applied sciences

MDPI

Article

Multi-Attribute NMS: An Enhanced Non-Maximum Suppression Algorithm for Pedestrian Detection in Crowded Scenes

Wei Wang [1,†], Xin Li [1,2,†], Xin Lyu [1,2,*], Tao Zeng [1], Jiale Chen [1] and Shangjing Chen [1]

[1] College of Computer and Information, Hohai University, Nanjing 211100, China; wang-wei@hhu.edu.cn (W.W.); li-xin@hhu.edu.cn (X.L.); tzeng.nj@hhu.edu.cn (T.Z.); 211307040002@hhu.edu.cn (J.C.); hhu_csj@hhu.edu.cn (S.C.)
[2] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China
* Correspondence: lvxin@hhu.edu.cn
† These authors contributed equally to this work.

Featured Application: In this paper, a Multi-Attribute Non-Maximum Suppression (MA-NMS) algorithm, which adaptively adjusts suppression based on density and count attributes, is proposed. It can help detectors to obtain more accurate predictions in crowded scenes, which will benefit subsequent tasks regarding pedestrian detection, like face recognition, pedestrian re-identification, and human interaction. The proposed MA-NMS and the attribute branch (ATTB) can be easily embedded into generic pedestrian detectors for performance improvement. Moreover, with the proposed ATTB, a pedestrian detector is proposed, based on the MA-NMS, which can be directly used for pedestrian detection in crowded scenes, such as shopping malls, streets, airports, etc.

Abstract: Removing duplicate proposals is a critical process in pedestrian detection, and is usually performed via Non-Maximum Suppression (NMS); however, in crowded scenes, the detection proposals of occluded pedestrians are hard to distinguish from duplicate proposals, making the detection results inaccurate. In order to address the above-mentioned problem, the authors of this paper propose a Multi-Attribute NMS (MA-NMS) algorithm, which combines density and count attributes in order to adaptively adjust suppression, effectively preserving the proposals of occluded pedestrians while removing duplicate proposals. In order to obtain the density and count attributes, an attribute branch (ATTB), which uses a context extraction module (CEM) to extract the context of pedestrians, and then, concatenates the context with the features of pedestrians in order to predict both the density and count attributes simultaneously, is also proposed. With the proposed ATTB, a pedestrian detector, based on MA-NMS, is constructed for pedestrian detection in crowded scenes. Extensive experiments are conducted using the CrowdHuman and CityPersons datasets, and the results show that the proposed method outperforms mainstream methods on AP (average precision), $Recall$, and MR^{-2} (log-average miss rate), sufficiently validating the effectiveness of the proposed MA-NMS algorithm.

Keywords: pedestrian detection; intra-class occlusion; non-maximum suppression; multi-attribute

Citation: Wang, W.; Li, X.; Lyu, X.; Zeng, T.; Chen, J.; Chen, S. Multi-Attribute NMS: An Enhanced Non-Maximum Suppression Algorithm for Pedestrian Detection in Crowded Scenes. *Appl. Sci.* **2023**, *13*, 8073. https://doi.org/10.3390/app13148073

Academic Editors: Junchi Yan and Minghao Guo

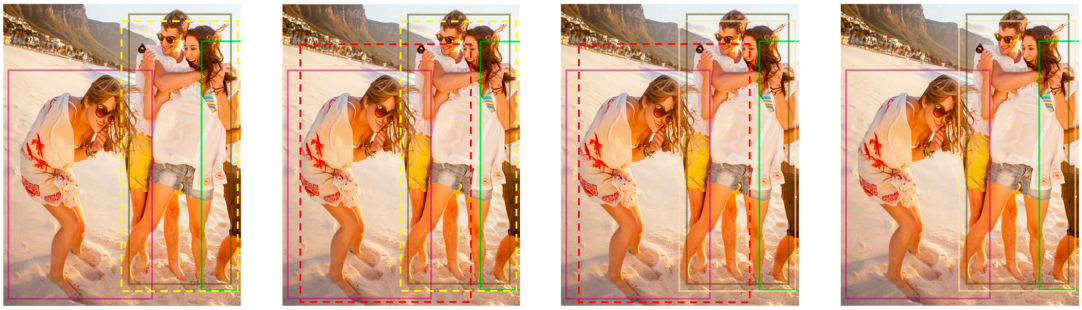
Received: 18 June 2023
Revised: 1 July 2023
Accepted: 3 July 2023
Published: 11 July 2023

1. Introduction

Pedestrian detection [1] is a popular research topic in computer vision that has been widely applied in automatic driving [2], video surveillance [3], robotics [4], etc. As a fundamental task, pedestrian detection drives the development of research, such as face recognition [5], pedestrian re-identification [6], and human interaction [7].

With the wide spread of convolutional neural networks, numerous object detectors [8–10] have been proposed based on deep learning features, some of which have been applied to pedestrian detection after fine-tuning. Furthermore, several proposed pedestrian detectors [11–13]

have attempted to detect pedestrians using high-level semantic features [14,15] from another perspective. In order to obtain all the pedestrians in the images, most of the above-mentioned pedestrian detectors are dense detectors, which generate multiple proposals for each pedestrian. However, each pedestrian is represented by only one optimal proposal in the final predictions, for which NMS is widely used to remove duplicates. NMS was originally implemented via the greedy algorithm, which is known as Greedy NMS. For each pedestrian, it uses a rigid threshold in order to divide a uniform suppression interval, and directly discards all other proposals within the interval to remove the duplicates, which satisfied the need for pedestrian detection in sparse scenes in early studies. Recently, as artificial intelligence [7] has gradually penetrated our lives, the demand for pedestrian detection in more crowded scenes, such as shopping malls, railway stations, airports, and streets, has increased. However, dense crowds and various shooting angles mean the pedestrians in the images are heavily occluded by each other, resulting in highly similar proposals between pedestrians and their surrounding occluded pedestrians, and leading Greedy NMS to mistakenly consider the proposals of occluded pedestrians as duplicate ones, and then, discard them. The above-mentioned problem finally means that the detectors fail to meet the demand for pedestrian detection in crowded scenes, as shown in Figure 1a.



(**a**) (**b**) (**c**) (**d**)

Figure 1. Results of different NMS algorithms. The yellow dashed boxes show the occluded pedestrians that are incorrectly removed, and the red dashed boxes highlight the false positives mistakenly retained by NMS. (**a**) The result of Greedy NMS; (**b**) the result of Soft NMS; (**c**) the result of Adaptive NMS; and (**d**) the result of MA-NMS (ours).

Recent studies [16–18] have attempted to soften suppression in order to decrease the tendency of NMS to incorrectly remove proposals belonging to occluded pedestrians. Moreover, the authors of [19–21] attempted to remove duplicate proposals according to the distance of their attributes from others. Though significant breakthroughs have been made, numerous false positives are generated, and highly occluded pedestrians continue to be missed, as shown in Figure 1. We compared the current NMS algorithms and found the following common drawbacks:

- For each pedestrian, only a rigid or a dynamic threshold is used to divide the single suppression interval, and all the proposals within the interval are considered duplicates;
- A uniform suppression operation is applied in the suppression interval, such as discarding, or a suppression weight function for re-scoring, making it more difficult to remove the highly similar duplicate proposals of pedestrians while preserving the proposals of occluded pedestrians.

The above-mentioned drawbacks make it difficult for the current NMS algorithms to remove various duplicate proposals that are too close to or distant from the annotations, and retain the proposals of occluded pedestrians at the same time. Thus, there is still room for research on making an NMS that can accurately preserve occluded pedestrians while removing duplicate proposals.

In this paper, we propose a Multi-Attribute NMS (MA-NMS), which adaptively adjusts suppression based on density and count attributes. It refines the traditional single

suppression interval into strong and weak suppression intervals in order to separate the proposals of potentially occluded pedestrians and duplicate ones. Within the corresponding interval, strong or weak suppression is adaptively assigned in order to enhance the penalties of duplicate proposals while preserving the proposals of occluded pedestrians. Furthermore, suppression factors are applied to adjust the intensity of the strong and weak suppression, in order to further suppress duplicate proposals. As shown in Figure 1d, our proposed MA-NMS can effectively preserve occluded pedestrians while removing duplicate proposals. A comparison of the flows of MA-NMS and other well-known NMS algorithms is shown in Figure 2. It clearly reflects that Multi-Attribute NMS demonstrates more adaptive treatments for proposals with different duplicate probabilities.

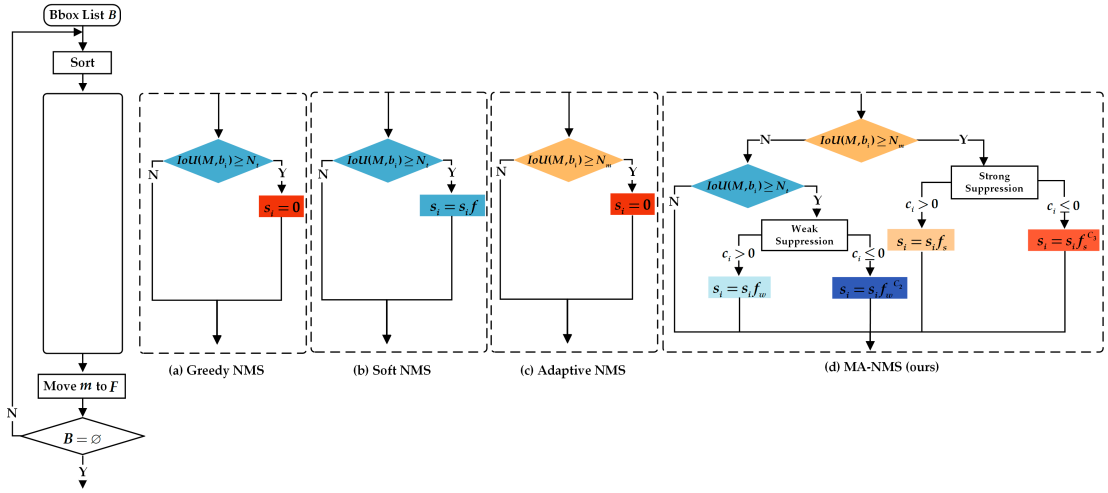


Figure 2. Comparison of the flows of different NMS algorithms. (**a**) The flow of Greedy NMS; (**b**) the flow of Soft NMS; (**c**) the flow of Adaptive NMS; and (**d**) the flow of MA-NMS, which is proposed in this paper.

In addition, an attribute branch (ATTB) was designed, with which to obtain the density and count attributes of pedestrians, and simultaneously guide the adjustment of suppression intervals and intensity of MA-NMS. Notably, ATTB can be easily embedded into generic pedestrian detectors, which can then be used to guide the adjustment of NMS to mitigate the impact of intra-class occlusion. With the proposed ATTB, a pedestrian detector for crowded scenes was constructed based on MA-NMS, which enables more accurate predictions for pedestrian detection in crowded scenes. Moreover, the annotations of count and density attributes required for training are generated on the basis of the existing full-body annotations, without additional annotations needed. Extensive experiments were conducted using the CrowdHuman [22] and CityPersons [23] datasets. Our method delivers promising progress in crowded pedestrian detection, most notably, a 6.5% improvement in *Recall* compared with the baseline in CrowdHuman.

Our contributions can be summarized as follows:

- In order to accurately remove duplicate detections, a Multi-Attribute NMS (MA-NMS) is proposed. Rather than using a uniform suppression interval, it refines the suppression intervals based on density attributes to perform adaptive suppression, which effectively preserves potentially occluded pedestrians, while substantially removing duplicate proposals. Additionally, the suppression intensity is further adjusted according to the count attributes, which further reduces the generation of false positives.
- To obtain the density and count attributes of pedestrians, an attribute branch (ATTB) is proposed. In ATTB, a context extraction module (CME) is designed to obtain the context of pedestrians. Furthermore, it concentrates the context with the feature of pedestrians from the generic detection branch to obtain more representative feature, which allows for a more comprehensive consideration of pedestrians and their surrounding occluded pedestrians.

- With the proposed ATTB, a pedestrian detector for crowded scenes is constructed based on MA-NMS. It simultaneously considers the density and count attributes of pedestrians and adjusts the NMS based on these two attributes for more accurate pedestrian predictions in crowded scenes.

The subsequent sections of the paper are organized as follows. Section 2 provides an overview of related works. Section 3 describes the fundamental flow of NMS and our method. Section 4 introduces the details and analysis of the experiments. Section 5 discusses the results, theoretical support, application scenarios, and future work. Section 6 summarizes the entire paper and possible future work.

2. Related Works

In this section, we briefly describe related works, covering pedestrian detection, intra-class occlusion handling and NMS in pedestrian detection.

2.1. Pedestrian Detection

Depending on the method of feature extraction, pedestrian detectors can be categorized into the following types: handcrafted feature-based and deep feature-based pedestrian detectors. In early studies, handcrafted feature-based pedestrian detectors used sliding windows to extract features, such as HOG [24], LBP [25], SIFT [26], and Haar [27]. Classifiers such as SVM, AdaBoost, and random forest were trained to filter out the background. Integral channel features [28] combine LUV channels, gradient magnitude, and gradient histograms to effectively capture diverse information from input images. In addition, a deformable part model (DPM) was used in [10] to handle object deformation by partitioning the human body into different parts.

Recently, a convolutional neural network [29] was proposed. Since then, numerous deep feature-based pedestrian detectors have been designed, which have gradually replaced handcrafted feature-based pedestrian detectors. Among these, the well-known object detector Faster R-CNN [8] is widely utilized for various computer vision tasks, and has become one of the most common pedestrian detectors after fine-tuning. In [30], a scale-aware weight mechanism is proposed to detect large and small pedestrians separately, and then, assign weights to them dynamically. CSP [13] treats pedestrian detection as a high-level semantic feature [31,32] detection task by predicting the centroids and scales of pedestrians. Following this idea, PP-Net [12] proposes a depth-guided module to capture higher-level information. CSANet [33] uses channel attention and spatial attention together to model context [34–36] dependencies and enhance pedestrian features. MAPD [19] optimizes a positive sample allocation strategy and utilizes a triplet loss function to learn the high-level ID features of pedestrians, which are then combined with the NMS algorithm. While mainstream detectors can meet the requirements of pedestrian detection in simple scenes, such as sparse scenes, their performance tends to degrade in crowded scenes due to ubiquitous intra-class occlusion.

2.2. Intra-Class Occlusion Handling

Intra-class occlusion is a significant challenge in pedestrian detection, especially in crowded scenes, and profoundly impacts the performance of pedestrian detectors. To address this problem, MGAN [37] utilizes masks to extract pedestrian features and guides the detector to focus on the visible parts of pedestrians. In [38], three attention networks were designed for the whole body, the visible parts and the body parts, to distinguish between different pedestrians. Furthermore, a joint attention mechanism [18] was proposed to extract more robust features of pedestrians using channel, spatial, and self-attention mechanisms in the lateral connections of an FPN. In addition, NMS-loss was proposed in [39], which pulls proposals belonging to the same pedestrian closer together while pushing those belonging to different pedestrians away from each other. These methods aim to detect more pedestrians, especially occluded ones, highlighting the significance of post-processing algorithms. Consequently, several studies have attempted to improve

NMS algorithms to accurately retain pedestrians, especially those occluded by others, and remove numerous duplicate proposals generated by pedestrian detectors.

2.3. NMS

The NMS algorithm, initially known as Greedy NMS, is widely utilized as a post-processing method in object detection. It uses a rigid threshold to divide a uniform suppression interval for each pedestrian and directly discards all the other proposals within the interval to remove duplicates. However, the rigid threshold and discarding operation utilized in Greedy NMS tend to incorrectly remove the proposals belonging to occluded pedestrians in crowded scenes. In order to handle this problem, Soft NMS [16] uses a re-scoring strategy that replaces the discarding operation used in Greedy NMS. Furthermore, Adaptive NMS [17] uses a dynamic threshold to replace the rigid threshold in the existing NMS algorithms, and demonstrates improved performance, especially combined with Soft NMS. Considering the sensitivity of hyperparameters in Soft NMS, a normalized suppression function [39] was proposed to enhance the robustness of NMS. In addition, MAPD [19] employs a triplet loss function to learn the high-level ID features of pedestrians and uses a new segmented NMS algorithm. In [40], each pedestrian was predicted to have a nearby object, and suppression was conducted based on the proximity between the current bounding box and the nearby object. Alternatively, a pedestrian detector [41] predicts two proposals for each pedestrian and conducts NMS within each branch. MB-CSP [42] predicts the upper, middle, and lower parts of each pedestrian, and applies NMS based on occlusion modes. Although the above-mentioned methods employ different approaches that attempt to retain occluded pedestrians, they still encounter challenges where partial false positives are mistakenly retained while highly occluded pedestrians are still removed.

3. The Proposed Method

In this section, to facilitate the understanding of our approach, we first review the flow of the NMS algorithm by revisiting Greedy NMS. Next, we provide a detailed introduction to the Multi-Attribute NMS (MA-NMS), the attribute branch (ATTB), and the pedestrian detector for crowded scenes, which are proposed in this paper. Finally, we describe the methods of obtaining annotations of density and count attributes with existing annotations.

3.1. Greedy NMS

Greedy NMS, as the initial Non-Maximum Suppression algorithm, plays a critical role in removing duplicate proposals generated by pedestrian detectors. Each proposal is represented by a pair of confidence scores and a bounding box. The algorithm follows the steps outlined below:

1. Sort all the bounding boxes in set B in descending order based on their confidence scores.
2. Calculate the intersection-over-union (IoU) of the first bounding box M, which has the highest confidence score, and the sequenced bounding boxes b_i. If $IoU(M, b_i)$ exceeds the rigid threshold N_t, the confidence score of b_i will be set to zero.
3. Move the proposal m, with bounding box M, into the set F, which is initialized with an empty set.
4. Repeat the above three steps for the remaining bounding boxes in B until complete traversal.

The set F represents the final predictions of the pedestrian detector. The above process can be expressed as the following re-scoring function:

$$s_i = \begin{cases} s_i, IoU(M, b_i) < N_t, \\ 0, IoU(M, b_i) \geq N_t, \end{cases} \tag{1}$$

where s_i and b_i represent the confidence score and bounding box of the ith proposal, and N_t is a constant rigid threshold that ranges between 0 and 1.

Despite several studies [14–16] aiming to decrease the incorrect removal of occluded pedestrians by adjusting the single threshold or weakening the suppression operation in

Greedy NMS, these approaches tend to generate more false positives while still mistakenly removing highly occluded pedestrians, as shown in Figure 1. Therefore, striking a balance between removing false positives and retaining occluded pedestrians in NMS remains a challenging task.

3.2. Multi-Attribute NMS

Based on the analysis above, we propose a Multi-Attribute NMS (MA-NMS) algorithm that adaptively adjusts suppression based on density and count attributes. MA-NMS takes into account that the proposals whose bounding boxes have moderate overlap with their neighboring bounding boxes may still represent potentially occluded pedestrians. Therefore, instead of using absolute operations like discarding or retaining, MA-NMS applies suppression with weak intensity to provide protection. Specifically, for each pedestrian, MA-NMS refines the traditional single suppression interval into strong and weak suppression intervals to adaptively handle duplicate proposals and proposals belonging to occluded pedestrians. The division between the strong suppression interval (SSI) and weak suppression interval (WSI) is determined as follows:

$$d_i := \underset{b_j \in G, i \neq j}{max} IoU(b_i, b_j), \tag{2}$$

$$N_m = max(d_i, N_t), \tag{3}$$

$$SSI = [N_m, 1], \tag{4}$$

$$WSI = [N_t, N_m), \tag{5}$$

where N_t represents the rigid threshold used in Greedy NMS. The dynamic threshold N_m is consistent with Adaptive NMS, and is calculated using Equation (3).

Within the SSI, numerous duplicate proposals that closely overlap with M are present, while only a few pedestrians occluded by M may be included. Therefore, strong suppression is applied within the interval to effectively remove duplicate proposals. In order to ensure the fair treatment of occluded pedestrians with different densities, the strong suppression weight function f_s incorporates the parameter N_m, as shown in Equation (6). Within the WSI, there are a number of occluded pedestrians, along with a few duplicate proposals that are distant from the annotations. Hence, weak suppression is applied to preserve the potentially occluded pedestrians, while mildly penalizing duplicate ones as a complement to the strong suppression. The weak suppression weight function f_w is expressed in Equation (7). By utilizing both the strong and weak suppression, the re-scoring strategy of MA-NMS is described in Equation (8).

$$f_s(M, b_i, d_M) = \left(\frac{iou(M, b_i)}{N_m} \right)^{C_1}, \tag{6}$$

$$f_w(M, b_i, d_M) = 1 - (IoU(M, b_i) - N_t), \tag{7}$$

$$s_i = \begin{cases} s_i, IoU(M, b_i) < N_t, \\ s_i f_w(M, b_i, d_M), N_t \leq IoU(M, b_i) < N_m, \\ s_i f_s(M, b_i, d_M), IoU(M, b_i) \geq N_m, \end{cases} \tag{8}$$

where C_1 is a constant, and the experiments show that the detector achieves optimal performance when a value of 4 is used.

Additionally, we note that the count of surrounding occluded pedestrians varies greatly between pedestrians, particularly for pedestrians located at the edge of the crowd and those in the center. For a given pedestrian, if all the occluded pedestrians surrounding

them have been retained, the remaining proposals within the occluded pedestrian's strong and weak suppression intervals would only be duplicates. In order to further remove duplicate proposals, MA-NMS takes into account the count attribute. Based on the count attribute, the suppression intensity within each interval is adjusted using the suppression factors. By considering both density and count attributes, MA-NMS can be described using Algorithm 1. The re-scoring function of MA-NMS is expressed in Equation (9).

$$s_i = \begin{cases} s_i, IoU(M, b_i) < \mathrm{N}_t, \\ s_i f_w(M, b_i, d_M, c_i), N_t \leq IoU(M, b_i) < N_m, \\ s_i f_s(M, b_i, d_M, c_i), IoU(M, b_i) \geq N_m, \end{cases} \tag{9}$$

$$f_w(M, b_i, d_M, c_i) = (f_w(M, b_i, d_M))^x, \tag{10}$$

$$f_s(M, b_i, d_M, c_i) = (f_s(M, b_i, d_M))^y, \tag{11}$$

where c_i is the count attribute of the ith pedestrian and represents the count of surrounding pedestrians occluded by them. Specifically, x and y are suppression factors, which are initially set to 1. If the surrounding occluded pedestrians of a pedestrian have been retained, the suppression factors x and y are assigned the respective constants C_2 and C_3 for further suppression. The experiments in the subsequent section show that values of 3 and 4 for C_2 and C_3 obtain the optimal performance.

Algorithm 1: The procedure of Multi-Attribute NMS.

Input: $B = b_1, \ldots, b_n, S = s_1, \ldots, s_n$,
$D = d_1, \ldots, d_n, C = c_1, \ldots, c_n, N_t$
B is the list of initial bounding boxes;
S is the list of corresponding confidence scores;
D is the list of corresponding density attributes;
C is the list of corresponding count attributes;
N_t is the rigid NMS threshold.
Output: F

```
begin:
  F ← ∅
  While B ≠ ∅ do
    m = argmaxS
    M = b_m
    N_m = max(N_t, d_m)
    F ← F ∪ (s_m, M); S ← S − s_m; B ← B − M
    for b_i in B do
      if IoU(M, b_i) ≥ N_m then
        if c_i ≥ 0 then
          s_i = s_i × f_s(M, b_i, d_M)
        else
          s_i = s_i × (f_s(M, b_i, d_M))^C3
        c_i = c_i − 1
      else if IoU(M, b_i) ≥ N_t then
        if c_i ≥ 0 then
          s_i = s_i × f_w(M, b_i, d_M)
        else
          s_i = s_i × (f_w(M, b_i, d_M))^C2
        c_i = c_i − 1
      end for
    end while
return F
end
```

3.3. Attribute Branch

In addition to category and location information, abundant attribute information is contained in the feature map generated by the backbone network. In contrast to [17], which solely focuses on learning the density attribute, an attribute branch (ATTB) is proposed to simultaneously extract both the density and count attributes of pedestrians, guiding the adjustment of the suppression intervals and intensity in MA-NMS. The structure of ATTB is illustrated in the orange box in Figure 3.

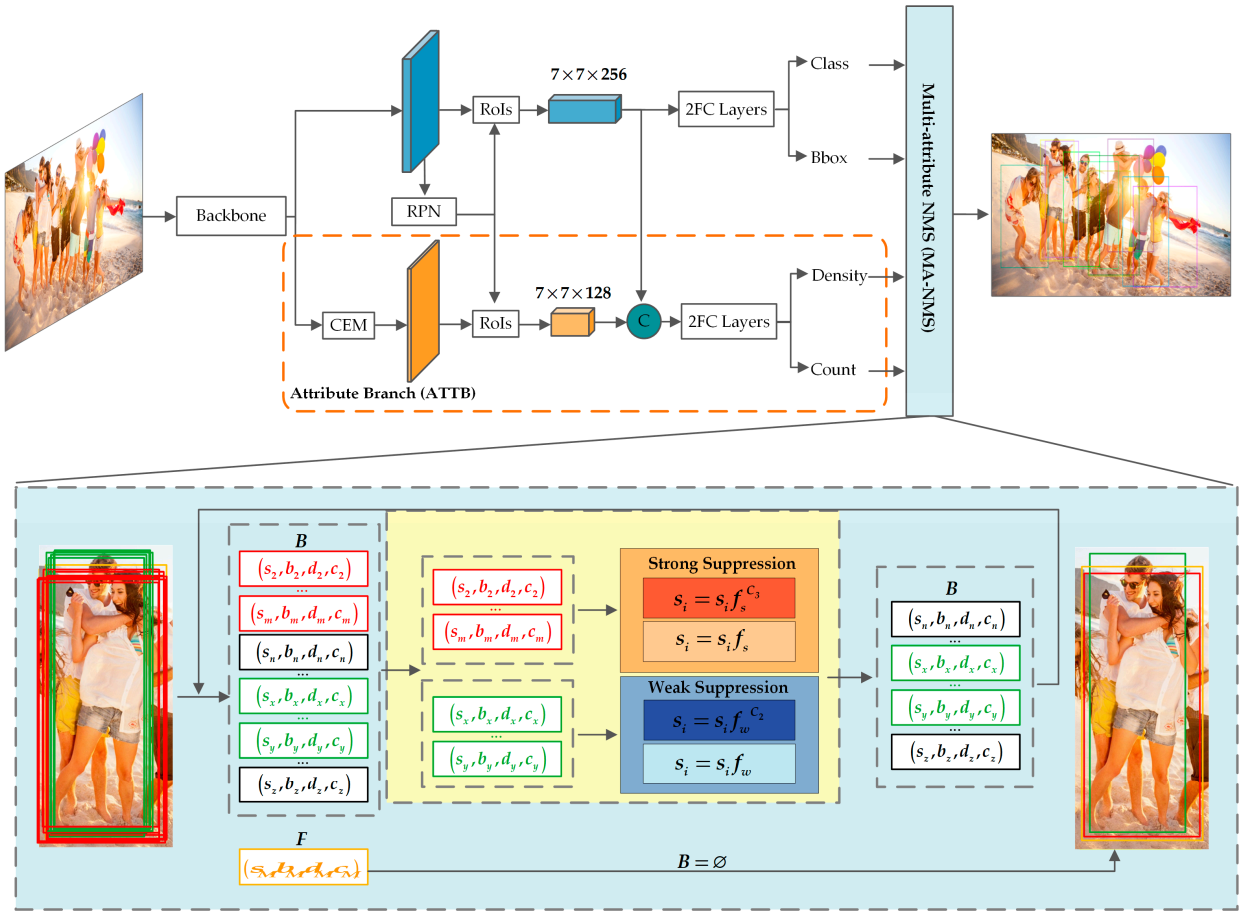


Figure 3. The structure of the pedestrian detector for crowded scenes. The context extraction module (CEM) consists of a 5 × 5 dilated convolution and a 1 × 1 standard convolution in series. The detector generates a tuple of attributes for each pedestrian, including confidence score (s), bounding box (b), density (d), and count (c). To remove duplicate proposals (highlighted in red), MA-NMS applies strong suppression using a score threshold of 0.01. Meanwhile, weak suppression is employed to retain the occluded pedestrians (highlighted in green). Color printing is recommended.

In the attribute branch (ATTB), a context extraction module (CEM), consisting of a 5 × 5 dilated convolution followed by a 1 × 1 standard convolution, is designed. The CEM is applied to capture the contextual information of pedestrians. The extracted context is then fed into the Region Proposal Network (RPN) to generate regions of interest (RoIs). These RoIs are subsequently concatenated with that from the detection branch, enabling a more comprehensive representation of pedestrians and their surrounding occluded pedestrians. Moreover, the concatenated ROIs are fed into two fully connected layers (FC layers) to obtain the density and count attributes. Notably, the proposed ATTB can be easily embedded into generic pedestrian detectors to acquire the density and count attributes, which, in turn, guide the adjustment of the NMS. Consequently, the incorporation of ATTB enhances the capabilities of NMS-based detectors in mitigating the impact of intra-class occlusion.

3.4. Pedestrian Detector for Crowded Scenes

Incorporating the proposed MA-NMS and ATTB, a pedestrian detector for crowded scenes is constructed on the basis of Faster R-CNN, as shown in Figure 3. Notably, the detection branch of the detector remains unchanged. Subsequently, the detector is trained on two datasets separately, and extensive experiments are conducted to verify the effectiveness of our proposed method.

During the training phase, the constructed detector optimizes the entire network using a weighted loss function L, which is expressed in Equation (12).

$$L = \lambda_1 L_{rpn\_cls} + \lambda_2 L_{rpn\_reg} + \lambda_3 L_{box\_cls} + \lambda_4 L_{box\_reg} + \lambda_5 L_{density} + \lambda_6 L_{count}, \quad (12)$$

where λ_i is employed to balance the gradient magnitude of the corresponding loss. L_{rpn_cls} and L_{box_cls} are computed using cross-entropy loss, and the L_{rpn_reg} and L_{box_reg} are calculated using smooth L_1 loss, which is in line with [8]. Furthermore, the detector treats the detection of density and count attributes as regression tasks, employing smooth L_1 loss to calculate $L_{density}$ and L_{count}.

During the inference phase, MA-NMS is utilized to suppress the numerous duplicate proposals based on the count and density attributes. As illustrated in Figure 3, proposals with bounding boxes that overlap with that of the current pedestrian will be removed by strong suppression. Additionally, proposals belonging to occluded pedestrians are retained by weak suppression. Eventually, our pedestrian detector obtains more accurate predictions for pedestrian detection in crowded scenes.

3.5. Ground Truth for Pedestrian Density and Count Attributes

The constructed pedestrian detector additionally detects the density and count attributes of pedestrians, which represent the maximum occlusion degree and the count of surrounding occluded pedestrians for each pedestrian. Unfortunately, no specific annotations for the density and count attributes are available in the benchmark datasets for pedestrian detection, and it is expensive to manually annotate a public benchmark dataset. In order to address this challenge, recent studies [13,42] have proposed methods to generate approximate annotations based on existing annotations. Inspired by these methods, we generate annotations of the density and count attributes using the existing full-body annotations.

Consistent with [17], we calculate the density attribute of each pedestrian by considering the maximum IoU value between their ground truth and that of the others, which is expressed in Equation (2). Additionally, inspired by Greedy NMS, we account for the occlusion between two pedestrians when their ground truths exhibit an IoU value above the rigid threshold N_t. For each pedestrian, the count attribute is determined by summing the number of pedestrians occluded by them. The quantization process is expressed in Equation (13).

$$c_i := \sum_{b_j \in G, i \neq j} h(iou(M, b_i) - N_t), \quad (13)$$

$$h(x) = \begin{cases} 0, x < 0, \\ 1, x \geq 0, \end{cases} \quad (14)$$

where c_i is the count attribute of the ith pedestrian and represents the count of surrounding pedestrians occluded by them. G denotes the ground truths and b_i represents the full-body annotation of the ith pedestrian.

4. Experiments

In this section, we first introduce the datasets and evaluation metrics. Subsequently, we provide a detailed description of the experimental setup, followed by a comprehensive analysis of our proposed method.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

CrowdHuman [22], a challenging pedestrian benchmark, is widely used to evaluate the performance of pedestrian detectors in crowded scenes. In this paper, we choose CrowdHuman to evaluate the performance of our proposed method in crowded scenes. CrowdHuman consists of images from over 40 different cities worldwide, which are crawled by Google. These images depict dense and diverse crowds, posing significant challenges for pedestrian detection. CrowdHuman comprises a total of 24,370 images, with 15,000 images allocated for training, 4370 images for validation, and 5000 images for testing.

CityPersons [23], a subset of Cityscapes [43], is a commonly used benchmark dataset for pedestrian detection. It is selected in the paper to evaluate the performance of MA-NMS in slightly crowded scenes. CityPersons exhibits a high level of diversity and depicts numerous cities and countries in Europe. CityPersons comprises 5000 images, with 2975 images allocated for training, 500 images for validation, and 1525 images for testing. Table 1 presents the statistics of the training sets for CrowdHuman and CityPersons.

Table 1. Statistics of CrowdHuman and CityPersons training sets.

Objects	CrowdHuman	CityPersons
Images	15,000	2975
Persons	339,565	19,238
Ignore regions	99,227	6768
Person/image	22.64	6.47
Unique persons	339,565	19,238

4.1.2. Evaluation Metrics

On CrowdHuman, we report the performance of our proposed method using various metrics, including average accuracy (AP), $Recall$, MR^{-2}, and FPS. On CityPersons, we evaluate the performance of our detector using MR^{-2} on four subsets: reasonable (R), bare (B), partial (P), and heavy occlusion (H). These subsets are divided based on the visibilities of pedestrians, as shown in Table 2. More details of the evaluation metrics are given below

- AP: Average precision, which summarizes a precision–recall curve of detection results, is one of the most popular evaluation metrics in generic object detection. In the subsequent experiments, we follow the AP metric in PASCAL VOC [44] (the larger, the better) and consider proposals with $IoU \geq 0.5$ to be positive. This metric effectively measures the accuracy of a detector.
- $Recall$: The maximum recall, for a fixed number of proposals, represents the proportion of true positives detected by a detector out of the total number ground truths. This metric evaluates the ability of a detector to accurately detect the true ground truths. Larger values indicate better performance.
- MR^{-2}: Log-average miss rate, which is calculated using false positives per image (FPPI) in the range of $\left[10^{-2}, 10^{0}\right]$, is a commonly used evaluation metric in pedestrian detection. This metric is particularly sensitive to false positives, especially those with high confidence scores. Smaller values of MR^{-2} indicate better performance of a pedestrian detector.
- FPS: Frames per second, which represents the number of frames processed per second, is a commonly used metric for measuring the speed of detectors. Larger values of FPS indicate faster processing speed of a detector.

Table 2. Subsets of CityPersons dataset based on visibility.

Subsets	Visibility
Reasonable (R)	$[0.65, 1]$
Bare (B)	$[0.9, 1.0]$
Partial (P)	$[0.65, 0.9]$
Heavy (H)	$[0.2, 0.65]$

4.2. Implementation Details

In the subsequent experiments, we choose Faster R-CNN with ResNet-50 and a Feature Pyramid Network (FPN) as the baseline, in which Greedy NMS is used for post-processing. In order to obtain more accurate features, RoI Align [45] is employed. Given the varying shapes of images in CrowdHuman, the input images are uniformly resized to 800 pixels on the shorter side and kept below 1400 pixels on the longer side. The aspect ratios of the anchors are resized to $\frac{H}{W} = \{1, 2, 3\}$. Following [8,23], the height and width of the images in CityPesons are enlarged by a factor of 1.3, and the aspect ratios of the anchors are adjusted to $\frac{H}{W} = \{2.44\}$ to accommodate pedestrian scale.

During the training phrase, ResNet-50 is pretrained on ImageNet [46]. The remaining parameters of our detector are initialized using Kaiming initialization [47]. Our detector is trained on 2 NVIDIA Ampere A40 GPUs, with a minimum batch size of 16 images. We employ Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 0.0001 as the optimizer. For CrowdHuman, our detector is trained for 39,979 iterations, while for the CityPersons dataset, our detector is trained for 5580 iterations. In the case of CrowdHuman, the initial learning rate is set to 0.04, and it is decreased by a factor of 10 after 18,750 and 28,125 iterations. For CityPersons, the initial learning rate is set to 0.02, and it is decreased by a factor of 10 after 3720 and 4650 iterations.

During the inference phase, each image is subjected to a maximum of 100 detections. The same resizing operation, as mentioned during training, is applied. To ensure a fair comparison with other NMS methods, the value of the rigid threshold N_t is set to 0.5, unless stated otherwise.

4.3. Ablation Study

To comprehensively evaluate the performance of the proposed method, detailed ablation experiments are conducted on CrowdHuman. These experiments involve the progressive application of strong suppression (SS), weak suppression (WS), and suppression factors (SF). Table 3 shows the results of the ablation study, with the best results shown in bold. This table clearly reflects the significant performance improvement of the pedestrian detector following the sequential application of strong suppression, weak suppression, and suppression factors in MA-NMS. Specifically, when strong and weak suppression are used, there is a notable improvement of 6.1% and 1.1% in terms of *Recall* and MR^{-2} compared to the baseline. This indicates that the adaptive application of strong and weak suppression contributes to the retention of more pedestrians, especially those occluded pedestrians that are incorrectly removed by the Greedy NMS used at the baseline, while effectively removing duplicate proposals. Moreover, the additional application of suppression factors results in further improvement of 1.1% and 1.7% for AP and MR^{-2}. This suggests that considering count attributes enables more precise suppression of duplicate proposals. Ultimately, with the complete application of MA-NMS, a substantial improvement of 4.2% and 6.5% is achieved for AP and *Recall*, strongly verifying the superiority of the proposed MA-NMS in enhancing the performance of generic pedestrian detectors in crowded scenes.

Table 3. Ablation experiments evaluated on CrowdHuman validation set. The baseline (the first line) corresponds to Faster R-CNN [8] with ResNet-50, FPN, and Greedy NMS. MA-NMS—Multi-Attribute NMS. SS—strong suppression. SW—weak suppression. SF—suppression factors.

Methods	SS	WS	SF	AP	Recall	MR^{-2}
Baseline				85.0	88.1	44.8
MA-NMS(ours)	√			88.6	92.7	45.1
	√	√		89.1	94.2	43.7
	√		√	89.3	93.5	42.5
	√	√	√	**90.2**	**94.6**	**42.0**

Better performance is indicated by the table's bold font. This also applies to subsequent tables.

4.4. Hyperparameters

4.4.1. Rigid Threshold

Our MA-NMS, along with well-known NMS algorithms, such as Greedy NMS, Soft NMS [16], and Adaptive NMS [17], incorporates the hyperparameter N_t as the rigid threshold. In order to achieve their optimal performances and sensitivities, different values of N_t are employed in the experiments. The evaluation metric AP is used to measure the average precision of the pedestrian detector. To ensure a fair comparison, all the NMS algorithms are implemented using the same programming language and separately applied to the baseline with the same settings. Figure 4 illustrates the results of NMS algorithms with varying values of N_t. Obviously, MA-NMS achieves the best performance across all values of N_t, and the optimal results for all the NMS algorithms in the experiments are obtained, when N_t is set to 0.5. Furthermore, the results shows that Greedy NMS remains consistent between 0.55 and 0.85, Adaptive NMS is stable between 0.80 and 0.87, while MA-NMS remains stable between 0.85 and 0.90, with only 16.67% and 71.42% of their AP values floating, which indicates that MA-NMS exhibits lower sensitivity to the hyperparameters N_t compared to these well-known NMS algorithms.

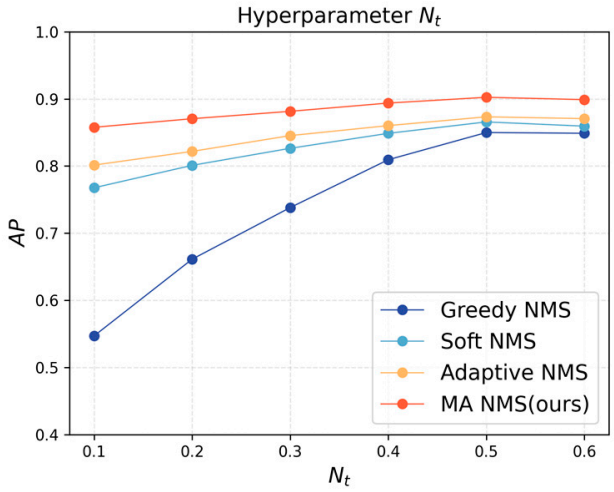


Figure 4. Results of NMS algorithms with varying values of rigid threshold N_t. The experiments are conducted on the CrowdHuman validation set. To ensure a fair comparison, these NMS algorithms are applied separately to the same baseline, with the same settings stated in Section 4.2. The higher the AP value, the better the result.

4.4.2. Exponential Constants

The strong suppression weight function of MA-NMS contains a hyperparameter C_1, which is used to enhance the suppression of duplicate proposals. Additionally, C_2 and C_3 are utilized as large suppression factors in strong and weak suppression to further remove duplicate proposals, respectively. Extensive experiments are conducted to evaluate the performance of MA-NMS using the control variables method. More concretely, when C_1 is

varied, C_2 and C_3 are set to 3 and 4. When C_2 is varied, C_1 and C_3 are set to 4. When C_3 is varied, C_1 and C_2 are set to 4 and 3. Figure 5 shows the results of MA-NMS with different values of these exponential constants. As depicted in Figure 5a, the best performance is obtained when C_1 is set to 4. Furthermore, it can be seen from Figure 5b,c that the optimal performance is obtained when C_2, C_3 are set to 3 and 4.

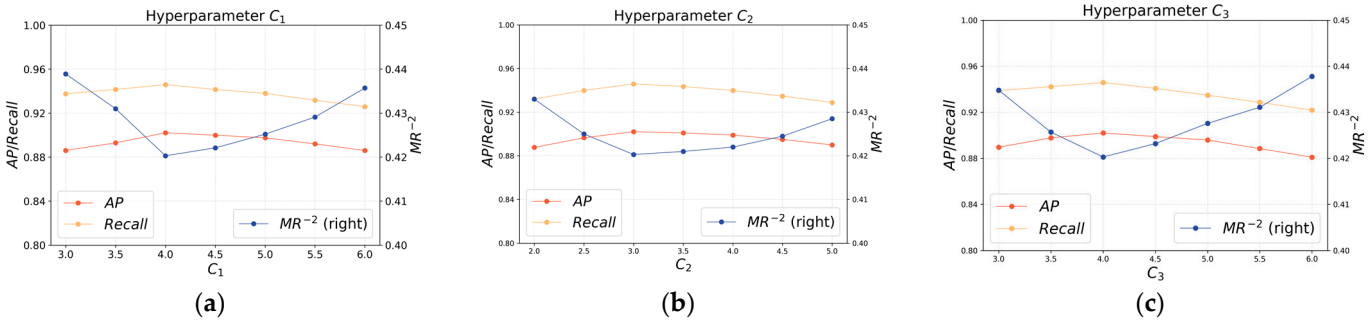


Figure 5. Results of MA-NMS with varying values of exponential constants. (**a**) Hyperparameter C_1; (**b**) hyperparameter C_2; (**c**) hyperparameter C_3. The higher the *AP* and *Recall*, the lower the MR^{-2}, and the better the result.

4.5. Speed

The inference speed is equally essential for NMS. In order to obtain a speed comparison, we apply the well-known NMS algorithms [16,17] to Faster R-CNN separately and conduct the experiments in the same environment. Table 4 shows the results of the accuracy and speed comparison. Despite exhibiting a slight decrease in speed, MA-NMS demonstrates significant improvements in other metrics, particularly a 3.8% *Recall* improvement and a 2.5% MR^{-2} improvement compared to Soft NMS [16]. The improvement in *Recall* implies that MA-NMS retains more pedestrians than other NMS algorithms, especially occluded pedestrians that are prone to being mistakenly removed, as no changes have been made to the generic pedestrian detection branch [8]. Additionally, the improved MR^{-2} value indicates that MA-NMS performs more potent suppression on duplicate proposals, resulting in fewer false positives. In summary, while the FPS of MA-NMS is slightly lower than that of the previously superior Greedy NMS, the decrease is only 3.9%. On the other hand, the accuracy is significantly increased. These findings imply that MA-NMS achieves a balance between accurate detection results and comparable speed for pedestrian detection in crowded scenes.

Table 4. Results of different NMS algorithms equipped on Faster R-CNN. MA-NMS—Multi-Attribute NMS.

Methods	N_t	AP	Recall	MR^{-2}	FPS
Greedy NMS	0.5	85.0	88.1	44.8	**10.75**
Soft NMS [16]	0.5	86.6	90.8	44.5	10.62
Adaptive NMS [17]	0.5	87.3	90.0	45.2	10.45
MA-NMS (ours)	0.5	**90.2**	**94.6**	**42.0**	10.33

4.6. Comparison

4.6.1. Results of CrowdHuman

The proposed detector is trained on the CrowdHuman training set, and then, compared with other state-of-the-art methods on the challenging crowded dataset. Table 4 shows the comparative results of the CrowdHuman verification set, where *w* (*w/o*) represents with (without) suppression factors, and the best results are shown in boldface. The results clearly show that our MA-NMS outperforms the other NMS methods [40,48] in terms of the *Recall* metric, even with only the application of strong and weak suppression. This is due to the fact that the adaptive application of strong and weak suppression allows for a

divide-and-conquer strategy that effectively suppresses duplicate proposals while retaining those of occluded pedestrians, which is challenging to achieve using the single uniform suppression operation employed in other NMS methods [40,48]. Furthermore, with the application of suppression factors, MA-NMS outperforms all the competitors, especially considering its 3.3% improvement in MR^{-2} compared to IDADA [49]. Despite IDADA [49] utilizing data augmentation to obtain high-quality proposals, more duplicate proposals are incorrectly retained by its Greedy NMS, making IDADA [49] more sensitive to false positives than MA-NMS, and a similar dilemma is also faced by method the method used in [50,51]. Moreover, the authors of [39,52] trained their models to more accurately generate bounding boxes to assist Greedy NMS in identifying duplicate proposals; however, their assistance is limited in crowded scenes due to ubiquitous intra-class occlusion, and they lack the adaptive suppression capability of MA-NMS, ultimately resulting in 3.6% and 6.2% lag in *Recall* compared to MA-NMS. Furthermore, MA-NMS gains 1.4% improvement in AP compared to JointDet [18]. This can be attributed to MA-NMS effectively leveraging attribute information in pedestrian features for suppression, which is absent in JointDet [18], despite its attempts to extract more robust pedestrian features. These results validate the effectiveness of MA-NMS in improving the performance of generic pedestrian detectors in crowded scenes, making it a competitive choice over other advanced models [18,49–51] in crowded scenes.

4.6.2. Results of CityPersons

We train the proposed detector on the training set of CityPersons, and compare its performance with the state-of-the-art detectors on the dataset. Table 5 shows the comparison results of the CityPersons validation set. It clearly demonstrates that MA-NMS achieves the best performance on the R, P, and B subsets, with the strong and weak suppression tokens. In particular, our detector shows a significant 3.5% MR^{-2} improvement compared to NOH NMS [40]. This improvement can be attributed to the adaptive application of strong and weak suppression in MA-NMS, which imposes a stronger penalty on duplicate proposals and better preserves numerous occluded pedestrians in the H subset. By additionally applying suppression factors, MA-NMS further enhances its performance on these four subsets, and outperforms CSP [13], with an 1.0% and 0.4% MR^{-2} improvement on the R and H subsets. This can be attributed to the adaptive adjustment of suppression intensity in MA-NMS, which is determined based on the density and count attributes of pedestrians, leading to further suppression of duplicate proposals, and reducing sensitivity to false positives. Moreover, MA-NMS gain a 5.5% and 2% MR^{-2} improvement compared to TLL [47] and ALFNet [48] due to the comprehensive consideration of pedestrian density and count attributes in MA-NMS, which enables more accurate suppression compared to methods that rely solely on bounding box and confidence score information. The results reported in Table 6 validate the superior performance of MA-NMS for pedestrian detection in slightly crowded scenes.

Table 5. Comparison with the state-of-the-art methods on CrowdHuman validation set. *w* (*w/o*) represents with (without) suppression factors.

Method	Backbone	AP	Recall	MR^{-2}
PBM + R2NMS [48]	ResNet-50	89.3	93.3	43.4
NOH-NMS [40]	ResNet-50	89.0	92.9	43.9
RepLoss [39]	ResNet-50	85.6	88.4	45.7
AutoPedestrian [50]	ResNet-50	87.7	93.0	46.9
LLA.FCOS [51]	ResNet-50	88.1	93.4	47.9
JointDet [18]	DarkNet-53	88.8	-	43.4
IDADA [49]	ResNet-50	88.0	93.6	45.3
CouLoss [52]	ResNet-50	89.8	91.0	42.4
MA-NMS (*w/o*)	ResNet-50	89.1	94.2	43.7
MA-NMS (*w*)	ResNet-50	**90.2**	**94.6**	**42.0**

Table 6. Comparison with the state-of-the-art methods on CityPersons validation set. w (w/o) represents with (without) suppression factors.

Methods	Backbone	R	H	P	B
RepLoss [39]	ResNet-50	13.2	56.9	16.8	7.6
TLL [53]	ResNet-50	15.5	53.6	17.2	10.0
TLL + MRF [53]	ResNet-50	14.4	52.0	15.9	9.2
ALFNet [54]	ResNet-50	12.0	51.9	11.4	8.4
PBM + R2NMS [48]	VGG16	11.1	53.3	-	-
NOH NMS [40]	ResNet-50	10.8	53.0	11.2	6.6
AutoPedestrian [50]	ResNet-50	11.5	56.7	-	-
CSP [13]	ResNet-50	11.0	49.4	10.4	7.3
MA-NMS (w/o)	ResNet-50	10.6	49.5	9.9	6.5
MA-NMS (w)	ResNet-50	**10.0**	**48.9**	**9.0**	**6.3**

4.7. Visualization

Ubiquitous intra-class occlusion [55] in crowded scenes poses a challenge for pedestrian detectors, as proposals belonging to occluded pedestrians share a high similarity with duplicate proposals, leading to inaccurate detection results. In order to provide a clear illustration, we visualize the detection results of our proposed detector on the CrowdHuman and CityPersons datasets. Figure 6 shows the visualization of our method on CrowdHuman. It clearly shows that crowded pedestrians captured from multiple angles are accurately detected by our proposed detector, even those who are occluded by others. This is owing to the adaptive suppression in MA-NMS, which effectively preserves occluded pedestrians while significantly suppressing duplicate proposals. Figure 7 presents additional visualization examples on CityPersons, where our detector achieves accurate pedestrian detection in slightly crowded scenes. The visualized results collectively demonstrate the superiority of our MA-NMS in improving the performance of generic pedestrian detectors in both crowded and slightly crowded scenes.

Figure 6. Visualization of our method on CrowdHuman.

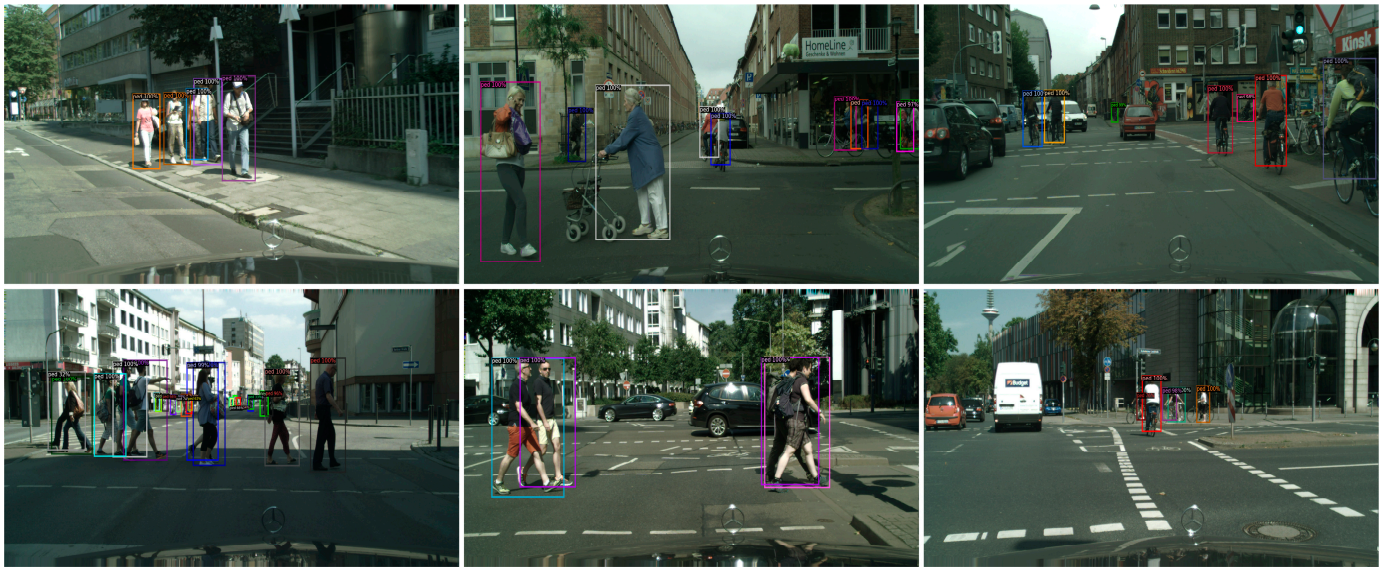


Figure 7. Visualization of our method on CityPersons.

5. Discussion

Dense crowds cause ubiquitous intra-class occlusion [55] between pedestrians, which is exacerbated by varying postures and different shooting angles. Consequently, dense and highly overlapped proposals are generated, posing a challenge for NMS algorithms to preserve the proposals of occluded pedestrians while removing duplicate ones. As a result, generic pedestrian detectors fail to meet the requirements of pedestrian detection in crowded scenes.

To address the above-mentioned problem, this paper proposes a Multi-Attribute NMS (MA-NMS), which enables adaptive suppression based on the density and count attributes of pedestrians. Additionally, an attribute branch (ATTB) is proposed to obtain the density and count attributes, which guides the adjustment of suppression intervals and intensity in MA-NMS. Furthermore, leveraging the proposed ATTB, a specialized detector based on MA-NMS is constructed for pedestrian detection in crowded scenes, which incorporates the density and count attributes of pedestrians and adjusts the NMS algorithm to enhance the accuracy of pedestrian detection in crowded scenes.

Extensive experiments are conducted on CrowdHuman and CityPersons benchmark datasets to evaluate the performance of our method. The experimental results show promising progress in crowded pedestrian detection, especially a noTable 5.2% and 6.5% improvement in *AP* and *Recall* compared to Greedy NMS on CrowdHuman. These findings strongly verify the advantages of the proposed MA-NMS in improving the performance of generic pedestrian detectors in crowded scenes.

The promising progress observed in these experiments are interpretable from the perspective of previous studies and hypotheses. Recent studies [17,18] have emphasized the importance of an appropriate suppression interval for more accurate predictions in crowded scenes. However, due to the presence of ubiquitous intra-class occlusion, determining a precise suppression interval for each pedestrian is challenging. In order to address this problem, MA-NMS introduces strong and weak suppression intervals, providing an approximation of the precise suppression interval. Additionally, inspired by Soft NMS [16], MA-NMS proposes strong and weak suppression weight functions for re-scoring in the corresponding interval. This adaptive approach enhances the adaptability of MA-NMS to complex crowded scenes, enabling more accurate predictions for pedestrian detection in crowded scenes. Moreover, inspired by [20,21], we propose an attribute branch (ATTB) to obtain the density and count attributes for post-refining the NMS-based pedestrian detectors. The ATTB guides the adjustment of suppression intervals and intensity of MA-NMS, benefiting the accuracy of predictions from an overall perspective.

MA-NMS can be easily integrated into generic pedestrian detectors based on the ATTB. By mitigating the influence of intra-class occlusion on pedestrian detectors, MA-NMS facilitates more accurate predictions for pedestrian detection in crowded scenes, in turn, having a positive impact on subsequent tasks, such as face recognition [5], pedestrian re-identification [6], and human interaction [7].

In the future, we will utilize MA-NMS to deal with other problems in pedestrian detection, such as inter-class occlusion and inter-class confusion.

6. Conclusions

With the increasing demand for pedestrian detection in crowded scenes, traditional pedestrian detectors face challenges due to their inability to handle ubiquitous intra-class occlusion, for which NMS struggles to accurately differentiate between the proposals of occluded pedestrians and duplicate proposals. To address the above-mentioned problem, we propose a Multi-Attribute NMS (MA-NMS) that adaptively adjusts suppression based on density and count attributes. MA-NMS strengthens suppression on duplicate proposals while weakening that for potentially occluded pedestrians, thus leading to more accurate predictions in crowded scenes. In addition, an attribute branch (ATTB) is designed to obtain the density and count attributes and simultaneously guides the adjustment of MA-NMS. ATTB uses a context extraction module (CEM) to extract the context of pedestrians, and then, concentrates the context with the feature of pedestrians for accurate attribute detection. Moreover, utilizing the proposed ATTB, a pedestrian detector based on MA-NMS is constructed, which enables more accurate predictions in crowded scenes. Extensive experiments on two challenging benchmarks are conducted, and the results demonstrate the superiority of our method compared to existing approaches, sufficiently validating the effectiveness of our model for pedestrian detection in crowded scenes. In the future, we plan to extend the application of MA-NMS to address inter-class occlusion and inter-class confusion in pedestrian detection. Additionally, a comprehensive evaluation of the sensitivity of the proposed MA-NMS is planned, and the corresponding solution is expected to be proposed in future work. Moreover, we will further attempt to use a lighter object detector to replace the two-stage Faster R-CNN, and design lightweight modules to improve the speed of the proposed MA-NMS algorithm.

Author Contributions: Conceptualization, W.W. and X.L. (Xin Lyu); methodology, W.W. and X.L. (Xin Li); software, W.W.; validation, W.W. and X.L. (Xin Lyu); formal analysis, W.W.; investigation, W.W. and X.L. (Xin Li); resources, W.W., T.Z. and X.L. (Xin Li); data curation, W.W., T.Z., J.C. and S.C.; writing—original draft preparation, W.W.; writing—review and editing, W.W., T.Z. and X.L. (Xin Lyu); visualization, W.W., J.C. and X.L. (Xin Lyu); supervision, X.L. (Xin Lyu) and X.L. (Xin Li); project administration, X.L. (Xin Lyu) and X.L. (Xin Li); funding acquisition, X.L. (Xin Lyu) and X.L. (Xin Li). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Excellent Post-doctoral Program of Jiangsu Province (Grant No. 2022ZB166), the Fundamental Research Funds for the Central Universities (Grant No. B230201007), the Project of Water Science and Technology of Jiangsu Province (Grant No. 2021080 and 2021063), the National Natural Science Foundation of China (Grant No. 42104033, 42101343, and 82004498), the Joint Fund of the Ministry of Education for Equipment Pre-research (Grant No. 8091B022123), the Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory (Grant No. 2021JCJQ-SYSJJ-LB06905), and the Qinglan Project of Jiangsu Province.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available on CrowdHuman at https://doi.org/10.48550/arXiv.1805.00123, reference [22], and CityPersons at https://doi.org/10.1109/CVPR.2017.474, reference [23].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cao, J.; Pang, Y.; Xie, J.; Khan, F.S.; Shao, L. From Handcrafted to Deep Features for Pedestrian Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4913–4934. [CrossRef]
2. Claussmann, L.; Revilloud, M.; Gruyer, D.; Glaser, S. A Review of Motion Planning for Highway Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1826–1848. [CrossRef]
3. Sikandar, T.; Ghazali, K.H.; Rabbi, M.F. ATM Crime Detection Using Image Processing Integrated Video Surveillance: A Systematic Review. *Multimed. Syst.* **2019**, *25*, 229–251. [CrossRef]
4. Lee, I. Service Robots: A Systematic Literature Review. *Electronics* **2021**, *10*, 2658. [CrossRef]
5. Sepas-Moghaddam, A.; Pereira, F.M.; Correia, P.L. Face Recognition: A Novel Multi-Level Taxonomy Based Survey. *IET Biom.* **2020**, *9*, 58–67. [CrossRef]
6. Wu, D.; Huang, H.; Zhao, Q.; Zhang, S.; Qi, J.; Hu, J. Overview of Deep Learning Based Pedestrian Attribute Recognition and Re-Identification. *Heliyon* **2022**, *8*, e12086. [CrossRef]
7. Harris, E.J.; Khoo, I.-H.; Demircan, E. A Survey of Human Gait-Based Artificial Intelligence Applications. *Front. Robot. AI* **2021**, *8*, 749274. [CrossRef]
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Neural Information Processing Systems (NIPS): La Jolla, CA, USA, 2015; Volume 28.
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 779–788.
10. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-Aware Fast R-CNN for Pedestrian Detection. *IEEE Trans. Multimed.* **2018**, *20*, 985–996. [CrossRef]
11. Liu, W.; Liao, S.; Ren, W.; Hu, W.; Yu, Y. High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 5182–5191.
12. Cai, J.; Lee, F.; Yang, S.; Lin, C.; Chen, H.; Kotani, K.; Chen, Q. Pedestrian as Points: An Improved Anchor-Free Method for Center-Based Pedestrian Detection. *IEEE Access* **2020**, *8*, 179666–179677. [CrossRef]
13. Liu, W.; Hasan, I.; Liao, S. Center and Scale Prediction: Anchor-Free Approach for Pedestrian and Face Detection. *Pattern Recognit.* **2023**, *135*, 109071. [CrossRef]
14. Li, X.; Xu, F.; Lyu, X.; Gao, H.; Tong, Y.; Cai, S.; Li, S.; Liu, D. Dual Attention Deep Fusion Semantic Segmentation Networks of Large-Scale Satellite Remote-Sensing Images. *Int. J. Remote Sens.* **2021**, *42*, 3583–3610. [CrossRef]
15. Li, X.; Li, T.; Chen, Z.; Zhang, K.; Xia, R. Attentively Learning Edge Distributions for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 102. [CrossRef]
16. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5562–5570.
17. Liu, S.; Huang, D.; Wang, Y. Adaptive NMS: Refining Pedestrian Detection in a Crowd. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6452–6461.
18. Ma, W.; Zhou, T.; Qin, J.; Zhou, Q.; Cai, Z. Joint-Attention Feature Fusion Network and Dual-Adaptive NMS for Object Detection. *Knowl. Based Syst.* **2022**, *241*, 108213. [CrossRef]
19. Wang, Y.; Han, C.; Yao, G.; Zhou, W. MAPD: An Improved Multi-Attribute Pedestrian Detection in a Crowd. *Neurocomputing* **2021**, *432*, 101–110. [CrossRef]
20. Zhang, J.; Lin, L.; Zhu, J.; Li, Y.; Chen, Y.; Hu, Y.; Hoi, S.C.H. Attribute-Aware Pedestrian Detection in a Crowd. *IEEE Trans. Multimed.* **2021**, *23*, 3085–3097. [CrossRef]
21. Zhang, H.; Yan, C.; Li, X.; Yang, Y.; Yuan, D. MSAGNet: Multi-Stream Attribute-Guided Network for Occluded Pedestrian Detection. *IEEE Signal Process. Lett.* **2022**, *29*, 2163–2167. [CrossRef]
22. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv* **2018**, arXiv:1805.00123.
23. Zhang, S.; Benenson, R.; Schiele, B. CityPersons: A Diverse Dataset for Pedestrian Detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 4457–4465.
24. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
25. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
26. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.

27. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
28. Zhang, H.; Zhao, L. Integral Channel Features for Particle Filter Based Object Tracking. In Proceedings of the 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 26–27 August 2013; Volume 2, pp. 190–193.
29. Bouwmans, T.; Jayed, S.; Sultana, M.; Jung, S.K. Deep Neural Network Concepts for Background Subtraction: A Systematic Review and Comparative Evaluation. *Neural Netw.* **2019**, *117*, 8–66. [CrossRef]
30. Wang, X. Intelligent Multi-Camera Video Surveillance: A Review. *Pattern Recognit. Lett.* **2013**, *34*, 3–19. [CrossRef]
31. Li, X.; Xu, F.; Liu, F.; Xia, R.; Tong, Y.; Li, L.; Xu, Z.; Lyu, X. Hybridizing Euclidean and Hyperbolic Similarities for Attentively Refining Representations in Semantic Segmentation of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5003605. [CrossRef]
32. Li, X.; Xu, F.; Liu, F.; Lyu, X.; Tong, Y.; Xu, Z.; Zhou, J. A Synergistical Attention Model for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5400916. [CrossRef]
33. Zhang, Y.; Yi, P.; Zhou, D.; Yang, X.; Yang, D.; Zhang, Q.; Wei, X. CSANet: Channel and Spatial Mixed Attention CNN for Pedestrian Detection. *IEEE Access* **2020**, *8*, 76243–76252. [CrossRef]
34. Liu, Z.; Song, X.; Feng, Z.; Xu, T.; Wu, X.; Kittler, J. Global Context-Aware Feature Extraction and Visible Feature Enhancement for Occlusion-Invariant Pedestrian Detection in Crowded Scenes. *Neural Process. Lett.* **2022**, *55*, 803–817. [CrossRef]
35. Li, X.; Xu, F.; Xia, R.; Lyu, X.; Gao, H.; Tong, Y. Hybridizing Cross-Level Contextual and Attentive Representations for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2021**, *13*, 2986. [CrossRef]
36. Li, X.; Xu, F.; Xia, R.; Li, T.; Chen, Z.; Wang, X.; Xu, Z.; Lyu, X. Encoding Contextual Information by Interlacing Transformer and Convolution for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2022**, *14*, 4065. [CrossRef]
37. Xie, J.; Pang, Y.; Khan, M.H.; Anwer, R.M.; Khan, F.S.; Shao, L. Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection. *IEEE Trans. Image Process.* **2021**, *30*, 3872–3884. [CrossRef]
38. Zhang, S.; Chen, D.; Yang, J.; Schiele, B. Guided Attention in CNNs for Occluded Pedestrian Detection and Re-Identification. *Int. J. Comput. Vis.* **2021**, *129*, 1875–1892. [CrossRef]
39. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion Loss: Detecting Pedestrians in a Crowd. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7774–7783.
40. Zhou, P.; Zhou, C.; Peng, P.; Du, J.; Sun, X.; Guo, X.; Huang, F. NOH-NMS: Improving Pedestrian Detection by Nearby Objects Hallucination. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1967–1975.
41. Chu, X.; Zheng, A.; Zhang, X.; Sun, J. Detection in Crowded Scenes: One Proposal, Multiple Predictions. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12211–12220.
42. Abdelmutalab, A.; Wang, C. Pedestrian Detection Using MB-CSP Model and Boosted Identity Aware Non-Maximum Suppression. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 24454–24463. [CrossRef]
43. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 3213–3223.
44. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
45. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
46. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: New York, NY, USA, 2015; pp. 1026–1034.
48. Huang, X.; Ge, Z.; Jie, Z.; Yoshie, O. NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10747–10756.
49. Zhou, S.; Tang, Y.; Liu, M.; Wang, Y.; Wen, H. Impartial Differentiable Automatic Data Augmentation Based on Finite Difference Approximation for Pedestrian Detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2510611. [CrossRef]
50. Tang, Y.; Li, B.; Liu, M.; Chen, B.; Wang, Y.; Ouyang, W. AutoPedestrian: An Automatic Data Augmentation and Loss Function Search Scheme for Pedestrian Detection. *IEEE Trans. Image Process.* **2021**, *30*, 8483–8496. [CrossRef]
51. Ge, Z.; Wang, J.; Huang, X.; Liu, S.; Yoshie, O. LLA: Loss-Aware Label Assignment for Dense Pedestrian Detection. *Neurocomputing* **2021**, *462*, 272–281. [CrossRef]
52. Wang, Z.; Wang, J.; Yang, Y.; Xing, J. A Coulomb Force Inspired Loss Function for High-Performance Pedestrian Detection. *IEEE Signal Process. Lett.* **2022**, *29*, 2318–2322. [CrossRef]

53. Song, T.; Sun, L.; Xie, D.; Sun, H.; Pu, S. Small-Scale Pedestrian Detection Based on Topological Line Localization and Temporal Feature Aggregation. In Proceedings of the Computer Vision—ECCV 2018, Pt Vii, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing Ag: Cham, Switzerland, 2018; Volume 11211, pp. 554–569.
54. Liu, W.; Liao, S.; Hu, W.; Liang, X.; Chen, X. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 643–659.
55. Li, F.; Li, X.; Liu, Q.; Li, Z. Occlusion Handling and Multi-Scale Pedestrian Detection Based on Deep Learning: A Review. *IEEE Access* **2022**, *10*, 19937–19957. [CrossRef]