Visible Region Enhancement Network for Occluded Pedestrian Detection

Fangwei Sun[1], Caidong Yang[1], Chengyang Li[1,2], Heng Zhou[1,3], Ziwei Du[1], Yongqiang Xie[1,*], and Zhongbo Li[1,*]

[1] *Institution of Systems Engineering, Academy of Military Sciences, Beijing, 100141, China*
[2] *School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*
[3] *School of Electronic Engineering, Xidian University, Xi'an, Shanxi, 710071, China*

Abstract

Occlusion is a big challenge in detecting pedestrians. In this paper, we propose a new network module named Visible Region Enhancement Network(VREN), which is consisted of a spatial attention network and a channel attention network. Given feature maps, our module infers attention maps from two dimensions, spatial and channel. In particular, compared with the previous attention mechanism, the acquisition of the two kinds of attention in VREN is interrelated, rather than independent. Based on attention maps, VREN can enhance the effective feature from different dimensions, while reducing the interference noise. Because VREN works in the feature extraction stage, it can be integrated into any Convolutional Neural Network(CNN) architecture and is end-to-end trainable along with base CNNs. We validate our VREN through extensive experiments on CrowdHuman datasets. Our experiments show VREN can effectively increase detection performances compared to the Faster R-CNN baseline.

Keywords

Pedestrian Detection; Occlusion Detection; Spatial Attention; Channel Attention

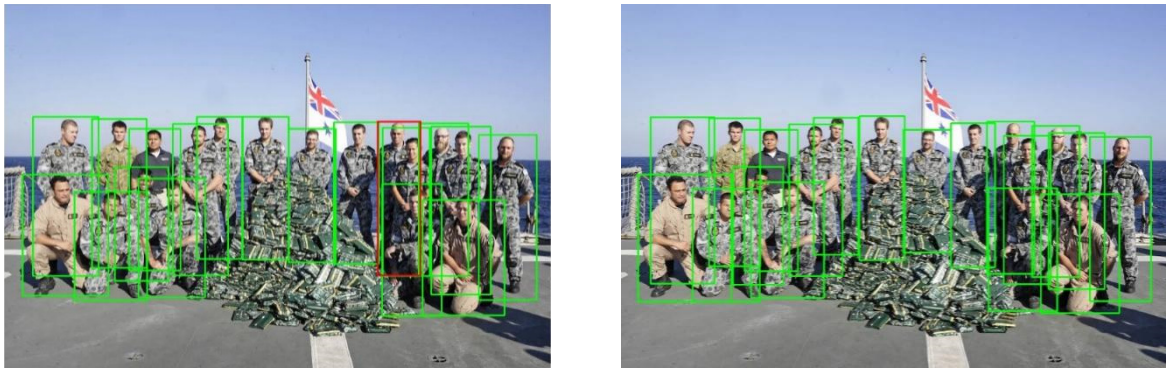
1. Introduction

Pedestrian detection, as a branch of object detection, is an important task in computer vision. It is widely used in various fields, such as autonomous driving, object tracking, video surveillance, and many other fields. In recent years, with the development of deep learning, especially CNN, the performance of pedestrian detection has obtained rapid improvement. According to the different generation modes of proposals, the CNN frames can be roughly divided into two types: one-stage detector[1][2][3][4] without independent to generate proposals, and two-stage detector[5][6][7][8][9][10] with independent network generating proposals. In contrast, the one-stage detector has a faster detection speed but a lower detection accuracy, while the two-stage detector has a higher detection accuracy but a slower detection speed. These advanced detectors have greatly promoted the research of pedestrian detection and made great breakthroughs.

However, in the real world, it is very common for the pedestrian to occlude each other or be occluded by other objects, which cause the body is not fully visible. The difficulties of occluded object detection are as follows: (a) Because of the influence of the datasets and the complexity of occlusion, Fawzi and Frossard[11] proved occlusion detector which based on CNN is not robust. (b) Occlusion interference feature extraction and occlusion of each other two objectives are likely to have very similar characteristics, which cause the detector cannot predict accurately distinguish. (c) During occlusion, the prediction boxes of different objects may be seriously overlapped, so the prediction boxes of different object may be regarded as the prediction of one object by the non-maximum suppression(NMS) algorithm, and the false suppression will lead to missed detection. From the above analysis, occlusion remains a big challenge in detecting pedestrians.

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China
*Corresponding author's e-mail: xyq_ams@outlook.com (Yongqiang Xie); lzb_ams@outlook.com (Zhongbo Li)

To handle occlusion, an effective solution is to use attention mechanism. Attention mechanisms not only tell us where to focus, but it also improve the representation of target feature information. In this paper, we propose a new network module, named "Visible Region Enhancement Network". Since CNN extract features by blending cross-spatial and channel information together, we adopt our module to emphasize meaningful features along spatial and channel dimensions. In addition, the two kinds of attention acquisition are closely related. As a result, our module efficiently helps the feature information transfer within the network by learning which information to enhance or suppress. Fig. 1 (a) shows the results predicted by Faster R-CNN[7] baseline: the detector fails to predict instances heavily overlapped with others. Fig. 1 (b) shows the prediction results of our method. In particular, our method also improves positioning accuracy.



(a) Baseline (b) Ours

Figure 1. Human detection in crowds. (a) Results predicted by Faster R-CNN[7] baseline. The red box indicates the missed detection. (b) Results of our method. All instances are correctly predicted.

In the CrowdHuman datasets[12], we obtain accuracy improvement from the baseline network by plugging our module, proving the efficacy of VREN. Since our module is designed to work in the feature extraction stage, in theory, both the one-stage model and the two-stage model can add VREN in most cases.

Contribution. Our main contribution is three-fold:

1. We propose an effective attention module (VREN), which can be integrated with any CNN architecture.
2. Compared with the existing attention mechanism, VREN combines spatial attention and channel attention and enhances the correlation between the two kinds of attention.
3. We evaluate the effectiveness of VREN through a large number of ablation experiments.

2. Related Work

As mentioned in the introduction, occlusion interference feature extraction will cause the feature map not to be able to effectively guide the classifier to make a correct judgment on the predicted box. Therefore, for the detection of the occlusion scene, the feature information should be distinguished. To make this purpose, the attention mechanism can re-weight the feature by adjusting the spatial dimension and the channel dimension.

Occluded Pedestrian Detection. Several studies have been proposed to handle occlusion in pedestrian detection. A common strategy is a part-based approach where a set of part detectors are learned with each part designed to handle a specific occlusion pattern. Some of these part-based methods, such as [13][14], divide pedestrians into different parts and then train several detectors to detect each part.

As a part of the whole, the component detector can effectively use the structural information of the visible part when dealing with the occlusion problem. However, training each component detector separately linearly increases computing resources consumed with the number of defined component detectors. In addition, some part-based methods, such as [15][16], integrate structural information of objects into a network and exploit visible body information to learn specific occlusion modes. Different from these methods, we propose a module that uses the attention mechanism to adjust the weight of the input

feature map and uses effective information to detect pedestrians.

Attention Mechanism. The attention method consists of spatial attention and channel attention, specifically, spatial attention helps us focus on where features are meaningful and channel attention helps us focus on what features are meaningful. Since Squeeze-and-Excitation Networks(SENet)[17] have demonstrated the effectiveness of the attention mechanisms, which are widely used in many computer vision tasks such as image classification, object detection, instance segmentation, and semantic segmentation. SENet[17] improves detection performance at a very low cost with MaxPool and AveragePool operations, but it ignores the importance of spatial information. Therefore, the Bottleneck Attention Module(BAM)[18], Double Attention Networks(DANet)[19], and Convolutional Block Attention Module(CBAM)[20] are proposed to obtain the attention map by combining the spatial and channel attention. Motivated by CBAM, to extract richer feature information, a new second-order pooling method was proposed in [21] based on Global Second-order Pooling(GSoP). Subsequently, [22] introduces a dynamic selection attention mechanism named Selective Kernel Networks(SKNet), which allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. The ResNeSt[23] proposes a similar Split-Attention block that applies channel-wise attention to different network branches to leverage their success in capturing cross-feature interactions and learning diverse representations. To reduce model complexity and improve detection efficiency, GCNet[24] introduces a simple spatial attention module and thus a long-range channel dependency is developed. The ECANet[25] employs the one-dimensional convolution layer to reduce the redundancy of fully connected layers. The FcaNet[26] proposes a novel multi-spectral channel attention that realizes the pre-processing of channel attention mechanism in the frequency domain. On the basis of SENet[17], EPSANet[27] groups the feature map to obtain a split attention block. To effectively combine two types of attention mechanisms and reduce the computational overhead, SA-Net[28] first groups channel dimensions into multiple sub-features before processing them in parallel.

For occlusion object detection, we propose a visible region enhancement network that combines spatial and channel attention, specifically, the acquisition of them is interrelated compared to the above-mentioned methods.

3. Method

In this section, we introduce the VREN, which consists of spatial and channel attention. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$ as input, VREN sequentially infers a 2D spatial attention $A_s \in \mathbb{R}^{1\times H\times W}$ and a 1D channel attention $A_c \in \mathbb{R}^{C\times 1\times 1}$, especially, the acquisition of the A_c is affected by A_s, the overall framework is shown in Fig.2. The overall process of VREN can be summarized as:

$$F' = F \times A_s$$

$$A_c = F \otimes A_s$$

$$F'' = F' \times A_c$$

F'' is the final refined feature map as output. The following describes the details of VREN and the attention module.

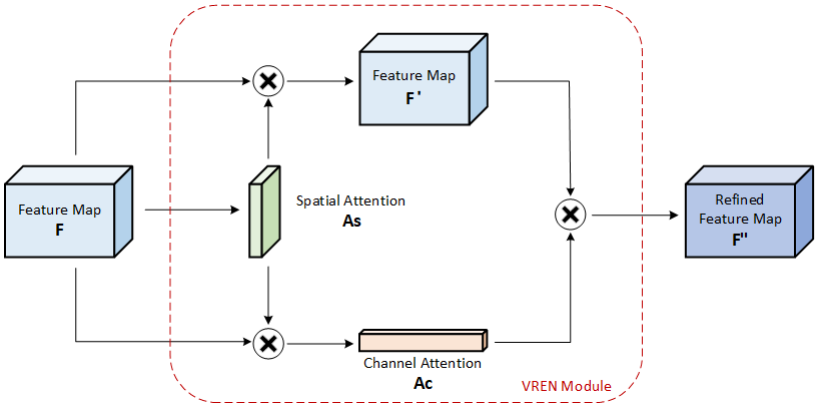


Figure 2. The architecture of VREN.

Visible Region Enhancement Network. As mentioned earlier, we design VREN to take into account the incompleteness of object information in the case of occlusion, and the missing information will reduce the overall confidence of the object. Therefore, VREN first obtains spatial attention to determine where are visible at the spatial level and then obtains channel attention by feature map convolve with spatial attention to determine what features are visible. Finally, we obtain the refined feature map after feature map sequentially through the processing of spatial attention and channel attention. Refined feature map makes the information of the object's visible region enhanced, and the irrelevant information is suppressed. The overall framework of VREN is shown in figure 2.

Spatial Attention. Spatial attention focuses on 'where' features of a given input image are visible, our method produces a spatial attention mask through three consecutive convolution operations. For aggregating attention feature information, Woo et al.[20] use both max-pooling and average-pooling operations, this operation is very simple and shows to be effective in highlighting informative regions[29]. In order to improve the learning ability of spatial attention and the nonlinear expression ability of VREN, we use three convolution operations to obtain spatial attention mask. Specifically, the first two convolutional layers continuously reduce the channel dimension to $\mathbb{R}^{1\times H\times W}$ as preliminary mask, and the last convolutional layer adjusts the mask with very few parameters as the final spatial attention mask. To reduce the complexity of the model, we set the size of the first convolution kernel to 1×1, and the second and the third to 3×3. In short, spatial attention is computed as:

$$A_s(F) = f^{3\times3}\left(f^{3\times3}\left(f^{1\times1}(F)\right)\right)$$

where $f^{1\times1}$ denotes a convolution operation, which has the filter with the size of 1×1. The $f^{3\times3}$ denotes a convolution operation, which has a filter with the size of 3×3. Figure 3 depicts the computation process of spatial attention.

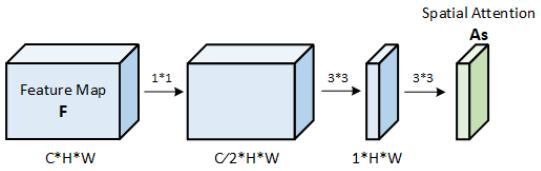


Figure 3. Computation process of spatial attention.

Channel Attention. Different channel represents different filter, channel attention focuses on 'what' features of a given input image are visible. Our method produces a channel attention map through feature map F convolve with spatial attention mask. For aggregating spatial feature information, common operations are to use max-pooling and average-pooling for dimensionality reduction. Hu et al. use it to design a simple attention module to obtain effectively channel information. However, we consider that only if the object characteristic information is visible, the channel filter should play a specific role. In other words, 'where' should guide the generation of 'what'. We first aggregate spatial information of a feature map by using spatial attention mask $A_s(F) \in \mathbb{R}^{1\times H\times W}$ convolves with feature map $F \in \mathbb{R}^{C\times H\times W}$, generating a spatial context descriptor $A \in \mathbb{R}^{C\times1\times1}$. The descriptor is then forwarded to a network to produce channel attention map $A_c \in \mathbb{R}^{C\times1\times1}$, the network is composed of fully connected layer(FC) with three hidden layers. In short, the channel attention is computed as:

$$A_c(F) = \sigma\left(FC\left(FC\left(FC\left(FC\left(F\otimes A_s(F)\right)\right)\right)\right)\right)$$

$$= \sigma\left(W_3\left(W_2\left(W_1\left(W_0(A_C)\right)\right)\right)\right)$$

Where σ denotes the sigmoid function, $W_0 \in \mathbb{R}^{\frac{C}{4}\times C}$, $W_1 \in \mathbb{R}^{\frac{C}{16}\times\frac{C}{4}}$, $W_2 \in \mathbb{R}^{\frac{C}{4}\times\frac{C}{16}}$, and $W_3 \in \mathbb{R}^{C\times\frac{C}{4}}$. Figure 4 depicts the computation process of spatial attention.

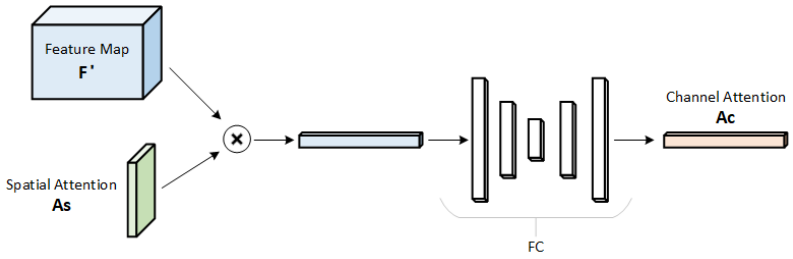


Figure 4. Computation process of channel attention.

4. Experiments

In this section, we evaluate VREN on the standard benchmarks: CrowdHuman datasets[12] for object detection. In order to perform better comparisons, we first reproduce the Faster R-CNN in the PyTorch framework and set it as our baseline. Then we perform extensive experiments to thoroughly evaluate the effectiveness of our module.

4.1. Datasets and Evaluation Metrics

Datasets. The quality of the datasets greatly affects the performance and generalization ability of the detector, so we chose CrowdHuman datasets[12] as our test data to simulate occlusion situations. CrowdHuman contains 15000 training images, 4370 validation images, and 5000 test images respectively. Especially, each picture has an average of 22.64 pedestrians, and the occlusion rate of 2.4 pedestrians exceeds 0.5[30]. We use the full-body benchmark in [24] to evaluate our model, and the results are evaluated on the validation dataset.

Evaluation Metrics. To better reflect the advantages of the proposed method, we use two metrics for comparison, including AP and MR^{-2}[31].

- AP, which is short for average precision, is the most popular metric for object detection. AP reflects both the precision and recall of detection results. The larger the AP, the better the performance of the detector.
- MR^{-2}[31], which is short for log-average Miss Rate on False Positive Per Image (FPPI) in [10^{-2},100], is a common metric used in pedestrian detection. MR^{-2} reflects false positives of detection results. The smaller the MR^{-2}[31], the better the performance of the detector.

4.2. Implementation Details

We use the open-source implementation of Faster R-CNN[7] for experiments. The models are trained on 2 NVIDIA Tesla V100 GPUs, and the batch size is 8 per GPU within 90 epochs. We use the SGD optimizer with a momentum of 0.9, the weight decay of 10^{-4}. The learning rate is initially set to 0.01 and is decreased by the factor of 10 at the 72th and the 81th epochs, respectively.

4.3. Ablation Study.

We perform the ablation experiments of the proposed module to evaluate the effectiveness of various parts, including spatial attention and channel attention. The baseline is Faster R-CNN using Resnet50 for feature extraction. It is clear that the best performance is achieved only when both spatial attention and channel attention act on the visible region enhancement network. Table 1 has shown the specific performance of our experiments. It is clear that our method consistently improves the detection performances by 3.5% in AP and 7.2% in MR^{-2}[31] compared to the baseline network Faster R-CNN[7]. To improve test efficiency, we only add each attention to the last feature map.

Table 1. Ablation experiments on CrowdHuman.

Spatial attention	Channel attention	AP(%)	MR^{-2}(%)
		84.61	47.34
√		86.84	43.10
	√	85.29	46.36
√	√	87.10	42.61

4.4. Comparisons with Other Attention Mechanism

To our knowledge, very few previous works of attention mechanisms on crowded detection report their results. To compare, we reproduce several attention algorithms. All methods use Faster R-CNN[7] as the base detector with the same implementation details. Table 2 lists the comparison results. In contrast, our method achieves the best results. The reason is that VREN guide the generation of channel attention through spatial attention. Spatial attention first filters out the interference information from the spatial level so that channel attention can focus more accurately on the selection of feature patterns by learning of FC.

Table 2. Comparison experiments on CrowdHuman.

Method	AP(%)	MR^{-2}(%)
baseline	84.61	47.34
+SENet	86.99	44.13
+BAM	87.39	41.60
+CBAM	87.49	41.28
+SKNet	87.05	43.76
+EPSANet	86.38	44.58
+SANet	87.48	42.55
+VREN(ours)	88.12	40.12

In order to better show the effect of our method, we visually compare the results of three algorithms, which are baseline, CBAM[20], and our method. The reason for choosing CBAM[20] is that it is the best result except for our method. Figure 5 shows the results of the visual comparison.

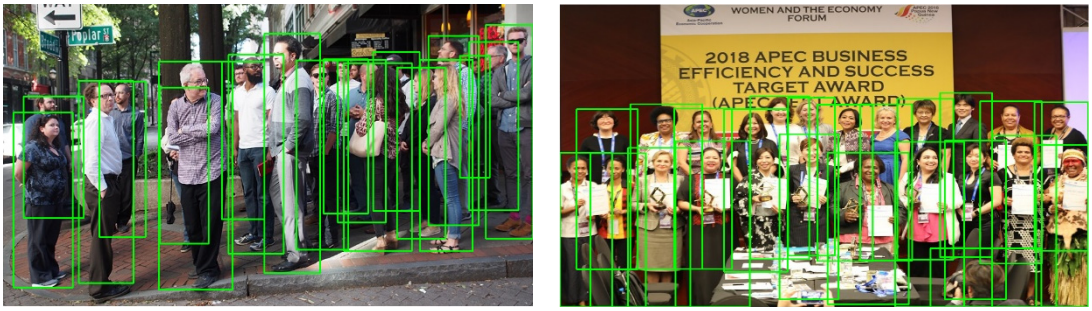


Figure 5. Visual comparison. The first row is the results of the baseline. The second row is the results of CBAM[20]. The last row is the results of our method. Red boxes are the missed detection ones.

5. Conclusion

In this paper, we have proposed the visible region enhancement network(VREN), a novel method to improve the representation power for occluded pedestrian detection. This method makes use of the concept of attention, designing new spatial attention and channel attention. Our approach is not only effective but also easy to combine with most existing state-of-the-art detection frameworks.

6. References

[1] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//LNCS 9905: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 8-16, 2016. Cham: Springer, 2016: 21-37.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 779-788.

[3] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2999-3007.

[4] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[5] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[6] GIRSHICK R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 13-16, 2015. Washington: IEEE Computer Society, 2015: 1440-1448.

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems 28, Dec 7-12, 2015. Red Hook: Curran Associates, 2015: 91-99.

[8] DAI J, LI Y, HE K, et al. R-FCN: object detection via region based fully convolutional networks[C]//Advances in Neural Information Processing Systems 29, Barcelona, Dec 5-10, 2016. Red Hook: Curran Associates, 2016: 379-387.

[9] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 2980-2988.

[10] Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 43(5): 1483-1498.

[11] Fawzi A, Frossard P. Measuring the effect of nuisance variables on classifiers[C]//British Machine Vision Conference (BMVC). 2016 (CONF).

[12] Shao S, Zhao Z, Li B, et al. Crowdhuman: A benchmark for detecting human in a crowd[J]. arXiv preprint arXiv:1805.00123, 2018.

[13] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[C]//Proceedings

of the IEEE international conference on computer vision. 2015: 1904-1912.
[14] Zhou C, Yuan J. Multi-label learning of part detectors for occluded pedestrian detection[J]. Pattern Recognition, 2019, 86: 99-111.
[15] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 637-653.
[16] Xie J, Pang Y, Cholakkal H, et al. PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection[J]. Science China Information Sciences, 2021, 64(2): 1-13.
[17] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
[18] Park J, Woo S, Lee J Y, et al. Bam: Bottleneck attention module[J]. arXiv preprint arXiv:1807.06514, 2018.
[19] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.
[20] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. 2018: 3-19.
[21] Gao Z, Xie J, Wang Q, et al. Global second-order pooling convolutional networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3024-3033.
[22] Li X, Wang W, Hu X, et al. Selective kernel networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 510-519.
[23] Zhang H, Wu C, Zhang Z, et al. Resnest: Split-attention networks[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2736-2746.
[24] Cao Y, Xu J, Lin S, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.
[25] Wang Q, Wu B, Zhu P, et al. Supplementary material for 'ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA. 2020: 13-19.
[26] Qin Z, Zhang P, Wu F, et al. Fcanet: Frequency channel attention networks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 783-792.
[27] Zhang H, Zu K, Lu J, et al. Epsanet: An efficient pyramid split attention block on convolutional neural network[J]. arXiv preprint arXiv:2105.14447, 2021.
[28] Zhang Q L, Yang Y B. Sa-net: Shuffle attention for deep convolutional neural networks[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2235-2239.
[29] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint arXiv:1612.03928, 2016.
[30] Chu X, Zheng A, Zhang X, et al. Detection in crowded scenes: One proposal, multiple predictions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12214-12223.
[31] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 34(4): 743-761.