SAF: Semantic Attention Fusion Mechanism for Pedestrian Detection

Ruizhe Yu, Shunzhou Wang, Yao Lu(✉), Huijun Di, Lin Zhang, and Lihua Lu

Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
{2120171090,shunzhouwang,vis_yl,ajon,zhanglin,lulihua}@bit.edu.cn

Abstract. Benefiting from deep learning methods, pedestrian detection has witnessed a great progress in recent years. However, many pedestrian detectors are prone to detect background instances, especially under urban scenes, which results in plenty of false positive detections. In this paper, we propose a semantic attention fusion mechanism (SAF) to increase the discriminability of detector. The SAF includes two key components, attention modules and reverse fusion blocks. Different from previous attention mechanisms which use attention modules for re-weighting the top features of network directly, the outputs of our attention modules are fused by reverse fusion blocks from high level layers to low level layers step by step, which aims at generating strong semantic features for pedestrian detections. Experiments on CityPersons dataset demonstrate the effectiveness of our SAF.

Keywords: Pedestrian detection · Semantic attention · Background errors

1 Introduction

Pedestrian detection is the fundament of many computer vision tasks, such as pose estimation, multiple pedestrian tracking, and action recognition. Benefitting from the deep learning method, pedestrian detection has got a significant progress such as [7,11–13,15,18,20,21]. However, there still exist many issues to be solved, such as the occluded pedestrian detections, the small scale pedestrian detections, background errors and so on [19]. Specially, background errors are the main source of false positive detections [19], especially under the urban street scenes.

There are some approaches for decreasing background errors. One way [18] is that adding the boosted forests after the top of network to refine the output detections, the other way [11] is that adding more predictors to improving detection results. The above methods all employ two or more classification steps for filtering out background errors, but some small background objects are still detected as pedestrian wrongly due to the low layers of network lacking of enough pedestrian semantic information. Attention mechanism can be used

A. C. Nayak and A. Sharma (Eds.): PRICAI 2019, LNAI 11671, pp. 523–533, 2019.
https://doi.org/10.1007/978-3-030-29911-8_40

for suppressing background information and Zhang et al. [21] develop different attention modules for guiding the network pay more attention to different pedestrian parts. However, they do not consider the fact that some background instances will also have high detection responses and are tend to be detected as pedestrians. In a word, the pedestrian semantic information is not utilized effectively in [11,18,21].

Fig. 1. The visualization results of ALFNet without (a) and with (b) the semantic attention fusion mechanism (SAF) from *CityPersons* validation set. The green and red rectangles represent detection results and the ground truth results respectively. The comparison indicates that our semantic attention fusion mechanism decreases background errors effectively. (Color figure online)

In order to decrease background errors effectively, we propose a Semantic Attention Fusion mechanism (SAF), which consists of attention modules(AM) and reverse fusion blocks(RFB) [6,8,10]. Different from previous attention mechanisms [21] which use attention modules for re-weighting the top features of network directly, the outputs of attention modules are fused by reverse fusion blocks from high level layers to low level layers step by step, which aims at generating strong semantic features for pedestrian detections. Moreover, we place our SAF on ALFNet [11] and Fig. 1 shows the visualization results of the ALFNet without and with our SAF. The number of background errors in Fig. 1(b) decreases indeed comparing with that in Fig. 1(a), which indicates that our SAF works effectively.

The rest of our paper is organized as follows. Section 2 introduces some related works about pedestrian detections and attention modules. Section 3 represents our proposed SAF and how to apply the SAF on ALFNet [11]. In Sect. 4, we perform some experiments on CityPersons dataset for evaluating effectiveness of our method. In Sect. 5, we conclude the paper.

2 Related Work

In this section, we will review some related works from the following.

Pedestrian Detection. In recent years, there are a lot of works improving pedestrian detection performance from different aspects. Zhang et al. [20] did five modifications of the Faster R-CNN and made the Faster R-CNN achieve the state of the art results on Caltech dataset [2]. Mao et al. [12] analysed how different features affect the final detection results and designed a HyperLearner combining with extra features. Zhang et al. [21] added attention mechanism to the Faster R-CNN for detecting occluded pedestrians. Lin et al. [7] designed a graininess attention mask and proposed a zoom-in-zoom-out mechanism to detect small and occluded pedestrians. Wang et al. [15] proposed the PCN network, which used part semantic information and contextual information to detect occluded pedestrians. While the above works [7,12,15,20,21] mainly focus on solving the small scale pedestrians detection and occluded pedestrians detection, our SAF is proposed for decreasing false background errors and increasing the ability of the network to distinguish pedestrians from hard background instances.

Attention Module. Some attention modules have been proposed in recent years. Hu et al. [5], Squeeze and Excitation module(SE) were proposed for reweighting channel weights of different layers and after that SE module were applied in different computer vision problems. Zhang et al. [21] chose the SE module as the network component for detecting occluded pedestrians accurately. Yu et al. [17] used the SE module for helping the network to select discriminative features for semantic segmentation. However, SE module only considered channel relationship of network layer and the spatial information of feature map was also important. Therefore, Woo et al. [16] designed convolutional block attention module(CBAM), which helped Faster-RCNN [14] to get performances improvement on MS COCO detection dataset and VOC 2007 detection dataset. In this work, CBAM [16] is chosen as our attention module and the outputs of different attention modules are fused step by step from high level layers to low level layers with the help of reverse fusion blocks for generating strong semantic features.

3 Proposed Method

We apply our proposed SAF on ALFNet [11] and the architecture of all network is shown in Fig. 2. It can be divided into three parts: Backbone Network, Semantic Attention Fusion (SAF) and Asymptotic Localization Fitting Module (ALF) [11]. Each input image is processed by the three parts sequentially to get final detection results. Backbone Network is used for extracting the convolutional features of each input image. The SAF is used for outputting strong semantic features for pedestrian detection and the ALF is used for outputting the final detection results.

3.1 Backbone Network

MobileNet [4], which has been pretrained on the ImageNet dataset, is chosen as our backbone. As shown in Fig. 2, we select feature maps from the backbone network for different scale pedestrian detections and these feature maps are named

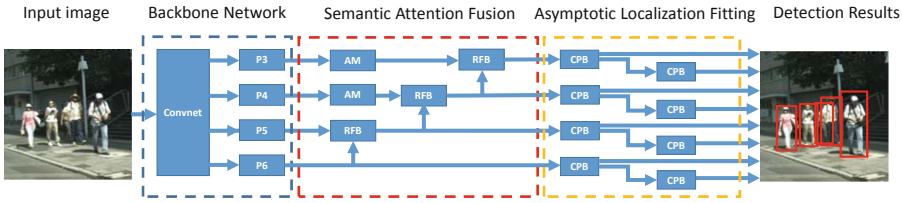

Fig. 2. Illusion of the ALFNet with SAF. Each input image is firstly processed by the backbone network (Sect. 3.1). Then, the output features are processed by the semantic attention fusion (Sect. 3.2) for generating more strong semantic features. Finally, asymptotic localization fitting modules (Sect. 3.3) receive all refined features and output the final pedestrian detections.

as P3, P4, and P5, whose size are 1/8, 1/16, and 1/32 of the original input image respectively. Especially, P6 is a new added convolutional layer, which is used for making pre-trained backbone adapt pedestrian detection task [11], and acquiring high level semantic information as well as global contextual information during training stage [17].

3.2 Semantic Attention Fusion

Different from the work [21] whose attention modules are directly used for re-weighting the top features of network, in this paper, our semantic attention fusion helps the network to learn more strong semantic features via fusing features from high level layers to low level layers step by step. The diagram of the semantic attention fusion is shown in Fig. 2.

The proposed semantic attention fusion includes two key components: attention modules and reverse fusion blocks. Attention modules are used for helping the network to concentrate on more pedestrian relevant features and suppressing background object responses. Reverse fusion blocks are employed to fuse features from high level layers to low level layers step by step for generating more strong semantic features. Combining with attention modules and reverse fusion blocks, our SAF outputs features with more semantic information layer by layer. Specially, our attention modules are only attached to P3 and P4. The reason is that P3 and P4 include lots of low level object detailed features and only some of the features represent pedestrians. Therefore, attention modules help P3 and P4 to filter out irrelevant background information and pay more attention to relevant pedestrian features. P5 and P6 have owned abundant high level semantic features for pedestrian detection so it is not necessary to add attention modules after P5 and P6.

Convolutional block attention module (CBAM) [16] is chosen as the AM and its structure is shown in Fig. 3. The CBAM has two parts, channel attention module and spatial attention module. The channel attention module focuses on finding whether the feature map including pedestrians or not and the spatial attention module focuses on looking for where pedestrians are. The channel

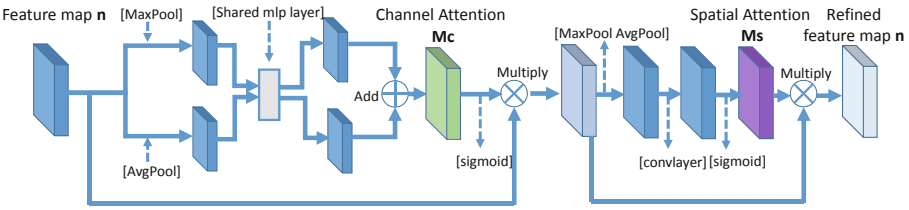

Fig. 3. Convolutional block attention module is chosen as our attention module. The feature maps are processed by the channel attention module and the spatial attention module in sequence.

attention module is followed by spatial attention module, which is the same as [16]. The attention process can be presented as follows:

$$\begin{aligned} F^{'} &= M_c(F) \otimes F \\ F^{''} &= M_s(F^{'}) \otimes F^{'} \end{aligned} \tag{1}$$

where $\otimes$ represents element-wise multiplication. M_c is the channel attention vector generated by the channel attention module and M_s is the spatial attention vector generated by the spatial attention module. F, $F^{'}$ and $F^{''}$ stand for original feature maps, feature maps processed by the channel attention module and feature maps processed by the CBAM module respectively.

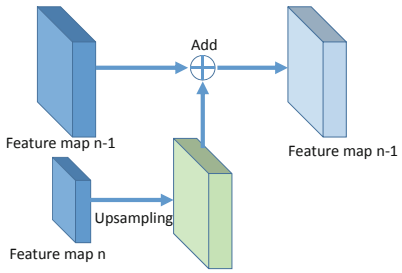

Fig. 4. The diagram of reverse fusion block. The feature map of Stage n is up-sampled to the same size of Stage n-1 and above feature maps are processed element-wise add with each other to generate new feature map for Stage n-1.

The kernel size of the spatial attention module should be set manually. We find that the detection accuracy is the best when the kernel size was set to 3. Effects of different spatial kernel size value on the final detection results will be discussed in Sect. 4.2.

Following AM, the RFB workflow is shown in Fig. 4. Feature maps of Stage n are firstly up-sampled to the same spatial size of Stage n-1. Then, the above two feature maps will do element-wise add to generate the new feature maps of Stage n-1.

3.3 Asymptotic Localization Fitting Module

The ALF module is the same with [11]. We also stacked two ALF modules in sequence and the two steps IoU thresholds are set as {0.3, 0.5} and {0.5, 0.7} respectively. Convolutional predictor block (CPB) [11], whose structure is shown in Fig. 5, is used for converting feature maps to the final detection results.

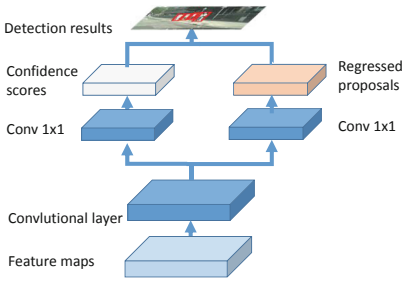

Fig. 5. The workflow of convolutional prediction block. Each feature map will be processed by the classification branch and regression branch for getting the final detection results.

3.4 Loss Functions

Being the same with [9,11], each convolutional predictor is optimized with the multi-task loss as follows:

$$L = l_{cls} + \lambda l_{loc} \tag{2}$$

where l_{cls} is the classification loss and l_{loc} is the localization loss. λ is the weight parameter to balance the above two losses. In order to decrease the impact of imbalanced amount of positive samples and negatives samples on the final detection accuracy, we also choose the focal loss as our classification loss. The focal loss is formulated as:

$$l_{cls} = -\alpha \sum_{i \in S_+} (1 - p_i)^{\gamma} log(p_i) - (1 - \alpha) \sum_{i \in S_-} p_i{}^{\gamma} log(1 - p_i) \tag{3}$$

where S_+ presents the samples whose IoUs with ground truth are larger than u_{high} and S_- presents the samples whose IoUs with ground truth are smaller than u_{lower}. The values of u_{high} and u_{high} are set according to the description in Sect. 3.3. Following [9], p_i stands for the likelihood of sample i to be positive. α and γ are hyper-parameters, which are set to 0.25 and 0.75 respectively.

The $smooth_{L_1}$ loss [3] is chosen as the localization loss, which is formulated as:

$$l_{loc} = \sum_{i \in x,y,w,h} smooth_{L_1}(t_i{}^{u} - v_i) \tag{4}$$

$$smooth_{L_1} = \begin{cases} 0.5x^2, & if|x| < 1. \\ |x| - 0.5, & otherwise. \end{cases} \tag{5}$$

where t^u is the predicted bounding box of pedestrian class and v is the ground truth bounding box. x and y represent the left corner location of the bounding box. w and h are the width and height of the bounding box.

4 Experiments

4.1 Experiments Setup

CityPersons dataset [20] is chosen to evaluate the performance of our proposed method. The CityPersons dataset is derived from the *Cityscapes* dataset [1], which is one of the most Challenging semantic segmentation dataset. The CityPersons dataset includes 5000 images recorded from 27 different cities. There are 35016 persons and 13172 ignore regions annotated. The CityPersons dataset is divided into training subset, validation subset and test subset, which includes 2975 images, 500 images and 1525 images respectively. MR^{-2} [20] is chosen as the evaluation metric. Data augumentation strategies are used, such as random cropping, horizontal flipping and color distortion. The Adam is chosen as our network optimizer. Less than 240k iterations of ALFNet training setting [11], our method only need 40k iterations to be convergent. The initial learning rate is set to 0.0001 and after 20k iterations the learning rate is divided by 10. All experiments is under the Ubuntu 14.04 system, with the Intel Core i5-4430, 32 G RAM and only one GTX Titan X card. In addition, all experiments are performed with the original image size, whose the width of image is 2048 and the height of image is 1024.

4.2 Ablation Studies

In this part, some experiments are performed on the validation set for analyzing the effect of different spatial kernel size of spatial attention module on the final detection accuracy. We also compare our method with ALFNet based on Mobilenet backbone [11] to prove effectiveness of our SAF.

Different Spatial Kernel Size. There is a hyper-parameter about kernel size of spatial attention module in the CBAM. To find the best value of the kernel size, we have done experiments on the validation set. The results are shown in Table 1. When the kernel size is set to 3, our method achieves the best result, no matter what different IoU thresholds are. The results of kernel size with 5 and kernel size with 1 are not better than the results of kernel size with 3, which means that too large or too small kernel size of the spatial attention module doesn't help the low layers for acquiring semantic information from the high layers. Small kernel size can't acquire enough features and large kernel size may introduce noise features, which harms the final pedestrian detection.

Table 1. Comparison of the different spatial kernel sizes under different IoU thresholds on CityPersons validation set.

Spatial kernel size	Test time	**R**	**HO**	**R+HO**
1	0.24 s/img	16.67	46.96	31.39
3	0.24 s/img	**16.25**	**45.12**	**29.81**
5	0.24 s/img	16.85	46.92	31.52

Semantic Attention Fusion Improvement. To validate effectiveness of our semantic attention fusion, we perform experiments on the validation set of CityPersons. The results are shown in Table 2. From Table 2, we find that when IoU threshold is equal to 0.5, the performances of ALFNet with RFB and our proposed method are lower than the original ALFNet, which decreases **1.12**% point and **0.80**% point respectively. But when IoU threshold is equal to 0.75, the results of above two methods are better than original ALFNet. Specially, our method gets a **3.55**% point gain under more restrict threshold comparing with the original ALFNet. The detection accuracy improvement can be attributed to the semantic attention fusion, which makes our network pay more attention to pedestrian relevant features.

Table 2. Semantic attention fusion improvement evaluated under different IoU thresholds on CityPersons validation set. *ALF* represents two stacked asymptotic localization fitting modules. *RFB* represents reverse fusion blocks and *SA* represents that attention modules used for semantic attention fusion.

Backbone	**ALF**	**RFB**	**SA**	MR^{-2}	
				IoU = 0.5	IoU = 0.75
ResNet-50 [11]				16.01	48.94
MobileNet [11]				18.88	56.26
MobileNet [11]	√			**15.45**	47.42
MobileNet	√	√		16.57	44.24
MobileNet	√	√	√	16.25	**43.87**

Meanwhile, the visualization results of different methods are shown in Fig. 6. The first row shows that original ALFNet detects instances from background, such as wall or car, as pedestrian by mistake. The detection results of ALFNet with RFB are shown in the second row and some background errors are reduced. The results of our proposed method are shown in the last row and we find that the number of background errors is the least among the above three methods. The above comparisons demonstrate the effectiveness of our semantic attention fusion and efficiency of our method in handling background errors.

Table 3. Comparison of our method with the other attention networks on CityPersons validation set. All experiments are done with the original image size of the validation set. **R** represents reasonable occlusion level. **HO** represents heavy occlusion level. **R+HO** represents occlusion level includes reasonable occlusion situation and heavy occlusion situation.

Method	Attention part	**R**	**HO**	**R+HO**
Faster-RCNN (VGG-16) [21]	-	15.52	64.83	41.45
Faster-RCNN (VGG-16)+ATT-self [21]	Self attention	20.93	58.33	40.83
Faster-RCNN (VGG-16)+ATT-vbb [21]	vbb supervision	16.40	57.31	39.49
Faster-RCNN (VGG-16)+ATT-part [21]	Part detections	15.96	56.66	38.23
ALFNet (MobileNet) [11]	-	**15.46**	47.40	31.11
Ours (MobileNet)	Semantic attention	16.25	**45.12**	**29.81**

Fig. 6. The visualization of detection results. The first row((a)) displays the outputs of original ALFNet [11]. The second row((b)) displays the outputs of ALFNet with RFB. The third row((c)) displays the outputs of our proposed method. The green and red rectangles represent detection results and the ground truth results respectively. (Color figure online)

4.3 Comparing with the Other Attention Networks

Our method is also compared with the other attention networks on the validation set of CityPersons. The results of different occlusion levels are shown in Table 3. Comparing with Faster-RCNN(VGG-16)+ATT-part [21], our proposed method achieves **11.54**% point improvement under heavy occlusion subset(**HO**) and **8.42**% point improvement under reasonable plus heavy occlusion subset(**R+HO**). Comparing with ALFNet [11], our method also achieve **2.28**% point and **1.30**% point improvement under **HO** subset and **R+HO** subset

respectively. The above comparisons indicate two facts. On one hand, our SAF is more practical than the other attention modules because we don't need external information as attention guidance. On the other hand, the results also show that semantic attention fusion mechanism is important for pedestrian detection, especially for occluded pedestrian detection.

5 Conclusions

In this paper, we propose a semantic attention fusion mechanism for pedestrian detection. We use the convolutional block attention module as our attention module and we choose the reverse fusion block to fuse and transmit semantic information from the high layers to the low layers step by step. The above two components construct the semantic attention fusion, which makes the network own more strong semantic features for pedestrian detection. To validate our method, some experiments are performed on CityPersons dataset. The effects of different kernel sizes of the spatial attention module are compared and our proposed method achieve a 6.85% point gain than the ALFNet without the SAF under the strict IoU threshold. We also compare our method with the other attention networks on the validation set of CityPersons and our method also achieves more superior results than theirs. In the future, we can extend our semantic attention fusion mechanism to another pedestrian detectors, such as the Faster R-CNN and so on.

Acknowledgments. This work is in part supported by the National Nature Science Foundation of China (No. 61273273), by the National Key Research and Development Plan, China (No. 2017YFC0112001).

References

1. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
2. 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 (2011)
3. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Howard, A.G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
6. Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., Chen, Y.: Ron: reverse connection with objectness prior networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5936–5944 (2017)

7. Lin, C., Lu, J., Wang, G., Zhou, J.: Graininess-aware deep feature learning for pedestrian detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 745–761. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_45
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
10. Liu, L., et al.: Deep learning for generic object detection: A survey. arXiv preprint arXiv:1809.02165 (2018)
11. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 643–659. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_38
12. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3127–3136 (2017)
13. Noh, J., Lee, S., Kim, B., Kim, G.: Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 966–974 (2018)
14. 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)
15. Wang, S., Cheng, J., Liu, H., Tang, M.: PCN: Part and context information for pedestrian detection with CNNs. arXiv preprint arXiv:1804.04483 (2018)
16. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
17. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1857–1866 (2018)
18. 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
19. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1259–1267 (2016)
20. Zhang, S., Benenson, R., Schiele, B.: Citypersons: a diverse dataset for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3221 (2017)
21. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6995–7003 (2018)