applied sciences

MDPI

Article

A Multi-Pedestrian Tracking Algorithm for Dense Scenes Based on an Attention Mechanism and Dual Data Association

Chang Li [1], Yiding Wang [2,*] and Xiaoming Liu [1]

[1] College of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China
[2] College of Information Science and Technology, North China University of Technology, Beijing 100144, China
* Correspondence: ydwang@ncut.edu.cn

Abstract: Aiming at the problems of frequent identity switches (IDs) and trajectory interruption of multi-pedestrian tracking algorithms in dense scenes, this paper proposes a multi-pedestrian tracking algorithm based on an attention mechanism and dual data association. First, the FairMOT algorithm is used as a baseline to introduce the feature pyramid network in the CenterNet detection network and up-sampling the output multi-scale fused feature maps, effectively reducing the rate of missed detection of small-sized and obscured pedestrians. The improved channel attention mechanism module is embedded in the CenterNet's backbone network to improve detection accuracy. Then, a re-identification (ReID) branch is embedded in the head of the detection network, and the two sub-tasks of pedestrian detection and pedestrian apparent feature extraction are combined in a multi-task joint learning approach to output the pedestrian apparent feature vectors while detecting pedestrians, which improves the computational efficiency and localization accuracy of the algorithm. Finally, we propose a dual data association tracking model that tracks by associating almost every detection box instead of only the high-scoring ones. For low-scoring detection boxes, we utilize their similarities with trajectories to recover obscured pedestrians. The experiment using the MOT17 dataset shows that the tracking accuracy is improved by 0.6% compared with the baseline FairMOT algorithm, and the number of switches decreases from 3303 to 2056, which indicates that the proposed algorithm can effectively reduce the number of trajectory interruptions and identity switching.

Keywords: multi-pedestrian tracking; attention mechanism; dual data association; FairMOT

Citation: Li, C.; Wang, Y.; Liu, X. A Multi-Pedestrian Tracking Algorithm for Dense Scenes Based on an Attention Mechanism and Dual Data Association. *Appl. Sci.* **2022**, *12*, 9597. https://doi.org/10.3390/app12199597

Academic Editor: Byung-Gyu Kim

Received: 17 August 2022
Accepted: 22 September 2022
Published: 24 September 2022

1. Introduction

In recent years, the multi-object tracking (MOT) algorithm has been a research hotspot in the field of computer vision, and it is widely used in intelligent transportation [1,2], automatic driving, surveillance, smart cities [3], and motion recognition [4]. Multi-pedestrian tracking [5–13], which is one of the most difficult and significant parts of MOT field research, has important application value for crowd congestion, safety hazards, and pedestrian flow statistics in traffic hubs, shopping malls, and other public places.

Object detection is one of the most active topics in computer vision and the basis of MOT. The continuous development of deep learning techniques has greatly improved the performance of MOT algorithms [14,15] and has made the tracking-by-detection (TBD) two-stage pedestrian tracking algorithms [16,17] the current mainstream framework. The TBD algorithm first detects the current frame image through the object detection network [17–22] and obtains multiple pedestrian detection boxes, and then correlates them with the pedestrian trajectories already established in the previous sequence of video frames by the Kalman filter and Hungarian algorithm. The DeepSORT algorithm is a classical TBD algorithm that was proposed by Wojke et al. [23] It uses YOLOv3 [24] as the pedestrian detection network and extracts pedestrian apparent features via a pedestrian re-identification module, but because pedestrian detection and apparent feature extraction are performed in two steps, there are many redundant calculations. Track-RCNN [25]

algorithm has difficulty meeting real-time requirements in pedestrian-dense scenarios, as it adds an ReID head on top of the Mask RCNN [26] and regresses a bounding box and ReID feature for each proposal. With the development of multi-task learning algorithms, one-shot pedestrian tracking algorithms with shared feature layers for pedestrian detection and apparent feature extraction have attracted the attention of scholars.

Wang et al. [27] proposed the first one-shot multi-pedestrian tracking algorithm, joint detecting embedding (JDE), which detects pedestrians while extracting their apparent features. Thanks to the high efficiency of one-stage object detection networks and the sharing of underlying features by multi-task joint networks, such algorithms can achieve near real-time tracking speeds. However, the YOLOv3 detection network JDEuse is anchor-based; it is very sensitive to hyperparameters and requires a lot of time for tuning, and the positive and negative samples are unbalanced during the training process. Therefore, in 2020, Zhang proposed the FairMOT multi-object tracking algorithm [28] based on the CenterNet [29] anchor-free detection network, which effectively makes up for the shortcomings of JDE. Because the pedestrian tracking algorithm requires extensive spatial location information, the detection network in current tracking algorithms undergoes multilayer down-sampling and convolution; a large amount of spatial location information is lost in the shallow pedestrian network, resulting in insufficient tracking accuracy and frequent IDs. Moreover, the data association method of the commonly used multi-object tracking algorithm only associates detection boxes that have scores higher than a certain threshold with the trajectories, ignoring the fact that those scores are lower than the threshold when the pedestrian is obscured, which leads to non-negligible true missing objects and fragmented trajectories. This is insufficient for depicting the pedestrian motion state and easily causes interrupted trajectories and IDs.

To solve this problem, we propose a simple and effective multi-pedestrian tracking algorithm, the attention mechanism and dual data association tracking, (AMDDATrack) algorithm, which enhances network recognition of obscured pedestrians by introducing feature pyramids networks (FPNs) [30] and high-resolution feature maps (HRs) into the neck part of the network, and an improved spatial attention mechanism module to improve the accuracy of the model in spatial location of pedestrians. Finally, we propose a dual data association method that associates almost every detection box instead of only the high-scoring ones:

(1) The two branches of pedestrian detection and apparent feature extraction are integrated and trained by multi-task learning, so that they can output pedestrian detection results and corresponding pedestrian apparent feature vectors at the same time, reducing redundant computations and improving the overall speed of the tracking algorithm.
(2) CenterNet uses multiple deformable convolutional and deconvolutional u-sampling of only one-quarter the size of the input image for prediction and does not fuse multi-layer pedestrian features, so the tracking network has a poor tracking effect for obscured pedestrians. Therefore, the FPN is introduced, and the obtained feature maps are up-sampled to obtain high-resolution feature maps after multi-layer features are fused, which effectively improves the stability of the network for tracking obscured pedestrians.
(3) An improved attention mechanism module is introduced in the backbone network of CenterNet to enhance its ability to extract spatial location information and apparent features of pedestrians, improve the tracking algorithm accuracy, and reduce the incidence of IDs.
(4) For pedestrian tracking in videos, the data association part is improved, and low-scoring detection boxes are associated with pedestrian trajectories for dual data association, which reduces the frequency of trajectory interruption when pedestrians are obscured and improves the robustness of the tracking module.

2. Materials and Methods

2.1. Dataset

To verify the effectiveness of the proposed AMDDATrack algorithm in multi-pedestrian tracking tasks, data from the publicly available CUHK-SYSU and PRW datasets were chosen as the training set and evaluated on MOTChallenge [31]. CUHK-SYSU is a large-scale benchmark dataset for people searches containing 18,184 images, 8432 pedestrian identities, and 99,809 well-labeled bounding boxes. The images are derived from two sources, films and TV dramas, and contain different perspectives, lighting, resolutions, occlusions, and backgrounds. Person re-identification in the wild (PRW), a dataset collected by Tsinghua University, is a 10-h video in which all pedestrians appearing in each frame are labeled with bounding boxes and assigned an ID at the same time. Multi-Object Tracking Challenge (MOTChallenge) is a publicly available benchmark platform for multi-object tracking. Taking MOT16 [32] as an example, 14 video sequences are provided, seven of which are used for training and seven for testing, all of which are still or moving images in an unconstrained environment. The tracking results are already given in the data, so there is no need to solve the problem of pedestrian detection during training, only tracking. In this paper, the proposed AMDDATrack algorithm is tested on the MOT16 and MOT17 datasets.

2.2. Structure of CenterNet

The first step of the multi-pedestrian tracking algorithm is to detect pedestrians, so the performance of the detection network directly affects the results of the tracking model. Since the anchor mechanism was proposed, most high-performance detection algorithms are based on anchoring. Although this is effective, the anchor-based detection network is very sensitive to the scale and aspect ratio hyperparameters of the anchor, which requires a lot of time to adjust., In order to ensure a high recall rate, a large number of anchors is required, which will lead to unbalanced positive and negative samples. In recent years, some studies have questioned the need for anchors and proposed a detection network that does not depend on anchors, CenterNet, which frees the network from the dependence on anchors.

Given that the anchor-based detection network has the problems of a large number of parameters and sensitivity to hyperparameters that directly affect the accuracy and speed of the tracking model, this paper adopts CenterNet, an anchor-free single-stage object detection network that transforms the object detection problem into an object centroid estimation problem, avoiding the introduction of a large number of parameters by presetting anchor frames. The local peak of the heatmap corresponds to the center point of the pedestrian being tracked, and the width and height of the pedestrian detection boxes are directly regressed according to the coordinates of the object's center point, which can accurately locate the pedestrian while reducing the number of parameters, providing a good basis for the subsequent tracking module to accurately match the tracking.

In the process of pedestrian detection with CenterNet, when the S_t frame passes through the detection network, three branches with different dimensions (heatmap, box_size, offset) are obtained, each of which outputs the location of all detected pedestrian centroids: $\boldsymbol{heatmap_t} = \{\boldsymbol{c_t^1}, \boldsymbol{c_t^2} \ldots \boldsymbol{c_t^N}\}$; width and height of all pedestrian detection boxes, $\boldsymbol{boxsize_t} = \{\boldsymbol{z_t^1}, \boldsymbol{z_t^2} \ldots \boldsymbol{z_t^N}\}$; the offset is used to refine the pedestrian centroids of the heatmap and improve the accuracy of the detection network. The three branches are shown in Figure 1.

During pedestrian tracking, due to camera angle and other factors, the size of pedestrians on the surveillance video screen varies widely, and the problem of mutual occlusion in tracking often occurs. Usually, detection convolutional neural networks tend to design the network very deep in order to extract more differentiated high-level semantic information about pedestrians, and as the convolutional neural network deepens, the perceptual field becomes larger and larger, which is very unfavorable to detecting small-sized pedestrians. For example, when a pedestrian turns his back to the camera and walks further and further away, the size presented on the screen becomes smaller and smaller; it will be a missed

detection, which directly leads to the interruption of the trajectory in the subsequent tracking process; in dense scenes where pedestrians block each other, the blocked pedestrians have often missed detection, which leads to ID switches in the subsequent tracking process when two pedestrians walk in opposite directions to produce blocking. It can be seen that the performance of the front-end detection network directly affects the subsequent tracking results. Therefore, we first improve the detection ability and localization accuracy of small-sized pedestrians by introducing a feature pyramid structure to the CenterNet network to fuse deep and shallow feature maps and then improve the accuracy of mutually occluded pedestrians by increasing the resolution of feature maps to reduce the probability of trajectory interruption and ID switches in the subsequent tracking process.

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

Figure 1. Three branches of CenterNet detection network: (**a**) heatmap; (**b**) center point offset; (**c**) size of box.

2.3. Improved CenterNet

2.3.1. Multi-Layer Feature Aggregation

CenterNet uses backbone networks with small sampling rates, and the module's backbone feature extraction networks, such as ResNet [33], Inception, and MobileNet [34] only use multiple deformable convolutions [35] and deconvolution to up-sample one-quarter of the input image for detection and do not achieve reuse of multi-layer features. The module is prone to missing some small pedestrians in the image and loses a large amount of pedestrian texture and location information contained in the shallow feature maps, which is essential for tracking algorithms. Therefore, multi-layer feature aggregation [36] is helpful to reduce identity switches with the one-shot algorithm due to its improved ability to handle scale variations.

As a result, we incorporated the FPN into CenterNet, as shown in Figure 2. The shallow feature map has a smaller sensory field and contains more location and feature information of the obscured pedestrians compared to the deep feature map, so its introduction makes the network directly incorporate the semantic information of obscured pedestrians, which is beneficial for extracting their features. The fusion of feature maps of different layers of the FPN can effectively improve the accuracy of the network for obscured pedestrian detection.

2.3.2. High-Resolution Feature Maps

As an important part of the tracking algorithm, pedestrian detection technology is often applied in dense pedestrian scenes, such as shopping malls, traffic hubs, road intersections, etc. Due to the high similarity in appearance of pedestrian targets and serious mutual occlusion, the visual pixels of occluded pedestrians in the image are limited, and the foreground target occupies part of the area of occluded pedestrians. In Figure 3b, the red box indicates the foreground pedestrian and the green box the obscured pedestrian; their centroids are indicated by red and green dots, respectively. According to the matching mechanism of CenterNet, centroids falling into the same sub-feature are sent to the training process as belonging to the same pedestrian, resulting in obscured pedestrians being directly ignored by the detection network and missed detection of pedestrians. Moreover, due to the effect of occlusion, most pixels are shared between occluding and occluded pedestrians, resulting in low differentiation between the two in the feature extraction process. As seen in Figure 3, compared with the one occluding, the occluded pedestrian occupies a very limited number of pixels, which results in the extracted features being mixed with a large

amount of information about the foreground target, which in turn affects the detection of the obscured pedestrian and easily causes missed detection, which leads to tracking trajectory interruptions and IDs.

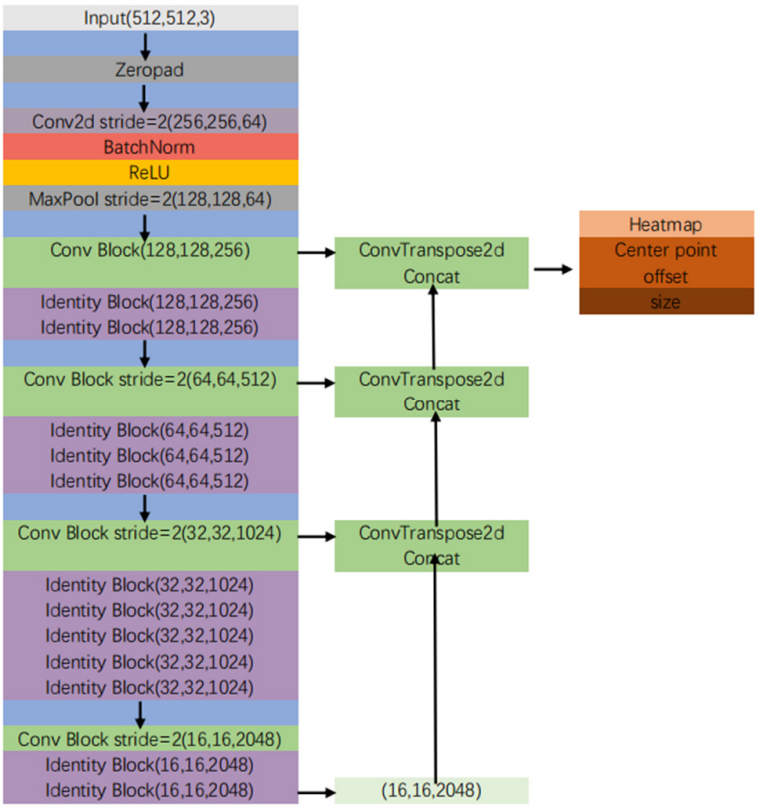

Figure 2. Multi-Layer feature aggregation network.

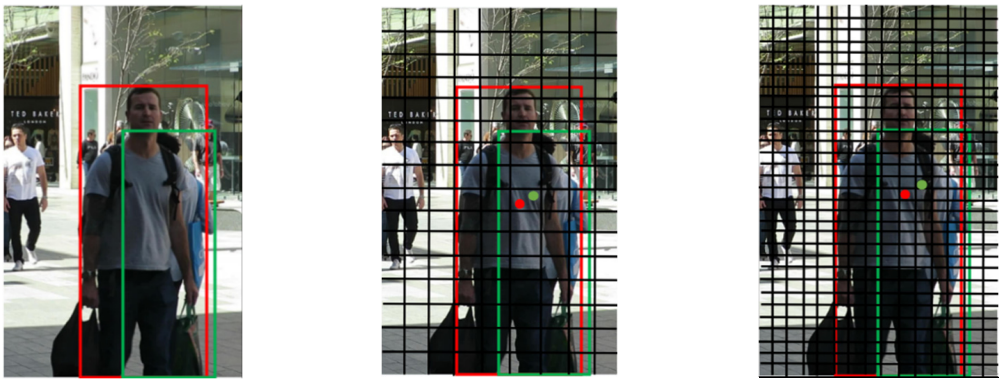

Figure 3. Difficulty in detecting obscured pedestrians: (**a**) high overlap of two pedestrians; (**b**) centroids falls into the same grid; (**c**) centroids falls into different grids.

Therefore, this paper introduces a high-resolution feature map to address the problem of missed detection of obscured pedestrians in the detection network; that is, up-sampling is performed on the feature map after the FPN, fusing the shallow pedestrian information extracted from the backbone network and the multi-layer pedestrian information. Then, as shown in Figure 3, we use up-sampling to increase the resolution of the FPN and make the centroids of highly overlapping pedestrians fall in different sub-feature regions as much as possible. The features of obscured pedestrians originate from their limited visible pixels, and the shallow features contain more details than the deep features due to their small perceptual field. Therefore, the introduction of FPN and high-resolution

feature maps not only enhances the ability of the network to extract information about the obscured pedestrians but also improves the accuracy of their localization in the subsequent tracking process.

2.4. Improved Channel Attention Mechanism Module

With the deepening of the neural network, the feature map size becomes smaller and smaller, resulting in the weakening or loss of some appearance features and spatial location information of small-sized pedestrians in the image; the relationship between feature channels is not fully considered in the upscaling and downscaling operations of the backbone network. The convolutional channel features correspond to different parts of the human body, which affects the accuracy of pedestrian localization. Therefore, the relationship between body parts and convolutional channels can be effectively utilized through the channel attention mechanism to deal with seriously occluded pedestrians. To address this, in this paper we propose an improved channel attention mechanism to enhance the performance of the ReID branch by extracting more expressive pedestrian apparent features and spatial location information in the feature channels, reducing the number of trajectory interruptions and IDs caused by missed detection or mutual occlusion between pedestrians, and improving the robustness of the multi-pedestrian tracking algorithm. In response to problems, such as loss of pedestrian spatial location information and destruction of image spatial structure caused by global average pooling (GAP) and fully connected (FC) operations in the classical SENet [37] channel attention mechanism, we designed a new module, AM-Block, shown in Figure 4b, which can effectively improve the accuracy of pedestrian localization in the subsequent tracking process.

The AM-Block structure is shown in Figure 4b, F_A represents the input feature map of the AM-Block; $W \times H$ is the width and height of F_A, and C represents the number of feature channels. After the feature map is input into the AM-Block module, to obtain rich pedestrian location information, the input feature maps are spatially compressed by GAP and global max pooling (GMP) parallel structures to obtain two feature map channel weights. For better characterization of global information, the spatially compressed feature maps are, respectively, dimensionalized by 1×1 convolution of the number of channels C/d (it is experimentally known that the best result is obtained when d is taken as 8). Replacing the fully connected operation in the common attention mechanism, the 1×1 convolution operation does not destroy the spatial structure of the image. Then the number of channels is dimensioned up by the 1×1 convolution of the number of channels as C, and finally, the feature map after the two branches are summed is subjected to the nonlinear sigmoid activation function to find the channel weights:

$$F_1 = f^1_{1\times1}(\delta(f^1_{1\times1}(f_{averagepool}(F_A)))) \tag{1}$$

$$F_2 = f^2_{1\times1}(\delta(f^2_{1\times1}(f_{\max pool}(F_A)))) \tag{2}$$

$$F_3 = \sigma(F_1 + F_2) \tag{3}$$

In Equation (1), δ represents the ReLU [38] activation function, σ represents the Sigmoid activation function, and F_3 represents the generated feature channel weights. Finally, the input feature map F_A is dotted with feature channel weights F_3 to obtain the feature map with channel attention F_B:

$$F_B = F_A \cdot F_3 \tag{4}$$

Feature map F_B after the AM-Block module can effectively increase the channel weights of important features and strengthen the important feature information of pedestrians, suppress the interference of useless information, and improve the characterization ability of the backbone feature extraction network.

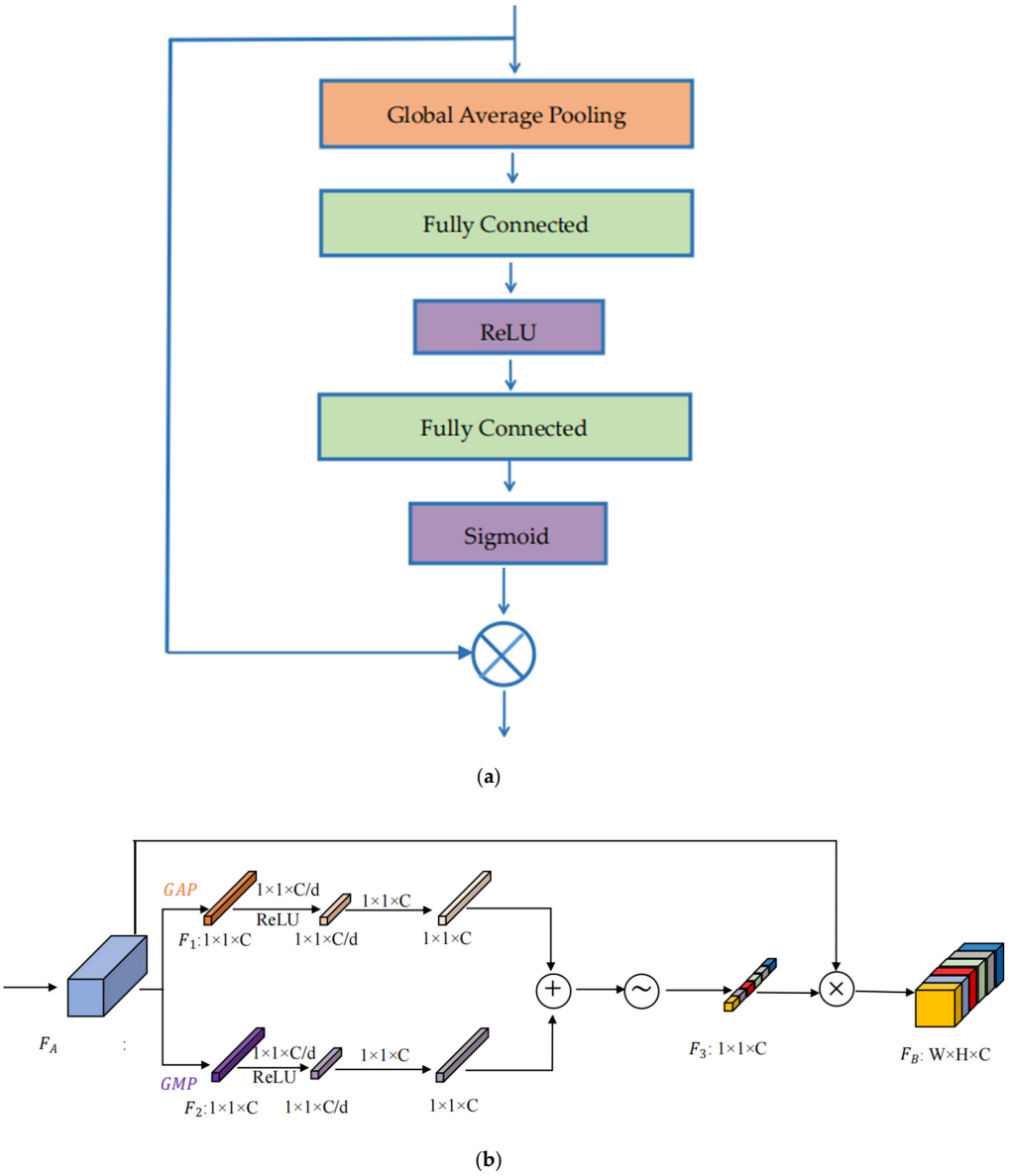

Figure 4. Improved channel attention mechanism module; (**a**) flowchart of SENet; (**b**) flowchart of AM-Block module.

2.5. Dual Data Association

The multi-pedestrian tracking algorithm is aimed at continuously localizing multiple pedestrians in a video sequence, maintaining the identity of each pedestrian consistently across video frames, and generating pedestrian motion trajectories. The most commonly used tracking algorithm uses the Kalman filter to process video data frame by frame, and the Hungarian matching algorithm matches the pedestrian detection boxes and appearance feature vectors output from the detection network in the current frame with the pedestrian position predicted by the Kalman filter in the current frame across frames. The cascade matching method is used for the problem of associating detected pedestrians in the current frame with already existing tracking trajectories, and pedestrians that appear more frequently are prioritized for matching to solve the probability dispersion problem of continuous prediction.

$$X_k = AX_{k-1} + BU_{k-1} + W_{k-1} \tag{5}$$

$$Z_k = HX_k + V_k \tag{6}$$

$$\check{X}_k = A\check{X}_{k-1} + BU_{k-1} \tag{7}$$

$$P_{\bar{k}} = AP_{k-1}A^T + \mathrm{Q} \tag{8}$$

$$K_k = \frac{P_{\bar{k}}H^T}{HP_{\bar{k}}H^T + R} \tag{9}$$

$$\hat{X_k} = \hat{X}_{\bar{k}} + K_k\left(Z_k - H\hat{X}_{\bar{k}}\right) \tag{10}$$

$$P_k = (I - K_kH)P_{\bar{k}} \tag{11}$$

The equations of the Kalman filter are given as Equations (5)–(10), where $\hat{X}_{\bar{k}}$ represents the a priori state estimate, which is predicted from the optimal estimate at the previous moment to the current moment; the Z_k in Equation (6) represents the actual detected pedestrian motion state by CenterNet at the moment t, Z_k is detected by the CenterNet detection network and X_k is calculated by the Kalman filter. Equation (7) is the covariance matrix of Equation (6), and Equation (8) is the covariance matrix of Equation (5). By using the covariance matrix of the theoretical value in Equation (5) and the actual measured values in Equation (6), we can know their stability, respectively. In the pedestrian tracking process, in order to determine whether the theoretical value in Equation (5) or the detected value in Equation (6) is accurate, the equation shown in Equation (10) is established. $\hat{X_k}$ represents the optimal estimate, $Z_k - H\hat{X}_{\bar{k}}$ represents the difference between the pedestrian motion state predicted by the Kalman filter and the motion state detected by the CenterNet detector, so K_k represents the weight, whose value is close to the true value, which value of the weight is greater. It can be seen in Equation (9) that K_k is larger when the covariance $P_{\bar{k}}$ of the theoretical value is larger, and at the same time, we can see by Equation (10) that the proportion of the theoretical value is smaller when K_k is larger; after obtaining the measured and theoretical values and their differences, we update the parameters of the Kalman filter using Equation (11).

However, there is an occlusion problem in real traffic scenarios, and mutual occlusion between pedestrians leads to a lower detection frame score for occluded pedestrians. In Figure 5, the pedestrian with ID number 4 is gradually blocked from frame 77 to frame 95, and the detection frame score gradually decreases from 0.8 to 0.1. Therefore, although the pedestrian is detected in frames 91 and 95, this pedestrian is discarded because the detection frame score of 0.1 is less than the confidence threshold, resulting in a trajectory interruption. Thus, it is not reasonable to discard the low-scoring pedestrian detection frame based only on a frame confidence threshold.

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

Figure 5. Dropping a low-scoring detection frame causes a trajectory break; (**a**) 77th frame; (**b**) 91st frame; (**c**) 95th frame.

In this paper, we set the confidence threshold of the detection boxes to 0.01 to keep the low-scoring detection boxes, and for the high-scoring detection boxes, we followed the current common data association method, Then, we used the Hungarian algorithm to match it optimally. The low-scoring detection boxes were associated with the trajectories that were not successfully matched with the high-scoring detection boxes in the previous frame for the second time; then, the Hungarian algorithm was used again to match the low-scoring pedestrian detection boxes with the existing trajectories. The purpose is to pull out partially occluded pedestrians from low-scoring detection boxes, optimize the trajectory interruption problem caused by occlusion during the tracking process, and maintain the continuity of the trajectories.

For the convenience of calculation, assume that the pedestrians in the video are moving in a uniform motion, and the flow chart of the improved tracking algorithm is shown in Figure 6, it has the following steps:

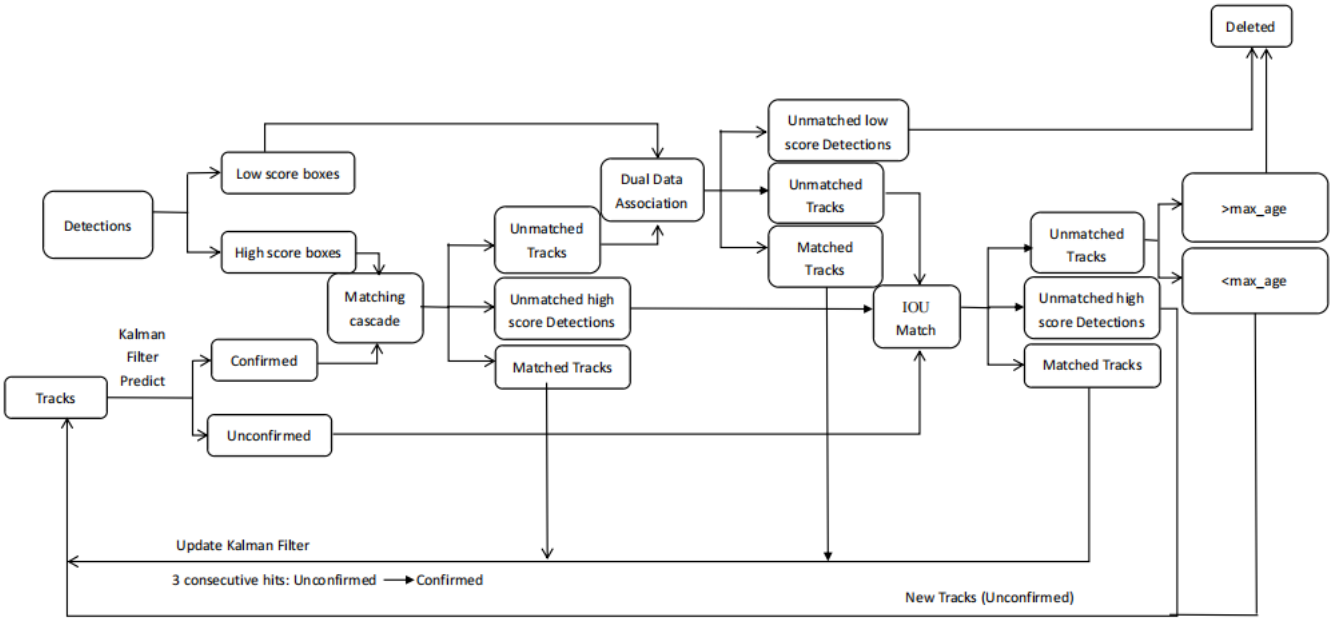

Figure 6. Flowchart of dual data association module.

Step 1: Output the detected boxes and apparent feature vectors of all pedestrians detected in the current frame through the anchor-free detection network, and set the confidence threshold to 0.01 to retain both high and low-scoring detected boxes;

Step 2: Use the Kalman filter to predict the pedestrian motion state of the current frame and set a tracker for each existing track. If the Kalman filter predicts the pedestrian motion state of the current frame with a matching high-scoring pedestrian detection box, the counter of the tracker is set to 0; conversely, if the tracker does not find matching detection boxes for a period of time (max_age > 30), the tracker is deleted and the line of people leaves the video screen by default.

Step 3: At the same time, when a new pedestrian appears in a frame, a new tracker is created for the pedestrian. If the motion state predicted by the Kalman filter can find a matching detection box for three consecutive frames, the new pedestrian is confirmed and the tracker state is set to "confirmed"; otherwise, it is "unconfirmed", which generally prioritizes matching the confirmed tracker, where the matching algorithm uses the Hungarian algorithm and considers the motion information and appearance feature association of the pedestrian at the same time.

Motion information correlation: Because of the continuity of pedestrian motion in the video, the degree of correlation between the high-scoring pedestrian detection box and the position predicted by the tracker is calculated using the Mahalanobis distance, as shown in Equation (12):

$$l^{(1)}(i,j) = (l_j - p_i)^T X_i^{-1} (l_j - p_i) \tag{12}$$

where $l^{(1)}(i,j)$ represents the match between the j-th pedestrian detection box in the current frame and the i-th trajectory predicted by the Kalman filter, X_i^{-1} represents the covariance matrix between the position of the pedestrian detection box and the average tracking position, l_j represents the position of the j-th pedestrian detection box in the current frame, and p_i represents the position of the trajectory predicted by the i-th tracker in the current frame.

Appearance information association: In order to enhance the robustness of the model when pedestrians are obscured and then reappear, pedestrian apparent feature vectors

are introduced for data association. First, for each high-scoring pedestrian detection box output by the detection network, a 128-dimensional feature vector is extracted by the ReID module, and a list is created for each tracked pedestrian, storing the apparent feature vectors of the last 100 frames that were successfully matched. Next, the minimum cosine distance between the last 100 successfully associated feature sets of the i-th tracker and the appearance feature vector of the j-th pedestrian detection box in the current frame is calculated. If this distance is less than the set threshold, indicating success, the cosine distance is calculated as shown in Equation (13):

$$l^{(2)}(i,j) = \min\{1 - r_j^T r_k | r_k^i \in R_i\} \tag{13}$$

The degree of matching between the detection boxes and the tracking trajectories is calculated by simultaneously considering the correlation between pedestrian motion information and appearance information, and then linearly weighting, as shown in Equation (14):

$$c_{i,j} = \lambda l^{(1)}(i,j) + (1-\lambda) l^{(2)}(i,j) \tag{14}$$

when $c_{i,j}$ is within the intersection of the two metric thresholds, the high-scoring pedestrian detection boxes are considered to be correctly matched with the existing trajectories, and the remaining detection boxes and trajectories are classified as unmatched.

Step 4: At this point, the unsuccessfully matched trackers and low-scoring pedestrian detection boxes in step 1 are again subjected to minimum cosine distance calculation, obscured pedestrians are excavated in the low-scoring detection boxes, successfully matched detection boxes and the track are updated with the Kalman filter, and unsuccessful low-scoring detection boxes are directly deleted.

Step 5: Finally, the unconfirmed trackers in step 4 are compared with the unconfirmed high-scoring detectors and unconfirmed trackers in step 3 to calculate the minimum cosine distance, and the successful ones are updated with the Kalman filter. For the unconfirmed trackers, if they are "confirmed" and max_age < 30, the tracker will be saved and continue to match with the next frame of the video. If it is "not-confirmed", it will be deleted directly.

3. Experimental Results and Analysis

3.1. Experimental Environment

The environment configuration for this experiment was as follows: Intel (R) Core (TM) i7-10870H CPU @2.20HZ and NVIDIA GeForce RTX 2060 Ti GPU; the software environment for experiments and testing includes Windows 10 operating system, CUDA 10.0 +cuDNN7.1 GPU gas pedal, and pytorch based deep learning framework.

3.2. Experimental Evaluation Criteria

In order to compare the performance of multi-pedestrian tracking algorithms more objectively, this paper uses the evaluation criteria commonly used in the field of multi-object tracking and object detection, and the main MOT indexes are shown in Table 1. In the table, an upward arrow indicates the higher the value, the better the algorithm performance, and a downward arrow indicates the lower the value, the better the algorithm performance.

Table 1. Evaluation Criteria and their meaning.

Evaluation Criteria	Meaning of Criteria
FP↓	Rate of being misidentified as a positive sample, i.e., false detection rate
FN↓	Rate of being mistaken for negative samples, i.e., missed detection rate
IDs↓	Number of pedestrian ID switches, i.e., pedestrian identity changes
MOTA↑	Tracking accuracy calculated by metrics, such as FP, FN, IDs
IDF1↑	Accuracy and recall of tracking with constant ID
ML↓	Tracking of failed pedestrians as a percentage of all pedestrians
MT↑	Track of successful pedestrians as a percentage of all pedestrians

Tracking accuracy (MOTA), shown in Table 1, is the main evaluation index of the multi-pedestrian tracking algorithm. It can directly reflect the performance of the algorithm, and the calculation process is given in Equation (15), where GT represents the real number of pedestrians in the video frame image, and the MOT value ranges from negative infinity to 1. Since the detection network is a very important part of the tracking algorithm and is improved in this paper, it needs to be evaluated as well. The performance of the detection network is evaluated based on precision, recall, average precision (AP), and mean average precision (mAP) as shown in Equations (16)–(18), in which TP stands for true positive, FP stands for false positive, FN stands for false negative.

$$MOTA = 1 - \frac{FN + FP + IDs}{GT} \in (-\infty, 1] \tag{15}$$

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{AllDetections} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{AllGroundTruths} \tag{17}$$

$$mAP = \frac{1}{|Q_R|} \sum_{q=Q_R} AP(q) \tag{18}$$

3.3. Ablation Experiments

3.3.1. Detection Performance

Because the detection performance largely determines the performance of the tracking algorithm, in order to verify the effectiveness of the improved CenterNet detection network proposed in this paper, first, we conducted two sets of ablation experiments for the two parts of the improvement and then selected four current mainstream detection networks to compare with the improved CenterNet. The first set of experiments tested the effectiveness of ablation of the feature pyramid network (FPN) and high-resolution feature map (HR) introduced in the neck part of the CenterNet The data in Table 2 show that the introduction of both the FPN and the HR can effectively improve the detection capability of the network, and the combination of the two results in a more significant improvement in detection capability.

Table 2. Impact of improvements to neck part of CenterNet on detection performance.

Algorithms	mAP(%)
CenterNet	68.5
CenterNet+FPN	68.9
CenterNet+HR	68.6
CenterNet+FPN+HR	70.3

The second group of experiments compared the effect of the improved channel attention (CA) mechanism in the backbone network on the detection capability of CenterNet by using the same dataset. As shown in Table 3, the improved channel attention mechanism significantly improves the performance of the detection network by adaptively improving the differentiation of pedestrian features extracted by the backbone network with almost no increase in computation and inference time.

Table 3. Impact of improved spatial attention mechanism on CenterNet detection capability.

Algorithms	mAP(%)
CenterNet	68.5
CenterNet+FPN+HR	70.3
CenterNet+FPN+HR+CA_ours	70.5

The third group of experiments was conducted to verify the performance of the proposed improved CenterNet detection network compared with four mainstream detection networks, among which SSD [39] and YOLOv4 [40] are single-stage detection networks, and Faster RCNN [41] and Mask RCNN are two-stage detection networks. CenterNet_ours is the improved network in this paper. The results in Table 4 show that the improved CenterNet proposed in this paper is significantly better than the mainstream networks in common use today, with a 4.2 percentage point improvement over YOLOv4, and a 13.1 percentage point improvement compared to Faster RCNN, the classical two-stage detection network.

Table 4. Comparison with mainstream detection networks.

Algorithms	AP(%)
SSD	31.2
YOLOV4	43.5
Faster RCNN	34.7
Mask-RCNN	39.8
CenterNet	47
CenterNet_ours	47.6

3.3.2. Tracking Performance

To verify the impact of the proposed improved CenterNet detection network and the dual data association module on the tracking algorithm, we experimentally compared the proposed AMDDATrack algorithm with the current mainstream tracking algorithms; the results are shown in Table 5. It can be seen that the improved detection network performance and retrieval of low-scoring detection boxes lead to a significant improvement in tracking algorithm accuracy with FairMOT as the baseline.

Table 5. Tracking results on the MOT17 dataset.

Algorithms	MOTA	IDF	IDs	FN	FP
Tube_TK(one-shot) [41]	63.0	58.6	413	177,483	27,060
CSTrack(one-shot) [42]	74.9	72.6	3567	114,303	23,847
DeepSORT(two-stage)	60.3	61.2	2442	185,301	36,111
TransTrack(one-shot) [43]	65.8	56.9	5355	163,683	24,000
FairMOT(one-shot)	73.7	72.3	3303	117,477	27,507
AMDDATrack(one-shot)	74.3	75.3	2056	84,932	26,773

As shown in Table 5, with the conf_thres set to 0.6, the tracking accuracy of the AMDDATrack algorithm proposed in this paper is 0.6 percentage points higher on the MOT17 dataset than the FairMOT multi-pedestrian single-stage tracking algorithm; in particular, the IDs index is significantly reduced, indicating that the AMDDATrack algorithm has improved robustness when dealing with severe occlusion and can significantly reduce the number of pedestrian trajectory interruptions and ID switches.

4. Discussion

In order to demonstrate the effectiveness of the algorithm proposed in this paper more intuitively, Figures 7 and 8 show the tracking results of FairMOT and AMDDATrack on the test sequences of MOT15, MOT16, and MOT17 datasets. Figure 7 shows a section of tracking results of the two algorithms on the ETH-Crossing sequence of the MOT15 dataset. The scene has complex pedestrian flow, changing backgrounds, a large number of pedestrians, and large appearance changes depending on the shooting angle. With regard to the numbers marked on the pedestrians in the video, when tracking with the FairMOT algorithm, the algorithm assigns ID numbers 1 and 10 to the middle-aged couple (Figure 7a), and when the boy wearing a helmet crosses through the pedestrians, an occlusion is gradually created (Figure 7b). Because of the occlusion, the trajectory is

interrupted, and the two pedestrians are reassigned new ID numbers, 16 and 19, at frame 126, (Figure 7d), that is, an ID switch occurs, and the couple has a total of six switches in the whole video sequence. The AMDDATrack algorithm initially assigns the couple ID numbers 1 and 7, (Figure 7e), and when the boy with the helmet gradually obscured the lady with number ID 1, there is a brief interruption in the trajectory, but the tracking frame is quickly retrieved at frame 120. The ID numbers of both people do not change after the end of masking, and no ID switches occur in the whole video sequence, which maintains accurate and stable tracking the whole time (Figure 7h).

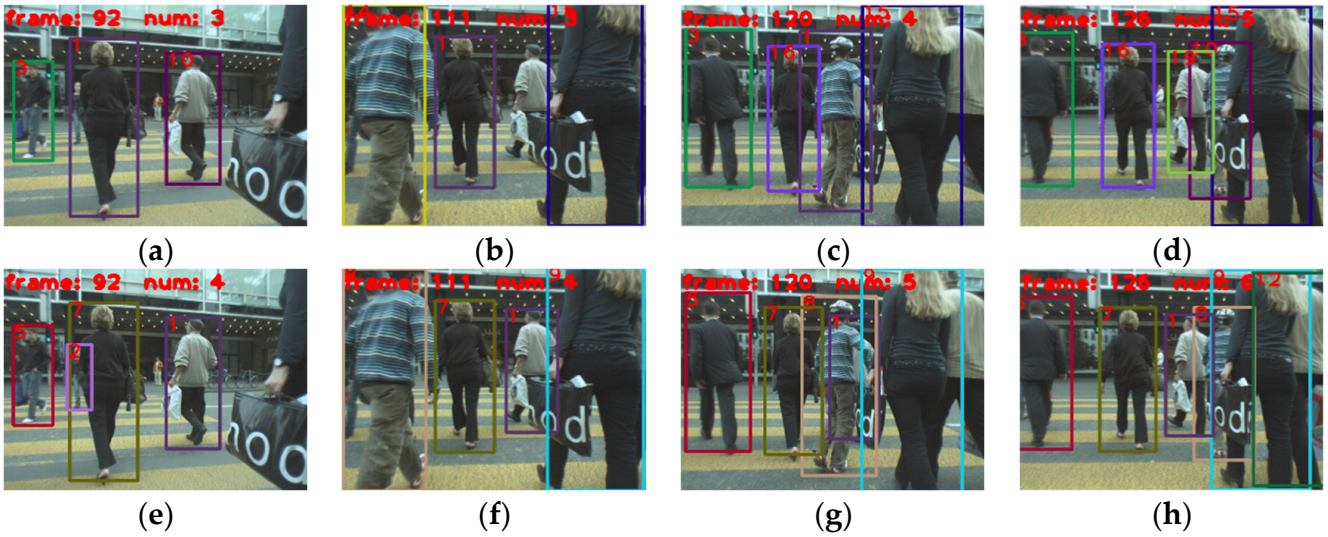

Figure 7. Tracking results of ETH-Crossing sequence. (**a**) FairMOT results in frame 92. (**b**) FairMOT results in frame 111. (**c**) FairMOT results in frame 120. (**d**) FairMOT results in frame 126. (**e**) AMDDATrack results in frame 92. (**f**) AMDDATrack results in frame 111. (**g**) AMDDATrack results in frame 120. (**h**) AMDDATrack results in frame 126.

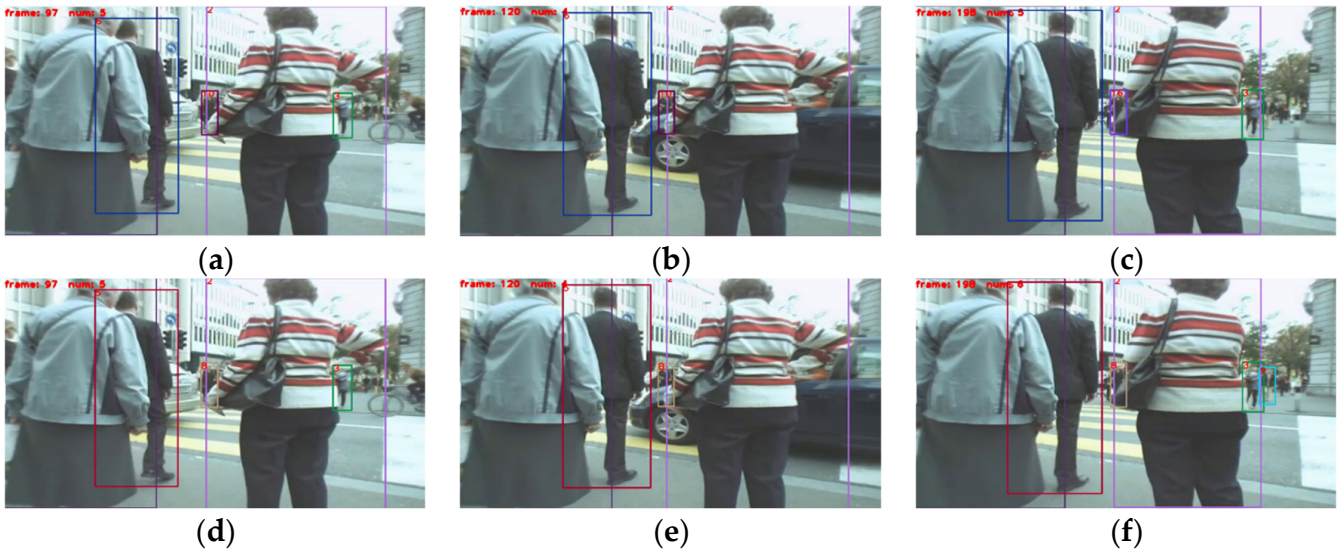

Figure 8. Tracking results of 05-SDP sequence. (**a**) FairMOT results in frame 97. (**b**) FairMOT results in frame 120. (**c**) FairMOT results in frame 198. (**d**) AMDDATrack results in frame 97. (**e**) AMDDATrack results in frame 120. (**f**) AMDDATrack results in frame 198.

The tracking effects of the FairMOT and AMDDATrack algorithms proposed in this paper on the 05-SDP video sequence in the MOT17 dataset are shown in Figure 8. When using the FairMOT algorithm to track the video sequence, the pedestrian in the center of

the frame is detected at frame 97, and the algorithm assigns him ID number 10 (Figure 8a). Subsequently, the video is continuously obscured by vehicles and pedestrians to different degrees (Figure 8b), and the pedestrian is reassigned ID number 16 at frame 198, i.e., the trajectory is interrupted, generating an ID switch, A total of three ID switches are generated in the whole video, and track interruption is generated at frame 214. When the AMDDATrack algorithm is used, the pedestrian in the center of the frame is initially assigned ID number 8, (Figure 8d); the algorithm can still track him continuously and stably after vehicles and other pedestrians obscure him, (Figure 8f). When the pedestrian in the center of the frame is nearly half obscured by the pedestrian on the left, the algorithm can still accurately associate him with ID number 8. The trajectory is associated, and no ID switches occur for the pedestrian with number 8 in the whole video, but the trajectory is interrupted at frame 238 due to complete occlusion. That is, the proposed algorithm can still correctly associate the pedestrian tracking frame with the pedestrian with ID number 8 in the case of severe occlusion, without generating track interruptions and ID switches.

From the experimental results of these two sets of video sequences, it can be seen that the algorithm in this paper has good robustness when dealing with multi-pedestrian tracking in complex scenes; especially when the pedestrians are severely obscured, it can still correctly associate the detection boxes with the tracking trajectories and the tracking performance is stable.

5. Conclusions

In this paper, we proposed a multi-pedestrian tracking algorithm called AMDDATrack based on the attention mechanism and dual data association to address the problem of frequent IDs and trajectory interruptions in the current mainstream multi-pedestrian tracking algorithms. The proposed algorithm is based on the FairMOT algorithm and enhances pedestrian apparent feature extraction ability, improves tracking algorithm accuracy, and significantly reduces the number of IDs; a dual data association method is designed for the characteristics of pedestrian tracking, which effectively reduces the number of trajectory interruptions in the case of pedestrian occlusion, with almost no additional parameters introduced to increase the computation. Moreover, the network complexity is low and it is easy to train compared with higher accuracy algorithms. However, there are still some shortcomings in this algorithm. One is that it did not deal with the problem of an anchor-free detection network judging more than one pedestrian as the same when dense pedestrian centroids overlap; the second is that there is still room to further improve the problem of missed and false detection of small-sized pedestrians in scenes with strong lighting changes.

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

Funding: This research is supported by the National Key Research and Development Program of China (2018YFB601003) and the Beijing Great Wall Scholar Training Program (CIT&TCD20190304).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Nama, M.K.; Nath, A.; Bechra., J. Machine learning-based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. *Int. J. Commun. Syst.* **2021**, *34*, e4814. [CrossRef]
2. Blome, D.S.; Beveridge, J.R.; Draper, B.A.; Liu, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
3. Coccoli, M.; De Francesco, V.; Fusco, A.; Maresca, P. A cloud-based cognitive computing solution with interoperable applications to counteract illegal dumping in smart cities. *Multimed. Tools Appl.* **2022**, *81*, 95–113. [CrossRef]
4. Shen, Y.; Lin, W.; Wang, Z.; Li, J.; Sun, X.; Wu, X.; Wang, S.; Huang, F. Rapid Detection of Camouflaged Artificial Target Based on Polarization Imaging and Deep Learning. *IEEE Photonics J.* **2021**, *13*, 1–9. [CrossRef]
5. Hong, H.H.; Yi, X.; Yan, J.H.; Qian, Y.; Zhi, G.Z. Pedestrian Tracking by Learning Deep Features. *J. Vis. Commun. Image Represent.* **2022**, *83*, 103428. [CrossRef]
6. Kim, S.J.; Nam, J.Y.; Ko, B.C. Online Tracker Optimization for Multi-Pedestrian Tracking Using a Moving Vehicle Camers. *IEEE Access* **2018**, *6*, 48675–48678. [CrossRef]
7. Kalun, H.; Janis, K.; Margret, K. Unsupervised Multiple Person Tracking using AutoEncoder-Based Lifted Multicuts. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 19 June 2020.
8. Yuan, G.; Jian, N.C.; Xiao, S.Y.; Cheng, D.W. A Modified Multi-Pedestrian Tracking System. In Proceedings of the Chinese Control Conference, Guangzhou, China, 27–30 July 2019.
9. Kai, C.; Xiao, S.; Xiang, Z. An Integrated Deep Learning Framework for Occluded Pedestrian Tracking. *IEEE Access* **2019**, *7*, 26060–26072.
10. Sungmin, Y.; Sungho, K. Recurrent YOLO and LSTM-based IR single pedestrian tracking. In Proceedings of the 19th International Conference on Control Automation and Systems (ICCAS), Jeju, Korea, 15–18 October 2019; pp. 94–96.
11. Guojiang, S.; Linfeng, Z.; Jihan, L. Infrared Multi-Pedestrian Tracking in Vertical View via Siamese Convolution Network. *IEEE Access* **2019**, *7*, 42718–42725.
12. Ge, Y.; Zihao, C. Pedestrian Tracking Algorithm for Dense Crowd based on Deep Learning. In Proceedings of the 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2–4 November 2019; pp. 568–572.
13. Li, C.; Li, G. Learning Multiple Instance Deep Representation for Objects Tracking. *J. Vis. Commun. Image Represent.* **2022**, *71*, 102737. [CrossRef]
14. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and real-time tracking. In Proceedings of the International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016.
15. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Zhao, X.; Kim, T.-K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. [CrossRef]
16. Mahmoudi, N.; Ahadi, S.M.; Rahmati, M. Multi-target tracking using CNN-based features. *Multimed. Tools Appl.* **2019**, *78*, 7077–7096. [CrossRef]
17. Chen, L.; Ai, H.; Zhuang, Z. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In Proceedings of the 26th IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
18. Xing, K.Z.; Shu, C.L.; Xu, W. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11 October 2021; pp. 2778–2788.
19. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
20. Ross, G. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. Tsung, Y.L.; Priya, G.; Ross, G. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
23. Wojke, N.; Bewley, A.; Paulus, D. Simple online and real-time tracking with a deep association metric. In Proceedings of the International Conference on Image Processing, Beijing, China, 17–20 September 2017.
24. Lin, J.Y.; Yu, C.F.; Ning, X. Video Instance Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5188–5197.
25. Kaiming, H.; Georgia, G.; Piotr, D. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
26. Zhong, D.W.; Liang, Z.; Yi, X.L. Towards Real-Time Multi-Objects Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 107–122.
27. Yi, Y.Z.; Chun, Y.W.; Xing, G.W. FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087.

28. Kai, W.D.; Song, B.; Ling, X.X. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
29. Tsung, Y.L.; Piotr, D.; Ross., G. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
30. Dendorfer, P.; Osep, A.; Milan, A.; Schindler, K.; Cremers, D.; Reid, I.; Roth, S.; Leal-Taixé, L. MOT Challenge: A Benchmark for Single-Camera Multiple Target Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 845–881. [CrossRef]
31. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
32. He, K.; Zhang, X.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
34. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
35. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412.
36. Jie, H.; Li, S.; Gang, S. Squeeze-and-Excitation Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the 14th International Conference on Artifical Intelligence and Statistics, Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
38. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherland, 8–16 October 2016; pp. 21–37.
39. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y. YOLOV4: Optimal Speed and Accuracy of Object Detection. In Proceedings of the Conference on Computer Vision and Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
40. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 28–32. [CrossRef] [PubMed]
41. Bo, P.; Yi, Z.L.; Yi, F.Z. Tubetk: Adopting tubes to tracks multi-object in a one-step training model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6308–6318.
42. Chao, L.; Zhi, P.Z.; Xue, Z. Rethinking the competition between detection and reid in multi-object tracking. *arXiv* **2020**, arXiv:2010.12138.
43. Pei, Z.S.; Jin, K.C.; Yi, R. Transtrack: Multiple-object tracking with transformer. In Proceedings of the Conference on Computer Vision and Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.