DIRECT MULTI-SCALE DUAL-STREAM NETWORK FOR PEDESTRIAN DETECTION

Sang-Il Jung and Ki-Sang Hong

POSTECH
Department of Electrical Engineering
Pohang, Korea

ABSTRACT

We propose Direct Multi-scale Dual-stream network (DMDnet) for pedestrian detection. DMDnet takes a full-size image as input and detects pedestrians of various sizes directly without extracting proposals or using resampling. To improve detection accuracy we adopt dual-stream architecture that combines two types of features that are branched off from two different layers. The lower layer observes the proper sizes of receptive fields depending on the scales of pedestrians; the upper layer contains contextual information. The whole network is trained end-to-end. In experiments DMDnet yields state-of-the-art detection accuracy on the Caltech pedestrian benchmark with new annotation, and has quite fast detection speed.

***Index Terms*—** Pedestrian detection, deep convolutional neural network

1. INTRODUCTION

Deep Convolutional Neural Networks (DCNNs) have been widely applied to various visual recognition problems, but for the task of pedestrian detection, DCNNs are too slow. For example, applying DCNN in a sliding-window fashion is computationally infeasible. To solve this problem, most of the previous pedestrian detection methods based on DCNN [1, 2, 3] adopt a proposal-and-classification strategy; the candidate regions (proposals) are extracted using fast AdaBoost-based pedestrian detector such as ACF [4] or LDCF [5], and only the extracted proposals are fed into a DCNN for classification and localization. This approach achieves a good trade-off between detection accuracy and detection speed. However, because the bounding boxes of the proposals tend to overlap one another, it causes redundant computation. To avoid this redundant computation, SA-FastRCNN [6] adopted the Fast R-CNN approach [7] that shares convolutional features by cropping proposal regions from the convolutional feature map instead of from the input image. MS-CNN [8] integrated multiple region-proposal networks (RPNs) with a detection network that is similar to Faster R-CNN [9]. RPN+BF [10] uses a DCNN to extract proposals, then classifies them using boosted forests with the trained DCNN features. All existing pedestrian detection methods based on DCNN adopt two-step processes of extracting proposals and classifying them.

In this paper, we designed a 'Direct' detector that takes a full-size image as input and detects pedestrians of various sizes directly without extracting proposals or using resampling. To the best of our knowledge, this is the first direct DCNN detector for pedestrian detection, although some methods [11, 12] use a DCNN to detect objects directly.

A certain layer of a DCNN has a specific receptive field; i.e., the size of the receptive field increases as the level of the layer in the hierarchy increases. It is important to see the appropriate sizes of receptive fields according to the sizes of the pedestrians [8]. MS-CNN [8] extracted proposals of small objects from lower layer of the DCNN and large objects from higher layers. Following the way of RPNs in MS-CNN, we configure 'Multi-scale' detection networks by branching off from multiple layers depending on the sizes of the pedestrians. Unlike MS-CNN, we used these networks as detectors for final detection, not as RPNs to extract proposals.

We also consider contextual information by also seeing wider receptive fields than the proper receptive fields. The size of receptive field and the amount of semantic information increase as the level of the layer in the hierarchy increases. Thus, the higher layers give more contextual and semantic information. The semantic information of the higher layers is proven to be effective for object detection [13, 14]. In our method, the contextual networks are branched off from the higher layers and integrated to the detection networks ('Dual-stream'). These additional contextual networks are nearly cost-free, but improve the detection accuracy greatly.

Combining these concepts, we propose Direct Multi-scale Dual-stream network (DMDnet) for pedestrian detection. The DMDnet has three benefits: 1) it detects pedestrians without extracting proposals or using resampling ('Direct'); 2) it detects pedestrians of various sizes by splitting networks depending on the scales of pedestrians ('Multi-scale'); and 3) it combines two types of features by branching off from two different layers ('Dual-Stream'). The whole network is trained end-to-end. DMDnet was evaluated on the Caltech pedestrian benchmark and achieved new state-of-the-art detection accuracy (with new annotation). The detection speed is also quite fast.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

2. DIRECT MULTI-SCALE DUAL-STREAM NETWORK

In this section, we describe in detail the proposed DMDnet for pedestrian detection.

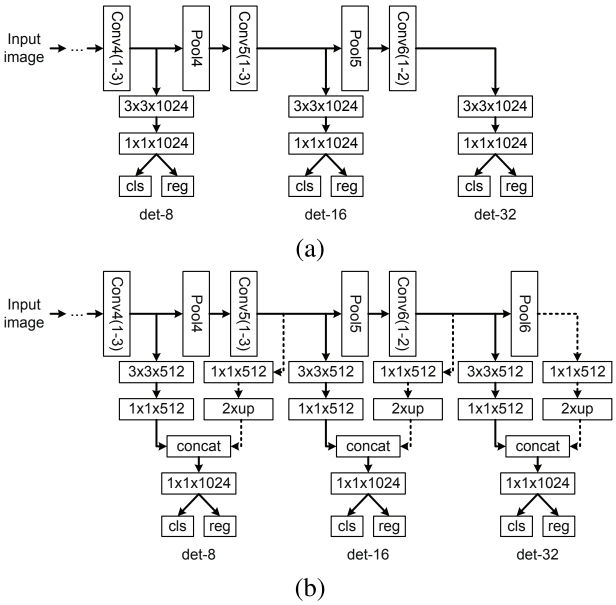
2.1. Overview

DMDnet resembles the RPNs of Faster R-CNN [9] and MS-CNN [8] that define anchor boxes and classify them. The anchor boxes [9] are a set of bounding boxes that have multiple scales and aspect ratios. The anchor boxes are located on the image with some strides where each position on the feature map corresponds to an anchor box on the image. For the pedestrian detection task, the aspect ratio is fixed to 0.41 [16], so we only consider the scales of the anchor boxes. We set the size of the smallest anchor box to 50×20.5 pixels, and generate anchor boxes of eight different scales with scale stride 1.3; i.e., the heights of the anchor boxes are $\{50, 65, 84.5, 109.8, 142.8, 185.6, 241.3, 313.7\}$. These anchor boxes are classified and localized to be aligned with pedestrians using the proposed network. Then, non-maximum suppression with an Intersection-Over-Union (IOU) threshold of 0.5 is conducted as post-processing to eliminate multiple detections of a pedestrian.

2.2. Network architecture

We configured DMDnet based on the VGG-16 network [17], which consists of five convolutional blocks and three fully-connected layers. Only the convolutional layers (up to the fifth pooling layer) are used for the proposed network. First, we designed a Direct Multi-scale network (DMnet) (Fig. 1a) that resembles the RPN of MS-CNN [8]. Based on the convolutional layers of the VGG-16 network ($\sim$ 'Pool5' layer), we added two more convolutional layers (Conv6(1-2) in Fig. 1a) after the 'Pool5' layer. Then, three detection networks branch off from the layers 'Conv4-3', 'Conv5-3' and 'Conv6-2' depending on the sizes of the anchor boxes. We named the detectors 'det-8', 'det-16' and 'det-32' following MS-CNN [8] where the number k of 'det-k' indicates the stride of the anchor boxes; e.g., the anchor boxes of 'det-8' slides the input image with 8-pixel stride that is caused by the three 2×2 pooling layers with stride 2. Each detection network contains convolutional layers of $3 \times 3 \times 1024$ and $1 \times 1 \times 1024$, and sibling layers for classification and bounding box regression. The features extracted from each branch see the region of the proper receptive fields according to the sizes of the anchor boxes; for brevity, we call these features the proper-receptive-field (PRF) features.

We designed the DMDnet (Fig. 1b) by adding contextual information to the DMnet. The networks for PRF features branch off in the same manner as in DMnet, but we reduce the dimensions of the convolutional layers from 1024 to 512. We additionally branch the layers for contextual features for each detector. The contextual networks for 'det-8' and 'det-16' branch off from the layers of 'Conv5-3' and 'Conv6-2', respectively. For the 'det-32', we added a sixth max-pooling layer ('Pool6'), and branch off from that layer. Each contextual network (dotted arrows in Fig. 1b) contains one $1 \times 1 \times 512$ convolutional layer and one upsampling layer, which conducts bilinear interpolation to generate a contextual feature map that is a spatially upscaled by a factor of 2 so that the spatial dimensions of contextual feature map match those of PRF features. The PRF features and the contextual features are concatenated and are fed into $1 \times 1 \times 1024$ convolutional layer followed by two sibling layers for classification and bounding box regression.



(a)

(b)

Fig. 1. Network architectures. (a) Direct Multi-scale network (DMnet) that resembles RPN of MS-CNN [8] and (b) Direct Multi-scale Dual-stream network (DMDnet).

The eight types of anchor boxes are assigned to the detection networks appropriately depending on the sizes of the receptive fields. For example, the branching layer of 'det-8' is 'Conv4-3', which has a receptive field of 92×92 pixels. Therefore, three types of anchor boxes of heights $\{50, 65, 84.5\}$ are assigned to 'det-8' and these anchor boxes slides the input image with stride of 8 pixels. In the same manner, we assigned $\{109.8, 142.8, 185.6\}$ to 'det-16' and $\{241.3, 313.7\}$ to 'det-32'.

2.3. Loss

Each detection network has two output layers, one for classification and one for bounding box regression. The classification output layer indicates whether or not the anchor boxes contain pedestrians; the bounding box regression layer estimates the transformation offsets to align the anchor boxes to the pedestrians. Two-way softmax-log loss is used for classification, and smooth$_{L1}$ loss is used for bounding box regression [7]. The two losses are combined with hyper-parameter γ and all the losses of the three detection networks that contain K = 8 types of anchors are summed:

$$L = \sum_{k=1}^{K} \sum_{\mathbf{b}_i \in B^k} \Big(l_{cls}^{k}(\hat{p}_i, y_i) + \gamma l_{loc}^{k}(\hat{\mathbf{t}}_i, \mathbf{t}_i) \Big), \qquad (1)$$

where B^k indicates the box that belongs to the k-th anchor box type; $\mathbf{b}_i$ indicates the i-th box; $l_{cls}^{k}(\hat{p}_i, y_i)$ is the classification loss of the i-th box, where $\hat{p}_i$ is the estimated probability and y_i is the ground-truth class label; $l_{loc}^{k}(\hat{\mathbf{t}}_i, \mathbf{t}_i)$ is the bounding box regression loss of the i-th box, where $\hat{\mathbf{t}}_i$ and $\mathbf{t}_i$ indicate the estimated offsets and target offsets. The parameterization of $\mathbf{t}$ is the same as used in [7], but because we fix the aspect ratio, we use a single scale transformation instead of two.

2.4. Sampling

The samples indicate the boxes $\{\mathbf{b}_i\}_{i=1}^{N}$ that include all types of anchor box at every location of the images in a mini-batch. In our methods, we used M = 2 images as a mini-batch for each iteration. For each box $\mathbf{b}_i$ we calculated IOUs with ground-truth bounding boxes; we set the class label of $\mathbf{b}_i$ as positive if the maximum IOU (maxIOU) $\geq$ 0.6, as negative if maxIOU $\leq$ 0.4, and ignore the samples if 0.4 $<$ maxIOU $<$ 0.6. Because the numbers of positive and negative samples are highly imbalanced, we select the samples for computing the classification loss as follows: we select $n_p \leq 32$ positive samples and $n_n = \min(32, 3n_p)$ negative samples. For the bounding box regression loss, we select $n_l \leq 32$ samples that have IOU $\geq$ 0.45 with a ground-truth bounding box. We used random sampling for the first few iterations (e.g., 12,000 iterations in our experiments), then bootstrap sampling for the rest of iterations following MS-CNN [8].

2.5. Learning details

A pedestrian dataset usually contains many more small pedestrians than large ones. To balance the number of positive samples depending on the anchor types, the training images are randomly resized, horizontally flipped and cropped for data augmentation.

The five convolutional blocks were initialized by the pre-trained parameters of the VGG-16 network [17] on the ImageNet dataset [18]. The parameters of all other additional layers were initialized randomly using a Gaussian distribution with mean = 0 and variance = 0.01. The stochastic gradient descent method was used to update parameters with momentum 0.9. The learning rates were 0.001 for 70,000 iterations, then 0.0001 for 10,000 iterations, then 0.00001 for 5,000 iterations. The hyper-parameter γ to balance classification loss and bounding box regression loss was set to 0.05 during first 12,000 iterations, and to 1 thereafter.

3. EXPERIMENTAL RESULTS

We evaluated the proposed network on the Caltech pedestrian benchmark [16]. For training, we used 42,782 training images that were obtained by sampling every 3rd frame from training videos, as was done for SCF [1] and RPN+BF [10]. We used the standard 4,204 testing images for testing. We evaluated the detection accuracy by measuring the average miss rate MR_{-2} over the false-positive-per-image (FPPI) range $[10^{-2}, 10^{0}]$. This evaluation process is the standard procedure for the Caltech dataset [16]. We also evaluated the proposed method on a new annotation [19] of the Caltech dataset that sanitizes the original annotation [16].

To evaluate the effectiveness of contextual network, both DMnet and DMDnet (Fig. 1a,b) were trained and tested on the Caltech dataset with the original annotation. MR_{-2}^{O}s were measured, where O stands for original annotation of testing set (Fig. 2). The DMnet achieved MR_{-2}^{O} = 12.97%, and DMDnet achieved MR_{-2}^{O} = 10.19%; the 2.8% reduction shows that the contextual network for DMDnet is effective.

We compared the detection accuracy of DMDnet with those of the state-of-the-art methods. The proposed method did not achieve the best detection accuracy, but achieved comparable results with the state-of-the-art methods (Table 1). The miss rate of our DMDnet (MR_{-2}^{O} = 10.19%) is slightly higher than the best-performing methods MS-CNN (9.95%), SA-FastRCNN (9.68%) and RPN+BF (9.58%).

To analyze the results, we observed the top-40 false positives of the detection results of DMDnet; these correspond to 10^{-2} FPPI because the number of test images is 4,204. Of these false positives, 11 were missing annotation and 10 had badly-localized annotation; i.e., more than half of the false positives occurred because of errors in original annotation. Thus, we also evaluated the methods with the new annotation.

All methods were trained with the original annotation, then tested on the Caltech testing set with the new annotation. DMDnet yielded a new state-of-the-art detection accuracy (MR_{-2}^{N} = 5.78%), compared to MS-CNN (MR_{-2}^{N} = 8.08%), SA-FastRCNN (MR_{-2}^{N} = 7.47%) and RPN+BF (MR_{-2}^{N} = 7.28%) (Table 2). With the original annotation, the miss rate of DMDnet (10.19%) was slightly higher than MS-CNN (9.95%), but with the new annotation DMDnet (5.78%) beat MS-CNN (8.08%) by a large margin (2.3%); DMDnet was also more accurate than the second-best RPN+BF (7.28%). The new annotation contains fewer errors than the original annotation, so evaluation is more reliable with the new annotation than with the original annotation.

We also trained DMDnet with the new annotation. The MR_{-2}^{N} of DMDnet trained with new annotation (DMDnet (New)) is 4.89% (Table 2); training with new annotation reduced MR_{-2}^{N} by $\sim$ 1% where the MR_{-2}^{N} of 'DMDnet (Ori)'

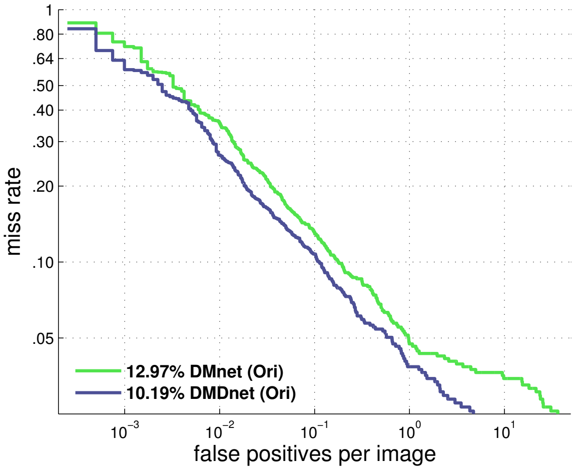


Fig. 2. The miss rates over FPPI of DMnet and DMDnet. The evaluation was conducted on the Caltech testing set with original annotation (MR^O_{-2}). The '(Ori)' indicates that the networks were trained with original annotation.

Table 1. Evaluation on the Caltech testing set with original annotation (MR^O_{-2}).

Methods	MR^O_{-2}	Methods	MR^O_{-2}
LDCF [5]	24.80	DeepParts [3]	11.89
SCF+AlexNet [1]	23.32	CompACT-Deep [20]	11.75
Katamari [21]	22.49	MS-CNN [8]	9.95
TA-CNN [2]	20.86	SA-FastRCNN [6]	9.68
CCF+CF [22]	17.32	RPN+BF [10]	9.58
Checkerboards+ [23]	17.10	**DMDnet (Ori)**	**10.19**

Table 2. Evaluation on the Caltech testing set with new annotation (MR^N_{-2}). The '(New)' indicates that the networks is trained with new annotation. All other methods except 'DMDnet (New)' were trained with the original annotation.

Methods	MR^N_{-2}	Methods	MR^N_{-2}
LDCF [5]	23.72	DeepParts [3]	12.90
CCF+CF [22]	22.34	CompACT-Deep [20]	9.15
Katamari [21]	22.18	MS-CNN [8]	8.08
SCF+AlexNet [1]	21.59	SA-FastRCNN [6]	7.47
TA-CNN [2]	18.75	RPN+BF [10]	7.28
Checkerboards+ [23]	16.33	**DMDnet (Ori)**	**5.78**
		DMDnet (New)	**4.89**

Table 3. Detection accuracy and speed. The detection speeds of other methods [8, 6, 10] were obtained from the original papers. Due to the differences in hardware and implementation details, direct comparison is invalid, but this roughly shows that DMDnet is computationally efficient.

Methods	MR^N_{-2}	Detection speed
MS-CNN [8]	8.08%	8 im/s
SA-FastRCNN [6]	7.47%	2.7 im/s
RPN+BF [10]	7.28%	2 im/s
DMDnet	**5.78%**	**8.4** im/s

was 5.78%. Comparing the detection accuracy of 'DMDnet (New)' with the other detection accuracies in Table 2 might be invalid, but we report it just for comparison for future studies.

We also measured detection speed on the Caltech Pedestrian testing images where the resolution is 640 × 480. We used the MatConvNet [24] library to design, train and test the proposed network on an Intel i7 3.60-GHz CPU and TitanX GPU. Because a DCNN is less discriminative for small pedestrians than for large ones, most DCNN-based methods upscale the sizes of the input images by a factor of 1.5 $\sim$ 2. The input images were resized, so that their short edges had 720 pixels for MS-CNN, 800 pixels for SA-FastRCNN, and 720 pixels for RPN+BF, while maintaining the aspect ratio. We also increased the sizes of the input images, but only the image width by a factor of 2. Because a bounding box of pedestrian is narrower than tall, increasing horizontal resolution is important for small pedestrians. The detection speeds of other methods were obtained from the original papers. Because the environments of implementation were not the same, direct comparision is invalid, but we can roughly see that DMDnet seems to be computationally efficient (Table 3) with high detection accuracy. The detection speed of DMDnet is 8.4 im/s with MR^N_{-2} = 5.78% whereas the detection speed of RPN+BF is 2 im/s with MR^N_{-2} = 7.28%. MS-CNN had comparable detection speed (8 im/s), but much higher MR^N_{-2} (8.08%) than DMDnet.

4. CONCLUSION

We proposed Direct Multi-scale Dual-stream network (DMDnet) for pedestrian detection. Unlike existing DCNN-based pedestrian detectors, DMDnet directly detects pedestrians of various sizes without extracting proposals, or applying resampling. To improve detection accuracy, we used a dual-stream architecture to combine proper-receptive-field features and the contextual features. DMDnet yields a new state-of-the-art detection accuracy on the Caltech dataset with new annotation and the detection speed is quite fast.

5. REFERENCES

[1] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele, "Taking a deeper look at pedestrians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4073–4082.

[2] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, 2015, pp. 5079–5087.

[3] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1904–1912.

[4] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 8, pp. 1532–1545, 2014.

[5] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[6] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, and Shuicheng Yan, "Scale-aware fast r-cnn for pedestrian detection," *arXiv preprint arXiv:1510.08160*, 2015.

[7] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[8] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, "Is faster r-cnn doing well for pedestrian detection?," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," *arXiv preprint arXiv:1612.03144*, 2016.

[14] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[15] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation," *arXiv preprint arXiv:1611.06612*, 2016.

[16] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[19] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele, "How far are we from solving pedestrian detection?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1259–1267.

[20] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3361–3369.

[21] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele, "Ten years of pedestrian detection, what have we learned?," in *European Conference on Computer Vision*. Springer, 2014, pp. 613–627.

[22] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li, "Convolutional channel features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 82–90.

[23] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, "Filtered channel features for pedestrian detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 1751–1760.

[24] Andrea Vedaldi and Karel Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, 2015, pp. 689–692.