MULTI-GRAINED DEEP FEATURE LEARNING FOR PEDESTRIAN DETECTION

Chunze Lin[1,2,3], *Jiwen Lu*[1,2,3,*], *Jie Zhou*[1,2,3]

[1]Department of Automation, Tsinghua University, Beijing, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
lcz16@mails.tsinghua.edu.cn; lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn

ABSTRACT

In this paper, we address the challenging problem of detecting pedestrians who are heavily occluded or far from camera. Unlike most existing pedestrian detection methods which only use coarse-resolution feature maps with fixed receptive field, our approach exploits multi-grained deep features to make the detector more robust to visible parts of occluded pedestrians and small-size targets. Specifically, we jointly train a scale-aware network and a human parsing network in a semi-supervised manner with only bounding box annotation. We carefully design the scale-aware network to predict pedestrians of particular scales using most appropriate feature maps, by matching their receptive field with the target sizes. The human parsing network generates a fine-grained attentional map which helps guide the detector to focus on the visible parts of occluded pedestrians and small-size instances. Both networks are computed in parallel and form an unified single stage pedestrian detector, which assures a great trade-off between accuracy and speed. Experiments on two challenging benchmarks, *Caltech* and *KITTI*, demonstrate the effectiveness of our proposed approach, which in addition, executes 2× faster than competitive methods.

***Index Terms*—** Pedestrian Detection, Human Parsing, Attention, Deep Learning

1. INTRODUCTION

Pedestrian detection is one of the most important topics in computer vision and has attracted great attention over past few years [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. It is a key technology in many practical applications such as automotive safety, intelligent video surveillance and human behavior analysis. Despite the recent progress, it is still a challenging problem to detect occluded pedestrians due to the noisy representation and small-size targets because of the low resolution.

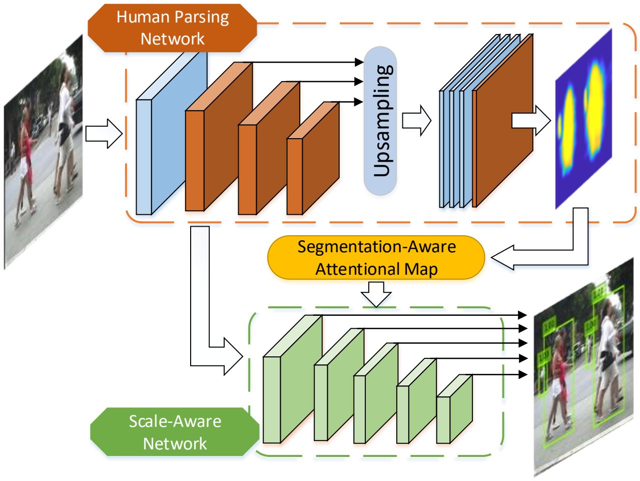


Fig. 1. Overview of the proposed framework. Given an input image, the human parsing network generates fine-grained features which are applied on the detection network as a segmentation-aware attentional map to help focus on visible parts of occluded targets and small-size pedestrians. The detection network is designed to be scale-aware, where multiple feature maps with different receptive fields are employed to detect pedestrian of specific scale.

Existing pedestrian detection methods can mainly be classified into two categories: hand-crafted features based [1, 2, 3] and deep learning features based [4, 5, 11]. For the first category, prior knowledge such as edges and human shape are considered to generate features and decision trees are usually learned by applying boosting to these features to form a pedestrian detector. For the second category, the features are learned via a series of convolutional and pooling layers according to the training data. The convolutional neural network (CNN) generates high-level semantic features which significantly improve the pedestrian detection performance.

While many CNN-based methods have been presented in

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant U1713214, Grant 61572271, and Grant 61527808, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564.

[*]Corresponding author.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

recent years, there are still two shortcomings: 1) most of them usually use feature maps with a single receptive field to deal multi-scale pedestrians. The mismatching between the sizes of targets and receptive fields limits the performance. The small-size instances especially suffer from this inconsistency, which are often ignored when the receptive field is too large; 2) most of them are full-body detectors which are not efficient when dealing with occlusion. Even if some methods learn a set of human part detectors to handle occlusion issue, often only a single feature map with fixed receptive field is employed for the detection.

In this paper, we propose a multi-grained deep features learning (MDFL) based detection system to simultaneously handle the occlusion and small-size problems in pedestrian detection task. Fig.1 illustrates an overview of the proposed framework. Unlike most existing deep learning based methods which only consider a single feature map for detection, we exploit multiple feature maps with different receptive fields and incorporate pixel-wise information to make the detector more robust to occluded and small-size pedestrians. Specifically, we jointly train a scale-aware network and a human parsing network. The scale-aware network is carefully designed to form a feature pyramid and detects pedestrian of specific size with most appropriate feature maps. Concretely, shallower feature maps with small receptive field are reserved in detecting small-size targets while deeper layers are used for large instances. The human parsing network is trained in a weakly supervised way, requiring only bounding box annotation. It generates a fine-grained human parsing mask where regions containing pedestrians are classed as foreground and the rest as background. This mask is then converted into an attention map to make the scale-aware network focus on the presence of pedestrians. Experiments results on challenging pedestrian detection datasets show the superiority of the proposed method. Moreover, since the both networks are computed in parallel, it can execute at least 2× faster than existing pedestrian detection approaches.

2. RELATED WORK

With the prevalence of deep convolutional neural network, most recent methods are CNN-based. In RPN+BF [4], given pedestrian candidates generated by a Region Proposal Network (RPN) [12], higher resolution features were extracted and fed into a boosted forest to handle small-size issue. Instead of boosted forest, SDS-RCNN [11] used the VGG16 net for classification and exploited an additional semantic segmentation loss to implicitly supervise the detector. More similar to our work, MS-CNN [5] integrated a multi-scale network into Faster-RCNN [12] to address the scale problem. While F-DNN [13] used SSD [14] for region proposals and a series of deep classifiers in parallel to post verify each candidate. Besides, some methods [15, 9, 16, 10] learned occlusion-specific detectors, where each one was responsive to detect a human part to handle occlusion issue. These detectors would give a high confidence score based on the parts which are still visible when the full-body detector is confused by the presence of background. Different from the above methods, most of which adopted two-stage pipeline [12], we propose a single stage framework such as in [14] and instead of part-level detection we exploit pixel-wise classification to deal with occlusion and small-size issues.

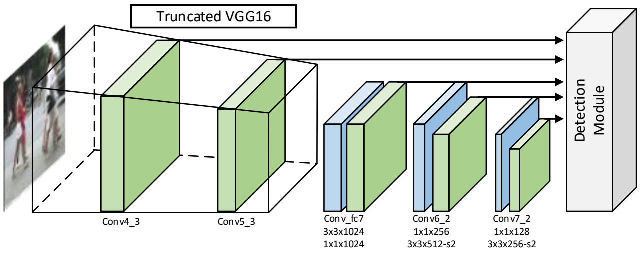


Fig. 2. Architecture of the scale-aware network. It mainly consists of a truncated VGG16 net and extra convolutional layers. The detection layers are presented in green, which are used for multi-scale pedestrian detection. Each detection layer is followed by a detection module for final prediction.

3. APPROACH

Our framework is composed of two key parts: a scale-aware network which detects pedestrian using multi-grained features and a human parsing network which generates a fine-grained attentional map to help the detector focus on regions that contain pedestrians. These two networks are computed in parallel and form a single stage detection framework [14], which offers a great trade-off between accuracy and speed. An overview of the proposed architecture is depicted in Fig.1.

3.1. Scale-Aware Network

The scale-aware network is designed to detect the targets of specific size using feature maps with appropriate resolution and receptive field. Specifically, high-resolution feature maps are used for smaller targets detection while feature maps with larger receptive field are extracted for large-size pedestrians detection.

Architecture of the network: The scale-aware network, as shown in Fig.2, is composed of the following structures:

- Trunk Network: The scale-aware network is based on a truncated VGG16 network where the fully connected layers are converted into convolutional layers. Extra convolutional layers are added to the end of the base network. These layers decrease in size and increase in receptive field progressively in order to cover multi-scale pedestrians.

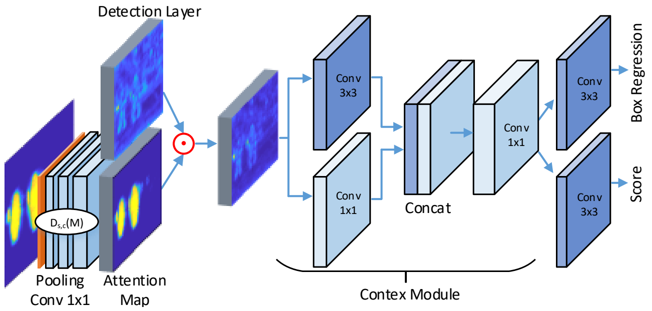


Fig. 3. Architecture of the detection module. The operator $D(M)$ converts the segmentation mask into an attention map which activates most relevant parts of the detection features with element-wise product operation. The concatenation of feature maps with different receptive fields allows to incorporate context information. An 1×1 filter then selects the best features before the prediction.

Table 1. The height of reference box in pixels associated to each prediction layer. The context module provides two different TRFs which mimics the context information incorporation.

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

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

Design of Reference Boxes: A series of reference boxes are placed at each prediction layer and the bounding box regression is based on the offsets with respect to these reference boxes. Since the reference boxes have an important effect on the regression performance, they are carefully designed based on receptive field of the prediction layers. According to [17], in the theoretical receptive field (TRF) of a convolutional layer, center pixels have much more impact comparing to the rest, and as a result, the effective area is in general of Gaussian form. Based on this observation, the height of our reference boxes is designed significantly smaller than the TRF in order to match the effective area (see Table 1). Once the height of the box is determined, the width is computed according to the aspect ratios of pedestrians: $\{0.25, 0.41, 0.52\}$. As the result, six reference boxes of different scales are considered at each location.

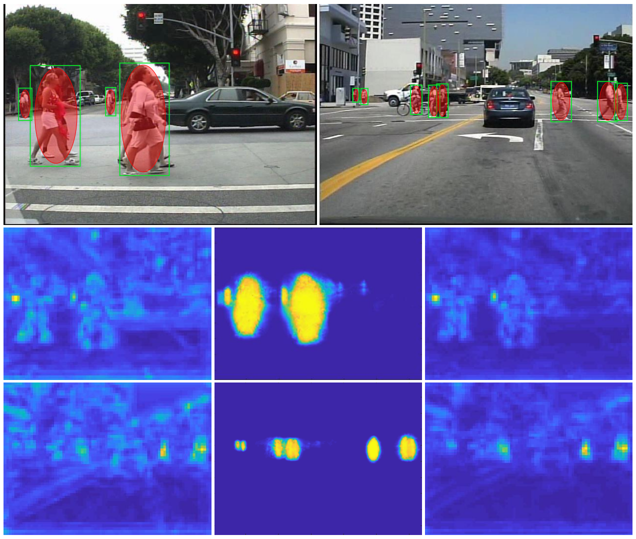


Fig. 4. Illustration of the segmentation results and the attention map effects. **First row:** images with the ground truth bounding boxes drawn in green and the artificial foreground areas presented in red which are used for training segmentation network. **Second row and third row:** Conv4_3 feature map visualization of the left and right images respectively. Representation of initial feature maps (left), the segmentation mask (middle), and the features with attention map (right). The resulting features highlight pedestrians while ignoring most background regions.

3.2. Human Parsing Network

In parallel with the detector, a human parsing network generates a semantic segmentation mask which classes the regions that contain pedestrians as foreground and the rest as background. This mask is then converted into a segmentation-aware attentional map to guide the detector.

Architecture of the Network: The human parsing network is based on the VGG16 network truncated at conv5_3. We change the layer pool4 from 2×2-$s2$ to 3×3-$s1$ and adopt the atrous algorithm [18] to compute more dense feature maps. Each convolutional stage (conv2_2, conv3_3 and conv5_3) is up-sampled to generate feature maps in the size of the input image. The concatenation of these hierarchical maps forms discriminative features which are then followed by an

1×1 convolutional layer and a sigmoid layer to output pedestrian segmentation. Note that the stem parts (conv1-conv2) are computationally expensive, we share these layers with the scale-aware network. The architecture is depicted in the top part of Fig.1.

Weakly Supervised Training: In general, only bounding box annotations are provided in pedestrian detection tasks. Therefore, to train our human parsing network, we follow a weakly supervised strategy by creating artificial foreground segmentation using bounding box information. In practice, we consider the center area of the bounding box (80% of pixels within the box) as foreground, as shown the first row in Fig.4. This process considerably eliminates background inside the bounding box while keeping the main parts of pedestrian. Some segmentation results are depicted in the middle column of Fig.4. We can verify that, despite the weak annotation, the small targets are effectively highlighted.

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

$$\mathrm{A}_{s,c} = D_{s,c}(\mathrm{M}) \odot \mathrm{F}_{s,c} \tag{1}$$

where $D_{s,c}(\mathrm{M})$ down-samples M by s times and outputs with c channels and $\odot$ is the Hadamard operator. Some results are depicted in Fig.4, which show that the conv4_3 feature maps with the attention mechanism become more focused on pedestrians and the background is significantly smoothed.

3.3. Training Objective

Our framework has two sibling output layers and an intermediate segmentation output layer. The first outputs bounding-box regression offsets, $\mathbf{d} = (d^x, d^y, d^w, d^h)$. The parameterization for $\mathbf{d}$ is as in [19], in which it specifies a scale-invariant translation and log-space height/width shift relative to a reference box. The second branch outputs the detection confidence score (c), computed by a softmax over two classes (pedestrian v.s. background). The intermediate output corresponds to the segmentation result, in which each pixel is classified as pedestrian or background. We use a multi-task loss L to train the scale-aware and human parsing networks:

$$L \quad = \quad L_{\text{box}} + \lambda_c L_{\text{conf}} + \lambda_s L_{\text{seg}} \tag{2}$$

The box regression loss L_{box} targets at minimizing the Smooth L1 loss $R(\mathbf{d}, \hat{\mathbf{g}})$ defined in [19], between the estimated parameters $(\mathbf{d})$ and the ground truth box regression targets $(\hat{\mathbf{g}})$, where $\hat{\mathbf{g}}$ has the same parametrization as $\mathbf{d}$.

$$L_{\text{box}} \quad = \quad \frac{1}{N} \sum_{i \in Pos}^{N} \sum_{k \in \{x,y,w,h\}} x_{ij} R(d_i^k - \hat{g}_j^k) \tag{3}$$

where $x_{ij} = \{1, 0\}$ is an indicator for matching the i-th reference box to the j-th ground truth box and N is the number of matched reference boxes. If $N = 0$, we set the loss of detection module to 0. In our implementation, we begin by matching each ground truth box to the reference box with the best intersection over union (IoU) and we then match reference boxes to any ground truth with IoU higher than 0.5.

The confidence score loss L_{conf} is the softmax loss and the cross-entropy loss L_{seg} is used for the segmentation. In our experiments, we regularize our multi-task loss by setting the weight terms $\lambda_c = \lambda_s = 1$.

3.4. Implementation Details

Training: Our scale-aware network and human parsing network were partially initialized with the detection model of [14] and the DeepLab segmentation model [20], respectively. All new additional layers were randomly initialized with the Xavier [21]. In order to facilitate the convergence, we first trained the two networks separately and then the both networks were jointly optimized. Specifically, the scale-aware network was fine-tuned for $50k$ iterations where we used 10^{-4} learning rate for the first $40k$ iterations then continued with 10^{-5} for the rest iterations. The human parsing network was fine-tuned for $80k$ iterations with a learning rate of 10^{-8}. Then the both networks are jointly optimized for $20k$ iterations. All our implementations were based on Caffe framework [22].

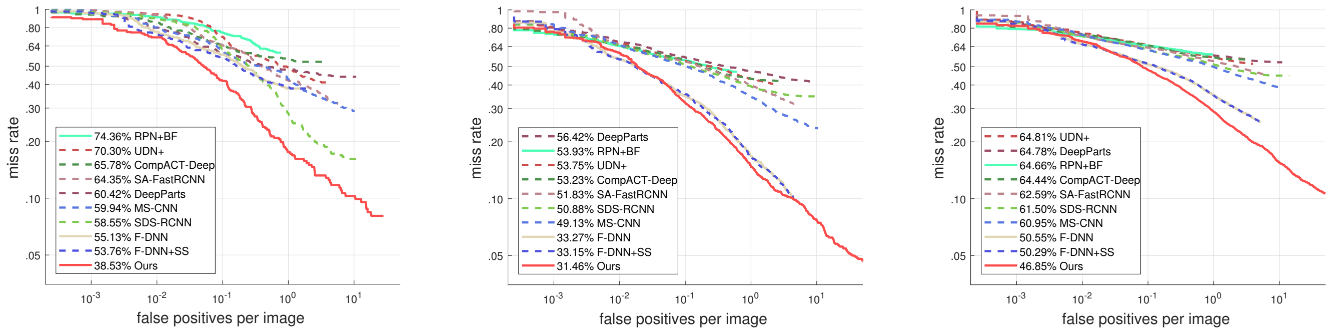
Hard negative mining: Our detector has to evaluate a considerable number of reference boxes, yet only a few locations contain pedestrians, which causes a significant class imbalance during training. For more stable training, instead of using all negative samples, we sorted them by the highest loss values and kept the top ones so that the ratio between the negatives and positives is at most 5:1.

Data augmentation: To make our model more robust to sizes and illumination variations, we adopted following data augmentation strategies: color distortion, random crop, expansion and horizontal flip.

4. EXPERIMENTS

4.1. Datasets and Evaluation Protocols

We conducted experiments on two challenging pedestrian detection datasets, Caltech [23] and KITTI [24] datasets, to eval-



(a) Heavy Occluded

(b) Medium

(c) Overall

Fig. 5. Comparison with the state-of-the-art methods on the Caltech dataset.

uate our proposed MDFL method and compared it with state-of-the-art pedestrian detection approaches. Here we give a brief description of these datasets.

Caltech [23]: Caltech dataset consists of ∼10 hours of urban driving video with $350K$ labeled bounding boxes. The log-average miss rate is used to evaluate the detection performance and is calculated by averaging miss rates at 9 false positive per-image (FPPI) points sampled within the range of $[10^{-2}, 10^{0}]$. In our experiments, three subsets were considered to demonstrate the performance on occlusion and small-size issues: *Heavy Occluded*, *Medium* and *Overall*. In the *Heavy Occluded* subset, pedestrians are 36-80% occluded, in the *Medium* subset, pedestrians are of 30-80 pixels height without occlusion and the *Overall* subset consists of all pedestrians taller than 20 pixels with or no occlusion.

KITTI [24]: KITTI dataset consists of 7481 training images and 7518 test images, comprising about 80K annotations of cars, pedestrians and cyclists. KITTI evaluates the PASCAL-style mean Average Precision (mAP) under three difficulty levels: easy, moderate and hard. Under moderate setting, which is used to rank the competing methods in the benchmark, the pedestrians taller than 25 pixels with or no occlusion are considered.

4.2. Results and Analysis

Caltech: We used the Caltech training set, which contains 42,782 training images, to train our detection system and evaluated it on the Caltech testing set. We compared our proposed MDFL with the methods that have achieved great performance on Caltech [4, 9, 26, 25, 16, 5, 11, 13, 10]. As shown in Fig.5, our method achieves an impressive 38.53% and 31.46% miss-rate on *Heavy Occluded* and *Medium* subsets respectively, which outperforms considerably the current methods. The comparison with the most recent part-detector based approach [10] (ROC plot not released) which has achieved 49.20% miss-rate on the *Heavy Occluded* subset, demonstrates the effectiveness of our MDFL model to handle occlusion issues. The performance on the *Medium* subsets shows the capability of our approach to deal with small-size pedestrians. We also evaluated our method on the *Overall* subset to analyze its performance on more general setting. A miss rate of 46.85% is observed (Fig.5(c)), which points out the generalization capability of our MDFL.

Table 2. Comparison of our method with the state-of-the-art approaches in terms of trade-off between accuracy and speed. Caltech-Heavy miss rate, KITTI mAP score and runtime are tabulated.

Method	Caltech	KITTI	Runtime
RPN+BF [4]	74.36	61.29	0.5s
SA-FastRCNN [25]	64.35	65.01	0.5s
DeepParts [16]	60.42	58.67	1s
MS-CNN [5]	59.94	73.70	0.14s
SDS-RCNN [11]	58.55	63.05	0.21s
F-DNN[13]	55.13	-	0.3s
F-DNN+SS [13]	53.76	-	2.48s
JL-Tops [10]	49.20	-	0.6s
Ours	38.53	66.32	0.07s

KITTI: We used the KITTI training set to train our pedestrian detector and evaluated on the test set, considering only the pedestrian class. Our method achieves 66.32mAP on the moderate setting for pedestrian class, which outperforms most approaches [4, 25, 16, 11]. The comparison results are shown in the third column of Table 2. Note that in the KITTI evaluation, cyclists are counted as false positives and person-sitting are ignored, while on Caltech these two classes are labeled as pedestrians. Since semantic segmentation information are useful for detecting person in unusual poses, this advantage is less helpful on KITTI.

Runtime Analysis: Efficiency is one of the advantages of our single stage framework, and here we give a short analy-

Table 3. Performance analysis when key components are successively disabled on the Caltech test set.

Component Disabled	Medium	Heavy	Overall
Context module	35.31	44.37	49.99
Segmentation mask	33.27	40.27	47.83
Our-MDFL	31.46	38.53	46.85

sis of runtime. Our method takes 0.07s/image with an input image of size 640×480 on a single Nvidia 1080Ti GPU. Compared to the most methods, our approach executes 2× faster (fourth column of Table 2). Specifically, compared to JL-Tops [10], which was proposed to handle occlusion issue, our method is 8× faster. The comparison shows the effectiveness of the proposed MDFL detector.

Ablation Study: In order to analyze the contribution of key components of our framework on performance, we successively removed each component and evaluated on Caltech, as summarized in Table 3. When the segmentation-aware attentional map was disabled, the performance degraded by 2% on the *Medium* and *Heavy Occluded* subsets. When we further removed the context module, the performance for detecting small-size targets and occluded pedestrians dropped by 2% and 4%, respectively. The ablation analysis confirms that the segmentation and context information effectively make the detector more robust to small-size and occluded targets.

5. CONCLUSION

In this paper, we have proposed a multi-grained deep feature learning based method for pedestrian detection. By jointly training a scale-aware network and a human parsing generator, our approach exploits pixel-wise segmentation information, background context and multi-scale property to handle simultaneously the occlusion and small-size issues. The whole detection system is a single stage framework, assuring a great accuracy/speed trade-off. The proposed method has achieved impressive performance on challenging pedestrian detection datasets, outperforming most existing approaches while executing 2× faster. How to efficiently convert our model into a video based method and incorporate temporal information to further boost the performance appears to be an interesting future work.

6. REFERENCES

[1] Piotr Dollár, Zhuowen Tu, Pietro Perona, and Serge Belongie, “Integral channel features,” in *BMVC*, 2009.

[2] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, “Fast feature pyramids for object detection,” *TPAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[3] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, “Filtered channel features for pedestrian detection,” in *CVPR*, 2015.

[4] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, “Is faster r-cnn doing well for pedestrian detection?,” in *ECCV*, 2016.

[5] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *ECCV*, 2016.

[6] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Pedestrian detection aided by deep learning semantic tasks,” in *CVPR*, 2015.

[7] Arthur Daniel Costea and Sergiu Nedevschi, “Semantic channels for fast pedestrian detection,” in *CVPR*, 2016.

[8] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe, “Learning cross-modal deep representations for robust pedestrian detection,” in *CVPR*, 2017.

[9] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang, “Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection,” *TPAMI, accepted*, 2017.

[10] Chunluan Zhou and Junsong Yuan, “Multi-label learning of part detectors for heavily occluded pedestrian detection,” in *ICCV*, 2017.

[11] Garrick Brazil, Xi Yin, and Xiaoming Liu, “Illuminating pedestrians via simultaneous detection and segmentation,” in *ICCV*, 2017.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015.

[13] Xianzhi Du, Mostafa El-Khamy, Jungwon Lee, and Larry Davis, “Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection,” in *WACV*, 2017.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, “Ssd: Single shot multibox detector,” in *ECCV*, 2016.

[15] Wanli Ouyang and Xiaogang Wang, “A discriminative deep model for pedestrian detection with occlusion handling,” in *CVPR*, 2012.

[16] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, “Deep learning strong parts for pedestrian detection,” in *CVPR*, 2015.

[17] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in *NIPS*, 2016.

[18] Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” in *ICLR*, 2015.

[19] Ross Girshick, “Fast r-cnn,” in *ICCV*, 2015.

[20] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *arXiv*, 2016.

[21] Xavier Glorot and Yoshua Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *AISTATS*, 2010.

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in *ACMMM*, 2014.

[23] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, “Pedestrian detection: A benchmark,” in *CVPR*, 2009.

[24] Andreas Geiger, Philip Lenz, and Raquel Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *CVPR*, 2012.

[25] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan, “Scale-aware fast r-cnn for pedestrian detection,” *arXiv*, 2015.

[26] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos, “Learning complexity-aware cascades for deep pedestrian detection,” in *ICCV*, 2015.