Multi-level temporal feature fusion with feature exchange strategy for multiple object tracking[*]

GE Yisu[1,2], YE Wenjie[1], ZHANG Guodao[3], and LIN Mengying[4]**

1. College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China

2. School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321019, China

3. Department of Digital Media Technology, Hangzhou Dianzi University, Hangzhou 310023, China

4. College of Intelligent Manufacturing, Wenzhou Polytechnic, Wenzhou 325035, China

(Received 12 January 2024; Revised 8 April 2024)

With the deepening of neural network research, object detection has been developed rapidly in recent years, and video object detection methods have gradually attracted the attention of scholars, especially frameworks including multiple object tracking and detection. Most current works prefer to build the paradigm for multiple object tracking and detection by multi-task learning. Different with others, a multi-level temporal feature fusion structure is proposed in this paper to improve the performance of framework by utilizing the constraint of video temporal consistency. For training the temporal network end-to-end, a feature exchange training strategy is put forward for training the temporal feature fusion structure efficiently. The proposed method is tested on several acknowledged benchmarks, and encouraging results are obtained compared with the famous joint detection and tracking framework. The ablation experiment answers the problem of a good position for temporal feature fusion.

Document code: A **Article ID:** 1673-1905(2024)08-0505-8

DOI https://doi.org/10.1007/s11801-024-4139-5

Video object detection and tracking is the basic problem in computer vision, and the cornerstone of visual analysis and understanding. The joint multiple object tracking and detection frameworks are proposed for efficient understanding the object movement in the video. In the practical application, the network input is a series of video frames, and the detection and tracking results of each object in frames are output. According to the tracking approach, these frameworks can be divided into two categories as follows.

Feature matching based tracking: The frameworks detect and extract object feature simultaneously, then associate targets in adjacent frames by feature matching. It finishes object detection and re-identification at the same time. In multiple object tracking, the frameworks based on feature matching can correctly pair off the targets in subsequent frames after the tracking loss in the current frame. So that the target motion trajectory can ultimately be traced, and the identification (ID) switch and the tracking loss problem caused by occlusion are released.

Position shift based tracking: Position shift based frameworks output the relative position shift of objects between the previous and current frames, and match the objects in different frames according to position matching. These frameworks integrate the information of correlative frames, and only utilize position matching in tracking, which is efficient and straightforward.

In the joint tracking and detection frameworks, how to effectively fuse the temporal feature between adjacent frames is an interesting issue. How can we improve the performance of detection and tracking by making good use of temporal consistency constraints between adjacent frames? Is it possible to fuse temporal feature by employing a simple fusion structure while maintaining the end-to-end network training? These are the main problems of this paper. To address this issue, a multi-level temporal feature fusion structure is advised, and a network training approach based on feature exchange is proposed for the end-to-end training of the temporal network, which improved the performance of multiple object tracking simply. The contributions of this paper are as follows.

The temporal feature are effectively utilized through the multi-level feature fusion structure, and the proposed method can boost the joint tracking and detection framework to reach better performance on public datasets.

[*] This work has been supported by the Zhejiang Provincial Natural Science Foundation of China (No.LQ24F020016), the General Scientific Research Fund of Zhejiang Provincial Education Department (No.Y202248776), the Wenzhou Science and Technology Plan Project (No.G20220035), and the Wenzhou Key Laboratory Construction Project (No.2022HZSY0048).

[**] LIN Mengying is a lecturer at the College of Intelligent Manufacturing, Wenzhou Polytechnic. She received her master degree in professional engineering (2019) from University of Sydney. Her research interests are mainly in smart materials, defect detection and artificial intelligence. E-mail: 2019291034@wzpt.edu.cn

The temporal network training approach based on feature exchange between adjacent frames is proposed, which makes the temporal feature fusion structure implement the end-to-end network training.

The multi-level feature fusion structure is studied, and the experiments search for the most suitable levels of feature fusion, which answer the problem of using the temporal feature appropriately to improve detection and tracking performance.

The temporal feature fusion in joint multiple object tracking and detection frameworks is discussed in this paper. Therefore, the joint multiple object tracking and detection frameworks and temporal feature fusion methods are introduced in the related works, respectively.

Joint multiple object tracking and detection frameworks have been paid more and more attention in recent years, due to the requirements of practical application in deep learning technology. FEICHTENHOFER et al[1] set up an architecture named ConvNet for simultaneous detection and tracking, using a multi-task structure for frame-based object detection and across-frame track regression. ZHANG et al[2] further researched the joint detection and tracking framework and put forward fairness of detection and re-identification in multiple object tracking (FairMOT). It was an anchor-free structure that made the object location more accurate in re-identification and helped to achieve state-of-the-art tracking performance. In the joint tracking and detection frameworks based on feature matching, another famous approach Chained-Tracker was proposed by PENG et al[3]. The siamese network structure was applied to extract the feature of adjacent frames, then multi-task learning was utilized with paired attentive regression. CenterTrack proposed by ZHOU et al[4] was a representative method in the frameworks based on position shift, which differed from the above methods. Some objects with low confidence might be discarded when occluded, leading to the loss of some tracks. To address this issue, ZHANG et al[5] proposed ByteTrack, an algorithm that sets detection boxes as tracking targets, and associates all detection boxes, not just high-scoring ones. With the development of transformer, the SeqTrack[6] utilized the sequence model in multi-object tracking and showed the great potential of this paradigm.

The detection based frameworks were efficient in practical application, but the temporal feature in network propagation are ignored. The transformer based methods linked the temporal relation by multi-head attention, but the calculation of transformer is not suitable for real-time application. Making good use of temporal feature is the key point in multi-object tracking. Therefore, we focus on the temporal feature fusion of joint multiple object tracking and detection frameworks, and attempt to fuse the temporal feature with an efficient structure.

In video processing methods, making use of the temporal feature can effectively improve video processing results due to temporal consistency constraints. Therefore, various temporal feature extraction and fusion methods were proposed in video processing and video object detection. LIU et al[7] combined the long short-term memory (LSTM) with single shot MultiBox detector (SSD) for video object detection, and ConvLSTM correlates the feature in each frame to improve the detection performance of the current frame. BERTASIUS et al[8] fused continuous video feature by deformable convolution to achieve state-of-the-art results on VID datasets. GUO et al[9] put forward a video object detection method that established the spatial correspondence between feature across frames by progressive sparse local attention (PSLA). TANG et al[10] paid attention to obtaining high quality object linking results for better classification, and extended prior methods by the cuboid proposal network with short tubelet detection and the short tubelet linking algorithm. Direct applying transformer in multi-object tracking can lead to high computational cost. XU et al[11] proposed a transformer-based tracking framework that incorporates dense representations in TransCenter, and aims to balance cross-frame context modeling and real-time inference.

These methods equipped the spatiotemporal feature fusion, extracting and utilizing the changing feature between frames effectively. However, the heavy computation of the feature fusion module and tedious training of the multi-step networks are unavoidable. Consequently, is there any solution with lower computation, and easy for training? Aiming at the above problems, a multi-level temporal feature fusion structure is advanced, which improves the multi-object tracking with end-to-end feature exchanging training strategy.

With further research of the convolution neural network, the multi-level feature improved the network detection and classification, widely used in detection and tracking. However, in the joint tracking and detection frameworks, the multi-level temporal feature fusion that extends the feature fusion in time dimension was less noticed. In order to further improve the tracking and detection in the frameworks, the multi-level temporal feature fusion structure is proposed. For training the temporal structure efficiently, the feature exchange training algorithm is studied. Because both CenterTrack and FairMOT methods are equipped with the DLA34 backbone[12], this paper focuses on the multi-level temporal feature fusion in DLA34 to improve the accuracy of multiple object tracking. The multi-level temporal feature fusion structure and feature exchange training method are introduced in detail as follows.

CenterTrack uses the image and heatmap of the previous frame as the input, then combines the feature from three sub networks, which completes the temporal feature fusion at the lower level. The FairMOT framework applies the heatmap for more accurate object location, and the embedded feature was extracted for feature matching, which fuses the temporal information at the higher level. So, where is the best position for temporal feature fusion, and how could the temporal feature be used efficiently?

The orange part in Fig.1 is the position using the temporal information. Similar to the chain structure in Chained-Tracker, the feature of the ($i-1$)-th frame in multi-level temporal feature fusion (MTFF) is transferred to the i-th fusion structure, adds to the feature in the i-th frame, and the fused feature in the i-th frame is obtained; the feature of the i-th frame delivers the feature to the (i+1)-th fusion structure and is added to the feature of the (i+1)-th frame, then the (i+1)-th fusion feature is gotten. Therefore, the previous features are reused in the current network, which merges the temporal information for better tracking performance. The output of the network feature in fusion structure $F_{(i-1)\to i}$ can be shown as

$$F_{(i-1)\to i}=f_i+f_{i-1}, \tag{1}$$

where f_i is the input feature in the i-th frame.

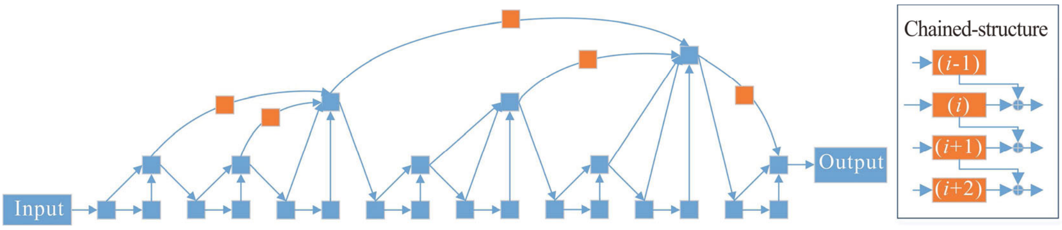

Fig.1 Multi-level temporal feature fusion structure

In the network with multiple branches, the feature from multiple branches will be aggregated by bottleneck structure; the bottleneck structure is also used to connect different sub networks. Therefore, a large amount of information in the network is aggregated and transmitted through the bottleneck structure, and temporal feature fusion in the bottleneck part can directly impact the performance. Hence, according to the 6-layer tree structure of DLA34, and considering the input and output of each tree structure, seven bottleneck positions should be considered. In the experiment, the temporal feature fusion positions were verified and discussed carefully, and we found that higher-level feature is more suitable for temporal fusion.

After the introduction of temporal feature fusion structure, how to train the above structure effectively should be solved, especially about how to realize the end-to-end network training with fast convergence. At present, most temporal feature fusion methods need to insert a special network layer for feature fusion, and train the fusion structure independently. Otherwise, like the Chained-Tracker, the temporal networks should be trained end-to-end by the siamese network structure. In the siamese network training, the backbone of network is trained twice in each propagation, while the feature fusion structure is trained only once, which leads to the imbalance of training between network structures. Different with the above methods, the proposed feature exchange training algorithm directly delivers the feature from the adjacent frames to fusion structures, and exchanges the adjacent feature for temporal feature fusion.

In the network training, batch samples were read at first, and the data was inputted into the network for forward propagation; then, the back propagation was carried out to correct the network parameters. Generally, the samples will be selected randomly to achieve better training results. However, for sequential data, random sampling will lose the correlation between video frames, which affects the temporal feature fusion. Therefore, it is necessary to provide sequential samples for training the temporal networks. Different sample selection and training strategies are illuminated in Fig.2.

For end-to-end training of the temporal network efficiently, the feature exchange training algorithm is proposed for temporal feature training. The algorithm divides the data in a batch into several groups; each group contains two adjacent frames, and the groups are selected randomly, which makes the data within each group have a strong correlation, and there is a big difference between each group. In training, the feature flows in the same group can be the fusion feature of each other, which looks like the feature exchange in each group. For example, in sequential training, the feature f_{i-1} is the temporal feature for f_i, and f_i should only be the temporal feature of f_{i+1}; but in the feature exchange algorithm, f_i can be the temporal feature for f_{i-1} as well, and the outputs of feature fusion are f_i+f_{i-1} and f_{i-1}+f_i, presented as

$$\begin{cases}F_{(i-1)\to(i)}=f_i+f_{i-1}\\F_{(i)\to(i-1)}=f_{i-1}+f_i\end{cases},\quad i=2\cdot j+1, j\in N. \tag{2}$$

Through feature exchange, the sample complexity in batch gets to the greatest extent, and the temporal feature fusion structure between adjacent frames can also be trained effectively. For example, when watching two adjacent frames, if there is no additional information or logical guidance, it is difficult to distinguish the order of these two frames. Therefore, after feature exchange training, the temporal feature in two different time directions can be trained simultaneously, improving the generalization ability of the network.

As we know, the data of detection and re-identification is discontinuous, which cannot be directly used in the sequential network training. However, in the feature exchange training algorithm, the discontinuous data are also applied in training, but the data loading is different from the sequential data. In training sequential datasets, the paired-off frames are the frames with close distance. However, the paired-off samples in discontinuous datasets use the same image with different augment parameters, as depicted in Fig.3. In Fig.3, the images on the left line are the paired-off samples in sequential datasets. The right line images are the data from discontinuous datasets, of which the paired-off images are the same. This little trick expands the training datasets, which makes the better performance in tracking.

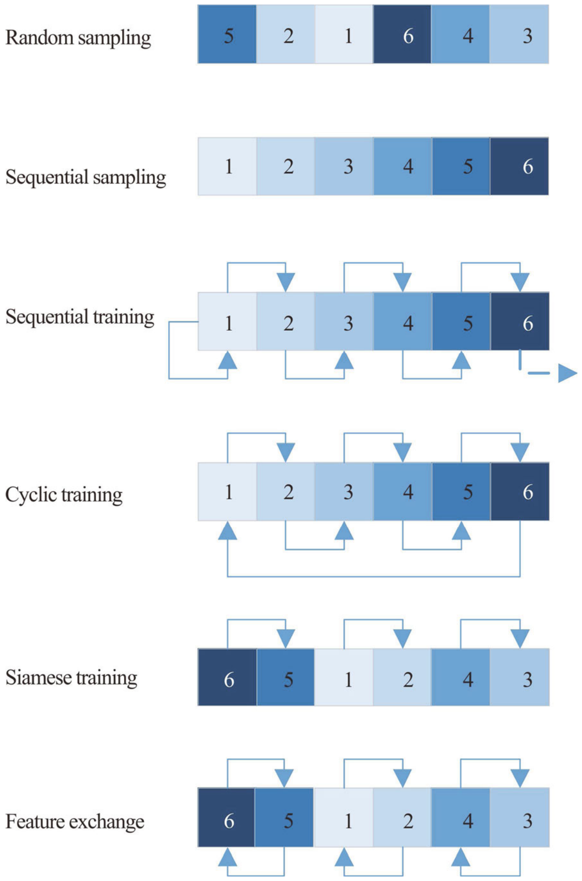

Fig.2 Different sample selection and training strategies

The experiments are mainly divided into three parts. Part one is the ablation experiment about the position of feature fusion, part two is the improvement of CenterTrack, which is tested on the public datasets, and part three is the verification of modified FairMOT. All the super parameter settings of training and testing are the same as those in CenterTrack and FairMOT, respectively. The hardware of the experiment is a computer with Intel i9-10900x central processing unit (CPU) and two NVIDIA Titan RTX graph processing unit (GPU). The experiment involves several datasets of multiple object tracking, pedestrian detection and pedestrian re-identification, including MOT15[13], MOT17[14], CrowdHuman[15], KITTI tracking[16], nuScenes[17], Caltech Pedestrian[18], CityPersons[19], CUHKSYSU[20], PRW[21] and ETHZ[22].

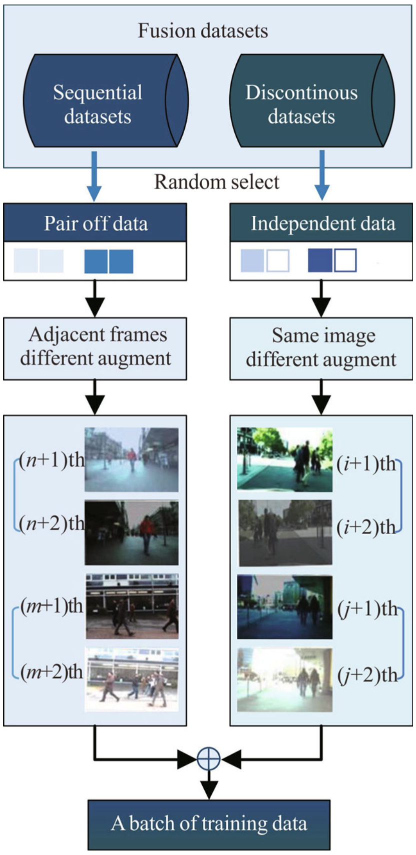

Fig.3 Paired-off samples on different training datasets

We use the official metrics for multiple object tracking, and the multi-object tracking accuracy (*MOTA*) is the common metric in each benchmark, which combines three error sources as

$$MOTA = 1 - \frac{\sum_t (FP_t + FN_t + IDSW_t)}{\sum_t GT_t}, \tag{3}$$

where FP_t means the false positives, FN_t denotes the false negatives, $IDSW_t$ represents the identity switch in frame t, and GT_t is the number of ground truth bounding boxes in frame t.

Multi-object tracking precision (*MOTP*) is the metric for evaluating the misalignment between the annotated and the predicted bounding boxes, which is an essential metric in joint multi-object tracking and detection frameworks. The higher *MOTP* means the tracker can locate the target better.

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t}, \tag{4}$$

where $d_{t,i}$ is the distance of two bounding boxes in frame t, and c_t is the correct matching number in frame t.

ID F1 score (IDF_1) is the ratio of correctly identified detections over the average ground-truth and computed detections.

$$IDF_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN}. \tag{5}$$

IDTP is the identification of true positives, *IDFP* means the identification of false positives, and *IDFN* represents the identification of false negatives.

In nuScenes dataset, the *AMOTA*[17] is proposed, which is a weighted average of *MOTA* across different output thresholds. The three parts of the main experiments are introduced as follows in detail.

In this part, MOT17 is used to find the answer about the most suitable position for temporal feature fusion. Half of the data in training sequences are selected as the training data, and the rest of the data in MOT17 is for validation. The networks are trained 30 epochs, and the initial learning rate is 0.000 1 which will be dropped by 10 in 20 epoch. The batch size is 12, the same as the original setting in CenterTrack. Part of the networks are pretrained, and part of them are trained from scratch. The position of feature fusion is presented in Fig.4, and each bottleneck is labeled for easy understanding. For making the biggest effect on feature fusion, the each paired-off bottlenecks is selected and tested respectively, as presented in Tab.1.

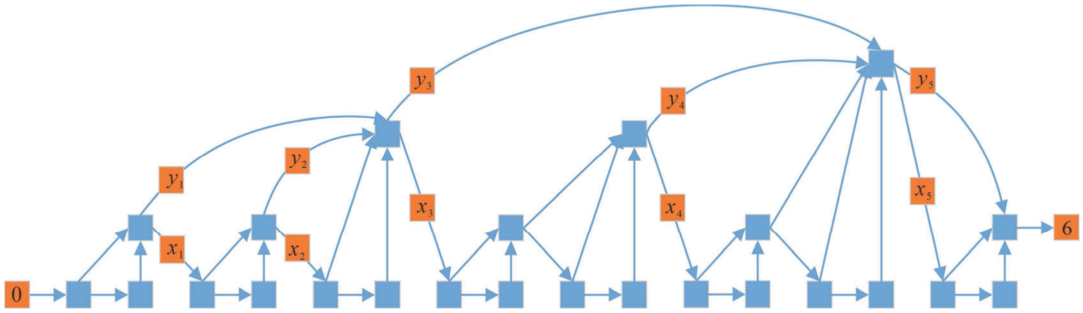

Fig.4 Position of feature fusion

Tab.1 Effect of feature fusion structure

Fusion position	Pre_image	Pre_hm	Pretrian	*MOTA*
0	√	√	×	39
x_1 and y_1	√	√	×	39
x_2 and y_2	√	√	×	40.7
x_3 and y_3	√	√	×	41.6
x_4 and y_4	√	√	×	60
x_5 and y_5	√	√	×	60.2
6	√	√	×	61.3
Null	√	√	×	60.7

There is an interesting result in Tab.1. In the lower level, the result of fusion shows negative optimization, and with the deepening of the network, the feature fusion structure improves the accuracy more obviously. Like the human visual system, overlapping two images, like the lower-level fusion, may confuse us. Still, watching two frames respectively, the correlation of objects in two frames can be obtained easily, which seems like the feature fusion at the higher level. Directly fusing the temporal feature at the lower level may affect the feature extraction in the rest of the network.

For evaluating the influence of each fusion structure on the bottlenecks, the ablation experiment on each fusion position is provided separately and the backbone is pretrained, as shown in Tab.2.

As we can see, the merge position in bottleneck x_1, x_2 and x_3 will decrease the performance of network, and x_4 and x_5 have little effect on tracking results. The fusion of these positions will affect further feature extraction, but the y_1 to y_5 and bottleneck 6 only aggregate the feature from multiple levels, showing excellent temporal feature fusion performance. Interestingly, the two additional sub networks proposed by CenterTrack can mostly improve the network performance. Then, based on the above observation, we find that CenterTrack combined with multi-level temporal feature can achieve the best performance, and the experimental results are shown in Tab.3. The original CenterTrack is compared with the CenterTrack equipped with the MTFF that fuses the feature of y_1 to y_5 and bottleneck 6. Most metrics in Tab.3 present that the CenterTrack with MTFF can get better performance. The above networks were pretrained on the dataset CrowdHuman, and were fine-tuned on half of the MOT17 dataset. The framework with MTFF is 1.6% higher than the original one on *MOTA*.

Tab.2 Influence of each bottleneck structure

Fusion position	Pre_image	Pre_hm	Pretrian	*MOTA*
0	×	×	√	39
x_1	×	×	√	39
x_2	×	×	√	40.7
x_3	×	×	√	41.6
x_4	×	×	√	60
x_5	×	×	√	60.2
y_1	×	×	√	64
y_2	×	×	√	63.8
y_3	×	×	√	63.7
y_4	×	×	√	64.4
y_5	×	×	√	63.8
6	×	×	√	63.6
DLA	×	×	√	63.8
Null	×	√	√	66.3
Null	√	×	√	64.5
Null	×	×	√	59.5

Following the experiments of CenterTrack, the experiment on benchmarks will be listed in this part. Same as the CenterTrack, the result of the MOT17 test benchmark will be submitted and presented in Tab.4. Most metrics on the MOT17 test show that the CenterTrack with MTFF is better than the original one, which means the optimization of MTFF is effective, and the *MOTA* is 1.3% higher.

In addition, the CenterTrack with MTFF structure is compared with the original CenterTrack on the KITTI tracking dataset as well. In the training process of CenterTrack, the nuScenes detection dataset is applied for pre training, then the pretrained network is fine-tuned on the training set of KITTI. The results are displayed in Tab.5. In this part, the pretrained network is provided by CenterTrack due to the large training consumption on nuScenes detection, and the MTFF is only trained on KITTI tracking. In Tab.5, the results of the two frameworks are within the same range, the MTFF-based network is 3.62% *MOTA* better on the pedestrian benchmark, but the original one is 1.88% *MOTA* higher on the car benchmark. Maybe the modified network can obtain better results after the pretraining on the nuScenes detection dataset.

In the same way as the experiments of CenterTrack, the nuScenes dataset was also selected as the comparison dataset. The network is pretrained on nuScenes detection at first, and MTFF is also trained on nuScenes tracking. Due to the different training equipment, the training parameters are a little different within CenterTrack, of which the batch size reduced to 32, the initial learning rate is 0.000 125, the learning rate will be dropped by 10 in 60 epochs, and the network is fine-tuned 70 epochs. As shown in Tab.4, the modified CenterTrack has improved on *AMOTA*, and is 0.4 better.

In addition to comparing the CenterTrack, the FairMOT was modified to prove the effect of the multi-level temporal feature fusion structure because DLA34 is also equipped in FairMOT. Similar to the previous experiments, the training parameter and data are the same as in FairMOT, only the multi-level temporal feature fusion of DLA34 and the training algorithm in FairMOT are modified. In network training, FairMOT inherits the training dataset of JDE and the mixed dataset named MIX which includes Caltech Pedestrian, CityPersons, CUHKSYSU, PRW, ETHZ, MOT16, and MOT17. Therefore, we also use the same experimental method to train the modified FairMOT on the MIX dataset. The network is pretrained on CrowdHuman with 60 epochs, then trained on MIX about 30 epochs to obtain the basic network.

We also compared two networks on public benchmarks MOT15, MOT16, and MOT17, the experimental results are encouraging. All the results are listed in Tab.7. Similar to the works on FairMOT, the MOT15 training dataset is used for training in MOT15 benchmarks, and the experiments on MOT16 and MOT17 do no more work. On MOT15, the modified FairMOT boosts a little on *MOTA*, *MOTP*, and IDF_1, of which the *MOTA* is 0.4% higher, and the original FairMOT is better on ID switch. However, the FairMOT+MTFF got the better performance on *MOTA* and IDF_1 in benchmark MOT15. On MOT16, the *MOTA* in FairMOT+MTFF is 5.2% better than the original one, and except for the ID switch, the other metrics on FairMOT+MTFF are still better than FairMOT. An interesting result on MOT17 shows that FairMOT is better on *MOTA* and *MOTP*, but the IDF_1 and ID switch of FairMOT+MTFF are better. And the IDF_1 of FairMOT+MTFF got the better score in the MOT17 benchmark.

Tab.3 CenterTrack-based results on half of MOT17 training dataset

Method	Benchmarks	Training 1	Training 2	*MOTA*↑	*MOTP*↑	IDF_1↑	*FP*↓	*FN*↓
CenterTrack	MOT17-half	CrowdHuman	MOT17-half	66.1	82.1	64.2	2 453	15 287
CenterTrack+MTFF	MOT17-half	CrowdHuman	MOT17-half	**67.7**	**82.4**	**65.1**	**2 317**	**14 543**

Tab.4 CenterTrack-based results on MOT17 test

Method	Benchmarks	Training 1	Training 2	*MOTA*↑	*MOTP*↑	IDF_1↑	*IDSW*↓
CenterTrack	MOT17	CrowdHuman	MOT17-half	67.8	78.4	**64.7**	**3 039**
CenterTrack+MTFF	MOT17	CrowdHuman	MOT17-half	**69.1**	**79.3**	59.6	5 409

Tab.5 CenterTrack-based results on KITTI tracking test

Method	Benchmarks	Training 1	Training 2	*MOTA*↑	*MOTP*↑	*IDSW*↓	*FRAG*↓	F1↑
CenterTrack	Pedestrian	nuScenes detection	KITTI tracking	55.34	74.02	**95**	751	74.98
CenterTrack	Car	nuScenes detection	KITTI tracking	**89.44**	85.05	**116**	**334**	**95.41**
CenterTrack+MTFF	Pedestrian	nuScenes detection	KITTI tracking	**58.96**	**75.02**	98	**744**	**76.47**
CenterTrack+MTFF	Car	nuScenes detection	KITTI tracking	87.56	**85.30**	203	480	94.54

Tab.6 CenterTrack-based results on nuScenes

Method	Benchmarks	Training 1	Training 2	*AMOTA*↑	*AMOTP*↑	*Recall*↑	*IDSW*↓
CenterTrack	nuScenes tracking	nuScenes detection	nuScenes tracking	6.8	1.543	0.222	**2 673**
CenterTrack+MTFF	nuScenes tracking	nuScenes detection	nuScenes tracking	**7.2**	**1.562**	**0.258**	2 986

Tab.7 FairMOT-based results on MOT

Method	Benchmark	Training 1	Training 2	Training 3	*MOTA*↑	*MOTP*↑	IDF_1↑	*IDSW*↓
FairMOT	MOT15	CrowdHuman	MIX	MOT15	60.6	76.5	64.7	**591**
FairMOT	MOT16	CrowdHuman	MIX	Null	69.3	80.2	72.3	**815**
FairMOT	MOT17	CrowdHuman	MIX	Null	**73.7**	**81.3**	72.3	3 303
FairMOT+MTFF	MOT15	CrowdHuman	MIX	MOT15	**61.0**	**77.1**	**65.6**	687
FairMOT+MTFF	MOT16	CrowdHuman	MIX	Null	**74.5**	**81.1**	**72.5**	957
FairMOT+MTFF	MOT17	CrowdHuman	MIX	Null	71.2	80.9	**74.4**	**3 051**

In the above experiments, most of the *MOTA* and IDF_1 on MTFF-based framework get the better performance, which means the MTFF with feature exchange training algorithm can help the joint multiple object tracking and detection frameworks to be more powerful, both on feature matching based tracking and position shift based tracking. But the ID switch of most MTFF-based frameworks is increased, which denotes that the MTFF structure helps the network to have a better tracking location. However, the link of tracking is blurred.

Aiming at the problem of temporal feature fusion in joint multiple object detection and tracking framework, a simple and efficient temporal feature fusion structure is proposed in this paper, which needs not too many additional modifications, and realizes the temporal feature fusion by delivering the feature from the previous frame. For training the temporal networks effectively, the feature exchange training algorithm is designed, which is an end-to-end training strategy. The experiments show that the deeper network feature is much more suitable for temporal feature fusion, and CenterTrack with the multi-level temporal feature fusion structure, can perform better after training the whole framework from scratch. This means the proposed optimal method is effective. The experiment of modified FairMOT obtains good results on acknowledged benchmarks, proving the effective of the proposed method.

The proposed structure in this paper is pretty simple, maybe it is efficient, but not good enough. Therefore, studying the better temporal feature fusion structure is our ongoing work.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Detect to track and track to detect[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 3038-3046.

[2] ZHANG Y, WANG C, WANG X, et al. FairMOT: on the fairness of detection and re-identification in multiple object tracking[J]. International journal of computer vision, 2021, 129: 3069-3087.

[3] PENG J L, WANG Q, WANG X. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//16th European Conference on Computer Vision, August 23-28, 2020, Glasgow, UK. Heidelberg: Springer, 2020: 145-161.

[4] ZHOU X, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]//16th European Conference on Computer Vision, August 23-28, 2020, Glasgow, UK. Heidelberg: Springer, 2020: 474-490.

[5] ZHANG Y, WANG C, WANG X, et al. Bytetrack: multi-object tracking by associating every detection box[C]//17th European Conference on Computer Vision, October 24-28, 2022, Tel Aviv, Israel. Heidelberg: Springer, 2022: 1-21.

[6] CHEN X, PENG H, WANG D, et al. SeqTrack: sequence to sequence learning for visual object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2023, Vancouver, Canada. New York: IEEE, 2023: 14572-14581.

[7] LIU M, ZHU M. Mobile video object detection with temporally-aware feature maps[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 5686-5695.

[8] BERTASIUS G, TORRESANI L, SHI J. Object detection in video with spatiotemporal sampling networks[C]//15th European Conference on Computer Vision, September 8-14, 2018, Munich, Germany. Heidelberg: Springer, 2018: 331-346.

[9] GUO C, ZHENG N, TAN Y, et al. Progressive sparse local attention for video object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019, Seoul, Korea. New York: IEEE, 2019: 3909-3918.

[10] TANG P, WANG C, WANG X, et al. Object detection in videos by high quality object linking[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(5): 1272-1278.

[11] XU Y, BAN Y, DELORME G, et al. TransCenter: transformers with dense representations for multiple-object tracking[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(6): 7820-7835.

[12] YU F, WANG D, SHELHAMER E, et al. Deep layer aggregation[C]//Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 2403-2412.

[13] LEAL-TAIXÉ L, MILAN A, REID I, et al. Motchallenge 2015: towards a benchmark for multi-target tracking[EB/OL]. (2015-04-01) [2023-12-23]. https://arxiv.org/abs/1504.01942.

[14] MILAN A, LEAL-TAIXÉ L, REID I, et al. MOT16: a benchmark for multi-object tracking[EB/OL]. (2016-03-01) [2023-12-23]. https://arxiv.org/abs/1603.00831.

[15] SHAO S, ZHANG Y, ZENG W, et al. Crowdhuman: a benchmark for detecting human in a crowd[EB/OL]. (2018-05-01) [2023-12-23]. https://arxiv.org/abs/1805.00123.

[16] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 3354-3361.

[17] CAESAR H, BANKITI V, LANG A, et al. Nuscenes: a multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 11621-11631.

[18] DOLLÁR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: a benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 304-311.

[19] ZHANG S, BENENSON R, SCHIELE B. Citypersons: a diverse dataset for pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3213-3221.

[20] XIAO T, LI S, WANG B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3415-3424.

[21] ZHENG L, ZHANG H, SUN S, et al. Person re-identification in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1367-1376.

[22] ESS A, LEIBE B, SCHINDLER K, et al. A mobile vision system for robust multi-person tracking [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE, 2008: 1-8.