Single-Stage Detector with Semantic Attention for Occluded Pedestrian Detection

Fang Wen[1], Zehang Lin[2], Zhenguo Yang[2,3(✉)], and Wenyin Liu[2(✉)]

[1] Department of Automation, Guangdong University of Technology, Guangzhou, China
l201070300 56w@gmail.com
[2] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
gdutlin@outlook.com, liuwy@gdut.cn
[3] Department of Computer Science, City University of Hong Kong, Hong Kong, China
zhengyang5-c@my.cityu.edu.hk

Abstract. In this paper, we propose a pedestrian detection method with semantic attention based on the single-stage detector architecture (i.e., RetinaNet) for occluded pedestrian detection, denoted as PDSA. PDSA contains a semantic segmentation component and a detector component. Specifically, the first component uses visible bounding boxes for semantic segmentation, aiming to obtain an attention map for pedestrians and the inter-class (non-pedestrian) occlusion. The second component utilizes the single-stage detector to locate the pedestrian from the features obtained previously. The single-stage detector adopts over-sampling of possible object locations, which is faster than two-stage detectors that train classifier to identify candidate object locations. In particular, we introduce the repulsion loss to deal with the intra-class occlusion. Extensive experiments conducted on the public CityPersons dataset demonstrate the effectiveness of PDSA for occluded pedestrian detection, which outperforms the state-of-the-art approaches.

Keywords: Occluded pedestrian detection · Single-stage detector
Repulsion loss · Semantic segmentation network

1 Introduction

Pedestrian detection is a significant research topic in object detection, which benefits many applications, e.g., driverless cars, intelligent robotics and intelligent transportation. It is quite common to utilize the methods proposed in object detection [1–3] to detect pedestrians directly. However, these methods can hardly obtain the optimal performance. The main reason is that pedestrians always gather together and are easily obscured by other objects in reality. Therefore, it is challenging and meaningful to deal with occlusion problems in pedestrian detection.

Quite a few researchers focus on the inter-class occlusion, i.e., pedestrians are occluded by non-pedestrian objects, e.g., buildings, trees and cars. It is difficult to locate the pedestrians based on parts of the bodies since there are rich categories of

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11296, pp. 414–425, 2019.
https://doi.org/10.1007/978-3-030-05716-9_34

obstruction, e.g., right-left and bottom-up occlusions. Intuitively, it is easy for detectors to learn features from the exposed parts compared with the heavily occluded pedestrians. In previous work, constructing pedestrian templates is the mainstream for pedestrian detection, which divides a pedestrian into different parts as templates, and then utilizes these templates to train different classifiers for various occlusions. However, it suffers from high computational cost. Recently, Zhang et al. [4] apply attention mechanism to handle different occlusion patterns, which achieves the state-of-the-art performance on heavy occlusion. However, their method only works on the two-stage models, i.e., Faster-RCNN [3], which consists of proposing regions and computing the confidences of object classes.

Recently, the advanced models are based on the single-stage models, e.g., YOLOv2 [5], DSSD [6] and RetinaNet [2], which directly calculate both bounding boxes and confidences of object classes. In this paper, we aim to use the single-stage detection model to handle different occlusion patterns on pedestrian detection, by designing a novel network named as pedestrian detection with semantic attention (PDSA). More specially, PDSA contains two components, i.e., a semantic segmentation component and a detector component. The semantic segmentation component is used to reduce the influence of the heavily occluded parts with the visible bounding boxes of pedestrians. It takes low-level features as input and try to learn a feature map supervised by the visible bounding boxes. Furthermore, the feature map will guide as attention to the input features. The detector component uses a single-stage detection model, i.e., RetinaNet [2], which combines feature pyramids [1] to predict the bounding boxes and the confidences of object classes. The input of the detector component is obtained from the semantic segmentation component, which helps the detection model to detect the heavily occluded pedestrians. In particular, we also consider the intra-class occlusion, which occurs when a pedestrian is occluded by other pedestrians, and introduce the repulsion loss [7] to improve the performance of our model.

The main contributions of our work in this paper are summarized as follows:

- We propose the PDSA model that exploits semantic segmentation to address the inter-class occlusion, and introduce the repulsion loss to deal with intra-class occlusion.
- PDSA utilizes semantic segmentation information to reduce the influence of the heavily occluded parts. To the best of our knowledge, it is the first attempt to utilize visible bounding boxes with semantic segmentation component to obtain the semantic attention for pedestrian detection.
- We conduct extensive experiments on the CityPersons dataset containing heavily occluded pedestrians, and outperforming the state-of-the-art approaches.

The rest of the paper is organized as follows. In Sect. 2, related work is reviewed. In Sect. 3, the motivation is introduced. In Sect. 4, the proposed PDSA model is presented. In Sect. 5, extensive experiments are conducted and analyzed. Finally, Sect. 6 offers some concluding remarks.

2 Related Work

In this section, we review some existing research works on the pedestrian detection and occlusion handling, respectively.

2.1 Pedestrian Detection

Recently, the convolutional neural network (CNN) has achieved great progress on pedestrian detection. In the early time, quite a few works [8–10] tried to apply CNN directly for pedestrian detection. Li et al. [11] proposed SA-Fast RCNN to detect pedestrians in different scales, and Cai et al. [12] used MS-CNN to obtain competitive performance on pedestrian detection. Meantime, Zhang et al. [13] refined the Faster R-CNN network by combining region proposal networks with Boosted Forest, which improves the performance on small objects and hard negative samples. However, these methods are based on the two-stage detectors (i.e., Faster R-CNN [3]), which suffer from high computational cost.

2.2 Occlusion Handling

In term of occlusion handling, part-based methods are one of the mainstream approaches. Ouyang et al. [14] designed a framework that models the part visibility as latent variables to predict the scores of part detectors. Mathias et al. [15] proposed the Franken-classifiers method, which utilized multiple classifiers to learn a specific type of occlusion for different occluded pedestrians. Tian et al. [16] proposed a DeepParts model to obtain competitive performance on occlusion handling. The authors constructed an extensive part pool and integrated these parts scores to the final score of the predicted results. However, the part-based methods usually require the part classifiers to learn corresponding occlusion pattern independently, which results in a lot of computations. Zhang et al. [4] applied channel-wise attention mechanism to handle the occlusion. However, their methods are only suitable for the two-stage detectors, while the single-stage detectors are faster with high performance.

3 Motivation

In the context of occluded pedestrian detection, detectors usually fail to detect the pedestrians due to that detectors learn features from the whole bounding boxes in the training stage. However, the bounding boxes not only contain pedestrians but also may include parts of other pedestrians (i.e., intra-class occlusion) or non-pedestrian objects (i.e., inter-class occlusion).

(a) For the intra-class occlusion, it happens commonly in the crowd, which results in high overlap rate between bounding boxes. The detectors are easy to predict only a single pedestrian. Inspired by Wang et al. [7], we introduce the repulsion loss to narrow down the gap between a proposal and its designated target, and keep it away from other ground-truth objects.

(b) For the inter-class occlusion, non-pedestrian objects occupy part of the bounding boxes. The features obtained by detectors may result in false detection when similar non-pedestrian objects appear, and fail to detect the pedestrians that the occlusion is heavy. Intuitively, if we reduce the weight of the non-pedestrian objects and emphasize the parts of pedestrians, the detectors will learn positive knowledge. Therefore, we introduce the semantic segmentation component to obtain the semantic attention map, which uses the information of the visible bounding boxes. Through the semantic attention map, the detectors tend to focus more on the parts of pedestrians. The details are presented in the next section.

4 Methodology

4.1 Overview

As shown in Fig. 1, our PDSA consists of two parts: a detector component and a sematic segmentation component. The first part adopts a single-stage detector (i.e., RetinaNet [2]) to predict the bounding boxes and the probability of pedestrians. The sematic segmentation component utilizes the visible bounding boxes as the input to train a sematic attention map for reducing the influence of the heavily occluded parts. We will introduce these components in more details subsequently.

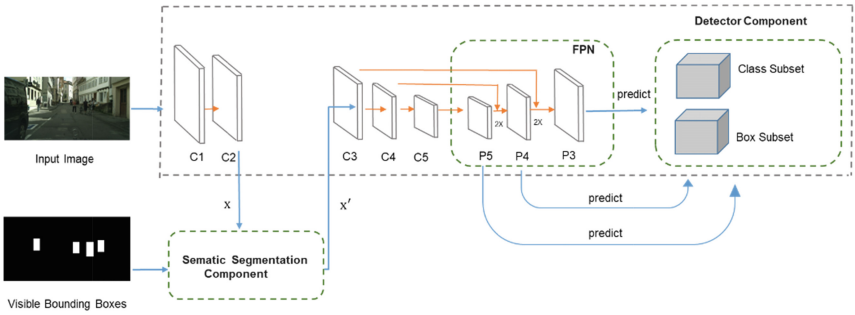

Fig. 1. Overview of our PDSA. FPN denotes the feature pyramid network.

4.2 Detector Component

Our detector component is a single-stage detector (i.e., RetinaNet). We replace the basic model ResNet [17] with VGG16 [18]. Therefore, our detector contains five blocks of convolution layers (i.e., C1, C2, C3, C4 and C5). In addition, RetinaNet utilizes the feature pyramid networks (FPN) [1] to adapt multi-scale pedestrians, which contains additional three convolution layers (i.e., P5, P4 and P3) that combine with the previous convolution layers (i.e., C5, C4 and C3). More specifically, P5, P4 and P3 utilize 1×1 convolutional layer with ReLU function [19], and the input of P5 comes from C5 directly. The input of P4 is the combination of P5 with upsampling method [1]

and C4, and the input of P3 is same combination of P4 and C3. In addition, we utilize the focal loss [2] to train the classification loss, which is defined as follows:

$$L_{Classification} = \begin{cases} -(1-p)^{\gamma}\log(p) \ if \ y = 1 \\ -p^{\gamma}\log(1-p) \ otherwise \end{cases} \tag{1}$$

where $y \in \{0, 1\}$ is a ground truth class, $p \in [0, 1]$ is the probability for the class with label $y = 1$.

In addition, in order to handle the intra-class occlusion, we introduce the repulsion loss [7] to optimize the detector. In particular, we only add the repulsion term loss, which repels the proposal from its neighboring ground truth objects. Here, we assume that P is the positive proposals set ($IoU \geq 0.5$), and B is the predicted bounding box regressed from proposal P, and G is the ground truth bounding box. The repulsion loss can be defined as:

$$L_{Rep} = \frac{\sum_{p \in P} Smooth_{L1}\left(IoG\left(B^p, G^p_{Rep}\right)\right)}{|P|} \tag{2}$$

$$G^P_{Rep} = max_{G \setminus max_G IoU(G,P)} IoU(G, P) \tag{3}$$

where $IoG(B, G) = \frac{area(B \cap G)}{area(G)}$ denotes the overlap between B and G, $Smooth_{L1}$ represents the smooth L_1 distance [3], and IoU denotes the Intersection over Union [20].

4.3 Sematic Segmentation Component

The inter-class occlusion for pedestrian detection is handled by the sematic segmentation component. As shown in Fig. 2, it takes the low-level detection layer (i.e., the output of C2 block, denoted as X) and the visible bounding boxes as input. In particular, the ground truth is generated by the visible bounding boxes. Furthermore, the semantic segmentation component generates a sematic attention map with the same dimension as the input layer. Finally, we utilize this map to activate the input layer by element-wise multiplication to obtain the output feature X'.

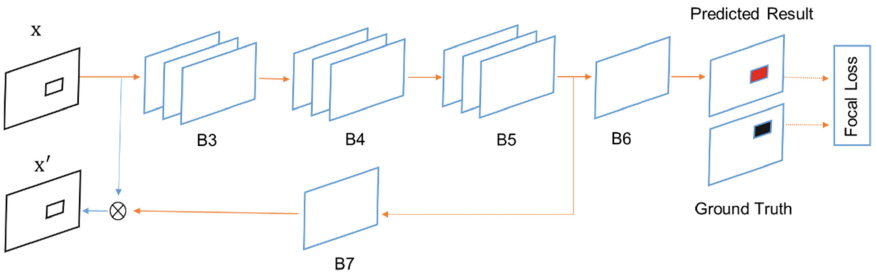

Fig. 2. The structure of sematic segmentation component.

For the segmentation network, we use the same structure with VGG16 but remove the pooling layers, and replace the subsequent two convolution blocks with the dilated convolution blocks (i.e., B4 and B5) [21]. Furthermore, we utilize two 1×1 convolutional layers with sigmoid function (i.e., B6 and B7) to generate the segmentation prediction and the sematic attention map, respectively. However, the ground truth is the visible bounding boxes, which are only four coordinate points and cannot be input directly. Here, we scale the visible bounding boxes into 1/4, which is the same size as the result of segmentation prediction. Furthermore, we set the whole pixels in the visible bounding boxes as 1 and the others as 0, which obtains the segmentation ground truth.

In addition, the visible part occupies a small area which leads to the imbalance between the positive and negative samples. To make the sematic segmentation task converged, we introduce the focal loss for optimization as follows:

$$L_{Segmentation} = \begin{cases} -(1-p)^{\gamma}\log(p) \ if \ y = 1 \\ -p^{\gamma}\log(1-p) \ otherwise \end{cases} \tag{4}$$

where $y \in \{0, 1\}$ is ground truth class for each pixel, $p \in [0, 1]$ is the probability for the class with label $y = 1$.

By adding the aforementioned loss function, the final objective function of PDSA is given follows:

$$L = L_{Regression} + L_{Classification} + \alpha L_{Rep} + \beta L_{Segmentation} \tag{5}$$

where $L_{Regression}$ is the original bounding box regression loss. The parameters α and β balance different tasks. The convergence of the segmentation loss is shown in Sect. 5.6.

5 Experiments

In this section, we introduce the dataset used for pedestrian detection, and evaluate the performance of the proposed approaches and baselines.

5.1 Dataset

The CityPersons dataset [22] consists of cityscape images containing persons, with backgrounds including Germany and some other surrounding countries. The ground truth of the images contains bounding box annotation, visible bounding box annotation, and five class labels (i.e., ignore regions, pedestrians, riders, sitting persons, other persons with unusual postures, and group of people). As show in Table 1, the dataset contains 3,475 images in total with rich annotations including 23k pedestrians and 9k ignored regions. The training set contains nearly 3,000 images, with an average of seven pedestrians per image. Only 30% of the pedestrians are visible completely, which shows that the CityPersons dataset have rich types of occlusion.

Table 1. Statistics of the CityPersons dataset

	Train	Val	Total
#Images	2,975	500	3,475
#Persons	19,654	3,938	23,592

5.2 Evaluation Metrics

We use a commonly used metric on the CityPersons dataset [22] for evaluations, i.e., the average value of miss rate for the false positive per image (MR), ranging from 10^{-2} to 10^{0} (the smaller the better). In this paper, we care more about the occlusion and only consider pedestrians with height $\in$ [50, *inf*]. We show the results across three different occlusion levels. In addition, we visualize the distribution of pedestrian at different occlusion level on CityPersons, as shown in Fig. 3.

(1) Reasonable (R): visibility $\in$ [0.65, *inf*];
(2) Heavy occlusion (HO): visibility $\in$ [0.2, 0.65];
(3) Reasonable + Heavy occlusion (R + HO): visibility $\in$ [0.2, *inf*].

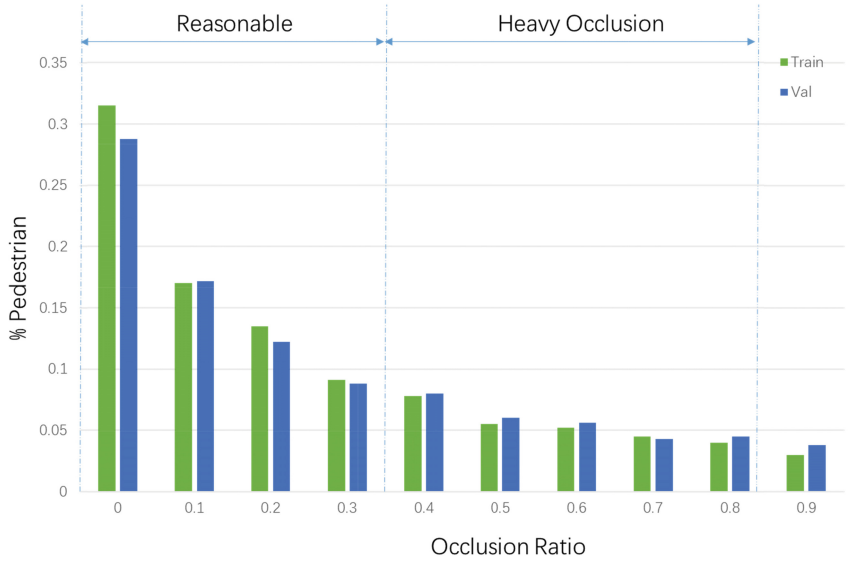

Fig. 3. Occlusion distribution on CityPersons dataset.

5.3 Implementation Details

In our experiment, we adopt VGG16 [18] as the fundamental network structure, and the other convolution layers in detector component are same as [2]. For the sematic segmentation component, the previous three convolution layers (i.e., B3, B4 and B5) are

the same with VGG16 [18], but we remove the pooling layers and utilize the dilated convolution in the last two layers (i.e., B4 and B5). In particular, the dilation rates of dilated convolution are set as 2 and 4, respectively. B6 is a 1×1 convolution with sigmoid function and the channel number is 1, while B7 is also a 1×1 convolution with sigmoid function and the channel number is same as the channel number of the input X.

For the optimizations, we use the parameters of the pre-trained VGG16 to initialize our model, and initialize the dilated convolution parameters in the segmentation component by Xavier initialization [23]. We adopt the Adam solver [24] with the learning rate of 10^{-4} for 14,000 iterations, and take the image in original size as the input. In addition, the balance parameter for repulsion loss α set as 0.5 following [7], and the balance parameter for the semantic segmentation loss β set as 0.5 since our main task is not semantic segmentation.

5.4 Comparing with the State-of-the-Art Methods

The baselines include a number of state-of-the-art methods on pedestrian detection, such as FasterRCNN [13], FasterRCNN + ATT-part [4], FasterRCNN + RepLoss [7], Somatic Topology Line Localization (TLL) [25] and RetinaNet [2]. The performances of the approaches are shown in Table 2. From the table, we can observe that the proposed PDSA achieves competitive performance for HO and R + HO, which outperforms the previous state-of-the-art detectors. The proposed PDSA benefits from the semantic attention map and the repulsion loss, which can detect heavily occluded pedestrians effectively. Note that our PDSA cannot outperform the best baselines for R. The reason is that we use the single-stage detector (i.e., RetinaNet), which is not fully optimized for the small-scale pedestrian detection, while the baselines use the two-stage detector (i.e., Faster RCNN).

Table 2. MR performance of the approaches on the CityPersons dataset.

Method	R	HO	R + HO
FasterRCNN	15.52%	64.83%	41.45%
FasterRCNN + ATT-part	15.96%	56.66%	38.23%
FasterRCNN + RepLoss	**13.20%**	56.90%	-
TLL	14.40%	52.00%	-
RetinaNet	17.92%	56.22%	36.61%
PDSA	16.51%	**48.24%**	**31.88%**

5.5 Evaluation on Different Strategies

PDSA utilizes the single-stage detector (i.e., RetinaNet). We adopt the repulsion loss to deal with the intra-class occlusion. Besides, we introduce a semantic segmentation component to deal with the inter-class occlusion. To evaluate the two components, we denote PDSA with repulsion loss as PDSA-r and PDSA with semantic segmentation component as PDSA-s, respectively. The results are shown in Table 3. We notice that

PDSA-r performs well for R since it is robust to the influence of intra-class occlusion. In addition, PDSA-s outperforms the RetinaNet on different occlusions, which demonstrates that the semantic segmentation component is effective for addressing occlusion. Furthermore, we combine repulsion loss with the semantic segmentation component and finally obtain the best performance, taking into account both intra-class and inter-class occlusions.

Table 3. Comparison of different strategies on the CityPersons dataset (lower is better).

Method	+Repulsion Loss	+Segmentation	R	HO	R + HO
RetinaNet	-	-	17.92%	56.22%	36.61%
PDSA-r	✓	-	16.73%	56.51%	35.80%
PDSA-s	-	✓	16.86%	48.77%	32.46%
PDSA	✓	✓	**16.51%**	**48.24%**	**31.88%**

5.6 Convergence of PDSA

PDSA consists of four loss terms, i.e., regression loss, classification loss, semantic segmentation loss, and repulsion loss. As shown in Fig. 4, we can see that all the losses are converged after 10,000 iterations. The experimental results demonstrate the effectiveness of our training procedures.

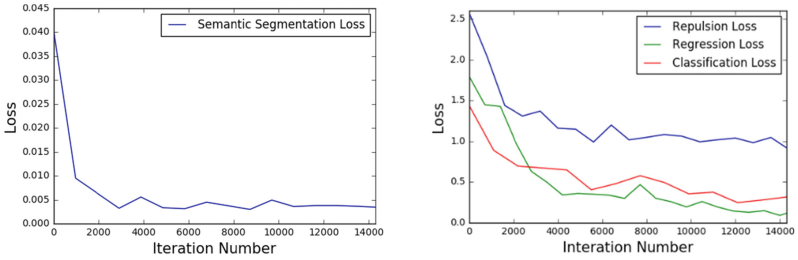

Fig. 4. Convergence of PDSA

5.7 Visualizations

As shown in Fig. 5, we visualize the semantic attention map trained by the semantic segmentation component. We can see that the full bodies and the visible parts of occluded persons result in obvious response on the heatmap. For instance, two pedestrians are occluded heavily by cars while their upper bodies still show obvious response. The heatmap demonstrates that our sematic segmentation component can extract features from heavily occluded pedestrians.

Furthermore, we visualize the bounding boxes predicted by RetinaNet and the proposed PDSA model in Fig. 6. The RetinaNet fails to detect pedestrians that are occluded by other non-pedestrian objects, while our pedestrian detector obviously

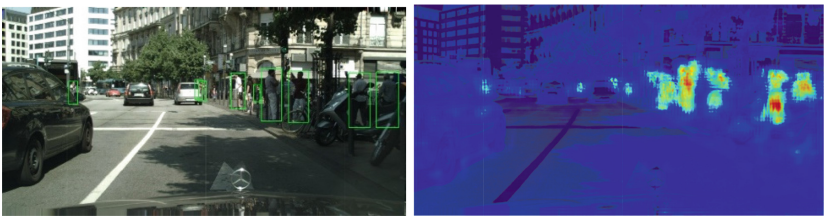

Fig. 5. Visualization of the visible parts of persons on the heatmap.

(a) RetinaNet

(b)PDSA

Fig. 6. Detected results of RetinaNet and PDSA. The red bounding boxes represent the detected results, and the green ones represent the ground truth. (Color figure online)

reduces the samples of false positives and missed detections. In addition, we find that our PDSA can locate the different pedestrians in the crowd, which demonstrates that our method is effective for both inter-class and intra-class occlusions.

6 Conclusion

In this paper, we propose a novel method PDSA for occluded pedestrian detection. In order to handle the inter-class pedestrian occlusion, we introduce a semantic segmentation component, which utilizes the visible bounding boxes to obtain the semantic attention map. This component helps the subsequent detector component to focus on the pedestrians when the occlusion happened. In particular, we introduce the repulsion

loss to deal with the intra-class occlusion, which helps to improve the performance of our PDSA. The experiment results have demonstrated the effectiveness of our proposed approach, which achieves the state-of-the-art performance on heavy occlusion.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61703109, No. 91748107), China Postdoctoral Science Foundation (No. 2018M643026), and the Guangdong Innovative Research Team Program (No. 2014ZT05G157).

References

1. Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125 (2017)
2. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: International Conference on Computer Vision (ICCV), pp. 2999–3007 (2017)
3. Girshick, R.: Fast R-CNN. In: Computer Vision and Pattern Recognition (CVPR), pp. 1440–1448 (2015)
4. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNs. In: Computer Vision and Pattern Recognition (CVPR), pp. 6995–7003 (2018)
5. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Computer Vision and Pattern Recognition (CVPR) (2017)
6. Fu, C., Liu, W., Ranga, A., Tyagi, A., Berg, A.: DSSD: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)
7. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: detecting pedestrians in a crowd. In: International Conference on Computer Vision (CVPR) (2018)
8. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: Computer Vision and Pattern Recognition (CVPR) (2014)
9. Hosang, J., Omran, M., Benenson, R., Schiele, B.: Taking a deeper look at pedestrians. In: Computer Vision and Pattern Recognition (CVPR), pp. 4073–4082 (2015)
10. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: Computer Vision and Pattern Recognition (CVPR) (2015)
11. Li, J., Liang, X., Shen, S., Xu, T., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. IEEE Trans. Multimedia **20**(4), 985–996 (2017)
12. Cai, Z., Fan, Q., Feris, Rogerio 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
13. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 443–457. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_28
14. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: Computer Vision and Pattern Recognition (CVPR) (2012)
15. Mathias, M., Benenson, R., Timofte, R., Van, L.: Handling occlusions with Franken-classifiers. In: International Conference on Computer Vision (ICCV) (2013)
16. 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 (ICCV), pp. 1904–1912 (2015)

17. 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)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2014)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, vol. 60, pp. 1097–1105 (2012)
20. Jiang, Y., Jiang, Y., Cao, Z., Cao, Z., Huang, T.: UnitBox: an advanced object detection network. In: ACM on Multimedia Conference, pp. 516–520 (2016)
21. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
22. Zhang, S., Benenson, R., Schiele, B.: CityPersons: a diverse dataset for pedestrian detection. In: Computer Vision and Pattern Recognition (CVPR) (2017)
23. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (ICAI), pp. 249–256 (2010)
24. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation. arXiv preprint arXiv:1807.01438 (2018)