Pedestrian Detection Aided by Scale-Discriminative Network

Zongqing Lu, Wenjian Zhang, Qingmin Liao
Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua University, China
Shenzhen Key Lab. of Information Sci & Tech / Shenzhen Engineering Lab. of IS & DRM

***Abstract*—Deep learning is greatly successful when used for pedestrian detection. However, we find that this method is barely satisfactory for multi-scale detection. Meanwhile, various solutions such as multi-scale classifiers have been developed (based on traditional methods) to handle this situation. Considering this, we propose a scale-discriminative classifier layer (SDC) that contains numerous classifiers to cope with different scales. To expand the capacity for small-scale pedestrian detection, we construct a full-scale layer that converges both high-level semantic features and low-level features. From the analysis above, a scale-discriminative network (SDN) for pedestrian detection was born. We apply this network to the Caltech pedestrian dataset, and the experimental results show that the SDN achieves state-of-the-art performance.**

1. Introduction

Pedestrian detection is a major topic in the computer vision research community. Numerous approaches have been brought forward over the past few decades. Researchers are primarily interested in pedestrian detection because of its usefulness in various applications, such as video surveillance, robotic navigation and driving safety.

Existing pedestrian detection methods fall into two categories: Models based on handcrafted features and those based on deep learning.

Models based on handcrafted features have achieved great success. VJ [1] employed multi-scale Haar wavelets to describe objects and presented integral images to detect objects. Afterwards, a framework [2] that combined histogram of gradients (HOG) with a linear support vector machine (SVM) was introduced to discriminate objects from background. This framework marked a breakthrough in pedestrian detection. The integral channel feature (ICF) [3] was proposed to combine various features such as LUV color channels, normalized gradient magnitude and HOG. The aggregated channel features (ACF) [4], InformedHaar [5], and Checkboards [6] adopted the same channel features as ICF. ACF calculated the pixel sum of each block in channels and the resulting lower resolution channels were smooth. InformedHaar was designed specifically for pedestrian detection using the statistical templates of upright pedestrian bodies. Checkboards used class 6 filters to compute features from channel images, which was the generalization from the ICF.

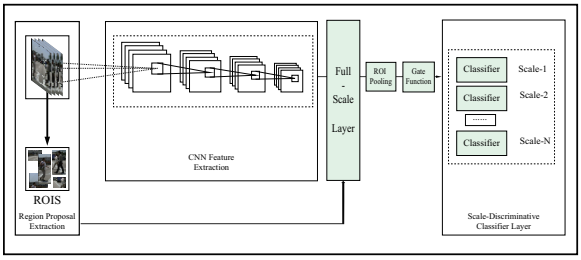


Figure 1. The framework of the scale-discriminative network (SDN). There are two key components, **full-scale layer** and **scale-discriminative classifier layer**.

In the second category, convolutional neural network (CNN) methods[1] have shown promising results. Task-assistant CNN (TA-CNN) [7] improved the performance of pedestrian detection by introducing and optimizing semantic tasks. DeepParts [8] constructed a wide part pool where different parts could be selected in a data-driven manner to handle the occlusion situation. Regions with CNN features (R-CNN) [9] first combined the region proposal extraction method with CNN features, achieving a 30% increase in PASCAL VOC 2012 [10]. Fast R-CNN [11] definitely enhanced the accuracy and efficiency of R-CNN. The region of interest (ROI) pooling process made the convolutional layer reusable and accelerated the high-level feature extraction of proposals. Developed from Fast R-CNN, scale-aware Fast R-CNN (SA-FastRCNN) [12] adopted the divide-and-conquer structure to solve the scale-variant problem.

A review of the approaches evaluated in the Caltech pedestrian dataset ([13], [14]) shows that all top performance methods (e.g. [8], [12]) are based on deep learning. CNN models can learn features directly from the pixels and two factors play a decisive role in its success: **1)** Feature representation proves to be sophisticated and more discriminative; **2)** CNN features possess a good generalization capability.

After this overview of existing methods, let us get back to the issue itself: The task of pedestrian detection. It can be defined as predicting bounding boxes for all pedestrians in images or video sequences, especially those taken from a monocular camera mounted on a vehicle [15]. The main challenges of this task are the intra-class variations of pedestrians in scale, background, lighting and occlusion.

1. In our paper, "models based on deep learning" has the same significance with "CNN models".

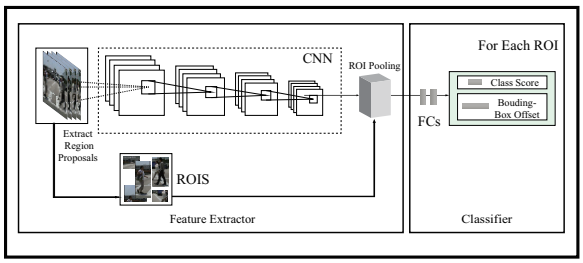


Figure 2. Framework of Fast R-CNN [11]. This structure can be divided into two parts: **feature extractor** and **classifier**.

The pedestrian scale[2], for instance, mainly lies in the distance between the observed pedestrian and the camera. Pedestrians who are close to the camera show more pixels than the ones far from the camera in image data. Features, extracted from the image data, of the large-scale and small-scale instances are obviously different when the range of pedestrian scale is large.

To solve the scale-variant problem, we propose a scale-discriminative network (SDN) that comprises two key components: Full-scale layer and scale-discriminative classifier layer. **Full-scale layer** combines high-level and low-level features of images to heighten the distinguishability of small-scale perdestrians. **Scale-discriminative classifier layer** contains several classifiers selectively activated according to the scale of the input proposals. The experimental results show that the SDN is robust in scale-variant pedestrian detection. Figure 1 shows an overview of the SDN.

Our contributions are multifold. **First**, we propose a scale-discriminative classifier layer that is sensitive to pedestrian scale, to improve performance in scale-variant pedestrian detection. **Second**, we introduce a full-scale layer that converges hierarchical features to enhance the capability of small-scale pedestrian detection. **Finally**, our method for multi-scale pedestrian detection demonstrates state-of-the-art performance in the Caltech pedestrian dataset ([13], [14]) and is faster than other close methods.

2. Related Work

Our work closely relates to three aspects: Fast R-CNN [11], fully convolutional network (FCN) [16] and implementation of multi-scale detection under the sliding windows paradigm.

2.1. Fast R-CNN

The main architecture for our method was developed from Fast R-CNN [11]. Fast R-CNN was derived from R-CNN [9] that has been a breakthrough in object detection. By employing a deep CNN model (e.g. VGG16 [17], AlexNet [18]), R-CNN has achieved a significant increase in PASCAL VOC 2012 [10]. However, R-CNN is a time-consuming model because it repeatedly applies the deep network to thousands of proposals per image. Fast R-CNN has enhanced the efficiency and accuracy of R-CNN by reusing the convolutional feature maps and calculating the bounding box regression offset for each proposal. Figure 2 shows the framework of Fast R-CNN.

The inputs of the net are a set of proposals and an entire image. Firstly, the net applies the convolutional layers to compute feature maps for the input image; Then, the ROI pooling layer maps each proposal from feature maps to a fixed-length feature vector; After that, each feature vector will be fed into a sequence of fully connected layers; Finally, two sibling output layers are output: One layer outputs the softmax probability scores and the other outputs the bounding box position offset.

2. The scale means the pixel scale which is shown in image or video sequences.

Figure 3. Various multi-scale handling approaches.

2.2. Fully Convolutional Network

When applying Fast R-CNN directly to pedestrian detection, it struggles with pedestrians in small-scale. We believe the issue has the following causes. First, the convolutional feature maps for detecting small-scale objects are at a low resolution. Second is that the ROI pooling layer implemented in low solution feature maps can produce blur feature maps. Those feature maps are not discriminative and therefore degrade the capability of the subsequent classifier. It is therefore necessary to find the best method to enhance the distinguishability of small-scale instances.

Recently, the excellent performance of the fully convolution network (FCN) in computer vision has received wide attention. In [16], it combined high-level information with low-level information for semantic segmentation and got a competitive result. HyperNet ([19]), which was derived from FCN, has achieved leading recall and state-of-the-art object detection accuracy on PASCAL VOC 2007 and 2012.

Inspired by FCN and HyperNet, we constructed a feature map layer with high resolution that combines high-level semantics features and low-level features of images. In this way, we can ensure that small-scale instances retain a suitable resolution feature maps before the ROI pooling process and become more distinguishable.

2.3. Multi-scale Handling Approaches

Most of the previous top performing pedestrian detection models used a sliding window at multi-scale over the input image. ICF [3] trained a single classifier for pedestrians and resized the image multiple times. Nevertheless, the feature computation at each scale was time-consuming. The fastest

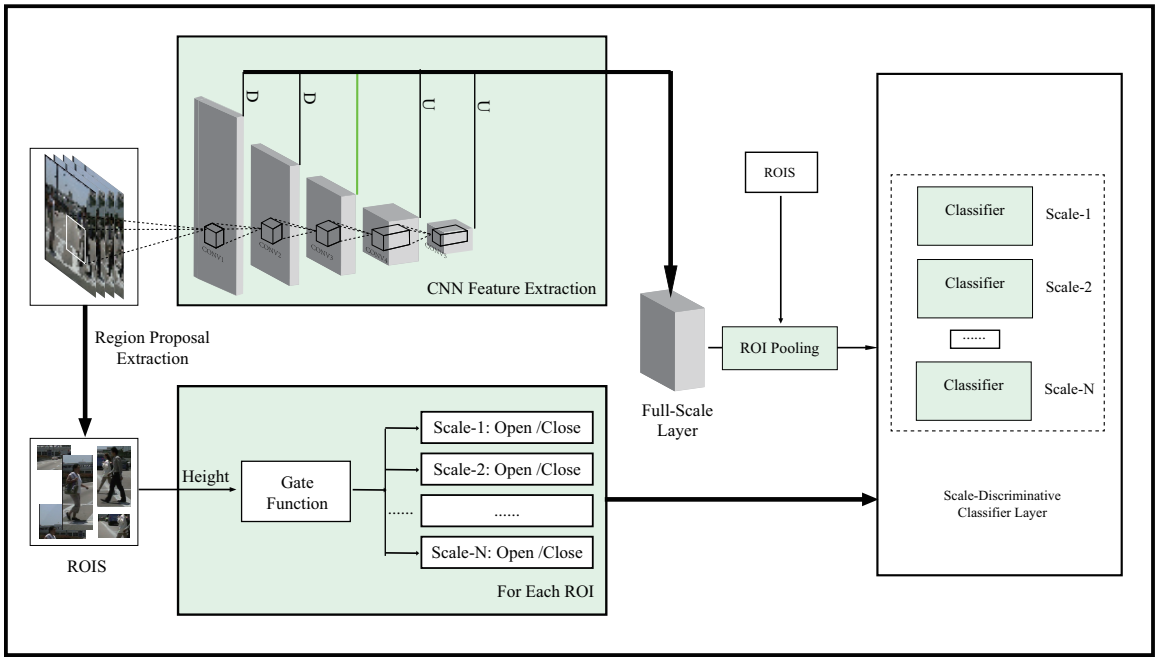


Figure 4. Basic architecture of the scale-discriminative network (SDN). The SDN can be divided into the following parts. **1) Region proposal extraction**: Generating perdestrian proposals; **2) CNN feature extraction**: Extracting feature representation by fine-tuning a pre-trained CNN model; **3) Gate function**: The role of this function is computing the scale-index according to the scale of proposal; **4) Full-scale layer**: A feature map layer that combines hierarchical features from different layers. The letter 'D' means downsampling and the letter 'U' means upsampling; **5) Scale-discriminative classifier layer**: Its function is to calculate detection results through the classifier as determined by the scale-index for each proposal.

pedestrian detector in the west (FPDW) [20] improved the speed of ICF without loss of accuracy. FPDW computed the features only for half scales and approximated the features on the intermediate scale. In addition, VeryFast [21] was proposed by using a single image scale and multiple classifiers on different scales. Figure 3 shows the different approaches to detect pedestrians at multi-scale.

Meanwhile, CNN models usually adopt two approaches to deal with the scale-variant problem. One is the "brute force" approach (e.g. [9], [11]), in which images are resized at a predefined scale. The other is the multi-scale learning approach (e.g. [11], [22]), which employs the image-pyramids as a way of data augmentation. However, both approaches have limitations. The "brute force" approach can not solve the issue effectively when the range of pedestrian scale reaches a certain threshold, and the multi-scale learning approach is time-consuming because it applies the deep network to images of different scales.

Inspired by [21], we constructed a scale-discriminative classifier layer containing multiple classifiers to make our model more robust in scale-variant pedestrian detection.

3. Proposed Scale-Discriminative Network

Figure 4 illustrates the architecture of the proposed method for pedestrian detection. At the beginning, an integral image is fed to convolutional layers to get the feature maps of each layer. And then, the hierarchical features are integrated from different layers into a unitary layer called the full-scale layer. Next, region proposals of image are generated, and we compute the scale-index[3] for each proposal. Finally, these proposals are classified and adjusted by the classifier based on the scale-index.

3.1. Network Structure

As shown in Figure 4, the scale-discriminative network (SDN) is composed of five concatenated individual components. **The region proposal extraction** extracts pedestrian proposals. **The CNN feature extraction** learns feature representation directly from the raw data by fine-tuning a pre-trained CNN model. We have compared several popular CNN models, such as AlexNet [18], VGG16 and VGG19 [17], and chose the VGG16 for our basic CNN model. The others are presented in detail in subsequent sections.

3.2. Full-Scale Layer

We find that Fast R-CNN [11] struggles with pedestrians in a small-scale. In order to solve this issue, we integrated hierarchical features at different depths into a unitary layer to enhance the distinguishability of small-scale objects. Figure 5 shows the construction process of this unitary layer, which is similar to the production of [19]. The new unitary layer is called the full-scale layer.

The process is divided into four steps: **1) Feature maps extraction**: Applying the convolutional layers to compute

3. Scale-index means the corresponding classifier will be open or closed.

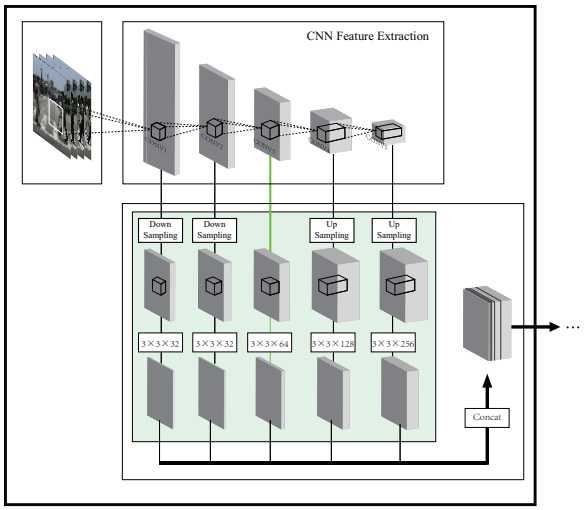


Figure 5. Construction process of the full-scale layer. We divide the process into four steps: **feature maps extraction**, **resolution normalization**, **balance adjustment** and **concatenation**. The green line means that this resolution meets the requirements.

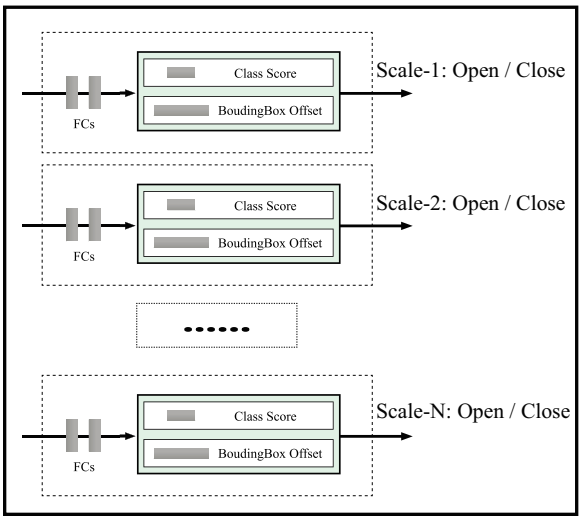


Figure 6. Structure of the scale-discriminative classifier layer. The SDC layer contains N classifiers, each of which is in respect to a fixed scale.

feature maps for the entire image; **2) Resolution normalization**: We adopt different sampling strategies specific to different layers, downsampling for low-level layers and upsampling for high-level layers. This normalizes the resolution of feature maps; **3) Balance adjustment**: To achieve balance between the high-level semantics features and the low-level features, we utilize a convolutional (conv) layer after sampling processing; **4) Concatenation**: We concatenate the processed feature maps into our full-scale layer.

3.3. Scale-Discriminative Classifier Layer

Generally, many object detection approaches can be taken as the combination of the feature extractor and classifier. We divide Fast R-CNN [11] into two parts as shown in Figure 2. The classifier of Fast R-CNN contains a sequence of fully connected layers, and ends up with two sibling output layers: One layer outputs the softmax probability scores and the other outputs the bounding box position offset.

Figure 6 illustrates the structure of our scale-discriminative classifier (SDC) layer which is constructed based on the Fast R-CNN classifier. The SDC layer contains N classifiers and each classifier respondeds to a fixed scale. We calculate the scale-index first before employing the index classifier to get the results of each proposal. Only one classifier is activated for each proposal. In this way, we have improved the performance of scale-variant pedestrian detection.

3.4. Gate Function

The SDC layer contains many classifiers, each of which is trained for a fixed scale. The gate function is defined as selecting a suitable classifier for each proposal. The key factor is scale.

The scale of proposals can be measured by either pixel height or pixel width. However, for pedestrian detection, while a pedestrian is standing in front of the camera, the pixel height of his bounding box may correlate more to the scale of the pedestrian than the pixel width [14]. We only trained three classifiers due to the GPU limit.

$Classifier\_L$ will be selected when pixel height is more than 80 pixels, $Classifier\_M$ will be chosen when pixel height is between 50 pixels and 80 pixels, and $Classifier\_S$ will be activated when pixel height is less than 50 pixels.

H denotes the pixel height of proposals, $Scale_L$, $Scale_M$ and $Scale_S$ respectively denote the scale-index of $Classifier\_L$, $Classifier\_M$ and $Classifier\_S$. The gate function is as follows:

$$Scale_L = \varepsilon(H - 80) \tag{1}$$

$$Scale_M = \varepsilon(H - 50) \times \varepsilon(80 - H) \tag{2}$$

$$Scale_S = \varepsilon(50 - H) \tag{3}$$

where:

$$\varepsilon(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{4}$$

3.5. Multi-Task Loss

Our SDN has two sibling output layers, one layer outputs the softmax probability scores $s = (s_0, s_1)$, and the other layer outputs the bounding box position offset of pedestrians. The offset of the pedestrian can be denoted as $t = (t_x, t_y, t_w, t_h)$.

TABLE 1. THE CONSTRUCTION DETAILS OF FULL-SCALE LAYER. IN THE FIRST ROW, THE EXTRACTED FEATURE MAPS ARE DENOTED AS "CONV(START_INDEX - END_INDEX)"; IN THE SECOND ROW, THE SAMPLING STRATEGY IS DENOTED AS "DOWN/UP_ < RECEPTIVE FIELD SIZE > OR HOLD"; IN THE THIRD ROW, THE CONVOLUTIONAL PARAMETERS ARE DENOTED AS "CON_ < RECEPTIVE FIELD SIZE > _ < NUMBER OF CHANNELS >"; IN THE FOURTH ROW, WE CONCATENATE THE PROCESSED FEATURE MAPS INTO THE FULL-SCALE LAYER.

Feature maps extraction	conv(1-2)	conv(3-4)	conv(5-7)	conv(8-10)	conv(11-13)
Resolution normalization	down_4×4	down_2×2	hold	up_2×2	up_4×4
Balance adjustment	CON_3×3_32	CON_3×3_32	CON_3×3_64	CON_3×3_128	CON_3×3_256
Concatenation	Full-scale layer				

TABLE 2. LEARNING RATE IN DIFFERENT STEPS. IN THE THIRD ROW, THE CONVOLUTIONAL LAYER PARAMETERS ARE DENOTED AS "CONV(START_INDEX - END_INDEX) OR NONE".

Step	**Step1**	**Step2**	**Step3**	**Step4**	**Step5**
Basic	0.001	0.00005	0.00001	0.00001	0.00001
Zero	none	conv(1-4)	conv(1-13)	conv(1-13)	conv(1-13)

TABLE 3. COMPARISION OF THE MISS RATE IN THE DISTANT-SET AND TESTING TIME WITH OTHER CLOSE METHODS, INCLUDING SA-FASTRCNN [12] AND COMPACT-DEEP [23].

Method	**SA-FastRCNN**	**CompACT-Deep**	**Ours**
Miss rate (%)	86%	86%	83%
Test time (S/im)	0.37	1	0.1

Each training proposal is labeled by ground truth class k and the bounding box regression target T. We minimize a multi-task loss function:

$$L(s, k, t, T) = L_{cls}(s, k) + L_{loc}(t, T) \tag{5}$$

Where L_{cls} is log loss for classification and L_{loc} is smooth L_1 [11] loss for the bounding box regression.

3.6. Joint Training

In practice, a multi-step training process jointly optimizes the SDN. The training process is as follows:

- **Step 1**: Adopting the VGG16 [17] as a basic CNN model to initially train a deep CNN model as the initializing layer in **step2**.
- **Step 2**: Training a SDN model 0 that only contains one classifier, for initializing main layers in **step 3**, **4** and **5**.
- **Step 3**: Training a SDN model 1 that contains $Classifier\_L$, by using proposals in large scale (more than 80 pixels).
- **Step 4**: Training a SDN model 2 that contains $Classifier\_M$, by using proposals in middle scale(between 50 pixels and 80 pixels).
- **Step 5**: Training a SDN model 3 that contains $Classifier\_S$, by using proposals in small scale(less than 50 pixels).
- **Step 6**: Concatenating the main layer of **step 2** and classifier of **step 3**, **4** and **5** into a unitary SDN model. This SDN model is the final model.

The learning rate varies with steps and layers. Table 2 shows the changes of learning rate in different steps. The second row shows the basic learning rate in each step, and the third row indicates the layer to be fixed to 0 in each step during training.

4. Experiments

The proposed scale-discriminative network (SDN) was evaluated in the Caltech pedestrian dataset ([13], [14]). In order to evaluate the performance of detectors, we plotted miss rate against false positives per image (FPPI) by changing the threshold on probability score (using log-log plots). We used the log-average miss rate to sum detector capability, computed by averaging FPPI in log-space in the range 10^{-2} to 10^{0}, as proposed in [14].

4.1. Datasets

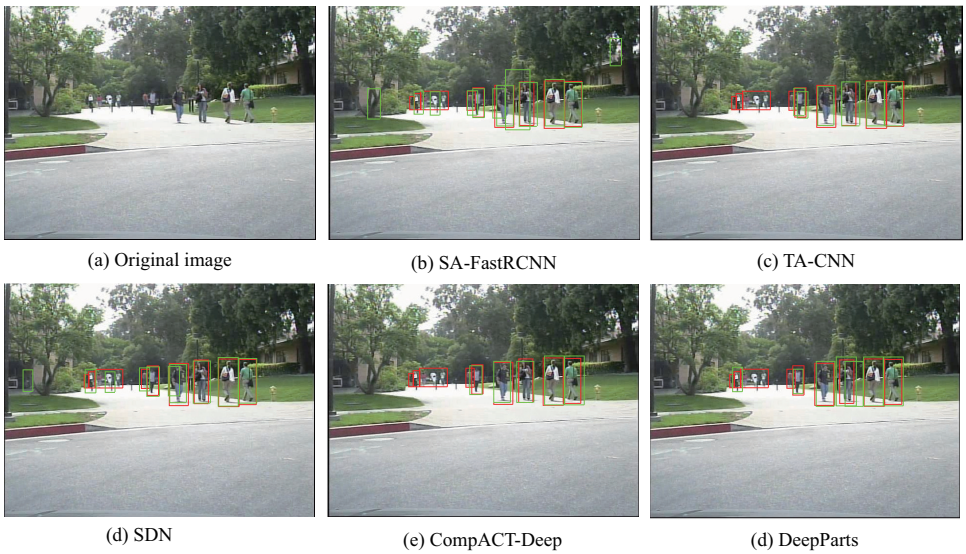
The Caltech pedestrian dataset is currently the most popular and largest for pedestrian detection. It consists of about 10 hours of 640×480, 30 Hz video taken from a moving vehicle in regular traffic. About 250,000 frames with 350,000 bounding boxes and 2,300 specific pedestrians are labeled. It also contains detailed occlusion labels, enabling researchers to analyze the performance of detectors at different occlusion levels.

The first six sets are defined as a training set, and the remaining five sets are defined as testing data. In order to evaluate our method practicability in a wide range of scales, we choose the following evaluation settings:

- All: Performance evaluated on pedestrians whose pixel height is over 20 pixels tall and showing at least 20 percent of visible body parts.
- Distant: Performance evaluated on pedestrians whose pixel height was between 20 and 50 and showing at least 20 percent of visible body parts.
- Close: Performance evaluated on pedestrians whose pixel height is over 50 pixels tall and showing at least 20 percent of visible body parts.

4.2. Implementation Details

To save the computational cost and make a fair comparison with state-of-the-art methods, we utilized the ACF



(a) Original image (b) SA-FastRCNN (c) TA-CNN

(d) SDN (e) CompACT-Deep (d) DeepParts

Figure 7. Comparison of pedestrian detection results with other state-of-the-art Methods. The red bounding box denotes the ground truth and the green bounding box denotes the detection result.

[4] detector to generate proposals. We used the VGG16 [17] model as the pre-trained model to initialize the main CNN of the SDN. Table 1 shows the construction details of full-scale layer.

Our method involves the weight decay of 0.0005 and Stochastic Gradient Descent (SGD) with a momentum of 0.9. Each mini-batch consists of 64 pedestrian proposals in one image. We maintained the ratio of positive proposals to negative proposals at one to three. Negative object proposals have intersection over union (IOU) with the ground truth box less than 0.1. Positive ones have an IOU with a ground truth bounding box larger than 0.5. During training and testing, the 640×480 scale of input images remained unchanged.

The SDN is trained and tested on the publicly available Caffe platform [24]. We adopt a multi-step training process (as presented in the previous section) to jointly optimize the SDN.

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

Figs. 7, 8, 9, 10 show the overall experimental results. We compared the results with those top performance methods, including TA-CNN [7], DeepParts [8], CompAct-Deep [23] and SA-FastRCNN [12].

The proposed method achieves top performance in the all-set with a miss rate of 63% and gets a competitive result in the close-set with a miss rate of 27%. Beyond that, our method has achieved a new state-of-the-art performance in the distant-set with a miss rate of 83%. This is significantly superior (by 3%) compared to the previous state-of-the-art approach SA-FastRCNN [12].

We also compared SDN with other close deep models of detection speed. As shown in table 3, the proposed method is significantly faster than the close methods ([12], [23]).

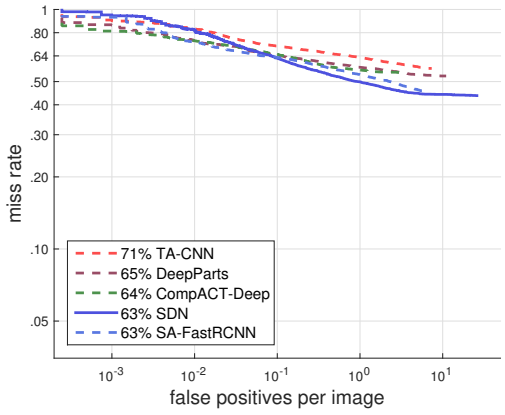


Figure 8. Results in the all Set.

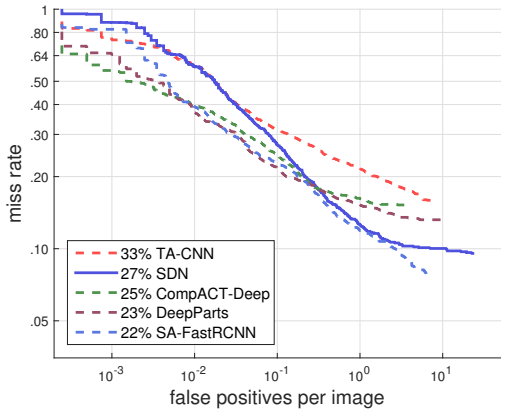


Figure 9. Results in the close Set.

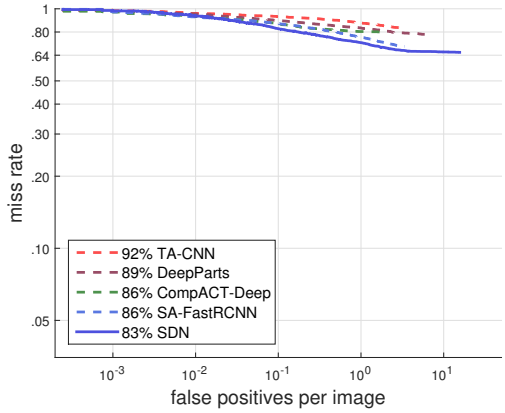


Figure 10. Results in the distant Set.

5. Conclusion

In this paper, we proposed a novel scale-discriminative network that is robust in scale-variant pedestrian detection. The proposed method improves pedestrian detection performance in two aspects:

- The full-scale layer is the convergence of high-level semantic features and low-level features of CNN. In this way we can ensure that small-scale objects retain a suitable resolution feature maps before ROI pooling process. The distinguishability of small-scale pedestrians has been enhanced.
- The scale-discriminative classifier layer contains N classifiers while each classifier is respect to a fixed scale, which realizes the robustness of multi-scale pedestrians.

In the future, we will take into account the structural information of pedestrian physiques to solve the occlusion situation in pedestrian detection.

Acknowledgments

This work was supported by Shenzhen STP (JCYJ20150331151358150).

References

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[3] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," 2009.

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

[5] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 947–954.

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

[7] Y. Tian, P. Luo, X. Wang, and X. 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.

[8] ——, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1904–1912.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[10] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

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

[12] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *arXiv preprint arXiv:1510.08160*, 2015.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 304–311.

[14] P. Dollar, C. Wojek, B. Schiele, and P. 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.

[15] S. Zhang, "Efficient pedestrian detection in urban traffic scenes," Ph.D. dissertation, Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Diss., 2015, 2015.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

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

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," *arXiv preprint arXiv:1604.00600*, 2016.

[20] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west." in *BMVC*, vol. 2, no. 3. Citeseer, 2010, p. 7.

[21] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2903–2910.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

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

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.