Research Article

Double Mask R-CNN for Pedestrian Detection in a Crowd

Congqiang Liu,1 Haosen Wang,1 and Chunjian Liu2

1School of Information Engineering, Hebei University of Architecture, Zhangjiakou 075000, Hebei, China
2School of Electrical and Electronic Engineering, Anhui Science and Technology University, Bengbu 233030, Anhui, China

Correspondence should be addressed to Haosen Wang; q13255109561@126.com

Received 16 October 2021; Revised 19 January 2022; Accepted 26 February 2022; Published 22 March 2022

Academic Editor: Yugen Yi

Copyright © 2022 Congqiang Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the difficulty of feature extraction and the limitation of NMS (nonmaximum suppression) in crowded pedestrian detection, a new detection network named Double Mask R-CNN based on Mask R-CNN with FPN (Feature Pyramid Network) is proposed in this article. The algorithm has two improvements: firstly, we add a semantic segmentation branch on the FPN to strengthen the feature extraction of crowded pedestrians; secondly, we design a rule to estimate the pedestrian visibility of detected image according to the human keypoints information, and this rule can cover binary mask on the image whose pedestrian visibility is less than a certain threshold. Then we input the masked image into the network to locate occluded pedestrians. Experimental results on the CrowdHuman dataset show that the log-average miss rate (MR) of Double Mask R-CNN is 13.12% lower than the best results of other mainstream networks. Similar improvements on WiderPerson dataset are also achieved by the Double Mask R-CNN.

1. Introduction

Pedestrian detection is a classical problem in the field of computer vision, which has a wide range of applications, such as unmanned driving, robots, intelligent monitoring, human behaviour analysis, and amblyopia assistive technology. [1, 2]. Traditional pedestrian detection methods mainly used HOG (Histogram of Oriented Gradient) to extract pedestrian features and then used SVM (Support Vector Machine) for classification [3]. However, HOG can only describe pedestrian features from gradient or texture, with poor discrimination [4], and the SVM is also not suitable for the increasing scale of pedestrian detection datasets. In recent years, the accuracy of pedestrian detection has been greatly improved [5–7] with the development of a deep convolutional neural network. However, pedestrian detection in crowded scenes is still difficult.

Crowded pedestrian detection primarily involves two problems. Firstly, the similarity between pedestrians is high, while current detection models focus on extracting overall features. This makes it difficult to distinguish between highly overlapped pedestrians. Secondly, there are limitations in the postprocessing methods of prediction boxes. Traditional detection models collect samples from feature maps to generate dense prediction boxes. Nonmaximum suppression (NMS) is adopted for removing overlapped prediction boxes. However, it is very difficult to set an NMS threshold when this method is applied to crowded pedestrian scenarios. If the NMS threshold is too low, a large number of missed detections will be generated. If the NMS threshold is too high, a large number of false detections will be generated.

The above two difficulties are mainly solved from two aspects. The first aspect is to strengthen the extraction of crowded pedestrian features. Zhang et al. and Wang et al. added additional loss items in the loss function to make the proposal boxes of the same object closer to each other and to make the proposal boxes of different objects more different from each other. Thus, the feature extraction of crowded pedestrians is strengthened [8, 9]. Ge et al. first used P-RCNN to detect less crowded pedestrians, artificially constructed binary mask to cover these pedestrians, and then used S-RCNN to detect the remaining crowded targets (P-RCNN and S-RCNN both use Faster-RCNN as the basic structure); the model is forced to pay attention to the crowded target by constructing a mask [10], but constructing a mask for all detected images will greatly increase the detection time. Lin
et al. proposed a pedestrian attention mechanism that encodes fine-grained attention masks into convolutional feature maps to enhance the extraction of pedestrian features [11]. Zhou et al. proposed a discriminative feature transformation branch to strengthen the discriminability of the network between pedestrians and nonpedestrians [12]. Chi et al. predicted the extra head mask of pedestrians during the training stage to enhance the extraction of pedestrian features [13]. The second aspect is to change the postprocessing method of the prediction box. Soft NMS [14] retains the prediction box with a close distance through linear weighting of adjacent prediction boxes, but this method will generate a large number of false detections for highly overlapping objects. The adaptive NMS [15] added a branch to the detection network to predict the density of each frame and used the predicted density to replace the NMS threshold to achieve the dynamic adjustment of the NMS threshold. However, there are still difficulties in density prediction and it is still doubtful whether the density can represent the best NMS threshold setting. In addition, the prediction box is often not fully matched with the real box, which may lead to IoU (intersection over union) and prediction density inconsistence between prediction boxes. Joint-NMS [16] is to simultaneously detect the full-body bounding box and the head bounding box that are not easily occluded and then perform NMS together. This method requires that the pedestrian head label exists in the dataset. $R^2$NMS [17] also has a similar idea. Both the visible body box and the full-body box are simultaneously used for NMS together, which requires that the dataset has the visible body label.

Current crowded pedestrian detection methods tend to focus on only one difficulty (strengthening feature extraction or changing postprocessing methods). In order to further improve the performance of pedestrian detection in crowded scenarios, it is necessary to deal with two difficulties simultaneously.

In this article, Double Mask R-CNN is proposed to process two difficulties simultaneously. Double Mask R-CNN is improved based on Mask R-CNN with FPN. For the first difficulty, in order to enhance the edge feature extraction ability of crowded pedestrians, we improved FPN and added semantic segmentation branches, which is named SFPN. In order to train SFPN, we use the head bounding box and body bounding box of pedestrians to construct more elaborate pseudolabels, which is different from [11, 12] that uses all pixels of the pedestrian bounding box as pseudosemantic labels. The labels obtained by this approach are relatively coarse. For the second difficulty, in order to avoid the limitation of NMS, we adopt a method similar to PS-RCNN [10] to construct a binary mask for the detected pedestrians and then reinput the detected image into the network to obtain the occluded pedestrians. The difference is that we design an effective rule to estimate the visibility of pedestrians in the image according to human keypoints. Only the detection images whose pedestrian visibility is less than a certain threshold will be covered with a mask at the location of detected pedestrians and then reinput into the network. This method can greatly reduce the detection time and is more in line with the actual needs. The method of constructing masks for all detection images has low efficiency.

To summarize, our contributions are as follows:

1. proposed SFPN module to strengthen the edge feature extraction ability of crowded pedestrians
2. combined head bounding box and body bounding box to construct more refined pseudosemantic segmentation labels
3. designed a rule used to estimate the image pedestrian visibility according to the human keypoints, only the image with low pedestrian visibility will reinput to the network to detect occluded pedestrians, which significantly reduced detection time.

2. Related Work

2.1. Generic Object Detection. The object detection model based on the deep convolutional neural network can be divided into one-stage and two-stage models according to the existence of RPN (Region Proposal Network). The purpose of the one-stage detector [18, 19] is to accelerate the detector’s responding process so as to meet the time efficiency requirements of various practical applications. Two-stage detector [20] refines detection results by adding a two-stage classification and regression network to obtain higher accuracy. Our work is based on the two-stage detector, which is the typical two-stage detector, Faster R-CNN. The Faster R-CNN [20] firstly generates a certain number of proposal boxes through RPN (Region Proposal Network) and then refines the proposal box through a two-stage classification and regression network. FPN [21] extended the Faster R-CNN by introducing a top-down feature pyramid network to deal with the scale changes of detection objects. Mask R-CNN [22] proposed Roi align to solve the problem of misalignment between proposal boxes and corresponding features in Roi Pooling [20].

2.2. Occlusion Handling. Zhang et al. [23] employed an attention mechanism across channels to represent various occlusion patterns. Song et al. [24] operated somatic topological line localization to reduce ambiguity. Stewart et al. [25] proposed a recurrent LSTM layer for sequence generation by using a Hungarian loss function. Hu et al. [26] proposed an object relation module that handles a set of objects simultaneously through the interaction of the appearance feature and geometry. Goldman et al. [27] proposed a layer for estimating the Jaccard index as a detection quality score and a novel EM merging unit and used these quality scores to resolve detection overlap ambiguities.

3. Method

The structure of Double Mask R-CNN proposed in this article is shown in Figure 1. Double Mask R-CNN consists of the following phases:

1. Input the detected image into the SFPN module to obtain the feature map and semantic segmentation map. The latter is only used in the training stage to
improve the network’s ability to extract edge features from crowded pedestrians. The semantic branch needs to be turned off during the evaluation stage.

(2) Input the feature map into the Region Proposal Network (RPN) to obtain the proposal boxes, which plays the same role as Faster R-CNN.

(3) Input the proposal boxes into MKFRCNN (Mask and Keypoint with Fast RCNN) module to obtain the prediction boxes and the corresponding instance segmentation map and human keypoints. The detailed structure will be described later. Only the prediction box and instance segmentation map are needed in the training stage to calculate loss and update the weights of the model. Human keypoints information is only needed in the evaluation stage to calculate the visibility of pedestrians of the detected image.

(4) According to the number of detected human keypoints, we can calculate the visibility of pedestrians of the detected image: \( \alpha \). If \( \alpha \) is less than the threshold \( T \), it indicates that the pedestrian density of the detected image is high, then a binary mask is added to the first detected pedestrians according to the instance segmentation map. After covering masks, the detected images will be reinput into the detection network in order to obtain the position of occluded pedestrians. Note that the instance segmentation and human keypoints branches should be closed during the second detection to decrease detection speed.

(5) Obtain the output results by merging two detection results.

Next, the specific structure of SFPN, MKFRCNN, and the calculation process of pedestrian body visibility are described. Finally, we describe the loss function used in the training detection network.

3.1. SFPN Module. The specific meaning of SFPN is the feature pyramid network with semantic segmentation branch, which is an extension of FPN proposed in 2017. Because the FPN structure is similar to the encoder-decoder structure of U-NET, the semantic segmentation branch can be easily constructed.

The construction process of SFPN is shown in Figure 2. The number above each bar chart is the number of channels. Firstly, we select ResNet50 [28] as the backbone, which had been pretrained on ImageNet [29] dataset. Then we extract the feature map obtained by conv1 \( 7 \times 7 \), conv2, conv3, conv4, and conv5, and name them as \{C1, C2, C3, C4, C5\}, respectively. M5 is obtained by \( 1 \times 1 \) convolution of C5, and then M5 is upsampled (bilinear interpolation method) to the same resolution as C4. We obtain M4 by adding convolution of C4 to M5. M3 and M2 are obtained by the same process. Then, we obtain P5, P4, P3, and P2 by Conv \( 3 \times 3 \) of M5, M4, M3, and M2 to generate proposal boxes in the RPN stage.

The establishment of the semantic segmentation branch starts from P2, and S1 is obtained by an upsampling of P2. Then, we obtain S2 through Conv \( 3 \times 3 \) and the Relu function of S1, and S2 has the same number of channels as C1. Next, we obtain S3 through Conv \( 1 \times 1 \) of S2 plus C1. Finally, the probability distribution function is obtained through the Sigmoid function. We do not carry out \( 1 \times 1 \) convolution of C1 to expand the number of channels to 256 because this method needs more GPU memory in the process of back-propagation gradient calculation. The method in this article can reduce the consumption of GPU memory and save computing resources.

Training of SFPN requires pixel-level annotations in the training dataset. Since pixel-level annotations do not exist in the CrowdHuman [30] dataset, we directly use the pedestrian annotation box to establish pseudosegmentation annotations on the basis of [30]. The conventional literature uses all pixels in the box as a pseudosegmentation annotation.
segmentation annotation. But in this article, we combine pedestrian head box with a visible box to construct pseudosemantic segmentation annotation to obtain more accurate annotation. The above two methods are shown in Figure 3.

The construction process of our pseudosemantic segmentation annotations is as follows: assume that the height and width of the upper left of the head labeling box of a pedestrian are \((X_1, Y_1)\) and \((W_1, H_1)\), respectively. Moreover, the height and width of the upper left of the visible body box are \((X_2, Y_2)\) and \((W_2, H_2)\), respectively. The polygon consists of 8 coordinates; the pseudosemantic segmentation annotation and the horizontal and vertical coordinates are represented by \(P\) and \(Q\), respectively. The calculation process is as follows:

\[
P_1 = X_1, Q_1 = Y_1, P_2 = X_1, Q_2 = Y_1 + H_1, P_3 = X_2, Q_3 = Q_2, P_4 = P_3, Q_4 = Q_3 + H_2 - H_1, P_5 = P_4 + W_2, Q_5 = Q_4, P_6 = P_5, Q_6 = Q_3, P_7 = P_1 + H_1, Q_7 = Q_6, P_8 = P_7, Q_8 = Q_1.
\]

3.2. MKFRCNN. MKFRCNN refers to the Fast R-CNN structure with instance segmentation and human keypoints detection branches. The branch structure is improved based on literature [22]. The upsampling method in the instance segmentation branch is changed from deconvolution to bilinear interpolation, and then feature aggregation is carried out through normal convolution. This is because the segmentation labeling mode of pseudo-instance segmentation annotation is relatively fixed, and the deconvolution may cause overfitting. And the spatial structure of the object is easier to be preserved through bilinear interpolation. The pseudo-instance segmentation annotation used in the training of MKFRCNN is converted from the pseudosemantic segmentation annotation constructed in 3.1. We only
need to give different values to the pixels in each annotation box. The structure of MKFRCNN is shown in Figure 4, which has three branches in total: box, mask, and keypoint, and these three branches predict the position of pedestrians, the instance segmentation map, and keypoints of the human body, respectively.

The number in the square represents the resolution and number of channels. For example, $7 \times 7 \times 256$ means that the resolution of the feature map is $7 \times 7$, and the number of channels is 256. Moreover, the number in the rectangle represents the number of nodes in the full connection layer. The number on the arrow represents the size of the convolution kernel and the number of channels of the convolution. For example, $4 \times 3 \times 3$ represents 4 kernels of $3 \times 3$ convolution and $K$ represents the number of human keypoints, which is determined by the pretrained dataset annotation. Only box and mask branches should be opened during training, and all three branches should be opened during testing. However, Mask and Keypoint branches should be closed after the construction of binary mask to decrease detection time.

Since the CrowdHuman pedestrian dataset used for training and evaluating is not labeled with human keypoints, the detection branch of human keypoints uses the pretrained key point detection branch on the COCO keypoints dataset. The pretrained key point detection branch can accurately detect human keypoints so as to accurately calculate the visibility of pedestrians in detected images. COCO keypoints dataset marks up to 17 human keypoints for each pedestrian: nose, left/right eye, left/right shoulder, left/right elbow, left/right wrist, left/right hip, left/right knee, and left/right ankle.

Figure 5 shows the detection of human keypoints. Green numbers represent the index of keypoints in the human body, such as 0 denotes nose and 1 denotes left eye. The red number represents the activation value by the keypoints detection branch. It can be found that the score of the occluded or incorrectly positioned keypoints is usually less than or equal to 0. Therefore, when calculating the visibility of pedestrians, we determine that the keypoints whose activation value is greater than 0 are successfully detected. Otherwise, detection fails.

### 3.3. Pedestrian Visibility

Before adding binary masks to pedestrian images, we need to calculate the visibility of pedestrian bodies for the purpose of reducing detection time and improving detection efficiency. This is because not all the images are severely occluded, and adding binary masks to all detected images will significantly increase detection time. The calculation process of pedestrian body visibility $a$ is as follows:

1. Calculate the ratio of the number of detected human keypoints to the total number of keypoints $R_i$:

$$R_i = \frac{k_i}{K}$$

where $i$ denotes the index of detected pedestrian, $k_i$ denotes the number of detected keypoints of a detected pedestrian, and $K$ denotes the number of labeled human keypoints in the training dataset. The detection results will give the score of each keypoint. If the score of a keypoint is greater than 0, the keypoint is estimated to be successfully detected. Different datasets need different numbers of human keypoints. For example, COCO [32] needs 17 human keypoints and MPII [33] needs 16 human keypoints.

2. Add all the ratios $R_i$ to obtain $M$:

$$M = \sum_{i=1}^{N} R_i$$

3. Divide $M$ by the number of detected pedestrians $N$ to obtain the pedestrian visibility $a$:

$$a = \frac{M}{N}$$

### 3.4. Loss Function

As semantic segmentation branches are added to the basic model, the loss function needs to be reset. The loss function is composed of classification loss, bounding box regression loss, instance segmentation loss, and semantic segmentation loss. The classification loss, instance segmentation loss, and semantic segmentation loss all use the cross-entropy loss function [34]. Regression loss uses smooth $\ell_1$ loss function [7]. The formula is as follows:

$$\text{Loss} = L_{\text{cls}}(p_i, p_i^*) + p_i^* L_{\text{box}} + p_i^* L_{\text{mask}} + L_{\text{seg}}$$

where $N$ denotes the number of the proposal boxes, $y_i$ denotes the label of the proposal box, and $p_i$ denotes the prediction probability of pedestrians. If the proposal box is labeled positive, $p_i^*$ is set to 1; otherwise $p_i^*$ is set to 0. $M$ represents the number of pixels in the semantic segmentation map, $l_j$ represents the label of the pixel, and $s_j$ represents the Sigmoid score of pixels. And $t_j^*$ denotes the offset between the prediction box and ground truth box and $t_j$ denotes the offset of the prediction box.

### 4. Experiments

In order to verify the effectiveness of Double Mask R-CNN, we conducted experiments on CrowdHuman and WiderPerson datasets, two crowded pedestrian detection datasets.
4.1. Datasets and Evaluation Metric

4.1.1. CrowdHuman Dataset. We use the CrowdHuman dataset for pedestrian detection; the dataset is designed for crowded pedestrian detection. The CrowdHuman dataset has become the evaluating benchmark of the crowded pedestrian detection algorithm, and the training set contains 15000 images, and the validation set contains 4370 images. CrowdHuman provides three categories of bounding boxes annotations for each human instance: person visible-region bounding-box and person full-body bounding-box, head bounding-box. Detecting visible-region is more difficult since the aspect ratios are more diverse than the full-body annotations. We just evaluate our method with the visible-region annotations. The comparison between the training set of CrowdHuman and other common pedestrian detection datasets is shown in Table 1.

It can be seen from Table 1 that the CrowdHuman dataset has the largest number of pedestrians in all datasets, and it has the largest number of pedestrian density (22.64) and pairwise overlaps between two pedestrians (IoU > 0.5, 2.40). Therefore, this dataset can better reflect the pedestrian detection performance of the detection network in crowded scenarios.

4.1.2. WiderPerson Dataset. WiderPerson [35] is another crowded pedestrian detection dataset collected from multiple scenarios, in which the training set contains 8000 images and the validation set contains 1000 images, and it contains five types of annotations: pedestrians, riders, partially visible persons, crowd, and ignored regions. In our experiment, we combined the former three categories into one category during the training and evaluating stage.

4.1.3. Evaluation Metric. The evaluation criterion is adopted from the literature [36]. The standard log-average miss rate (MR) is calculated in the false positive per image (FPPI) with a range of $[10^{-2}, 10^0]$. Besides, AP$_{50}$ is also evaluated following the standard COCO evaluation metric. For the CrowdHuman dataset, the validation set is divided into seven subsets: Small, Medium, and Large subsets, according to the height, to verify the robustness of our method to detect objects of different scales. According to the different occlusion ratios, Heavy, Partial, and Bare subsets verify the robustness of our method to detect objects with different occlusion ratios. Reasonable subset follows the same standard used in [37], a general subset to evaluate. All experiments are evaluated at IoU (intersection

![Figure 4: The structure of MKFRCNN.](image-url)
over union) threshold of 0.5. The subset of CrowdHuman is shown in Table 2.

The occlusion rate is defined as follows:

$$\text{occ} = 1 - \frac{A_v}{A_f}$$

(6)

where $A_v$ denotes the area of visible body box and $A_f$ denotes the area of full-body box.

4.2. Implementation Details. This article is based on PyTorch [30] deep learning framework, and the GPU is a single RTX 2070. In order to fairly compare the performance of different detection networks, the process of input images is consistent with the literature [28]. The short edge is at 800 pixels, while the long edge should be no more than 1400 pixels. The aspect ratios of the anchor are set to {0.5, 1.0, 2.0} without any data enhancement techniques. The batch size is set as 1, and the initial learning rate is $1 \times 10^{-3}$, with a total of 150 K iterations. After 105 K and 135 K iterations, the learning rate decreases to $1 \times 10^{-4}$ and $1 \times 10^{-5}$, respectively. Momentum is set as 0.9, and weight decay is set as $5 \times 10^{-4}$, and the threshold $T$ of pedestrian visibility $a$ is set as 0.6 in evaluation stage. In order to make the model capable of detecting human keypoints, we first trained 10 epochs on the COCO keypoints dataset to obtain the pretrained model and then trained on the CrowdHuman dataset. During training, the weights of the keypoint detection branch are not updated.

4.3. Experiments on CrowdHuman

4.3.1. Overall Performance. In order to show that the detection performance of Double Mask R-CNN is better than that of other detection networks, the detection results of Mask R-CNN are compared with those of other detection networks on CrowdHuman. All detection results are obtained on Reasonable subset, and the comparison is shown in Table 3.

As shown in Table 3 (* stands for our reimplemented results), Double Mask R-CNN has the lowest MR, which is 39.07%, 16.87% lower than the baseline. Compared with the

<table>
<thead>
<tr>
<th>Table 1: Comparison of different pedestrian detection datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images</td>
</tr>
<tr>
<td>Persons</td>
</tr>
<tr>
<td>Person/image</td>
</tr>
<tr>
<td>IoU &gt;0.5</td>
</tr>
</tbody>
</table>

Figure 5: Detection of human keypoints.
4.3.2. Upsample Influence. In Section 3.2, we mentioned that the upsampling method of instance segmentation branch in Mask R-CNN module is changed from deconvolution to bilinear interpolation. Because the annotation mode of pedestrian pseudo-instance segmentation is relatively fixed, the deconvolution may cause overfitting and thus achieve worse performance. In Table 4, we give the experimental results obtained using different upsampling methods for Double Mask R-CNN. As shown in Table 4, in the instance segmentation branch, bilinear interpolation instead of deconvolution can achieve better detection performance; MR decreases by 0.62% and AP increases by 0.42%.

4.3.3. Ablation Study. In order to verify the effectiveness of each module in Double Mask R-CNN, we gradually add the SFPN, MKFRCCN, and a (the module that adds binary mask according to pedestrian visibility) module to the model. The experimental results are shown in Table 5. The ‘Baseline’ stands for our reimplemented results for comparison.

As shown in Table 5, after the addition of the SFPN module, the MR of Reasonable decreases by 12.24%, Small, Medium, Large decreases by 8.18%, 9.05%, and 11.9%. Heavy, Partial, Bare decreases by 7.68%, 10.72%, and 10.69%. The significant improvement means that adding the semantic information can strengthen the feature extraction of crowded pedestrian, and effectively improves the performance of detection network with different scale and occlusion rate of pedestrian. The addition of instance segmentation branch MKFRCCN can further decreases the MR in different subsets of CrowdHuman: the MR of Reasonable decreases by 0.76%, Small, Medium, Large decreases by 0.65%, 0.59%, 0.59%. Heavy, Partial, Bare decreases by 0.22%, 0.29%, 0.49%. Finally, after the addition of the masking module according to pedestrian visibility a, the MR of different subsets decreases significantly, the MR of Reasonable decreases by 3.75%; Small, Medium, and Large decreases by 1.04%, 1.97%, and 3.85%; Heavy, Partial, and Bare decreases by 3.43%, 4.05%, and 2.72%. The maximum decrease on the Heavy and Partial subsets indicates that the second detection method, after adding a binary mask, can effectively locate the occluded pedestrians in the crowded scene. In addition, another reason for the significant decrease of the MR on different subsets is that the detection model pretrained on the COCO keypoints dataset. This indicates that the human keypoints information can also effectively improve the network’s feature extraction ability of crowded pedestrians.

As shown in Table 6, with the addition of SFPN, MKFRCCN, and a modules, the AP on all subsets keeps rising, compared with the ‘Baseline’; Reasonable AP increases by 5.54%, Small, Medium, and Large subsets improved by 13.24%, 7.83%, and 4.73%; Heavy, Partial, and Bare subsets improved by 12.75%, 8.98%, and 6.11%, respectively. The improvement in Small and Heavy subsets most obviously indicates that semantic information and human keypoints information effectively improve the feature extraction ability of the model for small objects and occluded objects and further indicate the robustness of Double Mask R-CNN in detecting pedestrians with different scales and different occluding ratios.

4.3.4. Detection Time. Table 7 shows the MR% and per image detection time when a threshold set is of 0.6 and 1.0 (covering masks with all detected images); it can be found that when a is 1.0, MR% is 2.3%, which is higher than that when a is 0.6, and the inference time is four times than that when a is 0.6, showing that the method of covering masks with all detected images will significantly increase inference time and introduce more false positives. It is more efficient and performs better to filter high pedestrian density images based on pedestrian visibility a.

4.3.5. Qualitative Analysis. In order to visualize the advantages and disadvantages of Double Mask R-CNN to locate crowded pedestrians, some detection results are shown in Figure 6. The picture in the left column represents the detection result of the bounding box, and the picture in the right column represents the detection result of human keypoints. The green box is the results of the first detection,
Table 5: Ablation experiments using MR.

<table>
<thead>
<tr>
<th>Method</th>
<th>SFPN</th>
<th>MK</th>
<th>MR%</th>
<th>Reas</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
<th>Heavy</th>
<th>Partial</th>
<th>Bare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td></td>
<td></td>
<td></td>
<td>55.82</td>
<td>81.71</td>
<td>73.12</td>
<td>67.28</td>
<td>85.05</td>
<td>63.58</td>
<td>48.88</td>
</tr>
<tr>
<td>Double Mask R-CNN</td>
<td>√</td>
<td></td>
<td></td>
<td>43.58</td>
<td>73.53</td>
<td>64.07</td>
<td>56.09</td>
<td>77.37</td>
<td>52.86</td>
<td>38.19</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
<td>42.82</td>
<td>72.88</td>
<td>63.48</td>
<td>55.50</td>
<td>77.15</td>
<td>52.57</td>
<td>37.70</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
<td>39.07</td>
<td>71.84</td>
<td>61.51</td>
<td>51.65</td>
<td>73.72</td>
<td>48.52</td>
<td>34.98</td>
</tr>
</tbody>
</table>

Table 6: Ablation experiments using AP_{50}.

<table>
<thead>
<tr>
<th>Method</th>
<th>SFPN</th>
<th>MK</th>
<th>AP_{50}%</th>
<th>Reas</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
<th>Heavy</th>
<th>Partial</th>
<th>Bare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td></td>
<td></td>
<td></td>
<td>81.26</td>
<td>34.13</td>
<td>50.88</td>
<td>69.61</td>
<td>49.82</td>
<td>70.81</td>
<td>80.93</td>
</tr>
<tr>
<td>Double Mask R-CNN</td>
<td>√</td>
<td></td>
<td></td>
<td>85.19</td>
<td>44.54</td>
<td>56.32</td>
<td>72.63</td>
<td>57.79</td>
<td>77.63</td>
<td>86.23</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td>√</td>
<td></td>
<td>86.05</td>
<td>45.96</td>
<td>57.92</td>
<td>72.79</td>
<td>59.97</td>
<td>78.45</td>
<td>86.42</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td>√</td>
<td>86.80</td>
<td>47.37</td>
<td>58.71</td>
<td>74.34</td>
<td>62.57</td>
<td>79.79</td>
<td>87.04</td>
</tr>
</tbody>
</table>

Table 7: Average inference time per image.

<table>
<thead>
<tr>
<th>a</th>
<th>MR%</th>
<th>Time(s/image)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>39.07</td>
<td>0.4</td>
</tr>
<tr>
<td>1.0 [10]</td>
<td>41.20</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Figure 6: Continued.
and the red box is the second detection results obtained by adding a binary mask to the image. As can be seen from the six detection result graphs given in Figure 6, the proposed method has good robustness to pedestrians of different scales, especially to achieve good detection results for most small pedestrians, as shown in Figure 6(a). Moreover, our method also has good robustness for pedestrians of different occlusion ratios, as shown in Figures 6(a) and 6(b), the image filtered according to the pedestrian visibility has a high pedestrian density, and it is quite common for people to occlude each other. However, most of the occluded pedestrians can be effectively detected by adding a binary mask. The calculation of pedestrian visibility depends on the detection accuracy of human keypoints. By observing the detection results of human keypoints in each group of images, it can be found that the keypoints detection branch pretrained on COCO human keypoints dataset can detect the most visible keypoints of pedestrians.

The method in this article has good robustness for pedestrians with different scales and different occlusion ratios, but there are still some false detections, as shown in Figure 6(c), which can detect objects similar to pedestrians. For this problem, hard negative examples mining may be used to alleviate it.

4.4. Experiments on WiderPerson. In order to verify the generalization performance of our method, we also conducted experiments on the WiderPerson dataset. Since the WiderPerson dataset only has a pedestrian visible body box, we can not divide the dataset according to the occlusion ratio. Therefore, we only present experimental results on All (unrestricted height and occlusion rate), Small, Medium, and Large subsets. The other settings follow our experiments on CrowdHuman. The experimental results are shown in Tables 8 and 9. As shown in Tables 8 and 9, with the addition of SFPN, MKFRCNN, and a module, MR% of all subsets keeps decreasing and AP50% keeps increasing. Compared with Baseline, All MR% decreased by 8.81%; Small, Medium, and Large decreased by 8.13%, 15.46%, and 13.3%, respectively. AP of All increased by 5.11%; Small, Medium, and Large increased by 7.94%, 3.98%, and 2.43%, respectively. This proves the generalization of our method.

5. Conclusions

In this article, we propose the Double Mask R-CNN to improve the detection performance of pedestrians in crowded scenarios. Experimental results on CrowdHuman and WiderPerson dataset show that model combined with

<table>
<thead>
<tr>
<th>Method</th>
<th>SFPN</th>
<th>MKFRCNN</th>
<th>$\alpha$</th>
<th>All</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td></td>
<td></td>
<td></td>
<td>71.4</td>
<td>88.81</td>
<td>59.68</td>
<td>39.43</td>
</tr>
<tr>
<td>Double Mask R-CNN</td>
<td>√</td>
<td>√</td>
<td></td>
<td>64.77</td>
<td>82.37</td>
<td>47.82</td>
<td>29.90</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td>$\alpha$</td>
<td>64.54</td>
<td>81.98</td>
<td>46.95</td>
<td>30.21</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td>62.59</td>
<td>80.68</td>
<td>44.22</td>
<td>26.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>SFPN</th>
<th>MKFRCNN</th>
<th>$\alpha$</th>
<th>All</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline*</td>
<td></td>
<td></td>
<td></td>
<td>64.28</td>
<td>33.08</td>
<td>75.61</td>
<td>88.27</td>
</tr>
<tr>
<td>Double Mask R-CNN</td>
<td>√</td>
<td>√</td>
<td></td>
<td>66.98</td>
<td>37.89</td>
<td>78.00</td>
<td>89.43</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td>$\alpha$</td>
<td>67.02</td>
<td>38.76</td>
<td>78.05</td>
<td>89.41</td>
</tr>
<tr>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td>69.39</td>
<td>41.02</td>
<td>79.59</td>
<td>90.70</td>
</tr>
</tbody>
</table>
the semantic segmentation and instance segmentation (SFPN and MKFRCNN module) effectively strengthens the feature extraction of crowded pedestrians and improves the detection accuracy for different scales and occlusion ratios of pedestrian. In addition, images with high pedestrian density can be successfully filtered out according to the effective rule of calculating pedestrian body visibility in the image according to human keypoints information so that the occluded pedestrians can be detected by adding a binary mask and then reinput detected image into the detection network, which significantly reduces the MR in crowded scenarios.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Science and technology research project of Higher education in Hebei Province (B2021054), Academic Team Innovation Ability Improvement Project of Hebei University of Architecture (TD202011), Hebei Higher Education Teaching Reform Research and Practice (No. 2018GJJ328), and Graduate Innovation Fund Project of Hebei Institute of Civil Engineering and Architecture (XY202154).

References


