Occluded pedestrian detection combined with semantic features

Binjie Ruan | Chongyang Zhang

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

Correspondence
Chongyang Zhang, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Email: zcy603@163.com

Abstract
The task of pedestrian detection is to identify the location and size of pedestrians in images or videos. However, occlusions are very common in real-life scenarios, which make pedestrian detection more difficult. In order to solve the occlusion problem in pedestrian detection, a semantic feature enhancement module to acquire more informative and richer semantic features is proposed. The detector enhances semantic features by fusing feature maps of different layers, and detects pedestrians based on their locations and scales. The Experiments performed on Caltech and CityPersons datasets show that the algorithm achieves superior performance for detecting occluded pedestrians, especially heavily occluded ones. 30.6% and 47.9% log-average missing rates are achieved in the heavily occluded subsets of Caltech and Citypersons, respectively. Moreover, this method is robust to the detection of heavily occluded pedestrians, and the module can be easily used by other detection frameworks.

1 | INTRODUCTION

With the rise of artificial intelligence technology, driverless technology is getting more and more attention from researchers. Pedestrian detection is an important part of driverless technology. Its main task is to identify the location and size of pedestrians in images or videos. However, factors such as diverse pedestrian postures, different clothes and appearances increase the difficulty of pedestrian detection. Low-level features such as edges, corners, and colours cannot detect pedestrians in complex scenes. If the detector cannot find the pedestrian in the image, then the safety of the pedestrian may be threatened during autonomous driving, so the performance of the pedestrian detection algorithm plays a critical role in real applications.

In recent years, the universal object detector has been developed rapidly, and has achieved great success in the detection task on the ImageNet [1], Pascal, and MS COCO datasets [2]. A growing number of researchers improve the universal object detectors to make them more suitable for pedestrian detection. Some of them customise architecture designs on the two-stage detector, Faster R-CNN [3], some study on the one-stage detector, SSD [4], and some improve the accuracy of the no-anchor detector, CenterNet [5]. We show here that the appropriately adjusted CSP [6] can also achieve promising detection results.

Although good performance has been achieved in detecting non-occluded or slightly occluded pedestrians on some pedestrian datasets, the performance of detecting heavily occluded pedestrians is still far from meeting the requirements of practical applications. Take the Cityperons [7] dataset as an example. One of the best-performing methods is PBM [8], achieves a miss rate of 11.1% for detecting non-occluded pedestrians, but the miss rate increased significantly to 53.3% for detecting heavily occluded pedestrians. However, occlusions are very common in real-life scenes. Pedestrians are blocked not only by other pedestrians, but also by other objects (such as cars, signs and buildings). In order to solve the problem of detecting different degrees of occluded pedestrians, we propose a semantic feature enhancement module, which can enhance the semantic information of extracted features. The detection head consists of three branches, one branch is used for size estimation, and the other two branches are used for position estimation. Our proposed method is called multi-features centre and scale prediction (MF-CSP) detector.

In summary, our main contributions are:

1. We propose a semantic feature enhancement module and prove that adding it in the feature extraction process can
2.2 | Methods using multiple layers

Many methods use different layers in convolution to improve the results of detection and semantic segmentation. [14] sums the predicted individual category scores on the multi-scale feature maps to obtain the semantic segmentation. [4] designs different scale object detectors for different feature layers to solve the detection problem of multiple scales. [15] uses a multi-scale pyramid hierarchy to build feature pyramids, developing a top-down architecture to extract feature maps that contain high-level semantic information.

A large amount of excellent works tell us that combining multi-layer outputs can effectively improve the detection of the model. We attempt to form more complex feature fusion on multi-layer feature maps in order to obtain more robust semantic information.

2.3 | Occlusion handling for pedestrian detection

Occlusion is very common in pedestrian detection, and for this reason, it is also one of the most widely concerned issues in the field of pedestrian detection. Some methods [16, 17] are specifically designed to handle situations where multiple pedestrians block each other. [18] solves pedestrian detection in dense scenes from the perspective of loss function. The new loss function can improve the detection of occluded pedestrians by reducing the distance between the predicted box and its corresponding real object, and increasing the distance between the predicted box and other surrounding objects. In order to solve the occlusion problem, [19] improves on the basis of faster R-CNN. For the same object, the network outputs two bounding boxes, one is a complete pedestrian box, and the other is the visible part of the pedestrian. A series of detectors [9, 20] use specific parts of the body to detect occluded pedestrians. [21] introduces an attention mechanism based on the original faster R-CNN, which assigns different weights to different channels, so as to improve the ability to detect occluded pedestrians. [22] proposes a multi-instance prediction concept to solve the problem of pedestrian detection in crowded scenes.

In this paper, we demonstrate that a properly adapted CSP can achieve excellent detection results in the detection of occluded pedestrians.

2.4 | Semantic features in pedestrian detection

Utilizing the semantic features of pedestrians in the images can help the model understand the concept of pedestrians and improve the detection ability. These features can be the head, torso and limbs of the pedestrian. [23] utilizes multi-resolution channels and semantic channels in the detection process. The detector benefits from semantic
information and can detect pedestrians better. [24] encodes the attention masks to convolutional feature maps, which helps the model to recognise body parts better and identify pedestrian areas effectively. [25] divides the human body into five parts. The model obtains semantic features of each part respectively, and then merges with the global features to obtain the final detection results. [6] eliminates the traditional anchor box settings and treats pedestrian centres as high-level semantic features of detection. [26] introduces semantic segmentation results as self-attention cues to improve the pedestrian detection performance.

We mainly focus on how to improve the semantic features extracted by the model, and use these features to locate pedestrians and predict their scales in the images.

3 | APPROACH

Pedestrian occlusion is divided into intra-class occlusion and inter-class occlusion [18], which may lead to different pose features and increase the difficulty of detection. We propose to introduce a semantic feature enhancement module in the feature extraction process, so that the network can extract semantic features that contain more information about human body structure.

3.1 | Overview

The CSP detector has been advanced in pedestrian detection. We use it as the base detector in our experiments, adjust the backbone network and add semantic feature enhancement modules. The adjusted model structure is shown in Figure 2. The overall structure of the pedestrian detection model is mainly composed of two parts: (1) Semantic feature extraction module. (2) Detection head module. The semantic feature extraction module should extract the semantic features of pedestrians. The detection head module determines the size and location of pedestrians based on the extracted semantic features. The network can be trained end-to-end by optimizing the following loss functions:

\[
L = \lambda_c \ast L_{cls} + \lambda_r \ast L_{reg} + \lambda_o \ast L_{offset},
\]  

(1)

where \(L_{cls}\) is the cross entropy loss used for pedestrian centre point classification, \(L_{reg}\) is the L1 loss for bounding box regression, and \(L_{offset}\) is the L1 loss for adjusting the centre point, the calculation of the three loss functions will be detailed in a later subsection. In the experiment, the values of \(\lambda_c, \lambda_r\) and \(\lambda_o\) are 0.01, 1, 0.1, respectively.

3.2 | Semantic feature extraction module

The backbone network of the semantic feature enhancement module is the ResNest [27] model. The Split-Attention module in the network groups the channel dimensions of the feature map and calculates the attention weights corresponding to each group of features, which effectively improves the quality of features extracted from the model without significantly increasing the model computation and parameters.
The semantic feature extraction module is shown in Figure 2. The downsampling of images is divided into five stages. Then the feature maps of the fifth stage and the sixth stage are upsampled to obtain the feature maps of the sixth and seventh stages, respectively.

Now the feature maps of the second stage to the seventh stage are respectively denoted as $\theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7$. After shallow convolution, the feature maps contain more low-level features such as target edges and corners, which are useful for locating pedestrians. As the number of convolution layers increasing, the size of receptive field also increases, which results in the feature maps containing more semantic features of the objects, and the extracted information is more abstract, which is suitable for pedestrian recognition. Therefore, in order to further improve the quality of the features extracted, the model adopts a more complex feature fusion strategy. We first fuse $\theta_3$ and $\theta_7$, $\theta_4$ and $\theta_6$ to obtain $\theta_{\text{sma}}, \theta_{\text{mid}}, \theta_5$ directly as $\theta_{\text{big}}$. $\theta_{\text{sma}}, \theta_{\text{mid}}$ are suitable for detecting small, medium and large objects, respectively. Then, perform further feature fusion on $\theta_{\text{sma}}, \theta_{\text{mid}}$ and $\theta_{\text{big}}$ to obtain the final feature map $\theta_{\text{det}}$ for detection. Since the resolution of each feature map of $\theta_{\text{sma}}, \theta_{\text{mid}}$ and $\theta_{\text{big}}$ is different, we use L2 standardization to adjust the standard deviation of each feature map to 10, and then use deconvolution to adjust the resolution of these feature maps to the same size.

Assuming that the size of the original image is $H \times W$, the size of the feature map finally input to the detection head is $H/r \times W/r$, where $r$ is the downsampling factor. The value of $r$ is 4 in the experiment.

We visualise feature maps with or without semantic feature enhancement modules. As shown in Figure 3, the first row is the detection result of some images in the Citypersons validation set, the second row is the feature map with the semantic feature enhancement module added, and the third row is the feature map without the semantic feature enhancement module. The highlighted areas in the feature map indicate the location of the pedestrian centre points. By comparing the second row with the three rows, we can find that the pedestrian centre points of the feature maps of the second row are more obvious, indicating that the semantic feature enhancement module can enhance the semantic features of the feature map.

### 3.3 Detection head module

The detection head module mainly solves two problems: (1) Where are the pedestrians? (2) What is the scale of each pedestrian? For the accuracy of pedestrian positioning, there are two branches of the detection module to solve the first problem, and another branch to solve the second problem.

The three branches parse the feature maps extracted by the previous feature extractor to obtain the detection results. First, we use a $3 \times 3$ convolution kernel to reduce the channel size to 256, which helps reduce subsequent calculations and parameters of the model, and then use three $1 \times 1$ convolution kernels to obtain centre heatmap, scale heatmap and offset heatmap, respectively.

Each value of the centre heatmap represents the probability that the point is the centre of the pedestrian. The position of this point on the feature map is multiplied by the downsampling factor $r$, combined with the offset predicted by another branch, that is, the position of the pedestrian centre point on the original
image. The formula for classifying losses is as follows:

\[
L_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \left[ y_i (1 - \hat{y}_i)^\gamma \log(\hat{y}_i) + (1 - y_i) \hat{y}_i^\gamma \log(1 - \hat{y}_i) \right],
\]

where \( N \) represents the number of points on the centre heatmap, that is, the resolution of the centre heatmap. \( y_i \) indicates whether the point is a centre point, \( y_i \in \{0, 1\} \). \( \hat{y}_i \) is the predicted value for that point, \( \hat{y}_i \in (0, 1) \). \( \gamma \) is a hyper-parameter for partitioning difficult or easy to classify samples. If the difference between the values of \( y_i \) and \( \hat{y}_i \) is too large, it means that the sample is difficult to distinguish, the classification loss increases, and the model should pay more attention to the learning of such samples. \( \gamma \) takes the value of 4 in the experiment.

Scale heatmap is mainly used to predict the logarithmic value of pedestrian height. Since the aspect ratio of upright pedestrians is roughly fixed, the width of pedestrians can be calculated by a fixed factor. On the one hand, the number of prediction branches can be reduced. On the other hand, the model parameters can be reduced. We perform K-means cluster analysis on all pedestrian samples in the Citypersons training set to obtain the cluster centroid. Based on the aspect ratio of the centre point, the width of the pedestrian in the sample is about 0.41 of the height. Thus, \( L_{\text{reg}} \) can be formulated as:

\[
L_{\text{reg}} = \text{SmoothL1}(\lg(\hat{h}), \lg(h)),
\]

In the experiment, the logarithm of height is used as the target of regression.

The model predicts the centre point in the downsampling feature map, which inevitably rounds the position in the original image. Therefore, the other branch is used to adjust the deviation value between the predicted centre point and the actual centre point to improve the accuracy of the predicted centre point. \( L_{\text{offset}} \) can be formulated as:

\[
L_{\text{offset}} = \text{SmoothL1}(x_{\text{offset}} - \hat{x}, y_{\text{offset}} - \hat{y}),
\]

\[
x_{\text{offset}} = \frac{x_1 + x_2}{2r} - \left\lfloor \frac{x_1 + x_2}{2r} \right\rfloor,
\]

\[
y_{\text{offset}} = \frac{y_1 + y_2}{2r} - \left\lfloor \frac{y_1 + y_2}{2r} \right\rfloor,
\]

where \((x_1, y_1)\) and \((x_2, y_2)\) are the coordinates of the upper left and lower right corners of the real boundary frame, respectively. \( r \) is the downsampling factor. \( \hat{x} \) and \( \hat{y} \) are the predicted values of the model.

3.4 Generate bounding box

For an original image with the size of \( H \times W \), the model can get the centre heatmap, scale heatmap and offset heatmap with resolutions of \( \frac{H}{c} \times \frac{W}{c} \), which means that each image has \( \frac{H \times W}{2r} \) grid points. The value of each point in the centre heatmap represents the probability that the point is a pedestrian centre. Since the predicted value will be activated by the sigmoid function, the prediction result is between 0 and 1. In the process of generating the bounding box, if the value of grid point is greater than the threshold, the model will treat it as a pedestrian centre point. In the experiment, the threshold is 0.5.

In order to visualise the centre heatmap, we multiply each value on the centre heatmap by 255 and perform rounding operations. The visualisation result of the centre heatmap is shown in Figure 4(b), and the original image is shown in Figure 4(a). Similarly, each grid point on the scale heatmap can predict the logarithm of pedestrian height. The predicted value of this point is valid only when the value of the corresponding position of the centre heatmap is greater than the threshold. The channel size of the offset heatmap is 2, which are the \( x \)-coordinate offset heatmap and the \( y \)-coordinate offset heatmap. Each grid point predicts the horizontal and vertical offset of the current centre point.

As shown in Figure 5, it shows the process of generating the bounding box. The red point is the pedestrian centre point predicted by the centre heatmap, and the green point is the pedestrian centre point corrected by the offsets. Then, according to the scale heatmap, we obtain the corresponding pedestrian height information. The height is multiplied by a fixed factor to get the width of the pedestrian, and finally generate the bounding box of this pedestrian. We sort the bounding boxes according to the probability of pedestrian centre points, and remove the boxes with high overlap rate by non-maximum suppression operation.

4 EXPERIMENTS

In this section, we first briefly describe the datasets used in the experiments, followed by the evaluation metrics we use. Then we show the detection performance of different algorithms on each dataset and compare the network of this paper with existing networks. Finally, we will show the results of the ablation experiments in order to demonstrate the usefulness of our proposed structure.

4.1 Datasets

The Caltech [28] dataset contains 10 h of driving videos in Los Angeles, and researchers have annotated pedestrians in these videos. In this paper, the training set uses a total of 42788 images extracted from set00-set05, and the test set is a total of 4024 images extracted from set06-set10. In the experiments, we use new annotation information of these images provided by [7].

Citypersons [7] is a pedestrian dataset extracted from the Cityspaces [29] dataset. These images are collected from many European countries or cities. The most important thing is that it contains a large number of pedestrian samples with varying
degrees of occlusion, and the data is rich in diversity. The experiment on Citypersons in this paper is to train on the official training set containing 2795 images, and test on the validation set containing 500 images.

4.2 Evaluation metrics

In our experiments, we use the standard average-log miss rate (MR), which is computed in the FPPI range of \([10^{-2}, 10^0]\) [28].

![Visualisations of centre heatmap, scale heatmap and offset heatmap](image)

**FIGURE 4** Visualisations of centre heatmap, scale heatmap and offset heatmap

The lower the log-average miss rate, the better the prediction performance of the model. Since the experiment focuses on the detection effect of each algorithm on occluded pedestrians, the occlusion situations are divided into different levels according to the height and occlusion rate of pedestrians. Table 1 shows the criteria for the classification of occlusion levels.

| Level   | Height (PXs) | Occluded (%) |
|---------|--------------|--------------|
| Reasonable | > 50        | < 35         |
| Partially | > 50 1<∼ 35 |            |
| Heavy   | > 50 35<∼ 80 |            |

**TABLE 1** Evaluation settings for each subset of the pedestrian dataset

4.3 Comparison with State-of-the-arts

4.3.1 Results on Caltech

In order to verify the validity of the proposed algorithm, five algorithms with good performance on Caltech are selected for comparison with the algorithm proposed in this paper. It can be seen from Table 2 that the detection algorithm proposed in this paper achieves the lowest missed detection rate of each subset, which is improved by 0.1%, 1.5% and 5.9%, respectively, compared with the second-ranked method.

At present, the detection accuracy of various algorithms on Caltech is close to saturation, so it is difficult to improve the detection ability on the reasonable subset. The log-average miss rate of MF-CSP in the reasonable subset is 5.8%, which is 0.8% higher than RepLoss. However, the log-average miss rate of the...
model pre-trained on CityPersons decreases by 2.1% and our model ranks first in the reasonable subset. MF-CSP pre-trained on CityPersons shows a significant improvement in occluded pedestrian detection, and the log-average miss rate decreases by 15.1%. The main reasons are: On the one hand, MF-CSP first combines low-level features that are conducive to object positioning and high-level features that are conducive to object recognition to obtain high-quality semantic features. Then the detector judges the centre of pedestrians point by point on the centre heatmap, which effectively improves missed rate. On the other hand, because CityPersons contains a wealth of occluded pedestrian samples, using CityPersons for pre-training models can effectively improve the ability to detect heavily occluded pedestrians.

### 4.3.2 Results on CityPersons

To verify the robustness of the algorithm, we perform comparison experiments on the CityPersons validation set. As can be seen from Table 3, MF-CSP achieves the best result on the partial and heavy subsets. In especially detecting heavily occluded pedestrian samples, the log-average miss rate decreases by 1.4% over the second-ranked CSP. The result shows that the detection algorithm proposed in this paper can indeed detect heavily occluded pedestrians.

However, the log-average miss rate of the Bare subset is not as good as OR-CNN. Since the occlusion rate of pedestrians in the Bare subset is less than 10%, the outline of the human body is relatively clear when there is a small amount of occlusion. OR-CNN divides pedestrian targets into five parts, and local features are obtained according to the prior information of the human body layout, so the missed detection rate in this case is low.

### 4.4 Ablation experiment

In order to verify the effective of the feature fusion method proposed in this paper on pedestrian detection, ResNest-50 is used as the backbone network, but after the model downsampling the image, it directly uses \(\theta_3, \theta_4\) and \(\theta_5\) to perform feature fusion to obtain \(\theta_{det}\). It can be seen from Table 4 that the model with the semantic feature enhancement module has lower results in each subset. Especially for heavily occluded pedestrian detection, the log-average miss rate decreases by 3.31%, indicating that the feature fusion method proposed in this paper can effectively improve the model’s detection performance for heavily occluded pedestrians.

We also compare the effects of different backbone networks on the performance of model detection. It can be seen from Table 5 that ResNet has the best detection performance in each subset. Especially in the heavily occluded subset, the miss rate of ResNet is 2.74% and 5.4% lower than that of ResNet [33] and SKNet [34], respectively. ResNet may be more suitable for extracting features in images than other backbone networks.

The fusion strategy of different feature layers will also affect the detection accuracy of the model. Table 6 shows the detection results of fusion of different layers. Since different feature maps are suitable for detecting objects of different scales, the lack of any feature maps will reduce the detection accuracy of the model. In particular, \(\theta_{sma}\) has the greatest impact on the detection results. If there is no \(\theta_{sma}\), the log-average miss rate in

### TABLE 2
Comparison to state-of-the-art detectors on the Caltech dataset; bold highlights the best results on each subset

| Method            | Reasonable MR (%) | All MR (%) | Heavy MR (%) |
|-------------------|-------------------|------------|--------------|
| RPN+BF [11]       | 7.3               | 59.9       | 54.6         |
| RepLoss [18]      | 5.0               | 59.0       | 47.9         |
| RepLoss+city [18] | 4.0               | 58.6       | 41.8         |
| OR-CNN+city [25]  | 4.1               | 58.8       | 45.0         |
| ALFNet [30]       | 6.1               | 59.1       | 51.0         |
| ALFNet+city [30]  | 4.5               | 56.8       | 43.4         |
| CSP [6]           | 4.5               | 56.9       | 45.8         |
| CSP+city [6]      | 3.8               | 54.4       | 36.5         |
| MF-CSP [ours]     | 5.8               | 54.8       | 45.7         |
| MF-CSP+city [ours]| 3.7               | 52.9       | 30.6         |

### TABLE 3
Comparison to state-of-the-art detectors on the CityPersons dataset; bold highlights the best results on each subset

| Method     | Backbone | Reasonable MR (%) | Bare MR (%) | Partial MR (%) | Heavy MR (%) |
|------------|----------|-------------------|-------------|----------------|--------------|
| OR-CNN [25]| VGG-16   | 12.8              | 6.7         | 15.3           | 55.7         |
| CC-CNN [31]| VGG-16   | 11.8              | 8.2         | 14.1           |              |
| RepLoss [18]| ResNet-50| 13.2              | 7.6         | 16.8           | 56.9         |
| TLL [32]   | ResNet-50| 15.5              | 10.0        | 17.2           | 53.6         |
| ALFNet [30]| ResNet-50| 12.0              | 8.4         | 11.4           | 51.9         |
| CSP [6]    | ResNet-50| 11.0              | 7.3         | 10.4           | 49.3         |
| MGAN [24]  | VGG-16   | 11.5              |             | -              | 51.7         |
| PBM [8]    | VGG-16   | 11.1              |             | -              | 53.3         |
| MF-CSP[ours]| ResNet-50| 11.2              | 7.2         | 10.4           | 47.9         |
TABLE 4  Comparisons of semantic feature enhancement (SFE) modules with or without

|       | Reasonable | Bare | Partial | Heavy |
|-------|------------|------|---------|-------|
| +SFE  | 12.74      | 8.16 | 11.99   | 51.23 |
| ✓     | 11.15      | 7.23 | 10.41   | 47.92 |

TABLE 5  Comparisons of different backbone networks

| Backbone | Reasonable | Bare | Partial | Heavy |
|----------|------------|------|---------|-------|
| ResNet   | 11.70      | 7.81 | 11.19   | 50.66 |
| ResNest  | 11.15      | 7.23 | 10.41   | 47.92 |
| SKNet    | 15.14      | 9.81 | 15.29   | 53.32 |

each subset increases significantly, and the log-average miss rate of the reasonable subset increases by 1.96%.

In order to verify the influence of the offset heatmap on the accuracy of the model, we compare the accuracy of model detection with or without offset adjustment. It can be seen from Table 7 that the detection accuracy of each subset has been improved after adjusting the predicted centre point.

4.5  Discussions

Note that MF-CSP detects pedestrians in the image based on the position of the pedestrian’s centre point and height of this pedestrian. However, when the centre points of two pedestrians in the image are very close, the model may be confused about the blurred centre and cannot detect the occluded pedestrian, as shown in Figure 6(a). We try to reduce the threshold of the centre point when generating the bounding box. When the threshold drops from 0.5 to 0.1, the model can detect occluded pedestrians, as shown in Figure 6(b).

However, it is not always feasible to lower the threshold to improve the model’s detection performance of occluded pedestrians, because if the threshold is too low, false detection may occur. We believe that effective methods should to be increasing the model’s confidence in occluded pedestrians. For example, the model can use the semantic features of other visible parts of the pedestrian’s body to further increase the confidence of occluded pedestrians, which will further reduce the missed detection rate.

5  CONCLUSION

For the problem of occluded pedestrian detection, we propose a detector combined with semantic features. By using multi-feature fusion, richer semantic features can be obtained, which is conducive to the positioning and recognition of the detector. Experiments show that the algorithm proposed in this paper achieves the lowest missed detection rate in the heavily occluded pedestrian detection of Caltech and CityPersons, and has a good application prospect due to its strong robustness. This paper has not considered the running speed of the model. Our next step will study how to reduce the number of model parameters and increase the running speed while ensuring the accuracy of model detection.

ORCID

Binjie Ruan  https://orcid.org/0000-0002-2257-8239
Chongyang Zhang  https://orcid.org/0000-0001-7325-7200

REFERENCES

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 60(6), 84–90 (2012)
2. Lin, T.Y., et al.: Microsoft coco: common objects in context. European Conference on Computer Vision, pp. 740–755. Springer, Cham (2014)
3. Ren, S., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(6), 1137–1149 (2017)
4. Liu, W., et al.: Ssd: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer, Cham (2016)
5. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv:1904.07850, (2019)
6. Liu, W., et al.: High-level semantic feature detection: a new perspective for pedestrian detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5187–5196. IEEE, Piscataway, NJ (2019)
7. Zhang, S., Benenson, R., Schiele, B.: Citypersons: a diverse dataset for pedestrian detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3213–3221. IEEE, Piscataway, NJ (2017)
8. Huang, X., et al.: NMS by representative region: towards crowded pedestrian detection by proposal pairing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 10750–10759. IEEE, Piscataway, NJ (2020)
9. Tian, Y., et al.: Deep learning strong parts for pedestrian detection. In: Proc. of the IEEE Int. Conf. on Computer Vision, pp. 1904–1912. IEEE, Piscataway, NJ (2015)
10. Tian, Y., et al.: Pedestrian detection aided by deep learning semantic tasks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 5079–5087. IEEE, Piscataway, NJ (2015)
11. Zhang, L., et al.: Is faster R-CNN doing well for pedestrian detection? In: European Conf. on Computer Vision, pp. 443–457. Springer, Cham (2016)
12. Cai, Z., et al.: A unified multi-scale deep convolutional neural network for fast object detection. In: European Conf. on Computer Vision, pp. 354–370. Springer, Cham (2016)
13. Li, J., et al.: Scale-aware fast R-CNN for pedestrian detection. IEEE Transactions on Multimedia 20(4), 985–996 (2017)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3431–3440. IEEE, Piscataway, NJ (2015)
15. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2117–2125. IEEE, Piscataway, NJ (2017)
16. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3198–3205. IEEE, Piscataway, NJ (2013)
17. Pepik, B., et al.: Occlusion patterns for object class detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3286–3293. IEEE, Piscataway, NJ (2013)
18. Wang, X., et al.: Repulsion loss: detecting pedestrians in a crowd. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 7774–7783. IEEE, Piscataway, NJ (2018)
19. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: Proc. of the European Conf. on Computer Vision, pp. 135–151. Springer, Cham (2018)
20. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: Proc. of the IEEE Int. Conf. on Computer Vision, pp. 3486–3495. IEEE, Piscataway, NJ (2017)
21. Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNs. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 6995–7003. IEEE, Piscataway, NJ (2018)
22. Chu, X., et al.: Detection in crowded scenes: one proposal, multiple predictions. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 12214–12223. IEEE, Piscataway, NJ (2020)
23. Daniel Costea, A., Nedevschi, S.: Semantic channels for fast pedestrian detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2360–2368. IEEE, Piscataway, NJ (2016)
24. Pang, Y., et al.: Mask-guided attention network for occluded pedestrian detection. In: Proc. of the IEEE Int. Conf. on Computer Vision, pp. 4967–4975. IEEE, Piscataway, NJ (2019)
25. Zhang, S., et al.: Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: Proc. of the European Conf. on Computer Vision, pp. 637–653. Springer, Cham (2018)
26. Zhou, C., Wu, M., Lam, S.K.: SSA-CNN: Semantic self-attention CNN for pedestrian detection. arXiv:1902.09080, (2019)
27. Zhang, H., et al.: Resnest: Split-attention networks. arXiv:2004.08955, (2020)
28. Dollar, P., et al.: Pedestrian detection: An evaluation of the state of the art. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 34(4), 743–761 (2011)
29. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3213–3223. IEEE, Piscataway, NJ (2016)
30. Liu, W., et al.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proc. of the European Conf. on Computer Vision, pp. 618–634. Springer, Cham (2018)
31. Zhao, Y., Yuan, Z., Chen, B.: Training cascade compact cnn with region- ious for accurate pedestrian detection. IEEE Transactions on Intelligent Transportation Systems 21(9), 3777–3787 (2019)
32. Song, T., et al.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proc. of the European Conf. on Computer Vision, pp. 536–551. Springer, Cham (2018)
33. He, K., et al.: Deep residual learning for image recognition. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway, NJ (2016)
34. Li, X., et al.: Selective kernel networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 510–519. IEEE, Piscataway, NJ (2019)

How to cite this article: Ruan, B., Zhang, C.: Occluded pedestrian detection combined with semantic features. IET Image Process. 15, 2292–2300 (2021). https://doi.org/10.1049/iper2.12196