




Context feature fusion and enhanced non-maximum suppression for pedestrian detection in crowded scenes

Yu Shao¹ · Jianhua Hu² · Lihua Hu¹  · Jifu Zhang¹ · Xinbo Wang²

Received: 30 October 2023 / Revised: 27 February 2024 / Accepted: 29 February 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Pedestrian detection has a wide range of applications in the field of multimedia, and significant progress has been made. However, in densely populated scenes, there are two problems: occlusion and mistake suppression of overlapping bounding boxes, which lead to false positives and false negatives, thereby degrading overall performance. To tackle these problems, firstly, by leveraging contextual information to capture correlations between pedestrians and backgrounds, we propose the Context Feature Fusion Module (CFFM), which alleviates the absence of crucial features caused by occlusion. Secondly, by combining the intersection over Union (IoU) and the distance between center points of overlapping bounding boxes, we propose Distance Set Non-Maximization Suppression (DSNMS), which tackles error suppression of overlapping bounding boxes. Finally, extensive experiments were conducted on the CrowdHuman dataset, yielding remarkable results for our method with an Average Precision (AP) of 91.22%, a Log average miss rate (MR^{-2}) of 40.26%, and a Jaccard Index (JI) of 83.54%. Furthermore, the visualization results of real-world scenes further validate the efficacy of our proposed method.

Keywords Densely populated · Pedestrian detection · Occlusion · Contextual information

✉ Lihua Hu
hlh@tyust.edu.cn

Yu Shao
S202120210775@stu.tyust.edu.cn

Jianhua Hu
jianhua.hu@ia.ac.cn

Jifu Zhang
zjf@tyust.edu.cn

Xinbo Wang
xinbo.wang@ia.ac.cn

¹ School of Computer Science and Technology, Taiyuan University of Science and Technology, Waliu, Taiyuan, Shanxi 030024, China

² Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

1 Introduction

Pedestrian detection is a computer vision task that involves accurately recognizing and locating pedestrians. It has various applications in the multimedia field, such as self-driving [1], video surveillance [2], multiple object tracking [20], and robotics [46]. Specifically, in multi-object tracking, accurate positioning information relies on effective pedestrian detection. Similarly, for self-driving systems to effectively take avoidance or deceleration measures, it is crucial to precisely detect and determine the position of pedestrians.

Currently, in low population density scenarios, general object detection [3, 5, 9, 11–14, 21, 24–26, 28, 29, 31, 32, 38, 39] has been proven to be excellent at detecting pedestrians. However, in crowded scenes with high pedestrian density, there will be serious occlusion and overlap between pedestrians. Consequently, the detector fails to accurately distinguish the instances, resulting in greatly reduced performance of general object detection.

Previous works [4, 7, 8, 18, 27, 33, 34, 36, 37, 40, 41, 44, 45] have tried to solve the above problems from different perspectives. However, they seem to overlook two crucial aspects: (1) The significance of contextual features in addressing the occlusion problem was overlooked by previous studies. Contextual information enables the exploration of relationships between pedestrians and their surroundings, thus helping to alleviate the problem of missing important features caused by occlusion. (2) In the post-processing stage, only relying on Intersection over Union (IoU) [19] as the suppression condition cannot effectively distinguish overlapping bounding boxes. This is because IoU mainly focuses on the degree of overlap, while in high-density crowded scenes, bounding boxes usually have the characteristic of overlapping each other.

In this paper, we propose a new method for pedestrian detection in crowded scenes. Firstly, we design the Context Feature Fusion Module (CFFM) to generate rich context information, which enhances a comprehensive understanding of occluded pedestrians by exploiting the relationship between pedestrians and surroundings. Secondly, inspired by Set NMS in CrowdDet [8], we propose Distance Set Non-Maximization Suppression (DSNMS), which combines the center distance and IoU of bounding boxes as new suppression conditions, to address the issue of falsely suppressing overlapping bounding boxes when relying solely on IoU. By designing CFFM and DSNMS, excellent performance is achieved in crowded pedestrian detection, which will also support us in playing a role in the field of dense crowd counting, dense crowd tracking, and autonomous driving in complex scenes.

The main contributions are summarised as follows:

- We propose the Context Feature Fusion Module (CFFM), which counteracts the impact of occlusion by leveraging contextual features to capture the relationship between pedestrians and backgrounds.
- The Distance Set Non-Maximization Suppression (DSNMS) is implemented to overcome incorrect suppression of overlapping bounding boxes.
- Our proposed method has been effectively demonstrated on the CrowdHuman [35] dataset and further verified by visualization in real-world scenes.

The paper is structured as follows: Section 2 reviews related work. Section 3 describes the CrowdDet [8] baseline. The proposed CFFM and DSNMS are described in Section 4. Section 5 shows the experimental results and visualizations of the proposed method. Finally, Section 6 is devoted to conclusion.

2 Related work

Here, we briefly review two types of pedestrian recognition algorithms in brief: the first type is intended to be applied in general scenarios, whereas the second type is designed for crowded scenarios.

2.1 General object detection

For sparsely populated scenes, general object detection methods exhibit effective capabilities, which can be divided into traditional methods and deep learning methods.

Traditional methods typically rely on manual feature extraction, such as haar features [24], histogram of oriented gradients (HOG) [9], local binary pattern (LBP) [31], and scale-invariant feature transform (SIFT) [29]. All of the above algorithms perform favorably in simple and sparse pedestrian scenarios. However, when confronted with crowded scenes, traditional methods struggle to handle complex extreme occlusion and overlapping.

As a result of the development of deep learning, there are now two types of detection algorithms: anchor-based and anchor-free. Among anchor-based detection algorithms, two categories can be further divided. The first one is the two-stage detection algorithms [5, 13, 14, 25, 32], which exhibit high accuracy but suffer from slower detection speeds. The other is one-stage algorithms [3, 26, 28, 39], which provide faster detection capabilities and are suitable for real-time tasks but are more prone to localization errors. Recently, anchor-free detection algorithms [11, 12, 21, 38] have also been developed. They remove the difficulty of setting hyperparameters caused by anchors. However, a new problem arises in accurately defining and distinguishing between positive and negative samples.

In summary, the aforementioned algorithms excel in simple and sparse pedestrian scenarios. Nonetheless, in scenarios with a high density of pedestrians, general methods may struggle to capture the relative position and response to occlusion between pedestrians, resulting in insufficient adaptation of the model to the complexities of crowded scenes.

2.2 Crowded pedestrian detection

In crowded scenes, the high pedestrian density and mutual occlusion pose novel challenges for pedestrian detection. Consequently, extensive research has been conducted in this specific domain, with the employed methodologies being presented as follows.

- **Part-based detection** Using prior knowledge and the visible parts of the pedestrians, the occlusion problem is solved by segmenting the human body into components. Typical methods include: Tian et al. [37] identified pedestrians based on the highest score in the part detector. Zhou et al. [44] accomplished detection by exploiting correlations between different body parts. Chi et al. [7] proposed the JointDet, which utilizes the structural relationship between the head and body for joint detection. These methods effectively mitigate occlusion effects and enhance detection performance. However, they exhibit high training complexity that necessitates the separate identification of different body parts.
- **Improving loss function** The loss function has received a lot of attention and is an essential part of pedestrian detection methods. Wang et al. [41] proposed the Repulsion Loss, which enhances localization ability by attracting its proposals and repulsing from surrounding ones. Earth Mover's Distance (EMD) Loss [8] proposed by CrowdDet,

which aims to ensure that multiple predictions of proposals can be matched with the best targets. Building upon the EMD Loss, the Repulsion Loss of Minimum (RLM) [36] was introduced to consider the inter-proposal relationships and further enhance detection performance. These methods are essential for improving pedestrian detection performance. However, current designs for loss functions do not adequately consider the complexity of occlusion conditions.

- **Improving NMS** Non-Maximization Suppression (NMS) [30] is used to remove the excess bounding boxes generated during the detection process. Bodla et al. [4] introduced Soft-NMS, which attenuates the scores of overlapping proposals and retains only the one with the highest score, thereby suppressing other redundant boxes. Liu et al. [27] proposed Adaptive NMS, which dynamically adjusts the threshold of NMS based on crowd density. Huang et al. [18] presented Representative Region NMS (R^2 NMS), which uses visible parts of targets as judgment criteria to suppress redundant boxes. Zhou et al. [45] introduced NOH-NMS, which employs a Gaussian distribution to locate objects near each box and enhance detection efficiency.
- **Other methods** Wang et al. [40] proposed the DeFCN, which achieves superior performance by employing a prediction-aware one-to-one label assignment strategy. Danila Rukhovich et al. [33] introduced IterDet, which is specifically designed to mitigate duplicate detection. Shang et al. [34] presented the V2F-Net, which enhances performance by decomposing occluded pedestrians into visible region detection and full-body estimation.

According to the above analysis, the complexity of the real world presents numerous problems for crowded pedestrian detection, even with the advancements made by earlier works. Therefore, further research and improvements are required to improve the accuracy and robustness of crowded pedestrian detection.

3 CrowdDet baseline

Our work builds upon the framework of CrowdDet. For the sake of completeness, we provide a brief description of the baseline in this section; interested readers can find more information in [8]. Next, we present the issue formulation.

3.1 CrowdDet baseline

CrowdDet is an anchor-based object detector, the core idea of CrowdDet is to generate multiple predictions from a single proposal, as depicted in Fig. 1 (cited from [8]). Firstly, EMD loss is suggested in order to improve the correspondence between the ground-truth and projections. Secondly, Set NMS is proposed to eliminate redundant boxes by determining whether two overlapping boxes originate from the same proposal, and if so, the suppression process is skipped. Finally, an optional Refinement Module (RM) is introduced to enhance detection results further.

3.2 Problem formulation

CrowdDet has achieved good performance in crowded pedestrian detection. However, it still faces the persistent challenge of false positives and false negatives. Figure 2 illustrates the problems in CrowdDet.

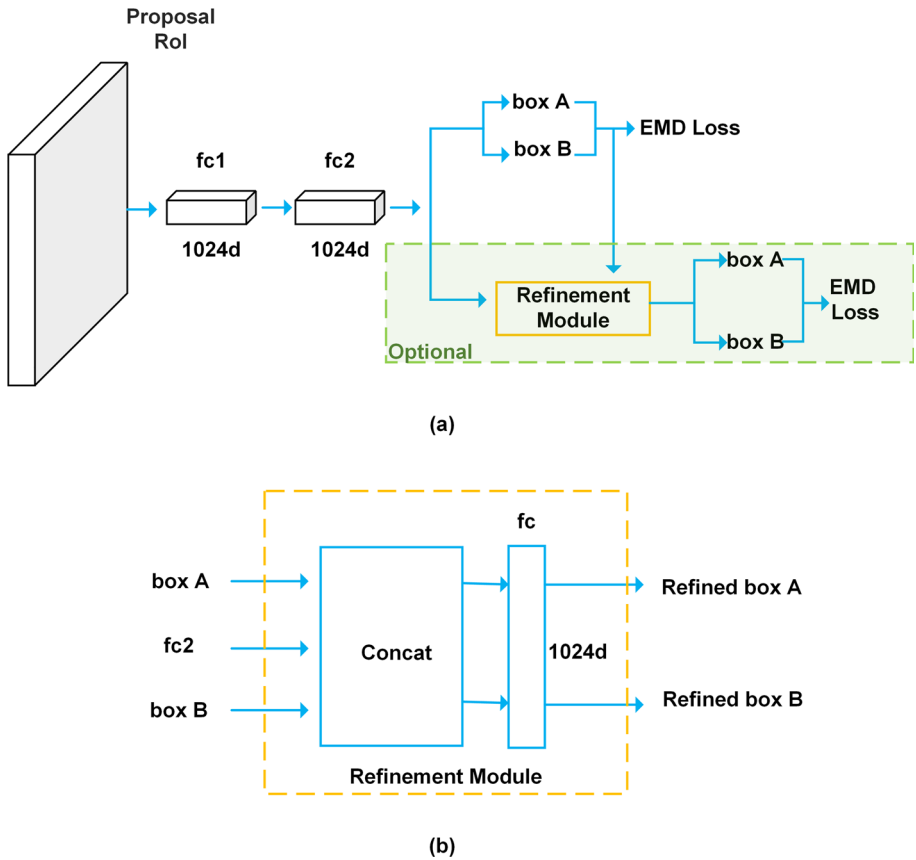


Fig. 1 CrowdDet baseline.(cited from [8])

In Fig. 2(a), The main features of the woman wearing red are conspicuously absent due to occlusion, and she has a comparable hair color with the man beside her, causing the detector to perceive them as a single entity, resulting in a false negative. Additionally, redundant bounding boxes are retained in the detection results shown in Fig. 2(b), causing false positives. To address the above issue, we adopt a different method. Regarding the problem in Fig. 2(a), we propose CFFM to leverage contextual information for mining the correlation between pedestrians and background, thereby mitigating errors caused by occlusion. As for Fig. 2(b), we solve this problem by proposing DSNMS, which no longer uses IoU as the condition for removing redundant bounding boxes.

4 Our method

In this section, the Network structure of our detector is described, and then the Context Feature Fusion Module (CFFM) and the Distance Set Non-Maximization Suppression (DSNMS) are introduced in detail.

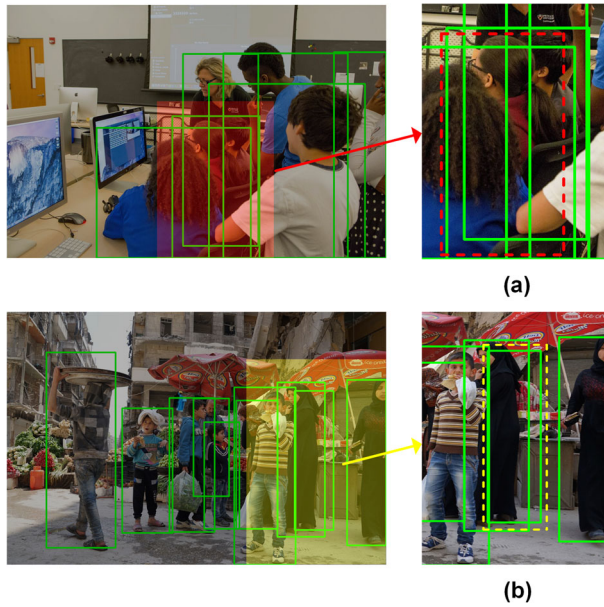


Fig. 2 Detection results of CrowdDet, where the red dashed line in (a) represents false negatives, and the yellow dashed line in (b) represents false positives

4.1 Network structure

The network structure of our detector is shown in Fig. 3. Firstly, we employ resnet50 [15] as the backbone, it is composed of five parts, namely conv1, conv.layer_2, conv.layer_3, conv.layer_4 and conv.layer_5. Downsampling is employed between each part to reduce the

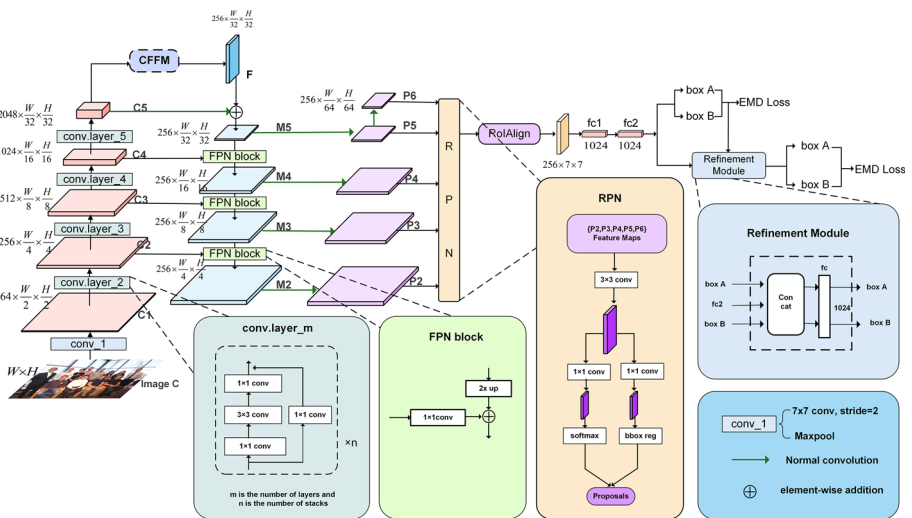


Fig. 3 Network structure of our detector

feature map size by half while doubling the number of channels. Max pooling is utilized for downsampling in conv1, while the subsequent four parts are built by stacking 3×3 and 1×1 convolutional layers with stack counts of 3, 4, 6 and 3, respectively. Input image C for feature extraction to obtain a series of feature layers $\{C1, C2, C3, C4, C5\}$.

Secondly, the feature layer $C5$ is subsequently fed into the Contextual Feature Fusion Module (CFFM) to facilitate contextual feature fusion, resulting in the generation of feature F . (The detailed structure of CFFM will be elaborated in Section 4.2) The features F and $C5$ are added and subsequently fed into the Feature Pyramid Network (FPN) [25] along with $\{C2, C3, C4\}$ for feature fusion. FPN performs upsampling on the higher layer feature map, applies a 1×1 convolution to adjust the channel dimensions of the lower layer feature map, and then element-wise adds the results of upsampling and convolution to obtain $\{M2, M3, M4, M5\}$. Finally, a 3×3 convolution operation is applied to $\{M2, M3, M4, M5\}$ to generate the feature layers $\{P2, P3, P4, P5\}$, while $P5$ undergoes an additional convolution operation, resulting in $P6$, which is half the size of $P5$. These features constitute the final output $\{P2, P3, P4, P5, P6\}$.

Subsequently, the Region Proposal Network (RPN) [32] is employed to generate proposals for each feature map. In RPN, there are two 1×1 conv branches, one is classified as target or background by softmax, and the other branch calculates accurate proposals by determining offsets with respect to the original image coordinates. To ensure effective supervision, RPN utilizes the following loss function:

$$L_0 = L_{rpn_cls} + L_{rpn_reg} \quad (1)$$

where L_{rpn_cls} is the Cross-Entropy classification loss and L_{rpn_reg} is the Smooth-L1 bounding box regression loss. RoIAlign [16] is used to unify the resulting proposals. After two fully connected layers, we predicted a set of instances using each proposal box proposed by CrowdDet. As shown in Fig. 1, for every proposal, two predictions are generated, and the prediction set and instance set of the proposal box was minimized using the EMD loss, which is shown in (2):

$$L(b_i) = \min_{\pi \in \Pi} \sum_{k=1}^k [L_{cls}(c_i^{(k)}, g_{\pi k}) + L_{reg}(I_i^{(k)}, g_{\pi k})] \quad (2)$$

Where b_i is the i th proposal box and π represents a particular permutation $1, 2, \dots, k$, the k th entry is π_k , $c_i^{(k)}$ and $I_i^{(k)}$ are the category and relative coordinates of the k th prediction of b_i respectively, and $g_{\pi k}$ is the groundtruth of π_k . $L_{cls}(\cdot)$ and $L_{reg}(\cdot)$ stand for classification and regression loss functions correspondingly. Simultaneously, we use the Refinement module to make a second prediction to improve the effect.

In the subsequent section, we will introduce the CFFM and the DSNMS in detail.

4.2 Context Feature Fusion Module (CFFM)

As mentioned in Section 3, CrowdDet also faces the challenge of error detection when dealing with occlusion. The presence of occlusion diminishes the effectiveness of feature extraction, resulting in the entire omission of some subjects and the incorrect treatment of two pedestrians as one entity. We think that combining multi-scale contextual information is useful for gaining a comprehensive understanding of the object, and can effectively alleviate the impact of occlusion. Therefore, inspired by earlier research [6] and [23], we propose the Context Feature Fusion Module (CFFM), whose structure is shown in Fig. 4.

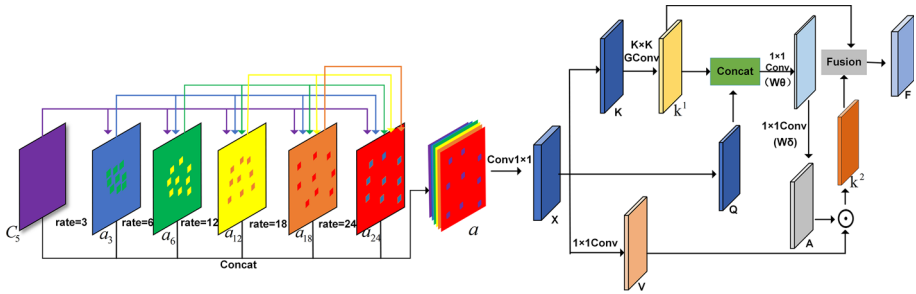


Fig. 4 Context Feature Fusion Module (CFFM). (W_θ with *Relu* activation function, W_δ without activation function, \odot denotes the local matrix multiplication operation)

Firstly, to obtain contextual information in different receptive fields, the feature layer C_5 is expanded by multi-path dilated convolutional layers [42] with different rates (e.g., rate = 3, 6, 12, 18, and 24). Secondly, in order to effectively integrate multi-scale information, the dense connections [17] are employed, where the output of each dilation layer is concatenated with the input feature maps and then fed into the next dilated layer. Thirdly, to maintain coarse-grained information, the outputs of all dilated layers are concatenated and inputted into a 1×1 convolutional layer, which fuses the coarse-grained and fine-grained features to generate the context feature X . In order to save computational resources while retaining precise object position information, we use normal convolutions instead of deformable convolutions. This process can be described as following equations:

$$a_3 = f_3(C_5) \tag{3}$$

$$a_6 = f_6(\text{concat}(a_3, C_5)) \tag{4}$$

$$a_{12} = f_{12}(\text{concat}(a_6, \text{concat}(a_3, C_5))) \tag{5}$$

$$a_{18} = f_{18}(\text{concat}(a_{12}, \text{concat}(a_6, \text{concat}(a_3, C_5)))) \tag{6}$$

$$a_{24} = f_{24}(\text{concat}(a_{18}, \text{concat}(a_{12}, \text{concat}(a_6, \text{concat}(a_3, C_5))))) \tag{7}$$

$$a = \text{concat}(a_3, a_6, a_{12}, a_{18}, a_{24}) \tag{8}$$

$$X = \text{GroupNorm}(\text{Conv}(a)) \tag{9}$$

Here, $\text{concat}(\cdot)$ represents the concatenation operation in the channel dimension. $f_i(\cdot)$ is dense block, which internal operation is shown in Fig. 5.

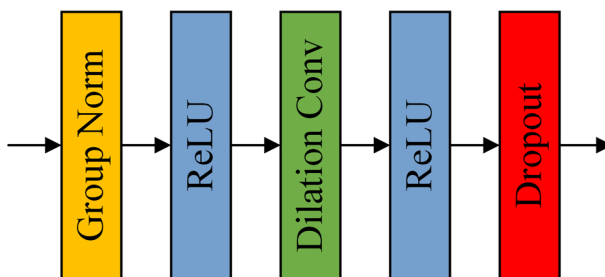


Fig. 5 The architecture of dense block. This includes the group normalization operation, ReLU activation function, dilation convolution, and dropout operation

Fourthly, the keys(K), queries(Q), and values(V) are defined as $K = X$, $Q = X$, and $V = XW_V$, respectively. The $k \times k$ group convolution [23] is used for neighbor keys within the $k \times k$ spatial grid and is performed to provide context for each key representation. K^1 is obtained, which naturally reflects the static context information between local adjacent keys and is the static context information representation of feature X . Next, concatenating K^1 and Q , performing two 1×1 convolutions (W_θ with *Relu* activation function, W_δ without activation function) to obtain the attention matrix A :

$$A = [K^1, Q] W_\theta W_\delta \tag{10}$$

Instead of using isolated query-key pairs, the attention matrix A is learned based on query feature Q and contextualized key features K^1 . This way enhances self-attention learning by mining additional guidance from the static context K^1 . Next, based on the context attention matrix A , the attended feature map K^2 is computed by aggregating all values V in typical self-attention (\odot denotes the local matrix multiplication operation):

$$K^2 = V \odot A \tag{11}$$

K^2 is a dynamic contextual representation of input feature X because it captures the dynamic feature interactions among inputs. Finally, the output F is obtained by integrating the static context K^1 and dynamic context K^2 through the attention mechanism [22]. At this point, F possesses rich multi-scale static and dynamic context information.

4.3 Distance Set Non-Maximization Suppression (DSNMS)

Non-Maximum Suppression (NMS) plays a key role in pedestrian detection. It is used twice in this task: first, in the network training phase, NMS is used to eliminate redundant proposals generated by RPN; the second time is in the prediction stage, which aims to eliminate redundant prediction boxes on the same target. Our proposed algorithm is only applied to the prediction process. Intersection over Union (IoU) is a crucial definition in NMS, which evaluates the overlap between two bounding boxes. IoU As originally defined, IoU quantifies the proportion of overlap between the proposal and the ground truth by calculating the ratio of their intersection to their union. This ratio is then compared against a predefined threshold; if it exceeds this threshold, the proposed box is classified as a positive sample; otherwise, it is considered negative. The mathematical expression for calculating IoU can be found in (12).

$$IoU = \frac{A \cap B}{A \cup B} \tag{12}$$

A and B denote the proposal box and the ground truth, respectively.

For CrowdDet, Set NMS employs IoU as a suppression condition to eliminate redundant bounding boxes during the post-processing stage. The expression of Set NMS is presented in (13), where S_i represents the confidence score corresponding to the i th bounding box, M denotes the bounding box with the highest confidence score, b_i denotes the i th bounding box among the remaining ones, C_M indicates the set index where M is located, C_{b_i} signifies the set index where b_i is located, N_t represents the threshold value. The specific process of Set NMS is as follows:

$$S_i = \begin{cases} S_i & IoU(M, b_i) < N_t \\ 0 & IoU(M, b_i) \geq N_t \text{ and } C_M \neq C_{b_i} \end{cases} \tag{13}$$

Firstly, the bounding boxes are sorted in descending order based on their confidence scores, and the box M with the highest confidence score is obtained. Subsequently, the IoU value is computed between M and each remaining bounding box b_i . If the IoU value exceeds or is equal to a predefined threshold and they do not belong to the same set (i.e., M and b_i do not come from the same proposal), then the confidence score for b_i is set to 0, and it is rejected for further consideration. This iterative process continues until all redundant boxes have been eliminated.

The Set NMS algorithm methodologies the similarity of bounding boxes by IoU, focusing on the extent of overlap. However, in crowded scenes, the bounding boxes themselves are very overlapping, therefore, relying solely on IoU becomes inadequate. To address this issue, DIOU [43] incorporates center distance, overlap rate, and scale size of the bounding box. In this study, we adopt DIOU [43] as the suppression criterion for Set NMS and propose a novel approach named DSNMS. The calculation process of DIOU is shown in (14).

$$DIOU = IoU - \frac{p^2(b, b^{gt})}{c^2} \quad (14)$$

Where b, b^{gt} represent the center of the bounding box and the ground truth, respectively; p represents the Euclidean distance between the two centers, and c^2 represents the diagonal distance of the minimum enclosed region that contains both bounding box and ground truth. The calculation function of the DSNMS algorithm is shown in (15):

$$S_i = \begin{cases} S_i & IoU - R_{DIOU}(M, b_i) < N_i \\ 0 & IoU - R_{DIOU}(M, b_i) \geq N_i \text{ and } C_M \neq C_{b_i} \end{cases} \quad (15)$$

The confidence score corresponding to the i th bounding box is denoted as S_i ; where M represents the bounding box with the highest confidence score; b_i denotes the i th bounding box in the remaining bounding box; C_M and C_{b_i} represent the set indices where M and b_i are located respectively; N_i is the threshold value; $R_{DIOU}(M, b_i)$ represents the DIOU between bounding boxes M and b_i . Assume that there are large IoU and distance values between bounding boxes, in this way, it becomes possible to identify multiple bounding boxes without rejecting them, thereby reducing the number of missed and false detections. The specific process is shown in Algorithm 1.

5 Experiments

The experiments conducted in this section aim to evaluate the performance of our method and assess its applicability in real-world scenarios. The subsequent subsections provide a comprehensive description of our experimental process, the obtained results, and its influence on practical pedestrian detection.

5.1 Datasets

To achieve a direct and fair comparison, we provide the experimental results in the following two datasets:

(1) CrowdHuman [35]: It consists of 15,000 training images, 4,370 validation images, and 5,000 test images. Notably, the dataset exhibits a high person density with an average of 23 instances per image. Moreover, it provides three types of bounding box annotations for each

Algorithm 1 Pseudo code of DSNMS.**Require:**

- 1: $B = \{b_1, b_2, \dots, b_n\}$, $S = \{s_1, s_2, \dots, s_n\}$, N_t
- 2: B is the list of initial detection boxes.
- 3: S contains corresponding detection scores.
- 4: C contains the set index.
- 5: N_t is the threshold of the DSNMS algorithm.

Ensure:

- 6: D is the final detection results.
- 7: $D \leftarrow \emptyset$
- 8: **while** $B \neq \emptyset$ **do**
- 9: $m \leftarrow \arg \max S$,
- 10: $M \leftarrow b_m$
- 11: $D \leftarrow D \cup M$; $B \leftarrow B - M$
- 12: **for** b_i in B **do**
- 13: **if** $DIoU(M, b_i) \geq N_t$ and $C_M \neq C_{b_i}$ **then**
- 14: $B \leftarrow B - b_i$, $S \leftarrow S - s_i$
- 15: **end if**
- 16: **end for**
- 17: **end while**

pedestrian instance: head bounding box, visible area bounding box, and full body bounding box.

(2) Our dataset: We took 100 pictures of crowded pedestrians in real situations and made predictions to further verify the actual effect of the proposed method. In addition, to verify the robustness of our method, we collected 200 images of severe weather for dense pedestrians in real-life challenging weather conditions, such as fog and snow.

5.2 Evaluation metrics

We utilize the following metrics for evaluation:

AP: Averaged Precision (AP) reflects the precision and recall of the detection results. The larger the AP, the better the performance.

MR^{-2} [10]: Log average miss rate (MR^{-2}) is a widely adopted metric for evaluating pedestrian detection, as shown in (16). False Negative (FN) refers to instances predicted as negative but actually positive. Conversely, True Positive (TP) denotes cases where both the prediction and actuality are positive. A smaller value of MR^{-2} value corresponds to superior performance.

$$MR^{-2} = \frac{FN}{TP + FN} \quad (16)$$

JJ [28]: The Jaccard index (JI) primarily assesses the level of overlap between the predicted set P and the ground truth label set G , as depicted in (17). A higher JI indicates superior performance.

$$JI = \frac{|P \cap G|}{|P \cup G|} \quad (17)$$

5.3 Implementation details

In this paper, we employ CrowdDet as the baseline and train our model on a single NVIDIA RTX 2080Ti GPU. To initialize the network weights, we utilize the ResNet50 pre-trained

Table 1 Ablation study. (The best results are highlighted in bold.)

Baseline	CFFM	DSNMS	AP%	MR ⁻² %	JI%
✓	×	×	90.48	41.68	82.47
✓	✓	×	90.99	40.48	83.33
✓	×	✓	90.74	41.44	82.72
✓	✓	✓	91.22	40.26	83.54

weights provided by the official CrowdDet. To normalize the size of the input image, we resize it to 800 pixels on the short side and 1400 pixels on the long side. We train for 50 epochs using SGD with momentum 0.9 as an optimizer. The initial learning rate is 1.25×10^{-3} , with decay rates of 0.1 and 0.01 at the 40th and 45th epochs, respectively. The batch size is set to 4, and the post-processing threshold is set to 0.5.

5.4 Ablation study

Ablation studies are conducted to understand better how different choices affect the performance of our proposed method. The experimental results are presented in Table 1. Analysis of the results reveals that:

- (1) When we adopted the CFFM, it obviously improved all the indicators. Notably, there was a 0.51% increase in AP and a 0.86% increase in JI, indicating enhanced detection capabilities for instances. What's more, we found that MR⁻² also reduced by a large margin, at 1.2%, suggesting that CFFM did not introduce more false predictions, and through the fusion of contextual information, enhanced the overall understanding of pedestrians and correctly detected pedestrians with occlusions and small scales.
- (2) When switching the Set NMS in the baseline to DSNMS, we observe a 0.26% boost in AP, a 0.24% reduction in MR⁻², and a 0.25% improvement in JI. Therefore, for this task, adopting Diou, which considers both the overlap area and center point distance between two bounding boxes during suppression, yields better performance than relying on IoU.
- (3) The proposed method achieves the best results by utilizing both CFFM and DSNMS, demonstrating excellent detection performance with AP, MR, and JI scores of 91.22%, 40.26%, and 83.54% respectively. Our experimental results indicate that compared to the baseline method, our approach yields a significant improvement in AP (0.74%) and JI (1.07%), while also reducing MR⁻² by 1.42%. These results validate the effectiveness of our proposed method.

5.5 Comparative experiments

In order to facilitate a comprehensive quantitative comparison with state-of-the-art competitors, we use the CrowdHuman dataset as the base dataset, and the results are presented in Table 2, clearly demonstrating that:

- (1) For crowded scenes with high density, our proposed method has a relative improvement over the general detection methods, such as FPN and Cascade R-CNN. As can be observed from Table 2, when compared to FPN, our method delivers 4.57% improvement in AP, 2.17% reduction in MR⁻², and 4.05% improvement in JI. Compared with Cascade R-CNN, our method improves by 5.62% on AP, 2.94% on JI, and reduces by 2.74% on

Table 2 Comparative experiments. (The best results are highlighted in bold.)

Methods	AP%	MR ⁻² %	JI%
FPN [25]	86.65	42.43	79.49
CascadeR-CNN [5]	85.6	43.0	80.6
JointDet [7]	–	46.5	–
CrowdDet [8]	90.7	41.4	82.3
MFPN [36]	90.96	40.24	83.12
Soft-NMS [4]	88.2	42.9	79.8
Adaptive NMS [27]	84.71	49.73	–
R ² NMS [18]	89.29	43.35	–
NOH-NMS [45]	89.0	43.9	–
DeFCN [40]	89.1	48.9	–
IterDet [33]	88.08	49.44	–
V2F-Net [34]	91.03	42.28	–
Ours	91.22	40.26	83.54

MR⁻². The rationale for these enhancements is that in conditions with dense crowds, occlusion, and relative position relationships among individuals, FPN and Cascade R-CNN might not adequately account for these intricate variables. However, our approach addresses the complications by making use of the interaction between pedestrians and the background, demonstrating more notable benefits in indications.

- (2) Compared with part-based detection methods, typically such as JointDet, our method significantly reduces MR⁻² indicating its superiority in false positives. Specifically, our method achieves a 6.24% reduction. We analyze the possible reasons as follows: the change of pedestrian pose has a significant impact on the performance of JointDet.
- (3) In contrast to the improving loss function method, such as MFPN (which enhances the EMD loss function proposed in CrowdDet), our method maintains similar performance in terms of MR⁻² but achieves an improvement of 0.26% in AP, and achieves an improvement of 0.42% in JI. This improvement is from deeper thinking about the problem, and we argue that MFPN does not fully consider the critical role of contextual features for crowded pedestrian detection in its design. Our method emphasizes contextual features, which improves detection accuracy and robustness by better capturing the interrelation and relative position of pedestrians in dense environments.
- (4) Compared with the improving NMS method, such as Soft-NMS, Our method improves 3.02% in AP, decreases 2.64% in MR⁻², and improves 3.74% in JI. The improvement can be attributed to the multiple advantages of our method, with one key factor being the enhancement of the NMS method and a particular emphasis on addressing false detections caused by occlusion and pose changes.
- (5) The AP, MR, and JI values of our method are higher than other methods, such as CrowdDet, MFPN, DeFCN, IterDet, and V2F-Net. Specifically, compared with the classical IterDet, the performance of our method is improved by 3.14% in AP and reduced by 9.18% in MR⁻². This indicates the good performance of the proposed method in crowded scenes.

Table 3 Robustness experiments with randomly erased pixels on the CrowdHuman-1 dataset. (The best results are highlighted in bold.)

Methods	AP%	MR ⁻² %	JI%
CrowdDet	88.14	44.61	80.39
FPN	84.49	45.44	77.48
RetinaNet	79.75	58.42	71.28
Ours	89.86	42.06	81.79

5.6 Robustness experiments

To evaluate the performance of our method under adverse factors such as poor image quality, we constructed two new datasets based on the CrowdHuman dataset, named CrowdHuman-1 and CrowdHuman-2, respectively. Our comparison methods include CrowdDet [8], FPN [25], RetinaNet [26], etc. Specifically:

Firstly, to verify the robustness of our method in severely occluded environments, we have constructed a new dataset CrowdHuman-1 based on the CrowdHuman dataset. Compared to the images in the CrowdHuman dataset, each image in the CrowdHuman-1 dataset is randomly pixel erased, which is 3% to 8% of the entire image, and the aspect ratio of the erased area varies between 0.3 and 1. The experimental results are presented in Table 3.

Compared with Table 2, we can conclude that:

- 1) Under the CrowdHuman-1 dataset, all evaluation metrics of these four methods have decreased. Specifically, our method reduces AP by 1.36%, MR⁻² decreases by 1.8%, and JI decreases by 1.75%. In contrast, the baseline method shows a larger decrease of 2.56% on AP, 3.21% on MR⁻², and 1.91% on JI. Furthermore, FPN demonstrates reductions of 2.16% on AP, 3.01% on MR⁻², and 2.01% on JI as well.
- 2) Compared with CrowdDet and FPN, our method exhibits minimal performance degradation.
- 3) Based on the above results, it is evident that CFFM has a preferable performance in severe occlusion environments, and it can effectively address occlusion problems. Section 5.7 provides visualizations of detection results for both the baseline and our method, further confirming the robustness of our method to adverse effects caused by the random erasure of pixels.

Secondly, to diminish the image quality, we have constructed a new dataset CrowdHuman-2 based on CrowdHuman dataset. In CrowdHuman-2, we add Gaussian noise with a mean of 0 and a standard deviation of 20 to each image. The experimental results are presented in Table 4.

Based on Table 4, we can conclude that:

- 1) Our method achieves an AP of 87.33%, MR⁻² of 45.10%, and JI of 78.18%.

Table 4 Robustness experiments with the addition of Gaussian noise on the CrowdHuman-2 dataset. (The best results are highlighted in bold.)

Methods	AP%	MR ⁻² %	JI%
CrowdDet	86.72	47.3	77.26
FPN	83.00	48.30	75.08
RetinaNet	78.75	60.83	69.78
Ours	87.33	45.10	78.18

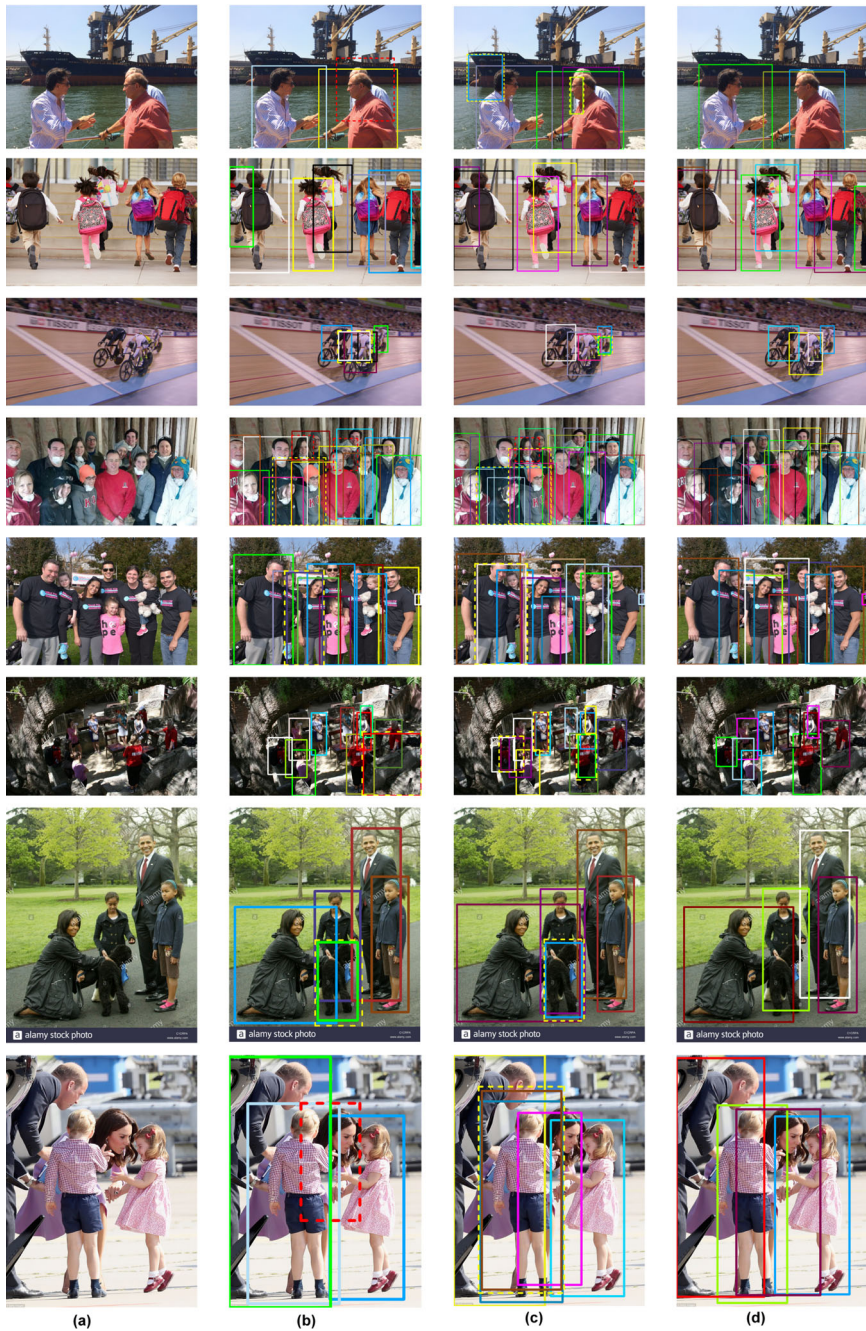


Fig. 6 Visualization of the CrowdHuman dataset. (a) stands for the original images, (b) stands for FPN detection results, (c) stands for CrowdDet baseline detection results, and (d) stands for our method detection results. The yellow dashed line indicates false positive instances, while the red dashed line indicates false negatives

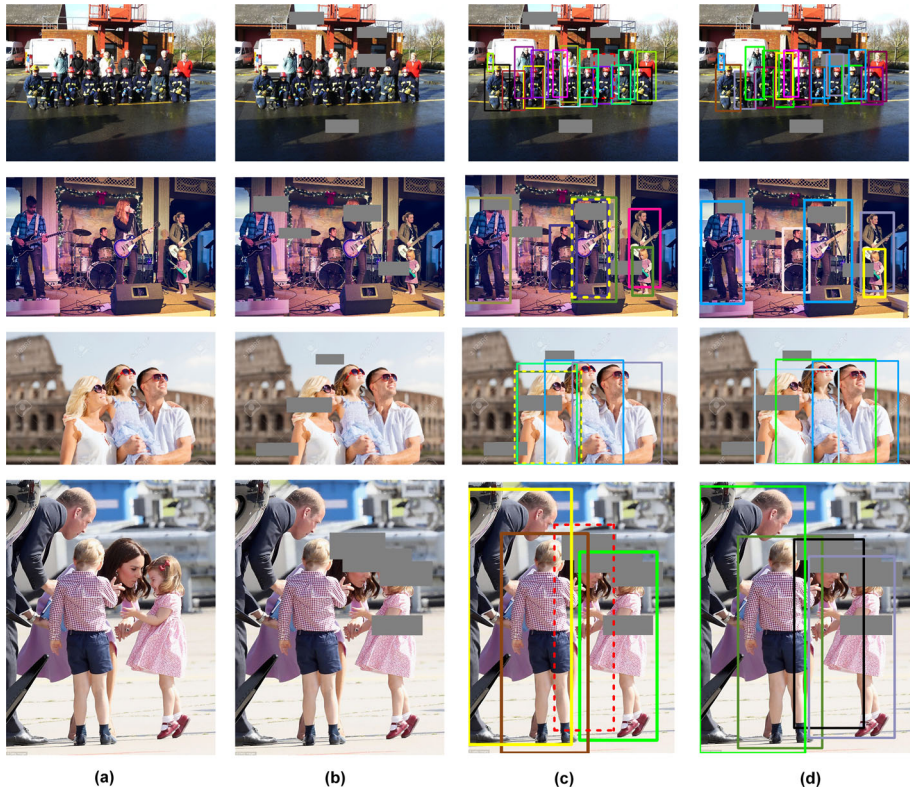


Fig. 7 Visualization results under the CrowdHuman-1 dataset. (a) stands for the original image, (b) stands for the image with randomly erased pixels, (c) stands for the CrowdDet baseline detection results, and (d) stands for the detection results of our method. Yellow dashed lines represent false positive instances and red dashed lines represent false negative instances

- 2) Compared to the baseline, our method improves by 0.61% on AP, 2.2% on MR^{-2} , and 0.92% on JI. There are fewer missed detections and false detections in our method, which also shows that our method has certain robustness in dealing with Gaussian noise.

5.7 Visualization of the crowdHuman dataset

In this section, we present partial visualizations based on the CrowdHuman dataset. The figures in Fig. 6 display the outcomes of FPN, CrowdDet baseline, and our method. under the CrowdHuman dataset. Figure 7 showcases the visualization results of the baseline and our method under the CrowdHuman-1 dataset. Furthermore, Fig. 8 demonstrates the visualization results of both the baseline and our method under the CrowdHuman-2 dataset. Visual analysis reveals that FPN and the baseline exhibit inaccuracies in detecting severely occluded pedestrians in certain instances, and are seriously affected by noise. Conversely, our method exhibits comprehensive and accurate pedestrian detection capabilities with notable robustness against severe occlusion and noise.

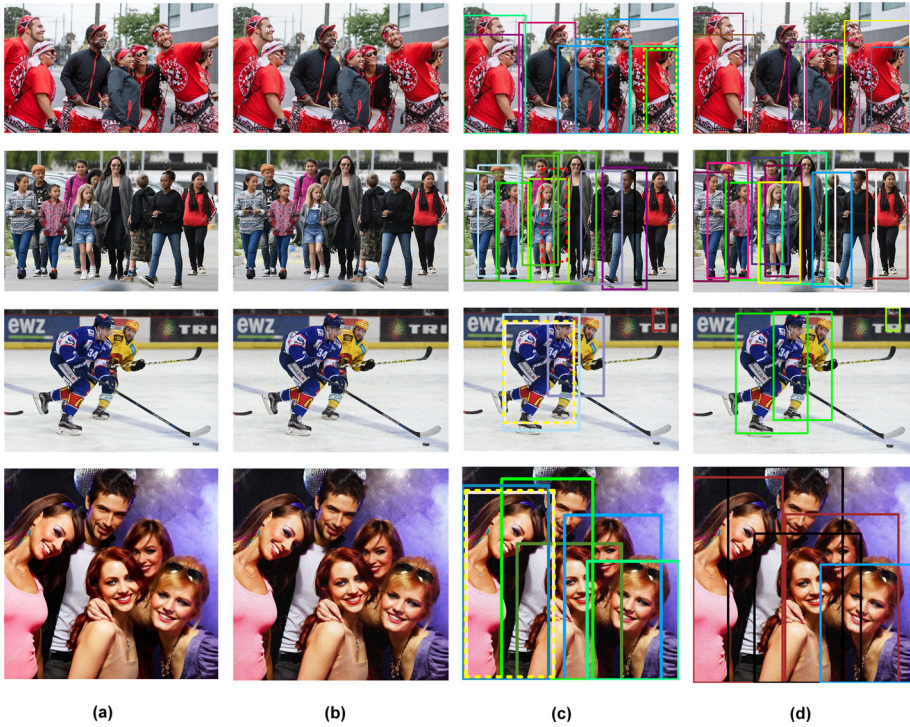


Fig. 8 Visualization results under the CrowdHuman-2 dataset. (a) stands for the original image, (b) stands for the image with Gaussian noise incorporated, (c) stands for the CrowdDet baseline detection results, and (d) stands for the detection results of our method. Yellow dashed lines represent false positive instances and red dashed lines represent false negative instances



Fig. 9 Visualization results of dense pedestrians in realistic scenarios. (a) stands for the original images, (b) stands for FPN detection results, (c) stands for CrowdDet baseline detection results, and (d) stands for our method detection results. The yellow dashed line indicates false positive instances, while the red dashed line indicates false negatives



Fig. 10 Visualization results of dense pedestrians in bad weather. (a) stands for the original images, (b) stands for FPN detection results, (c) stands for CrowdDet baseline detection results, and (d) stands for our method detection results. The yellow dashed line indicates false positive instances, while the red dashed line indicates false negatives

5.8 Visualization of our dataset

To verify the effectiveness and robustness of our method in real-world scenarios, 100 dense pedestrian images in common scenarios and 200 dense pedestrian images in bad weather are collected. Partial detection results of CrowdDet, FPN, and our method are shown in Figs. 9 and 10. It can be seen that our method can effectively reduce false positives and false negatives. This outcome not only signifies theoretical advancements but also holds significant practical implications for ensuring safety and efficiency within densely populated environments.

6 Conclusion

In this paper, we observe that occlusions and overlapping bounding boxes are wrongly suppressed in crowded scenes with high pedestrian density. Based on this observation, the Context Feature Fusion Module (CFFM) and Distance Set Non-Maximization Suppression (DSNMS) are proposed. Firstly, by exploiting the context information to fully explore the relationship between pedestrians and their backgrounds, CFFM effectively solves the problem of missing key features caused by occlusion. Secondly, the DSNMS improves error rejection in overlapping bounding boxes by combining IoU and the distance between their center points. Finally, extensive experiments and visualization results are presented to validate the performance of our proposed method. Through a large number of experiments and visualization results, our method shows excellent performance in the field of crowded pedestrian detection. Addition-

ally, our method also has good robustness under conditions of bad weather and poor picture quality. Therefore, our method is not only suitable for pedestrian detection in crowded scenarios, but also useful in other fields, such as intelligent traffic management, urban safety monitoring, and pedestrian flow analysis tasks.

However, there is no free lunch. Here, we do not claim that our method is better than other methods in all scenes, which is true. Since our method is based on two-stage, it has some shortcomings, such as long running time. Therefore, we are currently engaged in developing a lightweight architecture for future endeavors.

Acknowledgements This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62273248, the Computer Vision Joint Training Demonstration Base of Taiyuan University of Science and Technology (JD2022005).

Database availability statement The dataset in our study is based on CrowdHuman dataset (<https://www.crowdhuman.org/download.html>). Access to the visualization of our dataset can be obtained from the corresponding author upon reasonable request.

Declarations

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

References

1. Alfred Daniel J, Chandru Vignesh C, Muthu BA et al (2023) Fully convolutional neural networks for lidar–camera fusion for pedestrian detection in autonomous vehicle. *Multimedia Tools and Applications* pp 1–24
2. Ansari MA, Singh DK (2021) Human detection techniques for real time surveillance: a comprehensive survey. *Multimed Tools Appl* 80:8759–8808
3. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
4. Bodla N, Singh B, Chellappa R, et al (2017) Soft-nms–improving object detection with one line of code. In: *Proceedings of the IEEE international conference on computer vision*, pp 5561–5569
5. Cai Z, Vasconcelos N (2018) Cascade r-cnn: Delving into high quality object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 6154–6162
6. Cao J, Chen Q, Guo J et al (2020) Attention-guided context feature pyramid network for object detection. [arXiv:2005.11475](https://arxiv.org/abs/2005.11475)
7. Chi C, Zhang S, Xing J et al (2020) Relational learning for joint head and human detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 10647–10654
8. Chu X, Zheng A, Zhang X et al (2020) Detection in crowded scenes: one proposal, multiple predictions. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 12214–12223
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, pp 886–893
10. Dollar P, Wojek C, Schiele B et al (2011) Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell* 34(4):743–761
11. Duan K, Bai S, Xie L et al (2019) Centernet: keypoint triplets for object detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp 6569–6578
12. Ge Z, Liu S, Wang F et al (2021) Yolox: exceeding yolo series in 2021. [arXiv:2107.08430](https://arxiv.org/abs/2107.08430)
13. Girshick R (2015) Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp 1440–1448
14. Girshick R, Donahue J, Darrell T et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580–587
15. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778

16. He K, Gkioxari G, Dollár P et al (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
17. Huang G, Liu Z, Van Der Maaten L et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
18. Huang X, Ge Z, Jie Z et al (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
19. Jiang B, Luo R, Mao J et al (2018) Acquisition of localization confidence for accurate object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 784–799
20. Lahmyed R, El Ansari M, Kerkaou Z (2022) A novel visible spectrum images-based pedestrian detection and tracking system for surveillance in non-controlled environments. *Multimed Tools Appl* 81(27):39275–39309
21. Law H, Deng J (2018) Cornernet: detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750
22. Li X, Wang W, Hu X et al (2019) Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 510–519
23. Li Y, Yao T, Pan Y et al (2022) Contextual transformer networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 45(2):1489–1500
24. Lienhart R, Maydt J (2002) An extended set of haar-like features for rapid object detection. In: Proceedings. international conference on image processing, IEEE, pp I–I
25. Lin TY, Dollár P, Girshick R et al (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125
26. Lin TY, Goyal P, Girshick R et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
27. Liu S, Huang D, Wang Y (2019) Adaptive nms: refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6459–6468
28. Liu W, Anguelov D, Erhan D et al (2016) Ssd: single shot multibox detector. In: *Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I* 14, Springer, pp 21–37
29. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision, Ieee, pp 1150–1157
30. Neubeck A, Van Gool L (2006) Efficient non-maximum suppression. In: 18th international conference on pattern recognition (ICPR'06), IEEE, pp 850–855
31. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24(7):971–987
32. Ren S, He K, Girshick R et al (2015) Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process* 28
33. Rukhovich D, Sofiiuk K, Galeev D et al (2021) Iterdet: iterative scheme for object detection in crowded environments. In: *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, January 21–22, 2021, Proceedings, Springer*, pp 344–354
34. Shang M, Xiang D, Wang Z et al (2021) V2f-net: explicit decomposition of occluded pedestrian detection. [arXiv:2104.03106](https://arxiv.org/abs/2104.03106)
35. Shao S, Zhao Z, Li B et al (2018) Crowdhuman: a benchmark for detecting human in a crowd. [arXiv:1805.00123](https://arxiv.org/abs/1805.00123)
36. Shao X, Wang Q, Yang W et al (2021) Multi-scale feature pyramid network: a heavily occluded pedestrian detection network based on resnet. *Sensors* 21(5):1820
37. Tian Y, Luo P, Wang X et al (2015) Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision, pp 1904–1912
38. Tian Z, Shen C, Chen H et al (2019) Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636
39. Wang CY, Bochkovskiy A, Liao HYM (2023) Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7464–7475
40. Wang J, Song L, Li Z et al (2021) End-to-end object detection with fully convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15849–15858
41. Wang X, Xiao T, Jiang Y et al (2018) Repulsion loss: detecting pedestrians in a crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7774–7783
42. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. [arXiv:1511.07122](https://arxiv.org/abs/1511.07122)

43. Zheng Z, Wang P, Liu W et al (2020) Distance-iou loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI conference on artificial intelligence, pp 12993–13000
44. Zhou C, Yuan J (2019) Multi-label learning of part detectors for occluded pedestrian detection. *Pattern Recognit* 86:99–111
45. Zhou P, Zhou C, Peng P et al (2020) Noh-nms: improving pedestrian detection by nearby objects hallucination. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 1967–1975
46. Zou M, Yu J, Lu B et al (2022) Active pedestrian detection for excavator robots based on multi-sensor fusion. In: 2022 IEEE International Conference on Real-time Computing and Robotics (RCAR), IEEE, pp 255–260

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

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