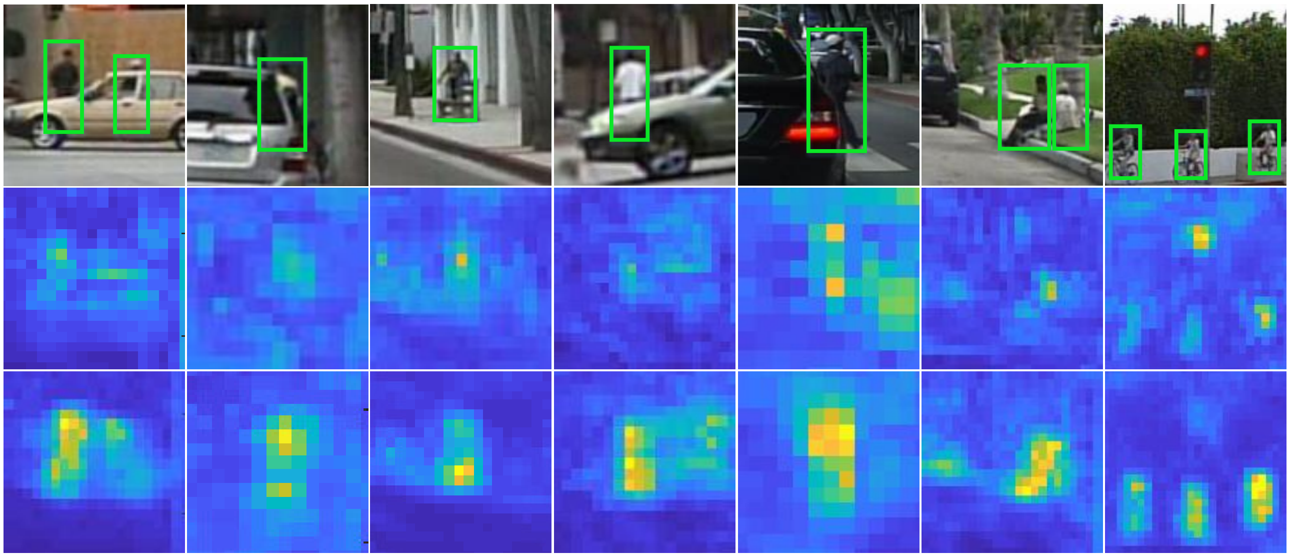


Fig. 9. Hard detection samples where box-based detector is often fooled due to noisy representation. The first row illustrates the images with pedestrians located by green bounding boxes. The second and third rows show the feature maps without attention maps and the graininess-aware feature maps, respectively. Best viewed in color.

TABLE VIII
ANALYSIS OF CHANNEL-WISE ATTENTION EFFECT IN TERMS OF MISS RATE (%) ON CITYPERSONS VALIDATION DATASET.

| Method | Reasonable | Heavy occlusion |
|---|---|---|
| GDFL | 14.84 | 44.17 |
| GDFL w channel att. | 14.56 | 43.43 |
| CA-GDFL w/o channel att. | 14.57 | 45.60 |
| CA-GDFL | 13.64 | 43.16 |

tion framework. The confusions such as box-in-box detection which can hardly be eliminated by NMS process and results in false alarms, are suppressed with our multi-scale attention maps. We observe an impressive improvement on the Caltech heavy occlusion subset, which demonstrates that the fine-grained masks better capture information of body parts. Some examples of occlusion cases are depicted in Fig. 9. We can see that the features without attention are unable to recognize human parts and tend to ignore occluded pedestrians. When we encode the pedestrian masks into these feature maps, human body parts are considerably highlighted. The detector becomes able to deduce the occluded parts by considering visible parts, which makes plausible the detection of occluded targets.

**Self multi-scale channel-spatial attention:** Our attention mechanism is trained with the pixel-level segmentation supervision. We aim to analyze the contribution of this supervisory signal. To this end, we developed a self multi-scale attention model which was optimize using only the usual detection loss functions. This model achieves 9.35% and 46.24% miss rate on Caltech reasonable and heavy occlusion subsets, respectively. We observe a drop of 1.02% and 4.9% on the two subsets compared to the multi-scale attention learned with the segmentation supervision. These results show that self-attention can not provide satisfactory guidance, while the fine-grained information provided by the segmentation supervisory signal are crucial for high quality attention generation.

**Channel attention:** We also investigated the effect of the channel-wise attention on the performance. On one hand, as GDFL model has only spatial attention, we added a channel-wise attention layer [56] after each spatial attention layer and call this model as "GDFL w channel att.". On the other hand, in CA-GDFL instead of producing the channel adaptive spatial attention maps that have the same channel and spatial dimension as the feature maps of the detection layer, we computed single channel spatial attention map by ignoring channel-wise guidance. Namely, we call this model as "CA-GDFL w/o channel att.". The experimental results of these models on CityPersons dataset are reported on Table VIII. We can see that without the channel-wise guidance, the performance drops by approximately 1% and more than 2% on reasonable and heavy occlusion subsets respectively for CA-GDFL framework. With the channel-wise attention, "GDFL w channel att." improves the performance on the heavy occluded subset by nearly 1%. Remark that without channel-wise attention, as GDFL has more precise spatial attention maps, it performs slightly better than "CA-GDFL w/o channel att." on heavy occlusion subset.

**Instance-sensitive weight in Softmax loss:** During the training stage, our attention module was supervised by a weighted Softmax loss and we examined how the instance-sensitive weight contributed to the performance. We compared two models trained with and without the weight term. As listed in the 7th column of Table VII, the performance drops on both two subsets of Caltech with the conventional Softmax loss. In particular, the miss rate increases from 44.68% to 47.69% in heavy occlusion case. The results point out that the instance-sensitive weight term is a key component for accurate attention generation. Instead of the reformulated instance-sensitive weight, the CA-GDFL trained with the standard instance-sensitive weight performs 8.91% and 41.44% miss rate. By considering the pedestrians of non-interested scale as background leads to a slight drop of performance.

**ZIZOM:** We further built the zoom-in-zoom-out module on our model with attention maps. Table VII shows that with the ZIZOM on top of the graininess-aware features $\tilde{F}_{conv4\_3}$, the performance is ameliorated by more than 1% and 2% on the Caltech heavy occlusion subsets for the GDFL and CA-GDLF models, respectively. We observe also a slight improvement on the reasonable subset. However, when we further constructed a ZIZOM on $\tilde{F}_{conv5\_3}$, the results were nearly the same. Since the feature maps $\tilde{F}_{conv5\_3}$ represent pedestrians with about 100 pixels tall, these results confirm our intuition that context information and local details are important for small targets but are less helpful for large ones. To better control the effectiveness of this module, we disabled the attention mechanism and considered a convolutional backbone with the ZIZOM on the original feature map $F_{conv4\_3}$ model. The comparison with the baseline shows a gain of 4% on the Caltech heavy occlusion subset. The results prove the effectiveness of the proposed zoom-in-zoom-out module.

## VI. Conclusion

In this paper, we have proposed a graininess-aware deep feature learning method for pedestrian detection. We trained a multi-scale attention mechanism with pixel-level supervision signal in a weakly-supervised manner. The attention maps with fine-grained information have higher capability to distinguish small-scale and occluded targets. By encoding the attention into the convolutional feature maps, we obtain more discriminative graininess-aware features which are more robust to background interference and focus on pedestrians. We further introduced a zoom-in-zoom-out module to enhance the feature maps of shallow layers by incorporating context and local information. Experimental results on five widely-used pedestrian benchmarks have validated the advantages of the proposed method on detection robustness and efficiency.

## 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 61672306, Grant U1713214, Grant 61572271, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564 and in part by Zhejiang Leading Innovation Research Program 2018R01017.

## References

[1] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[3] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.

[4] 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.

[5] 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.

[6] 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, 2017.

[7] 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 Winter Conf. Applications Comput. Vis.*, 2017, pp. 953–961.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] W. Liu, C. Zhang, H. Ma, and S. Li, "Learning efficient spatial-temporal gait features with deep learning for human identification," *Neuroinformatics*, vol. 16, no. 3-4, pp. 457–471, 2018.

[15] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-c3d: Temporal convolutional 3d network for real-time action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[16] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6995–7003.

[17] C. Lin, J. Lu, and J. Zhou, "Multi-grained deep feature learning for robust pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, 2018.

[18] ——, "Multi-grained deep feature learning for pedestrian detection," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2018, pp. 1–6.

[19] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 899–906.

[20] 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.

[21] 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.

[22] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[23] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3213–3221.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, p. 4.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[38] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 990–997.

[39] 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.

[40] C. Zhou and J. Yuan, "Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection," in *ACCV*, 2016, pp. 305–320.

[41] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 447–456.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

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

[44] X. Zhao, S. Liang, and Y. Wei, "Pseudo mask augmented object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4061–4070.

[45] 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.

[46] 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.

[47] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5813–5821.

[48] L. Jin, X. Shu, K. Li, Z. Li, G.-J. Qi, and J. Tang, "Deep ordinal hashing with spatial attention," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2173–2186, 2018.

[49] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph lstm for group activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[50] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *arXiv preprint arXiv:1909.13245*, 2019.

[51] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9637–9646.

[52] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.

[53] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, 2017.

[54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[55] C. Shen, G.-J. Qi, R. Jiang, Z. Jin, H. Yong, Y. Chen, and X.-S. Hua, "Sharp attention network via adaptive sampling for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, 2018.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[57] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3578–3587.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[59] 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.

[60] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *International Conference on Learning Representations*, 2015.

[61] Z. Fu, Z. Jin, G.-J. Qi, C. Shen, R. Jiang, Y. Chen, and X.-S. Hua, "Previewer for multi-scale object detector," in *ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 265–273.

[62] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," in *International Conference on Learning Representations*, 2016, p. 3.

[63] 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.

[64] 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. 2, no. 11, pp. 3148–3159, 2018.

[65] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–151.

[66] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 966–974.

[67] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[68] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010, pp. 249–256.

[69] 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.

[70] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2129–2137.

[71] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 36–42.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 536–551.

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

[79] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 657–674.

[80] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 618–634.

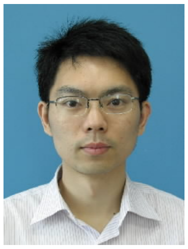
[81] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *arXiv preprint arXiv:1906.09756*, 2019.

[82] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

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

**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** Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xian University of Technology, Xian, 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 230 scientific papers in these areas, where over 70 of them are IEEE TRANSACTIONS Papers (including 14 T-PAMI papers) and 53 of them are CVPR/ICCV/ECCV papers. 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 the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He was a recipient of the National Science Fund of China for Excellent Young Scholars in 2018. He serves as the Co-Editor-of-Chief for the Pattern Recognition Letters, an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY and Pattern Recognition. He is the Program Co-Chair of IEEE ICME'2020 and AVSS'2020.



**Gang Wang** is currently a researcher of Alibaba, and a chief scientist of Alibaba AI Labs. He was an Associate Professor with the School of Electrical and Electronic Engineering at Nanyang Technological University (NTU). He had a joint appointment at the Advanced Digital Science Center (Singapore) as a research scientist from 2010 to 2014. He received his B.Eng. degree from Harbin Institute of Technology in Electrical Engineering and the PhD degree in Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. He is a recipient of MIT technology review innovator under 35 award (Asia). He is an associate editor of TPAMI and an area chair of ICCV 2017.

**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.