

---

# Pedestrian Tracking Algorithm Combining Contextual Information and Attention Mechanism

Shunliang Xiao, Zanxia Qiang, Weiguang Liu<sup>\*</sup>, Xianfu Bao

School of Computer Science, Zhongyuan University of Technology, Zhengzhou, China

## Email address:

1198928367@qq.com (Shunliang Xiao), Weiguang.liu@zut.edu.cn (Weiguang Liu)

<sup>\*</sup>Corresponding author

## To cite this article:

Shunliang Xiao, Zanxia Qiang, Weiguang Liu, Xianfu Bao. Pedestrian Tracking Algorithm Combining Contextual Information and Attention Mechanism. *American Journal of Computer Science and Technology*. Vol. 4, No. 4, 2021, pp. 111-118. doi: 10.11648/j.ajcst.20210404.14

**Received:** October 8, 2021; **Accepted:** October 25, 2021; **Published:** November 5, 2021

---

**Abstract:** In the real scene, because pedestrians are occluded or the size of pedestrians is small, the convolutional neural network cannot fully extract their features, resulting in poor detection results. In two adjacent frames, the same pedestrian is prone to errors when doing data association, which makes the pedestrian tracking effect unsatisfactory. In order to solve this problem, the pedestrian tracking algorithm based on Anchor-free idea is improved. A fusion context information module is proposed to enhance the model's feature extraction ability for different receptive fields, and improve the model's detection and tracking performance when the pedestrian size is small. In addition, in order to let the model learn to pay attention to the effective information of the feature layer. A coordinated attention mechanism is introduced to guide the model to learn the weights of different channels and different regions of the feature layer, and to improve the tracking performance of the model when pedestrians are occluded. In the experiment, the tracking performance of the model was verified on the MOT16 dataset. Experimental results show that compared with other main popular person tracking algorithms, the improved algorithm has higher tracking accuracy and lower pedestrian ID switching times. Its tracking accuracy is 70.74.

**Keywords:** Pedestrian Tracking, Anchor-Free, Context Information, Attention Mechanism, JDE

---

## 1. Introduction

Multi-object tracking is a basic task in computer vision. Pedestrian tracking is achieved by detecting pedestrians in the video and associating the same pedestrian in adjacent frames [1]. According to whether the information of future frames is needed during tracking, multi-object pedestrian tracking algorithms can be divided into two categories: online tracking and offline tracking. Because online tracking has better real-time performance and application prospects, it is favored by researchers. However, online tracking can only use the pedestrian information of the current frame and the previous frame. Pedestrian information in future frames cannot be used, so the optimal tracking effect cannot be achieved. Especially when pedestrians are occluded, the detector cannot fully obtain pedestrian information, resulting in tracking failure [2].

The current main solution is based on the Tracking by Detection (TBD), which divides pedestrian tracking into two stages: the first stage uses the detector to detect all

pedestrians in a single frame of image. The second stage extracts the apparent features of pedestrians, and associates the detection results with the pedestrians in the historical trajectory. This type of algorithm uses detection models and apparent feature extraction models to realize pedestrian detection and apparent feature extraction respectively. The apparent feature extraction model usually selects the pedestrian re-identification (Re-ID) model. With the development of deep learning, the accuracy of detection algorithms and re-identification algorithms has been improved. Therefore, the accuracy of this type of multi-object pedestrian tracking algorithm has also been improved. However, since the detection model and the apparent feature extraction model are two independent modules, the total inference time during tracking is the sum of the inference time of these two modules. Therefore, the real-time performance of tracking is poor. With the development of multi-task learning, the parallel realization of pedestrian detection and apparent feature extraction through a single network has attracted the attention of researchers.

VOIGTLAENDER et al. [3] proposed to add a Re-ID branch to Mask R-CNN to extract the apparent features of pedestrians. Wang et al. [4] extracted the apparent features of pedestrians by adding a Re-ID branch to the detection head part of YOLOv3. The above algorithm transforms the tracking task into multi-task learning. Use a shared backbone network to simultaneously complete pedestrian detection and apparent feature extraction, referred to as JDE (Joint Detection and Embedding) algorithm.

Zhang et al. [5] believe that the Anchor-based JDE algorithm will affect the final tracking effect due to the overlap of the Anchor when extracting the apparent features. In order to eliminate this effect, they proposed the FairMOT algorithm based on the Anchor-free idea. This algorithm improves the accuracy of pedestrian tracking and also reduces the number of pedestrian ID switching. However, the tracking effect of this algorithm is still not ideal when the pedestrian is occluded or the size is small. In order to solve the above problems, we designed a UFF (Upsampling Feature Fusion) module that integrates context information to enhance the model's feature fusion ability for different receptive fields. In addition, a coordinate attention mechanism is introduced to guide the model to pay attention to the effective area and useful content of the feature layer. We improve the performance of pedestrian tracking by improving the expressive ability of the model. The main contributions of this paper are as follows:

- 1) Propose a UFF module that integrates context information to enhance the model's feature extraction ability for different receptive fields.
- 2) Introduce the coordinate attention mechanism, and explore the effect of different combinations of the coordinate attention mechanism and the model on the tracking performance.

## 2. Related Work

### 2.1. FairMOT Model

The FairMOT multi-object pedestrian tracking algorithm proposed by ZHANG et al. includes two stages: one is to output the pedestrian detection results and the Re-ID pedestrian apparent feature extraction results in parallel. The second is to use the intersection ratio between the detection frame and the trajectory prediction frame, as well as the similarity of the pedestrian's apparent characteristics, to associate the detected pedestrians with the pedestrians in the historical trajectory. The pedestrian detection results and Re-ID feature extraction of the algorithm use the idea of multi-task learning, that is, feature extraction through a shared backbone network. Then the extracted features are sent to the detection branch and the Re-ID branch in parallel to complete the detection of pedestrians and the extraction of apparent features. In order to balance the detection branch and the Re-ID branch, FairMOT uses DLA-34 [6] as the backbone network for feature extraction. When the same pedestrian in adjacent frames is associated. First, the Kalman filter is used

to predict the trajectory of pedestrians. Then calculate the intersection ratio of the pedestrian detection frame and the trajectory prediction frame, and the cosine distance of the pedestrian's apparent feature. Finally, the Hungarian algorithm is used to complete the matching between the pedestrians in the current detection results and the pedestrians in the historical trajectory, so as to achieve pedestrian tracking.

### 2.2. Contextual Information

In actual scenes, pedestrians often do not exist alone, but are connected to the surrounding environment. The information related to pedestrians is called contextual information. In computer vision tasks, these contextual information can help the model classify and locate pedestrians more accurately. For example, when pedestrians are detected, numerous prediction boxes are generated. If the background information is a road, this prediction box has a higher probability of being a pedestrian. If the background is a bulletin board, the probability of this prediction box being a pedestrian is small. Making full use of context information helps pedestrian detection and apparent feature extraction, thereby improving the accuracy of pedestrian tracking.

In the actual tracking process, the feature information displayed by the occluded or small-sized pedestrians is not sufficient, and the down-sampling process of extracting pedestrian features by the convolutional neural network will cause part of the information to be lost. In the end, there are fewer effective features sent to the detector, resulting in a high rate of missed and false detections of pedestrians. How to effectively extract multi-scale contextual information is of great significance to solve the problem of pedestrian size change and occlusion. Inception [7] uses multiple branches and combines different sizes of convolution kernels to obtain multi-scale feature information. ASPP [8] uses dilated convolution with different dilated rates to capture the feature information of different receptive fields while retaining the resolution of the feature map. RFBNet [9] combines the ideas of Inception and ASPP. Use convolution kernels of different sizes and dilated ratios to obtain rich feature information. DAI et al. [10] believe that the geometric structure of the ordinary convolution kernel is fixed. When sampling the input feature layer with a fixed-size convolution kernel. The sampling position is fixed, and the geometric transformation of the object position cannot be handled well. Therefore, he proposed Deformable Convolution (DC). Deformable convolution learns the offset of sampling points to make it more accurately reflect the shape and position of the object in the feature map. HAASE et al. [11] pointed out that depthwise separable convolution is widely used in computer vision tasks, which reduces the amount of model calculation and improves the expressive ability of the model. Inspired by the above ideas, we propose the up-sampling feature fusion UFF module. The deformable convolution is combined with the depthwise separable convolution with different dilated rates to obtain the context feature information of different scales and receptive fields. To solve the problem of pedestrian size change and obscuration.

### 2.3. Attention Mechanism

The attention mechanism originated from the study of human vision. Inspired by its thoughts, researchers introduced the attention mechanism into computer vision tasks. It can make the model focus on important information and ignore irrelevant information. It has been widely proved that the use of attention mechanism can increase the calculation amount of the model within a reasonable range and make the model achieve better results. For example, in vision tasks such as object detection, semantic segmentation, and image classification. Adopting the attention mechanism can usually further improve the performance of the model. SENet [12] establishes the correlation between feature map channels through a 2-dimensional pooling operation. BAM [13] proposed an attention module that combines spatial attention mechanism and channel attention mechanism by focusing on the location information and channel information of the feature map. HOU *et al.* [14] proposed a coordinated attention (CA) mechanism by embedding position information into the channel attention mechanism on the basis of SENet. In computer vision tasks, the attention mechanism can be used to guide the model to pay attention to information that is beneficial to the task and ignore information that is not related to the task. Therefore, adding the attention mechanism can improve the expressive ability of the model. Inspired by the above ideas, we introduces the attention mechanism into the

model. The attention mechanism is used to guide the model to learn the weights of different regions and different channels of the fused feature layer. Thus, the effective information of the feature layer is fully utilized to improve the tracking effect.

## 3. Network Model

### 3.1. Structure of the Model

The multi-object pedestrian tracking model proposed in this paper mainly includes four parts: Backbone Network, Attention module, Detection branch and Re-ID branch. In the backbone network, in order to further integrate the feature information of feature layers D2, D3, and D4. Add the UFF module after the feature information of D2, D3, and D4 to obtain the context information of the three feature layers. An attention mechanism is added to the detection branch to guide the model to learn the importance of different information in the feature layer. The detection branch adopts three parallel sub-branches to predict the heat map of the pedestrian's center point, the size of the pedestrian, and the offset of the pedestrian's center point coordinate respectively. The Re-ID branch distinguishes different pedestrians by extracting the 128-dimensional apparent feature vector of the pedestrian center point. The network structure is shown in Figure 1. Then we will introduce the UFF module and the attention mechanism module respectively.

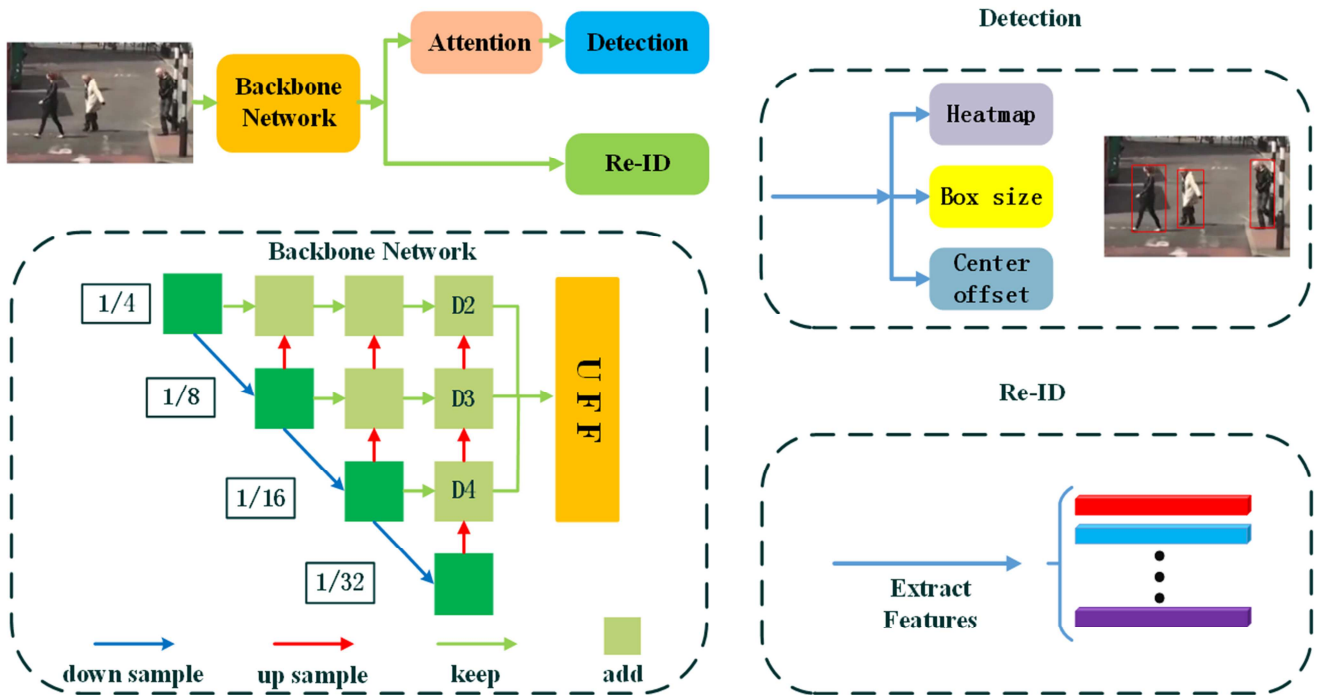


Figure 1. Structure of network model.

### 3.2. Structure of the UFF

As a non-rigid object, the shape of pedestrians changes all the time during the movement. In addition, the influence of factors such as mutual occlusion between pedestrians and

changes in light intensity brings serious challenges to the detector. To solve these problems, we proceed from two aspects. First, the geometric structure of the ordinary convolution kernel is fixed, and the receptive field size is the same in the same convolution layer. As a result, the

convolution kernel cannot adaptively adjust the receptive field area according to the scale of the pedestrian to accurately locate the position of the pedestrian in the feature layer. The deformable convolution proposed by DAI et al. [10] can accurately reflect the shape and position of the object in the feature map by learning the offset of the sampling point. Therefore, we use deformable convolution instead of ordinary convolution when adjusting the feature map channel. Second, the impact of occlusion and light intensity changes on tracking performance. We use convolution operations with different dilated rates to increase the receptive field of feature extraction, so as to obtain rich pedestrian feature information. In summary, we designed the UFF module. Its structure is shown in Figure 2. Then we will introduce its structure in detail.

The UFF module includes three stages: (a) Adjust the number of channels in the D2, D3, and D4 feature layers in the backbone network by deformable convolution (DC), so that the three feature layers have the same number of channels. (b) Through the linear interpolation operation, adjust the size of the D3 and D4 feature layers to make them the same size as the D2 feature layer. Then concatenate the three feature layers together along the channel direction. (c) For the feature layer after the concatenate in step (b), the depthwise separable convolution with a dilated rate of 1, 2, and 4 is used to capture the feature information of different receptive fields. Finally, the captured feature information is connected along the channel direction to form a fused feature layer.

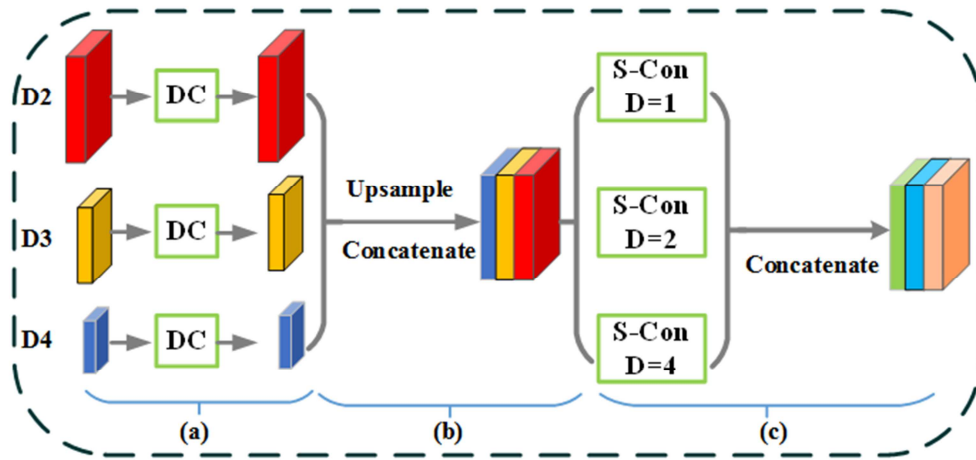


Figure 2. Structure of the UFF module, D represents dilation, S-Con represents depthwise separable convolution.

### 3.3. Coordinated Attention Mechanism

The feature layer extracted by the backbone network not only contains different pedestrian information in each area, but also each channel has a different importance. In order to allow the model to effectively use feature information when predicting pedestrians. This paper refers to the coordinated attention mechanism [14] to guide the model to learn the weights of different regions and different channels of the feature layer, and then adjust the weights of the feature layer according to the learned weights. The structure of the coordinated attention mechanism is shown in Figure 3. The module establishes the channel relationship and the spatial position information relationship through two steps: the embedding of coordinated information and the generation of coordinated attention. The embedding of coordination information is implemented as follows: for the input feature tensor  $X = [x_1, x_2, \dots, x_c] \in R^{C \times H \times W}$ , we use two spatial extents of pooling kernels  $(H, 1)$  and  $(1, W)$  to encode each channel along the horizontal coordinate and the vertical coordinate, respectively. Here  $W$ ,  $H$  and  $C$  are the width, height, and number of channels of the feature map, respectively. Thus, the output of the  $c$ -th channel at height  $h$  can be formulated as

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i) \quad (1)$$

Similarly, the output of the  $c$ -th channel at width  $w$  can be written as

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (2)$$

Coordinate Attention Generation:

Specifically, given the aggregated feature maps produced by Eqn. 1 and Eqn. 2, we first concatenate them and then send them to a shared  $1 \times 1$  convolutional transformation function  $F_1$ . Second, use normalization and activation operations to generate feature layer  $f \in R^{C/r \times (W+H)}$  ( $r$  is the reduction ratio in SENet, the value is 32). Therefore, the formula can be expressed as

$$f = \delta \left( F_1 \left( \left[ z^h, z^w \right] \right) \right) \quad (3)$$

Where  $[\cdot, \cdot]$  denotes the concatenation operation along the spatial dimension,  $\delta$  is a non-linear activation function.

Subsequently,  $f$  is divided into two tensors  $f^h \in R^{C/r \times H}$  and  $f^w \in R^{C/r \times W}$  along the spatial dimension. Another two  $1 \times 1$  convolutions  $F_h$  and  $F_w$  to act on  $f^h$  and  $f^w$  respectively to generate attention weight maps  $g^h$  and  $g^w$  with the same number of channels as the input. The formula can be expressed as

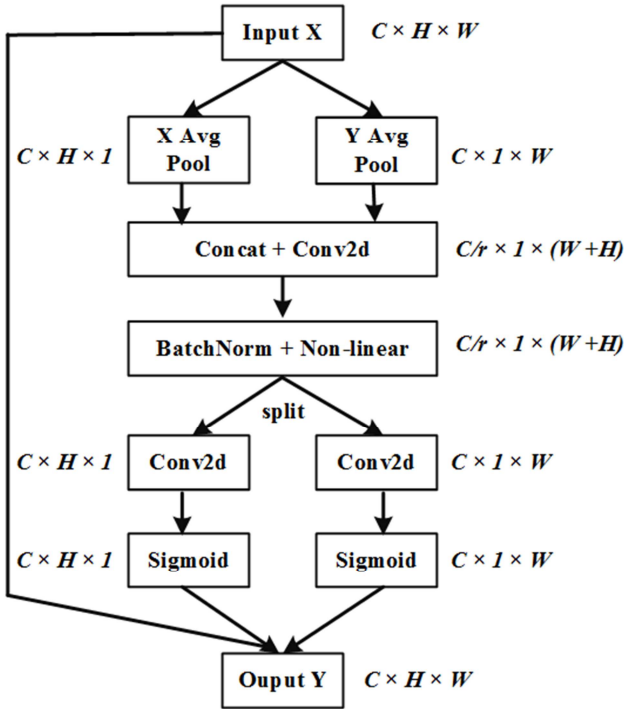
$$g^h = \sigma(F_h(f^h)) \quad (4)$$

$$g^w = \sigma(F_w(f^w)) \quad (5)$$

Where  $\sigma$  is the sigmoid function. The expression of the output result  $Y = [y_1, y_2, \dots, y_c]$  is:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (6)$$

The weight of the fused feature layer is adaptively adjusted through the coordinated attention mechanism module. Then the weighted feature layer is sent to the detection branch for pedestrian classification and regression.



**Figure 3.** Structure of the coordinate attention mechanisms. ‘X Avg Pool’ and ‘Y Avg Pool’ refer to 1D horizontal global pooling and 1D vertical global pooling, respectively.

## 4. Experimental

### 4.1. Experimental Setup and Dataset

The experimental environment in this paper is Windows 10 operating system, the hardware platform is configured with Intel Xeon W-2245 processor, 16G running memory, and a

workstation equipped with RTX 5000 GPU. The Adam optimizer was used to optimize the model during model training. In the experiment, the model was trained for total of 30 epochs, and the batch size was set to 8. The initial learning rate is  $10^{-4}$ , and the learning rate is adjusted to  $10^{-5}$  after the 20 epoch.

Our algorithm uses the model structure of joint detection and Re-ID. In order to prevent our algorithm from achieving better results on individual datasets, when the algorithm is applied to large-scale datasets, the tracking effect is not ideal. We combined six pedestrian datasets into one large dataset. These six datasets can be divided into two categories: the first type of dataset contains pedestrian identification and pedestrian location annotations, and the second type contains only pedestrian location annotations. Among them, CalTech (CT) [15], MOT17 (M17) [16], CUHK-SYSU (CS) [17], PRW [18] belong to the first type of dataset. ETH [19] and CityPerson (CP) [20] belong to the second type of dataset. Our training set is a combination of the training sets of these six datasets. Table 1 shows the statistical distribution of the joint training set.

**Table 1.** Statistics of the joint training set.

Dataset	ETH	CP	CT	M17	CS	PRW	Total
img	2K	3K	27K	53K	11K	6K	54K
box	17K	21K	46K	112K	55K	18K	270K
ID	-	-	0.6K	0.5K	7K	0.5K	8.7K

### 4.2. Evaluation Standard

In order to evaluate the tracking performance of the model, we choose to verify on the MOT16 [16] dataset. The standard MOT challenge evaluation mechanism is used to evaluate the tracking performance of the model. The meaning of each evaluation measure is shown in Table 2, where  $\uparrow$  means higher is better,  $\downarrow$  means lower is better.

**Table 2.** Evaluation index and meaning of the model.

Measure	Description
MOTA $\uparrow$	Multi-Object Tracking Accuracy
ML $\downarrow$	Mostly lost objects
MT $\uparrow$	Mostly tracked objects
IDS $\downarrow$	Number of Identity Switches
IDF1 $\uparrow$	The ratio of correctly identified detections over the average number of ground-truth and computed detections

### 4.3. Experimental Results and Analysis

In order to effectively verify the improvement of pedestrian tracking results by the UFF module and coordinated attention mechanism proposed in this paper. The FairMOT algorithm is reproduced in the same experimental environment, and the result is recorded as FairMOT (re). In order to eliminate the interference of other factors during the recurrence, only the basic model framework of the original algorithm is used, and irrelevant training skills are omitted.

First, verify our proposed UFF module that integrates context information. In order to prove that the module can obtain rich context information and improve the performance

of pedestrian tracking, DLA-34 and HRNet-W18 [21] are selected as the basic network structure. In the MOT16 verification set, the effect of adding UFF modules to the two networks on the performance of multi-object pedestrian tracking was compared. It can be seen from Table 3 that after adding UFF modules to the two networks. Both MOTA and IDF1 have improved, and IDSw has been significantly reduced during tracking. The experimental results show that adding UFF module can improve pedestrian tracking performance. The main reason is that the UFF module enriches the context information of the feature layer by fusing the feature layers of different scales and adopting the separable convolution operation with the depth of different dilated ratios. Thereby improving the performance of multi-object pedestrian tracking.

**Table 3.** Comparison experiment of UFF on different backbone networks.

Backbone	UFF	MOTA↑	IDF1↑	IDSw↓
DLA-34		85.81	84.97	447
DLA-34	√	86.32	85.90	425
HRNet-W18		83.49	82.08	560
HRNet-W18	√	83.59	82.92	506

Second, we explored the influence of the CA module on the effect of multi-object pedestrian tracking. In order to explore the improvement of tracking performance by the different combination of CA module and model. The CA modules were added to the Backbone Network (BN), the Re-ID branch and the Detection branch (DB) respectively, and comparative experiments were performed on the MOT16 verification set. It can be seen from Table 4 that when the CA module is added in Backbone Network, the number of ID switching times during pedestrian tracking is the lowest. When the CA module is added to the detection branch, the pedestrian tracking MOTA and IDF1 are the highest. MOTA is 87.21. The number of ID switching times for pedestrian tracking is reduced by 8.5% compared to the original model. Three ways of adding CA modules can improve the performance of pedestrian tracking. It can be concluded that after adding the CA module to the model. The model can automatically pay attention to the effective area and effective information of the feature layer. Therefore, the CA module can improve the expressive ability of the model.

**Table 4.** Ablation experiment of CA module.

BN	Re-ID	DB	MOTA↑	IDF1↑	IDSw↓
			85.81	84.97	447
√			86.97	85.70	385
	√		87.12	83.99	470
		√	87.21	86.60	409

In addition, we performed corresponding ablation experiments on the UFF module and the CA module on the MOT16 verification set, and explored the impact of each module on the performance of multi-object pedestrian tracking. As shown in Table 5, adding the UFF module and

the CA module separately can not only improve MOTA and IDF1, but also reduce IDSw. When the UFF module and the CA module are added to the model together. The values of MOTA and IDF1 have increased the most significantly. Their values are 87.21 and 86.6, respectively. Experimental results show that adding UFF module can effectively fuse context information. Adding a CA module can guide the model to focus on effective information areas and improve the expressiveness of the model. In the end, compared with FairMOT, our algorithm improved by 1.4 and 1.63 on MOTA and IDF1 respectively, and the number of times of pedestrian ID switching during the tracking process was reduced by 8.5%.

**Table 5.** Ablation experiment of UFF module and CA.

UFF	CA	MOTA↑	IDF1↑	IDSw↓
		85.81	84.97	447
√		86.32	85.90	425
	√	86.90	85.81	401
√	√	87.21	86.60	409

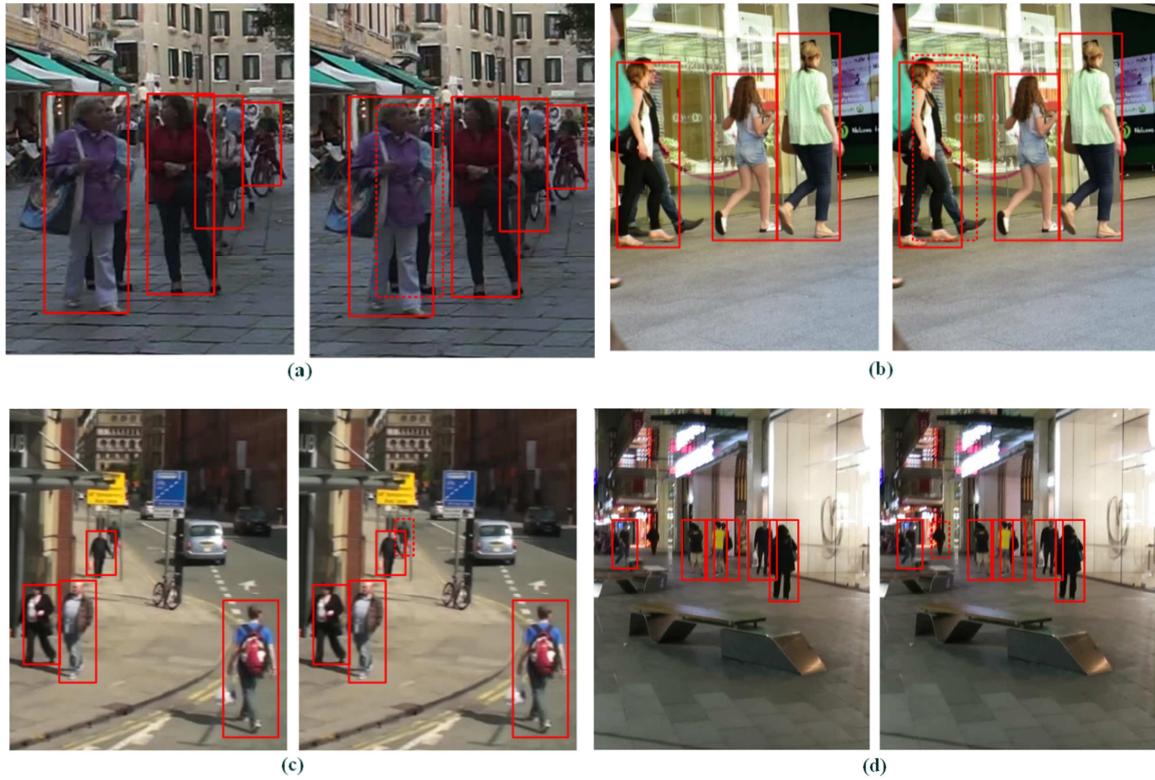
Table 6 shows the tracking results of our model and other pedestrian tracking algorithms on the MOT16 test set. It can be seen from Table 6 that the ML is only slightly lower than the TubeTK [22] algorithm in this paper, but good results have been achieved in other evaluation measure. Under the same experimental conditions, each evaluation measure is better than the FairMOT algorithm, which proves that the UFF module and attention mechanism are beneficial to improve the pedestrian tracking performance.

**Table 6.** Comparison of tracking algorithms on MOT16 test set.

Method	MOTA↑	IDF1↑	MT↑	ML↓	IDSw↓
TubeTK [22]	64.0	59.4	33.5%	19.4%	1117
CNNMTT [23]	65.2	62.2	32.4%	21.3%	946
CTracker [24]	67.6	57.2	32.9%	23.1%	1897
FairMOT (re)	69.52	70.72	34.78%	21.87%	924
Ours	70.74	70.95	36.50%	21.87%	824

The visualization of pedestrian detection results of our algorithm and FairMOT on the MOT16 verification set is shown in Figure 4. The detection results of pedestrians in crowded conditions are shown in 4(a) and 4(b). The left side is the FairMOT detection result, and the right side is our detection result. It can be seen from 4(a) and 4(b) that our algorithm can accurately detect the pedestrians that FairMOT missed (The dashed box in the picture represents the pedestrians that FairMOT missed, which was accurately detected by our algorithm). The detection results of small-scale pedestrians are shown in 4(c) and 4(d). It can be seen from 4(c) and 4(d) that the pedestrians missed by FairMOT (left) can be accurately detected by our algorithm (right). The accurate detection of pedestrians lays the foundation for the improvement of tracking performance. Therefore, the effectiveness of our algorithm improvement can be proved.





**Figure 4.** The visualization comparison between ours and FairMOT on MOT16 validation set. The left of each subgraph of (a), (b), (c) and (d) is the detection result of FairMOT, and the right is the detection result of ours. The dashed line boxes represent the pedestrians missed by FairMOT, which are accurately detected by ours.

## 5. Conclusion

This paper is improved on the basis of FairMOT algorithm. We propose a UFF module that integrates contextual information. The deformable convolution is used to learn the offset of the sampling point when extracting the pedestrian feature, so that it can adaptively adjust the position of the sampling point to accurately locate the pedestrian's position in the feature map. Use the depthwise separable convolution of different dilated rates to obtain the context information of different receptive fields, and enhance the model's feature extraction ability for different receptive fields. Finally, it effectively improves the tracking accuracy when pedestrians are occluded or small in size. In addition, we introduce a coordinated attention mechanism into the model. Guide the model to pay attention to the effective information of the feature layer and ignore irrelevant information to improve the expressive ability of the model. Experimental results show that our improved algorithm can achieve higher MOTA, while reducing the number of ID switching during pedestrian tracking. It has better tracking performance. Our next work will optimize the current model. Let the detection branch and the Re-ID branch independently learn the characteristic information of the backbone network, so that the finally learned characteristic information can adapt to its own task. To further improve the tracking performance of pedestrians.

## Acknowledgements

This work is supported by the National Natural Science Foundation Grant 61802454 of China.

## References

- [1] Claparrone G, Sanchez F L, Tabik S. Deep learning in video multi-object tracking: A survey [J]. *Neurocomputing*, 2020, 381: 61-88.
- [2] Zhang Y, Lu H Z, Zhang L P. Overview of Visual Multi-object Tracking Algorithms with Deep Learning [J]. *Computer Engineering and Applications*, 2021, 57 (13): 55-66.
- [3] Voigtlaender P, Krause M, Osep A. Mots: Multi-object tracking and segmentation [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE Press, 2019: 7942-7951.
- [4] Wang Z, Zheng L, Liu Y. Towards real-time multi-object tracking [C]//*European Conference on Computer Vision*. Glasgow: Springer, 2020: 107-122.
- [5] Zhang Y, Wang C, Wang X. Fairmot: On the fairness of detection and re-identification in multiple object tracking [J]. *International Journal of Computer Vision*, 2021: 1-19.
- [6] Zhou X, Koltun V, Krahenbuhl P. Tracking objects as points [C]//*European Conference on Computer Vision*. Springer, Cham, 2020: 474-490.

- [7] Szegedy C, Loffe S, Vanhoucke V. Inception-v4, inception-resnet and the impact of residual connections on learning [C]//The AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017.
- [8] Liu W, Lei H, Xie H. Multi-level Light U-Net and Atrous Spatial Pyramid Pooling for Optic Disc Segmentation on Fundus Image [C]//International Workshop on Ophthalmic Medical Image Analysis. Springer, Cham, 2020: 104-113.
- [9] Liu S, Huang D, Wang Y. Receptive field block net for accurate and fast object detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 385-400.
- [10] Dai J, Qi H, Xiong Y. Deformable convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2017: 764-773.
- [11] Haase D, Amthor M. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved Mobile Nets [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 14600-14609.
- [12] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks [C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE Press, 2018: 7132-7141.
- [13] Park J, Woo S, Lee J Y. A simple and light-weight attention module for convolutional neural networks [J]. International Journal of Computer Vision, 2020, 128 (4): 783-798.
- [14] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 13713-13722.
- [15] Dollar P, Wojek C, Schiele B. Pedestrian detection: A benchmark [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2009: 304-311.
- [16] Milan A, Leal-taixel L, Reid I. MOT16: A benchmark for multi-object tracking [J]. arXiv preprint arXiv: 1603. 00831, 2016.
- [17] Xiao T, Li S, Wang B. Joint detection and identification feature learning for person search [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 3415-3424.
- [18] Zheng L, Zhang H, Sun S. Person re-identification in the wild [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 1367-1376.
- [19] Ess A, Leibe B, Schindler K. A mobile vision system for robust multi-person tracking [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2008: 1-8.
- [20] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 3213-3221.
- [21] Cheng B, Xiao B, Wang J. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation [C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 5386-5395.
- [22] Pang B, Li Y, Zhang Y. Tubetk: Adopting tubes to track multi-object in a one-step training model [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6308-6318.
- [23] Mahmoudi N, Ahadi S M, Rahmati M. Multi-target tracking using CNN-based features: CNNMTT [J]. Multimedia Tools and Applications, 2019, 78 (6): 7077-7096.
- [24] Peng J, Wang C, Wan F. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking [C]//European Conference on Computer Vision. Springer, Cham, 2020: 145-161.