



FE-CSP: a fast and efficient pedestrian detector with center and scale prediction

Yugang Qin¹ · Yurong Qian^{1,2,3} · Hongyang Wei¹ · Yingying Fan² · Peiyun Feng¹

Accepted: 7 September 2022 / Published online: 21 September 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

There are still many pressing problems in pedestrian detection, such as difficulty in detection due to severe pedestrian occlusion, difficulty in detecting small objects and low detection speed. In this paper, we propose A Fast and Efficient Pedestrian Detector with Center and Scale Prediction (FE-CSP). We combine channel attention with spatial attention, replace the traditional convolution with deformable convolution, and embed the backbone network to propose CSANet (Channel and Spatial Attention Network), which efficiently extracts the semantic features of the object, and then propose a feature pyramid network to replace the traditional concatenation to perform multi-scale feature detection, which effectively improves the detection speed. By conducting experiments on CityPersons, our method achieves 10.1%, 13.7% and 47.4% MR^{-2} at a speed of 0.21 s/img on the reasonable setting, small setting and heavy setting, respectively. On Caltech, our method achieves 5.2% MR^{-2} at a speed of 0.06 s/img on the Reasonable setting, further demonstrating the superiority and generalization ability of the proposed method.

Keywords Pedestrian detection · Channel attention · Spatial attention · Deformable convolution · Feature pyramid network

1 Introduction

Pedestrian detection has a large number of real-life applications, from object tracking [1] and video surveillance [2] to the recent research hotspot of autonomous driving [3]. In autonomous driving, cars need to complete identification and tracking of pedestrians to avoid collisions. In addition, since the computational power inside the

✉ Yurong Qian
qyr@xju.edu.cn

Extended author information available on the last page of the article

car is limited, this requires a lightweight pedestrian detection model with high accuracy and speed to enable better perception of the surrounding environment.

Pedestrian detection is one of the long-standing problems in computer vision and an essential problem for object detection. With the development of deep learning, pedestrian detection has made great progress. However, there are still some areas for improvement. For example, pedestrians are severely impeded, causing low detection efficiency; small objects are challenging to detect; as in real-time detection, pedestrian detection still needs significant improvement in detection speed.

With the development of object detection techniques, some general object detectors, such as two-stage detectors and their improved algorithms [4–6] and one-stage detectors and their improved algorithms [7–9], have been proposed to achieve state-of-the-art performance on benchmark Pascal VOC [10] or MS COCO [11]. These one-stage detection algorithms are faster but less accurate, while the two-stage detection algorithms have higher accuracy but lower detection speed.

Since CornerNet [12] was proposed, researchers have successively proposed keypoint-based detectors. For instance, FCOS [13] proposed an end-to-end object detector using the center point and the distance from the center point to the ground truth bounding box. Note that both CenterNet [14] and CSP [15] use a center point for object detection and achieve good results in terms of speed and accuracy. Unlike other keypoint predictions, center point prediction does not require grouping the predicted keypoints, which dramatically saves computational cost and inspires us to conduct further research in pedestrian detection using center point prediction. Therefore, we propose the FE-CSP pedestrian detection algorithm, where we simplify pedestrian detection to a center and scale prediction task. Unlike the above studies, we try to go into how the use of attention mechanisms in pedestrian detection allows the backbone network to extract features more efficiently, especially for small and heavily occluded objects, and apply it to a pedestrian detection framework to achieve more efficient detection performance. The main contributions of this work are as follows:

- (1) To solve the problem that small objects are difficult to detect, we combine the Global Context Block with the Transformer Attention module and replace the traditional convolution with deformable convolution to propose CSANet, effectively improving the feature extraction capability of the backbone network, making the small object detection performance to be improved.
- (2) To make better and more effective use of the output of the feature from the backbone network, we explored the effect of concatenation operation and feature pyramid network. We replaced the traditional concatenation with a feature pyramid network, which facilitated the detection of multi-scale objects and effectively improved the detection speed of the model.
- (3) In this paper, we propose a one-stage detection algorithm, FE-CSP, and conduct extensive comparison experiments with the other state of the arts on the CityPersons and Caltech datasets to fully indicate that the algorithm effectively

improves the pedestrian detection performance and efficiency and achieves better trade-off, demonstrating strong robustness and generalizability.

The rest of the paper is organized as follows. Section 2 presents the work related to the paper. Section 3 details the methods proposed in this paper. Section 4 describes the experimental comparative analysis of the method proposed in this paper with classical and cutting-edge algorithms. Finally, Sect. 5 is on the overall conclusion of the paper.

2 Related work

With the rapid development of deep learning in computer vision, pedestrian detection has been in the deep learning stage since Girshick et al. proposed the RCNN [16] in 2014. Generally speaking, deep learning-based detection algorithms are divided into two main detection frameworks, one is two-stage detection method, and the other one is the one-stage detection method. Also, the application of attention mechanisms to object detection algorithms has achieved remarkable results in recent years. In the following, pedestrian detection is described in terms of these two detection frameworks and attention mechanisms.

2.1 Two-stage detection framework

The two-stage detection framework is mainly divided into two parts for detection. First, a series of regional proposals are generated on the image, and then further prediction of the regional proposals is performed to get the final result. Some general object detectors such as two-stage based detector RCNN [16], Fast RCNN [17], Faster RCNN [18] and other algorithms are proposed to improve the object detection performance significantly.

In the field of pedestrian detection, Shanshan Zhang et al. [19] produced a challenging benchmark CityPersons and experimented with the Faster RCNN model and obtained good results. Shifeng Zhang et al. [20] designed an aggregation loss to force proposals to approach and closely localize to the corresponding objects to solve the occlusion problem in pedestrian crowding situations. Xinlong Wang et al. [21] designed repulsion loss to achieve detection by repelling the surrounding objects and applying it to a two-stage detector to solve the pedestrian occlusion problem effectively. Irtiza Hasan et al. [22] argue that most pedestrian detectors do not generalize well. When evaluated by cross-validation sets, their performance degrades, so they propose a fine-tuning strategy to improve generalization and get good results on some benchmark datasets using a two-stage detector Cascade RCNN. However, the drawback of these two-stage pedestrian detectors is also evident in that the detection speed is not high and still needs further research.

2.2 One-stage detection framework

One-stage detectors remove the generation of a prior bounding box compared to two-stage detectors and directly predict objects and bounding boxes. With the development of deep learning in the field of object detection, one-stage based detectors SSD [23], YOLO [24], RetinaNet [25], EfficientDet [26], YOLOF [27] and other algorithms have been proposed one after another. One-stage detectors include both anchor-based methods, where the detector generates bounding boxes based on a pre-determined number of anchors with fixed scales and aspect ratios, such as the YOLO series [28–30], and anchor-free methods, such as CornerNet, FCOS, and other key-point detection-based methods.

In pedestrian detection, Tao Song et al. [31] proposed a multi-scale pedestrian detection method using topological line localization and temporal feature aggregation to solve small-scale object detection and acted on a one-stage detector to achieve competitive results. Wei Liu et al. [32] designed an efficient one-stage pedestrian detection framework ALFNet, a simple structure in this network but effective module, namely Asymptotic Localization Fitting (ALF), the default anchor boxes are evolved by stacking a series of prediction modules of SSD to obtain both the speed of SSD and the accuracy of a two-stage detector. CornerNet [12] solves the object detection problem as a key point detection problem by detecting the upper left and lower right corners of the objects to predict the bounding boxes, which is a pioneering work in the field of object detection. Later Wei Liu et al. [33] proposed a one-stage detection framework CSP, which simplifies pedestrian detection to center and scale prediction using convolution, and obtained competitive results. Compared with the two-stage pedestrian detector, the one-stage pedestrian detector has a higher detection speed but is still inferior to the two-stage detector in terms of accuracy.

2.3 Attention-based detection algorithm

It is well known that the attention mechanism plays an integral part in human perception [34–36], and the human visual system will choose to focus on the salient parts according to its needs. In recent years of deep learning research, attention mechanism algorithms have also been the subject of keen research by many scholars. Good results have been achieved by introducing attention mechanisms into computer vision.

Not all regions in an image are necessary; only those that are relevant to the task are important. Many research scholars proposed spatial attention mechanisms and improved algorithms [37, 38]. Max Jaderberg et al. [39] resented Spatial Transformer Networks, which can actively transform the feature map spatially without additional supervised training, achieving state-of-the-art performance on multiple benchmarks. One dimension is the aspect ratio for two-dimensional images, and the other is the channel, so researchers have proposed many channel attention mechanisms to improve network performance [40, 41]. Hu et al. [42] focused on channel relations. They proposed an architectural unit called squeeze-and-excitation (SE) block, which improves network performance by modeling the interdependencies

between the channels and adaptively calibrating the channel response characteristics, bringing significant performance gains at a very small additional computational cost.

The above approaches result from many researchers who have studied channel attention and spatial attention separately, while some others have combined channel attention and spatial attention by taking advantage of both and achieved very competitive results [43–45]. Sanghyun Woo et al. [46] proposed the Convolutional Block Attention Module (CBAM), which sequentially inferred attention maps along both channel and spatial dimensions, and then attention maps were multiplied with the input feature maps for adaptive feature refinement and can be trained end-to-end with CNNs. Introducing the attention mechanism in the above study improves the detection accuracy. However, it brings additional computational cost, which affects the detection speed and still has much room for improvement.

3 Proposed method

In this section, we detail the proposed one-stage pedestrian detection method FE-CSP. The framework of FE-CSP is shown in Fig. 1. First, in Sect. 3.1, we introduce the CSANet (Channel and Spatial Attention Network). Then the Channel Attention Module and Spatial Attention Module are described in two subsections, respectively. Section 3.1.1 introduces the Global Context Block (GCB), a channel attention mechanism that can establish long-range functional dependencies. In Sect. 3.1.2, we detail the Deformable Convolution Network (DCN) and form the Spatial Attention Mechanism by combining Transformer Attention with deformable convolution, which facilitates object localization and is also more effective in

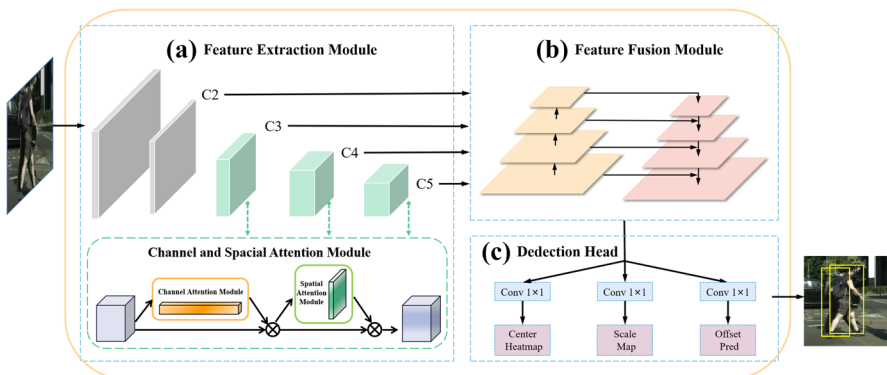


Fig. 1 The framework of FE-CSP. (a) CSANet. A feature extraction network with Channel Attention Module (CAM) and Spacial Attention Module (SAM). (b) Feature fusion module. Feature fusion on the feature level of the backbone network. (c) Detection head. There are three branches: Center Heatmap, Scale Map, and Offset Pred

acquiring object features. In Sect. 3.2, we introduce the feature pyramid structure for feature fusion, which effectively fuses some feature maps to generate multi-scale high-level semantic information and improves detection speed by sharing parameters in each layer of the feature pyramid.

3.1 CSANet

We choose ResNet-50 as the backbone, and we improve the feature extraction capability of the network by embedding the Channel and Spatial Attention Module (CSAM) into ResNet-50. The CSAM is shown in Fig. 1. The CSAM is divided into two submodules, channel attention and spatial attention, and channel attention is performed before spatial attention. First, we introduce Global Context Block (GCB), which can effectively model the global context and establish effective long-range dependencies. Then we replace the regular convolution with deformable convolution in the network with feature level {C3, C4, C5}, which can accurately improve the feature extraction ability of the network, followed by the introduction of the Transformer attention module, which can extract the semantic information we want by combining with deformable convolution. CSANet is proposed by improving ResNet-50, and the architecture of CSANet is shown in Fig. 2. It effectively enhances the ability of the network to extract semantic information without adding much overhead.

3.1.1 Channel attention module

Traditional convolution deals with adjacent pixels, models adjacent pixels and achieves long-range dependence by deepening the convolution layers [40], but it introduces many drawbacks. The first is that the network is not refined enough, which can bring a lot of parameters and computation. And the deeper the network is, the more difficult it is to optimize. And because of the maximum distance limitation

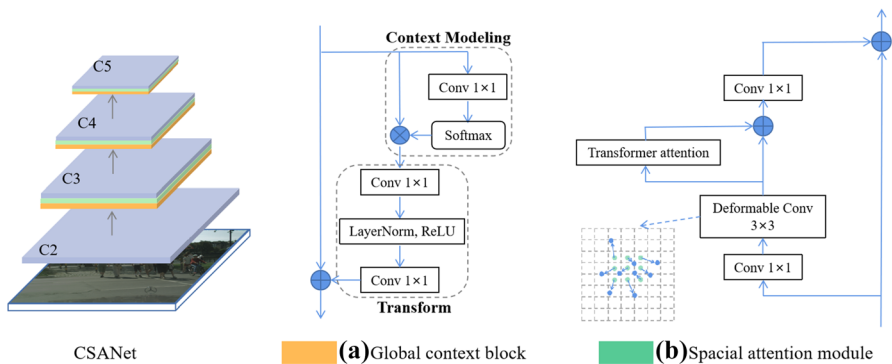


Fig. 2 CSANet, where C2, C3, C4, and C5 denote the feature level of the backbone network, the orange block represents the global context block used to establish effective long-range dependencies, and the green block represents spacial attention, which highlights spatially effective information by combining deformable convolution with Transformer attention

of the convolution operation, long-range dependence is not fully established [41]. Therefore, we introduce Global Context Block (GCB), as shown in Fig. 2(a), an alternative view of image processing, analyzing the image from a global perspective, effectively establishing long-range dependencies and integrating the advantages of lightweight.

As shown in Fig. 2(a), GCB is divided into three main modules: a) a context modeling module, which aggregates features from all locations to form a global context feature; b) a feature transformation module, which is used to capture interdependencies between channels; and c) a feature fusion module, which fuses global context features with features from all locations. We abstract this as the global context modeling framework, defined as

$$Z_i = F \left(X_i, \delta \left(\sum_{j=1}^{N_p} \alpha_j X_j \right) \right) \quad (1)$$

Where $\sum_j \alpha_j X_j$ denotes the context modeling module, which combines all features by weighted averaging and weighting α_j to obtain global context features. $\delta(\cdot)$ represents features that capture inter-channel correlation, and $F(\cdot)$ means the fusion function, aggregating global context features to each location.

Specifically, a 1×1 convolution and softmax function is first used to obtain attention weights. Attention pooling is used to build global contextual features, and then a 1×1 convolution operation is used to perform the feature transformation. Notably, the layer normalization is added between the two layers of bottleneck transform (before ReLU) to ease optimization and as a regularizer, also facilitates generalization. Finally, the obtained features are aggregated with the original input features, and the global context features are aggregated to the features at each location using addition, formulated as

$$Z_i = X_i + \omega_{v2} \text{ReLU} \left(\text{LN} \left(\omega_{v1} \sum_{j=1}^{N_p} \frac{e^{W_k X_j}}{\sum_{m=1}^{N_p} e^{W_k X_m}} X_j \right) \right) \quad (2)$$

where $\alpha_j = \frac{e^{W_k X_j}}{\sum_m e^{W_k X_m}}$ indicates global attention pooling and $\delta(\cdot) = \omega_{v2} \text{ReLU}(\text{LN}(\omega_{v1}(\cdot)))$ denotes bottleneck transformation.

3.1.2 Spatial attention module

To highlight the useful spatial information of the feature map, we again propose introducing the spatial attention module, which is composed of deformable convolution and Transformer attention, as in Fig. 2(b).

Instead of changing the computational operation of the convolution, deformable convolution adds a learnable parameter Δp_n to the region of convolution operation, formulated as

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad (3)$$

where R denotes the receptive field, p_0 is each location on the feature map, p_n enumerates the locations in R , Δp_n means the relative position to p_n , and $w(p_n)$ indicates the weight of the location p_n .

Assuming a 3×3 convolution kernel is used, for each output $y(p_0)$, the convolution kernel samples nine positions from the feature map, each position being obtained by diffusing the central position $x(p_0)$ in all directions, but with an extra Δp_n that allows the sampled points to diffuse into a non-gird shape. The offsets are obtained by convolving the original feature layer. It can be a floating-point number. In the learning of offsets, the gradient is backpropagated by bilinear interpolation.

Transformer attention considers a small set of sample locations to highlight all key features of the feature map and can be naturally extended to fuse multi-scale features, which, combined with the sparse spatial sampling capability of deformable convolution, improves the representational power of the feature map without adding a great deal of complexity.

It is possible to decompose the attention matrix into a product of random non-linear functions of the original query and key, the so-called random feature so that the similarity information can be encoded more efficiently. Transformer attention has four possible attention factors: the query and key content, the query content and relative position, the key content only, and the relative position only. We selected the third factor based on the findings of the literature [47], the key content only, and combined it with the deformable convolution to complement each other to achieve a good trade-off in terms of accuracy and efficiency.

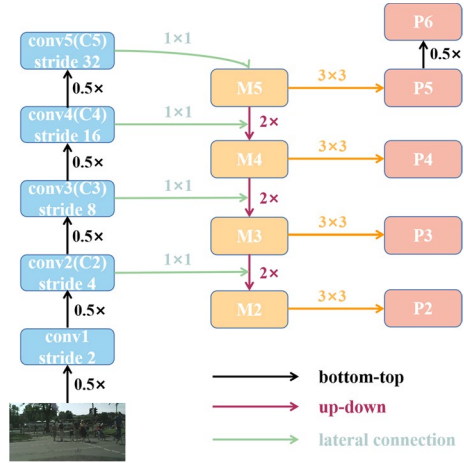
Specifically, a 3×3 deformable convolution operation is to achieve sparse space, making the features more focused on content and position offsets. A Transformer attention operation is to further focus on the content features and perform the feature fusion with the deformable convolution features. Finally, a skip connection with the original input feature map after another 1×1 convolution to obtain a feature map with more substantial representational power.

3.2 Feature fusion module

To effectively fuse the different scale feature maps of the backbone network, we adopt the FPN [48] (Feature Pyramid Network) structure, the detailed design is shown in Fig. 3. It uses a top-down approach and lateral connections to complete the fusion of the whole feature maps.

Taking ResNet-50 as an example, feature levels conv2, conv3, conv4, and conv5 are selected as the input features of the FPN, denoted as $\{C2, C3, C4, C5\}$, and the strides of these feature levels relative to the original map are 4, 8, 16, and 32, respectively. In the top-down process, the smaller feature maps of the upper levels are expanded to the corresponding sizes by an upsampling operation and fuse with the original feature level, which has the advantage of utilizing both the high-resolution information of the bottom layer, which is beneficial for object localization,

Fig. 3 Feature pyramid network, where C2 to C5 are the input feature levels and P2 to P6 are used as output feature levels



especially for small objects and the stronger semantic information of the lower resolution of the top layer, which is beneficial for object classification.

Specifically, first, a 1×1 convolution operation is performed to change the number of channels in the current layer. Then it is fused with the upper layer feature maps that have been upsampled by addition to obtain the feature levels {M2, M3, M4, M5}. Finally, 3×3 convolution operations will get the final set of feature maps {P2, P3, P4, P5} to reduce the confounding effect brought by the upsampling process. It is worth noting that P6 is obtained by downsampling P5 with a stride of 2. Thus, the feature pyramid finally outputs P2 to P6 feature maps, each with 256 channels. While many design choices are not critical, we emphasize that the FPN backbone is crucial. From preliminary experiments, using only one layer of features for detection can lead to poor performance.

3.3 Loss function

We follow the detection head of the CSP [15] model, which is divided into three parts, center heatmap, scale map and offset prediction. The corresponding loss function we use also has three parts. The classification loss is shown as:

$$L_{\text{center}} = -\frac{1}{K} \sum_{i=1}^{W/r} \sum_{j=1}^{H/r} \alpha_{ij} (1 - \hat{p}_{ij})^\gamma \log(\hat{p}_{ij}) \quad (4)$$

where

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij} = 1 \\ 1 - p_{ij} & \text{otherwise,} \end{cases} \quad \alpha_{ij} = \begin{cases} 1 & \text{if } y_{ij} = 1 \\ (1 - M_{ij})^\beta & \text{otherwise.} \end{cases} \quad (5)$$

where p_{ij} indicates the probability of the presence of a center at location (i, j) and $y_{ij} = 1$ indicates the presence of a center in the label. We set the γ to 2 according to the suggestion in [25], and α_{ij} is based on a Gaussian mask to reduce the contribution to the total loss.

For scale prediction, we apply Smooth L1 loss:

$$L_{scale} = \frac{1}{K} \sum_{k=1}^K \text{SmoothL1}(s_k, t_k) \quad (6)$$

where s_k denotes the predicted value and t_k denotes the true value of the positive sample label.

Offset prediction uses the same Smooth L1 loss as scale prediction. The loss function of the whole network can be given as follows:

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset} \quad (7)$$

where λ_c , λ_s , and λ_o are the weights of center prediction, scale prediction and offset prediction, which are set to 0.01, 1 and 0.1, respectively.

4 Experiments

4.1 Datasets and evaluation criteria

Datasets to demonstrate the effectiveness of the proposed method, we select the challenging pedestrian detection benchmarks CityPersons [19] and Caltech [49] for evaluation and to verify the generalization of the model. CityPersons is a subset of Cityscapes, a collection of 5000 images taken in 27 cities in Germany and surrounding countries, with many scenes containing high-quality pedestrian bounding box annotations. On average, there are seven people in each image. The training set, validation set and test set have 2975, 500 and 1575, respectively. Each image resolution is 2048×1024. The Caltech pedestrian dataset exists approximately 13K persons extracted from a 10-hour video, containing 42784 training images and 4024 test images, each normalized to 640×480 resolution. We trained our model on the official training set and tested it on the official validation set. In the testing, we keep the original image resolution as input.

Evaluation Criteria in this paper, to evaluate the model's performance, the evaluation follows the standard Caltech evaluation metrics [49], which we mainly use as the evaluation metric, that is log-average Miss Rate over False Positive Per Image (FPPI), the smaller, the better. Miss Rate (MR) denotes the miss rate metric of the detection result, formulated as Eq. 4. FPPI indicates the average miss rate per image, formulated as Eq. 5. which is further divided into several subsets according to the different visibility of pedestrians, such as reasonable setting, small setting, heavy setting, etc., as shown in Table 1. In addition, to evaluate the model's efficiency, we use training memory (MB) and test time (s/img) as evaluation metrics. Also, the smaller, the better.

Table 1 Subset division on citypersons

Setting	Height	Visibility	Images
All	[20, inf]	[0.2, inf]	2875
Reasonable	[50, inf]	[0.65, inf]	1579
Large	[100, inf]	[0.65, inf]	941
Medium	[75, 100]	[0.65, inf]	311
Small	[50, 75]	[0.65, inf]	351
Bare	[50, inf]	[0.9, 1]	769
Partial	[50, inf]	[0.65, 0.9]	814
Heavy	[50, inf]	[0, 0.65]	972
Heavy occlusion	[50, inf]	[0.2, 0.65]	735

$$MR = \frac{FN}{TP + FP} \quad (8)$$

$$FPPI = \frac{FP}{N} \quad (9)$$

where FN (False Negative) means the prediction result is a negative sample, but it is wrong. TP (True Positive) means the prediction result is a positive sample and correct. FP (False Positive) means the prediction result is a positive sample but wrong. N represents the total number of pictures.

4.2 Experimental settings

The experimental system environment in this paper is 64-bit Ubuntu 16.04, NVIDIA GeForce GTX2080 with 8 GB memory, implemented under the deep learning framework PyTorch 1.4. The backbone network we use is ResNet-50 trained on ImageNet. For Citypersons, the learning rate is 0.0002, Adam is used as the optimizer, batch_size is set as 2, and training is stopped after 80 epochs. For Caltech, the learning rate is 0.0001, batch_size is set as 4, and training stops after 40 epochs. We performed standard data enhancement on the training set to get sample diversity, mainly using data enhancement methods such as scaling, flipping, and random cropping on the images. Note that the aspect ratio of the images was kept constant during this process.

4.3 Comparison with the state of the arts

In Table 2, we compare FE-CSP with some state of the arts on the CityPersons dataset. For the comparison experiments with other algorithms, we selected the state of the arts FRCNN [19], OR-CNN [20], RepLoss [21], TLL [31], ALFNet [32], Adaptive NMS [50], CSP [15], APD [51], AEVB [52], AutoPedestrian [53], PRNet [54] and PRNet++ [55]. We compared the performance on the same dataset with all

Table 2 Comparison of FE-CSP with some state of the arts on benchmark CityPersons.

Method	Backbone	Reasonable	Small	Medium	Large	Heavy	Partial	Bare	Test Time
FRCNN[19]	VGG-16	15.4	25.6	7.2	7.9	–	–	–	–
OR-CNN[20]	VGG-16	12.8	–	–	–	55.7	15.3	6.7	–
RepLoss[21]	ResNet-50	13.2	–	–	–	56.9	16.8	7.6	–
TLL[31]	ResNet-50	15.5	–	–	–	53.6	17.2	10.0	–
ALFNet[32]	ResNet-50	12.0	–	5.7	6.6	51.9	11.4	8.4	0.27s/img
Ada. NMS[50]	ResNet-50	11.9	–	–	–	54.0	11.4	6.2	–
CSP[15]	ResNet-50	11.0	16.0	3.7	6.5	49.3	10.4	7.3	0.33s/img
APD[51]	ResNet-50	10.6	–	–	–	49.8	9.5	7.1	0.12s/img
AEVB[52]	ResNet-50	12.7	–	–	–	54.6	–	–	–
AutoPedes.[53]	VGG-16	11.3	–	–	–	50.5	–	–	–
PRNet[54]	ResNet-50	10.8	–	–	–	53.3	10.0	6.8	0.22s/img
PRNet++[55]	ResNet-50	10.7	–	–	–	51.2	9.9	6.9	0.31s/img
FE-CSP	ResNet-50	10.1	13.7	3.7	6.3	47.4	9.6	6.7	0.21s/img

Bold numbers indicate the best results

experimental settings kept as consistent as possible to keep fairness. IoU was set as 0.5 for testing if not otherwise stated.

From Table 2, it can be seen that FE-CSP has different degrees of improvement compared to many other state of the arts, both in terms of accuracy and speed. Specifically, in terms of accuracy, FE-CSP outperforms baseline CSP in Reasonable setting, Small setting, Medium setting, Large setting, Heavy setting, Partial setting and Bare setting by 0.84%, 2.31%, 0.02%, 0.19%, 1.8%, 0.82%, and 0.59%, respectively. Experimentally, it is noteworthy that the FE-CSP has significantly enhanced detection performance on small and heavily occluded objects, which is because small and heavily occluded objects lose a lot of useful information when downsampling and our addition of channel attention and spatial attention enhance the retention of useful information to the extent that this information can be detected. In terms of speed, FE-CSP improves by 36.3% compared to baseline CSP, which is because we replace concatenation with a feature pyramid network for feature fusion, which effectively eliminates a series of complicated operations in the concatenation process, such as using L2 normalization to focus feature maps at different scales on the same scale.

To visualize the detection effectiveness of FE-CSP, we show the detection quality of FE-CSP on CityPersons and demonstrate it in comparison with the state of the arts, as shown in Fig. 4. It can be seen that the other methods have poor detection performance, suffer from missed detection, low detection accuracy and poor quality of bounding boxes, while the bounding boxes of FE-CSP tightly surround the objects with no redundant bounding box and high precision.

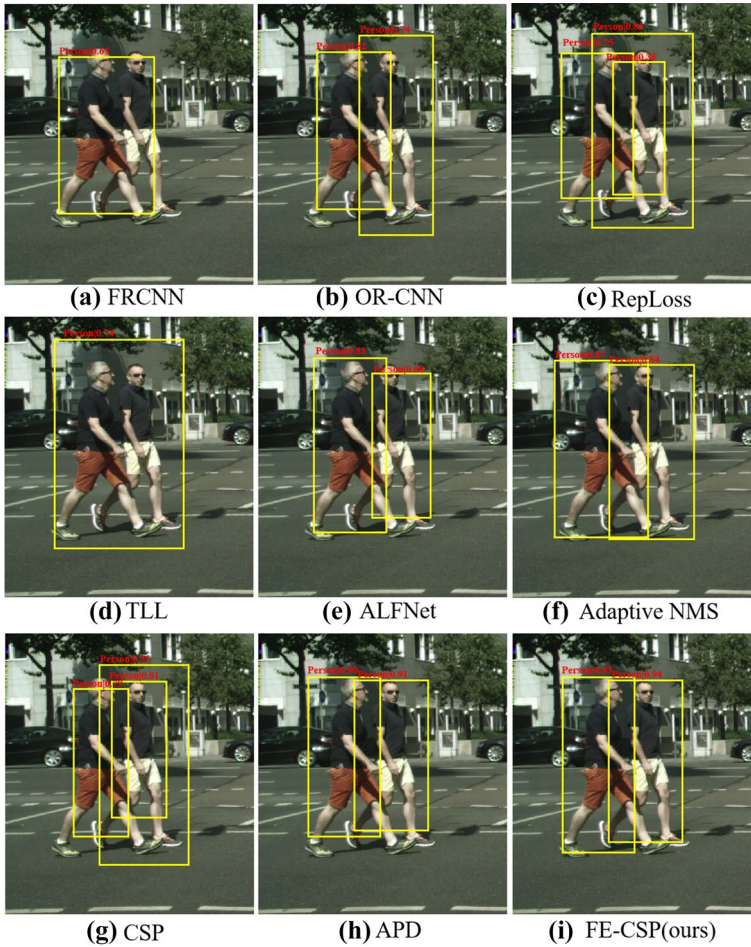


Fig. 4 Visualization comparison of detection performance on the state of the arts

4.4 Ablation experiments

In this section, we analyze the results of the ablation experiments performed on the FE-CSP model. First, we detail the effect of CSAM on FE-CSP, exploring the impact of the Channel Attention Module and Spatial Attention Module on extracting features. Then we analyze the effect of CSAM embedding positions in ResNet-50. Next, we perform comparative experiments and detailed analysis on the way of feature fusion, and finally, we explore the quality of the bounding box under a more stringent IoU.

Table 3 Effect of CSAM on the model.

Method	CAM	SAM	Memory	Reasonable	Small	Heavy
FE-CSP			2032	11.37	15.69	49.44
		✓	3662	11.50	14.18	49.61
	✓		4106	11.05	16.21	48.17
	✓	✓	4261	10.14	13.69	47.42

The CAM and SAM are embedded in the C3, C4, and C5 feature levels

Bold numbers indicate the best results

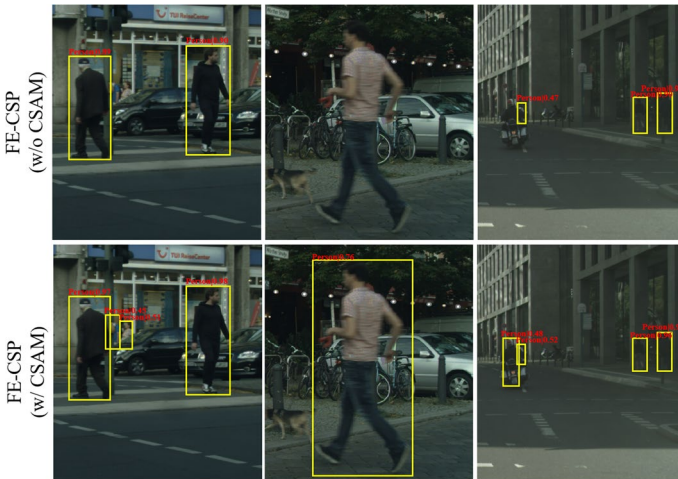


Fig. 5 Visualization of FE-CSP (w/o CSAM) and FE-CSP (w/ CSAM) detection performance

4.4.1 Validity verification of CSAM

From Table 2, we can see that the accuracy of FE-CSP on the CityPersons dataset has been improved to different degrees compared to the CSP algorithm. The most prominent is the accuracy improvement on small setting and heavy setting, which is 2.31% and 1.88%, respectively. After analysis, we believe there are several reasons for this. First, we selected ResNet-50 as the backbone for feature extraction. The localization ability of the objects decreases as the feature map scale becomes smaller in the process of downsampling, although it has advanced semantic features. Secondly, the convolution operation only models the relationship with local proximity pixels. The pooling operation also loses valuable information, ignoring the correlation between the whole and local ones, so we introduce the global context block to understand the image from a global perspective. And again, the purpose of introducing the spatial attention mechanism is to highlight the object features and preserve the image key information when performing spatial transformations. We have combined these two modules and achieved good results.

To verify the effect of CSAM on the final result, we conduct ablation experiments, as shown in Table 3. From the table, we can see that when only the spatial attention module is added, the accuracy of small objects is improved, but the accuracy of other metrics decreases. When only the global context block is added, there is an increase in accuracy on heavy setting and reasonable setting, but a decrease in accuracy on small setting, while when both the spatial attention module and the global context block are added, the performance of each setting then has a significant performance improvement. However, the parameters of the model increase, and later the two modules can be further investigated for a more effective combination and reduction of model parameters.

To see the effect of CSAM on the model more intuitively, we show the detection quality of FE-CSP (w/o CSAM) and FE-CSP(w/ CSAM) on CityPersons, as shown in Fig 5. We can observe that the FE-CSP model with CSAM has good pedestrian detection performance. FE-CSP(w/o CSAM) fails to detect small and heavily occluded objects, fails to see blurred people, and has low detection accuracy. In contrast, FE-CSP (w/ CSAM) has strong robustness and high detection performance for dense people, and FE-CSP containing CSAM can detect small objects and heavily obscured objects with high accuracy.

4.4.2 Effect of CSAM embedding location

In the proposed method, we combine channel attention with spatial attention to improve the feature extraction capability of the backbone network. At the same time, ResNet-50 produces four feature levels {C2, C3, C4, C5} when downsampling is performed. Hence, we must explore which stage of feature levels to add CSAM to make it work the best. Based on the literature [49]'s practical experience, we conducted the ablation experiment, as shown in Table 4.

From the table, we can see that when we only apply CSAM to C5, the model has a low number of parameters, but the accuracy will also be low, which is because the feature map size in C5 is small, which is not conducive to the detection of small objects, the use of CSAM does not play a vital role, and the long-range dependence is not utilized well. In contrast, when the CSAM is applied at C4 and C5, each setting is improved by ranging from 1% to 2%. When CSAM is applied at C4, some of the features of small objects are enhanced, and long-range dependence is also established relatively effectively. Then when it is used again at C5, the features are

Table 4 Effect of CSAM embedding location on model performance.

Method	Stage			Memory	Reasonable	Small	Heavy
	C3	C4	C5				
FE-CSP (w/ FPN)			√	3841	12.17	17.52	50.76
		√	√	4045	11.37	15.60	49.96
	√	√	√	4261	10.14	13.69	47.42

Bold numbers indicate the best results

Table 5 Exploring the effect of different feature fusion methods on model performance

Method	DCN	Feature fusion	Memory	Reasonable	All	Test Time
FE-CSP		FPN	3502	11.5	38.06	0.23s/img
		CONCAT	4826	11	37.57	0.33s/img
	✓	FPN	3511	11.37	37.69	0.23s/img
	✓	CONCAT	4841	10.9	37.13	0.33s/img

further enhanced, thus effectively improving the pedestrian detection accuracy. Furthermore, it is worth noting that when CSAM is added to the {C3, C4, C5}, the accuracy of each setting works the best, and the attention mechanism can fully function. More importantly, GCB can maintain inter-channel dependence and improve the detection of large and small objects.

4.4.3 Effect of feature fusion approach

To enable the output features from the backbone network to be more powerful, a feature fusion module is usually added to address the inefficiency of single feature map detection. We explore the impact of the concatenation operation and the feature pyramid network on the model during the processing of feature maps, mainly training memory, MR^{-2} and test time for comparison experiments, as shown in Table 5. The table shows that the training memory is significantly lower when using the feature pyramid network. The test time is also considerably lower, but the accuracy is also reduced. It shows that concatenation can improve some accuracy but is too complex and brings many parameters. FPN does not improve accuracy but brings a significant speed improvement. Whereas, after adding the Deformable Convolutional Network (DCN) [56], the situation is the same as before, although each setting has improved. Therefore, concatenation and FPN have advantages and disadvantages, but we finally chose to use FPN for feature fusion to make the model faster.

4.4.4 Quality of the bounding box

A good pedestrian detector should generate high-quality bounding boxes, which should tightly surround the pedestrian. To better understand the quality of generated bounding boxes, we evaluate the performance of FE-CSP with IoU set to 0.75 and compare it with baseline CSP.

Table 6 Comparison of FE-CSP with baseline CSP algorithm at IoU of 0.75.

Method	IoU	Reasonable	Small	Heavy Occ.	Heavy
CSP	0.75	15.02	19.19	48.23	56.86
FE-CSP		13.37	17.7	43.79	52.79

Bold numbers indicate the best results

Table 7 Cross dataset evaluation on caltech.

Method	Backbone	Citypersons→ Caltech	Caltech→Caltech	Test Time
FRCNN [19]	VGG-16	21.1	8.7	0.20s/img
Vanilla FRCNN [19]	VGG-16	17.6	12.2	0.20s/img
Faster R-CNN [18]	ResNext-101	11.8	9.7	–
Cascade R-CNN [22]	Swin Transformer	9.1	8.0	–
Cascade R-CNN [22]	HRNet	8.8	6.2	0.20s/img
ALFNet [32]	ResNet-50	17.8	6.1	0.05s/img
CSP [15]	ResNet-50	12.1	5.6	0.12s/img
PRNet [54]	ResNet-50	13.5	5.8	–
F-CSP [57]	ResNet-50	15.3	6.0	0.09s/img
FE-CSP	ResNet-50	10.3	5.2	0.06s/img

A→B refers to training on A and testing on B

Bold numbers indicate the best results

IoU is the ratio of the intersection and the union of the prediction box and the ground truth box, which measures the degree of overlap between the prediction box and the label box. the higher the IoU, the better the quality of the predicted bounding box. Table 6 shows that FE-CSP performs better than baseline CSP at higher IoU, outperforming the CSP algorithm on the reasonable setting, small setting, heavy occlusion setting and heavy setting by 1.65%, 1.49%, 4.44% and 4.07%. It is worth noting that at higher IoU, FE-CSP can produce higher quality bounding boxes than CSP at severe pedestrian occlusions.

4.5 Generalizability evaluation of the proposed method

To verify the model's generalization ability, we performed across datasets evaluation. We selected some state of the arts FRCNN [19], Vanilla FRCNN [19], Faster R-CNN [18], Cascade RCNN [22], ALFNet [32], CSP [15], PRNet [54] and F-CSP [57] for comparison experiments, and we present the results of the experiments on Caltech in Table 7.

First, we trained each detector on Citypersons and tested it on Caltech. As expected, all detectors suffered from performance degradation. However, except for the two-stage detector Cascade RCNN, FE-CSP has a better generalization ability than most the state of the arts. Specifically, FE-CSP is 7.3%, 1.5%, 7.5%, 1.8%, 3.2% and 5% higher on reasonable setting compared to Vanilla Faster RCNN, ALT-Net, CSP, PRNet and F-CSP respectively. Cascade RCNN uses Swin Transformer and HRNet as its backbone, which is a stronger backbone network, and secondly, Cascade RCNN is a two-stage detector, which is more refined, so it is robust, but the detection speed is very low. As shown in column 4 of Table 7, we trained and tested on Caltech and obtained 5.2% MR^{-2} , outperforming most detectors. In terms of detection speed, FE-CSP is slightly lower than ALFNet, because ALFNet is based on the improved SSD model, which has an inherent advantage in speed.

However, the accuracy is still lower. Overall, FE-CSP achieves a better speed-accuracy balance.

5 Conclusion

In this paper, to solve the problems of severe pedestrian occlusion, difficult detection of small objects and low detection speed, we propose to introduce a backbone network using a combination of channel attention and spatial attention to effectively establish long-range dependence and highlight useful features, which reduces the loss of a large amount of information. And we use the feature pyramid network to fuse the extracted underlying and high-layer information to obtain high-level semantic features and improve the multi-scale feature representation capability. Experiments demonstrate that FE-CSP effectively improves pedestrian detection performance and outperforms most pedestrian detectors on the challenging pedestrian detection benchmarks CityPersons and Caltech, achieving very competitive performance. However, there are still many areas where the model can be improved and optimized. When the pedestrians are very crowded, it is difficult for the model to accurately detect individual pedestrians, which may be due to the shortcomings of NMS post-processing. In the future, the model can be further explored for other detection tasks such as vehicle detection and face detection.

Funding This work is supported by the National Natural Science Foundation of China (Grant No. 61966035), the International Cooperation Project of the Science and Technology Department of the Autonomous Region (Grant No. 2020E01023), the Joint Foundation of the National Natural Science Foundation of China (Grant No. U1803261), the Autonomous Region Natural Science Foundation of China (Grant No. 2021D01C083) and Autonomous Region Science and Technology Program Youth Science Fund Project (Grant No. 2022D01C83).

Data Availability All data generated or analysed during this study are included in this published article.

Declarations

Conflict of interest The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

1. Huang L, Zhao X, Huang K (2019) Bridging the gap between detection and tracking: A unified approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3999–4009
2. Hattori H, Naresh Boddeti V, Kitani KM, Kanade T (2015) Learning scene-specific pedestrian detectors without real data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3819–3827

3. Hbaieb A, Rezgui J, Chaari L (2019) Pedestrian detection for autonomous driving within cooperative communication system. In: 2019 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6. IEEE
4. Wei H, Zhang Q, Qian Y, Xu Z, Han J (2022) Mtsdet: multi-scale traffic sign detection with attention and path aggregation. *Appl. Intell.* 64:1–13
5. Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. *Adv. Neural Informat. Process. Syst.* 29:1–5
6. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969
7. Huang R, Pedoem J, Chen C (2018) Yolo-lite: a real-time object detection algorithm optimized for non-gpu computers. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 2503–2510. IEEE
8. Loey M, Manogaran G, Taha MHN, Khalifa NEM (2021) Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection. *Sustain. Cities Soc.* 65:10260
9. Heuer F, Mantowsky S, Bukhari S, Schneider G (2021) Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 997–1005
10. Everingham M, Eslami S, Van Gool L, Williams CK, Winn J, Zisserman A (2015) The pascal visual object classes challenge: A retrospective. *Int J Comput Vis* 111(1):98–136
11. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. *European Conference on Computer Vision*. Springer, London, pp 740–755
12. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750
13. Tian Z, Shen C, Chen H, He T (2019) Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636
14. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578
15. Liu W, Liao S, Ren W, Hu W, Yu Y (2019) High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5187–5196
16. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587
17. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448
18. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv Neural Informat Process syst* 28:11–27
19. Zhang S, Benenson R, Schiele B (2017) Citypersons: A diverse dataset for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3221
20. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Occlusion-aware r-cnn: detecting pedestrians in a crowd. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 637–653
21. Wang X, Xiao T, Jiang, Y, Shao S, Sun J, Shen C (2018) Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7774–7783
22. Cai Z, Vasconcelos N (2019) Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Trans Patt Anal Mach Intell* 43(5):1483–1498
23. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. *European Conference on Computer Vision*. Springer, Berlin, pp 21–37
24. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788
25. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988
26. Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790


27. Chen Q, Wang Y, Yang T, Zhang X, Cheng J, Sun J (2021) You only look one-level feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13039–13048
28. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271
29. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
30. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
31. Song T, Sun L, Xie D, Sun H, Pu S (2018) Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 536–551
32. Liu W, Liao S, Hu W, Liang X, Chen X (2018) Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 618–634
33. Liu W, Liao S, Ren W, Hu W, Yu Y (2019) High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5187–5196
34. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Patt Anal Mach Intell* 20(11):1254–1259
35. Rensink RA (2000) The dynamic representation of scenes. *Visual Cognit* 7(1–3):17–42
36. Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Rev Neurosci* 3(3):201–215
37. Zhu X, Cheng D, Zhang Z, Lin S, Dai J (2019) An empirical study of spatial attention mechanisms in deep networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6688–6697
38. Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, Jia J (2018) Psanet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 267–283
39. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial transformer networks. *Adv Neural Inform Process Syst* 28:1–7
40. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803
41. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0
42. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141
43. Lu E, Hu X (2021) Image super-resolution via channel attention and spatial attention. *Appl Intell* 90:1–9
44. Lu Z, Xu B, Sun L, Zhan T, Tang S (2020) 3-d channel and spatial attention based multiscale spatial-spectral residual network for hyperspectral image classification. *IEEE J Select Topics Appl Earth Observat Remote Sens* 13:4311–4324
45. Chen J, Chen Y, Li W, Ning G, Tong M, Hilton A (2021) Channel and spatial attention based deep object co-segmentation. *Knowledge-Based Systems* 211:106
46. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19
47. Zhu X, Cheng D, Zhang Z, Lin S, Dai J (2019) An empirical study of spatial attention mechanisms in deep networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6688–6697
48. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125
49. Dollar P, Wojek C, Schiele B, Perona P (2011) Pedestrian detection: An evaluation of the state of the art. *IEEE Trans Patt Anal Mach Intell* 34(4):743–761
50. Liu S, Huang D, Wang Y (2019) Adaptive nms: Refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6459–6468
51. Zhang J, Lin L, Zhu J, Li Y, Chen Y-C, Hu Y, Hoi SC (2020) Attribute-aware pedestrian detection in a crowd. *IEEE Trans Multimed* 23:3085–3097

52. Zhang, Y, He H, Li J, Li Y, See J, Lin W (2021) Variational pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11622–11631
53. Tang Y, Li B, Liu M, Chen B, Wang Y, Ouyang W (2021) Autopedestrian: an automatic data augmentation and loss function search scheme for pedestrian detection. *IEEE Trans Image Process* 30:8483–8496
54. Song X, Zhao K, Chu W-S, Zhang H, Guo J (2020) Progressive refinement network for occluded pedestrian detection. *European conference on computer vision*. Springer, Berlin, pp 32–48
55. Song X, Chen B, Li P, Wang B, Zhang H (2022) Prnet++: Learning towards generalized occluded pedestrian detection via progressive refinement network. *Neurocomputing* 482:98–115
56. Dai J, Qi H, Xiong, Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773
57. Zhang T, Cao Y, Zhang L, Li X (2022) Efficient feature fusion network based on center and scale prediction for pedestrian detection. *Visu Comput* 6:1–8

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Yugang Qin¹ · Yurong Qian^{1,2,3}  · Hongyang Wei¹ · Yingying Fan² · Peiyun Feng¹

Yugang Qin
qinyugang1998@gmail.com

Hongyang Wei
weihy@stu.xju.edu.cn

Yingying Fan
fyy0327@stu.xju.edu.cn

Peiyun Feng
peiyun_feng@stu.xju.edu.cn

¹ School of Software, Xinjiang University, Urumqi 830000, China

² Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Urumqi 830000, China

³ College of Information Science and Engineering, Xinjiang University, Urumqi 830000, China