Lightweight high-precision pedestrian tracking algorithm in complex occlusion scenarios

Qiang Gao[1], Zhicheng He[2*], Xu Jia[3], Yinghong Xie[2] and Xiaowei Han[1]
[1] Institute of Science and Technology Innovation, Shenyang University
Shenyang, CHN
[e-mail: tommy_06@163.com]
[2] School of Information Engineering, Shenyang University
Shenyang, CHN
[e-mail: 471516959@qq.com]
[3] School of Information Engineering, Liaoning University of Technology
Shenyang, CHN
[e-mail: jiaxu@lnut.edu.cn]
[*]Corresponding author: Zhicheng He

Received July 26, 2022; revised January 27, 2023; accepted February 12, 2023; published March 31, 2023

Abstract

Aiming at the serious occlusion and slow tracking speed in pedestrian target tracking and recognition in complex scenes, a target tracking method based on improved YOLO v5 combined with Deep SORT is proposed. By merging the attention mechanism ECA-Net with the Neck part of the YOLO v5 network, using the CIoU loss function and the method of CIoU non-maximum value suppression, connecting the Deep SORT model using Shuffle Net V2 as the appearance feature extraction network to achieve lightweight and fast speed tracking and the purpose of improving tracking under occlusion. A large number of experiments show that the improved YOLO v5 increases the average precision by 1.3% compared with other algorithms. The improved tracking model, MOTA reaches 54.3% on the MOT17 pedestrian tracking data, and the tracking accuracy is 3.7% higher than the related algorithms and The model presented in this paper improves the FPS by nearly 5 on the fps indicator.

Keywords: object tracking; attention mechanism; object detection; non-maximum suppression; lightweight neural network

http://doi.org/10.3837/tiis.2023.03.009 **ISSN : 1976-7277**

1 Introduction

Multi-target recognition and tracking is a research hotspot in the field of deep learning, and it has extensive research and scientific research value in many engineering practices. In actual complex scenes, there are still many difficulties in multi-target tracking, such as mutual occlusion between pedestrians caused by crowd congestion, frequent changes of pedestrian poses, missed detections, false detections caused by small target pedestrians in the distance, inaccurate target bounding boxes etc. Therefore, it is a challenging task to propose a robust online multi-target tracking algorithm with strong generalization ability.

In the multi-target tracking task, the appearance, color, texture and other features of the target are extracted for correlation matching. There are currently two popular multiple target tracking methods, the multiple target tracking based on filtering algorithm and the multiple target tracking based on deep learning. The commonly used multiple target tracking method based on filtering algorithms include correlation filtering[1,2], particle filtering[3,4], Kalman filtering [5-7]and so on. Such methods are prone to target loss when the target is occluded.

Compared with filtering algorithms, tracking methods using deep learning can achieve relative balance and improvement in the tracking precision and speed. Such methods are roughly divided into models based on joint target detection and tracking and models based on graph convolutional neural networks. The SORT algorithm proposed by Bewley uses convolutional neural network as target detection network instead of traditional detector, and combines specific tracker and appropriate matching algorithm to greatly improve the speed and precision of multiple target tracking[8]. Wojke et al., “used REID pedestrian re-recognition network to extract the apparent features of the target and predict the target position of the next frame in combination with the target location features, thus improving the problem of target ID interchange in SORT algorithm,this efficient method show that the number of identity switches have been reduced, achieving high frame rates [9]. Wang Z's team proposed a MOT model with shared structure combining object detection and embedded appearance features. This method can effectively improve the efficiency of MOT system by experimental results [10]. This model incorporates the embedded appearance features into the single-shot detector, and chooses Feature Pyramid Network (FPN) as the underlying framework, so that the running speed and tracking accuracy of the algorithm are comparable to the SOTA model using independent detection and embedding learning.Zhang et al pointed out that existing methods have problems such as how to share motion calculation on an infinite number of targets and how to fuse tracking targets and detection targets. An end-to-end DNN tracer method, which is called FFT was designed to solve the above problems through two effective techniques: target flow and target fusion. [11] Chu et al., “proposed a novel tracking network-FAMNet. All layers in the network model are differentiable, so that joint optimization can be used to learn the distinguishing features of robust MOTs and high-order affinity and force modes, which are directly supervised by the loss of the distributed ground facts.[12]The first network framework for computing object affinity proposed by Sun et al., deep affinity network DAN, learns pre-detected objects at multiple levels of abstractionIn order to infer the affinity of objects, the characteristics are arranged by exhaustive pairing in any two frames.[13] Xingyi Zho proposed a simultaneous detection and tracking algorithm, CenterTrack[14]. Given a small input, the detector infers objects by extracting heatmaps, correlating them between frames, and the tracker locates objects and predicts their relative Associating a frame, CenterTrack can easily learn to repeat the predictions from the previous frame, thereby refusing to track those cases that would cause large training errors. Jialian Wu et al., “proposed a new model which is called TraDeS[15]. This is a online object tracking algorithm, which used tracking cues to assist object detection. This model infers offsets for object tracking, with computational cost, for propagating previous object features to improve current object detection and segmentation. In this way, this framework can simultaneously complete detection, multi-target tracking, instance multi-target tracking and segmentation.Zhang et al.[16] proposed the SiamDW algorithm, which eliminates the negative impact of the deep network by designing a residual structure, and at the same time adjusts the step size and receptive field of the backbone network to ResNetThe twin network was introduced and verified by experiments, and better performance than the original model was achieved.Fan et al., “propose a C-RPN using a cascading regional recommendation network (RPN). [17]. C-RPN is used in a multi-layer RPN network, and the anchors belonging to the negative sample are selected layer by layer, and the anchors of the model are input to the next layer as positive samples.RPN networks can achieve more robust performance in complex backgrounds such as similar semantic obstacles.Yu et al., “[18]perform self-attention operations in SiamAttn's target branch and search branch respectively to achieve attention to channels and special locations. Search for branches and targetsCross-attention calculations are performed between branches so that the search branch learns the target information. In addition, similar twin network models such as TranT [19]and STARK[20] can show good results.Qiang Wang et al., “proposed to use a local correlation module to model the topological relationship between the target and the surrounding environment [21], the modified model can effectively improve the pedestrian identification ability in complex and occlusion scenes. Jiawei He proposed a new learnable graph matching method for target tracking task, which turns the association problem into an undirected graph matching problem[22]. It effectively solves the problem of inconsistency between training and reasoning caused by feature extraction of single neural network.

The above algorithms all have good performance in tracking but there is still room for improvement. Therefore, this paper combines the above work and proposes to use the one-stage fast target detection model YOLO v5 as the detector, combined with Deep SORT's pedestrian target recognition and tracking algorithm, for lightweight detection and tracking. The innovation research of this paper mainly includes three aspects as follows:

(1) Replacing the feature extraction network of Deep SORT with Shuffle Net, reduce the parameters of the network model to achieve the effect of lightweight tracking

(2) Aiming at the problem of low target recognition accuracy, the original SPP module is first replaced with the Atrous Spatial Pyramid Pooling Module (ASPP), and the attention mechanism ECA-Net is integrated with the detection network to enhance the model's ability to extract features, so that the model pays more attention to the detected target.

(3) Using the non-maximum suppression method of CIoU-NMS, combined with the CIoU loss function,the problem of missed detection and ID jump caused by occlusion during detection and tracking is solved.

2. The model proposed in this paper

In order to improve the pedestrian tracking speed and the problem of low detection precision caused by large-area occlusion of pedestrians in complex scenes, this paper proposes a YOLO v5 algorithm fused with attention mechanism combined with Deep SORT to efficiently track pedestrians. First, on the detector side, the ECA-Net attention mechanism and the spatial hole pyramid convolution module are used to improve the feature extraction ability of the network and reduce the loss of contextual information due to upsampling. The non-maximum suppression method of CIoU is used to improve the missed detection caused by large-area occlusion and the ID jump problem during tracking. In terms of tracker, Shuffle Net V2 is selected as the Deep SORT appearance feature extraction network, which greatly reduces the model parameters, improves the tracking speed, and achieves the effect of lightweight tracking. The overall model framework is shown in **Fig. 1** below.

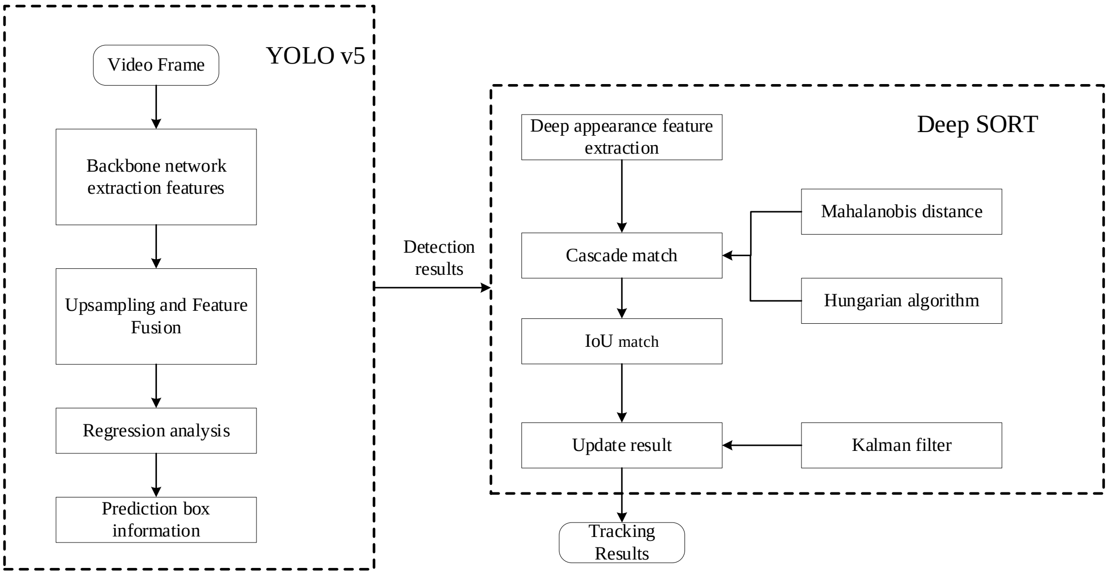


Fig. 1. Overall model framework

2.1 YOLO v5 model with attention mechanism

YOLO V5 consists of four parts: input part, CSP backbone feature extraction network, feature fusion network and output. The input part mainly includes data preprocessing, including Mosaic data enhancement, adaptive image filling and anchor frame calculation to automatically size the initial anchor frame when changing data sets. The CSP feature extraction network extracts features at different levels from images through deep convolution operations, combined with spatial pyramid pooling SPP [23], to reduce model parameters and increase model reasoning speed and precision. The feature fusion network layer includes the feature pyramid FPN and the path aggregation structure PAN [24], which realizes the information fusion of each network layer in the backbone feature extraction network and further improves the object detection capability. The output part is used to predict targets of different sizes on feature maps of different sizes. The YOLO v5 network structure is shown in **Fig. 2**.

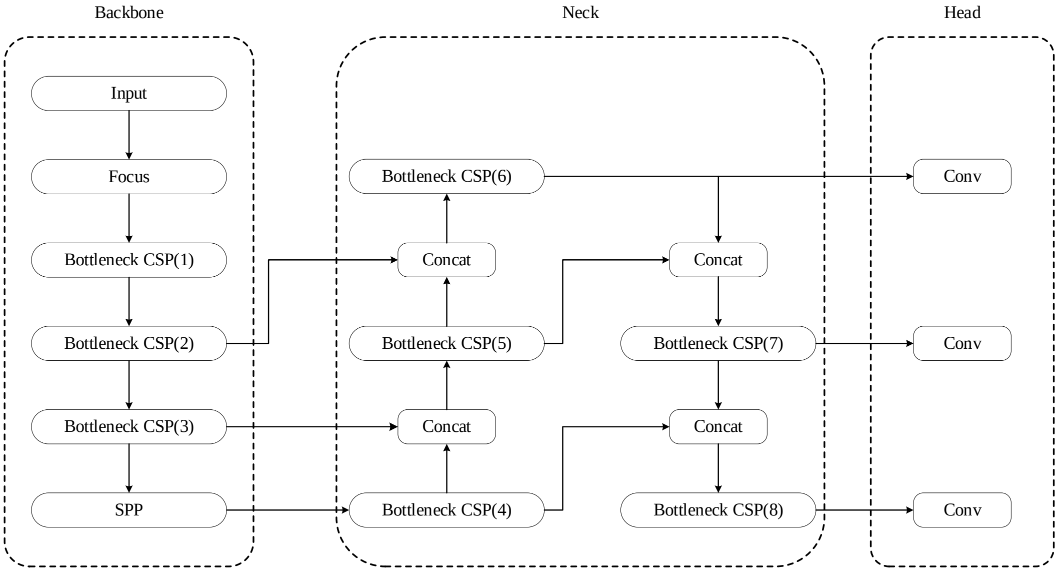


Fig. 2. YOLO V5 network structure

2.1.1 ECA-Net attention mechanism

As a new generation of target detection network, YOLO v5 has greatly improved detection speed and precision compared with other YOLO series networks, but considering the complexity of the environment, the original YOLO v5 does not work in crowded scenes and when the characteristics of distant pedestrians are small.so this paper proposes to introduce the ECA-Net lightweight attention mechanism [25] in the Neck feature fusion part. The improved structure is shown in **Fig. 3**.

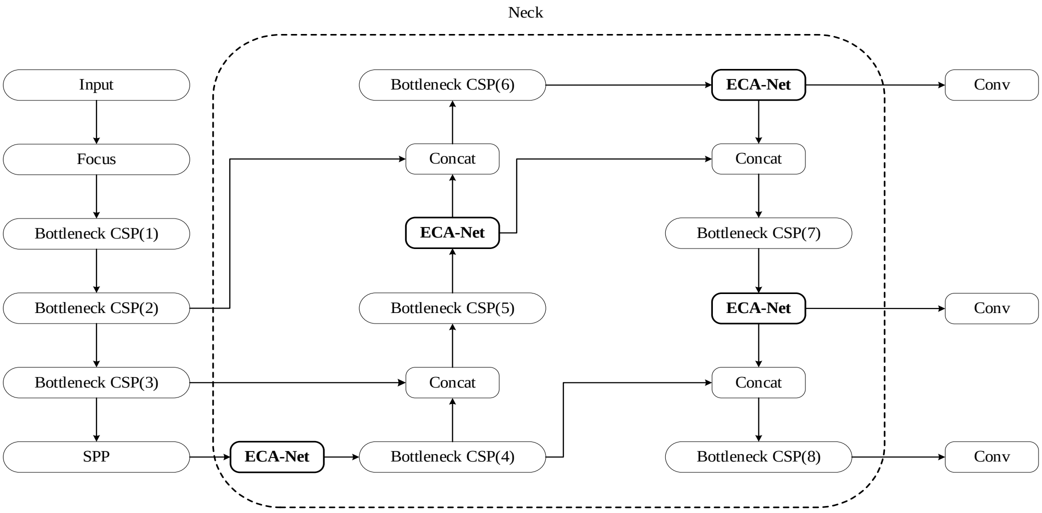


Fig. 3. ECA-YOLO V5 network structure

ECA-Net is an efficient channel attention mechanism. Compared with other model, this module effectively solves the problem of dimensionality reduction of neural network and can effectively collect the information of channel interaction between network layers. The ECA-Net is shown in **Fig. 4**. After ECA-Net performs channel-by-channel global average pooling without reducing the network dimension, ECA collects channel interaction information by considering each network layer channel and its K neighbors. This process can be effectively implemented by one-dimensional convolution with convolution kernel size k, where the parameter K is determined by the coverage of the local channel interaction information, that is, the number of neighboring neurons participating in the attention prediction of a channel.

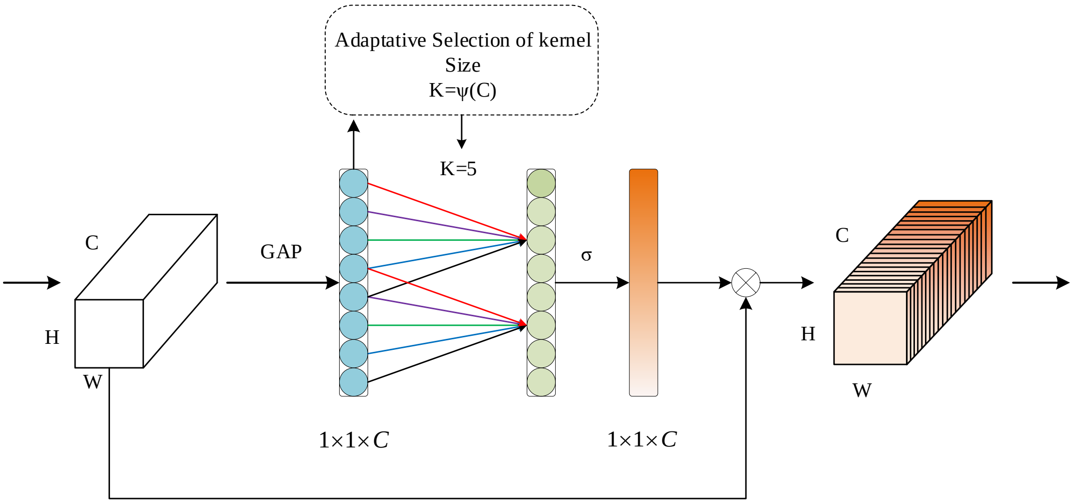


Fig. 4. ECA-Net structure

Since the ECA-Net module only involves k parameters, and the number of parameters is smaller than SE-Net, the ECA module does not add additional computational cost to the network, and the model complexity is reduced.

2.1.2 Spatial Hole Pyramid Pooling

The tail of YOLO V5 trunk feature extraction network uses the spatial pyramid pooling module to realize feature extraction of the same feature map at different scales, which is helpful to increase the detection precision. To further reduce the loss of network parameters and location information during downsampling, the original SPP module is replaced by the Spatial Hole Pyramid Pooling (ASPP) module[26]to improve the detection accuracy in this paper. The improved YOLO V5 structure is shown in **Fig. 5**.

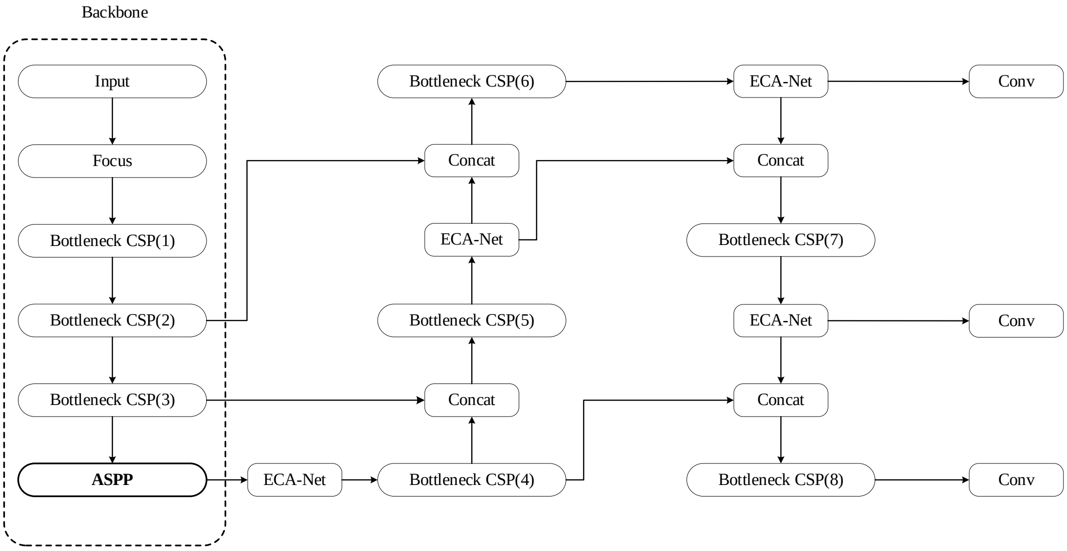


Fig. 5. Introducing the YOLO v5 model of ASPP

The Atrous Spatial Pyramid Pooling (ASPP) module consists of multiple parallel void convolution layers with different sampling rates. Each branch uses different sampling rates to extract image features. Further feature processing is carried out in this branch. After processing, concat fusion is performed to generate the final result. This module uses convolution kernels with different dilation rates to construct receptive fields of different sizes to obtain multiple scale information of images. The specific structure is shown in **Fig. 6** below:

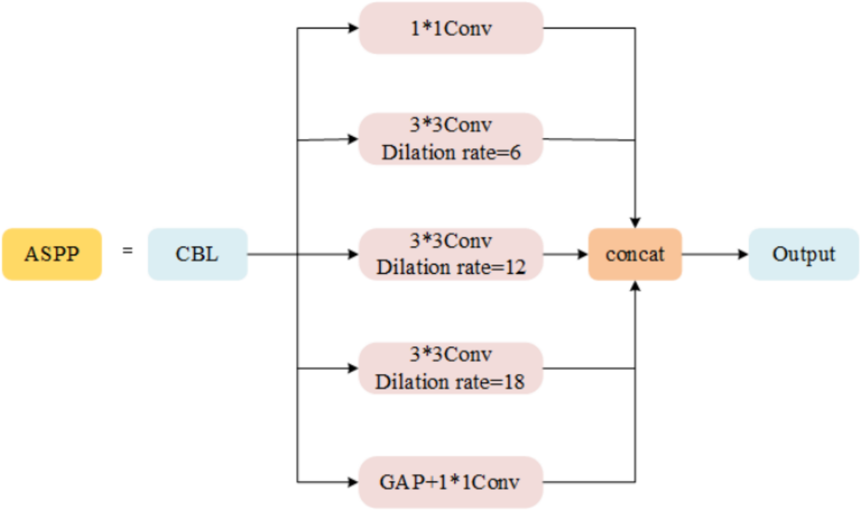


Fig. 6. Spatial Hole Pyramid Pooling Module

The depthwise convolution in the original depthwise separable convolution is replaced by a holey convolution to form a holey depthwise separable convolution in this paper. And by changing the hole rate of the hole convolution to change the receptive field in the pooling process, the amount of parameters is reduced and the calculation speed is improved.

2.1.3 Non-maximum suppression method combined with CIoU

In the prediction stage, the non-maximum suppression (NMS) method is usually used to remove redundant detection boxes. The criterion is the IoU ratio of the intersection of the prediction box and the detection box with the highest score., if the IoU value is better than the given threshold value , the prediction box will be removed. In general scenes, this method is effective, but in scenes with crowded targets and serious occlusion, due to the mutual occlusion of each target, the detection boxes of different objects are very close, and the overlapping area is large, so the prediction frame will be iremoved incorrectly, causing object detection fails. CIoU[27] considers the overlap area, center distance and length-width ratio of predicted and real boundary boxes, which can effectively improve the IoU error removal defect when the predicted box overlaps with the real box. CIoU is shown in Equation (1).

$$CIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} - \alpha v \tag{1}$$

where b and b^{gt} represent the center points of the predicted bounding box and the true bounding box, ρ^2 represents the Euclidean distance between the two points, and c represents the diagonal length of the two boxes. α represents a balance parameter, and v represents the consistency of the aspect ratio of the two boxes. The formula is shown in (2)

$$\begin{cases} v = \dfrac{4}{\pi^2}(\arctan\dfrac{w^{gt}}{h^{gt}} - \arctan\dfrac{w}{h})^2 \\ \alpha = \dfrac{v}{(1-IoU)+v} \end{cases} \tag{2}$$

where w^{gt}, h^{gt}, w, h represents the width and height of the ground-truth bounding box and the predicted bounding box, respectively. At the same time, considering that the GIoU in the original network heavily depends on the IoU item, resulting in a slow convergence rate in actual training and low accuracy of the predicted bounding box, This paper considers replacing the GIoU loss function with the CIoU loss function which the formula is shown in (3):

$$L_{CIoU} = 1 - CIoU \tag{3}$$

2.2 Lightweight target tracking model

The Deep SORT algorithm[28]uses the Kalman filter to process data correlation frame by frame and uses the Hungarian algorithm to perform target screening and cross-frame matching on the output of the detector. On the basis of the SORT algorithm, the apparent feature extraction network performs the re-recognition task of pedestrians. The apparent features of the tracking target are extracted for nearest neighbor matching, and the ID switch problem caused by occlusion is improved. The overall framework of the algorithm is shown in the following **Fig. 7**:

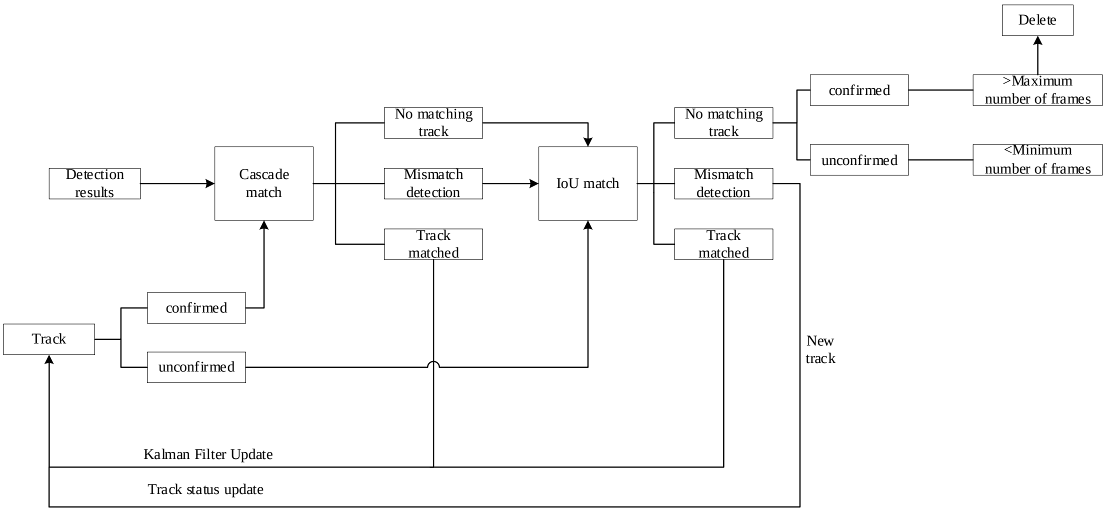


Fig. 7. The overall framework of the Deep SORT algorithm

2.2.1 Lightweight apparent feature extraction network

The DeepSORT algorithm uses a simple ReID network to extract the features of the bojectand trains the model on a large pedestrian re-recognition dataset, making it suitable for pedestrian detection and tracking. In order to speed up feature extraction, reduce the parameters of the network model, and improve the overall speed of tracking, this paper replaces the original residual convolutional neural network of Deep SORT with Shuffle Net V2 [29], and its overall structure is as follows **Fig. 8**:

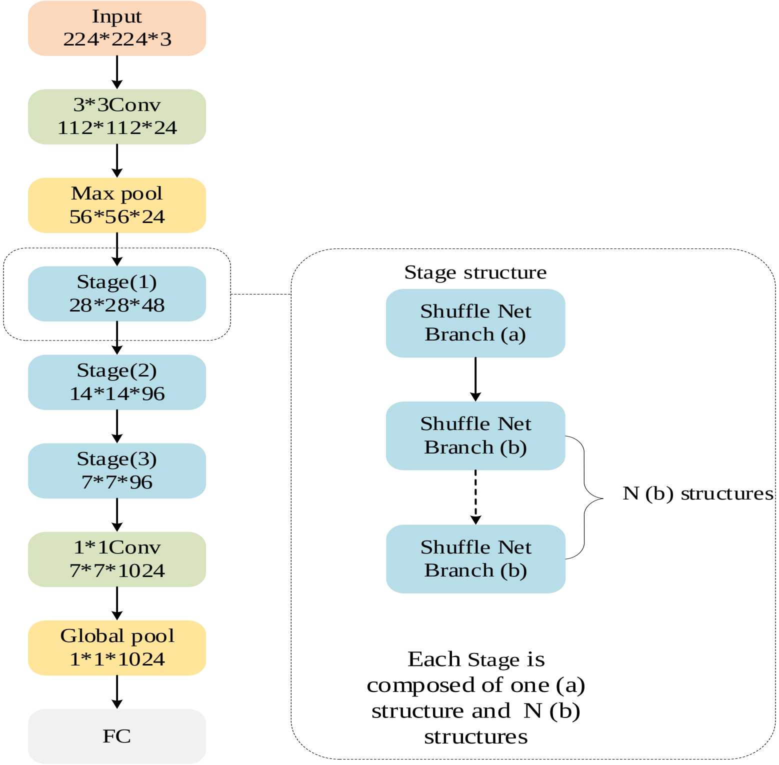
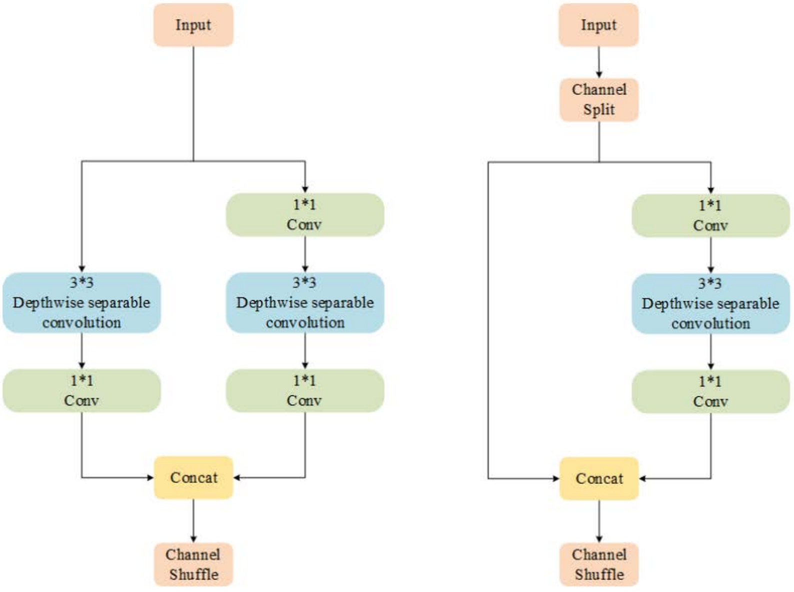


Fig. 8. Shuffle Net V2 overall structure

The branch structure (a) and branch structure (b) are shown in **Fig. 9**


Input
1*1 Conv
3*3 Depthwise separable convolution
3*3 Depthwise separable convolution
1*1 Conv
1*1 Conv
Concat
Channel Shuffle
Input
Channel Split
1*1 Conv
3*3 Depthwise separable convolution
1*1 Conv
Concat
Channel Shuffle

(a) (b)

Fig. 9. Shuffle Net V2 branch structure

The network structure **Fig. 9** (a)performs convolution operations on the left and right sides at the same time, the purpose is to increase the network depth and perform downsampling feature extraction. The channel split in the network structure **Fig. 9** (b) is to divide the number of input channels into two branches to replace the original grouped convolution structure, and the convolution layer in each branch keeps the same number of input and output channels, and the left branch does not take any operation. The purpose is to reduce the model parameters after convolution, and finally, concat the left and right branches by channel, and then channel shuffle is added to increase the information exchange between channels. In this way, the two structures (a) and (b) are used in combination to form the whole Shuffle Net. The overall Shuffle Net V2 network parameters are shown in **Table 1** below:

Table 1. Applications in each class

Network layer	kernel size	output size	Number of channels
Input	----	224x224	3
Conv Max pool	3x3 3x3	112x112 56x56	24
Stage1	structure (a) *1 structure (b) *3	28x28	48
Stage2	structure (a) *1 structure (b) *7	14x14	96
Stage3	structure (a) *1 structure (b) *3	7x7	192
Conv	1x1	7x7	1024
Global Pool	7x7	1x1	1024
FC	----	----	----

Note: The structures of (a) and (b) in the table are shown in **Fig. 9**

2.2.2 Association and Cascade Matching

The tracking algorithm data association uses the combination of target motion information and feature information to improve the accuracy of the association The Mahalanobis distance is used to match the Kalman predicted value with the actual measured value, the covariance matrix is normalized, the uncertainty of the state estimation of the detection and the average orbit deviation evaluation is calculated, and the motion information matching is realized, such as the formula:

$$d^{(1)}\left(i,j\right)=(d_j-y_i)^T \mathrm{S}_i^{-1}(d_j-y_i) \tag{4}$$

where d_j and y_i represent the state vector between the jth detection result and the ith prediction result, and S_i represents the covariance matrix between the detection result and the average tracking result. The Mahalanobis distance measures the standard deviation of the detection results from the average tracking results, taking into account the uncertainty of the state estimation, which can exclude low-probability associations. When the uncertainty of target motion information is low, Mahalanobis distance is a suitable correlation factor, but when the target is occluded or the camera perspective is shaken, only using Mahalanobis distance association will lead to target identity switching. Therefore, consider adding appearance information, calculate the corresponding appearance feature descriptor r_j for each detection frame, and set up $\left\|r_j\right\|=1$ a feature warehouse for each tracking trajectory k, which is used to save the feature descriptors $R_k=\left\{r_k^{(i)}\right\}_{k=1}^{L_k}$ that are successfully associated with the last 100 targets, $L_k=100$,and calculate the i-th feature descriptor. The minimum cosine distance between the tracking frame and the jth detection frame is as follows:

$$d^{(2)}(i,j)=\min\left\{1-r_k^{(i)} \mid r_k^{(i)}\in R_i\right\} \tag{5}$$

If $d^{(2)}(i,j)$ is less than the given threshold, the association is considered successful. Martens distance provides reliable target position information in short-term prediction cases, and the cosine similarity of appearance features can be used to recover the target ID when the target occlusion reappears, and in order to make the advantages of the two measures complement each other, a linearly weighted approach is used to as follows :

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j) \tag{6}$$

The information is then entered into the matching algorithm. The matching algorithm first uses cascading matching to obtain a preliminary matching pair, unmatched trajectory and unmatch detection, the framework of cascaded matching is shown in the following **Fig. 10**, and then the remaining unmatched objects are matched again through the IoU matching process to complete the entire tracking process.

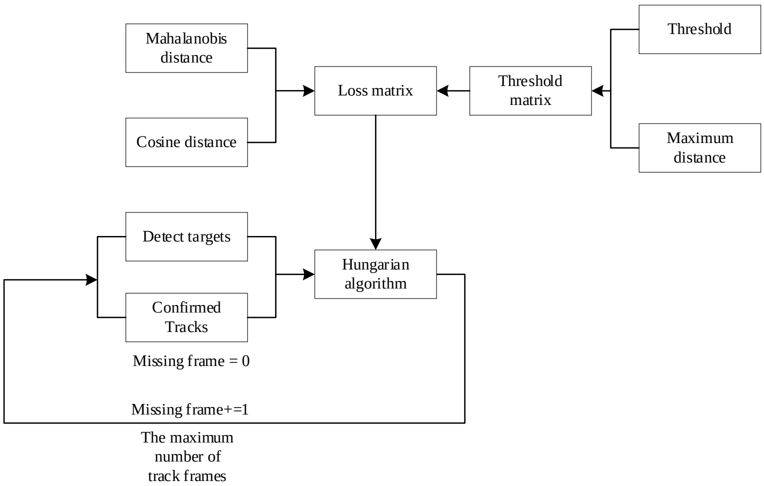


Fig. 10. Cascading Matching Flowchart

2.2.3 Tracking Processing and State Estimation

Deep SORT uses the detection results of YOLO V5 to initialize the trackers. Each tracker will set a counter. After the Kalman filter, the counter is accumulated. When the prediction result and the detection result are successfully matched, the counter is set to 0. If there is no valid match with the detection result within a period of time, the tracker is deleted. Deep SORT assigns a tracker to the new detection results in each frame. When the prediction results of the tracker in three consecutive frames can be effectively matched, it is confirmed that a new trajectory has appeared. Otherwise, this tracker will be deleted.

3. Experiments and Conclusions

3.1 Experimental dataset

(1) The pedestrian detection dataset selected in this experiment is the PASCAL VOC dataset. The main research is pedestrian detection, so the total training data is cleaned, only the Person class is retained, and six thousand pieces of training data are obtained, and a total of two thousand pieces of test data are obtained. Some of the figures are shown in **Fig. 11**.

Fig. 11. PASCAL VOC pedestrian dataset

(2)Pedestrian Re-ID dataset: Market-1501 dataset was collected by Tsinghua University, mainly for campus scenes and includes 1501 pedestrians, 32668 detected pedestrians Rectangle. There are 751 types of pedestrians in the whole data set, 12,936 images in the training set and 19,732 images in the test set. Some of the figures are shown in **Fig. 12**.

Fig. 12. Market-1501 dataset

(3) MOT17 Det pedestrian tracking dataset: MOT is a public benchmark dataset for the MOT Challenge multi-target detection and tracking method. Subsequent MOT algorithms will basically give the performance on MOT16, even if MOT17, MOT17Det (among them only pedestrians) Annotation) MOT17Det dataset contains 7 training sets and 7 test sets. Some of the figures are shown in **Fig. 13**.

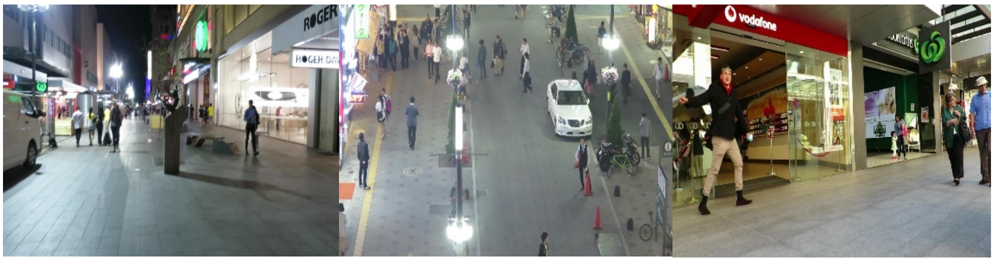


Fig. 13. MOT17 Det pedestrian tracking dataset

3.2 Evaluation indicators

(1) In terms of target detection, precision, Recall and mean Average precision (mAP) values are used to evaluate network performance, and the formula is shown in (7) - (10) :

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \sum_{i=1}^{N} AP_i / N \tag{10}$$

(2) Target tracking evaluation indicators

In terms of target tracking, IDs, MOTA, MOTP, and FPS values are used to judge network performance.

1) IDs: IDs indicate that at time t, the similarity between the real labeled target i and the hypothetical target j is less than a certain threshold, and the misjudgment of the tracking algorithm leads to the mismatch between the ID and the labeled information.

2) MOTA: It is used to represent the tracking precision of multiple targets, and its formula is as follows:

$$MOTA = 1 - \frac{FNv + FPV + IDs}{GT} \tag{11}$$

3)MOTP: It is used to represent the multiple target tracking precision, which intuitively expresses the error between the real frame and the predicted frame. The formula is shown in (12):

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_{t} c_{t}} \tag{12}$$

4)FPS: Indicates how many frames of pictures the algorithm can process per second, which is used to visually express the processing speed of the algorithm. The formula is shown in (13):

$$FPS = \frac{1}{t} \tag{13}$$

3.3 Experimental Analysis

3.3.1 Experimental Environment

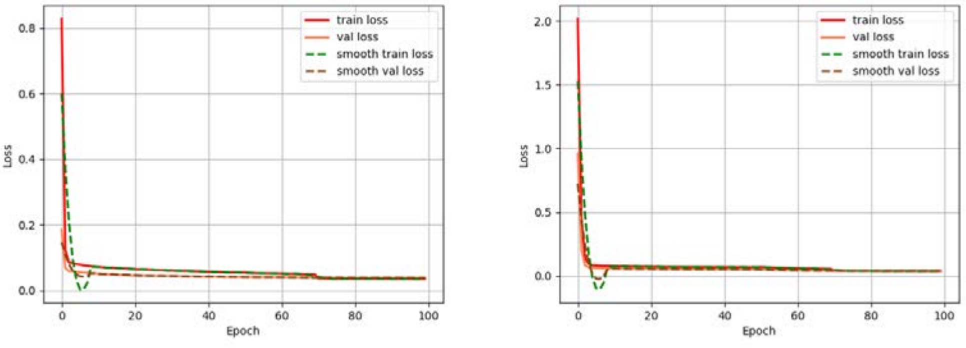
The experiment is implemented using the pytorch framework, the pytorch version is 1.2.0, the experimental platform uses the ubuntu 16.04 operating system, the GPU is NVIDIA Tesla v100, the CUDA version is 10.0.130, and the Python version is 3.7.0

3.3.2 YOLO v5 ablation experiment

So as to verify the effectiveness of the YOLOv5 improvement strategy proposed in this paper, ablation experiments were conducted on the PASCAL dataset to judge the effectiveness of each improvement point, and ECA-Net and CIoU Loss were added to the initial YOLOv5s in turn. The training process uses the same parameter configuration, the input image is 640×640, epoch=100, and the Adam optimizer is used. The ablation experiment results are shown in the **Table 2**.

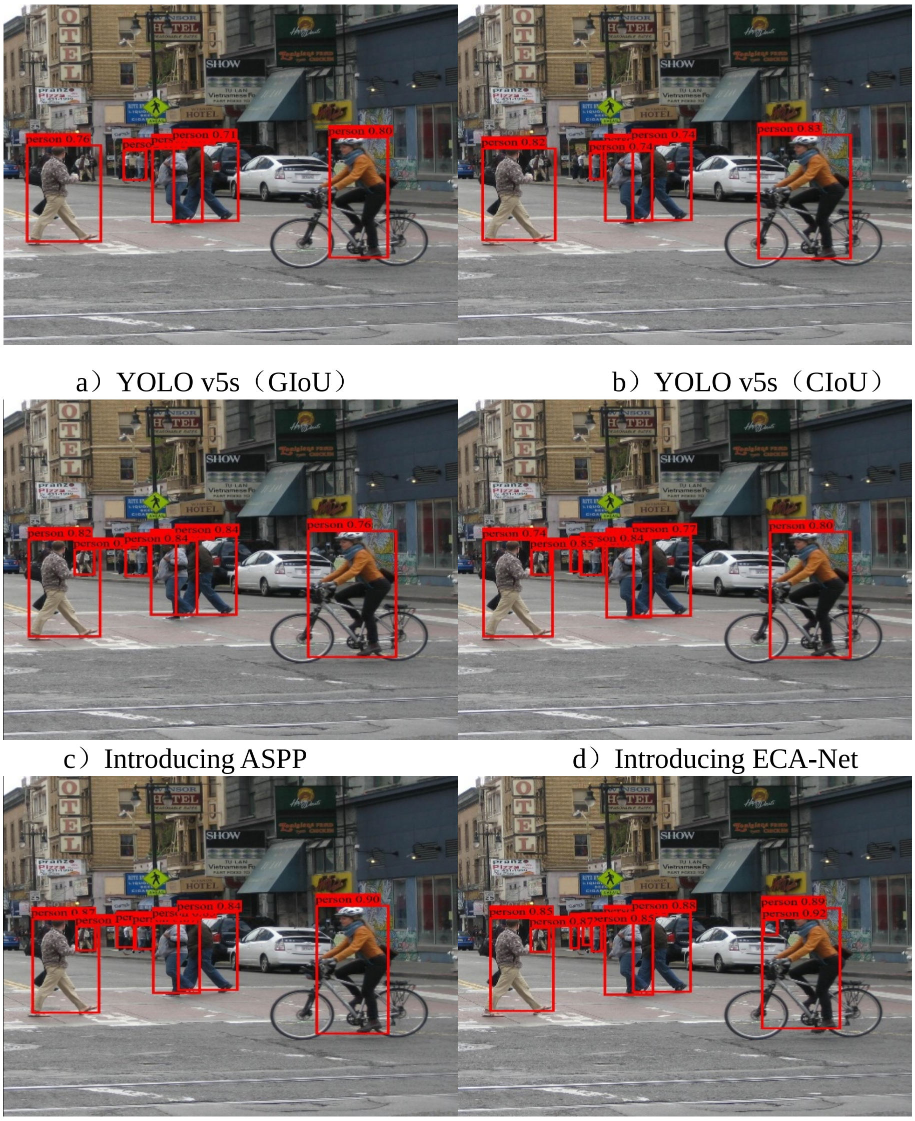
Table 2. YOLOv5s ablation experiments

CIoU Loss	ASPP	ECA-Net	Precision	Recall	mAP@0.5
×	×	×	93.2	78.0	92.1
√	×	×	93.7	78.5	92.2
√	√	×	94.3	77.4	92.2
√	×	√	95.4	77.6	94.5



(a) (b)

Fig. 14. Comparison of loss curves (a) GIoU (b)CIoU



a）YOLO v5s（GIoU） b）YOLO v5s（CIoU）

c）Introducing ASPP d）Introducing ECA-Net

e）ASPP+ECA-Net f）ASPP+ECA-Net(CIoU-NMS)

Fig. 15. Ablation experiment comparison diagram

After adding the CIoU loss function to the YOLO v5 network, it is found by judging the loss value that the CIoU loss can converge faster than the GIoU loss, and the detection accuracy is improved by 0.5%. When only the ASPP module is added to the network , the detection accuracy has increased by 0.6%, and after adding the ECA attention mechanism, the detection accuracy has increased by 1.7% compared with the original network model, and the map@0.5 benchmark has increased by 2.3%, and successfully detected To the small target located in the distance. When the ASPP module and ECA-Net are added to the network at the same time, compared with the original YOLO v5 network, the detection accuracy is increased by 3.1%, the recall rate is increased by 1.5%, and the map@0.5 benchmark is increased by 3.3%. Finally, after adding CIoU-NMS, the detection effect of the model in the face of occlusion has strong robustness. The correlation pairs are shown in the Fig. 14 and Fig. 15.

3.3.3 YOLO v5 Comparative Experiment

Table 3. Object detection comparison experiment

Method	**FPS**	**Model size**	**Precision**	**mAP@0.5**
Fast R-CNN	5.2	109.2M	94.7	94.5
YOLO v3	20.7	230 M	93.5	82.7
CA-YOLO v5	33.6	67.5 M	94.8	83.9
YOLO v4+SE	13	121.5M	94.8	94.4
Mobilenetv3+YOLO v4+ASPP	18	44.7M	93.8	94.2
Ours	21	42.9M	96.1	95.4

In the comparative experiment, the proposed algorithm was compared with the network after the introduction of fast R-CNN and YOLO v4 into the SE module, and the network after mobilenetv3-YOLO v4 was introduced into the ASPP module. The correlation pairs are shown in the **Table 3** and the experiment found that the improved algorithm in this paper improved the accuracy rate by 1.4% compared with Fast R-CNN, compared with Mobiletv3-YOLO v4, the accuracy rate increased by 2.3%, and the benchmark of map@0.5 was improved by 1.2%. Compared with YOLO v4-SE, the accuracy rate is improved by 1.4%, and the benchmark of map@0.5 is improved by 1.0%, compared with the other three models, the model proposed in this paper is smaller in size, has the effect of lightweight detection, and the real-time detection speed is faster. The correlation pairs are shown in the **Fig. 16**.

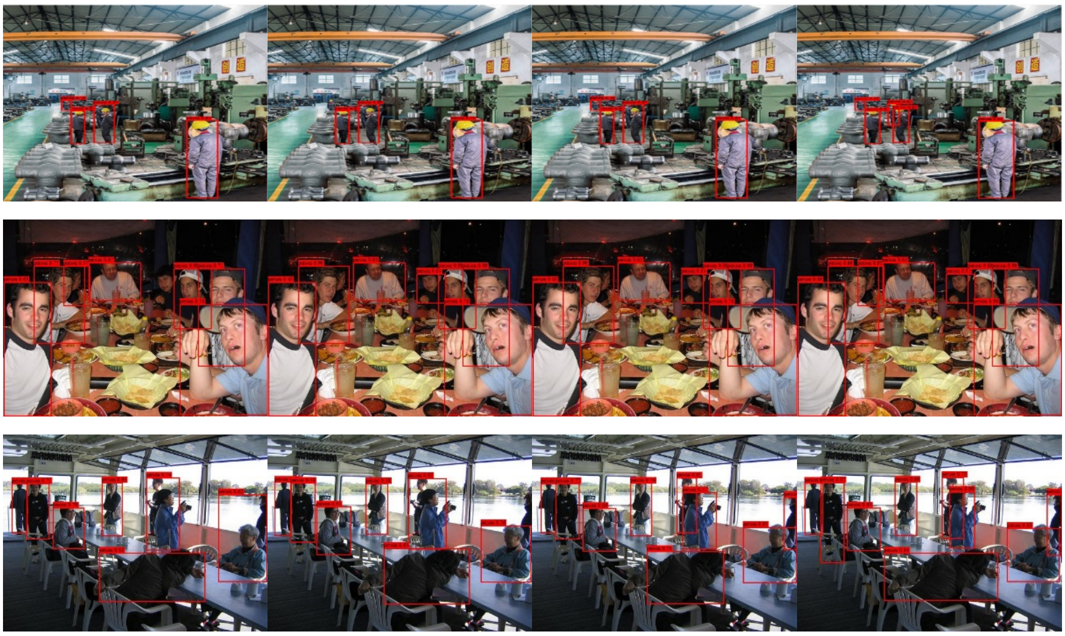

a) Fast R-CNN b) Tiny-YOLO v4 c) CA-YOLO v5 d) Ours

Fig. 16. Algorithm comparison diagram

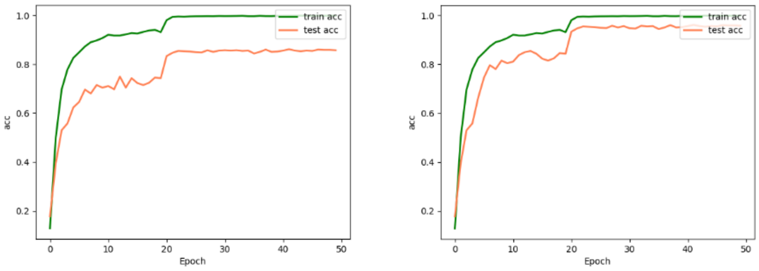
3.3.4 Target tracking experiments

(1) Shuffle-Net impact on the trace module

This article analyzes the model size, classification accuracy, and Shuffle Net's perspective on tracking module speed.

Table 4. Feature extraction network experiments

Network model	Model size	Precision
Original model	51.7	87.6
Shuffle Net	20.3	95.3



(a) (b)

Fig. 17. Training accuracy curve

As can be seen from the **Table 4** and **Fig. 17**, after replacing the original network with Shuffle Net, the model size has been greatly reduced, and the classification accuracy has been improved by nearly 8%.

Table 5. Effects of Shuffle Net on the Trace Module

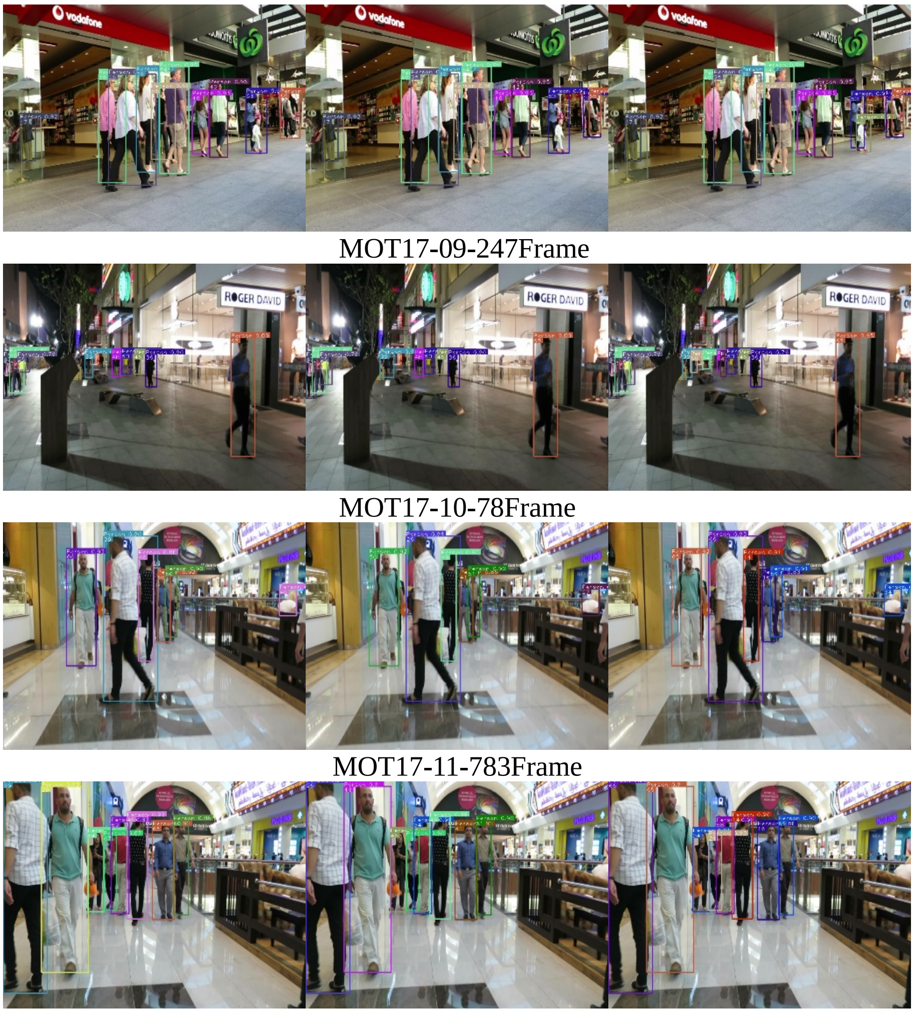
Method	FPS	MOTA	MOTP
Original model	20.7	54.1	79.0
Shuffle Net	26.3	54.3	79.1

Through **Table 5**, the tracking model using Shuffle Net has increased the FPS index by 4FPS compared with the original model, which improves the tracking speed, and after replacing Shuffle Net, it has not decreased on THE MOTA and MOTP indicators, and the accuracy of the original tracking can still be maintained.

(2) Comparison of tracking algorithms

Table 6. Target tracking comparison experiments

Method	FPS	MOTA	MOTP	IDs
MOTDT	21.3	49.6	77.1	941
SORT	14.3	46.7	76.1	1423
Deepsort+YOLOv3	12.5	49.4	76.2	877
Deepsort+YOLOv4	14.3	49.8	78.3	968
Deepsort+YOLOv5+ECA	21.7	54.2	79.0	721
SiamCNN	8.0	51.2	73.9	854
MDP	15.7	52.4	74.6	836
Ours	26.3	54.3	79.1	668



MOT17-09-247Frame

MOT17-10-78Frame

MOT17-11-783Frame

MOT17-11-785Frame

a) Faster R-CNN b)YOLO v4-tiny c) Ours

Fig. 18. Comparison of tracking effects

Through **Fig. 18** and **Table 6**, using DeepSORT combined with the improved YOLO v5 algorithm, after the introduction of the attention mechanism and ASPP module, the MOTA indicator is increased by 4.9% and 4.5% compared with YOLO v3 and YOLO v4, respectively, and the MOTP indicator is increased by 2.9% and 0.8% compared with the original YOLO v3 and YOLO v4, respectively, and the algorithm after the introduction of CIoU-NMS has a significant reduction in ID jump compared with the original algorithm. After the introduction of ShuffleNet, the proposed algorithm effectively reduces the size of the network model, and the tracking speed is greatly improved compared with other algorithms using Deep SORT, compared with the MOTDT algorithm, the tracking speed is also improved by 4.0FPS, which can basically meet the requirements of lightweight tracking.

4 Conclusion

The proposed algorithm uses YOLO v5 as the detector, combined with Deep SORT to identify and track pedestrian targets. Through the improvement of the Deep SORT model, the parameters used by the model are effectively reduced, and the tracking speed is improved; after the ECA-Net and ASPP modules are introduced into the YOLO v5 detector, the detection is further improved without expanding the network model. After the introduction of CIoU-NMS, the phenomenon of missed detection of pedestrians in occlusion and crowded scenes is effectively improved. Finally, a large number of experiments show that the algorithm proposed in this paper outperforms the existing ones for crowded scenes and pedestrians are greatly occluded. All of the relatively novel algorithms show good performance in the aspects of speed and accuracy, and achieve the purpose of lightweight fast tracking. In order to achieve a more accurate tracking task, certain improvements can be made in the calculation and design of the anchor frame in the future to make the prediction frame more accurate and further improve the overall robustness.

Acknowledgements

This research was supported by the Foundation of Liaoning Educational Committee under grant No. LJKMZ20221827.

We thank the anonymous referees for their helpful comments and suggestions on the initial version of this paper.

References

[1] Bolme DS, Beveridge J R, Draper B A, et al., “Visual object tracking using adaptive correlation filters,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, IEEE, 2544-2550, 2010. Article (CrossRef Link)

[2] Henriques J F, Caseiro Riro R, Martins P, et al., “High-speed tracking with kernelized correlation filters,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 583–596, 2015. Article (CrossRef Link)

[3] QiY J, WangY J, LiuY C., “Object tracking based on deep CNN feature and color feature,” in *Proc. of 2018 14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, 469-473, 2018. Article (CrossRef Link)

[4] LiL Q, YanM Y., “High-order cumulant-based particle filtering algorithm for pedestrian object tracking,” in *Proc. of 2018 14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, 304-307, 2018. Article (CrossRef Link)

[5] Sun Y J, Zhang L Y, Yun X, “Visual tracking algorithm based on region estimation and adaptive classification,” *Laser & Optoelectronics Progress*, 56(18), 181001, 2019.

[6] LiY B, JiuM Y, SunQ, et al., “An improved target tracking algorithm based on extended Kalman filter for UAV,” in *Proc. of 2018 IEEE Asia-Pacific Conference on Antennas and Propagation (APCAP)*, Auckland, New Zealand, 435-437, 2018. Article (CrossRef Link)

[7] HeimbachM, EbadiK, WoodS, “Resolving occlusion ambiguity by combining Kalman tracking with feature tracking for image sequences,” in *Proc. of 2017 51st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 144-147, 2017. Article (CrossRef Link)

[8] Bewley A, Ge Z, Ott L, Ramos F, Upcroft B, “Simple Online and Realtime Tracking,” in *Proc. of 2016 IEEE International Conference on Image Processing (ICIP)*, 3464-3468, 2016. Article (CrossRef Link)

[9] Wojke N, Bewley A, Paulus D., "Simple Online and Realtime Tracking with a Deep Association Metric," in *Proc. of 2017 IEEE International Conference on Image Processing (ICIP)*, 3645-3649, 2017. Article (CrossRef Link)

[10] Wang Z, Zheng L, Liu Y, et al., "Towards real-time multi-object tracking," in *Proc. of European Conference on Computer Vision*, Springer, Cham, 107-122, 2020. Article (CrossRef Link)

[11] Lu Z, Rathod V, Votel R, et al., "Retinatrack: Online single stage joint detection and tracking," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 14668-14678, 2020. Article (CrossRef Link)

[12] Chu P, Ling H., "Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 6172-6181, 2019. Article (CrossRef Link)

[13] Sun S J, Akhtar N, Song H S, et al., "Deep affinity network for multiple object tracking," *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 104-119, 2021. Article (CrossRef Link)

[14] Zhou X, Koltun V, Krähenbühl P., "Tracking objects as points," in *Proc. of European Conference on Computer Vision*, Springer, Cham, 474-490, 2020. Article (CrossRef Link)

[15] Wu J, Cao J, Song L, et al., "Track to detect and segment: An online multi-object tracker," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 12352-12361, 2021. Article (CrossRef Link)

[16] Zhang Z, Peng H., "Deeper and wider siamese networks for real-time visual tracking," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 4591-4600, 2019. Article (CrossRef Link)

[17] Fan H, Ling H., "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 7952-7961, 2018. Article (CrossRef Link)

[18] Yu Y, Xiong Y, Huang W, et al., "Deformable siamese attention networks for visual object tracking," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 6728-6737, 2020. Article (CrossRef Link)

[19] Chen X, Yan B, Zhu J, et al., "Transformer tracking," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 8126-8135, 2021. Article (CrossRef Link)

[20] Yan B, Peng H, Fu J, et al., "Learning spatio-temporal transformer for visual tracking," in *Proc. of the IEEE/CVF international conference on computer vision*, 10448-10457, 2021. Article (CrossRef Link)

[21] Wang Q, Zheng Y, Pan P, et al., "Multiple object tracking with correlation learning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3876-3886, 2021. Article (CrossRef Link)

[22] He J, Huang Z, Wang N, et al., "Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5299-5309, 2021. Article (CrossRef Link)

[23] Hu J, Shen L, Sun G., "Squeeze-and-excitation networks," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 7132-7141, 2018. Article (CrossRef Link)

[24] Liu S, Qi L, Qin H, et al., "Path aggregation network for instance segmentation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 8759-8768, 2018. Article (CrossRef Link)

[25] Qilong W, Banggu W, Pengfei Z, et al., "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11534-11542, 2020. Article (CrossRef Link)

[26] Chen L C, Zhu Y, Papandreou G, et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of the European conference on computer vision (ECCV)*, 801-818, 2018. Article (CrossRef Link)

[27] Rezatofighi H, Tsoi N, Gwak J Y, et al., "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition*, 658-666, 2019. Article (CrossRef Link)

[28] Wojke N, Bewley A, Paulus D., "Simple online and realtime tracking with a deep association metric," in *Proc. of 2017 IEEE international conference on image processing (ICIP)*, IEEE, 3645-3649, 2017. Article (CrossRef Link)

[29] Ma N, Zhang X, Zheng H T, et al., "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. of the European conference on computer vision (ECCV)*, 116-131, 2018. Article (CrossRef Link)

Qiang Gao received the Ph. D. degree in computer application technology from Northeastern University, China, in 2016. Since 2018, he has been with Shenyang University, where he is currently an associate professor with Institute of Innovation Science and Technology. He is the first author of more than 20 articles. His main research interests include computer game, artificial intelligence, and computer vision.

Zhicheng He is a master's student in control science and engineering at Shenyang University. His main research interests are artificial intelligence, video image processing, and deep learning.

Xiaowei Han received the Ph.D. degree in control theory and control engineering from Northeastern University, in 2005. He is currently a Professor and the President of Scientific and Technological Innovation Institute, Shenyang University. He has presided over or undertaken more than ten research projects supported by national, provincial and municipal funds, completed a number of horizontal engineering projects, compiled two monographs, published more than 40 articles, and obtained more than 50 invention patents, utility model patents. His current research interests include computer vision, artificial intelligence, and wireless sensor networks.

Yinghong Xie received the Ph.D. degree in pattern recognition and artificial intelligence from Northeastern University, China, in 2014. Since2005, she has been with Shenyang University, where she is currently a Professor with Information and Engineering Institute. From 2014 to 2016, she was a Postdoctoral Researcher with Tianjin University. She was a Scholar with the University of Michigan–Dearborn, in 2017. She is the first author of more than 20 articles, and the Host of Natural Science Foundation of China, in 2015. Her main research interests include artificial intelligence, video image processing, and pattern recognition.

Xu JIA was born in Kaiyuan City, Liaoning Province, China in 1983. He received the B.S. in automation from Shenyang Aerospace University, Liaoning Province, China, in 2005 and the M.S. and the Ph.D. degrees in pattern recognition and intelligent system from Northeastern University, Liaoning Province, China, in 2009 and 2012.

Since 2021, he has been a professor with School of Electronics and Information Engineering, Liaoning University of Technology. He is the author of more than 30 articles. His research interests include machine learning and image processing.