2022 11th International Conference of Information and Communication Technology (ICTech) | 978-1-6654-9694-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICTECH55460.2022.00096

Towards Better Pedestrian Detection Using Multi-Scale CSPN and Dual Attention

XinXin Huang[1,2,3]
[1]University of Chinese Academy of Sciences, Beijing 100049, China
[2]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China
[3]Liaoning Key Laboratory of Domestic Industrial Control Platform Technology on Basic Hardware & Software, Shenyang 110168, China

ZhenYu Yin[2,3,*]
[2]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China
[3]Liaoning Key Laboratory of Domestic Industrial Control Platform Technology on Basic Hardware & Software, Shenyang 110168, China
congmy@163.com

Chao Fan[1,2,3]
[1]University of Chinese Academy of Sciences, Beijing 100049, China
[2]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China
[3]Liaoning Key Laboratory of Domestic Industrial Control Platform Technology on Basic Hardware & Software, Shenyang 110168, China

***Abstract*—Pedestrian detection is a significant research direction in the computer vision, but the detection performance of existing pedestrian detection algorithms is inadequate. Therefore, this article proposes a novel algorithm to improve the anchor-free pedestrian detection algorithm. First, the multi-scale CSPN module is used to deepen the network depth, further extract semantic information on multiple scales, and improve detection performance. Moreover, the dual attention module based on feature fusion is used to effectively fuse features of different scales, assigning new weights to the fused features in the two dimensions of space and channel. Experiments show our method reduces MR^{-2} by 0.10%, 2.60% and 0.98% on the Reasonable, Heavy Occlusion and ALL of the Caltech pedestrian dataset, which is better than the existing algorithms.**

Keywords-pedestrian detection; multi-SCALE; dual attention; CSP; Anchor-free

I. INTRODUCTION

Pedestrian detection is a significant research direction in the computer vision. It is applied to road monitoring and robotics, and is an indispensable part of other computer vision tasks (person re-identification, pedestrian tracking). Pedestrian detection is to judge whether there is a pedestrian in the picture or the video, and locate the position of the pedestrian.

Traditional pedestrian detection methods mainly use manual feature extraction. The main representative methods include HOG[1](Histogram of Oriented Gradients), ICF[2](Integral Channel feature), DPM[3](Deformable part Models). However, it is vulnerable to the external environment and has weak robustness. Moreover, it has a poor detection effect on pedestrians with different poses, so the detection effect is not good.

Since AlexNet[4] network was proposed in 2012, the Deep Convolutional Neural Networks (DCNNs) method has been used for computer vision, achieving success in many fields, such as image classification and object detection. With the triumph of deep learning in object detection, scholars began to apply DCNNs to pedestrian detection. Many pedestrian detection methods based on DCNNs emerged, opening a new chapter in pedestrian detection.

At present, there are two classification methods for pedestrian detection based on DCNN. One classification method is classified as one-stage and two-stage methods, which are anchor-based methods. The two-stage methods are also called region-based convolutional neural Networks (RCNNs). Such methods first use CNN to generate feature maps, then Region Proposal Network (RPN) generate many candidate regions, and finally use CNN for further classification and regression to complete pedestrian detection. Typical representatives include the R-CNN series[5-7], Mask R-CNN[8]. Unlike the two-stage method, the one-stage methods do not use RPN to generate candidate regions, which means these one-stage methods have a higher detection speed, but the detection accuracy is sacrificed simultaneously. The one-stage methods directly use feature maps generated by CNN to predict object confidence and bounding boxes. Typical representatives include YOLO[9], SSD[10], RetianNet[11], etc.

Another classification method is divided into anchor-based and anchor-free methods. The setting of the anchor has a great influence on the performance of anchor-based methods. A good anchor setting can greatly improve the detection performance, but a bad anchor setting will negatively affect the detection performance. Furthermore, different datasets need to set different anchors. Because of this, CornerNet[12] proposes an anchor-free approach that does not require anchors. CSP[13] (Center and Scale Prediction), CenterNet[14], and FoveaBox[15] are typical examples of this approach. In comparison to anchor-based methods, anchor-free methods are more flexible because they do not need an additional anchor super-parameter and have

978-1-6654-9694-0/22/$31.00 ©2022 IEEE
DOI 10.1109/ICTech55460.2022.00096

stronger applicability to all datasets. However, anchor-free methods have the problem of low detection accuracy.

The existing pedestrian detection algorithms have various problems, and the detection performance is not ideal. Therefore, this paper is based on an anchor-free pedestrian detection algorithm CSP. On this basis, the multi-scale CSPN (Cross Stage Partial Network) module and the dual attention module based on feature fusion are used to improve CSP algorithm and pedestrian detection performance. The chief contributions of this article include:

- Using the multi-scale CSPN module to deepen the network depth, extracting multi-scales semantic information, and improving the performance of network detection.
- Introducing the dual attention module based on feature fusion, effectively fusing features of different scales, assigning new weights to the fused features in the two dimensions of space and channel, enhancing the distinguishability and robustness of features.
- Obtaining 43.21% MR^{-2} on the Caltech dataset.

II. Related works

A. CSP network

CSP algorithm is the first application of anchor-free to the field of pedestrian detection, which solves the problem of designing anchors for specific datasets in previous anchor-based pedestrian detection algorithms. It reduces the calculation of related parameters of the anchor. The CSP algorithm considers pedestrians' center points and scales as a high-level semantic feature, and pedestrian detection can be transformed into the detection of semantic features.

The network structure of CSP algorithm consists of the feature extraction and the detection head part. In feature extraction part, the backbone network is ResNet50. Deconvolution and normalization are used on the 3rd, 4th and 5th stage feature maps of ResNet50 to increase the size of the feature map to the same size as the second stage feature map. Then, channel concatenate is carried out to generate the feature maps for detection. In the detection head part, the 3×3 convolution is first used to reduce dimension and computation, and then use three 1×1 convolutions to predict object centre coordinates, scale size and coordinate offset, respectively, and finally generate pedestrian detection frame.

B. CSPN structure

As neural networks become deeper and wider, the performance of the neural network becomes more and more powerful, but with it comes the increase in the amount of calculation. Sometimes, computation costs are higher than the benefit of performance improvement, so it is necessary to measure between the two. Therefore, CSPNet[16] proposes a architecture, which can maintain or even improve network performance while reducing computation by 20%. CSPNet believes that the repetition of gradient information in network optimization will lead to a large amount of inference calculations, so the feature map integrates the gradient changes, reducing the number of calculations and ensuring accuracy. The amount of calculation can be reduced, and at the same time, richer gradient combination information can be achieved. This paper draws on the structure of CSPNet to construct the multi-scale CSPN module, and while deepening the network depth, it obtains more semantic information at a smaller computational cost.

C. Attention mechanism

In cognitive science, when humans observe an object, they will only attend to important pieces of information and ignore others, which is often called the attention mechanism. In computer vision, it has been popularly used in various domains, and the attention mechanism model's performance is greatly improved. SENet[17] performs feature recalibration in the channel dimensions by learning the correlation between channels. CBAM[18] proposed a dual attention mechanism that connects channel and spatial attention in series. BAM[19] proposes a dual attention mechanism that connects dual attention in parallel. The attention mechanism will help network pay attention to important information and ignore irrelevant information to learn features that can help improve network performance. Influenced by the above methods, this paper introduces the dual attention module based on feature fusion, which effectively integrates multi-scale features, distributes the weight of features in spatial and channel dimensions so that the feature map can be adjusted adaptively.

III. The proposed method

A. Overall architecture

Figure 1 is the overall architecture of this paper. The backbone network uses ResNet50. The model is composed of the detection head and the feature extraction two parts. Similar to the CSP algorithm, the detection head part first uses 3×3 convolution to decrease the dimensionality of the feature map generated by the feature extraction part and uses three parallel 1×1 convolutions to predict the center coordinates, scale information and offset of the coordinates, respectively. ResNet50, the multi-scale CSPN module and the dual attention module based on feature fusion constitute the feature extraction part. First, the image is input into ResNet50 for feature extraction. ResNet50 consists of five layers. The feature map output from the second, third, fourth and fifth layers are denoted as σ_2, σ_3, σ_4, σ_5 respectively. σ_3, σ_4 and σ_5 are fed into the multi-scale CSPN module to generate feature map η_3, η_4, η_5 at different scales. For η_3, η_4, η_5, the resolution is improved to be consistent with σ_2 by deconvolution and normalization. The feature maps after deconvolution are fed into the dual attention module based on feature fusion for feature fusion and weight adjustment of channel and space, and the feature map σ_{conv} is generated. Finally, σ_{conv} is sent to the detection head part for pedestrian detection. Next, the multi-scale CSPN module and

the dual attention module based on feature fusion are introduced in detail.

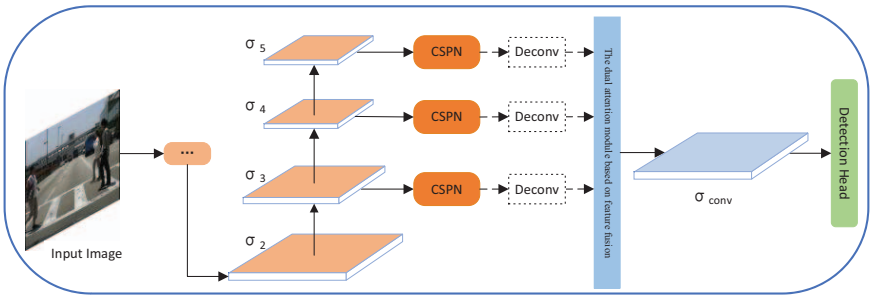


Figure 1. Overall architecture.

B. *The multi-scale CSPN module*

In the neural network, the deep feature map contains more semantic information. CSP algorithm is essentially a detection of high-level semantic features, while ResNet50 lacks the semantic information from the input image. Since strengthening network depth can generate more semantic information, the multi-scale CSPN module is added after the feature maps σ_3, σ_4, σ_5 (where the sizes of σ_4 and σ_5 are the same) to deepen the network depth and further extract multiple scales. The semantic information on the network strengthens the network's ability to detect pedestrians of different scales.

The network structure of the multi-scale CSPN module is shown in Figure 2(a). After the feature maps of different scales σ_3, σ_4, σ_5, the CSPN is added to generate feature maps η_3, η_4, η_5. The main component of the CSPN structure is CBM, shown in Figure 2(c). It composes Conv, BN and Mish activation functions. The formula of Mish is:

$$Mish = x * tanh\left(ln\left(1+e^{x}\right)\right) \tag{1}$$

The structure of a single CSPN is shown in Figure 2(b). First, it passes through a 3×3 CBM to generate a new feature map. Then divide the number of channels by two 1×1 CBMs to generate two feature maps Q and F, which improves the reusability of features and reduces the amount of calculation. After 3×3 CBM and 3×3 convolution operation of feature map Q, the generated feature map and the feature map G perform an element-wise addition operation to generate a feature map T. The feature map T and F concatenate on the channel dimension. Finally, a 3×3 CBM is used to generate the output feature map.

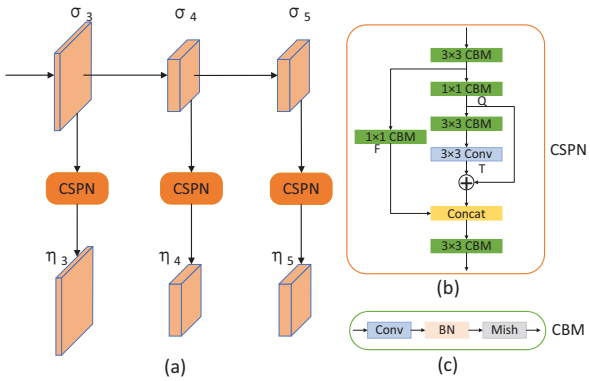


Figure 2. (a) The multi-scale CSPN module; (b) CSPN; (c) CBM.

C. *The dual attention module based on feature fusion*

After the multi-scale CSPN module, the multi-scale feature maps contain richer semantic information. However not all the information is helpful to the performance of pedestrian detection, and some information will also cause the performance of pedestrian detection to decrease. Therefore, to let the model pay more attention to useful information for pedestrian, the dual attention module based on feature fusion is designed to fuse multi-scales features and improve feature reusability effectively. Additionally, the dual attention module conducts reweight operations in the spatiality and channel, improving the ability of the pedestrian detector to detect pedestrians.

The main structure of the dual attention module based on feature fusion is shown in Figure 3(a). First, the feature maps generated after deconvolution and normalization of the feature maps η_3, η_4, η_5 concatenate on the channel dimension. The feature map P_1 is generated after feature fusion, which can improve feature reusability. Then the feature map P_1 is sent to the Channel Attention Module to generate channel attention feature, that is, M_C. The element-wise multiplication operation of M_C and the feature map P_1 is performed to generate the feature map P_2, and the appropriate weights are assigned to the features of different scales in the channel dimension. Finally, the feature map P_2 is fed into the Spatial Attention Module to generate the spatial attention feature, namely M_S. M_S and the feature map P_2 are element-wise multiplied to generate the feature

map σ_{conv}, and the features of different scales in the spatial dimension give the appropriate weight. Figure 3(b) and 3(c) respectively show the structure of Channel Attention Module and Spatial Attention Module.

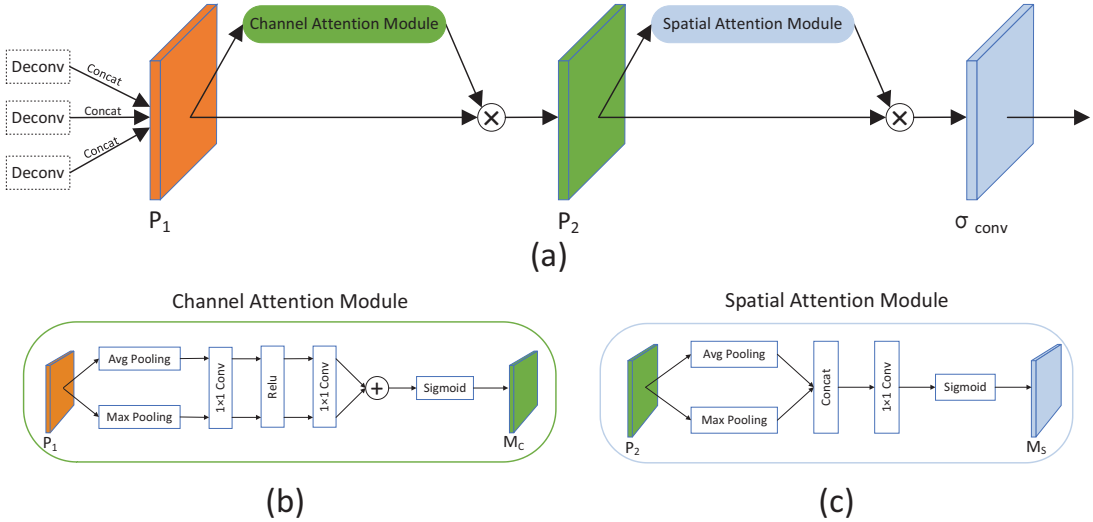


Figure 3. (a) The dual attention module based on feature fusion; (b) Channel Attention Module; (c) Spatial Attention Module.

IV. EXPERIMENTS

A. Dataset

This paper uses the Caltech pedestrian dataset for experiments. Caltech is currently the popular pedestrian datasets. Caltech contains 11 video sets captured by the vehicle-mounted camera, with a video resolution of 640×480. The set00~set05 are used for training, and the set06~set10 are used for testing. 42782 images are taken as the training set, and 4024 images are taken as the test set. Our algorithm is evaluated on three settings: Reasonable, Heavy Occlusion and ALL.

1.1. Evaluation metric

This paper uses log-average missing rate MR^{-2} as the evaluation metric for pedestrian detection. The lower the metric value is, the better the detection performance our network has. To calculate MR^{-2}, first draw the $FPPI-MR$ curve. The calculation formulas of $FPPI$ and MR are as follows. Then 9 values of $FPPI$ evenly distributed in logarithmic space $\left[10^{-2},10^{0}\right]$ are taken, and the corresponding MR values of these 9 values are averaged. The calculation methods of $FPPI$ and MR are as follows:

$$MR=\frac{FN}{TP+FN} \tag{2}$$

$$FPPI=\frac{FP}{N} \tag{3}$$

Where FN is the number of negative cases in the positive sample , TP is the number of positive cases in the positive sample, FP is the number of positive cases in the negative sample, and N is the number of pictures.

B. Training details

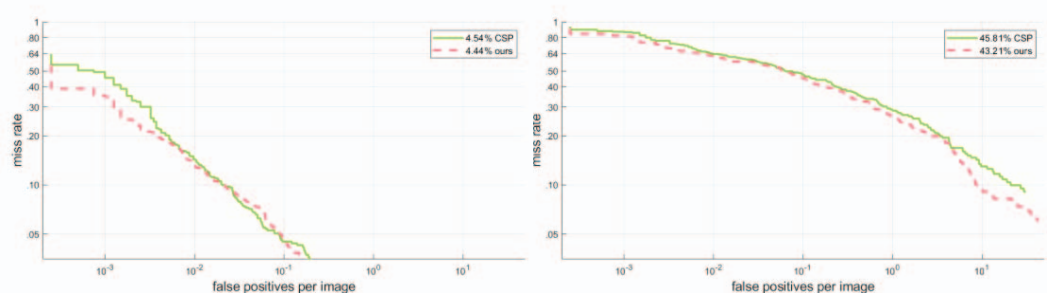
The experiment uses the PyTorch to implement the algorithm in this article. Adam is applied, and the moving average weights[20] is applied to gain more stable training. The backbone is ResNet50, pre-trained by ImageNet. For Caltech, the image resolution is scaled to 336×448 during training, and the batch size is set to 16, the learning rate is set to 10^{-4}, and the training stops after 15k iterations. The inference is made on the original picture resolution of 480×640.

C. Experimental results and analysis

To prove the validity of our method, we first compare it with the CSP algorithm on Caltech. The $FPPI-MR$ curves of the two algorithms are shown in Figure 4, where 4(a), 4(b), and 4(c) respectively represent the detection results in Reasonable, Heavy Occlusion and ALL. The occlusion ratio of the Reasonable does not exceed 35%, and the occlusion ratio of Heavy Occlusion is 35%~80%. The experimental results indicate that our proposed algorithm improved the original algorithm on Heavy Occlusion, and MR^{-2} is reduced by 2.60%. That is because pedestrians with a large occlusion ratio need more advanced semantic representations. The multi-scale CSPN module deepens the network depth and helps the network extract multi-scale pedestrian semantic representations. At the same time, it cooperates with the dual attention module based on feature fusion in the two dimensions of space and channel. To suppress redundant semantic information. Compared with Heavy Occlusion, MR^{-2} of Reasonable is reduced from 4.54% to 4.44%. That is because the CSP algorithm has achieved excellent detection results on this subset, so the improvement is tiny.

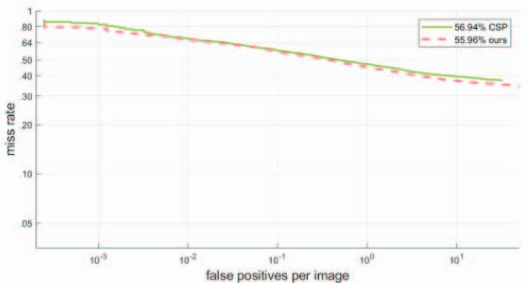
On ALL, MR^{-2} is reduced by 0.98%, indicating that effectively enhance the detection performance of our model.

Finally, our method is compared with the current state-of-the-art model on Caltech, including SDS-RCNN[21], MS-CNN[22], RepLoss[23] and ALFNet[24]. The experimental comparison results are shown in Table I. It demonstrated that our algorithm had achieved better results on Caltech. Our algorithm achieves better results because we use an anchor-free algorithm to convert pedestrian detection into semantic feature detection and uses high-resolution feature maps, which are less affected by occlusion and scale. In addition, the use of the multi-scale CSPN module to extract multi-scale semantic information helps the network detect pedestrians at different scales, and the dual attention module based on feature fusion distributes better weight to the feature that is conducive to detection in the channel and spatial dimensions.



(a)Reasonable

(b)Heavy Occlusion



(c)ALL

Figure 4. FPPI-MR curve.

TABLE I. COMPARISONS WITH THE STATE-OF-THE-ART ON CALTECH.

Model	Reasonable	Heavy Occlusion	ALL
RPN+BF	9.58	74.36	64.70
MS-CNN	9.95	59.94	60.90
SDS-RCNN	7.36	58.55	61.50
ALFNet	6.10	51.00	59.10
RepLoss	5.00	47.90	59.00
CSP	4.54	45.81	56.94
ours	**4.44**	**43.21**	**55.96**

V. CONCLUSIONS

Aiming at the lack of performance of existing pedestrian detection algorithms, two improvements are made to the anchor-free based CSP algorithm. First, the multi-scale CSPN module is used to deepen the network depth to extract semantic information on multiple scales, enabling better network detection performance. Secondly, the dual attention module based on feature fusion is used to effectively fuse features of different scales, adjust the fused features in the dimensions of spatiality and channel, and enhance the distinguishability and robustness of features. Compared with the original algorithm, the proposed method performs better on Caltech, especially on the Heavy Occlusion, MR^{-2} is reduced to 43.21%, which verifies the validity of our method. In future work, it is planned to conduct research from the backbone and consider proving the availability of our method on other pedestrian detection datasets.

ACKNOWLEDGMENT

This paper is supported by National Key R&D Program of China – Research on Key Technologies of Real-time Fault Diagnosis in Intelligent Production System Based on Industrial IoT (2017YFE0125300).

REFERENCES

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[2] Dollár P, Tu Z, Perona P, et al. Integral channel features[J]. 2009.

[3] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(9): 1627-1645.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[8] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[10] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[12] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[13] Liu W, Liao S, Ren W, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5187-5196.

[14] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6569-6578.

[15] Kong T, Sun F, Liu H, et al. Foveabox: Beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[16] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[18] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[19] Park J, Woo S, Lee J Y, et al. Bam: Bottleneck attention module[J]. arXiv preprint arXiv:1807.06514, 2018.

[20] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv preprint arXiv:1703.01780, 2017.

[21] Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection & segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4950-4959.

[22] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//European conference on computer vision. Springer, Cham, 2016: 354-370.

[23] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7774-7783.

[24] Liu W, Liao S, Hu W, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 618-634.