A Hybrid Self-Attention Model for Pedestrians Detection

Yuan Wang, Chao Zhu(✉), and Xu-Cheng Yin

School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing, China
LHMY599@163.com, {chaozhu,xuchengyin}@ustb.edu.cn

Abstract. In recent years, with the research enthusiasm of deep learning, pedestrian detection has made significant progress. However, the performance of state-of-the-art algorithms are still limited due to the high complexity of the detection scene. Therefore, in order to better distinguish between pedestrians and background, we propose a novel hybrid attention module which is capable of obtaining inter-dependencies between features from both channel and spatial dimensions through local convolution and dual-pass pooling, and guiding the network to focus on better pedestrians' feature representation while suppressing background noise. Further, we complement the information of channel attention and spatial attention through an effective fusion mechanism. To validate the effectiveness of the proposed hybrid attention module, we embed it into a representative pedestrian detection framework named Center and Scale Prediction (CSP) based detector. The experimental results on the Caltech Pedestrians Benchmark, one of the largest pedestrian detection datasets, show that the proposed method outperform not only the baseline framework but also several state-of-the-arts.

Keywords: Pedestrian detection · Channel attention · Spatial attention · Fusion mechanism

1 Introduction

Pedestrian detection has been an important research hotspot in the field of computer vision for many years and has attracted widespread attention in both academia and industry. In recent years, due to the great success of deep learning in general object detection, many effective algorithms [1–3] have been adjusted and transplanted into pedestrian detection, which greatly improves the performance of pedestrian detection. However, the performance of the state-of-the-art pedestrian detection algorithm still does not reach human standards [4]. Due to the complexity of the pedestrian detection environment in real scenes, it is difficult to distinguish between pedestrians and backgrounds, resulting in many network detection performance limitations. Therefore, the way to enhance the capability of pedestrians' feature representation is the key to further improving the performance of pedestrian detection.

© Springer Nature Switzerland AG 2020
H. Yang et al. (Eds.): ICONIP 2020, LNCS 12532, pp. 62–74, 2020.
https://doi.org/10.1007/978-3-030-63830-6_6

To achieve this, [5] reduces false positive samples from backgrounds by fusing multiple parallel networks, but it is very time-consuming. [6] gradually locates pedestrians for better classification through multi-step predictions that raise the IOU threshold multiple times, but it does not perform well at low IOU thresholds. [7,8] proposes two new loss functions to guide the network to optimize pedestrian feature representation, but they require complex hyperparameter adjustments.

Recently, attention mechanism is applied to pedestrian detection with excellent feature enhancement and discrimination ability. [9] proposes a channel attention model to enhance the feature representation of the visible part of the human body through different feature channels. [10] integrates bounding-box level segmentation information of pedestrian visible areas into detection branches to guide detector to better distinguish background and pedestrian. However, these models require additional data sets for pre-training or extra information for supervision. Also, few methods in the literature of pedestrian detection use both channel and spatial attention to guide the network with multi-dimensional information fusion for better pedestrian detection performance.

Therefore, in this paper, we propose a novel hybrid attention module with both channel and spatial attention to boost pedestrian detection. Different from common attention models to obtain global dependencies, our channel attention and spatial attention use one-dimensional convolution and stacked dilated convolution to obtain local dependencies. Such an operation can not only avoid the interference of obtaining long-distance irrelevant information, but also improve the efficiency of attention map calculation. These two attention mechanism can guide the network to focus on better pedestrian feature learning while suppressing background noise after an effective fusion strategy. Our attention model requires neither additional guidance information nor additional database, which makes our model easier to train and optimize. To validate the effectiveness of the proposed method for pedestrian detection, we embed our attention model into a representative pedestrian detection framework named Center and Scale Prediction (CSP) based detector [11], which is a detector that achieves state-of-the-art performance for pedestrian detection. The Caltech Pedestrian Benchmark, one of the largest pedestrian detection databases, is adopted to conduct experimental evaluation and comparison of the proposed method against other state-of-the-arts.

In summary, our main contributions are as follows: (1) We propose a novel hybrid self-attention model to obtain local dependencies through channel attention and spatial attention to boost pedestrian detection. (2) The proposed hybrid attention model is based on self-attention information acquisition and thus requires no additional supervision information or pre-trained databases, which makes our model easier to train and optimize. (3) The experimental results demonstrate that being embedded into a representative baseline (CSP) framework, our proposed method can achieve superior detection performance than not only the baseline detector but also several state-of-the-arts.

2 Related Work

2.1 Pedestrian Detection

In recent years, Faster-RCNN [1] is widely used as backbone network in deep learning pedestrians detection algorithms because it's high detection accuracy. RPN+BF [12] adopt the RPN sub-network in Faster-RCNN to generate proposals and then use cascaded boosted forest to refine the proposals in RPN. MS-CNN [13] propose multi-scale detection under different feature maps to enhance pedestrian' feature representation capabilities of Faster-RCNN at multiple scales. SDS-RCNN [14] uses bounding-box level semantic segmentation as additional supervision information to enhance the classification ability of Faster-RCNN. SSD [3] is a another backbone for pedestrian detection because of its high speed. ALFNet [6] use multi-step prediction SSD framework to gradually locate the pedestrian for better accuracy. In order for the network to adaptively enhance pedestrian representation capabilities, OR-CNN [7] and RepLoss [8] proposes two novel loss function to ensure the bounding-box distance between proposals and ground truths as short as possible. Adaptive NMS [15] and CSID [16] shows the new NMS strategy can reduce the possibility of targets filtered by NMS. CSP [11] is an anchor-free detector, it doesn't need tedious design of anchor boxes. CSP can obtain sufficient high-level semantic information by fusing feature maps of different resolutions in the backbone network. CSP reaches the new state-of-art performance of pedestrian detection at that time and thus is chosen as the baseline framework in this work.

Recently, the application of attention mechanism has become a new perspective to further improve the performance of pedestrian detection. Faster-RCNN+ATT [9] designs three subnet with attention methods plugged into Faster-RCNN to guide the network to focus on the visible parts of pedestrians. But Faster-RCNN+ATT requires additional datasets for pre-training, which does not meet the end-to-end manner. MGAN [10] uses two branches to predict the visible part and the whole body of the pedestrian, respectively, then uses the segmentation of the visible part bounding-boxes as a spatial attention mask to multiply into the whole body branch to enhance the learning of advanced semantic features of pedestrians. SSA-CNN [17] share the similar ideas with MGAN [10], but SSA-CNN uses the whole-body segmentation box as a guide and directly cascades the spatial attention mask and detection branch. These methods all require additional supervision information, which increases the difficulty of network optimization. Differently, our goal is to propose a self-attention module that is constructed directly from the inter-dependencies between channels and pixels without additional supervision information.

2.2 Attention Mechanism

Attention mechanism can be divided according to weighted approach, channel attention is given different weights to different feature channels; spatial attention is given different weights according to the importance between pixels in feature maps.

For Channel attention, SENet [18] integrates information between channel levels through global average pooling, then uses two fully connected layers for transformation operations, and finally uses the obtained vector to re-weight the original feature map. Similar to SENet, GENet [19] defines a gather operator with parameters using depthwise separable convolution to learn the context information and a excite operator to adjust the importance of different channels. LCT [20] uses group normalization and $1 * 1$ convolution for normalization operator and transform operator. SKNet [21] consists of multiple branches using different convolution kernel sizes and softmax fusion method, which can dynamically adjust the receptive field according to the target size. Different from them, we apply 1D convolution for channel attention, which can reduce the parameters in information transform. And we adopt max pooling and average pooling as two path to sum together for information aggregation and complementarity.

For spatial attention, STN [22] helps the network to correct image distortion by letting the network learn to calculate the spatial mapping of input map to output map. RAN [23] weights high-level semantic feature maps to low-level detailed feature maps in order to enrich the feature representation ability for lower layer. Non-Local Block [24] calculates the weighted sum of all position on the feature map as the response of a position to obtain the long-distance dependency. Although Non-Local Block doing pretty well on video classification task, but the calculation of NL is very time consuming and the memory consumption is very large. So Several algorithms [25–27] aims to reduce the computational complexity through different computational decomposition methods. Specifically, we stack multiple efficient dilation convolution with different dilation rates for spatial attention, which can well balance the calculation speed and accuracy for spatial attention.

The fusion mechanism of spatial attention and channel attention is an important part of designing an attention mechanism. CBAM [28] uses a serial structure to connect the two in the order of CA (channel attention) and SA (spatial attention). GCNet [29] also adopt a serial structure, but GCNet lets SA replace the pooling operations in CA, reducing the spatial information lost by channel attention when integrating channel information. DANet [30] treats CA and SA as two parallel branches, then fuses them through convolution layers and element-wise summation. In our work, we design CA and SA as independent modules for efficiency and then connect them in series.

3 Proposed Method

3.1 Revisiting the CSP Detector

The CSP detector is based on the idea of anchor-free detection, which can abandon the complexity of the anchor boxes and sliding windows design. It regards pedestrian detection as a high-level semantic feature detection task, and uses the pedestrian center point and height as abstract high-level semantic information features.

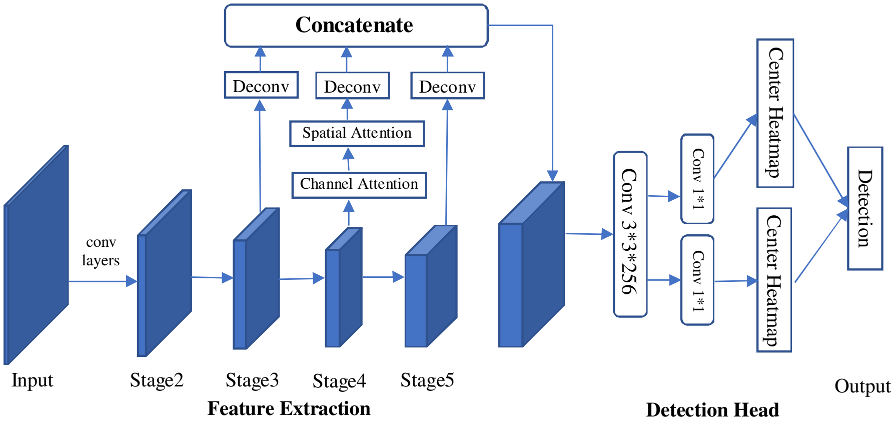

Fig. 1. The architecture after embedding our hybrid attention model into CSP.

The CSP framework consists of two modules: feature extraction module and detection head. For feature extraction, CSP uses ResNet-50 and MobileNet as backbone networks. Taking ResNet-50 as an example, its convolutional layer can be divided into 5 stages with down-sampling rates of the input image of 2, 4, 8, 16, and 16, respectively. The fifth stage uses dilated convolution to keep the size of the feature map unchanged and increase the receptive field of the convolution layers. Then the CSP uses deconvolution to up-sample the feature maps of the stage 3, 4, and 5 to the same size as stage 2, and concatenate them to feed into the detection head. For detection head, CSP reduces the channel dimension through a $3 * 3$ convolution layer, and then uses the two $1 * 1$ convolution layers to generate two branches of the heat map that predict the center point and scale, respectively. Although CSP achieves the state-of-the-art performance of pedestrian detection at that time, we demonstrate that it still can achieve better detection results when enhanced with more discriminative information on pedestrians and backgrounds through our hybrid self-attention module, as shown in Fig. 1. Note that our attention module is added after stage 4 of the CSP backbone network, because stage 4 has enough information for learning channel discrimination and spatial context.

3.2 Channel Attention

Channel attention is designed to help the network acquire inter-dependencies between channels. The first step for Channel attention is compresses spatial dimensions to simplify the calculation of global context information integration for spatial. However, unlike global average pooling adopted in SENet [18] that preserves background information, global maximum pooling can extract pedestrian texture information that is helpful for detection. So we use both global average pooling and global maximum pooling for global context information integration. Different from SENet [18], which uses two fully connected layers

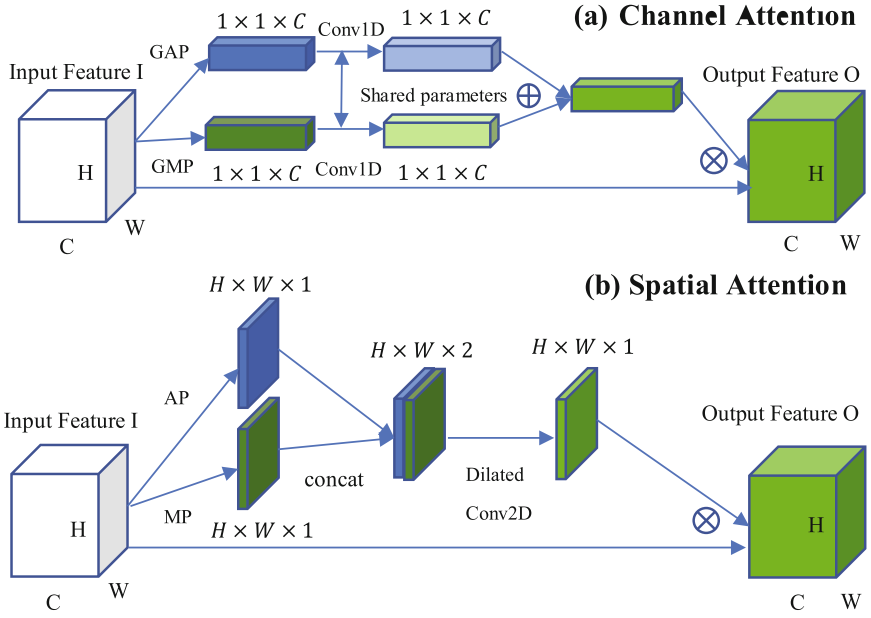

Fig. 2. The architecture of our channel attention and spatial attention.

to obtain the internal dependencies between all channels, we believe that this relationship can be calculated only locally in deep networks because of the huge differences between non local channels in deep layers. So we just use a single 1D convolution layer with a larger kernel size to capture this relationship. This can make the design of the attention module more concise, while reducing the amounts of parameters. Finally, we rescale the input features by obtaining the weighted vector from the previous step. Figure 2(a) shows the specific processing flow of our channel attention module.

Giving an input feature map as $I \in R^{(H \times W \times C)}$, we use two branches to obtain global context information from spatial by global average pooling (GAP) and global maximum pooling (GMP), then we generate two one-dimensional vectors: $I_{GAP} \in R^{(1 \times 1 \times C)}$ and $I_{GMP} \in R^{(1 \times 1 \times C)}$. We use these two vectors to pass a same 1D convolution layer with kernel size of $9 * 9(Conv1D^9)$ meanwhile keep the dimensions unchanged. This shared convolution layer can reduce the amounts of parameters and the possibility of overfitting. Then we add the two vectors by element-wise summation and go through the sigmoid activation function, and finally multiply the merged attention vector with the input feature map to get the final output $O \in R^{(H \times W \times C)}$. This process can be expressed by the following formula:

$$O = I \bigotimes (S(Conv1D^9(I_{GAP}) \bigoplus Conv1D^9(I_{GMP}))) \tag{1}$$

where S means sigmoid activation function, $\bigotimes$ and $\bigoplus$ represent multiply per channel and element-wise summation. Note that we do not use activation functions in Conv1D layer to ensure that channel information is not lost or suppressed during the global information transfer process.

3.3 Spatial Attention

Spatial attention is used to obtain long-range dependencies between pixels, it can also complement the channel attention information to help pedestrian detection. Similar to the channel attention we designed, we do average pooling and maximum pooling for each pixel along the channel axis, this operation can obtain the abstraction of global context information in the channel dimension, then we concatenate the two pooled feature maps together for subsequent convolution layers. Different from Non-Local Block [24], which calculates the relationship between any two pixels, we intend to expand spatial contextual information and receptive field to distinguish the pedestrian from the background in high-level features. Dilated convolution is a very intuitive and lightweight way to achieve this purpose. Meanwhile, we stack multiple dilated convolutions to capture multi-scale spatial contextual information under different receptive fields. At last, we use the obtained weighted feature map to recalibrate the input features. Figure 2(b) shows the specific details of our spatial attention module.

Giving an input feature map as $I \in R^{(H \times W \times C)}$, we use average pooling and maximum pooling along the channel axis to obtain two feature maps $I_{AP} \in R^{(H \times W \times 1)}$ and $I_{MP} \in R^{(H \times W \times 1)}$ containing global spatial information. Then the two feature maps are concatenated together and pass 3 layers of dilated convolution, the filters are all set to 1, the kernel sizes are all set to 3 * 3 and the dilated rates are 1, 2, and $5(DilaConv2D^{1,2,5})$, respectively. Finally we compute the output feature map $O \in R^{(H \times W \times C)}$ by multiplying the obtained attention map and the input feature map. Spatial attention can be calculated by the following formula:

$$O = I \bigotimes (S(DilaConv2D^{1,2,5}(I_{AP}, I_{MP}))) \tag{2}$$

where [,] means the concatenation of two feature maps, $\bigotimes$ represent multiply per pixel.

3.4 Hybrid Attention Fusion Strategy

Fusion of channel attention and spatial attention helps aggregate and complement global context information in different dimensions. Since our channel attention module and spatial attention module are two independent embeddable models with the same input and output, a very efficient fusion strategy is to connect two modules in the order of either channel attention first or spatial attention first. Another way of fusion is parallel structures, where channel attention and spatial attention modules are processed separately and added by element-wise summation. The experimental evaluation demonstrates that the first fusion strategy is better than the second one.

4 Experiments

4.1 Dataset and Evaluation Metrics

The Caltech Pedestrians Benchmark [31] is one of the largest and most widely studied pedestrian detection datasets. It contains a 10-h of 30HZ 640 * 480 vehicle driving video through regular traffic in an urban environment. This dataset has a total of 350,000 bounding boxes for about 2300 unique pedestrians. For the training set, we extract a picture from set00-05 every 3 frames (42,782 images in total). For the test set, we extract a picture from set06-10 every 30 frames (4024 images in total). Note that during training and testing phases we use the Caltech new annotations provided by [4].

We follow the average log missing rate(MR) of false positive per image(FPPI) between 10^{-2} and 10^{0} as the evaluation standard. Particularly, the pedestrians from three subsets of Reasonable, All and Heavy Occlusion are used for evaluation, and they are set as follows: The pedestrians height range for the three subsets are [50, inf], [20, inf], [50, inf]. (inf means infinite). The visible portion of pedestrians are [0.65, inf], [0.2, inf], and [0.2, 0.65] respectively.

4.2 Ablation Study

Channel Attention. The key to the design of channel attention is the size of the convolution kernel in the one-dimensional convolution, because it represents how wide the correlation is obtained between channels. Table 1 shows the influence of different convolution kernel sizes on channel attention performance. The first row of Table 1 is the original MR on the Reasonable set in the CSP [11] paper, while the second row is our own reproduced result by using its opensource code[1]. It's worth noting that our reproduced results are always worse than the original results in the CSP [11] paper by about 0.5 % on the Reasonable set. Therefore in the subsequent experiments, we will take our reproduced results as baseline for fair comparisons. From the results in Table 1, we can observe that as the size of the convolution kernel increases, a larger range of channel interdependencies can be obtained, and the performance of channel attention continues to get improved. In addition, different kernel sizes have improved pedestrian detection performance, validating the design rationality of our channel attention module.

Spatial Attention. The setting of the dilated rates affects the scope of contextual information acquisition in spatial. As shown in Table 2, we set up three combinations of dilated rates to verify its influence on spatial attention. We first set the dilated rates to 1-2-1, which slightly improves detection performance. Then we further expand the dilated rates to 1-2-5 and 5-2-1 in a way of increasing and decreasing. These two ways prove that getting more contextual information and a larger receptive field through a larger dilated rates setting is crucial to

[1] https://github.com/liuwei16/CSP.

Table 1. Comparison of different convolution kernel sizes for channel attention.

Algorithm	Reasonable set	
	$MR(\%)$	$\Delta MR(\%)$
CSP [11]	4.54	–
CSP (reproduced)	5.02	–
CSP+CA (k5)	4.94	+0.08
CSP+CA (k7)	4.85	+0.17
CSP+CA (k9)	4.37	+0.65

Table 2. Comparison of different dilated rate settings for spatial attention.

Algorithm	Reasonable set	
	$MR(\%)$	$\Delta MR(\%)$
CSP [11]	4.54	–
CSP (reproduced)	5.02	–
CSP+SA (dr 1-2-1)	4.84	+0.18
CSP+SA (dr 1-2-5)	4.42	+0.60
CSP+SA (dr 5-2-1)	4.43	+0.59

improve the performance of spatial attention. However, setting the dilated rates either in increasing way or in decreasing way has little affect on the results.

Fusion Strategies. After fixing the size of the one-dimensional convolution kernel to 9 and the dilated rates setting to 1-2-5, we further study the effects of different fusion strategies for integrating channel attention and spatial attention. As seen the results in Table 3, the sequential combination is significantly better than the parallel combination. This may be due to the fact that channel attention and spatial attention give weights to features in different dimensions, and adding them directly will cause chaos in the weight distribution. For the sequential fusion structure, we find that putting channel attention first is better than putting spatial attention first, thus in subsequent experiments, we will apply sequential fusion strategy in the order of first channel attention and then spatial attention. Moreover, we can see that the results of sequential fusion are better than the result of either channel attention or spatial attention alone, proving that the proposed hybrid attention model can provide complementary information of two single attention module and further improve pedestrian detection performance.

4.3 Comparison with State of the Arts

Finally, we compare the proposed method with several state-of-the-art pedestrian detection approaches on three Caltech subsets of Reasonable, All, and

Table 3. Comparison of different hybrid attention fusion strategies.

Algorithm	Reasonable set	
	$MR(\%)$	$\Delta MR(\%)$
CSP [11]	4.54	–
CSP (reproduced)	5.02	–
CSP+CA+SA	3.84	+1.18
CSP+SA+CA	4.01	+1.01
CSP+CASA (add)	4.75	+0.27

Table 4. Comparison with state of the arts (the best results are in bold).

Algorithm	$MR\%$		
	Reasonable	All	Heavt Occlusion
CSP [11]	4.5	56.9	45.8
CSP (reproduced)	5.0	57.9	49.1
RPN+BF [12]	7.3	–	54.6
ALFNet [6]	6.1	59.1	51.0
RepLoss [8]	5.0	59.0	47.9
OR-CNN (city) [7]	4.1	58.8	**45.0**
Ours	**3.8**	**56.8**	46.4

Heavy Occlusion. Note that we only compare algorithms trained and tested with the Caltech new annotations [4], the results are shown in Table 4. On Reasonable subset, compared with our reproduced result, we have significantly improved the performance by 24% after embedding our hybrid attention module. Even compared with the original results in [11], we still achieve a 15% improvement.

The performances of the proposed method are also better than other state-of-the-art algorithms, even better than OR-CNN [7] for 4.1% MR, which is a method for pre-training with additional Citypersons dataset [32]. For All subset, our hybrid attention module also achieved the best results, showing that it can guide the network to enhance detection performance in various sizes of pedestrians. Considering that our reproduced results on the Heavy Occlusion subset are far from the results in the CSP paper, we achieved slightly lower results after embedding the proposed attention module. However, our results are still competitive with other state-of-the-arts. Since we have significantly improved the detection performance on heavy occlusion pedestrians compared with the reproduced baseline result, it has been revealed that our method can pay more attention to the feature representation of pedestrians' visible parts.

5 Conclusion

In this paper, we propose a novel hybrid attention model for pedestrian detection from both channel attention and spatial attention aspects to obtain feature inter-dependencies in different dimensions. Our model can provide the detection network with more discriminative guidance information for pedestrians and backgrounds to enhance the capabilities of pedestrians' feature representation. By embedding the proposed attention module into the CSP baseline framework, the detection performance has been further improved on the standard Caltech pedestrian detection benchmark. The ablation study demonstrates the effectiveness of the proposed channel attention, spatial attention, and their hybrid fusion strategy. Our model also achieves superior performances than several state-of-the-art algorithms on the subsets named Reasonable, All, and Heavy Occlusion of the Caltech benchmark.

Acknowledgment. This work was supported by National Natural Science Foundation of China under Grant 61703039, and Beijing Natural Science Foundation under Grant 4174095.

References

1. 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)
2. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271 (2017)
3. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)
4. 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)
5. Du, X., El-Khamy, M., Lee, J., et al.: Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE 953–961 (2017)
6. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 618–634 (2018)
7. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware R-CNN: detecting pedestrians in a crowd. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 637–653 (2018)
8. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 7774–7783 (2018)
9. 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)

10. Pang, Y., Xie, J., Khan, M.H., Anwer, R.M., Khan, F.S., Shao, L.: Mask-guided attention network for occluded pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4967–4975 (2019)
11. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: a new perspective for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 5187–5196 (2019)
12. Zhang, L., Lin, L., Liang, X., et al.: Is Faster R-CNN doing well for pedestrian detection?. In: European Conference on Computer Vision, pp. 443–457. Springer, Cham (2016)
13. Cai, Z., Fan, Q., Feris, R.S., et al.: A unified multi-scale deep convolutional neural network for fast object detection. In: European Conference on Computer Vision, pp. 354–370. Springer, Cham (2016)
14. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4950–4959 (2017)
15. Liu, S., Huang, D., Wang, Y.: Adaptive nms: refining pedestrian detection in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 6459-6468 (2019)
16. Zhang, J., Lin, L., Chen, Y., et al.: CSID: center, scale, identity and density-aware pedestrian detection in a crowd[OL]. arXiv preprint arXiv:1910.09188 (2019)
17. Zhou, C., Wu, M., Lam, S.K.: SSA-CNN: Semantic self-attention CNN for pedestrian detection[OL]. arXiv preprint arXiv:1902.09080 (2019)
18. 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)
19. Hu, J., Shen, L., Albanie, S., Sun, G., Vedaldi, A.: Gather-excite: exploiting feature context in convolutional neural networks. In Advances in Neural Information Processing Systems. pp. 9401–9411 (2018)
20. Dongsheng, R., Jun, W., Nenggan, Z.: Linear context transform block. arXiv preprint arXiv:1909.03834 (2019)
21. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 510–519 (2019)
22. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. pp. 2017–2025 (2015)
23. Wang, F., Jiang, M., Qian, C., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2017)
24. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
25. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 603–612 (2019)
26. Yue, K., Sun, M., Yuan, Y., et al.: Compact generalized non-local network. In: Advances in Neural Information Processing Systems. pp. 6510–6519 (2018)
27. Chen, Y., Kalantidis, Y., Li, J., et al.: A^2-nets: Double attention networks. In: Advances in Neural Information Processing Systems. pp. 352–361 (2018)
28. Woo, S., Park, J., Lee, J.Y., et al.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)

29. Cao, Y., Xu, J., Lin, S., et al.: GCNet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)
30. Fu, J., Liu, J., Tian, H., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
31. Dollar, P., Wojek, C., Schiele, B., et al.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2011)
32. 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)