Graininess-Aware Deep Feature Learning for Pedestrian Detection

Chunze Lin[1], Jiwen Lu[1(✉)], Gang Wang[2], and Jie Zhou[1]

[1] Tsinghua University, Beijing, China
lcz16@mails.tsinghua.edu.cn, {lujiwen,jzhou}@tsinghua.edu.cn
[2] Alibaba AI Labs, Hangzhou, China
wg134231@alibaba-inc.com

Abstract. In this paper, we propose a graininess-aware deep feature learning method for pedestrian detection. Unlike most existing pedestrian detection methods which only consider low resolution feature maps, we incorporate fine-grained information into convolutional features to make them more discriminative for human body parts. Specifically, we propose a pedestrian attention mechanism which efficiently identifies pedestrian regions. Our method encodes fine-grained attention masks into convolutional feature maps, which significantly suppresses background interference and highlights pedestrians. Hence, our graininess-aware features become more focused on pedestrians, in particular those of small size and with occlusion. We further introduce a zoom-in-zoom-out module, which enhances the features by incorporating local details and context information. We integrate these two modules into a deep neural network, forming an end-to-end trainable pedestrian detector. Comprehensive experimental results on four challenging pedestrian benchmarks demonstrate the effectiveness of the proposed approach.

Keywords: Pedestrian detection · Attention · Deep learning
Graininess

1 Introduction

Pedestrian detection is an important research topic in computer vision and has attracted a considerable attention over past few years [4,7,9,11,18,32,37,39,43,45,48]. 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 variations, low resolution and occlusion issues.

Existing methods for pedestrian detection can mainly be grouped into two categories: hand-crafted features based [7,9,40,44] and deep learning features based [4,11,18,48]. In the first category, human shape based features such as Haar [39] and HOG [7] are extracted to train SVM [7] or boosting classifiers [9]. While these methods are sufficient for simple applications, these hand-crafted

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11213, pp. 745–761, 2018.
https://doi.org/10.1007/978-3-030-01240-3_45

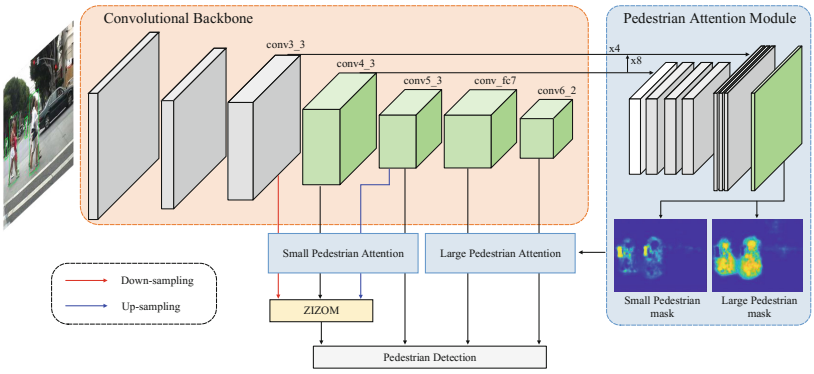


Fig. 1. Overview of our proposed framework. The model includes three key parts: convolutional backbone, pedestrian attention module and zoom-in-zoom-out module (ZIZOM). Given an image, the backbone generates multiple features representing pedestrians of different scales. The attention masks are encoded into backbone feature maps to highlight pedestrians and suppress background interference. ZIZOM incorporates local details and context information to further enhance the feature maps.

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, which shows more discriminative capability to recognize pedestrian with complex poses from noisy background. Deep learning features have considerably improved pedestrian detection performance. While many CNN based methods have been proposed [4,11,18,26,48], there are still some shortcomings for methods in this category. On one hand, 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. On the other hand, feature maps of coarse resolution and fixed receptive field are often used for prediction, which is inefficient for distinguishing targets of small size from background.

In this paper, we propose a graininess-aware deep feature learning (GDFL) based detector for pedestrian detection. We exploit fine-grained details into deep convolutional features for robust pedestrian detection. Specifically, we propose a scale-aware pedestrian attention module to guide the detector to focus on pedestrian regions. It generates pedestrian attentional masks which indicate the probability of human at each pixel location. With its fine-grained property, the attention module has high capability to recognize small size target and human body parts. By encoding these masks into the convolutional feature maps, they significantly eliminate background interference while highlight pedestrians. The resulting graininess-aware deep features have much more discriminative capability to distinguish pedestrians, especially the small-size and occluded ones from complex background. In addition, we introduce a zoom-in-zoom-out module to further alleviate the detection of targets at small size. It mimics our intuitive zoom in and zoom out processes, when we aim to locate an object in an image.

The module incorporates local details and context information in a convolutional manner to enhance the graininess-aware deep features for small size target detection. Figure 1 illustrates the overview of our proposed framework. The proposed two modules can be easily integrated into a basic deep network, leading to an end-to-end trainable model. This results in a fast and robust single stage pedestrian detector, without any extra refinement steps. Extensive experimental results on four widely used pedestrian detection benchmarks demonstrate the effectiveness of the proposed method. Our GDFL approach achieves competitive performance on Caltech [10], INRIA [7], KITTI [14] and MOT17Det [29] datasets and executes about 4 times faster than competitive methods.

2 Related Work

Pedestrian Detection: 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 [35] which has shown great accuracy in general object detection. RPN+BF [43] 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 [19] and MS-CNN [5] extended Fast and Faster R-CNN [15,35] with a multi-scale network to deal with the scale variations problem, respectively. Instead of a single downstream classifier, F-DNN [11] 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 [23,34]. Apart the above full-body detectors, several human part based methods [12,31,32,37,47,48] 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. 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. Instead of part-level classification, we explore pixel-level masks which guide the detector to pay more attention to human body parts.

Segmentation in Detection: Since our pedestrian attention masks are generated in a segmentation manner [17,25], we present here some methods that have also exploited semantic segmentation information. Tian *et al.* [38] 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.* [27] have demonstrated that fusing semantic segmentation features with detection features improves the performance. Du *et al.* [11] exploited segmentation as a strong cue in their F-DNN+SS framework. The segmentation mask was used in

Fig. 2. 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.

a post-processing manner to suppress prediction bounding boxes without any pedestrian. Brazil *et al.* [4] extended Faster R-CNN [35] 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. Instead of exploiting segmentation mask for post-processing or implicit supervision, our attention mechanism directly encodes into feature maps and explicitly highlights pedestrians.

3 Approach

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 masks which are encoded into these convolutional feature maps. This 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 3×3 convolutional layers over the resulting feature maps to output a detection score and a shape offset relative to the default box at each location [23].

3.1 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. 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 [22]. We use multiple feature maps to detect pedestrians at different scales. Specifically, we tailor the VGG16 network [36] for detection, by removing all classification layers and converting the fully connected layers into

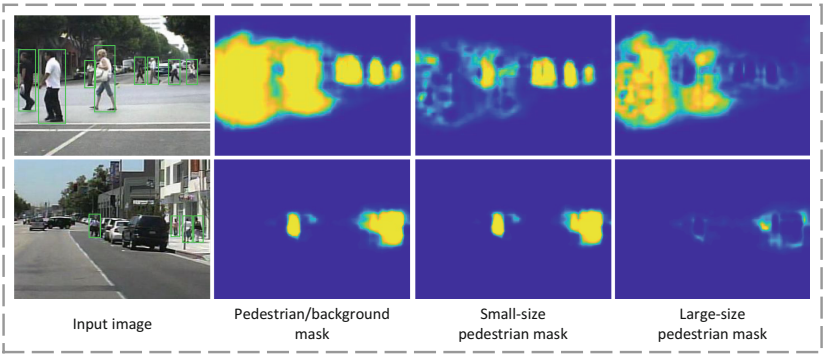


Fig. 3. Visualization of pedestrian attention masks generated from Caltech test images. From left to right are illustrated: images with the ground truth bounding boxes, pedestrian v.s. background mask, small-size pedestrian mask, and large-size pedestrian mask. The pedestrian/background mask corresponds to the sum of the last two masks and can be seen as a single scale pedestrian mask. Best viewed in color.

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. 1. 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, conv_fc7, conv6_2} as detection layers for multi-scale detection. As illustrated in Fig. 2, 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 [23] 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 conv_fc7 and conv6_2 are designed for large pedestrian detection.

3.2 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 attentional maps. By encoding the fine-grained attention masks into the convolutional feature maps, the features

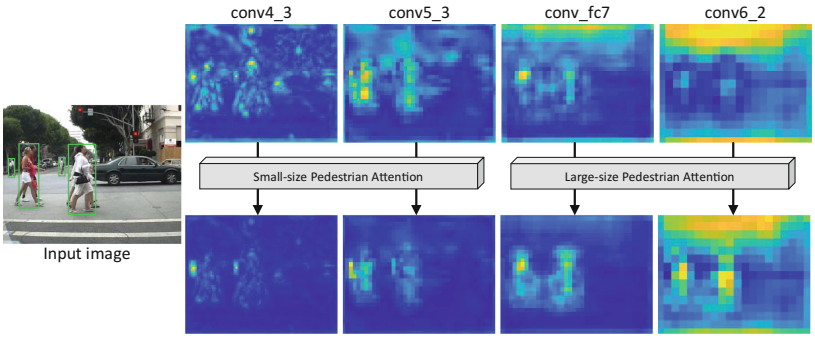


Fig. 4. Visualization of feature maps from detection layers of the backbone network (top), and visualization of feature maps with pedestrian attention (bottom). With our attention mechanism, the background interference is significantly attenuated and each detection layer is more focused on pedestrians of specific size. Best viewed in color.

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

The attention module is built 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 in Fig. 1. We construct a max-pooling layer and three atrous convolutional layers [20] on top of conv4_3 to get conv_*mask* layer which has high resolution and large receptive field. Each of 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 and followed by a 1×1 convolution and softmax layer to output attention maps. Where S_c corresponds to the number of scale-class. In default, we distinguish small and large pedestrians according to a height threshold of 120 pixels and set $S_c = 2$. Figure 3 illustrates some examples of pedestrian masks, which effectively highlight pedestrian regions.

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

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

$$\tilde{F}_j = F_j \odot R(M_L, j), \quad j \in \{\text{conv\_fc7}, \text{conv6}\} \tag{2}$$

where $M_S \in \mathcal{R}^{W \times H \times 1}$ and $M_L \in \mathcal{R}^{W \times H \times 1}$ correspond to attention masks 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 channel element-wise dot product 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

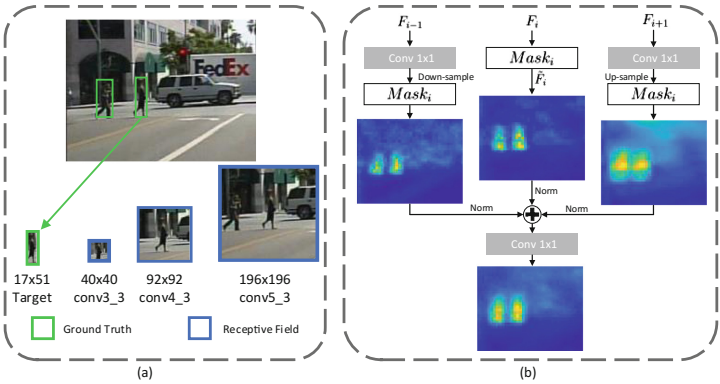


Fig. 5. Zoom-in-zoom-out module. (a) According to their receptive fields, the layer conv5_3 has more capability to get context information while the layer conv3_3 is able to get more local details. (b) Architecture of the module. Features from adjacent detection layers are re-sampled and encoded with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

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 conv_fc7 and conv6_2, which are used for large pedestrian detection. The feature maps with and without attention masks are shown in Fig. 4, where pedestrian information is highlighted while background is smoothed with masks.

3.3 Zoom-In-Zoom-Out Module

When our human annotators try to find and recognize a small object in an image, we 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. 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 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. Figure 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 an 1×1 kernel convolution to select features and an up- and down-sampling operation to harmonize the spatial size of feature maps. The sampling

operations consist of max-pooling and bi-linear interpolation without learning parameters for simplicity. The attention mask 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 1×1 convolutional layer for dimension reduction as well as features recombination. Since the feature maps from different layers have different scales, we use L2-normalization [24] to rescale their norm to 10 and learn the scale during the back propagation.

Figure 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 import 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).

3.4 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_{conf} is the confidence loss, L_{loc} corresponds to the localization loss and L_{mask} is the loss function of pedestrian attention masks. λ_l and λ_m are two parameters to balance the importance of different tasks. In our experiments we empirically set λ_l to 2 and λ_m to 1.

The confidence score branch is supervised by a Softmax loss over two classes (pedestrian vs. background). The box regression loss L_{loc} targets at minimizing the Smooth L1 loss [15], 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 a instance-sensitive weight $\omega_i = \alpha + \beta \frac{1}{h_i}$ and define the attention mask loss L_{mask} as a weighted Softmax loss:

$$L_{\text{mask}} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{l_s=0}^{S_c} \mathbb{1}\{y_i = l_s\} \omega_i^{\mathbb{1}\{l_s \neq 0\}} \log(c_i^{l_s}) \tag{4}$$

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. y_i is the ground truth label, $l_s = 0$ corresponds to the background label and $c_i^{l_s}$ is the predicted score of i^{th} pixel for l_s class. The constants α and β are set to 3 and 10 by cross validation.

4 Experiments and Analysis

4.1 Datasets and Evaluation Protocols

We comprehensively evaluated our proposed method on 3 benchmarks: Caltech [10], INRIA [7] and KITTI [14]. Here we give a brief description of these benchmarks.

The Caltech dataset [10] consists of ∼10 h 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 $[10^{-2}, 10^{0}]$. As the purpose of our approach is to alleviate occlusion and small-size issues, we evaluated our GDFL on three subsets: *Heavy Occlusion*, *Medium* and *Reasonable*. In the *Heavy Occlusion* subset, pedestrians are taller than 50 pixels and 36 to 80% occluded. In the *Medium* subset, people are between 30 and 80 pixels tall, with partial occlusion. The *Reasonable* subset consists of pedestrians taller than 50 pixels with partial occlusion.

The INRIA dataset [7] includes 614 positive and 1,218 negative training images. There are 288 test images available for evaluating pedestrian detection methods. The evaluation metric is the log-average miss rate on FPPI. Due to limited available annotations, we only considered the *Reasonable* subset for comparison with state-of-the-art methods.

The KITTI dataset [14] 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 based on minimum pedestrian height, occlusion and truncation level.

The MOT17Det dataset [29] consists of 14 video sequences in unconstrained environments, which results in 11,235 images. The dataset is split into two parts for training and testing, which are composed of 7 video sequences respectively. The Average Precision (AP) is used for evaluating different methods.

4.2 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. 3, despite the weak supervised training, our generated pedestrian masks carry significant semantic segmentation information.

Training: Our network is trained end-to-end using stochastic gradient descent algorithm (SGD). We partially initialize our model with the pre-trained model

Table 1. Comparison with the state-of-the-art methods on the Caltech heavy occlusion subset in terms of speed and miss rate.

Method	Miss rate (%)	Computing time (s)
FPDW [8]	95.56	0.2
DeepCascade+ [1]	82.19	0.06
RPN+BF [43]	74.36	0.36
SA-FastRCNN [19]	64.35	0.59
DeepParts [37]	60.42	1
MS-CNN [5]	59.94	0.10
SDS-RCNN [4]	58.55	0.26
F-DNN+SS [11]	53.76	2.48
JL-TopS [48]	49.20	0.6
Our GDFL	43.18	0.05

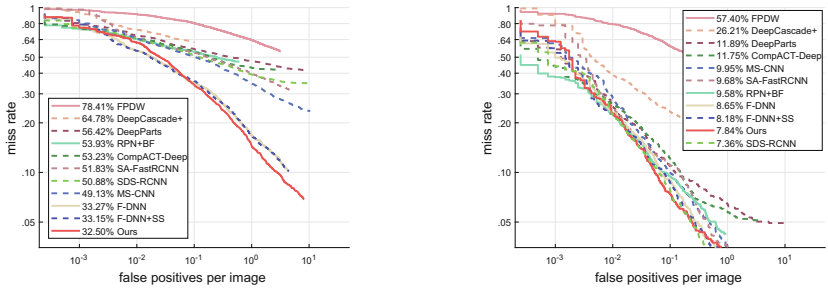
in [23], and all new additional layers are randomly initialized with the "xavier" method [16]. We adopt the data augmentation strategies as in [23] to make our model more robust to scale and illumination variations. Besides, during the training phase, negative samples largely over-dominate positive samples, 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: 480×640 for Caltech and INRIA, and 384×1280 for KITTI. 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 480×640 pixels.

4.3 Results and Analysis

Comparison with State-of-the-Art Methods: We evaluated our proposed GDFL method on four challenging pedestrian detection benchmarks, Caltech [10], INRIA [7], KITTI [14] and MOT17Det [29].

Caltech: We trained our model on the Caltech training set and evaluated on the Caltech testing set. Table 1 lists the comparison with state-of-the-art methods on Caltech heavy occlusion subset in terms of execution time and miss rate. Figure 6 illustrates the ROC plot of miss rate against FPPI for the available top performing methods reported on Caltech medium and reasonable subsets [1,4–6,8,11,19,37,43]. In heavy occlusion case, our GDFL achieves 43.18% miss rate, which is significantly better than the existing occlusion-specific detectors.



(a) Medium (b) Reasonable

Fig. 6. Comparison with state-of-the-art methods on the Caltech dataset.

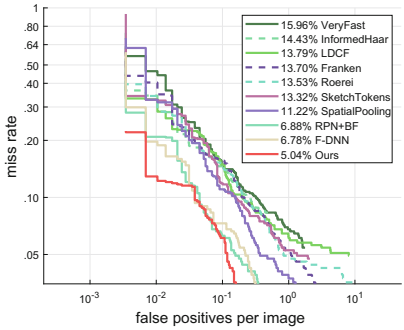


Fig. 7. Comparison with state-of-the-art methods on the INRIA dataset using the *reasonable* setting.

This performance suggests that our detector, guided by fine-grained information, has better capability to identify human body parts and thus to locate occluded pedestrians. In Caltech medium subset, our method has a miss rate of 32.50% which is slightly better than the previous best method [11]. In more reasonable scenarios, our approach achieves comparable performance with the method that achieves best results on Caltech reasonable subset [4].

Since our goal is to propose a fast and accurate pedestrian detector, we have also examined the efficiency of our method. Table 1 compares the running time on Caltech dataset. Our GDFL method is much faster than F-DNN+SS [11] and is about 10× faster than the previous best method on Caltech heavy occlusion subset, JL-TopS [48]. While SDS-RCNN [4] performs slightly better than our method on Caltech reasonable subset (7.36% vs. 7.84%), it needs 4× more inference times than our approach. The comparison shows that our pedestrian detector achieves a favorable trade-off between speed and accuracy.

INRIA: We trained our model with 614 positive images by excluding the negative images and evaluated on the test set. Figure 7 illustrates the results of our approach and the methods that perform best on the INRIA set [2,3,21,28,30,33,44].

Table 2. Comparison with published pedestrian detection methods on the KITTI dataset. The mAP (%) and running time are collected from the KITTI leaderboard.

Method	mAP on easy	mAP on moderate	mAP on hard	Time (s)
FilteredICF [46]	69.05	57.12	51.46	2
DeepParts [37]	70.49	58.68	52.73	1
CompACT-deep [6]	69.70	58.73	52.73	1
RPN+BF [43]	77.12	61.15	55.12	0.6
SDS-RCNN [4]	-	63.05	-	0.21
CFM [18]	74.21	63.26	56.44	2
MS-CNN [5]	83.70	73.62	68.28	0.4
Ours (384 × 1280)	83.78	67.73	60.07	0.15
Ours (576 × 1920)	84.61	68.62	66.86	0.27

Table 3. Comparison with published state-of-the-art methods on MOT17Det benchmark. The symbol * means that external data are used for training.

Method	KDNT* [42]	Our GDFL	SDP [41]	FRCNN [35]	DPM [13]
Average precision	0.89	0.81	0.81	0.72	0.61

Our detector yields the state-of-the-art performance with 5.04% miss rate, outperforming the competitive methods by more than 1%. It proves that our method can achieve great results even if the training set is limited.

KITTI: We trained our model on the KITTI training set and evaluated on the designated test set. We compared our proposed GDFL approach with the current pedestrian detection methods on KITTI [4–6,18,37,43,46]. The results are listed in Table 2. Our detector achieves competitive performance with MS-CNN [5] yet executes about 3× faster with the original input size. Apart its scale-specific property, MS-CNN [5] has explored input and feature up-sampling strategies which are crucial for improving the small objects detection performance. Following this process, we up-sampled 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 and compared with existing methods. Table 3 tabulates the detection results of our method and the state-of-the-art approaches. Our proposed detector achieves competitive 0.81 average precision without using

Table 4. Ablation experiments evaluated on the Caltech test set. Analysis show the effects of various components and design choices on detection performance.

Component	Choice							
Single-layer detection	✓							
Multi-layer detection		✓	✓	✓	✓	✓	✓	✓
Instance-sensitive weight			✓	✓		✓	✓	
Single scale attention			✓					
Scale-aware 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	9.59	**7.36**	8.01	8.86
Miss rate on medium	42.96	36.49	34.36	33.45	34.40	**32.50**	32.99	35.74
Miss rate on heavy occlusion	53.44	50.21	47.60	44.68	47.69	43.18	**42.86**	45.73

external datasets for training. This performance demonstrates the generalization capability of our model.

Ablation Experiments: To better understand our model, we conducted ablation experiments using the Caltech dataset. We considered our convolutional backbone as baseline and successively added different key components to examine their contributions on performance. Table 4 summarizes our comprehensive ablation experiments.

Multi-layer Detection: We first analyzed the advantages of using multiple detection layers. To this end, instead of multi-layer representation, we only used conv_fc7 layer to predict pedestrians of all scales. The experimental results of these two architectures demonstrate the superiority of multi-layer detection with a notable gain of 7% on Caltech Reasonable subset.

Attention Mechanism: We analyzed the effects of our attention mechanism, in particular the difference between single scale attention mask and multiple scale-aware attention masks. To control this, we compared two models with these two attention designs. From Table 4, we can see that both models improve the results, but the model with scale-aware attention has clearly better results. The confusions, such as box-in-box detection, are suppressed with our scale-aware attention masks. We observe an impressive improvement on the Caltech heavy occlusion subset, which demonstrates that the fine-grained masks better capture body parts. Some examples of occlusion cases are depicted in Fig. 8. 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

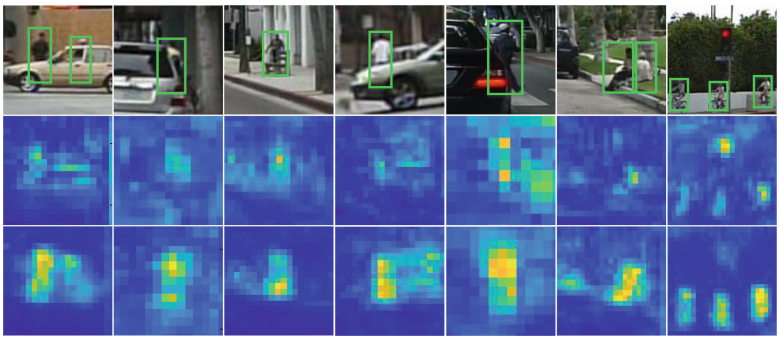


Fig. 8. Hard detection samples where box-based detector is often fooled due to noisy representation. The first row illustrates the images with pedestrians located by green bounding boxes. The second and third rows show the feature maps without attention masks and the graininess-aware feature maps, respectively. Best viewed in color.

becomes able to deduce the occluded parts by considering visible parts, which makes plausible the detection of occluded targets.

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 5^{th} column of Table 4, the performance drops on all three 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 masks generation.

ZIZOM: We further built the zoom-in-zoom-out module on our model with attention masks. Table 4 shows that with the ZIZOM on top of the graininess-aware features $\tilde{F}_{conv4\_3}$, the performance is ameliorated by 1% on all subsets of Caltech. 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 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.

5 Conclusion

In this paper, we have proposed a framework which incorporates pixel-wise information into deep convolutional feature maps for pedestrian detection. We

have introduced scale-aware pedestrian attention masks and a zoom-in-zoom-out module to improve the capability of the feature maps to identify small and occluded pedestrians. Experimental results on three widely used pedestrian benchmarks have validated the advantages on detection robustness and efficiency of the proposed method.

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 U1713214, Grant 61672306, Grant 61572271, and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564.

References

1. Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A.S., Ferguson, D.: Real-time pedestrian detection with deep network cascades. In: BMVC, vol. 2, p. 4 (2015)
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: CVPR, pp. 2903–2910 (2012)
3. Benenson, R., Mathias, M., Tuytelaars, T., Van Gool, L.: Seeking the strongest rigid detector. In: CVPR, pp. 3666–3673 (2013)
4. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection and segmentation. In: ICCV, pp. 4950–4959 (2017)
5. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 354–370. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_22
6. Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: ICCV, pp. 3361–3369 (2015)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
8. Dollár, P., Belongie, S.J., Perona, P.: The fastest pedestrian detector in the west. In: BMVC, vol. 2, p. 7 (2010)
9. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC, pp. 91.1–91.11 (2009)
10. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: CVPR, pp. 304–311 (2009)
11. Du, X., El-Khamy, M., Lee, J., Davis, L.: Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection. In: WACV, pp. 953–961 (2017)
12. Enzweiler, M., Eigenstetter, A., Schiele, B., Gavrila, D.M.: Multi-cue pedestrian classification with partial occlusion handling. In: CVPR, pp. 990–997 (2010)
13. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2010)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR, pp. 3354–3361 (2012)
15. Girshick, R.: Fast R-CNN. In: ICCV, pp. 1440–1448 (2015)
16. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS, pp. 249–256 (2010)
17. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR, pp. 447–456 (2015)

18. Hu, Q., Wang, P., Shen, C., van den Hengel, A., Porikli, F.: Pushing the limits of deep CNNs for pedestrian detection. TCSVT **28**, 1358–1368 (2017)
19. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. TMM **20**, 985–996 (2017)
20. Liang-Chieh, C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
21. Lim, J.J., Zitnick, C.L., Dollár, P.: Sketch tokens: a learned mid-level representation for contour and object detection. In: CVPR, pp. 3158–3165 (2013)
22. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, p. 4 (2017)
23. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
24. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: looking wider to see better. In: ICLR, p. 3 (2016)
25. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
26. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: CVPR, pp. 899–906 (2014)
27. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: CVPR, pp. 3127–3136 (2017)
28. Mathias, M., Benenson, R., Timofte, R., Van Gool, L.: Handling occlusions with franken-classifiers. In: ICCV, pp. 1505–1512 (2013)
29. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: a benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
30. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: NIPS, pp. 424–432 (2014)
31. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR, pp. 3258–3265 (2012)
32. Ouyang, W., Zhou, H., Li, H., Li, Q., Yan, J., Wang, X.: Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. TPAMI **40**, 1874–1887 (2017)
33. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 546–561. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_36
34. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: CVPR, pp. 779–788 (2016)
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
37. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: CVPR, pp. 1904–1912 (2015)
38. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: CVPR, pp. 5079–5087 (2015)
39. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: IJCV, p. 734 (2003)
40. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: ICCV, pp. 32–39 (2009)

41. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: CVPR, pp. 2129–2137 (2016)
42. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: POI: multiple object tracking with high performance detection and appearance feature. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 36–42. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_3
43. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 443–457. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_28
44. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed Haar-like features improve pedestrian detection. In: CVPR, pp. 947–954 (2014)
45. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR, pp. 1259–1267 (2016)
46. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR, p. 4 (2015)
47. Zhou, C., Yuan, J.: Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 305–320. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_19
48. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: ICCV, pp. 3486–3495 (2017)