https://doi.org/10.1007/s40747-022-00728-3

ORIGINAL ARTICLE

Self-attention-guided scale-refined detector for pedestrian detection

Xinchen Lin[1] · Chaoqiang Zhao[1] · Chen Zhang[2] · Feng Qian[1]

Received: 15 November 2021 / Accepted: 3 February 2022 / Published online: 22 April 2022

Abstract

Pedestrian detection has been researched for decades. Recently, an anchor-free method CSP is proposed to generate the pedestrian bounding box directly. When the predicted center deviates from the ground truth in the testing phase, the CSP model generates deviated pedestrian bounding box, which leads to false detection in occlusion situations. To handle this problem, we refine the scale regression branch of the CSP model to generate a more accurate prediction. The new scale regression branch outputs the distances between the center and the four edges of the pedestrian bounding box. Even if the predicted center deviates from the ground truth, an accurate bounding box can still be obtained. Moreover, we integrate a self-attention module into our model to take full advantage of the features in different depth layers. Our proposed model achieves better performance than the state-of-the-art detectors in comparison experiments on the two datasets, i.e., Citypersons and Caltech.

Keywords Pedestrian detection · Anchor-free detection · Center scale prediction · Attention mechanism

Introduction

Pedestrian detection is a fundamental task in computer vision and has been researched for decades [1]. Pedestrian detection has wide applications in real life, such as autonomous driving [2,3], intelligent video surveillance [4,5] and human behavioral analysis [6,7]. Pedestrian detection is also appropriate for automatic monitoring and control in process of chemical engineering, such as hazardous area monitoring and operator behavior monitoring. Benefiting from the development of cloud computing [8,9], pedestrian detectors can be deployed on different mobile devices. With the advancement of deep learning in recent years, the research on pedestrian detection has progressed rapidly [10].

R-CNN framework [11] is utilized in many pedestrian detection models for its good performance in object detection. However, the models based on R-CNN framework have some disadvantages. First, proposals are regressed from predefined anchors in the RPN [12]. To achieve a good result, the anchor settings should be adjusted on the basis of data. As the scale of pedestrians varies in a large scope, the anchors should be designed carefully with experience. In practical applications, different anchor settings influence the result obviously. Second, the negative proposals are much more than the positive ones. The imbalance problem hinders the optimization of the model and slows down the training process. Thirdly, the testing speed of methods based on R-CNN is unsatisfied.

To address the above problems, CSP [13] constructs a fully convolutional detector, which utilizes high-level semantic features to generate the pedestrian bounding box directly. No anchors are pre-defined and the detecting speed is improved obviously. However, the predicted center cannot be entirely correct in the testing phase. The predicted center probably falls near the true center. The model should be trained to obtain a suboptimal prediction when the predicted center drifts around the ground truth. An approximation is used in the training phase of the scale regression branch. The locations around the center are all defined as positives. The

✉ Feng Qian
fqian@ecust.edu.cn

Xinchen Lin
linxinchenupup@163.com

Chaoqiang Zhao
zhaocqilc@gmail.com

Chen Zhang
Zhang.chen@secco.com.cn

[1] Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

[2] Shanghai SECCO Petrochemical Co., Ltd., Shanghai 201507, China

ground truth of these locations is set to the height of the pedestrian as the true center location. By training the scale regression branch in this manner, the model can obtain an approximate bounding box when the predicted center falls near the true center. This approximation generates a few offsets for the bounding box prediction and probably leads to false detection. Moreover, the CSP model only predicts the height of the pedestrian and calculates the width by multiplying a fixed aspect ratio (0.41). This approximation also reduces the detection accuracy in complex situations. To make a better prediction, the APD [14] model refines the center location branch by defining new ground truth. The four pixels around the real center in the heatmaps are all defined as positive examples. The positive examples used to regress the real center are expanded three times larger than the CSP model. Although the APD model predicts the center location more accurately, the false prediction caused by the deviated center is not solved. The improvement by refining the center location branch of the CSP model is limited.

In this paper, we construct a new scale regression branch, which outputs a scale heatmap with four channels. The channels correspond to the distances between the center and the four edges of the pedestrian bounding box. The motivation is illustrated in Fig. 1. The refined scale regression branch makes an accurate prediction when the predicted center falls near the true center. To detect pedestrians on various scales, the CSP model concatenates the features at different levels in the channel dimension. Although the concatenation operation is simple and effective, the features from different depth layers are not utilized sufficiently. To make full use of the concatenated futures, we attach a self-attention module [15] after the fused features. The attention module re-weights the concatenated features in spatial and channel dimension. To sum up, the main contributions of our paper are listed as follows.

1. A scale-refined CSP model (SASR-CSP) is proposed. A new scale regression branch is constructed to make an accurate prediction when the predicted center deviates from the ground truth. The detection accuracy of occluded pedestrians can be improved.
2. A self-attention module is integrated into our model to take full advantage of the concatenated features at different levels. The re-weighted features can further increase the detection accuracy.
3. We conduct comprehensive experiments to evaluate our proposed model on two standard pedestrian datasets, i.e., Caltech [16] and Citypersons [17]. Our model achieves better performance than the state-of-the-art detectors on the two datasets at different occlusion levels.

The remainder of this manuscript is organized as follows. "Related work" presents some related work in pedestrian detection. "Methods" gives a detailed description of our proposed SASR-CSP model. Then, we make ablation study and comparative experiments in "Experiments". Conclusions are drawn in "Conclusion and future work".

Related work

A literature review is made in this section. First, traditional pedestrian detection methods are introduced. Then, models based on deep learning are presented. Some of them are proposed to handle multi-scale and occlusion detection. Finally, anchor-free detection methods are reviewed.

Traditional detection methods

Traditional pedestrian detectors extract well-designed handcrafted features. HOG [18] constructs a histogram of oriented gradient in different cells. A detection window slides on the image to extract the HOG feature, and a SVM is used as the classifier. ICF [19] combines gradient histogram, gradient magnitude and color channel features. ICF can be computed efficiently based on the integral channel images, which are generated by linear and non-linear transformations. SIDF and SSF [20] are proposed to model the inherent attributes of pedestrians. SIDF captures the boundary between the foreground and background from the inner and outer sides. SSF characterizes the symmetrical property of pedestrians. However, hand-crafted features can only extract low-level information. The lack of high-level semantic features hinders the traditional models to achieve good performance.

Deep learning detection methods

In the past decade, deep learning pedestrian detectors have achieved good performance. He et al. [21] explore to use Faster R-CNN as a pedestrian detector. A Boosted Forest (BF) is employed to classify the proposals generated by the RPN [12]. The anchor setting of the RPN is adjusted to fit pedestrian detection. Similarly, Faster R-CNN achieves a good result by several modifications, such as RPN refining, image up-scaling and stride reduction of feature maps [17].

To further improve the performance, some extra information is utilized to train the model. HyperLearner [22] constructs a multi-task learning framework to detect pedestrians and generate distinct channel features simultaneously. However, the model cannot be trained without the ground truth of semantic segmentation. To handle this problem, SDS-RCNN [23] is proposed to realize simultaneous detection and segmentation for pedestrians. The areas inside the bounding box are viewed as positives, and the areas outside the bounding box are viewed as negatives. The approximate ground

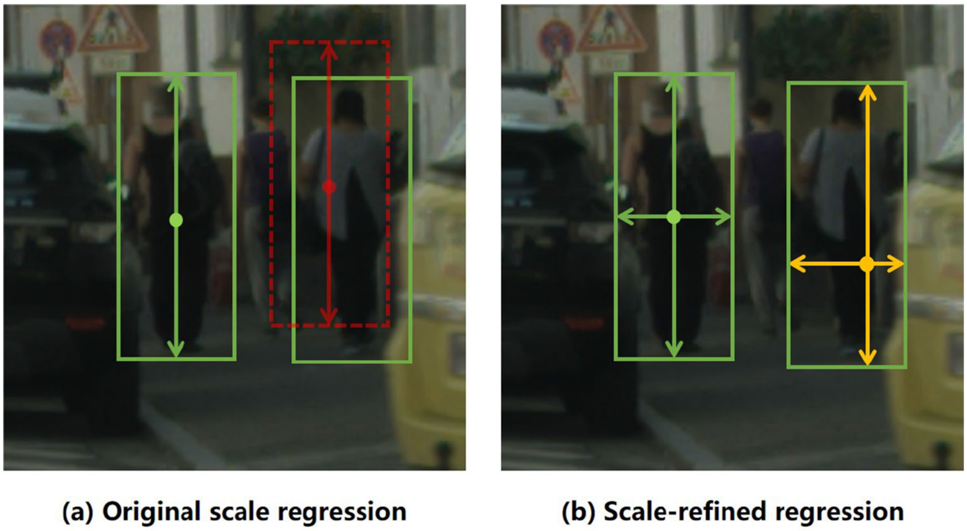

Fig. 1 The motivation of our scale-refined model. The CSP scale regression is illustrated in **a**. The CSP model predicts the height of pedestrians. When the predicted center deviates from the ground truth, the model makes an approximation to output the original height value for the current location. Our scale-refined regression is illustrated in **b**. Our model predicts the distances between the predicted center and the four edges of the bounding box. The green dot and bounding box represent the ground truth. The red ones in **a** represent the false detection caused by the deviated center. The yellow ones in **b** show that our model can make an accurate prediction even if the center is falsely predicted

truth of segmentation is used to train the model. The model achieves good performance without any other information.

Several models have been proposed to address the large-scale variation in pedestrian detection. SAF R-CNN [24] constructs two sub-networks to classify large and small proposals. The confidence scores obtained by the two sub-networks are merged by a learnable weighting layer. Some other models detect pedestrians on various scales in different depth layers of the network. This is because shallow features are more sensitive to small objects, and deep features are more sensitive to large objects. Following this idea, MS-CNN [25] and PAMS-FCN [26] generate large and small proposals in different depth layers. However, the receptive field cannot cover enough scales for the large and small proposals generated in a single depth layer. To solve this problem, GDFL [27] combines features in several adjacent stages of the backbone network by up-sampling and down-sampling. Moreover, a segmentation attention module is integrated into the model to suppress small or large objects in corresponding layers.

Occlusion is another challenge for pedestrian detection [28]. Pedestrians are usually occluded in complex situations. The occluded pedestrians are more likely to be neglected by detectors. A single-stage detector [29] is proposed based on DSSD [30] detection framework. The human body is divided into different cells. The part visibility scores can be estimated along with the pedestrian detection simultaneously. The part visibility scores of the cells are used to improve the detection accuracy. FC-net [31] generates an activation map by weighted summation of the convolutional features along the channel dimension. Then, calibrated features are generated by the activation map and improves the detection accuracy of occluded pedestrians. Bi-box [32] proposes a novel framework based on Fast R-CNN, which generates the bounding box and visible bounding box simultaneously. Bi-box first generates proposals by the RPN. Then, the proposals are sent to two paralleled branches with the same structure, i.e., the full body prediction branch and the visibility estimation branch. PBM [33] also utilizes visible information to train the model. Unlike Bi-box, PBM generates paired proposals. The paired proposals are used to predict the pedestrian and its visible region. The above two models are both trained by minimizing a multi-task loss, which prompts the models to detect occluded pedestrians more accurately. MGAN [34] is constructed by inserting a mask-guided attention (MGA) module into the Faster R-CNN framework. MGA takes the RoI features as input and outputs a visible probability mask. Then the mask is merged into the original features by an element-wise product. The attention module can stimulate MGAN to concentrate on the visible regions of pedestrians and suppress the background. Most of the deep learning methods are based on R-CNN framework. The anchor setting of the RPN should be carefully designed with experience. Moreover, the testing speed cannot meet the needs for real-time detection in practical applications.

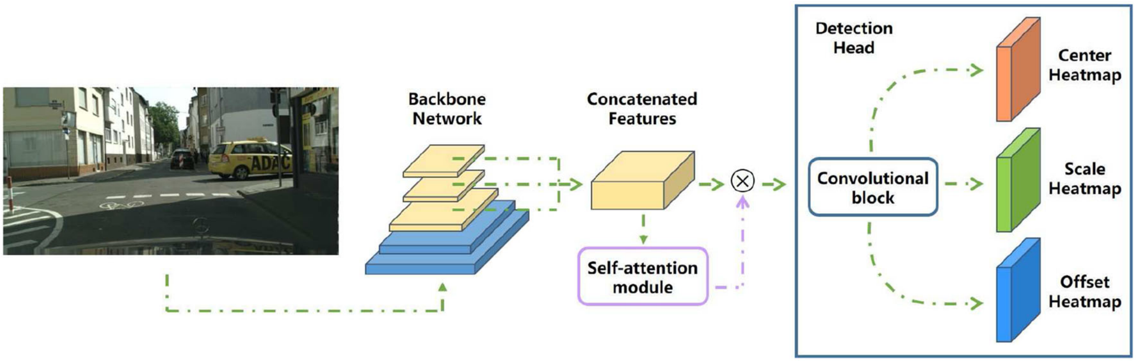

Fig. 2 The framework of our proposed SASR-CSP model. A ResNet-50 is selected as the backbone network. The features of the last three convolutional blocks are concatenated in channel dimension. The concatenated features are re-weighted by the self-attention module. Then, the features are sent to the detection head. The detection head predicts the bounding box by three branches, i.e., the center location branch, the scale regression branch, the down-sampling offset branch

Anchor-free detection

To address the problems of models based on the R-CNN framework, anchor-free detection is proposed as a new perspective of detection. Without defining any anchors, the anchor-free methods can generate the pedestrian bounding box directly. CornerNet [35] predicts the object by locating its left-top corner and bottom-right corner. Two prediction modules are used to outputs corner heatmaps and embeddings. The heatmaps represent the confidence score of each location to be a corner. The embeddings are used to match the paired corners which belong to the same object. To detect small-scale pedestrians more accurately, TLL [36] selects the somatic topological line as the annotation instead of the bounding box annotation. The topological line of the pedestrian is obtained by calculating the top and bottom vertex of the pedestrian bounding box. The new annotation can eliminate redundant background information and improve detection accuracy for small pedestrians. Tian et al. [37] propose FCOS, a single-stage anchor-free object detector. The model generates heatmaps to predict the locations and scales of the objects. Unlike TLL and CorNet, FCOS considers the region around the center as positive. All the pixels in the region are used to generate the bounding box. Similarly, CSP [13] constructs a simple fully convolutional pedestrian detector. Features in distinct levels are fused, and sent to a detection head. The detection head outputs heatmaps to predict the center and scale for the pedestrian. The location with a high confidence score in the center heatmap will be considered as the center of the pedestrian. The value of the same location in the scale heatmap represents the height of the pedestrian. The APD [14] model refines the CSP model by refining the center location branch. More positive examples are defined to regress the real center. A new branch is constructed to predict the density of pedestrians. Multi-task learning prompts the model to learn extra semantic features.

Since the CSP model makes an approximation when predicting the scales of pedestrians, the detection accuracy for occluded pedestrians is unsatisfactory. This approximation generates offsets between the predicted bounding boxes and ground truth. Moreover, the CSP model only predicts the height of the pedestrian, the width is calculated by a fixed aspect ratio approximately. The two approximations hinder the CSP model to achieve good performance in occlusion situations. Besides, the CSP model extracts features by concatenating features in different stages. The concatenation operation cannot utilize the information in different depth features sufficiently. We refine the scale regression branch to achieve better performance when occlusion happens. The feature extraction is also improved by a self-attention module. Compared to the APD model, our SASR-CSP model refines the scale regression branch instead of the center location branch to predict more accurate bounding boxes for pedestrians. Since the accuracy of predicted center locations for the CSP model is enough, refining the scale regression branch achieves better performance than the center location branch. Moreover, the APD model constructs a new density branch, which increases the number of parameters and model complexity.

Methods

In this section, we give a detailed introduction of our self-attention-guided scale-refined CSP (SASR-CSP) model. Following the CSP framework, our model extracts features by a backbone network. Then, bounding boxes are predicted by a detection head. The framework of our model is shown in Fig. 2. We first make a description of the backbone network in “Backbone network and feature extraction”. The feature extraction and self-attention module are introduced in this subsection. Then, the structure of the detection head and loss function are described in "Detection head and Loss function".

Backbone network and feature extraction

We use ResNet-50 [38] as the backbone network of our model. The weights are initialized by pretraining on the ImageNet. The ResNet-50 contains five convolutional blocks. Each block consists of several residual bottleneck modules. Considering pedestrians are usually small in size, the resolution of the feature maps should be increased. The down-sampling stride of our model is reduced to 16 as in [17]. Dilated convolution is employed to increase the receptive field. To detect pedestrians in multi-scales, the features of the last three convolutional blocks are concatenated in the channel dimension. Before concatenation, the features are first up-sampled to the same size by the transposed convolution. The kernel size is 4 and the stride is 4 and 2 for different depth features. Then, the features are normalized to the same scale. The concatenation operation is simple but will drop some valuable information. To take full advantage of the features in different depth blocks, we attach a self-attention module [15] after the concatenated features. The self-attention module consists of a channel attention module and a spatial attention module. The channel attention module explores the relationship between the channels. The spatial attention module captures the region of interest in the feature maps. The concatenated features pass through the two modules in sequence. The detailed structures of the two modules are shown in Fig. 3.

We make a brief description of the two modules. The details can be obtained in [15]. The tensor of the concatenated features is denoted as $\text{Feat}_{\text{cat}} \in R^{C\times H\times W}$. The C represents the number of channels after the concatenation. The H and W represent the height and width of the feature map. The channel attention module first aggregates features in spatial dimension by average and max pooling. A 1D vector $\text{Feat}_{\text{channel}} \in R^{C}$ can be obtained. Then, the features are re-weighted in channel dimension by element-wise product with $\text{Feat}_{\text{channel}}$. The diagram of the channel attention module is shown in Fig. 3a. Similarly, the spatial attention module first squeezes the features in channel dimension by average and max pooling. A 2D spatial attention map $\text{Feat}_{\text{spatial}} \in R^{H\times W}$ can be obtained. Then, the features are re-weighted in spatial dimension by element-wise product with $\text{Feat}_{\text{spatial}}$. The diagram of the spatial attention module is shown in Fig. 3b. The formula of the two modules can be denoted as follows:

$$A_{\text{channel}} = \text{Sigmoid}(\text{FC}(\text{Avg}(F))\oplus\text{FC}(\text{Max}(F))), \tag{1}$$

$$A_{\text{spatial}} = \text{Sigmoid}(\text{Conv}(\text{Avg}(F))(\text{cat})\text{Conv}(\text{Max}(F))), \tag{2}$$

$$F_A = \left(F \otimes \text{Broadcast}_{\text{spatial}}\ (A_{\text{channel}}\ )\right) \otimes \text{Broadcast}_{\text{channel}}\ \left(A_{\text{spatial}}\ \right), \tag{3}$$

where FC represents the fully connect layer. Avg and Max represent the average pooling layer and the max pooling layer. $\oplus$ and *cat* represent the element-wise addition and concatenation operation, receptively. Sigmoid and Conv represent the sigmoid activation layer and convolution layer.

The self-attention module only requires convolutional features without any other information. Moreover, the training and testing speed will not be impacted by the module. Benefit from the self-attention module, the concatenated features emphasize the meaningful regions or channels and suppress the irrelevant ones. The detection accuracy is improved on various scales and in occlusion situations.

Detection head and loss function

In this section, we present the detailed architecture of the detection head and the loss function. The detection head consists of a convolutional block and three paralleled branches, i.e., the center location branch, the scale regression branch and the down-sampling offset branch. The convolutional block reduces the channel dimension of the concatenated features. Then, the features are sent to the three branches. The branches output different heatmaps to predict the pedestrian bounding box. The details of the three branches are described as follows.

The center location branch

The center location branch generates a center heatmap by a convolutional layer. A sigmoid function is employed to confine the predicted values in the range of 0–1. The value represents the probability of each location to be the center of a pedestrian. The center location branch is trained as a binary classification task. The center of the pedestrian bounding box is viewed as positive. Other locations are viewed as negatives. However, the number of negatives is much more than the positives. The imbalance problem slows down the training process. Considering the characteristics of the center prediction task, most of the locations are far from the true center of the pedestrian. They are easy negative examples, which can be classified easily. Thus, we employ focal loss [14] to train the center location branch. The function is listed as follows:

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

where

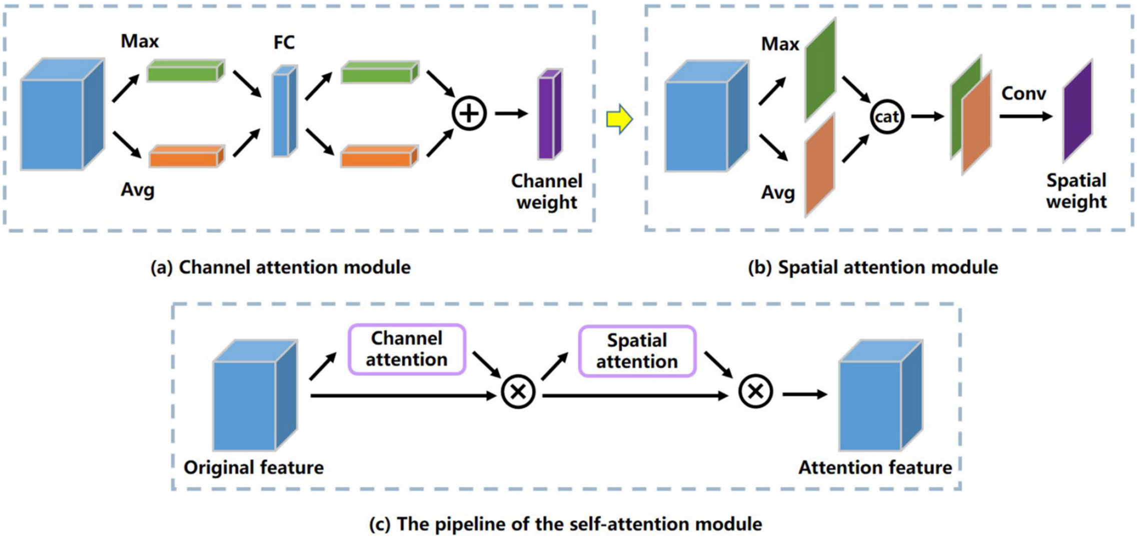

Fig. 3 The self-attention module consists of the channel attention module and the spatial attention module. The two modules are concatenated in sequence. The structure of the channel attention module is shown in **a**. The ⊕ represents the element-wise addition. The structure of the spatial attention module is shown in **b**. The cat means the concatenation in the channel dimension. The features are squeezed by MaxPooling (green) and AvgPooling (yellow). The pipeline of the self-attention module is shown in **c**

$$\hat{p}_{ij} = \begin{cases} p_{ij} & \text{if } y_{ij} = 1 \\ 1 - p_{ij} & \text{otherwise,} \end{cases} \tag{5}$$

$$\alpha_{ij} = \begin{cases} 1 & \text{if } y_{ij} = 1 \\ \left(1 - M_{ij}\right)^{\beta} & \text{otherwise.} \end{cases} \tag{6}$$

Specifically, the focal weight, i.e., $(1-\hat{p}_{ij})^{\gamma}$ assigns small weights to these easy examples, which contribute less to the final loss than the hard examples. As for the hard negative examples, i.e., the locations surrounding the center of the pedestrian, it is acceptable for these points to be predicted as the center. This false prediction only generates a few drifts from the ground truth. A mask M is introduced to reduce the contributions of these locations to the final loss. More details can be obtained in [13]. To make an accurate prediction for the bounding box when these hard negatives are selected as the center location, we refine the scale regression branch in the next subsection.

The scale regression branch

Since the focal loss used in the center location branch reduces the contributions of the hard negatives to the final loss, the surrounding locations of the center may be falsely predicted as positive examples. To address this problem, the original CSP model makes an approximation when predicting the scale of the pedestrian. These surrounding locations are defined as positives in the scale regression branch, and the ground truth values are the same as the height of the pedestrian. This approximation generates offsets for the obtained pedestrian bounding box, which reduces the detection accuracy in occlusion situations.

To achieve more accurate detection, our proposed scale-refined model constructs a new scale regression branch. A scale heatmap with four channels is generated. Each channel corresponds to the distance between the center and the edge of the pedestrian bounding box instead of the height. A ReLU activation layer is attached to confine the value to be positive. In our new branch, the center of the pedestrian and its surrounding pixels in radius 2 are defined as positive examples. First, all the positive examples in the heatmap are projected to the original image. The center pixel in the heatmap is projected to the real center of the pedestrian in the image. Since each of the surrounding pixels corresponds to a square region in the image, these surrounding pixels are projected to the center of the regions in the image. Then, the ground truth is calculated as the distances between these pixels and the four edges of the bounding box in the original image. The distances between these surrounding pixels and four edges can be calculated precisely. When the predicted center deviates from the ground truth, the correct bounding box can still be regressed by the hard negatives, i.e., locations surrounding the real center. Moreover, the new branch predicts the height and width of the pedestrian simultaneously instead of estimating the width by the height. The loss function of our new scale branch uses the GIoU loss [39]. The formula is presented as as follows:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}, \tag{7}$$

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|C \setminus (A \cup B)|}{|C|}, \quad (8)$$

where C is the smallest enclosing convex bounding box of A and B. $\setminus$ means the subtraction between the two areas. The GIoU means the generalized intersection over union of the two bounding boxes. The loss function is appropriate to the IoU based object detection metric.

Since pedestrians are occluded by each other in crowds, pedestrian bounding boxes are usually overlapped. The original CSP model probably predicts a pedestrian bounding box deviates from the ground truth. Then, the predicted bounding box will overlap with some other pedestrians. In this situation, some pedestrians are missed. Our proposed scale-refined regression branch can handle this problem effectively. By predicting the distances between the center and four edges, the correct bounding box can be obtained whenever the predicted center location is precise. More comprehensively, the new branch predicts the height and width simultaneously, which helps detect irregular pedestrian bounding boxes. The new scale regression branch can improve the detection accuracy for pedestrians in occlusion.

The down-sampling offset branch

Since the detection is based on the feature map generated by the backbone network, the ground truth bounding box in the original image is mapped to the feature map. The detection feature map is down-sampled from the image, and the down-sampling ratio is 4. When the center of the pedestrian in the image is mapped to the feature map, the coordinate of the center becomes a floating number. Then, the coordinate is rounded up to the integer location, which is the approximate center of the pedestrian. As in the center location branch, these approximate centers of pedestrians are also defined as the positive examples in the down-sampling offset branch. The ground truth of these examples is calculated as the distances between the real centers and approximate integral centers. The down-sampling offset branch is constructed to predict this offset value. The loss is calculated by the smooth L1 loss:

$$\mathrm{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise.} \end{cases} \quad (9)$$

The final loss is the weighted summation of the losses of the three branches:

$$L = \alpha L_{\mathrm{cen}} + \beta L_{\mathrm{sca}} + \gamma L_{\mathrm{off}}, \quad (10)$$

where the L_{cen}, L_{sca} and L_{off} are the loss of the center location branch, scale regression branch and down-sampling offset branch. The weight item α, β and γ are set to 0.01, 0.05 and 0.1 according to experience. The model can be trained end-to-end by optimizing the loss efficiently.

Our proposed SASR-CSP model refines the scale regression branch of the original CSP model. The new branch predicts the distances between the center and the edges of the bounding box. Considering the false prediction of the center location, the new scale branch generates a more accurate detection of the bounding box. The scale-refined branch can improve the detection accuracy of occluded pedestrians. The self-attention module is integrated into our model to take full advantage of the features in different depth layers, and further improves the performance.

Experiments

We first make a brief description of two standard datasets, i.e., Caltech [16] and Citypersons [17]. The evaluation metric is presented. Then, the implementation details of our SASR-CSP model are described. To demonstrate the effectiveness of our scale-refined regression branch and the self-attention module, we make some ablative study. Finally, our model is compared with the state-of-the-art detectors.

Datasets and metrics

Caltech [16] dataset contains a 10 h video at 30 Hz, which was collected in Los Angeles. The resolution of the video is 640 × 480. 250,000 frames are annotated with approximately 350,000 bounding boxes. Moreover, the visible region is also annotated for each pedestrian. The set00-set05 are used as the training set, and the set06-set10 are used as the testing set. Every 30th per frame in the dataset is selected. The training set contains 4250 images, and the testing set contains 4024 images. New annotations proposed in [40] are used for training and testing. Citypersons [17] dataset are collected in Germany and some surrounding countries. The resolution of the image is 1024 × 2048. The training set consists of 2975 images, which are collected in 18 cities. The validating set consists of 500 images, which are collected in 3 cities. The comparison of the two datasets is shown in Table 1. Compared to Caltech dataset, the density of pedestrians in Citypersons dataset is larger. Moreover, the occlusion pattern of pedestrians in Citypersons dataset is more complex than Caltech dataset. Thus, we select Citypersons dataset to make the ablative study.

We evaluate the model performance by the log-average miss rate MR^{-2} [16]. The metric is calculated by averaging miss rate at nine false positive per image rates in log-space in the range of 10^{-2} to 10^{0}. Comprehensive experiments are conducted on different occlusions levels. The datasets are divided into several subsets: "Bare", "Partial" and "Heavy". The occlusion ratios of the three subsets are: 0, (0, 0.35],

Table 1 Comparison between the statistics of the Caltech dataset and the Citypersons dataset

	Images	Bounding box	Person/image	Unique pedestrian
Training sets				
Caltech [16]	128,419	192,000	1.5	1300
Citypersons [17]	2975	19,654	6.6	20,000
Testing sets				
Caltech [16]	121,465	155,000	1.3	1000
Citypersons [17]	500	3938	7.9	4000

Table 2 Ablative experiments on Citypersons dataset

Method	Reasonable	Bare	Partial	Heavy
CSP	11.0	7.3	10.4	49.3
SR-CSP	10.6	7.2	9.8	**45.6**
SR-CSP+CA	10.3	7.2	9.3	46.8
SR-CSP+CA+SA	**9.8**	**6.8**	**9.2**	46.0

The bold contents are the best results

Table 3 Comparison of the model size and testing time

Method	Model size (MB)	Testing time (s)
CSP	160.1	0.0083
Our model (without attention)	160.1	0.0084
Our model (with attention)	160.4	0.0089

(0.35, 0.8]. The “Bare” subset and “Partial” subset are merged to “Reasonable” subset, which is usually used to represent the general performance of slightly occluded pedestrian detection. Following [16], only the pedestrians taller than 50 pixels are considered in evaluation.

Implementation details

Our model is constructed on the PyTorch [41] deep learning framework. The training process and testing process are carried a Quadro RTX 8000 GPU. The Adam [42] is utilized as the optimizer and a moving average strategy [43] is employed. In the training process of Caltech dataset, the model is trained for 120 epochs. The batch size is set to 16, and the learning rate is set to 1e−4. In the training process of Citypersons dataset, the model is trained for 150 epochs. The batch size is set to 8, and the learning rate is set to 2e−4. The random brightness jitter and horizontal flip are used to augment the dataset. To increase the robustness of the model to pedestrians on various scales, the image is resized to a random scale with a factor in range of [0.4, 1.5]. Then, the image is cropped/paved to the fixed scale of 640 × 1280 for Citypersons dataset, and 334 × 448 for Caltech dataset. The threshold for the center location prediction is set to 0.1.

Ablative experiments

In this section, we make an ablative study on our proposed SASR-CSP model. First, we test the effect of the new scale-refined regression branch. Then, the experiment is conducted to test the effect of the self-attention module.

Effect of the scale-refined regression branch

Since the approximation used in the scale regression branch, the CSP model probably generates false prediction of the pedestrian bounding box. The detection accuracy of the CSP model is unsatisfactory, especially in heavy occlusion. Our proposed new scale-refine regression branch predicts the distances between the center and four edges of the bounding box. Even if the predicted center deviates from the ground truth, our new branch can still output the correct pedestrian bounding box. To demonstrate the effectiveness of the new branch, we replace our proposed scale-refined branch with the original scale branch in the CSP model. The result is shown in Table 2. It can be seen that the new branch can improve the detection accuracy on all occlusion levels. The MR^{-2} values on the “Reasonable”, “Bare”, “Partial” and “Heavy” subsets are all improved. Especially on the “Heavy” subset, the MR^{-2} value is improved from 49.3 to 45.6. The new scale-refined branch makes more accurate prediction for the occluded pedestrian.

Effect of the self-attention module

Since features in different depth focus on various objects or regions, the CSP model concatenates the features in the last three convolutional blocks. Although the concatenation is simple, the operation cannot utilize the features in different depth sufficiently. The self-attention module re-weights the channels and spatial regions without any other information. To test the effect of the two modules, we select the scale-refined CSP model in the last subsection as a base detector. A channel attention module is first inserted before the detection head to re-weight the channels of the concatenated features. Then, a spatial attention module is attached in sequence. The comparison results are shown in Table 2. We can see that the

Fig. 4 Qualitative comparison between SASR-CSP and some state-of-the-art detectors. The last column in the second row is the ground truth. The middle column in the second row is the exemplar detection of our proposed SASR-CSP. The red bounding box represents the correct detection. The green bounding box represents the miss detection. The yellow bounding box represents false-positive detection

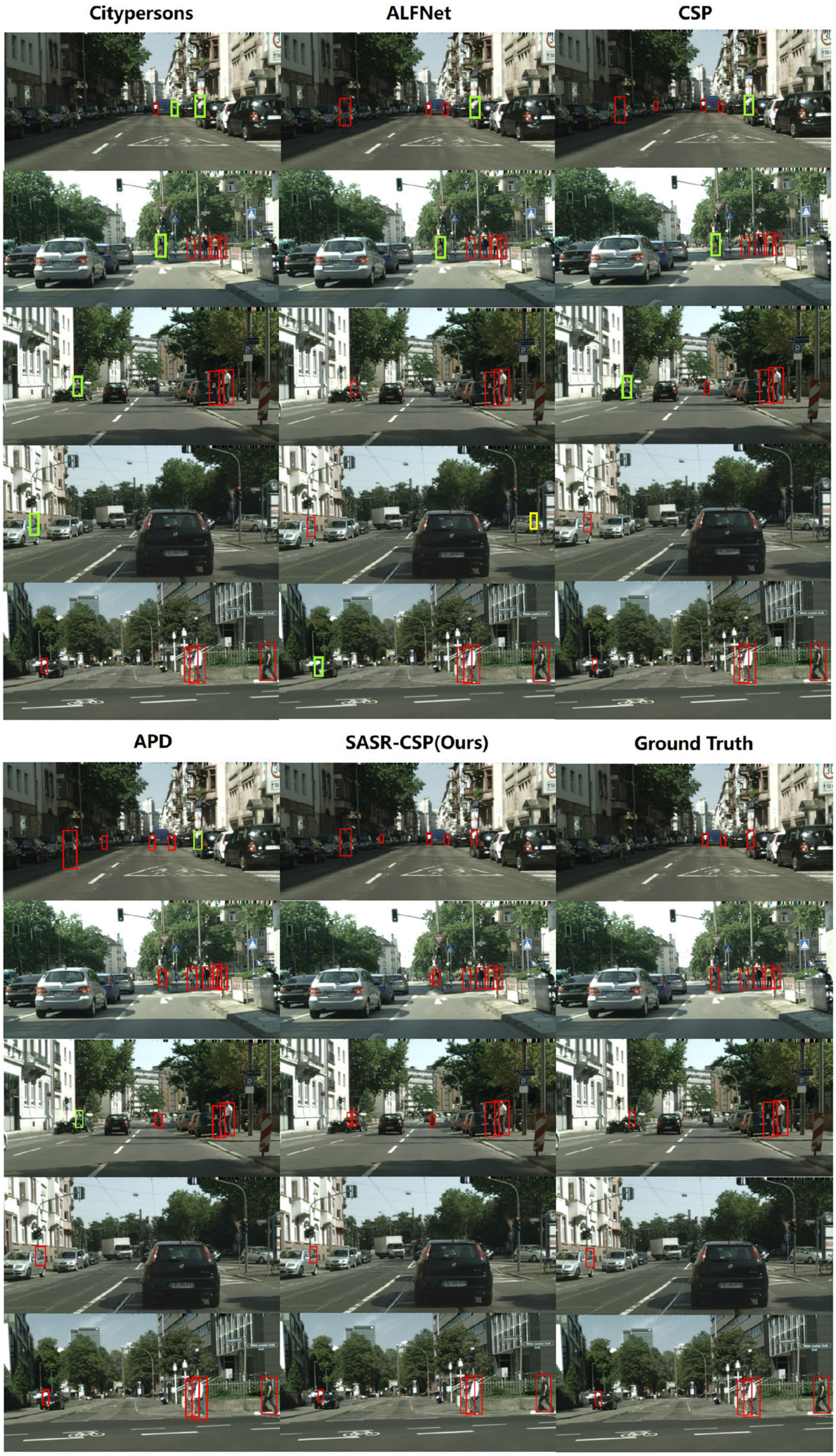

Table 4 Experiments on Citypersons dataset

Method	Reasonable	Bare	Partial	Heavy
Citypersons [17]	15.4	–	–	–
ALFNetV2 [44]	11.8	8.1	10.8	48.9
PBM [33]	11.1	–	–	53.3
CSP [13]	11.0	7.3	10.4	49.3
APD [14]	10.6	7.1	9.5	49.8
SSAM-RCNN [45]	10.9	7.4	10.8	47.5
PEN [46]	10.4	7.0	9.4	47.4
SASR-CSP	**9.8**	**6.8**	**9.2**	**46.0**

The bold contents are the best results

Table 5 Experiments on Caltech dataset

Method	Reasonable	Heavy
FasterRCNN+ATT [47]	8.6	39.1
SA-FastRCNN [24]	7.1	46.3
MS-CNN [25]	6.9	45.3
HyperLearner [22]	5.5	49.2
GDFL [27]	6.4	37.0
PAMS-FCN [26]	4.5	–
PEN [46]	4.2	–
SASR-CSP	**3.8**	**35.7**

The bold contents are the best results

channel attention module can improve the detection accuracy on the "Reasonable" and "Partial" subsets. The MR^{-2} values of the two subsets are improved to 10.3 and 9.3. The spatial attention module further improves the accuracy. The MR^{-2} values are improved to 9.8, 6.8 and 9.2 on the "Reasonable", "Bare" and "Partial" subsets. It should be noted that the self-attention module will decrease the detection accuracy on the "Heavy" subset to a certain extent. We suppose that the re-weighted features generated by the self-attention module will suppress some occluded body parts of the pedestrian.

To make a fair comparison, we compare the model size and testing speed of the CSP model and our proposed model (with/without the self-attention module). The results are shown in Table 3. It can be seen that the testing speed of our SASR-CSP model is almost the same as the CSP model. The self-attention module has little influence on the model size and testing speed of our model.

Some examples are selected to make a quantitative comparison between our proposed SASR-CSP and some state-of-the-art detectors in Fig. 4. The last column in the second row is the ground truth. The middle column in the second row is our model. The red bounding box represents the correct detection. The green bounding box and yellow bounding box represent the miss detection and false-positive detection, respectively. We can see that our model achieves better performance than the competitors. Our model detects occluded pedestrians more accurately than the competitors. This can be further proved by the quantitative comparison results in the next subsection.

Comparison with the state-of-the-arts

In this section, several state-of-the-art detectors are selected to compare with our proposed SASR-CSP model, i.e., MS-CNN [25], SA-FastRCNN [24], FasterRCNN+ATT [47], CSP [13], GDFL [27], HyperLearner [22], PAMS-FCN [26], SSAM-RCNN [45], ALFNetV2 [44], PBM [33], APD [14] and PEN [46]. Our SASR-CSP is compared with these detectors on Citypersons dataset and Caltech dataset. The MR^{-2} values on different occlusion subsets are calculated.

The detection results on Citypersons are listed in Table 4. Our model achieves better performance than all the state-of-the-art detectors on all the occlusion subsets. The MR^{-2} values of the four occlusion subsets are 9.8, 6.8, 9.2 and 46.0, respectively. The detection results on Caltech are listed in Table 5. It can be seen that our proposed SASR-CSP model outperforms the state-of-the-art detectors on the "Reasonable" and "Heavy" subsets. The MR^{-2} values are 3.3 and 33.6.

According to the results of comparison experiments, we can conclude that our proposed scale-refined regression branch is effective to handle occlusion. The new branch can improve detection accuracy for heavy occluded pedestrians obviously. The self-attention module takes full advantage of the concatenated features in different depth layers, and further improves detection accuracy.

Some representative images of Citypersons are selected to show the performance of our model in Fig. 5. It can be seen that our model can detect pedestrians accurately at different occlusion levels. Our model performs well in crowd situations (row 1). Some heavily occluded pedestrians can also be detected (row 2). In complex situations, i.e., small pedestrians in heavy occlusion or low brightness environment (row 3 and 4), our model probably misses some objects.

Conclusion and future work

In this paper, we propose a self-attention-guided scale-refined CSP model (SASR-CSP), which is based on the CSP framework. We construct a scale-refined regression branch by predicting the distances between the center and edges of the bounding box. The new branch obtains a precise prediction to replace the approximation prediction of the original scale regression branch in the CSP model. A self-attention module is employed in our model to make full use of the features in different depth layers. We conduct comprehensive

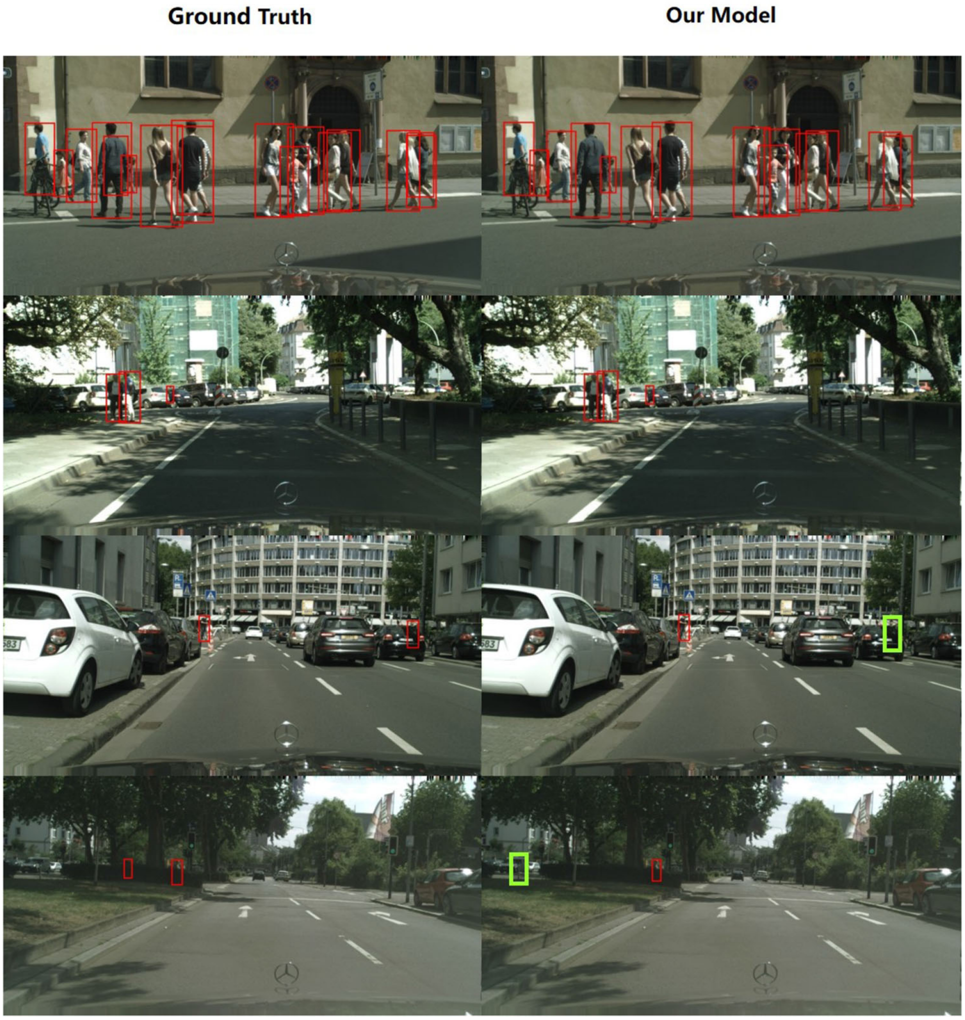

Fig. 5 Exemplar detections of our proposed SASR-CSP on Citypersons dataset. The first column is the ground truth. The red bounding box represents the correct detection. The green bounding box represents the miss detection. It can be seen that our proposed model performs well at different occlusion levels. The crowded pedestrians and small pedestrians can be detected accurately. Some heavily occluded pedestrians are missing

experiments on two datasets. Our proposed model achieves good performance on the two datasets on all the occlusion subsets. In future work, we will employ our proposed pedestrian detector to monitor the hazardous areas of chemical plants and recognize the behaviors of operators.

Funding Funding was provided by National Natural Science Foundation of China (61988101), National Natural Science Fund for Distinguished Young Scholars (61725301, 61925305), Programme of Introducing Talents of Discipline to Universities (B17017).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Nam W, Dollár P, Han JH (2014) Local decorrelation for improved pedestrian detection. In: Proceedings of Advances in neural information processing systems, pp 424–432
2. Liu L, Lu S, Zhong R, Wu B, Yao Y, Zhang Q, Shi W (2020) Computing systems for autonomous driving: state of the art and challenges. IEEE Internet Things J 8(8):6469–6486
3. Ghanem S, Kanungo P, Panda G, Satapathy SC, Sharma R (2021) Lane detection under artificial colored light in tunnels and on

highways: an iot-based framework for smart city infrastructure. Complex Intell Syst

4. Varga D, Szirányi T (2017) Robust real-time pedestrian detection in surveillance videos. J Ambient Intell Humaniz Comput 8(1):79–85
5. Han Q, Yin Q, Zheng X, Chen Z (2021) Remote sensing image building detection method based on mask r-cnn. Complex & Intelligent Systems
6. Khan MA, Kadry S, Parwekar P, Damasevicius R, Naqvi SR (2021) Human gait analysis for osteoarthritis prediction: a framework of deep learning and kernel extreme learning machine. Complex Intell Syst
7. Kareem Z, Zaidan A, Ahmed M, Zaidan B, Albahri O, Alamoodi A, Malik R, Albahri A, Ameen H, Garfan S et al (2021) An approach to pedestrian walking behaviour classification in wireless communication and network failure contexts. Complex Intell Syst
8. Fang W, Yao X, Zhao X, Yin J, Xiong N (2018) A stochastic control approach to maximize profit on service provisioning for mobile cloudlet platforms. IEEE Trans Syst Man Cybern Syst 48(4):522–534
9. Lin B, Zhu F, Zhang J, Chen J, Chen X, Xiong NN, Lloret Mauri J (2019) A time-driven data placement strategy for a scientific workflow combining edge computing and cloud computing. IEEE Trans Ind Inform 15(7):4254–4265
10. Hosang J, Omran M, Benenson R, Schiele B (2015) Taking a deeper look at pedestrians. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4073–4082
11. 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
12. Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
13. 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 5182–5191
14. Zhang J, Lin L, Zhu J, Li Y, Chen Y-c, Hu Y, Hoi CS (2020) Attribute-aware pedestrian detection in a crowd. IEEE Trans Multimed 23:3085–3097
15. 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
16. Dollar P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
17. 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 4457–4465
18. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 886–893
19. Dollár P, Tu Z, Perona P, Belongie S (2009) Integral channel features. In: Proceedings of the British machine cision conference, pp 91.1–91.11
20. Cao J, Pang Y, Li X (2016) Pedestrian detection inspired by appearance constancy and shape symmetry. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1316–1324
21. Zhang L, Lin L, Liang X, He K (2016) Is faster r-cnn doing well for pedestrian detection? In: Proceedings of the European conference on computer vision. Springer, pp 443–457
22. Mao J, Xiao T, Jiang Y, Cao Z (2017) What can help pedestrian detection? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3127–3136
23. Brazil G, Yin X, Liu X (2017) Illuminating pedestrians via simultaneous detection and segmentation. In: Proceedings of the IEEE international conference on computer vision. pp 4950–4959
24. Li J, Liang X, Shen S, Xu T, Feng J, Yan S (2018) Scale-aware fast r-cnn for pedestrian detection. IEEE Trans Multimed 20(4):985–996
25. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of the European conference on computer vision. Springer, pp 354–370
26. Yang P, Zhang G, Wang L, Xu L, Deng Q, Yang M-H (2021) A part-aware multi-scale fully convolutional network for pedestrian detection. IEEE Trans Intell Transport Syst 22(2):1125–1137. https://doi.org/10.1109/TITS.2019.2963700
27. Lin C, Lu J, Wang G, Zhou J (2018) Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of the European conference on computer vision
28. Ning C, Menglu L, Hao Y, Xueping S, Yunhong L (2021) Survey of pedestrian detection with occlusion. Complex Intell Syst 7(1):577–587
29. Noh J, Lee S, Kim B, Kim G (2018) Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 966–974
30. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: deconvolutional single shot detector. arXiv preprint arXiv:1701.06659
31. Zhang T, Ye Q, Zhang B, Liu J, Zhang X, Tian Q (2020) Feature calibration network for occluded pedestrian detection. IEEE Trans Intell Transport Syst
32. Zhou C, Yuan J (2018) Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of the European conference on computer vision. pp 135–151
33. Huang X, Ge Z, Jie Z, Yoshie O (2020) Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 10747–10756
34. Pang Y, Xie J, Khan MH, Anwer RM, Khan FS, Shao L (2019) Mask-guided attention network for occluded pedestrian detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 4967–4975
35. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision. pp 734–750
36. 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. pp 536–551
37. 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
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
39. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 658–666
40. Zhang S, Benenson R, Omran M, Hosang J, Schiele B (2016) How far are we from solving pedestrian detection? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1259–1267
41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8026–8037

42. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
43. Tarvainen A, Valpola H (2017) Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780
44. Liu W, Liao S, Hu W (2020) Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding. IEEE Trans Image Process 29:1413–1425. https://doi.org/10.1109/TIP.2019.2938877
45. Zhou C, Wu M, Lam S-K (2022) A unified multi-task learning architecture for fast and accurate pedestrian detection. IEEE Trans Intell Transport Syst 23(2):982–996
46. Jiao Y, Yao H, Xu C (2021) Pen: pose-embedding network for pedestrian detection. IEEE Trans Circuits Syst Video Technol 31(3):1150–1162. https://doi.org/10.1109/TCSVT.2020.3000223
47. Zhang S, Yang J, Schiele B (2018) Occluded pedestrian detection through guided attention in cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 6995–7003