Multimedia Tools and Applications (2024) 83:65687–65706
https://doi.org/10.1007/s11042-023-18019-5

Crowded pedestrian detection with optimal bounding box relocation

Ren Han[1] · Meiqi Xu[1] · Songwen Pei[1,2]

Received: 17 September 2023 / Revised: 21 November 2023 / Accepted: 24 December 2023 /
Published online: 16 January 2024

Abstract
In crowded pedestrian detection, occlusion situations are common challenges that seriously impact detection performance. These occlusions are usually classified into pedestrian-to-pedestrian occlusions and object-to-pedestrian occlusions which result in false detection and missed detection. In this paper, we propose a novel model to address the crowded pedestrian detections in the cases of occlusions, which can generate an optimal bounding box containing the pedestrian instance with accurate position information. Firstly, Distance-Intersection over Union loss is introduced in Region Proposal Networks module for network training to generate proposal boxes, considering both the position and area of the region where the pedestrian is occluded. Secondly, a refinement module is added in Region Convolutional Neural Network to eliminate false positive proposal boxes, Earth Mover's Distance Loss is used to re-predict the pedestrian in these boxes. Finally, Relocation Non-Maximum Suppression is employed to select the optimal bounding box. Considering the parts of the pedestrian contained by its adjacent proposal boxes, the optimal bounding box is located in order to achieve the complete pedestrian instance. The proposed model is evaluated on CrowdHuman and CityPersons datasets respectively. On CrowdHuman dataset, the proposed model improves AP by 5.6% and JI by 5.2%, while reducing MR^{-2} by 3.8% compared to the baseline. Compared to the state-of-the-art model, the proposed model reduces 0.4% on MR^{-2}, which shows its effectiveness for pedestrian detection in crowded scenes. On CityPersons dataset, the proposed model obtains the AP with 96.8% among all the evaluated models, which indicates its generalization for pedestrian detections in various crowded scenes.

Keywords Pedestrian detection · Non-maximum suppression · Distance-Intersection over Union · Image processing

1 Introduction

Pedestrian detection is an attractive issue in computer vision as it can identify pedestrians and marks their positions in an image. Pedestrian detection is already applied in security monitoring, Autonomous Driving [1, 2], smart home, etc. The demands for the technology

Extended author information available on the last page of the article are still increasing. Especially in modern society, there exist a large number of crowded pedestrian scenes, such as bus stations, shopping malls, gatherings, etc. The applications of pedestrian detection in such scenes not only gives people more conveniences, but also ensures their safety. Consequently, it is appealing to investigate the effective pedestrian detection approaches specifically suitable for crowded scenes.

The performance of pedestrian detection in crowded scenes can be affected by some factors. For example, the pixel size of individual pedestrian, the multiple posture variation of pedestrian, the degree of occlusion, etc. Among these mentioned issues, the occlusion degree is a problem that seriously affects the detection performance, so it is necessary for us to further explore the solution for this issue. The pedestrian occlusions are mainly classified into intra-class occlusion and inter-class occlusion [3, 4]. The intra-class occlusion is defined as the mutual occlusion between individual pedestrians, which often introduces a large amount of interference information and leads to false detection. The inter-class occlusion is the occlusion of pedestrians by other objects, which often brings about information loss of the detected pedestrians and thus leads to missed detection. In crowded scenes, these two issues result in the inability to detect pedestrians efficiently and locate their positions accurately. Therefore, to address the above problem, we need to conduct related researches and propose a novel model to improve the performances of pedestrian detection.

To obtain high detection performance, the anchor based two-stage pedestrian detection model is proposed, which displays the excellent performance on COCO [5], PASCAL VOC [6], etc. Furthermore, the researchers are devoted to enhance the modules of the two-stage pedestrian detection algorithm. For instance, Gao et al. [7] propose feature fusion model to improve pedestrian detection performance by capturing high quality features. Zhou et al. [8] address pedestrian detection in crowded scenes in terms of part detection. Bodla et al. [9] and Liu et al. [10] address the problem of low detection performance by improving the Non-Maximum Suppression (NMS) algorithm. Liu [11] et al. solve the weaknesses of FPN by a novel tripartite feature-enhanced pyramidal network (TFPN), which speeds up the encoding capability and generates more robust representations. The existing works contribute some effective detection methods, but they rarely involve the positions information for overlapping parts of prediction boxes which may be helpful in eliminating the effect of occlusion and determine the complete pedestrian instance precisely.

In this paper, a novel anchor-based two-stage pedestrian detection model is employed to solve the severe occlusions in crowded scenes. Firstly, to address the false detection due to occlusion scenes, we introduce Distance-Intersection over Union (DIoU) loss [12] to train the network model so as to improve the accuracy of the generated proposal boxes. During the training process, the presence of occluded pedestrian instances in the image results in one instance can be contained in more than one proposal box. An efficient method is needed to determine which proposal box has the best match with the Ground-truth box (Gt box) containing the instance. DIoU loss algorithm takes the center point distance between the proposals box and the Gt box as the basis for calculating the loss, and directly regresses the Euclidean distance between the center point of the two boxes to accelerate the convergence. Secondly, a refinement module is added to the Region Convolutional Neural Network (RCNN). Due to the occlusions of pedestrian instances, some proposal boxes generated by the module mentioned above contain several instances. Mover's Distance (EMD) loss [13] is introduced as the metric to determine which instance is preserved in the proposal box. Finally, we utilize Relocation Non-Maximum Suppression (RNMS) [14] as the post-processing operation. Compared to other NMS algorithms, RNMS not only selects the bounding box from a series of proposal boxes, but also relocates the box, so as to achieve the optimal bounding box. Our main contributions can be summarized as follows:

- We propose DIoU-RPN module to retrain the feature extraction network. The core of the module is to use the new loss algorithm to calculate the center point distance and the overlapping area between proposal boxes to distinguish occluded pedestrians.
- We introduce a refinement module to exclude false positives from the proposal boxes. The refinement module mainly uses EMD loss to minimize the loss values generated during training, thus optimizing the detection performance.
- RNMS is introduced as a post-processing operation. For a pedestrian instance, the position information of all proposal boxes containing it parts is obtained. Then, the optimal bounding box is relocated based on such position information, so that it contains the complete instance.

With the combination of these modules, the proposed model can eliminate the effect of the occlusion and achieve the better detection performance.

This paper is organized as follows: Related work is reviewed in Section 2. The details of our pedestrian detection model are described in Section 3. While in Section 4, the experimental results of our pedestrian detection model on the relevant data sets will be shown. Finally, the conclusions are discussed in Section 5.

2 Relate work

2.1 Pedestrian detection

Some methods have been proposed to detect pedestrians in various situations, such as size variation, occlusion, etc. Proposal boxes and prior boxes are commonly used in existing algorithms.

Based on the use of proposal boxes in detection, pedestrian detection algorithms can be classified into one-stage object detection algorithms [15–19] and two-stage object detection algorithms [3, 13, 20–26]. Instead of extracting features of candidate regions, the one-stage object detection algorithm directly uses the detection network to classify and regress objects in an image. These one-stage object detection algorithms characterize low computational cost and high real-time performance, but low accuracy when detecting dense objects. Detection networks include YOLO [15, 16], RetinaNet [17], and SSD [18]. The primary difference between the two-stage object detection algorithms and the one-stage object detection algorithms is that the first layer network model is used exclusively to extract the proposal boxes, and the second layer network model classifies and regresses the proposals boxes. Compared to the one-stage object algorithms, the two-stage object detection algorithms have higher accuracy, but consume larger resources and time, which results in poor real-time performance. Detection networks of two-stage object detection algorithms include the RCNN [20, 25], SPPNet [26], etc.

According to whether a priori boxes are used for detection, detection algorithms can be divided into anchor-based object detection algorithms [3, 13, 25] and anchor-free object detection algorithms [22, 23]. In the anchor-based algorithms, a set of anchor boxes at different scales are generated and then these anchor boxes contain the pedestrians are selected as the candidates. Most of the mentioned above two-stage object detection algorithms are also anchor-based. The central region and key point are the major approaches to implement anchor-free object detection algorithms, which eliminate the anchor box generation mechanism and speed up the detection. Nonetheless, the accuracy of anchor-free methods is lower than that of anchor-based methods. The common networks used by this type of algorithm are YOLO, CenterNet [27], Fcos [28], etc.

Furthermore, partial detectors [29] and novel detection models have been specifically designed to deal with the occluded scenes. Recently, convolutional neural networks have dominated the crowded pedestrian detection and showed the excellent performances. Shang et al. [24] propose that by supervising the visibility for each part, the network is encouraged to extract features with essential part information. Chu et al. [13] propose that a proposal can predict multiple instances, thereby improving the detection performance. Wang et al. [3] mainly use the proposed dual-region feature generation model to generate high-quality proposal features. Liu [30] et al. propose a feature blender to generate stronger features by fusing initially obtained rough features.

Despite these advances, the challenges posed by environmental changes in real-world scenes continue to persist, necessitating further research. Our model differs from existing models, as it not only detects the positions of pedestrians but also refines the generated proposal box information to address the occlusion issue in crowded scenes.

2.2 IoU loss

IoU can reflect the accuracy of the prediction results in object detection tasks. It shows the detection performance mainly by calculating the similarity between two boxes. IoU and IoU loss [31] equations are as follows:

$$\begin{gathered} IoU = \frac{(A \cap B)}{(A \cup B)} \\ L_{IoU} = 1 - IoU \end{gathered} \tag{1}$$

where $A \cap B$ represents the area of the overlapping part of two boxes, $A \cup B$ represents the area of the union of two boxes. The smaller the IoU loss value is, the larger the overlap area and the closer the position of the two boxes and the better the detection performance of the model.

In the object detection model, IoU loss is utilized in the RPN module to calculate the model training loss value. The specific application is to select and adjust the proposals box based on the IoU loss calculated from the anchor box and the Gt box. However, IoU indicates the area of the overlapping area between two boxes and fails to provide information about the position of the boxes. There are various cases where two boxes overlap, but there may be overlapping areas with the same area size but different overlapping positions in these cases.

In order to solve such problems, we use DIoU. DIoU is the improvement of IoU and adjusts and determines the position of the proposal box by adding a penalty item which minimizes the normalized distance between the center point of the proposal box and the Gt box [12]. DIoU not only optimizes the convergence speed in regression, but also involves several important factors for detection in the calculation, including overlapping area and center point distance. During the training, DIoU loss makes the proposals box regression more stable.

The calculation results of IoU and DIoU for the overlapping parts of two boxes are displayed in Fig. 1. It can be observed that DIoU is more sensitive to the overlapping position of boxes, which is benefit to achieve higher accuracy in object detection.

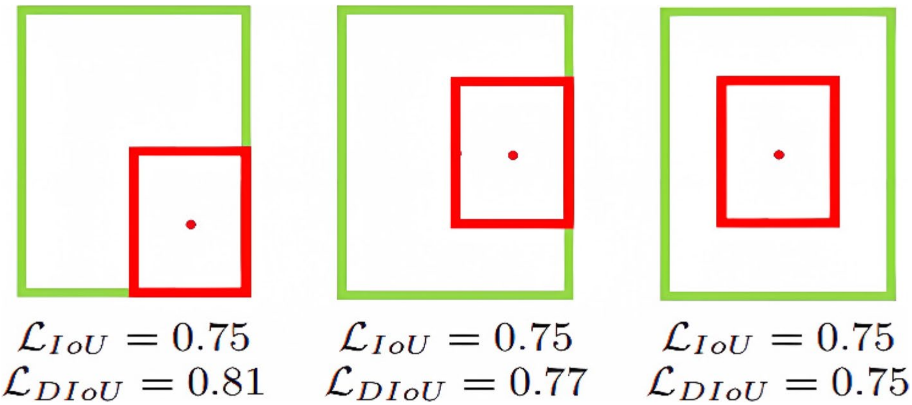

$\mathcal{L}_{IoU} = 0.75$
$\mathcal{L}_{DIoU} = 0.81$
$\mathcal{L}_{IoU} = 0.75$
$\mathcal{L}_{DIoU} = 0.77$
$\mathcal{L}_{IoU} = 0.75$
$\mathcal{L}_{DIoU} = 0.75$

Fig. 1 Comparisons between IoU and DIoU

2.3 NMS

NMS algorithm is commonly applied in post-processing operation in object detection, aiming at selecting the optimal bounding box from a set of proposal boxes.

The main steps of the traditional NMS algorithm are as follows: 1) select the proposal box with the highest category confidence as the optimal bounding box and remove it from the proposal box set; 2) calculate the IoU between the remaining proposals boxes and the currently selected optimal bounding box; 3) compare the calculated IoU values with the NMS threshold and suppress the proposal boxes larger than the threshold; 4) repeat the above operations until the proposal box set is empty.

In recent years, various NMS improvement algorithms have been proposed. Soft-NMS [9] is to reduce the detection scores instead of directly removing the highly overlapping proposal boxes. Adaptive-NMS [10] algorithm is to automatically set the confidence threshold based on the pedestrian density. Set-NMS [13] algorithm add an additional evaluation process to check whether two proposal boxes are coming from the same proposal before removing a box.

These algorithms can improve the recall rate to a certain extent, but it is insufficient to select the optimal bounding box based on the category confidence. The reason is that the proposal box generated by depth-based algorithms contains coordinate information and the category confidence. The coordinate information fails to provide any useful information about the proposal box. The category confidence indicates the probability of instance existence. The higher the category confidence, the higher the probability that pedestrian instance exists in the proposal box. The box selected based on the category confidence is the one containing the largest part of the pedestrian instance in all of the candidate boxes, but it does not necessarily contain complete pedestrian instances. Therefore, in this paper, RNMS [14] is proposed as a post-processing operation in the pedestrian detection algorithm to improve the reliability and accuracy in pedestrian detection in crowded scenes.

3 Method

The overall architecture of the proposed model is depicted in Fig. 2. The foundational network model is established based on Feature Pyramid Network (FPN) [32] and Resnet-50 [33], which is employed as the backbone for feature mapping. The feature maps generated by FPN are marked blue, the Gt boxes are marked yellow in Fig. The model presented in this paper comprises four primary processing steps. 1) The input image is pre-processed and rough features are generated by FPN. 2) DIoU-RPN module is proposed to train the network weights to generate proposal boxes. Compared with the original IoU loss, DIoU loss converges faster. Furthermore, considering not only the overlapping area between the proposal box and the Gt box but also the position of the overlapping part of the proposal boxes, DIoU-RPN effectively obtains the ideal proposal box set. which is marked purple in Fig. 3) In RM-RCNN, the refinement module is incorporated into RCNN to verify the legitimacy of the instance, thereby enhancing the accuracy and reliability of the model. 4) RNMS is introduced as a post-processing operation, which is more comprehensive than other NMS. RNMS relocates the optimal bounding box, which makes the final optimal bounding box obtained be the one containing the most information about the object instance boxes.

3.1 Distance-IoU region proposal networks

The existing works aim to enhance the quality of image features by improving RPN. RPN module is utilized to generate proposal boxes in Faster Region Convolutional Neural Network (Faster-RCNN) model [25]. The main purpose of RPN is to preliminarily adjust the anchor box, get the proposal box, and lay the foundation for the subsequent fine adjustment. There are the following steps in the RPN process: 1) A series of convolutions are applied in FPN module to obtain the common feature map, and predefined anchor boxes are performed on the common feature map to generate suggestion frames. These anchors have different sizes and shapes and the purpose is to frame objects of different sizes and shapes. 2) The rough proposal boxes are generated by performing 3×3 convolution and

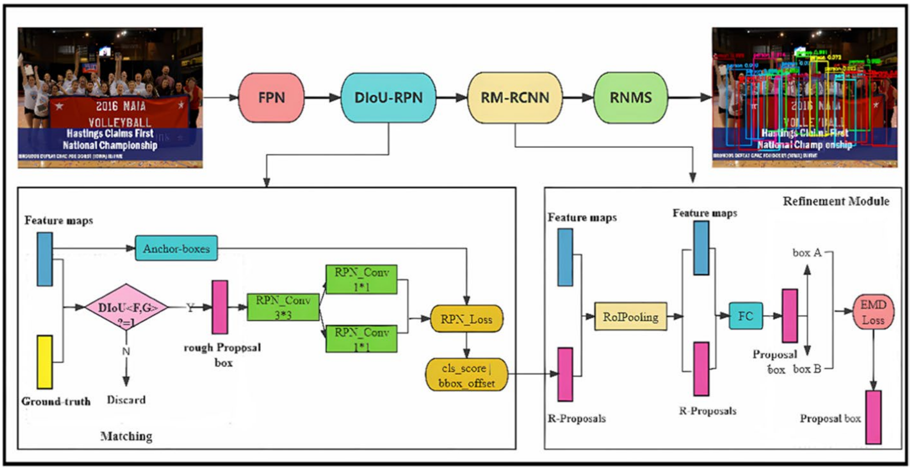


Fig. 2 The Architecture of the model

1×1 convolution on the common feature map, respectively. The 3×3 convolution is to determine whether the anchor box contains pedestrians or not, and the 1×1 convolution is to adjust the position of each anchor box. 3) Since there is a large number of rough proposal boxes, it is necessary to filter them. Firstly, according to the probability of whether it contains pedestrians or not, some proposal boxes with higher scores are filtered out, then the NMS is used to remove some boxes with more overlapping to get the final proposal boxes. 4) During the training process, IoU between the Gt box and anchor is calculated and the anchor box is selected according to IoU value. 5) The loss value between the chosen reliable box and the Gt box is computed, and the gradient descent of the network weight is determined from the loss value.

Proposal box regression is generally used in pedestrian detection to identify and locate the target object, so the results of this stage play a crucial role in our overall pedestrian detection performance. Our proposed improvement method focuses on improving the accuracy of the proposed box prediction during the training phase of the RPN module. RPN module calculates IoU loss mainly based on the anchor and the Gt box, the final proposal boxes are selected and adjusted according to the loss value. According to the discussion in SubSect. 2.2, IoU loss only reflects the overlapping area between the two boxes and fails to provide information on the relative positions of the two boxes. Besides, there exists a case where the overlapping area between two boxes is the same, but the overlapping positions are different. Therefore, it is doubtful to select a suitable proposed box by only relying on the overlapping area value. Based on the above, we employ DIoU loss to achieve higher precision detection performance as follows:

$$\begin{gathered} DIoU = IoU - \frac{\rho^2\left(b,b^{Gt}\right)}{d^2} = IoU - \frac{c^2}{d^2} \\ L_{DIoU} = 1 - DIoU \end{gathered} \tag{2}$$

where b represents the central position of the proposal box, b^{gt} represents the center of the Gt box, ρ is the Euclidean distance between the two center points which is also denoted as c, and d represents the diagonal length of minimum outer rectangle for the two boxes. Figure 3 provides an example that details these parameters for calculating DIoU. The proposal box and its center point are marked blue, the Gt box and its center point are marked green, and the minimum outer rectangle of these two boxes are marked red.

As shown in Fig. 1 and Fig. 3, compared to IoU, DIoU not only focuses on the overlapping area between multiple boxes, but also considers the distance between the center point of two boxes. IoU algorithm is likely to regard the occluded part of a masked instance and the unmasked part as two instances, which will increase the number of false positive samples. However, introducing DIoU method, based on the center distance between the

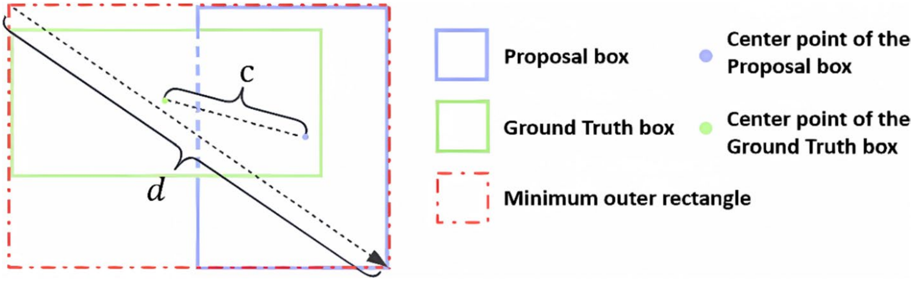

c
d
Proposal box
Center point of the Proposal box
Ground Truth box
Center point of the Ground Truth box
Minimum outer rectangle

Fig. 3 The parameters for calculating DIoU

overlapping boxes and calculating the loss, it is subsequently possible to better predict the presence of multiple instances and thus improve the performance of pedestrian detection. If the two boxes overlap perfectly, it means$c = 0, \mathrm{IoU} = 1, DIoU = 1$. Conversely, if There is no overlapping of the two boxes, $\frac{c^2}{d^2}$ tends to 1,IoU = 0, and,$DIoU = -1$. Therefore, the range of DIoU is$[-1, 1]$.

Ultimately, DIoU-RPN generates proposal boxes that are better suited, with lower category confidence loss and position loss relative to the original model.

3.2 Refinement region-CNN

Faster-RCNN model is comprised of two key components: RPN and Fast Region Convolutional Neural Network [34] detection module. RPN has been elaborated in SubSection 3.1. In the RCNN module of pedestrian detection, the tasks of classification and localization are performed.

RCNN is a simple and scalable object detection algorithm and has the following characteristics: 1) The regions of interests with varying sizes from RPN and FPN are mapped into candidate boxes with fixed size $w * h$ by using the pooling method. 2) RCNN assumes that there are multiple instances in each proposal boxes and records the class confidence and corresponding location information for the pedestrian instance. 3) The loss of category confidence and location information with respect to the Gt box is calculated for multiple pairs in the proposal box.

We assume that there are two instances in each proposal box, but in fact, there is only one instance in some proposals. So, these proposals need to be verified and confirmed. Therefore, our model deals with the problem by adding a refinement module. In this module, the prediction result of the RCNN is taken as input and combined with the proposal features to perform a second round of prediction in order to correct possible mispredictions.

In the RCNN module, there exists the matching problem between GT box and proposal box. As shown in Fig. 4, Ground-truth boxes Gt_0, are displayed in red, proposal boxes P_0, P_1 are displayed in green. Both P_0 and P_1 have intersecting regions with Gt_0, respectively. We introduce EMD Loss to match the optimal proposal box for a Gt box. EMD loss is a measure of the distance in one of the two multidimensional matrices in the feature space, which is utilized to minimize the loss incurred during multiple training runs. EMD loss can be calculated as follows:

$$\mathcal{L}_{loss} = \min \sum_{k=1}^{k} \left[\mathcal{L}_{cls}\left(c_i^{(k)}, g_{\pi k}\right) + \mathcal{L}_{reg}\left(l_i^{(k)}, g_{\pi k}\right) \right] \tag{3}$$

where π represents a certain permutation of (1, 2,..., K), whose π_k-th item isπ_k; $g_{\pi_k} \in G(b_i)$ is the π_k-th Gt box; $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ are classification loss and box regression loss respectively.

3.3 Relocation non-maximum suppression

In object detection, NMS algorithm is frequently employed as a post-processing operation. The performance of NMS algorithm works in object detection is not only related to the algorithm itself, but also often closely related to the threshold value it sets, especially in crowded scenes. If the NMS threshold is set small, the algorithm fails to distinguish all of the pedestrian. If the NMS threshold is set too large, the model detects the other objects as pedestrians, which leads to an increase number of false positive samples. Therefore, not

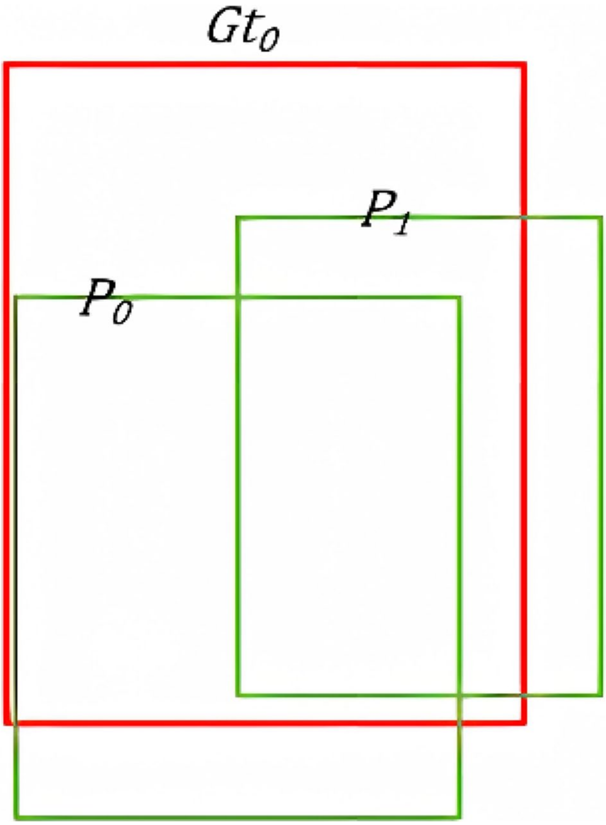


Fig. 4 Matching problem between Ground-truth box and prediction bounding box

only the adaptability of the algorithm should be considered in the selection of post-processing operation, but also its threshold value should be trained. In this paper, RNMS is proposed as a post-processing operation in the pedestrian detection algorithm in order to improve the reliability and accuracy in pedestrian detection in crowded scenes.

RNMS not only considers the proposal box with high category confidence score as the optimal bounding box as well as relocates the location of the optimal bounding box using the position relationship between the optimal bounding box and the surrounding proposal boxes [14]. Furthermore, RNMS employs the distance length instead of the IoU to measure the positional relationship between proposal boxes. The localization accuracy of the optimal bounding box is improved by RNMS.

RNMS methodology comprises two primary components: determining the optimal bounding box among the proposal boxes and relocation of the optimal bounding boxes. 1) Select the proposal box with the highest category confidence score as the bounding box *bi*, and subsequently calculate the Proximity (P) between the bounding box *bi* and other proposal boxes. Compare P with the proximity threshold. Proposal boxes above the proximity threshold are added to the set of localization references and deleted from the set of proposal boxes. Then the offset O between the bounding box *bi* and the proposed box in the set of localization references is computed. 2) Relocate bounding box *bi* using the offset O to get a higher quality optimal candidate box. 3) Repeat the above steps for the proposal boxes smaller than the proximity threshold until the proposal boxes set is empty.

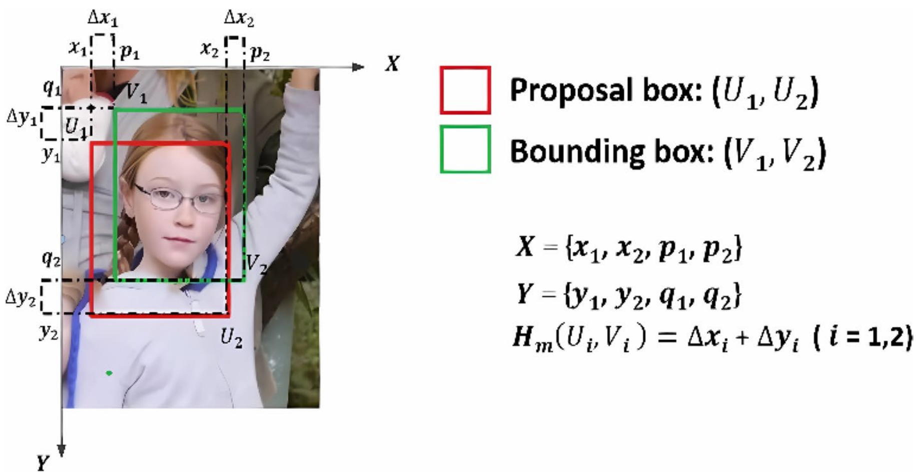

Δx_1
Δx_2
x_1
p_1
x_2
p_2
X
q_1
V_1
Δy_1
U_1
y_1
q_2
V_2
Δy_2
y_2
U_2
Y
Proposal box: (U_1, U_2)
Bounding box: (V_1, V_2)
$X = \{x_1, x_2, p_1, p_2\}$
$Y = \{y_1, y_2, q_1, q_2\}$
$H_m(U_i, V_i) = \Delta x_i + \Delta y_i \ (i = 1,2)$

Fig. 5 Manhattan distance

Algorithm 1 RNMS.

```
Require: threshold, boxes
Ensure: index, new_boxes
new_boxes ←[ ], index ←[ ]
order← Sort the boxes[:, 4] in descending order
while length (order) > 1
    i ← order[0]
    index.append(i)
    P ← proximity(boxes[i], boxes[order[1:]])
    inds ← np.where(P < threshold)[0]
    indices ← np.where(P >= threshold)[0]
    order ← order[indices + 1]
    reorder ← order[inds + 1]
    O ← sum(abs(boxes[i,0:4]-boxes[reorder[:],0:4])) / boxes.shape[0]
    M ← boxes[i].reshape(-1,7)
    M[:,0:4] ← M[:,0:4]+O[:]
    box_new.append(M)
end while
Return index, new_boxes
```

There is a new variable introduced in RNMS, P. P can be expressed by Manhattan distance between bounding box *bi* and the proposal box. The computation of P in the RNMS involves coordinate transformation and computation. Figure 5 illustrates the parameters used to calculate Manhattan distance. proposal box and bounding box are marked red and green, respectively. X and Y represent the set of horizontal and vertical coordinates of the two boxes respectively. Hm represent the Manhattan distance.

P can effectively represent the distance relationship between boxes when the size of box is similar. In the post-processing operation, there will be a large number of proposal boxes with different sizes. When the sizes of two boxes are obviously different, the P can not accurately measure the degree of their overlap [14]. In order to solve this problem, we introduced the method of normalizing the proposal box coordinate. This method makes the coordinates range between 0 and 1 and maintains their original positional relationship between the boxes.

$$\begin{aligned} norm\left(x_i, y_i\right) &= \left(x_i\prime, y_i\prime\right) \\ &= \left(\frac{x_i - min(X)}{max(X) - min(X)}, \frac{y_i - min(Y)}{max(Y) - min(Y)}\right) \end{aligned} \tag{4}$$

X and Y are the set of horizontal and vertical coordinates shown in Fig. 5.max(·) and min(·) represent the maximum and minimum values in set ·, respectively. The P of two boxes is calculated using coordinate normalization, the formula is as follows:

$$\begin{aligned} P = H_m\left(U_1, V_1\right) + H_m(U_2, V_2) = \left|y_1\prime - q_1\prime\right| + \left|y_2\prime - q_2\prime\right| + \\ \left|x_1\prime - p_1\prime\right| + \left|x_2\prime - p_2\prime\right| \end{aligned} \tag{5}$$

In order to implement the relocation operation of the bounding boxes, the offset O is utilized. The offset O is obtained by calculating the distance between the proposal boxes whose P are larger the proximity threshold and the bounding box *bi*, the formula is as follows:

$$O = \frac{\sum_{i=1}^{n} \left|B_i - M\right|}{n} \tag{6}$$

$$M_R = M + O \tag{7}$$

where: B_i is the proposal box less than the threshold, M is the bounding box, and O represents the offset between the bounding box and all proposal boxes. Finally, the optimal bounding box M_R is obtained by adding the offset O to the optimal bounding box M.

Finally, the execution steps of RNMS are shown in Algorithm 1:

4 Experiments and discussions

In this section, we perform experiments on CrowdHuman dataset [35] and CityPersons dataset [36] to evaluate the proposed model. We introduce the two datasets, assessment metrics and experimental setup in the experiments. Then we report the experimental results and discuss the performances of the proposed model.

4.1 Datasets

The datasets employed in this paper are CrowdHuman and CityPersons which are commonly used to evaluate the performances of pedestrian detection algorithms. It is essential to solve the complex occlusion problem to improve the pedestrian detection accuracy. If annotated example can reflect these aspects to a significant extent, it is anticipated that the pedestrian detection performance will witness substantial improvement. The CrowdHuman dataset provides three annotation labels for each pedestrian: Head Bounding-Box, Visible Bounding-Box, and Full Bounding-Box. A detailed picture can be seen [35].

We further investigate the robustness of the proposed model on CityPersons dataset. CityPersons, a subset of cityscape, is a lightly occluded pedestrian dataset with varying levels of occlusion. The dataset contains annotations for the region bounding boxes and full-body bounding box of pedestrians. There are 2,975 images for training, 500 images for validation and 1575 images for test.

Table 1 displays the crowding levels and the average number of pedestrians in each image. The value of the overlaps indicates IoU value greater than 0.5 between two pedestrian instances in the image. The average overlaps on CrowdHuman dataset are 2.4, and 0.32 on CityPersons dataset. We can thoroughly evaluate the robustness of the proposed model across multiple scenes with diverse crowded levels.

4.2 Evaluation metric

This paper mainly uses the following three indicators to evaluate the performance of the model:

AP: Average Precision is a measure that is jointly determined by Recall and Precision. With the value of log-average Miss Rate (MR^{-2}) as the threshold, the maximum Precision value is established for each MR^{-2} value, and the average value of all the Precision is the AP value. In object detection algorithms, AP serves as a reliable indicator of the model's Precision and Recall. The Eq. 9 to Eq. 11 used to determine AP incorporates two critical parameters. The accuracy rate expresses the ratio of correctly identified targets in the detection result to all targets detected by the detector. The recall rate denotes the ratio of correctly identified targets detected by the detector to the total number of real-world targets. The higher numerical value of AP signifies a superior performance of the detector.

$$Precision = \frac{N(Positive\ samples\ by\ detector)}{N(All\ samples\ by\ detector)} \tag{8}$$

$$Recall = \frac{N(Positive\ samples\ by\ detector)}{N(Positive\ samples\ in\ the\ label)} \tag{9}$$

$$AP = \frac{\sum_1^N \mathrm{Precision}}{N} \tag{10}$$

MR^{-2}: log-average Miss Rate [38]. MR^{-2} refers to the miss rate of false positives per image and is commonly used as an evaluation of the performance of object detection algorithm. It mainly calculates the false positive samples in the proposal box, and the lower value indicates the better detection performance of the model.

$$\mathrm{MR}^{-2} = \frac{N(\text{False positive})}{N(\text{True positive}) + N(\text{False positive})} \tag{11}$$

Table 1 Instance density of CrowdHuman and CityPersons datasets. The threshold for overlap statistics is IoU > 0:5 [13]

Dataset	# objects/ img	# overlaps/ img
CrowdHuman	22.64	2.40
CityPersons	6.47	0.32

JI: Jaccard Index [18]. JI is mainly evaluated the degree of overlap between the predicted set and the Ground-truth label set. The larger the value of JI, the closer the predicted result is to the Ground-truth.

$$JI = \frac{|DT \cap GT|}{|DT \cup GT|} \tag{12}$$

4.3 Implementation details

The backbone network we use is a Resnet-50 model pretrained on ImageNet dataset [39], using Faster RCNN with FPN as the baseline model, and the initial RoI Pooling [25] is replaced with RoI Align [40]. On CrowdHuman dataset, an aspect ratio of H:W = {1:1; 2:1; 3:1} anchor point scale is employed, while on CityPersons dataset, an aspect ratio of H:W = {0.5:1; 1:1; 2:1} anchor point scale is employed. Since the images on CrowdHuman dataset have a wide variety of sizes, these images need to be preprocessed to a unified size. In contrast, all images on CityPersons dataset are the same size, so this step can be omitted. We trained CrowdHuman dataset for a total of 30 Epochs, where the learning rate is set to 10% of the original at the 24th Epoch to the 27th Epoch, as well as the learning rate is set to 100% of the original at the 28th Epoch to the 30th Epoch. On CityPersons dataset, we train the proposed model for 25 epochs, where the learning rate is set to 10% of the original at the18th Epoch to the 21th Epoch, and at the 22th Epoch to the 25th Epoch, the learning rate is set to 100% of the original. For each proposal, we assume that there are two instances.

4.4 Detection results on CrowdHuman dataset

Ablation experiments To comprehensively evaluation the performance of the methods expounded in Section 3, a substantial number of experiments are carried out on CrowdHuman dataset. The effectiveness of the model is evaluated by the three evaluation indices mentioned in SubSection 4.2, with AP as the primary evaluation metric. Table 2 presents the results of the comparison between the methods mentioned in Section 3 and the baseline model. The Faster RCNN is used as the baseline model, IoU method is employed for the loss calculation in RPN, and the post-processing operation utilizes the NMS algorithm with a threshold of 0.5. To analyze the contribution of the proposed module separately, the components in the baseline model are gradually replaced with our module. The results of the experiments clearly demonstrate that the proposed module significantly enhances the detection performance. In particular, compared to the baseline, our model has increased 5.6% in AP metric and 5.2% in JI. More importantly, the ratio of MR^{-2} is reduced by 3.8%, providing evidence that the model does not generate false predictions. Although the

Table 2 The results of ablation experiments on CrowdHuman dataset

DIoU	RM	RNMS	AP/%	MR^{-2}/%	JI/%
Baseline (Faster RCNN)			86.2	46.6	78.5
√			89.3	43.1	80.4
√	√		89.5	42.1	80.8
√	√	√	91.8	41.4	82.3

refinement module has a little effect on AP and JI, its introduction results in a 1% reduction in MR^{-2}, demonstrating that the module mainly reduces false positives. Figure 6 shows the detection results of both our model and the baseline model on CrowdHuman dataset. For comparison purposes, detection results of the baseline model and our model are presented on the left and right, respectively. The number of Ground-truth boxes (GT) and the number of predicted boxes generated by the model (DT) are given under each result. Each

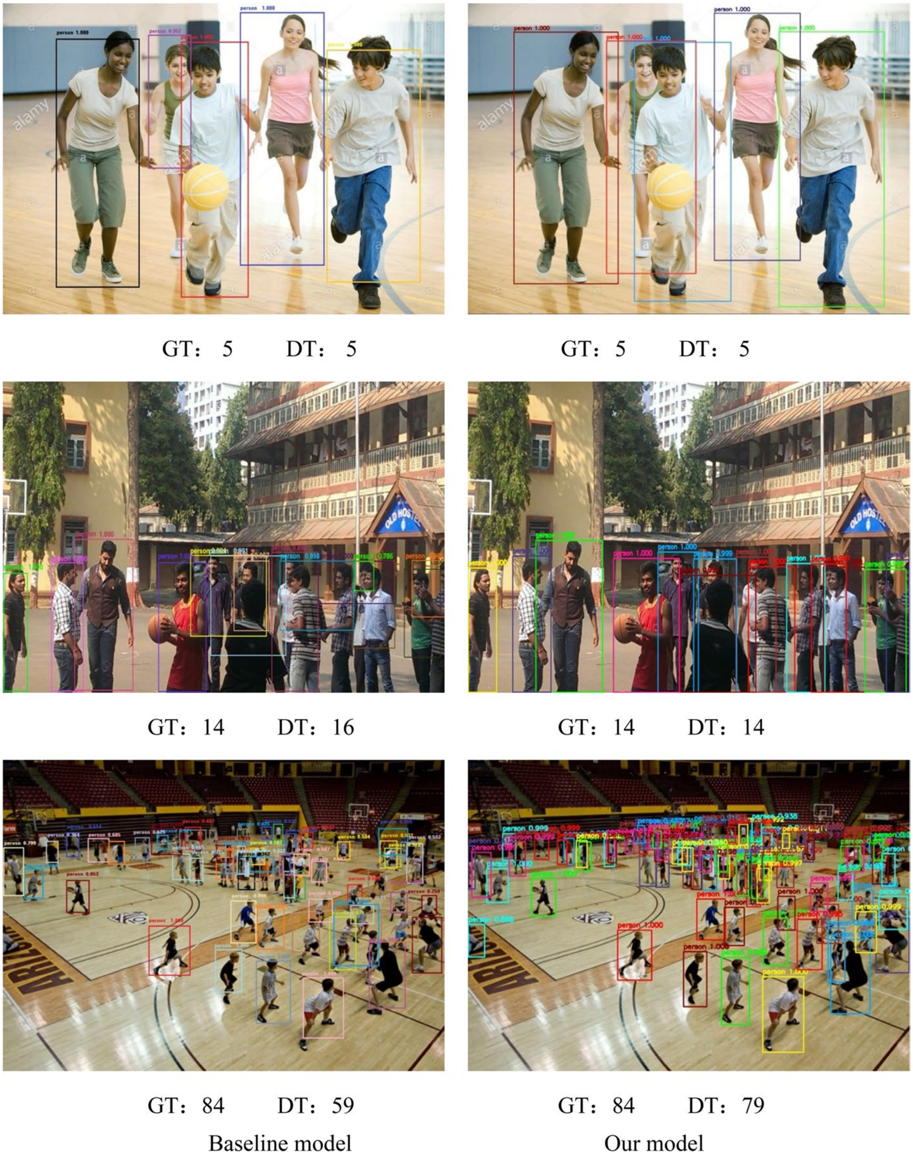
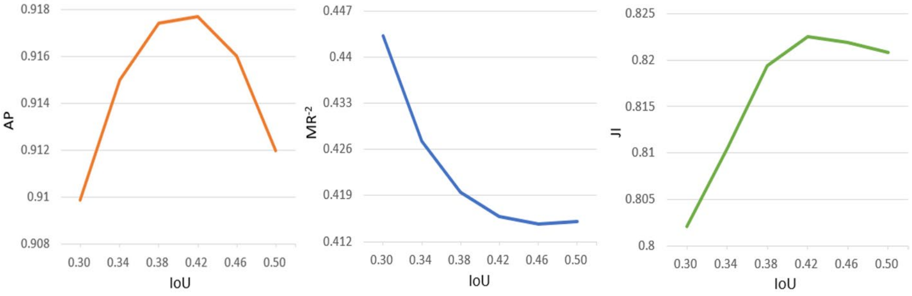


Fig. 6 Visualization of detection results. The detection results of the baseline model are on the left and the detection results of our model are on the right. The GT represents the number of Ground-truth boxes and the DT represents the number of prediction boxes

predicted box is labeled with the confidence value of the instance it contains and identified with a different color so as to be distinguished in a crowded scene. The result comparison in light occlusion scene is in the first row. Both methods detect all instances. But predicted boxes generated by our model can contain full pedestrian instances. The second row represents the result comparison in dense occlusion scene. Our model still detects all instances. There exist false detections in baseline model. The third row represents the result comparison in high crowding and heavy occlusion scene. Our model detects 79 out of 84 instances, while the baseline model detects only 59 instances. It can be clearly seen that our model is effective in detecting pedestrian instances under various crowding and occlusion levels.

Introduction of DIoU loss**.** Occlusion is the most challenging of pedestrian detection. Occlusion scenes are either pedestrian occluding each other or pedestrians being obscured by objects in the environment, which increases the number of false positive samples or loses information about pedestrians. To address the problem of low performance caused by occlusion in pedestrian detection, we introduce DIoU loss. Specifically, DIoU predicts whether the target is a different instance by the overlapping area and center distance between multiple proposal boxes, which in turn suppresses the false positive samples to solve the occlusion problem. In Table 2, our baseline is FPN with ResNet-50, DIoU is the loss calculation used for training in RPN. We are able to find that the AP value increases by 3.1% after DIoU is adopted compared to the baseline. This proves that our DIoU method can improve the accuracy of detection.

Impact of different hyperparameter Settings in RNMS In pedestrian detection algorithm, the setting of NMS threshold plays an important role in the performance of NMS algorithm. If the threshold setting is small, the algorithm cannot distinguish all of the pedestrians. If the threshold setting is large, the model considers other objects as pedestrians and increases the false positive samples. In order to analyze the optimal NMS threshold, it is necessary to conduct relevant experiments for validation. According to the existing work [14], the threshold of RNMS has a better performance in [0.3,0.5], so we also take the values in this interval. Figure 7 shows the changes of AP, MR^{-2} and JI values in the interval range of [0.3,0.5]. Combining the data of the three metrics, we found that a threshold value of 0.4 is the most comprehensive detection performance that meets our expectation.



(a) (b) (c)

Fig. 7 Setting threshold parameters, (**a**) is the AP with the threshold between 0.3 and 0.5. (**b**) is the MR^{-2} with the threshold between 0.3 and 0.5. (**c**) is the JI with the threshold between 0.3 and 0.5

Table 3 The various NMS algorithms are compared on CrowdHuman dataset. The baseline model is Faster RCNN

Method	IoU	AP/%	MR^{-2}/%	JI/%
baseline (NMS) [25]	0.5	89.5	42.1	80.7
Adaptive-NMS [10]	0.5	84.7	49.7	-
Soft-NMS [9]	0.5	90.1	42.1	80.9
Set-NMS [13]	0.5	91.0	41.8	82.1
Ours (RNMS)	0.4	**91.8**	**41.4**	**82.3**

Table 4 The results of different models on CrowdHuman dataset

Method	AP/%	MR^{-2}/%	JI/%
Baseline	86.2	46.6	78.5
Soft-NMS [9]	88.1	42.9	79.8
V2F-Net [24]	91.0	42.3	-
OAF-Net [22]	89.8	45.0	-
R2NMS [37]	89.2	43.3	-
OPLA + H-NMS [41]	90.15	49.41	-
Dual-Region Feature Extraction Networks [3]	**92.2**	41.8	**83.3**
Ours	91.8	**41.4**	82.3

Comparison with various NMS algorithms NMS algorithms are frequently treated as post-processing operation in object detection. In order to solve the problem of low object detection accuracy, there are some different NMS algorithms proposed by researchers. It is significant to select a suitable NMS algorithm in order to improve the pedestrian detection performance, and we propose RNMS as a post-processing operation to improve the detection accuracy. The conventional NMS algorithms filter the proposal boxes based on the category confidence, while RNMS determines the optimal bounding boxes based on the category confidence and the location information of the proposal boxes. In Table 3, RNMS is compared with NMS, Soft-NMS, Adaption-NMS and Set-NMS, the IoU value of each algorithm is set to the best performing value. Apparently, it can be found that RNMS shows the best performance in all of AP, MR^{-2} and JI, which demonstrates the ability of RNMS as a post-processing operation to improve the accuracy of the detection while reducing the introduction of false positive samples.

Comparison with existing work For a comprehensive evaluation of our model, we choose three types of detection models for comparisons, which are listed in Table 4. The first type is the baseline model, such as Faster RCNN, Soft-NMS, which are widely employed in detection performance evaluations. The second type is the detection model proposed in the last three years, such as R2NMS (2020), V2F-Net (2021), OAF-Net (2022), OPLA (2023). The third type is the state-of-the-art model, such as Dual-Region Feature Extraction. By comparing the performances with these models, the advantage in detection accuracy of our model can be effectively verified. It can be seen from Table 4, among all the models compared, our model shows the best performance in MR^{-2} with 41.4%. In AP and JI, our model is only slightly inferior to Dual-Region Feature Extraction Networks, but superior to any of other models. These results confirm that our model plays a positive role in improving pedestrian detection in crowded scenes.

Table 5 The results of different models on CityPersons dataset

Method	IoU*	AP/%	Δ*AP*
Baseline	0.5	95.2	-
Soft-NMS [9]	0.5	95.3	+0.1
CrowdDet [13]	0.5	94.7	-0.5
Repulsion Loss [42]	0.5	96.1	+0.9
V2F-Net [24]	0.5	96.2	+1.0
Dual-Region Feature Extraction Networks [3]	0.5	96.4	+1.2
Ours	0.4	**96.8**	+1.6

4.5 Detection results on CityPerson dataset

In order to further evaluate the performance of our model, we perform experiments on the CityPersons dataset as well. CityPersons is a dataset containing moderately crowded scenes.

Comparison with existing methods on CityPersons To further evaluate the performance of our model, we perform experiments on the CityPersons dataset as well. CityPersons is a dataset containing moderately crowded scenes. In Table 5, our model is compared with three types of models. They are baseline models such as Faster RCNN, Soft-NMS, recent models V2F-Net, CrowdDet, Repulsion Loss, and the state-of-the-art model Dual-Region Feature Extraction Network. As displayed in Table 5, our model performs the best AP with 96.8%, which 1.6% higher than the baseline model and 0.4% higher than Dual-Region Feature Extraction Networks. It can be demonstrated that our model is robust for various crowded scenes in pedestrian detection.

5 Conclusion

In crowded scenes, the occlusion degree is an important factor affecting the pedestrian detection performance. To improve the detection accuracy, we propose a novel model to relocate the optimal bounding box according to the location information of proposal boxes, which includes DIoU-RPN module, refinement module and RNMS. DIoU-RPN module and refinement module solve the false detection problem and improve the detection accuracy. RNMS solves the missed detection problem and relocates the optimal bounding box so that contains the complete instances. Our model is evaluated on two datasets with different crowded levels and shows great improvements in AP, MR^{-2} and JI compared to the existing models. However, our model can still be further improved. Our model is a two-stage detection model, which is characterized by the advantage in detection accuracy. Our model does not have a significant advantage in terms of speed of detection. In low-light environments, it is difficult to achieve high-quality pedestrian features, which results in a decrease of detection accuracy in our model. These issues will be considered in our future work.

Acknowledgements This research is supported by Engineering Re-search Center of Software/Hardware Co-design Technology and Application, Ministry of Education, East China Normal University under Grant No. OP202202.

Data availability Data will be made available on request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Cui YM, Cao ZW, Xie YX, Jiang XY, Tao F, Chen YJV, Li L, Liu DF (2022) DG-labeler and DGL-MOTS dataset: Boost the autonomous driving perception. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 58–67
2. Liu DF, Cui YM, Chen YJ, Zhang JY, Fan B (2020) Video object detection for autonomous driving: motion-aid feature calibration. Neurocomputing 409(7):1–11
3. Wang J, Zhao C, Huo Z, Qiao Y, Sima H (2022) High quality proposal feature generation for crowded pedestrian detection. Pattern Recognit 128:108605
4. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: Proceedings of the European conference on computer vision (ECCV), pp 637–653. https://doi.org/10.1007/978-3-030-01219-9_39
5. Lin TY, Maire M, Belongie S et al (2014) Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, pp 740–755
6. Everingham M, Eslami SMA, Gool LV, Williams CKI, Winn J, Zisserman A (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vis 111(1):98e136
7. Gao X, Xiong Y, Zhang G, Deng H, Kou K (2022) Exploiting key points supervision and grouped feature fusion for multiview pedestrian detection. Pattern Recognit 131:108866
8. Zhou C, Yuan J (2019) Multi-label learning of part detectors for occluded pedestrian detection. Pattern Recognit 86:99–111
9. Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-NMS--improving object detection with one line of code. Proc IEEE Int Conf Comput Vis 2017:5561–5569
10. Liu S, Huang D, Wang Y (2019) Adaptive NMS: Refining pedestrian detection in a crowd. Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2019:6459–6468
11. Liu DF, Liang JM, Geng TY, Loui A, Zhou TF (2023) Tripartite feature enhanced pyramid network for dense prediction. IEEE Trans Image Process 32:2678–2692. https://doi.org/10.1109/TIP.2023.3272826
12. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-IoU loss: Faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 34(07):12993–13000
13. Chu X, Zheng A, Zhang X, Sun J (2020) Detection in crowded scenes: One proposal, multiple predictions. Proc IEEE/CVF Conf Comput Vis Pattern Recognit 2020:12214–12223. https://doi.org/10.1109/CVPR42600.2020.01223
14 Su S, Chen R, Zhu R, Jiang B (2022) Relocation non-maximum suppression algorithm. Optics Precis Eng 30(13):1620–1630
15. Redmon J, Farhadi A (2018) YOLOv3: An incremental improvement. CoRR. https://arxiv.org/abs/1804.02767
16. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934
17. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. Proc IEEE Int Conf Comput Vis 2017:2980–2988. https://doi.org/10.1109/ICCV.2017.324
18. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, vol 2016. Springer International Publishing, pp 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

19. Liu DF, Cui YM, Tan WB, Chen YJ (2021) Sg-net: Spatial granularity network for one-stage video instance segmentation[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9816–9825. https://doi.org/10.1109/CVPR46437.2021.00969
20. Cai Z, Vasconcelos N (2019) Cascade R-CNN: High quality object detection and instance segmentation[J]. IEEE Trans Pattern Anal Mach Intell 43(5):1483–1498. https://doi.org/10.1109/TPAMI.2019.2956516
21. Jin Y, Zhang Y, Cen Y, Li Y, Voronin V (2021) Pedestrian detection with super-resolution reconstruction for low-quality image. Pattern Recognit 115:107846. https://doi.org/10.1016/j.patcog.2021.107846
22. Li Q, Su Y, Gao Y (2022) OAF-Net: An occlusion-aware anchor-free network for pedestrian detection in a crowd[J]. IEEE Trans Intell Transp Syst 23(11):21291–21300. https://doi.org/10.1109/TITS.2022.3171250
23. Wang Y, Han C, Yao G, Zhou W (2021) Mapd: an improved multi-attribute pedestrian detection in a crowd. Neurocom-puting 432:101–110
24. Shang M, Xiang D, Wang Z, Zhou E (2021) V2f-net: Explicit decomposition of occluded pedestrian detection[J]. arXiv preprint arXiv:2104.03106. http://arxiv.org/abs/2104.03106
25. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:91–99
26. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
27. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, 2019:6569–6578. https://doi.org/10.1109/ICCV.2019.00667
28. Tian Z, Shen C, Chen H, He T (2019) FCOS: Fully convolutional one-stage object detection[C]. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636. https://doi.org/10.1109/ICCV.2019.00972
29. Lin C, Lu J, Wang G, Jie Z (2018) Graininess-aware deep feature learning for pedestrian detection. In: Proceedings of the European conference on computer vision (ECCV), pp 732–747. https://doi.org/10.1109/TIP.2020.2966371
30. Cui YM, LQ Y, Cao ZW, Liu DF (2021) Tf-blender: Temporal feature blender for video object detection[C]. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 8138–8147. https://doi.org/10.1109/ICCV48922.2021.00803
31. Yu J, Jiang Y, Wang Z, Cao Z, Huang T (2016) Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia, pp 516–520. https://doi.org/10.1145/2964284.2967274
32. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125. https://doi.org/10.1109/CVPR.2017.106
33. 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. https://doi.org/10.3390/app12188972
34. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448. https://doi.org/10.1109/ICCV.2015.169
35. Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, Sun J (2018) Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123. https://doi.org/10.48550/arXiv.1805.00123
36. Zhang S, Benenson R, Schiele B (2017) Citypersons: A diverse dataset for pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3221. https://doi.org/10.1109/CVPR.2017.474
37. 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 10750–10759. https://doi.org/10.1109/CVPR42600.2020.01076
38. 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
39. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
40. He K, Gkioxari G, Dollr P, Girshick R (2017) Mask R-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969. https://doi.org/10.1109/ICCV.2017.322

41. He HY, Li ZS, Tian GZ, Chen HX, Xie L, Lu S, Su HY (2023) Towards accurate dense pedestrian detection via occlusion-prediction aware label assignment and hierarchical-NMS. Pattern Recogn Lett 174:78–84
42. Wang XL, Xiao TT, Jiang YN, Shao S, Sun J, Shen CH (2018) Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7774–7783. https://doi.org/10.1109/CVPR.2018.00811

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

Authors and Affiliations

Ren Han[1] · Meiqi Xu[1] · Songwen Pei[1,2]

✉ Ren Han
ren.han@usst.edu.cn

Meiqi Xu
yoke_xu1997@126.com

Songwen Pei
swpei@usst.edu.cn

[1] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China

[2] Engineering Research Center of Software/Hardware Co-Design Technology and Application, Ministry of Education (East China Normal University), Shanghai, China