 doi:10.20944/preprints202204.0254.v1

Article

Improved DeepSORT Algorithm Based on Multi Feature Fusion

Haiying Liu [1,†,‡]*, Yuncheng Pei [1,‡], Qiancheng Bei [1,] and Lixia Deng [1,]

[1] School of Information and Automation Qilu University of Technology (Shandong Academy of Sciences) Jinan, Shandong Province, 250353, China; haiyingliu2019@qlu.edu.cn
* Correspondence: haiyingliu2019@qlu.edu.cn;

Abstract: Pedestrian multi-target tracking technology plays an important role in artificial intelligence, driverless, virtual reality and other fields. The pedestrian multi-target tracking algorithm DeepSORT based on detection is widely used in industry. It mainly tracks multiple pedestrian targets continuously and keeps their ID unchanged. In order to improve the applicability and tracking accuracy of DeepSORT algorithm, this paper improved the IOU distance measurement in the matching process. At the same time, ResNet50 is used as the feature extraction backbone network, and combined with FPN (Feature Pyramid Network), the appearance features of multi-layer pedestrians are fused to improve the tracking accuracy of DeepSORT algorithm. The proposed algorithm is verified on the public data set MOT-16 and it's tracking accuracy is enhanced to 4.1%.

Keywords: multi-target tracking; DeepSORT; feature extraction; target detection

1. Introduction

With the great breakthrough of deep learning in computer vision in recent years, applying deep learning to pedestrian multi-target tracking and improving the accuracy of target tracking is the mainstream of multi-target tracking research[1]. The key steps of multi-target tracking algorithm are target apparent feature extraction, calculation of appearance similarity measure or distance measure between the newly detected target and the target in the trajectory, and prediction of motion trajectory[2]. At present, the most mature research direction of the application of deep learning technology in the field of pedestrian multi-target tracking is the extraction of the apparent features of tracked pedestrians[3].

AlexNet Network, proposed by Alex Krizhvsky[4], first applied convolution network to image feature extraction, and won the ImageNet competition that year, promoting people to enter the era of deep feature extraction. Alex Bewley et al[5]. proposed SORT algorithm, which brought deep learning into multi-target tracking for the first t ime. This method mainly uses Faster R-CNN[6] algorithm based on deep learning to detect pedestrians and get the pedestrian's position in the current video sequence. By using a deep learning based detector, combined with simple track prediction and data association algorithm, the accuracy of target tracking is greatly improved. At the same time, the algorithm achieves a speed of 60 Hz in the tracking process. Nicolai Wojke et al. proposed DeepSORT[7], which adds a pedestrian appearance feature similarity measure based on SORT algorithm, and combines cascade matching module to reduce the number of ID switching when pedestrians are occluded, and to improve the robustness of the model. Wang et al. embedded the feature extraction module of the multi-target tracker into the network of YOLOv3[8]. During the training process, the weights are directly shared with the detection network so that the tracker can use the target border output by the detector and the corresponding appearance features. This method improves the real-time performance in the tracking process. Zhang et al. proposed the FairMOT detection algorithm [9] based on the idea of JDE[8]. Because JDE directly uses the feature extraction network module of YOLOv3[10] algorithm, which leads to the existence of anchor frame. Meanwhile, the detection network and tracking algorithm have different features for pedestrian appearance, which makes them less accurate in dealing with dense pedestrians. FairMOT network uses the design of anchorless frame

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

of CenterNet[11]. At the same time, downsampling is used to extract the pedestrian's appearances, which greatly improves the tracking accuracy. To sum up, it can be seen that improving the accuracy of pedestrian appearance feature extraction is of great significance to improve the accuracy of pedestrian multi-target tracking.

This paper improves the tracking accuracy of deepsort algorithm by improving the feature extraction network model.

2. Analysis of related algorithms

2.1. Pedestrian Multiobjective Tracking Algorithm

DeepSORT tracking method is a single hypothesis target tracking algorithm based on Hungarian algorithm and Kalman filter algorithm. The state information of the target is represented by $(x, y, \gamma, h, v_x, v_y, v_\gamma, v_h)$, where (x, y) is the target center coordinate, γ is the target width-to-height ratio, v_x, v_y, v_γ, v_h is the moving speed of x, y, γ, h, Kalman filter algorithm is introduced to predict the location of the track in the detection space, and Hungarian matching algorithm is adopted to match the detection with the track.

Once the target tracking and detection are matched, the tracking and detection are matched using cascade matching. In the matching process, the appearance and motion information are combined to form a new measurement matrix for judging the matching degree between the aforementioned process . Appearance information uses a simple convolution neural network to extract the feature matrix r_j in each pedestrian detection frame, and stores all the appearance information in the last 100 frames of the track in R_i , with the feature matrix in each frame of the track being r_i . The minimum distance between the $i-th$ pedestrian trajectory feature and the$j-th$ pedestrian detection feature is calculated using the equation 1:

$$d_{(i,j)}^{(2)} = \min\left\{1 - r_j^T r_i^{(k)} \middle| r_i^{(k)} \in R_i, k \in (1, 100)\right\} \tag{1}$$

Motion information is expressed as the square value of the Mahalanobis distance between the predicted and detected locations of the Kalman trajectory, which is calculated by the equation 2:

$$d_{(i,j)}^{(1)} = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \tag{2}$$

Where$y_i$represents the projection of the predicted value of track $i-th$ in the detection space, d_j represents the $j-th$ detection target in the current detection space, S_i represents the covariance matrix of track $i-th$ in the detection space. The weighted sum of Mahalanobis distance based on motion information and cosine distance based on pedestrian appearance characteristics is used to form a new measurement. The formula for calculation is as follows equation 3:

$$c_{i,j} = \lambda d_{(i,j)}^{(1)} + (1 - \lambda) d_{(i,j)}^{(2)} \tag{3}$$

The DeepSORT algorithm uses cascade matching and crossover matching to match the new detection method to the trajectory predicted by the Kalman filter algorithm, and iterates over and over again to complete the target matching process.

2.2. Feature Pyramid Network

Feature Pyramid Network (FPN)[12] combines features from multiple levels to output the final feature map information, named after the feature pyramid network because of its pyramid-like structure and appearance. FPN is widely used in target detection, semantics segmentation, behavior recognition and other fields. It is important to improve model performance. Because of the feature fusion at different levels, the output feature map has multi-layer semantic information, which can extract a variety of semantic information when pedestrians extract multi-target features. The network structure is shown in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

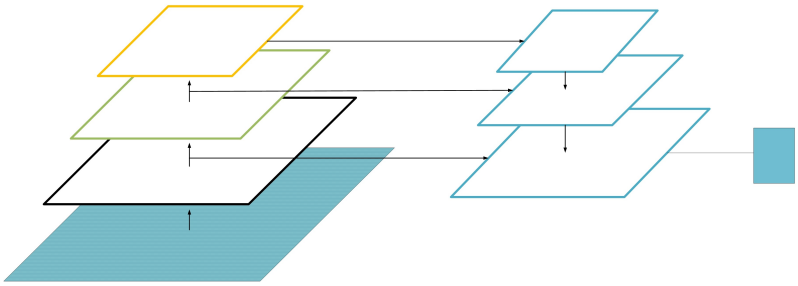


Figure 1. Feature pyramid network structure.

From Figure 1 of FPN network structure, it can be seen that when feature extraction is performed in the backbone network, feature information of different levels is finally fused by top-down fusion of feature maps. When extracting pedestrian target features, the size of the feature size has a great influence on the expression of pedestrian feature information. Using different depths of network structure to extract pedestrian appearance information with different sizes has different accuracy. Fusing pedestrian appearance feature information extracted from different depths of network can reduce the difference of feature information caused by scale changes in the matching process[13]. At the same time, the shallow feature extraction network is sensitive to the location information, while the deep feature extraction network is sensitive to the appearance information [14]. Combining the two, it can improve the probability of successful pedestrian detection and track matching in the tracking process.

2.3. Residual Network Structure

In the feature extraction network, as the network depth and structure complexity increase, the new layer will learn the same model parameters as the previous layer. Continuing to increase the network layer will not significantly increase the training network error, sometimes it will reduce, and increase the consumption of computing resources. In order to improve the training accuracy of the network and the diversity of feature information extraction, the residual network structure (ResNet)[15]] can be used to solve the problem of identical mapping in the network model. Its network module is shown in Figure 2.

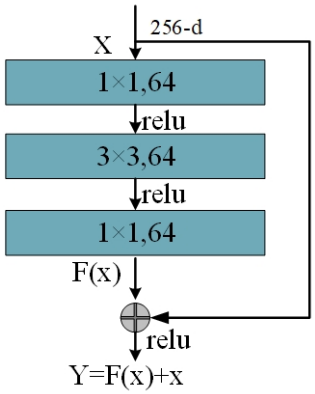


Figure 2. Structure diagram of residual module.

From Figure 2, the main idea of residual block is to sum the input x and $F(x)$ of the module as the output of Y the whole network, where $F(x)$ is the difference between the input and output, that is, residual. Compared with traditional directly connected convolution networks, ResNet have many input side branches connected to the output, making the result of network learning the residual information of the input and output. This network structure can better solve the problem of information loss in convolution calculation of traditional convolution neural network, protect the information integrity,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

simplify the difficulty of network learning, and improve the accuracy and accuracy of training.

3. Generalized Intersection over Union(GIOU)

The original intersection union ratio matching algorithm of DeepSORT algorithm may appear in the matching process. The IOU value between the trajectory prediction box and the detection box is the same, but there are different overlap positions. As shown in Figure 3.

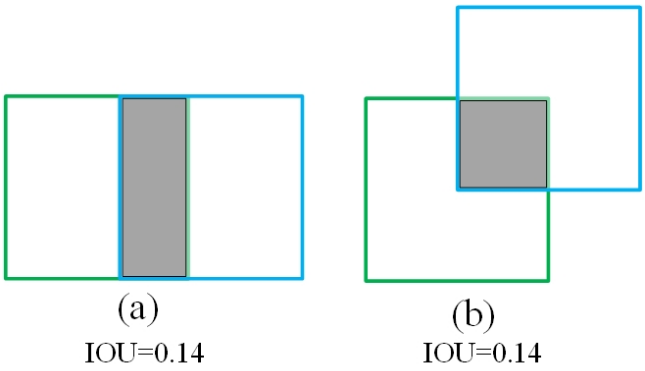


Figure 3. Coincidence position relationship between detection frame and prediction frame.

From Figure 3(a) and Figure 3(b), the Intersection over Union(IOU) values between the detection frame and the prediction frame are exactly the same, but the positional relationship between them is different, so the matching degree cannot be judged from the IOU distance.

For the purpose of overcoming this problem, position information is added to the distance measurement to judge how the detection frame and prediction frame intersect. Through the GIOU distance, the minimum bounding box of detection frame and prediction frame is introduced to solve the spatial position relationship between the two frames. The calculation process is shown in equation 4.

$$GIOU = IOU - \frac{C - A \cup B}{C} \tag{4}$$

Where, C is the minimum frame area used to surround the detection frame and prediction frame, A is the area of target trajectory prediction frame, and B is the area of pedestrian detection frame.

When the IOU values are the same and the overlap between the detection frame and the prediction frame is different, the GIOU measurement effect is shown in Figure 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

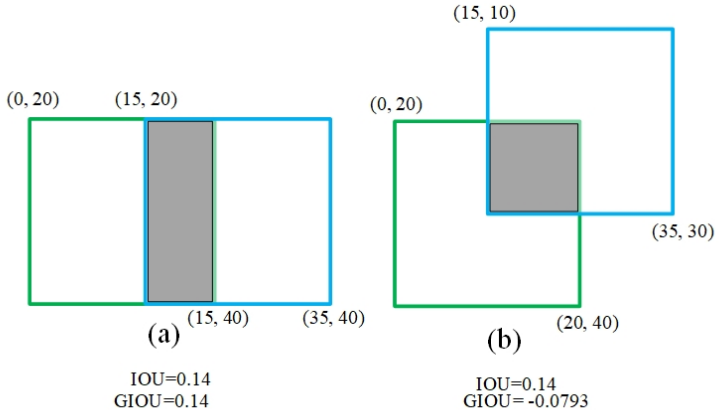


Figure 4. Effect drawing of GIOU measurement in different overlapping modes.

From Figure 4 that under different overlapping effects, the IOU value between the detection frame and prediction frame in Figure 4(a) and Figure 4(b) is 0.14, and the difference of matching measurement between them cannot be judged by IOU distance measurement. GIOU is different, GIOU in Figure 4(a) is 0.14 and GIOU in Figure 4(b) is -0.0793, which is better than that in Figure 4(a) below. As can be seen from Figure 4(a) and Figure 4(b), when IOU distance measurement cannot distinguish the matching degree between detection frame and prediction frame, GIOU distance measurement can well solve such problems.

In the pedestrian multi-target tracking algorithm, the GIOU distance measure is used as the cost matrix between detection and trajectory prediction. the phenomenon of false matching between detection and trajectory prediction can be reduced by introducing GIOU, and the phenomenon of pedestrian switching in the whole tracking process can be reduced.

4. The Improved the feature extraction network

In order to improve the performance of feature extraction network, ResNet50 is selected as backbone network. With the deepening of the network layer, the overall content of pedestrian appearance is richer, and the semantic information is more and more accurate. At the same time, for the case of integrating the features of different levels, improve the richness of pedestrian features output from the feature extraction network, improve the accuracy of the tracking process, and deal with the complex external environment. The backbone network structure of feature fusion is shown in Figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

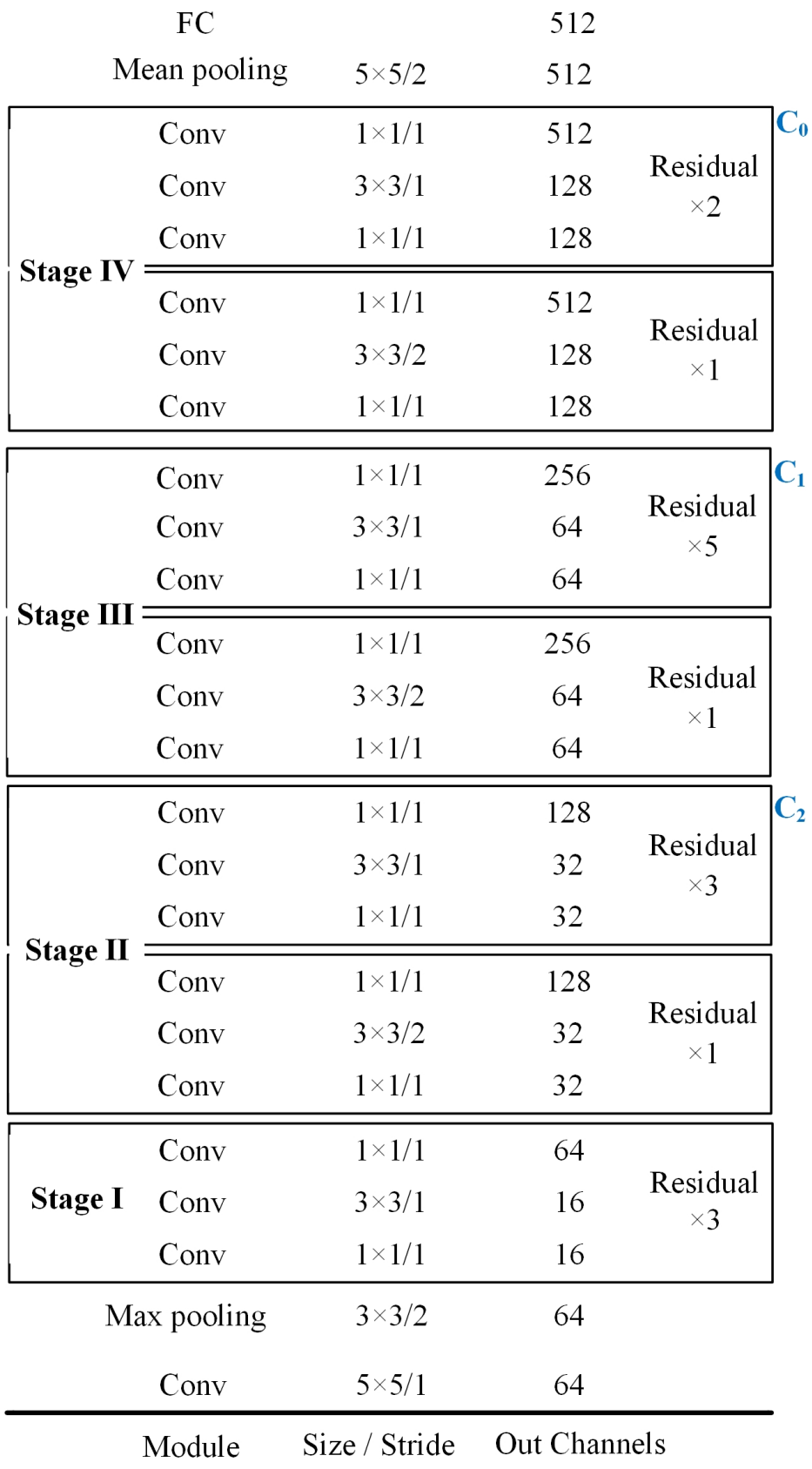


Figure 5. Feature fusion backbone network structures.

From Figure 5, the residual module of the backbone network is divided into four stages according to the number of output channels. The first stage is composed of three identical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

residual structures. In the second to fourth stages, the convolution with size 3×3 in the residual module of the first layer in each stage has a stride of 2, all other convolution stride are 1.The dimension of the final output feature of the whole pedestrian feature extraction network is 512, which can express more abundant pedestrian feature information than the 128 dimension of the original network. At the same time, at the end of the main network, at the end of the third and second stages, the output feature maps C_0, C_1, C_2 of each layer are extracted and passed to the FPN network. Among them, the output feature information of C_0layer has rich high-level semantic information, which has a good effect on the large target pedestrian feature extraction, but has a poor effect on the smaller pedestrian target feature extraction. Although the feature map information of C_1andC_2 output is not as rich as that of C_0, the output feature map information of C_0 is not as rich. However, it has a better extraction effect for smaller pedestrian targets. It may also retain some spatial information.

The extracted C0, C1, C2 are passed into the FPN module, and the network implementation process is shown in Figure 6.

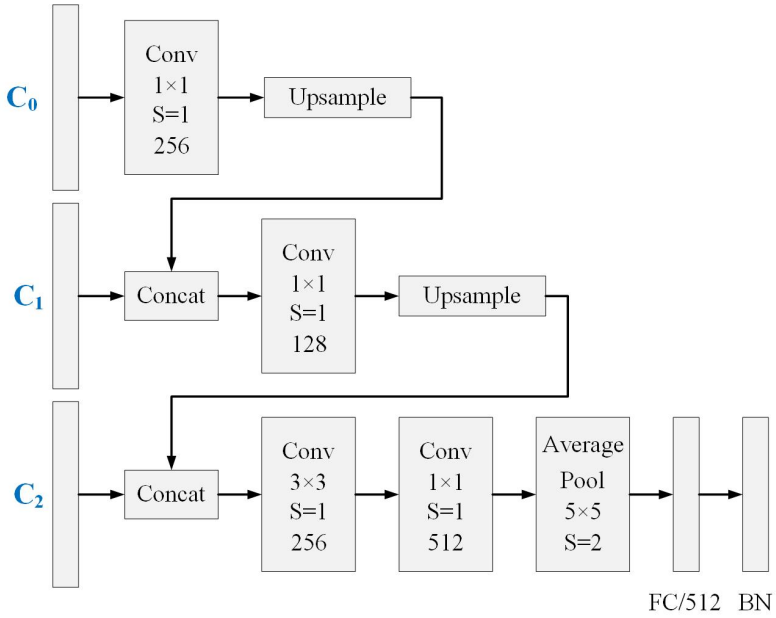


Figure 6. Multi-layer Feature Fusion Structure Diagram.

As shown in Figure 6, the feature maps C_0, C_1 and C_2 of different levels output by the backbone network are fused through the FPN network. C_0 is converted from 512 dimension to 256 dimension through a convolution operation with convolution size of 1×1 , stride of 1 and 256 output channels. After upsampling operation, it is spliced with C_1. After splicing, a convolution with the size of 1×1 , stride of 1 and the number of output channels of 128 is used for operation, and then it is spliced with C_2 after upsampling operation. After splicing, a convolution operation is performed with a convolution size of 3×3 , stride of 1, and an output channel of 256. After another convolution operation with a size of 1×1 , the output dimension is converted to 512. After the completion of the convolution calculation, through a mean pool layer with a size of 5×5 and a stride of 2, the final pedestrian appearance feature vector is obtained through full connection and batch normalization operations, which is used to calculate the appearance similarity measurement between the track and the detection.

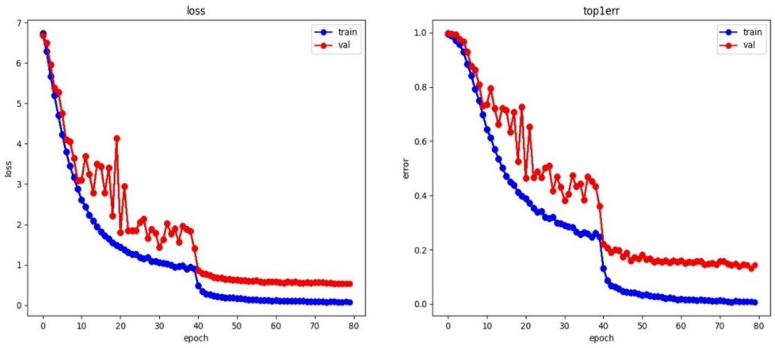
ResNet50 is used as the backbone network to extract the appearance information of pedestrian targets. In order to solve the unfriendly feature extraction of deep features for small target objects and the insensitive feature extraction of location information, this paper combines the FPN network to extract multi-layer features to solve the problem of deep network in feature extraction. After the whole feature extraction network has extracted pedestrian appearance features, The cost matrix is calculated by calculating the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 April 2022 doi:10.20944/preprints202204.0254.v1

feature similarity between pedestrian trajectory and detection combined with the motion feature information. The multi-target pedestrian tracking process is completed by cascade matching and GIOU matching. The improved feature extraction network can significantly improve the accuracy and accuracy of the tracking process.

5. Experimental results and analysis

Firstly, the improved feature extraction model, Market-1501 pedestrian recognition dataset, was trained for 80 epoch, with 64 samples in each batch. The training process is shown in Figure 7.



(a) Training Loss (b) Top1 Error

Figure 7. Feature Extraction Network Training Results.

In Figure 7(a) depicted the loss curve during feature extraction network training and prediction, and Figure 7(b) represents the Top-1 error rate change during feature extraction network training and prediction, where the blue segment represents the training phase and the red segment represents the prediction phase. As you can see from Figure 7(a), in the first 20-th epoch of training, the loss decreases very quickly, slows down gradually thereafter, accelerates suddenly in the 40-th epoch, then flattens out until the training loss in the 70-th epoch stabilizes, and then ends at the 80-th epoch. In the prediction stage, the downward trend of loss is similar to that in the training phase, but the first 40-th epoch of training will experience severe fluctuations due to inadequate model accuracy at the beginning of training. In Figure 7(b), the curve change law of Top1 error rate is similar to that of loss curve in Figure 7(a), and the training is also completed in the 80-th epoch.

The performance of the improved pedestrian multi-target tracking algorithm is verified under the MOT-16 data set, and the results are shown in Table I:

Table 1. Algorithmic performance comparison.

Algorithm	FP↓	FN↓	IDS↓	MOTA↑	MOTP↑
SORT	7318	32615	1423	33.4	72.1
DeepSORT	12852	36747	781	61.4	81.7
Our	11343	27874	627	63.9	81.0

From Table 1, it can be seen that MOTA and MOTP in SORT algorithm without feature extraction network are much lower than DeepSORT pedestrian tracking algorithm with feature extraction network, and the number of IDS is much higher than DeepSORT pedestrian tracking algorithm. Therefore, it can be seen that feature extraction network has a huge impact on improving the tracking effect. Compared with DeepSORT, the Our algorithm in the table reduced the number of FP by 1509, 11.7%, the number of FN by 36747 and 24.1%, which is the best evaluation parameter to improve, and the number of IDS by 154 and 19.7%. The final MOTA of Our's algorithm increases by 2.5 and 4.1% relative to DeepSORT, but the MOTP decreases by 0.86% and 0.7%. Overall, the improved tracking algorithm Our improved the accuracy of pedestrian recognition compared with DeepSORT algorithm.

6. Conclusions

This paper proposed a novel algorithm to enhance the accuracy, applicability and robust for pedestrian multi-target tracking. Specifically, we introduced the improved GIOU distance measure as the matching measure between the detection frame and the trajectory prediction frame, and used resnet50 as the feature extraction backbone network. Combined with FPN, it integrates the appearance features of multi-layer pedestrians to improve the tracking accuracy of DeepSORT algorithm by 4.1%.

References

1. Gong , X.; Le, Z.; Wu, Y.;et al. Real-Time Multiobject Tracking Based on Multiway Concurrency. *Sensors (Basel)* **2021**, *21*, 1–18.
2. Tathe, S.V.; Narote, A.S.; Narote, S.P. Face Recognition and Tracking in Videos. *Advances in Science Technology and Engineering Systems Journal* **2017**, *2*, 1238-1244.
3. Luo, W.; Xing, J.; Milan, A.; et al. Multiple object tracking: A literature review. *Artificial Intelligence: An International Journal* **2021**, *1293*, 103448.
4. Alex, K.; Ilya, S.; Geoffrey, E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **2017**, *60*, 84–90.
5. Bewley, A.; Zongyuan, G.; Ramos, F.; et al. Simple online and realtime tracking. *International Conference on Image Processing* **2016**, 3464–3468.
6. Ren, S.; He, K.; Girshick, R.; et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2017**, *39*, 1137–1149.
7. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. *2017 IEEE International Conference on Image Processing (ICIP)* **2017**,3645–3649.
8. Yu, F.; Li, W.; Li, Q.; et al. POI: Multiple object tracking with high performance detection and appearance feature. *Proceedings of European Conference on Computer Vision* **2016**, 36–42.
9. Jiaqian, C.; Minhong, J.; Wenyuan, W.; et al. Traffic flow detection based on yolov3 and deepsort. *Journal of metrology* **2021**, *42*, 718–723.
10. Jun, L.; Yaoru, W.; Guokang, F.; et al. Real-time detection tracking and recognition algorithm based on multi-target faces. *Multimedia Tools and Applications* **2021**, *80*, 17223–17238.
11. Yuqiao, G.; Weiyang, H., Zilong, Z. Pedestrian Target Tracking Based on DeepSORT with YOLOv5. *Proceedings - 2021 2nd International Conference on Computer Engineering and Intelligent Control, ICCEIC 2021.* **2021**, 1–5.
12. Ben, W.; Shuhan, C.; Jian, W.; et al. Residual feature pyramid networks for salient object detection. *The visual computer* **2020**, *36*, 1897–1908.
13. Kuang, Q. Face Image Feature Extraction based on Deep Learning Algorithm. *Journal of Physics* **2021**, *1852*, 032040.
14. Yu, C.; Cai, Q.; Huang, Q.; et al. An Image Defog Network Based on Multi-scale Feature Extraction and Weighting. *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)* **2021**, 423-427.
15. He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2016**, 770-778.