Multi-Target Pedestrian Tracking Method in Complex Environments Based on Improved YOLOv7 and BoT-SORT

Yanhui Lv, Qichao Guo, and Xudong Jia

Abstract—**In complex scenarios, pedestrian target scale variations are significant, which makes pedestrian target detection difficult. To address the issues of complex background interference and mutual occlusion among pedestrians, an improved strategy for pedestrian target detection and tracking is proposed, combining YOLOv7 and BoT-SORT. Firstly, the backbone feature extraction network of YOLOv7 is improved to enhance the feature extraction capability. Secondly, an improved SimAM attention mechanism is introduced into the network model to extract the main features, enhancing the feature extraction ability of small targets. An improved region-based BoT-SORT tracking algorithm is developed by refining BoT-SORT. Constraints are added to the locations where monitored targets appear, reducing the problem of ID switches caused by target occlusion. Experimental results show that the optimized target detection algorithm achieves a 2.5% increase in average detection accuracy on the WiderPerson public dataset, and the tracking algorithm reduces the number of IDs compared to the original algorithm by 11 on average on the MOT17 dataset.**

Index Terms—**target detection, target tracking, YOLOv7, BoT-SORT**

I. Introduction

The detection and tracking of objects are important image processing techniques that provide data support for complex tasks. They have been playing a crucial role in various industries, especially in the fields of intelligent vehicles and video surveillance. This paper focuses on pedestrian monitoring in complex scenarios. Security monitoring is widely used in various aspects of daily life, and intelligent monitoring technology[1] greatly improves work efficiency while reducing the workload of people, making it of significant research value.

The Detection-Based-Tracking[2] (DBT) mode divides object detection and tracking into two aspects, and targeted improvements can achieve complementarity between target detection and tracking technologies. Chen et al.[3] used the PA-ResNeXt network as the backbone network for feature extraction and introduced an attention feature fusion module to improve the model's simultaneous attention to local and global features. Wang et al.[4] used K-Means++ clustering technique to re-cluster the dataset to obtain the best anchor frames. And Focal-EIOU loss function was introduced to speed up the convergence and improve the regression results. Sun et al.[5] ensured efficient multi-scale feature fusion by integrating a small target detection layer and modifying the feature pyramid using a feature fusion technique in a weighted bi-directional feature pyramid network (BIFPN). Wang et al.[6] proposed a pedestrian tracking algorithm based on Scale Adaptive Kernel Correlation Filter (SAKCF). Median filtering is used to suppress the background information and improve the signal-to-noise ratio. Luo et al.[7] reconstructed the feature extraction module in the detector using an improved CSPDarkNet53[8] backbone network to achieve the goal of improving the detection and tracking performance with a higher-quality detector. He et al.[9] added the CBAM[10] attention mechanism to ResNext[11] to make the model focus on key information. Fan et al.[12] introduced position-constrained matching into the DeepSORT[13] algorithm, re-matching unmatched targets using Euclidean distance. Tu et al.[14]added behavior recognition to the DeepSORT algorithm, creating a multi-category division to make detection results more accurate. Zhao et al. [15] added direction differences to the cost matrix of the DeepSORT algorithm to make tracking more accurate.

Because pedestrian objects exhibit diversity and mobility, and are prone to crossings and occlusions in complex backgrounds, the above-mentioned studies have made relevant improvements in enhancing the accuracy of object detection and the number of associations. However, there are still issues with insufficient object detection accuracy and significant ID switches[16] caused by target occlusions. Based on previous research experience, this paper adopts YOLOv7[17] and BoT-SORT[18] as the fundamental strategies, primarily conducting the following research:

(1) In YOLOv7, the ConvNeXtv2[19] feature extraction structure is employed, along with the SimAM[20] attention mechanism to create a feature transfer module. This enhances the feature extraction capabilities for the main features, improving the ability to extract features from targets.

(2) The target tracking algorithm is optimized for specific scenarios that make a judgment on the location where the target appears to resolve the issue of unreasonable IDs.

The rest of the paper is organized as follows: section 2 describes the detection module constructed in this paper. Section 3 details the tracking algorithm proposed in this

Manuscript received January 7, 2024; revised September 23, 2024.

This work was supported by the Basic Research Projects of Liaoning Provincial Department of Education(JYTMS20230192).

Yanhui Lv is a professor of the College of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China (phone: 86-24-24682228; e-mail: yanhuilv@ 126.com).

Qichao Guo is a postgraduate student of Shenyang Ligong University, Shenyang 110159, China (e-mail: 1040431896@qq.com).

Jiaxu Dong is a postgraduate student of Shenyang Ligong University, Shenyang 110159, China (e-mail: jiaxd32312@foxmail.com).

paper. Section 4 reports the simulation results. Section 5 contains concluding remarks and future directions.

II. Improvement of YOLOv7

A. Principle of YOLOv7 Algorithm

YOLOv7 is a type of target detection algorithm YOLO with an adapted structure relative to YOLOv5. The algorithm is a real-time target detection algorithm designed to improve detection speed while maintaining high accuracy. The basic principle of the algorithm is to treat the target detection problem as a single regression problem by dividing a grid on the input image and predicting the bounding frames and categories of targets present in each grid.

Although YOLOv7 pushes the balance of accuracy and speed to a new height in the detection of pedestrian targets, there are still some problems. Firstly, when the number of pedestrian targets increases, the pedestrians obscure and overlap with each other resulting in a more complex detection scene, which often makes the detection results inaccurate. At the same time, the small targets in the scene also need more accurate feature extraction. Secondly, the feature fusion process does not make full use of the features of different scale feature maps, while the computational volume and accuracy of the model also need to be balanced. To address the above problems, the following improvements are made to YOLOv7.

B. Improvement of Backbone Network

The target detection task in this paper involves complex scenes, mainly including occlusion and deformation. Improvement of the backbone network can cope with these challenges through more sophisticated feature extraction, thus improving the accuracy of detection. This section will introduce some terms and design options for the module.

1) Convnextv2Block

The ConvNeXt network[21] improves the performance of purely convolutional architectures. ConvNeXtv2 is an improved version of ConvNeXt, which introduces a Full Convolutional Mask Auto-Encoder (FCMAE) framework and a new Global Response Normalization (GRN) layer .

The GRN layer is a feature processing layer for convolutional neural networks designed to enhance feature competition between channels. It is originally designed to solve the problem of feature collapse caused by FCMAE. GRN can perform global feature aggregation, normalization of features, and feature calibration, and the GRN layer is very simple to implement, requiring very little code and no learnable parameters. Added to convolutional neural networks, it enhances feature competition and improves model performance, which is effective for some visual tasks. The GRN is computed in three steps:

First, the spatial feature map X is aggregated into a vector gx with a global function $G(x)$, as shown in (1).

$$G(x) = X \in R^{H*W*C} \rightarrow gx \in R^{C} \tag{1}$$

Where, H, W, and C represent the height, width, and the number of channels of the output features, respectively. $G(x)$ represents a scalar that aggregates statistics for that channel.

Next, for the i-th channel, the relative importance compared to all other channels is calculated as shown in (2).

$$N(||X_i||) = ||X_i|| \in R \rightarrow \frac{||X_i||}{\Sigma_{j=1,2,3...,C}||X_j||} \rightarrow R \tag{2}$$

Where $||X_i||$ represents the L2 norm of the i-th channel.

Finally, a normalization calibration is performed to obtain the spatial feature Y_i as shown in (3).

$$Y_i = X_i * N(G(X)_i) \in R^{H*W} \tag{3}$$

Where N represents normalization.

The specific structure of the ConvNeXtv2 block is shown in Fig. 1.

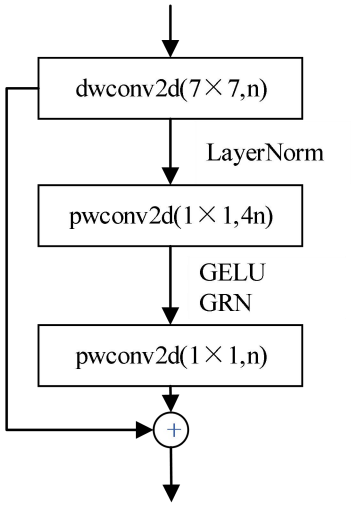

Fig. 1. ConvNeXtv2 block structure

2) GhostConv

GhostConv[22] is a structure used for lightweight models, which achieves the goal of reducing the number of parameters by introducing a linear operation on the ordinary convolution operation.

The computational effort of ordinary convolution and GhostConv can be compared by the ratio r of the two, as shown in (4).

$$r = \frac{c * h * w * n * H * W}{\frac{n}{s} * H^2 * W^2 * k^2 * c * \frac{(s-1)n}{s} * d^2} \approx s \tag{4}$$

Where c is the number of channels, h, w is the length and width of the input features, n is the number of convolution kernels, H, W is the length and width of the output features, k is the size of the convolution kernel, s is the number of transformations, and d is the size of the transformed convolution kernel. Using GhostConv its FLOPs can generally be reduced to 1/s of the original.

3) ELAN-CN module

The backbone network structure in YOLOv7 is mainly composed of ELAN and MP modules. The traditional Conv structure is used in ELAN, while the structure of ConvNeXtv2block can solve the problem of feature saturation, which can differentiate features. In this paper, the ELAN module is designed in conjunction with ConvNeXtv2block.

Firstly, on the basis of ConvNeXtv2block module, along with the idea of MP branching design, a transition module for feature extraction is designed, which is named CN in this paper, and its structure is shown in Fig. 2.

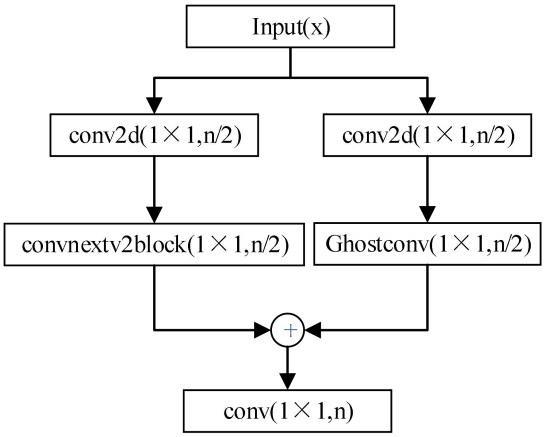

Fig. 2. CN module structure

The module is modeled on the MP design idea, where the features are processed in two parts, one part of the features is extracted using ConvNeXtv2block, and the other part retains the original feature information. Finally these features are fused.

Then the CN module is introduced into the ELAN module and the convolutional structure of GhostConv is adopted. The structure of the final feature extraction module ELAN-CN is shown in Fig. 3.

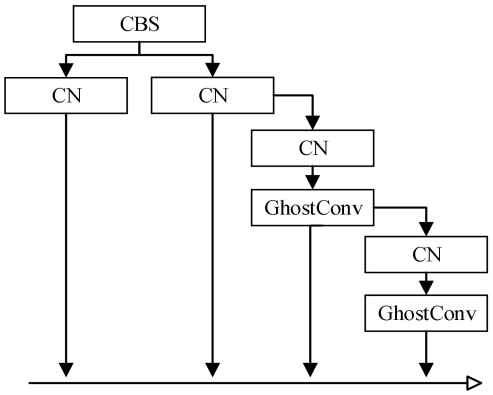

Fig. 3. ELAN-CN module structure.

ELAN-CN divides the network into four distinct connecti ons. At the beginning of each connection, the CN module is used to extract features. To reduce the parameters, the secon d feature extraction module utilizes GhostConv. Assuming t he input feature matrix is X, and the feature inputs for each g roup are X_1, X_2, X_3, and X_4, where X_3 and X_4 can both be repr esented by X_2. The features of four different depths are obtai ned sequentially through the CN convolution and GhostCon v convolution functions. Finally, these features are fused usi ng the Concat structure. The output features are obtained as shown in (5).

$$Xout = C(X_1) + C(X_2) + CG(X_2) + CG(CG(X_2)) \quad (5)$$

Where C represents CN operation and G represents Ghostconv operation. The architecture can extract features from a feature map with dimensions n and channel count C of (n, n, T) as a feature map with (n, n, $4T$). This enhances feature extraction with Conv and yields more precise features. Simultaneously, the use of GhostConv helps strike a balance between accuracy and computational complexity.

C. Design of the Feature Transfer Module

A well-designed feature transfer module is crucial for improving the performance of deep learning models. Effective feature transfer helps the model capture critical features of the data, thereby enhancing the accuracy of the model. In YOLOv7, the part connecting the backbone network and the head is only composed of CBS. Therefore, in this section the feature transfer module is designed in conjunction with the SimAM attention mechanism.

1) SimAM Attention Mechanism

Attention can emphasize the extraction of important features of interest. SimAM is a highly effective attention module. In existing attention mechanisms, this mechanism has the advantage of surpassing single spatial and channel attention. This module can derive 3D attention weights without the need for additional parameters.

2) SimAMCN

SimAM combines neuroscience to define an energy function. To guide and manage attention more effectively, it is necessary to evaluate the importance of each neuron and assign corresponding weights. The energy function *et* defined by SimAM for each neuron is shown in (6). The estimated value $\hat{t}$ of the target neuron and the estimated value $\hat{x_i}$ of the adjacent neuron are shown in (7) and (8), respectively.

$$e_t(w_t, b_t, y_i, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_0 - \hat{x_i})^2 \quad (6)$$

$$\hat{t} = w_t\, t + b_t \quad (7)$$

$$\hat{x_i} = w_t\, x_i + b_t \quad (8)$$

Where, w_t represents weights, b_t represents biases, y_i is the binarization parameter for y_t and y_0, x_i is the neuron estimation parameter for $\hat{t}$ and $\hat{x_i}$, y_t is the true value of the target neuron, y_0 is the true value of the adjacent neuron, and M is the number of neurons in channel i.

In neuroscience, information-rich neurons inhibit other unit neurons. To find this neuron, by setting y_t to 1 and y_0 to 0, and adding a regularization term, (9) can be obtained.

$$e_t(w_t, b_t, y_i, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - w_i x_i - b_t)^2 + (-1 - w_t\, x_i - b_t)^2 + \rho w_i^2 \quad (9)$$

Where w_i is the weight of neuron i and ρ is the regularization coefficient. Finally, by substituting w_t and b_t into (9), the minimum energy e_t* can be obtained, as shown in (10).

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (10)$$

Where λ is the introduced hyper-parameter, $\hat{\sigma}^2$ is the neurons variance, and $\hat{u}$ is the mean value of neurons in that channel, as shown in (11) and (12).

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{u})^2 \quad (11)$$

$$\hat{u} = \frac{1}{M}\sum_{i=1}^{M} x_i \quad (12)$$

The importance of neurons can be obtained by taking the reciprocal of e_t* and assigning different weights accordingly. Finally, the calculation of the weight W is as follows:

$$W = sigmoid(1/e_t^*) \quad (13)$$

In this paper, based on the SimAM attention mechanism, we design the feature transfer module, called SimAMCN, In this structure, there are three modules are involved in feature transfer. It improves the feature transfer module from the backbone to the HEAD and extends the simple CBS structure into a feature transfer module with SimAM as the core. The structure of the module is shown in Fig. 4.

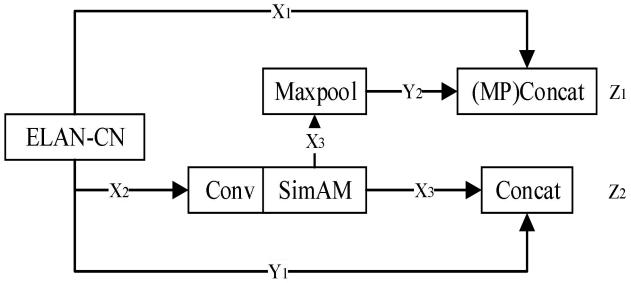

Fig. 4. SimAMCN structure.

Assuming the size of the feature map obtained after ELAN's feature extraction is (n, n) with T channels. Let X_1 represent the features passed to MP, X_2 represent the features passed to SimAM, and Y_1 represent the features passed to the neck. X_2 constructs a 3D weight by SimAM to obtain the feature map X_3. After applying max pooling to X_3, the resulting feature is Y_2 with dimensions (n/2, n/2). In the MP module, X_1 is fused with Y_2 to obtain Z_1. Simultaneously, Y_1 and X_3 are passed backward and fused with features from the neck to obtain Z_2.

The feature transfer module fuses the features of different scales, extracts the important features and connects the output feature maps with different sizes to get a structure that can fuse the main features and transfer the features backward at the same time. On the one hand, the selection results of SimAM as residual edges are fused into the backbone network, which can achieve cross-layer information transmission. On the other hand, the selection results of SimAM can be fused with the HEAD part, which can pass backward the features made significant by SimAM.

III. Optimization of the BoT-SORT algorithm

A. Principle of BoT-SORT algorithm

BoT-SORT, as a top algorithm for target tracking, follows the processing flow of the original bytetrack[23], which divides the detection frames into two confidence levels, and prioritizes the matching of high-confidence detection frames, while filtering and removing part of the low-confidence detection frames, to ensure the tracking is carried out efficiently. BoT-SORT uses the camera motion compensation, and at the same time, it incorporates the reID[24] and iou matching in the trajectory correlation.

The SORT algorithm, when extended to bytetrack, has a well-established process for trajectory association. However, the issue of ID discontinuity caused by occlusion remains one of the main reasons for tracking failures. To address the problem of ID discontinuity caused by occlusion, this paper proposes a tracking strategy based on region partitioning. By focusing on targets within specific regions, this strategy aims to handle ID discontinuities resulting from occlusion. As a result, it becomes possible to determine the reasonableness of trajectories even after a certain number of frames, referred to as max frames.

B. BoT-SORT tracking algorithm based on region classification

Inspired by the byte algorithm detection frames classification association, the appearance of the target in the video image often follows certain regularity, and different positions also required different attention. For example, the position where the target often appears can be prioritized and focused on. Usually, targets in surveillance video tend to first appear at the edge of the video, rather than at the center of the current frame. For specific scenarios, such as underground entrances, the center of the video tends to be the location with high foot traffic, and the entrance position will be arranged at the edge of the video, whereas traditional tracking algorithms do not take into account the reason for ID discontinuity due to occlusion.

IDs is an important metric for multi-target tracking, which indicates the total number of ID switches and can be used to measure the accuracy of a tracking algorithm trajectory. And there exists a situation of ID discontinuity, when the occluded target reappears, a new ID is given, which will be mistaken for the appearance of a new target. But in fact this target is the previous occluded target. In order to solve this problem, this paper uses a trajectory judgment based on region division, and adopts a new processing flow in the generation of new trajectories. The specific steps are as follows:

(1) The algorithm performs a total of three matches before and after tracking. The first one matches the high scoring target *high_det* with the confirmed state track *tracked_stracks*; the second one matches the low scoring target *low_det* with the remaining target track *u_track*0; and the third one matches the remaining high scoring target *u_det*0 with the unconfirmed state target track *unconfirm_track* for matching.

(2) The second matching will get the remaining target detection track *u_track*1. After the operation of the third IOU matching is completed, it gets the remaining high score detection frames *u_det*2 that is unsuccessful in this matching, and the non-confirmation state track *u_tracks*2 that are left behind, and these tracks and detection frames will be used for the judgments of the track identity and the generation of new tracks.

(3) Divide a surveillance video of 1920x1080 pixels into two parts including central region C and edge region B according to the crowd distribution characteristics, as shown in Fig. 5. Equally divide the video into 4x4 raster images, where the 960x540 part is used as the central region C and the rest is used as the edge region B. The specific range of region B can be given in the function according to the actual application scenario.

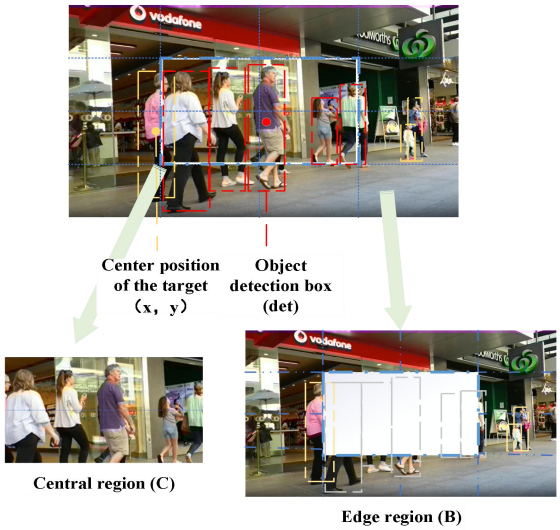

Fig. 5. Schematic diagram of BoT-SORT algorithm optimization

(4)The vertex position of the detected target is obtained through the checkbox properties function in the high scoring detection frame det, and the coordinates of the target center position (x, y) can be calculated by $x=l+w/2$ and $y=t+h/2$. Where, t and l represent the horizontal and vertical distance from the upper left corner of the image, respectively. And w and h represent the width and height of the detection frame, respectively.

(5) Treat the center of the target as a particle, use it to identify the position of the target in the image and use it as an indication of whether the target enters the observation field of view. In the illustration, the target is divided into a central observation target (indicated by a red detection frame) and an edge-to-be-observed target (indicated by a yellow detection frame), and their central positions are marked respectively. Determine whether the center position coordinate (x, y) has entered the central observation region. If the target enters the central region C and the track exists in the remove_tracks list, the track is merged into the indeterminate state and is listed as *lost_tracks*, and the survival time of the track is extended by 1 frame after the max frame until the target moves into the edge region B. This can keep the target alive in the region.

(6) When initializing a new trajectory, the target center position (x, y) needs to be located within region B, that is, the target will only be allowed to create a new trajectory when it moves outside the central observation area.

In this way trajectories in the region can survive continuously and unreasonable IDs due to occlusion are suppressed. The improved BoT-SORT algorithm is shown below.

ALgorithm: BoT-SORT tracking algorithm based on region segmentation

input: video sequence *V*; target detector *D*; number of surviving frames *m_thresh*; high scoring frame threshold *h_d_thresh*; new trajectory generation threshold *d_thresh*; region boundary *B*

output: video track sequence *T*

intialization: T←∅ D←∅
for frame in V:
  for det in D:
    if det.score>h_d_thresh:
      high_det←det
    else:
      low_det←det
  */*First match*/*
  ciou&re←min{iou_distance(high_det,tracked_stracks),reid_distance (high_det,tracked_stracks)}
  Linear_assigment()
  u_tracks0←remain from tracked_stracks
  u_dets0←remain from high_det
  */*Second match*/*
  ciou←iou_distance(low_det,u_tracks0) and Linear_assigment()
  u_tracks1←remain from U_tracks0
  u_dets1←remain from low_det
  */*Third Match*/*
  ciou←iou_distance(u_det0,unconfirm_track()) and Linear_assigment()
  u_tracks2←remain from su_unconfirm_track
  u_dets2←remain from u_det0
  */*Update Tracks*/*
  update_tracked(tracksed,activate,refind)
  if det in B and Max_time_out==m_thresh:
    max_time_out+=1
  else:
    update_Remove(lost_strack)
  update_lost_tracks()
  */*Initialising new trajectories*/*
  if det not in B and det.score>d_thresh+0.1:
    T←strack(det)
Return T

IV. Experimental Results and Analysis

A. Experimental Environment

In this paper, 7999 images are extracted from the WiderPerson[25] pedestrian public dataset and divided into training set, validation set and test set in the ratio of 6:2:2. The multi-objective tracking public dataset MOT17 is selected to test the algorithm proposed in this paper.

The operating system environment for the experiment is Ubuntu, the graphics card is Nvidia RTX 3090Ti, the running memory is 24G, the deep learning framework is Pytorch1.11.0, and the server Python version is 3.7.

B. Experimental evaluation metrics

mAP is a commonly used metric for evaluating the performance of target detection and image segmentation. It combines accuracy and recall to comprehensively evaluate the performance of the model in different categories. The calculation formula is as follows:

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{14}$$

Where AP_i represents the area under the PR curve for category i within all predicted pictures and k is the category.

MOTA is a measure of the multi-target tracker in terms of trajectory accuracy.

MOTA is an indicator that measures the trajectory accuracy of the multi-objective tracker. It takes into account three factors: mismatch, missed detection, false detection. Its calculation formula is as follows:

$$MOTA = 1 - \frac{Mmatches + Misses + FPositives}{TotalGT} \tag{15}$$

Mmatches denotes the number of mismatches, i.e., associating a real target with an incorrect tracker; *Misses* denotes the number of missed targets, i.e., real targets that failed to be detected correctly; *FPositives* denotes the number of false detections, i.e., mistakenly reporting the background or other objects as targets; and *TotalGT* denotes the total number of real targets.

MOTP is an indicator that measures the position accuracy of a multi-objective tracker. It calculates the average error between the predicted position and the actual position of the tracker. The calculation formula is shown in (16).

$$MOTP = 1 - \frac{sum(d)}{N} \quad (16)$$

Where d represents the error of each matching distance, sum (d) represents the sum of errors and N represents the total number of matches.

C. Experimental Results and Analysis

Both modules proposed in this paper have an enhancement effect on the YOLOv7 model, where SimAMCN improves 1.3% mAP over the use of attention alone. The improved object detection algorithm achieved a 2.5% increase in mAP compared to the original algorithm, as shown in Table I.

TABLE I
RESULTS BEFORE AND AFTER IMPROVEMENT

algorithm	P	R	mAP@0.5	GFLOPs
baseline	0.809	0.678	0.780	105.1
+ELAN-CN	0.794	0.695	0.784	104.0
+SimAM	0.806	0.685	0.785	105.1
+SimAMCN	0.813	0.700	0.798	114.7
+all	0.814	0.708	0.805	112.1

Additionally, two images from the WiderPerson test set are selected for testing, and the detection results are shown in Fig. 6. Among them, figure (a) shows the detection results of the original algorithm, and figure (b) shows the detection results of the improved algorithm. It can be seen that the improved algorithm generally enhances the confidence level for detecting the same target.

(a) Detection results of original YOLOv7

(b) Detection results of the improved algorithm

Fig. 6. Detection comparison results

The experiments compare the performance of different advanced algorithms on the WiderPerson dataset, and the results are shown in Table II. It can be seen that the mAP of the improved YOLOv7 target detection algorithm reaches 80.5%, which achieves the best detection performance in the crowd detection tasks in complex scenes.

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS

algorithm	mAP@0.5	algorithm	mAP@0.5
Faster RCNN	0.708	YOLOv3	0.728
Cascade RCNN	0.729	YOLOv4	0.752
FoveaBox	0.721	YOLOv5	0.746
FCOS	0.709	YOLOv7	0.780
FASF	0.737	Improved YOLOv7	0.805

Load the BoT-SORT algorithm before and after the improvement on the improved YOLOv7 algorithm, and the final experimental results are shown in Table III. The BoT-SORT algorithm before and after improvement can reduce the number of IDs in each video sequence, which reduces a total of 83 IDs, with an average reduction of 11 IDs.

TABLE III
TABLE OF CHANGES IN IDS

Name	Original algorithm IDs	Improved algorithm IDs
MOT17-02	81	61
MOT17-04	77	61
MOT17-05	44	36
MOT17-09	44	34
MOT17-10	60	54
MOT17-11	31	27
MOT17-13	39	20
Sum	376	293

The detailed results of the 5th sequence are shown in Table IV. The improved BoTSORT algorithm reduces the number of IDs by 8, increases the MOTA by 1.8%, and reduces the MOTP by 0.2% compared to the original tracking algorithm on the improved YOLOv7 algorithm.

TABLE IV
COMBINED EXPERIMENTAL RESULTS TABLE

Algorithm	IDs	MOTA	MOTP
YOLOv7+BoTSORT	60	0.472	0.329
YOLOv7+improved BoTSORT	56	0.483	0.342
improved YOLOv7+BoTSORT	44	0.562	0.335
improved YOLOv7+improved BoTSORT	36	0.580	0.333

In order to more intuitively demonstrate the effectiveness of our algorithm in tracking tasks, we conducted tracking detection on fixed lens MOT-09 sequences using a comparison method of the combination of YOLOv7+BoTSORT and improved YOLOv7+improved BoTSORT. The detection results are shown in Fig. 7 and Fig. 8, respectively.

Among them, figure (a) represents before the meeting, and figure (b) represents after the meeting. The target ID of trajectory 153 using the combination of improved YOLOv7+improved BoTSORT algorithm has not changed. The same target combined with YOLOv7+BoTSORT algorithm experienced an ID switch.

(a)before the meeting

(b)after the meeting

Fig. 7. Tracking results graph of YOLOv7+BoTSORT

(a)before the meeting

(b)after the meeting

Fig. 8. Tracking results graph of improved YOLOv7+improved BoTSORT

As shown in Fig. 9, the object detection area is divided into a center and an edge, and the area is divided in a 1:4 ratio. In figure (a), target 153 has already entered the central area, and at this time, target 153 is not allowed to create new trajectories. After the occlusion of the target ends, there is a high probability that the original trajectory will match the original correct detection box, thereby reducing ID switch in such situations.

Loading different advanced target tracking algorithms on the original YOLOv7 algorithm to track 7 video sequences of MOT17, the experimental results are shown in Table V. It can be seen that the MOTA of the improved BoTSORT target tracking algorithm reaches 42.2% and the IDs decreases to 410, achieving the best performance among similar tracking algorithms.

(a)before the meeting

(b)after the meeting

Fig. 9 Results analysis

TABLE V
COMPARATIVE EXPERIMENTAL RESULTS TABLE

Algorithm	IDF1	Rell	Prcn	IDs	MOTA	MOTP
SORT	0.496	0.496	0.840	708	0.395	0.330
DeepSORT	0.447	0.484	0.800	647	0.358	0.333
Bytetrack	0.500	0.531	0.820	458	0.410	0.333
BoTSORT	0.514	0.545	0.813	484	0.416	0.333
Improved BoTSORT	0.519	0.546	0.812	410	0.422	0.332

V. CONCLUSION

Aiming at the problem of pedestrian detection in complex traffic scenarios, an improved YOLOv7 combined with BoT-SORT method is proposed in this paper. A two-stage target tracking method is used for testing through the DBT paradigm. The YOLOv7 backbone network is improved and a feature transfer module is added, so as to improve the network detection accuracy and reduce the missed detection during the detection process. At the same time, the BoT-SORT tracking algorithm based on region segmentation is proposed, which optimizes generation process for tracking new trajectories and reduces the number of changes in IDs. The experimental results show that the combined detection effect of the improved YOLOv7 network and BoT-SORT is better than the original network model combination.

REFERENCES

[1] R. Barbosa, O. D. Ogobuchi, O. O. Joy, et al. “IoT based real-time traffic monitoring system using images sensors by sparse deep learning algorithm,” *Computer Communications*, vol. 2023, no. 210, pp. 321-330, 2023.

[2] Y. Zhang, H. Z. Lu, L. P. Zhang, et al., “Overview of visual multi-object tracking algorithms with deep learning,” *Computer Engineering and Applications*, vol. 57, no. 13, pp. 55-66, 2021.

[3] X. P. Chen, Y. Xu, “A multi-dimensional attention feature fusion method for pedestrian re-identification,” *Engineering Letters*, vol. 31, no. 4, pp. 1365-1373, 2023.

[4] Y. Wang, Y. Tian, “Research on pedestrian detection algorithm based on deep learning,” *IAENG International Journal of Computer Science*, vol. 50, no. 4, pp. 1446-1452, 2023.

[5] J. Sun, Z. Wang, "Vehicle and pedestrian detection algorithm based on improved YOLOv5," *IAENG International Journal of Computer Science*, vol. 50, no. 4, pp. 1401-1409, 2023.
[6] Y. Wang, Y. Li, Q. Han, "Vehicle-mounted infrared pedestrian tracking based on scale adaptive kernel correlation filter," *IAENG International Journal of Computer Science*, vol. 49, no. 2, pp. 349-356, 2022.
[7] X. Luo, R. Zhao, H. S. Zhuang, et al., "UAV multi-target tracking alg orithm jointly optimized by YOLOv5 and Deep-SORT," *Journal of Si gnal Processing*, vol. 38, no. 12, pp. 2628-2638, 2022.
[8] A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," arXiv preprint arXiv: 2004. 10934, 2020.
[9] Y. T. He, J. Che, J. M. Wu, "Pedestrian multi-target tracking method based on YOLOv5 and person re-identification," *Chinese Journal of Liquid Crystals and Displays*, vol. 37, no. 07, pp. 880-890, 2022.
[10] S. Woo, J. Park, J. Y. Lee, et al., "Cbam: convolutional block attention module," in *Proc. of the European Conference on Computer Vision*, 2018, pp. 3-19.
[11] S. Xie, R. Girshick, P. Dollár, et al., "Aggregated residual transformations for deep neural networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492-1500.
[12] W. S. Fan, J. Z. Lai, P. Lv , et al., "Vision/lidar object tracking and lo calization method based on improved DeepSORT," *Navigation Positi oning &. Timing*, vol. 9, no. 04, pp. 77-84, 2022.
[13] N. Wojke, A. Bewley, D. Paulus, "Simple online and real time tracking with a deep association metric," in *Proc. of the IEEE International Conference on Image Processing*, 2017, pp. 3645-3649.
[14] S. Q. Tu, X. L. Liu, Y. Liang, et al., "Behavior recognition and tracking of group-housed pigs based on improved DeepSORT Algorithm," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 53, no. 08, pp. 345-352, 2022.
[15] Y. L. Zhao, Y. G. Shan, J. Yuan, "Wearing mask pedestrian tracking based on improved YOLOv7 and DeepSORT," *Computer Engineering and Applications*, vol. 59, no. 06, pp. 221-230, 2023.
[16] K. Bernardin, R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1-10, 2008.
[17] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464-7475.
[18] N. Aharon, R. Orfaig, B. Z. Bobrovsky, "BoT-SORT: robust associations multi-pedestrian tracking," arXiv preprint arXiv:2206.14651,2022.
[19] S. Woo, S. Debnath, R. Hu, et al., "Convnext v2: co-designing and scaling convnets with masked autoencoders," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16133-16142.
[20] L. Yang, R. Y. Zhang, L. Li, et al., "Simam: a simple, parameter-free attention module for convolutional neural networks," in *Proc. International Conference on Machine Learning*, 2021, pp. 11863-11874.
[21] Z. Liu, H. Mao, C. Y. Wu, et al., "A convnet for the 2020s," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11976-11986.
[22] K. Han, Y. Wang, Q. Tian, et al., "Ghostnet: more features from cheap operations," in *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580-1589.
[23] Y. Zhang, P. Sun, Y. Jiang, et al., "Bytetrack: multi-object tracking by associating every detection box," in *European Conference on Computer Vision. Cham: Springer Nature Switzerland*, pp. 1-21, 2022.
[24] L. Zheng, L. Shen, L. Tian, et al., "Scalable person re-identification: a benchmark," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 1116-1124, 2015.
[25] S. Zhang, Y. Xie, J. Wan, et al., "Widerperson: a diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380-393, 2019.

Yanhui Lv was born in Changchun, Jilin, China in 1971. She received the B.S. and M.S. degrees in computer application technology from Shenyang Ligong University in 1993 and 2005, and the Ph.D. degree in computer application technology from Northeastern University in 2010.

Since 2012, she has been a professor with the College of Information Science and Engineering, Shenyang Ligong University. She is the author of two books, more than 30 articles, and more than 4 software copyrights. Her research interests include computer vision and artificial intelligence.

Her awards and honors include 3 Liaoning Province Natural Science Academic Achievement Awards. She was a recipient of 6 ministerial research projects fund and 2 provincial research project fund.

Qichao Guo was born in Zhangzi, Shanxi Province, China in 1995. He received the B.S. degrees in weapons systems engineering from North University of China in 2018. Now, he is a postgraduate student of Shenyang Ligong University. His research interests include computer vision and artificial intelligence.

Jiaxu Dong was born in Yuxian, Shanxi Province, China in 1995. He received the B.S. degrees in network engineering from Shenyang Ligong University in 2018. Now, he is a postgraduate student of Shenyang Ligong University. His research interests include computer vision and artificial intelligence.