# Pedestrian Detection by Using CNN Features with Skip Connection

Peng Zhang[1,2] and Zengfu Wang[1,2](✉)

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[2] University of Science and Technology of China, Hefei 230026, China
hizhangp@mail.ustc.edu.cn, zfwang@ustc.edu.cn

**Abstract.** The CNN based pedestrian detection is developing rapidly in recent years. Compared to the features used in former pedestrian detection models, the features from deep CNN have outperformed in many aspects. In this paper, we focus on the problem of pedestrian detection by using CNN features with skip connections and mainly address the corresponding issues in urban scenes: different spatial scales and optical blur of pedestrian. We propose an effective end-to-end pedestrian detector, which fuses different features from multi-layers to recover the coarse features of small scale pedestrians and optical blur pedestrians. The experimental results show that our method achieves state-of-the-art performance on Caltech Pedestrian Detection Benchmark.

**Keywords:** Pedestrian detection · Skip connection · Normalization

## 1 Introduction

Recently, pedestrian detection which aims to locate pedestrian instances in an image has drawn much attention beyond general object detection, because of its attractive applications in automatic driving, video surveillance, person re-identification and robotics.

Extensive research on pedestrian detection have been put forward [18,30], these papers achieved state-of-the-art performance on well-established benchmark dataset, such as Caltech Pedestrian Detection Benchmark [7] and KITTI Vision Benchmark [8]. However, with the decelerated improvement on detection quality (miss rate), it seems that we are reaching the bottleneck of pedestrian detection.

Currently, there remains two main challenging points of pedestrian detection in urban scenes: (1) different spatial scales and (2) optical blur. The scale of pedestrian instance in Caltech dataset ranges from 16 pixels to about 128 pixels, and nearly 70% pedestrian instances lies in between 30 to 80 pixels [7]. Compared to large-scale pedestrian instances, small-scale ones suffer from blurred boundaries and breezing appearance as a result of low resolution, which makes it difficult to distinguish them from background and similar objects (such as

J. Yang et al. (Eds.): CCCV 2017, Part III, CCIS 773, pp. 73–83, 2017.
https://doi.org/10.1007/978-981-10-7305-2_7

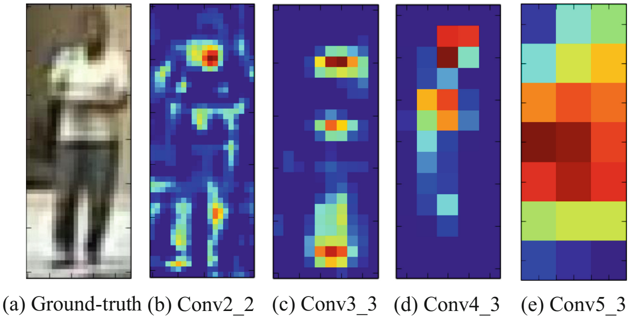

**Fig. 1.** The spatial feature distribution of pedestrian instance in each convolutional layer. For a typical ground-truth input, fine, detailed feature evolves to coarse, abstract feature from low layer to high layer.

street lamp and trash can). Meanwhile, observing the feature maps from the same convolutional layer, pedestrian instances with large-scale can provide more detailed information than small-scale ones. These difference will make the model less sensitive to small-scale pedestrian instances. Optical blur mainly results from the movement of the camera in the car. Fast relative displacement between camera and pedestrian instances will fuzzy the optical features, only coarse feature maps are obtained compared to other ones with similar scale.

Faster R-CNN [22] is a quite successful method in the field of general object detection, and has been the base of many recent popular pedestrian detection methods [5,18,27,30]. Instead of using hand-craft features, such as HOG+SVM rigid and deformable part detector (DPM), Faster R-CNN is based on deep CNN. Faster R-CNN consists of two parts: a fully convolutional region proposal network (RPN) and followed by an R-CNN [10] classifier. Convolutional neural network feature generated by RPN shows strong robustness to pedestrian instances in pedestrian detection models. However, native Faster R-CNN architecture cannot handle small-scale pedestrian instances and optical blur problems. Following the architecture of Faster R-CNN, we propose a skip connection feature model for pedestrian detection to overcome above two challenges. As shown in Fig. 1, lower convolutional layers provide discriminative features to capture intra edge variations, higher layers encode semantic concepts for object category classification. With appropriate combination of feature maps from these convolutional layers, synthetic feature maps can provide more expressive and distinguishable features for small-scale or blurry pedestrian instances. Using region proposals generated by RPN from conv5_3 layer in VGG16 [24], we extract specific convolutional feature maps in conv2_2, conv3_3, conv4_3 and conv5_3 layer for classifier. Unlike prior works using multi-stage pipeline for detection, we construct an end-to-end pedestrian detection model for consistent learning. Our model achieves 8.91% miss rate on Caltech Pedestrian Detection Benchmark, at the speed of 0.24 s/image with Titan X (Maxwell).

Our method has three contributions: (1) We introduce skip connection feature for pedestrian detection which combines different feature maps from different convolutional layers. (2) We prove that batch normalization is more suitable for skip connection feature combination. (3) We show that an end-to-end detection system is much easier to train compared to multi-stage system.

## 2 Related Work

### 2.1 Pedestrian Detection

Considering the prior works on pedestrian detection, we can divide them into three families [3]: DPM variants (MT-DPM [28]), Deep Neural Network (CompACT-Deep [5], SAF R-CNN [18]) and Decision forests (ChnFtrs [6], Katamari [3]).

Feature extraction is a very important step for pedestrian detection. Handcraft features such as Integral Channel Feature (ICF) [6] and Aggregated Channel Feature (ACF) [29] are extracted from the combination of three types of 6 channels (LUV color channels, normalized gradient magnitude, and histogram of oriented gradients), and then used to generate features of different levels [6], local sum of squares [2], or decorated LDCF (Linear Discriminant Channel Feature) [21]. These are very polular feature extraction methods before region-based convolutional neural network (R-CNN) achieved the top performance for general object detection based on AlexNet model. Deep convolutional neural network (DCNN) has been proved to be a powerful feature extractor on a wide variety of computer vision, SCF+AlexNet [13], DeepParts [25] and many other tasks [5,18,27,31] have significantly out-performed hand-crafted feature models using deep convolutional feature.

The most recent surveys [5,30] indicate that Convolutional Neural Network (CNN) performs a good feature extractor, but downstream classifier (the *fc* layers) degrades the results due to the low resolution of feature maps and lack of bootstrapping strategy. To overcome above drawbacks they learn boosted classifiers on top of hybrid feature maps which extracted from several layers, these improvements build an expressive feature map and powerful classifier, which help to achieve top performance. For example, RPN+BF [30] uses RPN in Faster R-CNN as a class-agnostic detector to generate region proposals, then adopt RoI Pooling to extract fixed-length features from regions, finally, these features are used to train a cascaded boosted forest classifier. This multi-stage strategy divides the detection process into several modules, making it more complex, time-consuming and hard to train, while our method is much simpler and easier for consistent learning.

### 2.2 Skip-Layer Connections

Skip-layer connections come from Fully Convolution Network (FCN) [20], which has an excellent performance on semantic segmentation. In this paper, information in the coarse, high layer is combined with that in the fine, low layer to generate a impressive feature map for semantic segmentation.

For a pre-trained CNN model, CNN features at different layers have different properties for an object. Low convolutional layer provides detailed local edge features which is helpful to distinguish targets from the background and similar object, but they are not robust to the change of appearance, such as deformation and blur. High convolutional layer captures more abstract and semantic information, the are more robust to different appearance and easy to classify each object into different classes, however, they suffer from the low resolution of feature maps, restricting their spatial distinguishing ability.

Recent HyperNet [17] extracts feature maps from several layers, applies max pooling or deconvolution to each feature maps to obtain a single output cube which is call Hyper Feature and leads to a more accurate object detection result.

## 3 Skip Connections Feature

In this section, we introduce our skip connection features (illustrated in Fig. 2), which are extracted from different convolutional layers of the neural network model. With the proper fusion of these normalized features, the model is able to achieve the state-of-the-art result on Caltech Pedestrian Detection Benchmark.

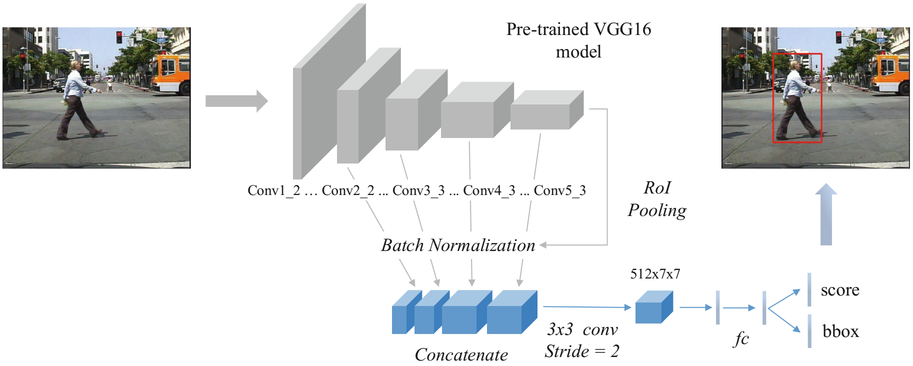

**Fig. 2.** Skip connections pedestrian detection architecture. Our model takes in an input image, uses pretrained VGG16 model to extract feature maps and evaluates 128 RoI generated by RPN, then RoI Pooling is applied to several layers to extract multiple level's abstraction, after batch-normalizing each layer's feature map, we apply a convolution layer to the concatenated feature maps, classification and adjustment are predicted for each proposal.

### 3.1 Feature Extraction

Exploring recent successful object detectors such as SPPnet, Faster R-CNN [22] and R-FCN [16], they all extract feature maps from the last convolution layer (conv5_3 in VGG16 [24] and res5c in ResNet-101 [12]), then feed feature maps into the *fc* layers for classification scores and adjusted bounding boxes.

Given an image, we keep the image's aspect and resize the short side into 800 pixels. As a result of 4 pooling layers in VGG16 model, size of feature maps is 50 pixels. And for a typical pedestrian instance with a median height of 48 pixels in the image [7], size of its feature maps is only about 6 pixels (after the input image resized into 800 pixels). RPN, which is based on Fixed-size sliding windows, can avoid collapsing bins [30]. However, RoI pooling [9] may suffer from the low resolution and insufficient information.

Inspired by the design of HyperNet [17] and Inside-Outside Net [1], using region proposals generated by RPN, we adopt RoI pooling to several layers (conv2_2, conv3_3, conv4_3 and conv5_3) to extract feature maps with different resolution. Thanks to the design of RoI pooling, original *fc* layers will have no limitations to the size of feature maps. For example, we can extract feature maps from RoIs on conv2_2 (of a stride = 2), then pool the feature maps into a fixed-size of $C \times 13 \times 13$, where $C$ depends on the number of channels in each layer.

We downsample or upsample feature maps from each feature maps to keep a $C \times 13 \times 13$ fixed-size before fed into the *fc* layers, it is also needed for feature fusion in the next phase. It's worth noting that we leave the number of feature maps' channels $C$ unchanged, because accuracy of classification is much more important than that of location in pedestrian detection. We keep feature maps in higher layers with more channels and less channels in lower layers, aiming to allow higher layers weigh bigger influence than lower ones.

### 3.2 Layer Fusion

Feature maps extracted from different layers have coarse-to-fine information across CNN model, through a careful fusion method, we can provide more expressive spatial information for the *fc* layers.

Since feature maps after RoI Pooling are down sampled or up sampled to the same size, a straightforward idea would be to concatenate these feature maps together and reduce the dimensionality using convolutional layer. However, feature maps at different layers may have very different amplitudes, feature map with high amplitude can suppress one with low amplitude and result in unstable learning. To combine coarse-to-fine information efficiently, normalizing the amplitude of different feature maps is needed such that each layer has similar distribution.

Instead of using L2 normalization in [1], we apply Batch Normalization [14] to each feature map. Batch Normalization makes training easier and brings in several advantages:

1. Faster training: Learning rate can be accelerated by Batch Normalization compared to other pedestrian detection models, and leads to faster convergence.
2. Amplitude Normalization: Batch Normalization layer can efficiently normalize data in feature maps (zeros means, unit variances, and decorrelated), and take care of backward propagation. On the contrast, feature normalization needs to be carefully addressed when applied to the *fc* layers using L2 normalization [19].

3. More accurate: In a batch-normalized model, each feature map has similar amplitude, we can apply unified learning rate to each layer and lead to a better learning, hence accuracy of the model.

In order to gain high resolution of feature maps for R-CNN classifier, we set the size of RoI Pooling kernel to 13 × 13. For the feature maps whose size less than 13 × 13, we add a deconvolution operation to perform upsampling and a convolutional operation is applied to the other. Then we concatenate them together to create a skip-connections-feature cube.

To match the *fc* layers in VGG16, a convolution layer with 3 × 3 kernel and stride of 2 is applied at the top. This operation not only reduces 13 × 13 feature maps into 7 × 7, but also selects class independent and location independent information from skip-connections-feature cube.

## 4 Experiments

We train and evaluate our model on the popular Caltech Pedestrian Dataset [7], and compare our results with other algorithms.

### 4.1 Datasets

Caltech Pedestrian Dataset and its detection benchmark [7] is one of the most popular pedestrian detection dataset. The dataset contains about 10 h of 640 × 480 30 Hz video of urban traffic collected from a vehicle camera. There are about 250,000 frames, which consist of 350,000 bounding boxes and 2300 unique annotated pedestrian instances.

To avoid the weakness of lacking training data, we prepare our training images one out of each 3 frames in original training data (set00–set05), instead of one out of each 30 frames for standard training set and test set. Then training images are augmented by horizontal flip with probability 0.5 and images with empty bounding boxes or occluded pedestrian are removed.

The standard testing data (set06–set10) are evaluated under the “reasonable” setting (on 50-pixel or taller, unoccluded or partially occluded pedestrians). We compare our performance using log-average miss rate (MR) which is calculated by averaging each miss rate on False Positive Per Image (FPPI) in $[10^{-2}, 10^{0}]$.

### 4.2 Implementation Details

We use the VGG16 model weights pre-trained on ImageNet [23] dataset to extract feature maps for detection. Since we need to extract feature maps from conv2_2 layer to conv5_3 layer, it is necessary to fine-tune starting from conv2_1 layer and keep previous layers frozen.

Following the setting of Faster R-CNN [22], the pool5 layer is removed and feature maps extracted by conv5_3 layer are feed into RPN directly to generate region proposals. As mentioned above, to get more spacial information for RCNN layer, we set the size of RoI Pooling kernel to 13 × 13, and a 3 × 3 kernel convolutional layer with the stride of 2 is followed to match the size of the *fc* layers.

One of the key points in improvement of pedestrian detection is enlarging the resolution, so an image is resized such that its short side has 800 pixels before feed into neural network. For RPN training, we set an anchor to be positive if it has an Intersection-over-Union (IoU) ratio greater than 0.5 with one of the ground-truth bounding box, we also adjust the anchor according to the dataset [7], the aspect ratios of anchor are set to 1.0, 2.0 and 3.0, and the scales are starting from 20 pixels height with a scaling stride of 2x to 160 pixels. The usage of a wide range of scales specific to pedestrian can avoid many false positive instances and make it easy for detecting pedestrian of different size.

Our model is implemented on the public available platform caffe [15]. We finetune our model with a weight decay of 0.0005 and a momentum of 0.9 and use single-scale training by default. We keep the first two layers frozen and update the parameters of other layers with a learning rate of 0.001 for 30 k iterations and 0.0001 for next 30 k iterations on Caltech Pedestrian Dataset.

### 4.3 Result Evaluation

To investigate the influence of our model on pedestrian detection, we conduct several groups of experiments on various factors and architectures.

**Table 1.** Comparison of detection performance on different layers' feature maps on the Caltech Reasonable test set. Note that all region proposals are generated from conv5_3 layer.

| Convolutional layer | Channels | Ratio | Miss rate (%) |
|---|---|---|---|
| Conv3_3 | 256 | 4 | 13.84 |
| Conv4_3 | 512 | 8 | **11.63** |
| Conv5_1 | 512 | 16 | 13.16 |
| Conv5_2 | 512 | 16 | 12.07 |
| Conv5_3 | 512 | 16 | 11.65 |

**Different Convolutional Layers.** Fusion of multi-layers' feature maps has been proven beneficial in detection [17,30] and other computer vision applications [11]. In our model, we combine several feature maps from selected convolution layers. To verify the various performance of feature maps in different convolution layers, we adopt RoI Pooling on different convolutional layers and generate region proposals from the same convolutional layers (conv5_3) for fair comparisons. Table 1 shows the results of using different RoI pooling feature maps in our method. Conv4_3 and conv5_3 perform good results (11.63% and 11.65%), indicating that high resolution (conv4_3) and semantic feature (conv5_3) are the main effect factors of the final performance. The decline trend after an initial ascent from conv4_3 to conv5_3 means that intermediate convolutional layers may suffer from low resolution and insufficient semantic features (Fig. 3).

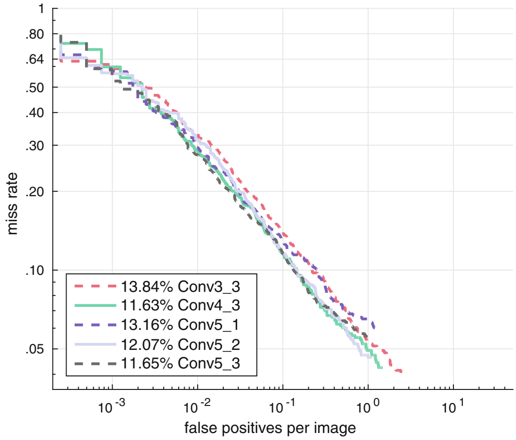

**Fig. 3.** Comparison on the Caltech Reasonable test set, the performance of each layer varies from $10^{-2}$ to $10^{0}$ on False Positive Per Image (FPPI).

**Normalization.** Normalizing the amplitude of different feature maps is an effective way to combine coarse-to-fine information from different layers. To analysis the performance of this strategy, we compare the results under different normalization method in Table 2. It can be observed that unnormalized concatenation with each feature map leads to unstable learning and not very good result (13.68%), LRN normalization decreases the miss rate by 3.68% and Batch Normalization gains the state-of-the-art performance (8.91%), showing that normalization are benificial for stable learning in skip-layer feature fusion and batch normalization is the key to get better performance.

**Comparison with state-of-the-art.** Figure 4 shows the result of our model trained on the Caltech training set and evaluated on the Caltech Reasonable test set, which achieves a miss rate of 8.91%. We compare our result with several former state-of-the-art models on Caltech Pedestrian Benchmark, including RPN+BF [30], SAF R-CNN [18], CompACT-Deep [5], DeepParts [25], Checkerboards+ [32], MS-RCNN [4], and TA-CNN [26]. It can be observed that our model outperforms other models and achieves state-of-the-art performance.

We also compare our model with Faster R-CNN [22], CompACT-Deep [5] and RPN+BF [30] in Table 3. With the design of RPN, Faster R-CNN is capable of

**Table 2.** Comparison of detection performance on unnormalized, LRN and BatchNorm concatenation on the Caltech Reasonable test set.

| Convolutional layer | Normalization | Miss rate (%) |
|---|---|---|
| Conv2_2, ..., Conv5_3 | None | 13.68 |
| Conv2_2, ..., Conv5_3 | LRN | 10.00 |
| Conv2_2, ..., Conv5_3 | BatchNorm | **8.91** |

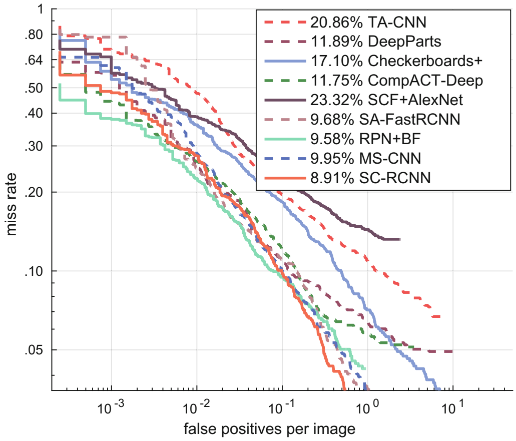

**Fig. 4.** Log-average miss rate on the Caltech Reasonable test set.

**Table 3.** Comparison of detection performance of our method with several prior state-of-the-art methods on the Caltech Reasonable test set, including test speed and miss rate.

| Method | Test rate (s/image) | Miss rate (%) |
|---|---|---|
| Faster R-CNN [22] | **0.23** | 12.65 |
| CompACT-Deep [5] | 0.37 | 9.68 |
| RPN+BF [30] | 0.50 | 9.60 |
| **Proposed** | 0.24 | **8.91** |

handling different size of pedestrian instances, but the feature maps generated by RoI Pooling is not expressive enough for efficient discrimination. Our model and RPN+BF both try to fuse feature maps from different convolutional layers, unlike RPN+BF, they trained a cascaded boosted forest as classifier, we directly feed these normalized feature map into the *fc* layers for classification and regression. The result shows that training an end-to-end system leads to smaller miss rate with faster test speed (0.24 s/image in our model compared to 0.50 s/image in RPN+BF).

## 5 Conclusion

In this paper, we propose an effective end-to-end pedestrian detector, which combines feature maps from different convolutional layers generated by RPN in Faster R-CNN. By fusing coarse features from high layers with fine features from low layers, we produce a more expressive feature map for the *fc* layers classifier. Our model is capable of handling pedestrian instances with different spatial scales and optical blur. Evaluation result in Caltech Pedestrian Dataset shows that our model achieves state-of-the-art performance on its benchmark and also has fastest test rate.

## References

1. Bell, S., Zitnick, C.L., Bala, K., Girshick, R.: Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks. arXiv preprint arXiv:1512.04143 (2015)
2. Benenson, R., Mathias, M., Tuytelaars, T., Van Gool, L.: Seeking the strongest rigid detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3666–3673 (2013)
3. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 613–627. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_47
4. 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
5. Cai, Z., Saberian, M., Vasconcelos, N.: Learning complexity-aware cascades for deep pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3361–3369 (2015)
6. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features (2009)
7. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361. IEEE (2012)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 346–361. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_23
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Hosang, J., Omran, M., Benenson, R., Schiele, B.: Taking a deeper look at pedestrians. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4073–4082 (2015)
14. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
16. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. arXiv preprint arXiv:1605.06409 (2016)

17. Kong, T., Yao, A., Chen, Y., Sun, F.: HyperNet: towards accurate region proposal generation and joint object detection. arXiv preprint arXiv:1604.00600 (2016)
18. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. arXiv preprint arXiv:1510.08160 (2015)
19. Liu, W., Rabinovich, A., Berg, A.C.: ParseNet: looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
21. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved pedestrian detection. In: Advances in Neural Information Processing Systems, pp. 424–432 (2014)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
25. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1904–1912 (2015)
26. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5079–5087 (2015)
27. Xiang, Y., Choi, W., Lin, Y., Savarese, S.: Subcategory-aware convolutional neural networks for object proposals and detection. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 924–933. IEEE (2017)
28. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3033–3040 (2013)
29. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Aggregate channel features for multi-view face detection. In: 2014 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2014)
30. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? arXiv preprint arXiv:1607.07032 (2016)
31. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? arXiv preprint arXiv:1602.01237 (2016)
32. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1751–1760. IEEE (2015)