Neurocomputing 476 (2022) 75–86

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

MAT: Motion-aware multi-object tracking

Shoudong Han *, Piao Huang, Hongwei Wang, En Yu, Donghaisheng Liu, Xiaofeng Pan

National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, 430074 Wuhan, China

ARTICLE INFO

Article history:
Received 16 February 2021
Revised 3 September 2021
Accepted 29 December 2021
Available online 3 January 2022
Communicated by Zidong Wang

Keywords:
Multi-object tracking
Motion-based prediction
Trajectory reconnection
Data association

ABSTRACT

Modern multi-object tracking (MOT) systems usually build trajectories through associating per-frame detections. However, facing the challenges of camera motion, fast motion, and occlusion, it is difficult to ensure the quality of long-range tracking or even the tracklet purity, especially for small objects. Most of tracking frameworks depend heavily on the performance of re-identification (ReID) for the data association. Unfortunately, the ReID-based association is not only unreliable and time-consuming, but still cannot address the false negatives for occluded and blurred objects, due to noisy partial-detections, similar appearances, and lack of temporal-spatial constraints. In this paper, we propose an enhanced MOT paradigm, namely Motion-Aware Tracker (MAT). Our MAT is a plug-and-play solution, it mainly focuses on high-performance motion-based prediction, reconnection, and association. First, the nonrigid pedestrian motion and rigid camera motion are blended seamlessly to develop the Integrated Motion Localization (IML) module. Second, the Dynamic Reconnection Context (DRC) module is devised to guarantee the robustness for long-range motion-based reconnection. The core ideas in DRC are the motion-based dynamic-window and cyclic pseudo-observation trajectory filling strategy, which can smoothly fill in the tracking fragments caused by occlusion or blur. At last, we present the 3D Integral Image (3DII) module to efficiently cut off useless track-detection association connections using temporal-spatial constraints. Extensive experiments are conducted on the MOT16&17 challenging benchmarks. The results demonstrate that our MAT can achieve superior performance and surpass other state-of-the-art trackers by a large margin with high efficiency.

1. Introduction

Multi-object tracking (MOT) plays a crucial role in scene understanding tasks for video analysis. It aims to estimate the trajectories of objects and associate them with per-frame detection results in either online or offline way. With the rapid development of object detection task, tracking-by-detection becomes the preferred MOT paradigm.

Proverbially, the core part of tracking-by-detection paradigm is data association [44]. Some methods like Tracktor++ [2], with the assumption of high frame rate and low pedestrian speed, tending to simply use the adjacent-frame detections for spatial-scale association. However, most tracking scenarios contain a variety of challenges, such as camera motion, fast motion, and occlusion. When that happens, the detector may fail in outputting high-quality detections steadily, and cause large displacements of objects between adjacent frames. Due to these challenges, the simple spatial-scale association is difficult to guarantee the quality of long-range tracking or even the tracklet purity, especially for small objects.

To alleviate the spatial-scale inconsistency from the large displacement or long occlusion, most tracking frameworks adopt the re-identification (ReID) for long-range feature-level association. For example, DeepSort [27] used a trained ReID model to perform appearance feature extraction, and FairMOT [33] included a 512-Dim ReID embedding head after backbone. But the ReID calculation is usually time-consuming, and ReID-based trackers rely too much on the performance of detector. The extracted appearance features may be very unreliable in some complex scenes, such as the noisy background interferences, partial-detections, blur or similar appearances, and lack of temporal-spatial constraints. Moreover, these issues may further cause a mass of false associations in terms of the Identity Switches, and the false negatives from occlusion or blur still cannot be addressed by ReID.

In this paper, we abandon the ReID-based association, due to its requirement of high-quality detections. Instead, we propose an enhanced MOT paradigm named as Motion-Aware Tracker

* Corresponding author.
E-mail address: shoudonghan@hust.edu.cn (S. Han).

https://doi.org/10.1016/j.neucom.2021.12.104
0925-2312/

(MAT), which mainly focuses on the motion-aware designs of prediction, reconnection, and association. Our MAT is a plug-and-play solution, and its tracking framework consists of three key modules, namely the Integrated Motion Localization (IML), Dynamic Reconnection Context (DRC), and 3D Integral Image (3DII). In detail, the IML module is designed for the joint prediction of nonrigid pedestrian motion and rigid camera motion. To ensure the robustness of long-range tracking, the DRC module determines the motion-based reconnection windows dynamically for different objects, based on the estimated Kalman velocity and our defined intensity of camera motion. Besides, when existing tracking fragments caused by occlusion or blur, the DRC adopts the cyclic pseudo-observation trajectory filling strategy to fill in them by smoothing the predictions of IML nonlinearly. The 3DII module is presented to efficiently cut off useless connections among all tracks and detections during the data association stage, by imposing temporal-spatial constraints on each connection. Only the tracks within partial-covered region of each detection will be regarded as the matched association candidates.

Comprehensively, our contributions are summarized as follows:

- We develop the IML module, which blends the nonrigid pedestrian motion and rigid camera motion seamlessly to balance their compatible issues.
- We devise the DRC module, which contains motion-based dynamic-window and cyclic pseudo-observation trajectory filling strategy. It can ensure the robustness for long-range motion-based reconnection.
- We present the 3DII module, which can impose temporal-spatial constraints to filter useless track-detection association connections efficiently.
- Incorporating the power of the IML, DRC, and 3DII modules, we propose the Motion-Aware Tracker (MAT). It is a plug-and-play solution, which can be easily embedded into other tracking-by-detection MOT trackers with the state-of-the-art performance. Hence our MAT is also an enhanced MOT paradigm.

2. Related work

Existing tracking-by-detection MOT methods can be divided as follows:

ReID-based MOT. Thanks to the great progresses of object detection over the past few years, most tracking-by-detection MOT methods first locate the positions of objects through detector, then extract the corresponding appearance features via ReID embedding for long-range matching or per-frame data association. Tracktor++ utilized the bounding box regression of Faster R-CNN [24] detector to modify the predicted position of a trajectory in next frame, and associated the deactivated tracks with detections using a ReID sub-network. CTracker [23] extended single-frame regression to adjacent-frame paired regression, and treated a ReID module as identity-attention to combine the detection and tracking into an end-to-end chained framework. CenterTrack [34] incorporated CenterNet [35] as detector, and used the spatial distance of track and detection center points to calculate the similarity for data association. Furthermore, FairMOT integrated the CenterNet and a ReID branch into one framework for joint spatial-embedding association, which greatly reduced the Identity Switches and still maintained a very high tracking speed.

However, most tracking-by-detection trackers fail to model the temporal-spatial relationships of large displacement detections when camera motion, fast motion, and occlusion challenges occur. Besides, due to noisy detections and similar appearances, the ReID-based data association may be highly unreliable, and the missing detections still cannot be retrieved. In contrast, our MAT mainly focuses on the motion-based prediction, reconnection and association, so long-range tracking can be more reliable when facing above mentioned challenges.

Motion-based MOT. Several trackers utilize motion to predict temporal-spatial variations of trajectories and compensate the detector failures in complex scenes. The motion patterns in video can be roughly summarized as the nonrigid motion (objects like pedestrians) and rigid motion (changing camera pose). The nonrigid motion is usually modeled by constant velocity model [8]. In [30], trajectories were smoothed by the observation-based Gaussian distributions. Recently, the Kalman Filter tends to be more popular for prediction, by using the provided detections as observations [4,27]. Moreover, the social force models were applied due to the complex pedestrian motion in crowded scenarios [20]. For rigid motion caused by changing camera pose, researchers study it in two main directions. One is based on 3D-information, such as Ego-motion [25] and SFM [8]. The other is based on affine transformation. Besides, the conditional probability model with recurrent neural network [13] was proposed to predict the positions and shapes of objects in next frame. In [40], a probabilistic autoregressive motion model was introduced to score tracklet proposals for detection assignments and inpaint tracklets containing missing detections. Moreover, the single object tracking (SOT) methods [42,43] were widely adopted or extended to search multiple objects directly in some MOT trackers like [9,10,15,17,28,36]. For instance, the SiamMOT [41] included a region-based Siamese tracker in SOT to model instance-level motion across frames.

While in above motion-based methods, the pedestrian motion and camera motion were modeled and used separately, and the motion states of objects cannot be accurately estimated. Differently, our MAT designs the IML module to blend the pedestrian motion and camera motion seamlessly to balance their compatible issues. Besides, the camera motion state can be estimated by calculating our defined metric, namely the intensity of camera motion.

3. Proposed Method

In this work, we propose an enhanced MOT paradigm named as MAT, which can be easily embedded into any tracking-by-detection tracker. Our MAT architecture is illustrated in Fig. 1. It contains three motion-based modules, namely the IML module for track prediction, the DRC module for trajectory reconnection, and the 3DII module for fast temporal-spatial association. In addition, like the regression-based tracker Tracktor++, we employ a re-implemented Cascaded RCNN as detector. Notice that our MAT mainly devotes to enhancing and cooperating with the tracking part of a tracker, and it can be completely independent from the detector. To improve the recall rate of detector, here we treat the IML predictions as self-defined proposals and pass them into detector to obtain corrected detections.

3.1. Integrated motion Localization (IML)

If the detection results are of high-quality and the motion is simple, the data association using IoU (Intersection-over-Union) metric can obtain admirable tracking results. However, if there are camera motion, fast motion, occlusion, or low frame rate challenges, without advanced motion compensation, the IoU-based trackers may fail when facing long-range tracking, especially for small objects.

The observable individual motion can be expressed as the combination of rigid camera motion and nonrigid pedestrian motion. For the different motion patterns, we should consider their temporal-spatial consistency, and balance the compatible issues.

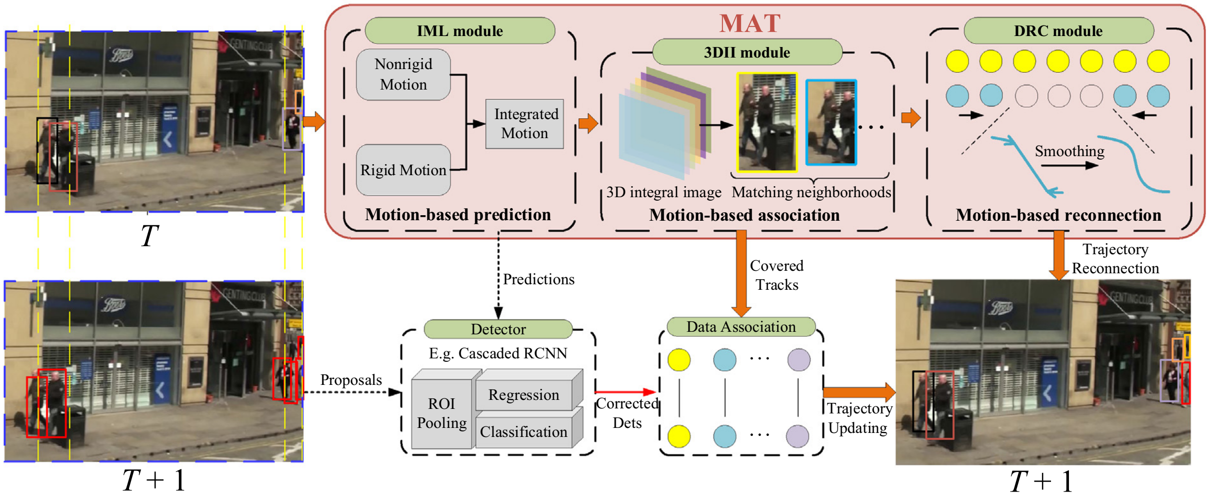

Fig. 1. Illustration of our MAT architecture. For a given frame T, the IML module is applied to predict each track's position in frame $T + 1$ considering both pedestrian and camera motions. Then, based on the 3DII module, all tracks within partial-covered region of each corrected detection will be efficiently matched. Finally, for deactivated trajectories, the DRC module determines their reliable reconnection windows and fills in the tracking fragments smoothly.

Therefore, the aim of our IML module is to blend the camera and pedestrian motions seamlessly and efficiently.

For camera motion, the pixel alignment among sequential frames can usually be established by the epipolar geometry (Ego) constraints or affine transformations. With the assumption that the objects in adjacent frames have slow motions and constant shapes, the camera states can be formulated as an optimization problem using Ego constrains. However, the fundamental matrix in Ego model needs to be estimated by feature point matching. It will be seriously disturbed by textural pedestrian parts, since most feature points are located in the areas full of gradient information. In contrast, for approaching the changes of exterior parameters caused by camera motion, the global affine transformation is more suitable and robust. Thus, the lightweight Enhanced Correlation Coefficient Maximization (ECC) [12] model is employed to estimate the global rotation and translation. Note that some methods [37–39] used relative motion among objects to alleviate the impact of camera motion, their networks are complex and time-consuming, not nearly as simple and efficient as ECC. Additionally, for the high efficiency and flexibility, the Kalman Filter is our best choice to model pedestrian motion.

Consequently, our IML module adopts the ECC to estimate the camera motion, and considers both the Kalman Filter prediction and global ECC alignment. Since pedestrian motion often plays the principal role in most MOT scenarios, it is always processed by default before the camera motion. In detail, each object's position needs to be predicted by the Kalman Filter first and then aligned by the ECC model. Thus, our IML can be simply established as below.

$$\begin{cases} s_{t+1} = warp(Fs_t) \\ P_{t+1} = FP_tF^T + Q \end{cases} \tag{1}$$

where F, Q, P, s respectively denote the state transition matrix, process uncertainty, covariance matrix, and state estimation of the Kalman Filter. Here $s = [Box, V_{Box}]^T$, Box represents the predicted bounding box, and V_{Box} contains the estimated velocities of all elements in Box. Besides, the $warp$ operation denotes the ECC alignment, notice that it is only applied to the Box part of Fs_t.

3.2. Dynamic reconnection context (DRC)

When severe occlusion or blur happens, it will inevitably lead to plenty of missing detections, and force the trajectories to be deactivated temporally. To ensure the robustness of long-range tracking, in this subsection, we devise the DRC module for motion-based reconnection, based on the estimated motion states and predictions of IML module. Our DCR module consists of two parts, namely the motion-based dynamic-window and cyclic pseudo-observation trajectory filling.

4. Motion-based dynamic-window

In crowded multi-object scenarios, there are lots of interactions and occlusions among different pedestrians, which will cause trajectories to be broken or overlapped.

The normal solution is to use ReID embedding metric to reconnect deactivated trajectories. But this way cannot work well for discriminating small-scale, occluded, blurred, and similar objects. Moreover, ReID-based association is a fully connected matching strategy, which is time-consuming and unreliable due to the absence of temporal-spatial constraints. Worst of all, it still cannot recover any piece of the tracking fragments even though successfully reconnected.

To alleviate the above problems, we design the motion-based dynamic-window mechanism. In detail, we suggest to enable the full-lifecycle predictions for all trajectories using our IML module, even though the trajectory is deactivated temporally due to association failure. These continuous predictions can be employed as the temporal-spatial cues to wait for the following track-detection association.

In addition, because of various motion patterns of different objects, such as the diverse moving speeds of pedestrian and camera, it is unsuitable to set a unified reconnection window for all deactivated trajectories. Our IML can work steadily with continuous observation updating. But during deactivated period, without observation updating, pure inertia prediction may become more and more unreliable, especially for trajectories with fast-moving pedestrian or camera. Therefore, to ensure the robustness of long-range tracking, we propose an adaptive mechanism to determine the motion-based reconnection window dynamically. As shown in formula (2), the dynamic-window of each object is based on its current motion states.

$$L_{rec} = L_{max} * e^{-[\alpha * I_{Cam} + (1-\alpha) * |V_{Box}|]} \quad (2)$$

where L_{rec} denotes the dynamic length of reliable reconnection window, L_{max} represents the maximal length of possible reconnection, $|V_{Box}|$ denotes the estimated Kalman velocity as defined in the description of formula (1), and α is the coefficient describing the weights of different motion patterns. Besides, I_{Cam} is the intensity of camera motion that can be defined as formula (3).

$$I_{Cam} = 1 - \frac{W \times R}{\parallel W_2 \parallel \times \parallel R_2 \parallel}, R = [I; O] \quad (3)$$

here W denotes the vectorization of affine matrix in ECC model, and R means the affine matrix of static frames. I is the identity matrix and O is the all-zero matrix.

5. Cyclic pseudo-observation trajectory filling

With the help of motion-based dynamic-window mechanism, most deactivated trajectories can be reconnected successfully in longer windows with highly reliable temporal-spatial constraints. As shown in Fig. 2, the predictions of IML denote the inertia IML. During the deactivated period, although they can be directly retained to construct the missing tracking fragments, they may still be far away from the true trajectory, and not smooth enough with a saltation at reconnection point **B**.

Therefore, we present the cyclic pseudo-observation trajectory filling strategy to smooth the predictions of IML nonlinearly, obtaining final filled tracking fragments. In Fig. 2, some trajectory is interrupted from active to deactivated at point **A**, and reconnected successfully at point **B** through the track-detection association. Then our trajectory filling strategy can be described as the following three steps.

Step 1: Linear initialization, utilize the linear interpolation algorithm to generate the initial boxes for all frames between point **A** and **B**, with the uniform changes in position and scale.

Step 2: Forward IML updating, treat the initial boxes obtained at Step 1 as the pseudo-observations, and forward update the trained IML model at point **A** frame by frame until point **B**.

Step 3: Backward IML updating, train a new IML model backward with several frames of the tracklet after point **B**. Then use the predictions obtained at Step 2 as the pseudo-observations, and backward update the trained IML model at point **B** frame by frame until point **A**.

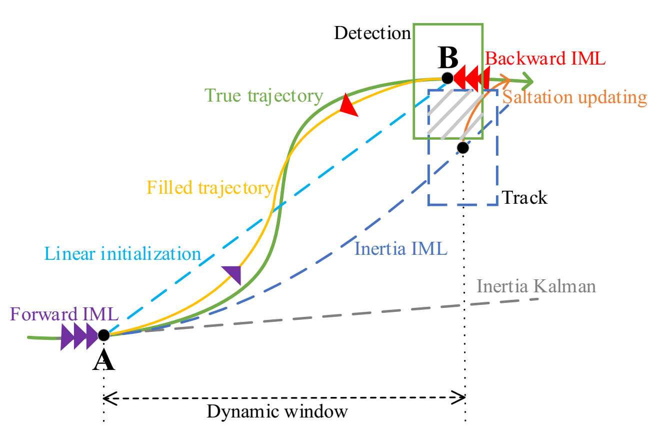

Fig. 2. Illustration of the cyclic pseudo-observation trajectory filling process.

After these three steps, the final corrected predictions are employed to fill in the tracking fragment. Consequently, the final filled tracking fragment will take the linear initialization of Step 1 as the baseline, the first half is mainly forward smoothed by the IML model at point **A**, and the latter half is mainly backward smoothed by the IML model at point **B**. This kind of cyclic nonlinear smoothing can further reduce the offset between the filled trajectory and true trajectory.

Notice that if we do not apply the nonlinear cyclic pseudo-observation trajectory filling strategy to smooth the inertia IML predictions after successful reconnection, but provide the continuous online predictions through our IML module, the inference process of our MAT is exactly an online MOT paradigm. Besides, the complete MAT with nonlinear cyclic smoothing only uses the reconnection point to correct the online IML predictions in the tracking fragments, and does not use a lot of information of future frames like the offline methods. Therefore, our proposed MAT can be regarded as a **near-online** MOT method.

5.1. 3D integral Image (3DII)

During the data association stage, we need to calculate the cost matrix among predicted tracks and corrected detections. It is usually performed in a fully connected manner. Unfortunately, in a dense multi-object scenario, the number of objects can be dozens or even hundreds, this calculation manner means a significant time cost. So, we present the 3DII module to efficiently cut off useless track-detection connections, by transferring all detections into the multilayered encoding maps.

As shown in Fig. 3, the input frame with K detections is divided into $M \times N$ cells, then the K-layered encoding maps will be constructed with size of $M \times N$ for each layer. Each detection corresponds to one layer, the map for this layer is initialized with all zeros, and the cells within partial-covered region of this detection bounding box will be encoded as ones. Therefore, the 3DII of the K-layered encoding maps at location (m, n) can be modeled as the sum of one-valued cells from $(0, 0)$ to (m, n) for each layer as:

$$I_{3D}(m,n) = \sum\nolimits_{m' \le m, n' \le n} f(m', n') \quad (4)$$

where I_{3D} represents the K-dim integral image, and f denotes the K-layered encoding maps with binary values.

To accelerate the calculation of formula (4), we can simplify its calculating process using dynamic programming.

$$\begin{aligned} I_{3D}(m,n) = & I_{3D}(m,n-1) + I_{3D}(m-1,n) \\ & - I_{3D}(m-1,n-1) + f(m,n) \end{aligned} \quad (5)$$

For each track, first, the diagonal coordinates of its bounding box can be mapped to the coordinate system in f as $[x_1, x_2, y_1, y_2]$. Then, utilizing the above 3D integral image in formula (5), the detections within partial-covered region of this track will be obtained directly with nearly constant time cost.

$$\begin{aligned} I_{3D}(x_1 : x_2, y_1 : y_2) = & I_{3D}(x_2, y_2) + I_{3D}(x_1 - 1, y_1 - 1) \\ & - I_{3D}(x_1 - 1, y_2) - I_{3D}(x_2, y_1 - 1) \end{aligned} \quad (6)$$

where the non-zero value in computed $I_{3D}(x_1 : x_2, y_1 : y_2)$ denotes the found nearby detection.

Notice that most of the operations in 3DII are assignment, addition or subtraction, so the time cost for filtering useless track-detection connections can be dramatically reduced. Besides, the 3DII is easy to enable the temporal-spatial constraints at any scale-level by just changing the cell size.

5.2. Algorithm representation of MAT

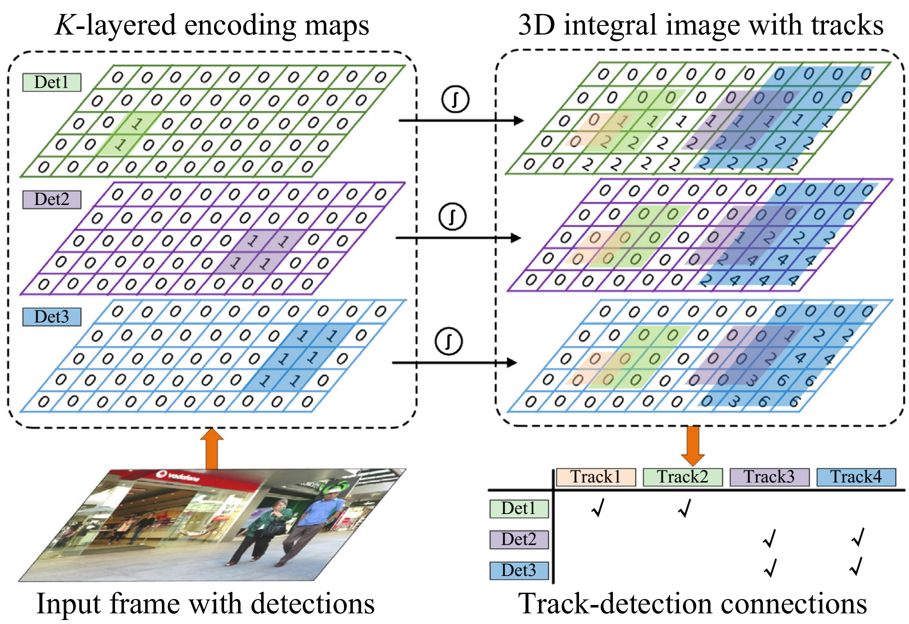

Fig. 3. The encoding and filtering process of 3DII module.

Algorithm 1. Tracking process of our MAT approach using public or private detector

Input: Image sequence $I = \{I_1, I_2, \cdots, I_S\}$, *public* or *private* detections $\mathscr{D} = \{D_1, D_2, \cdots, D_S\}$.

Output: Trajectories $\mathscr{T}$.
1.Initialize trajectories with null that $\mathscr{T} \leftarrow \phi$;
2. $\mathscr{L}_k$: Deactivated length of $\mathscr{T}_k$, initialized by 0;
3. **for** $t = 1, \cdots, S$ **do**
4. $T_t \leftarrow$ Obtain tracks by applying IML to $\mathscr{T}$ referring to formula (1);
5. **if** *private* **then**
6. $B_t \leftarrow D_t$;
7. **else**
8. $B_t \leftarrow NMS(Regress\_and\_Classify(\{D_t, T_t\}))$;
9. $Cost \leftarrow$ Calculate cost matrix of $3DII(B_t, T_t)$ by IoU metric;
10.Associate the B_t with T_t using KM algorithm according to $Cost$;
11.**for** $\mathscr{T}_k \in \mathscr{T}$ **do**
12. Update the L_{rec} of $\mathscr{T}_k$ referring to formula (2);
13. **if** $T_{t,k}$ is associated with $B_{t,i}$ **then**
14. $\mathscr{T}_k \leftarrow \mathscr{T}_k + B_{t,i}$;
15. Fill in the tracking fragment of $\mathscr{T}_k$ using DRC;
16. $\mathscr{L}_k = 0$;
17. **else**
18. **if** $\mathscr{L}_k > L_{rec}$ **then**
19. Stop further tracking of $\mathscr{T}_k$;
20. **else**
21. Hold $T_{t,k}$ with $\mathscr{T}_k$ temporarily for further IML;
22. $\mathscr{L}_k{+}{+}$;
23.**for** $B_{t,j} \in B_t$ **do**
24. **if** $B_{t,j}$ is not associated with anyone of T_t **then**
25. $\mathscr{T} \leftarrow \mathscr{T} + B_{t,j}$;
26. Remove ultra-short trajectories from $\mathscr{T}$ for post-processing.

As shown in Algorithm 1, we merge the pipeline of our MAT approach with public or private detector, and provide the corresponding pseudocode representation of Fig. 1 and Sections 3.1-3.3. As illustrated at line 9, we use the simplest IoU metric to evaluate the temporal-spatial similarity of the candidate track-detection connections. Moreover, based on 3DII module, we can efficiently cut off the unreliable connections with low IoU, before performing the Kuhn-Munkres (KM) algorithm [19] at line 10. Here the IoU threshold of valid candidate connection can be set as 0.3.

6. Experiments

6.1. Experimental settings

Experiments are conducted on two widely used challenging MOT benchmarks: MOT16 and MOT17 [1]. Although they contain the same 7 training sequences and 7 test sequences, they provide different public detections and ground truth labels. Such as, MOT17 uses three public detectors, namely the DPM [14], Faster R-CNN, and SDP [31], here their performance increases in turn. But MOT16 only includes the DPM detector. These benchmarks consist of many challenges, for instance frequent occlusion, and the scenes are heavily crowed and vary in the camera poses, object scales and frame rates.

Evaluation Metrics. Evaluation is carried out according to the widely accepted CLEAR MOT metrics [3], including the Multiple Object Tracking Accuracy (MOTA), ID F1 Score (IDF1), False Positives (FP), False Negatives (FN), Identity Switches (IDS), and Tracker Speed (Hz), et al. Among these performance metrics, the MOTA and IDS can quantify the main two aspects as trajectory accuracy and purity.

Implementation Details. All the experiments are implemented using PyTorch and run on a desktop with a CPU of 10 cores@2.2 GHz and a RTX2080Ti GPU. As for the detector, based on pre-trained network parameters on COCO datasets [7], we re-implement Cascaded RCNN [6] by changing the anchor scales to {32, 64, 128, 256, 512} and anchor aspects to {1.0, 2.0, 3.0}. The detector is simply trained on the training set of MOT17, by taking the samples with label 1 and visibility above 0.1 as positives, and the others as negatives. Random flipping is adopted to realize the data augmentation, and all the clip operations for ground truth boxes, anchors, and proposals are canceled to obtain complete detections even beyond the boundaries. The detector is trained for 30,000 iterations with batch size of 8, using the multi-step SGD optimizer with an initial learning rate of e^{-8}. The vanilla NMS [11] is employed with the confidence threshold of 0.05 and IoU threshold of 0.5. For MAT, the simple IoU metric is used to calculate the cost matrix of Kuhn-Munkres (KM) algorithm [19] for motion-based association, the Kalman Filter is simplified as a constant velocity model [22], L_{max} and α in formula (2) are fixed as 120 and 0.95, the 3DII map size is set as 16×8, and the trajectories whose length less than 5 frames are removed during post-processing.

6.2. Ablation study

As is well-known, the MOTChallenge official organization limits the number of times that researchers can submit results to the test server. Our ablation study mainly aims to verify the performance of our motion-aware designs, which do not depend on any training on ground truths. Therefore, to avoid the possible influence caused by the quality of detector, we evaluate the following implementations on MOT17 training set for convenience. So did other researchers.

As shown in Table 1, the Baseline represents our re-implemented Detector with Kalman, without any guidance of camera motion and dynamic reconnection, which is the simplest version of our MAT. Notice that the Baseline still includes a simple reconnection mechanism like Tracktor++, the deactivated trajectories can be reconnected within a fixed window of 10 frames and filled using the inertia predictions. The Baseline + means to replace some modules in Baseline, such as the prediction or reconnection mechanism. For instance, Baseline + Ego, Baseline + ECC, and Baseline + IML replace the Kalman prediction of Baseline with the Ego, ECC, and our IML, respectively. The Baseline + IML + DW just replaces the fixed reconnection window of Baseline + IML with our dynamic-window (DW) referring to formula (2). The Baseline + IML + DRC is the non-accelerated implementation of our MAT without 3DII module.

The results in Table 1 can verify that:

1) The Ego model cannot handle the motion alignments well just like the analysis given in Section 3.1. The ECC model can achieve nearly the same performance as Kalman Filter with much lower IDS. In contrast, our IML model can bring Baseline tracker an overall promotion on all metrics especially the IDS, which demonstrates the powerful motion pattern compatibility of the IML module.

2) Based on our IML module, it is also shown that the further employment of our DW or complete DRC, can improve the tracking performance of Baseline tracker with great margins on MOTA, IDF1, FN, and IDS, but small sacrifice of FP. In other words, for motion-based reconnection, it is essential to design a dynamic-window according to various motion patterns. In addition, for trajectory filling, our proposed cyclic pseudo-observation trajectory filling strategy (Baseline + IML + DRC) is proved to be more robust than just using inertia predictions (Baseline + IML + DW), with lower FP, FN and higher MOTA, IDF1. Note that the IDS of DW is slightly smaller than DRC since much more long-range trajectories are tracked by DRC.

Dynamic window. The above ablation experiments (Baseline + IML + DW vs. Baseline + IML) have confirmed that our design of dynamic reconnection window can obtain greater improvement than fixed one, even using inertia predictions to fill in the tracking fragments. Fig. 4 gives more details about dynamic vs. fixed window. Here they both use the powerful cyclic pseudo-observation trajectory filling, and share the same maximal length of possible reconnection L_{max}. The curves in Fig. 4 reveal that with the increase of L_{max}, the DRC module based on dynamic-window gradually outperforms the fixed window with higher MOTA, and the peak point is up to 120 frames to enable the long-range tracking tasks.

3DII module. During the data association stage, it is necessary to cut off useless track-detection connections using temporal-spatial constraints. If a detection has some overlap with the region of a surrounding track's bounding box, then the detection will be assigned as a candidate of the track. To verify the acceleration performance of our 3DII module, as a contrast, we also present an IoU-based fully-connected method to filter the track-detection connections without IoU, named as IoU-filter. The speed comparison is illustrated in Fig. 5, where the number of tracks and detections is simply set to the same for each track-detection association. The larger track-detection number denotes the denser multi-object scenario. The results in Fig. 5 demonstrate the significant advantage of our 3DII module in acceleration with nearly the constant time cost, against the IoU-filter, especially when there exist a large

Table 1
Ablation study in terms of different prediction models and motion-based reconnection mechanisms.

Method	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
Baseline	68.0	69.2	3749	102,709	1407
Baseline + Ego	45.5	41.4	14,983	158,850	9937
Baseline + ECC	68.2	69.1	3949	102,094	1083
Baseline + IML	68.7	71.1	**3430**	101,017	850
Baseline + IML + DW	71.1	74.6	10,422	86,360	**682**
Baseline + IML + DRC	**71.5**	**74.8**	9711	**85,653**	701

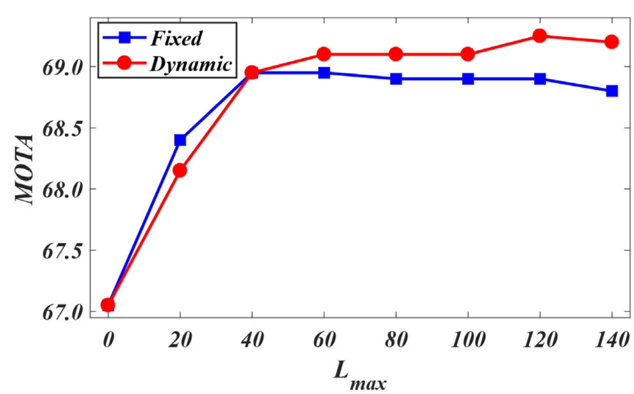

Fig. 4. Dynamic vs. fixed window motion-based reconnection with different maximal lengths.

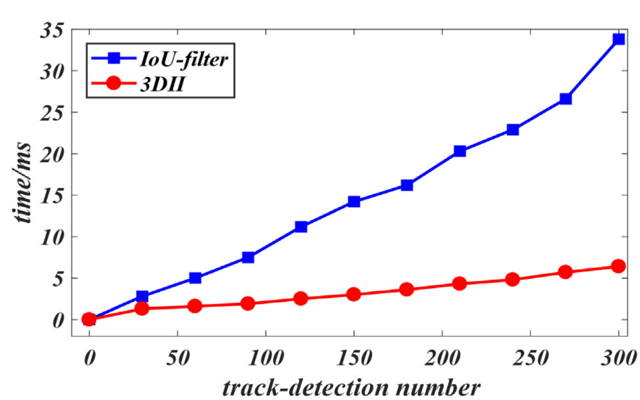

Fig. 5. Speed comparison between our 3DII module and IoU-filter.

number of pedestrians per frame. Note that our 3DII module can easily enable the temporal-spatial constraints at any scale-level by just changing the cell size, while IoU-filter can only judge whether there is an intersection.

Portability analysis. Our proposed MAT is an enhanced MOT paradigm, and it mainly devotes to enhance and cooperate with the tracking part of a tracker. It can be completely independent of the detector. In other words, our MAT is a plug-and-play solution, which can be easily embedded into any tracking-by-detection MOT tracker. As shown in Table 2, we embed our MAT into the FairMOT, which is by far the latest open source state-of-the-art tracker. Note that the following experimental verifications are also performed on the MOT17 training set. In detail, we maintain the detector part of FairMOT, and replace its tracking strategy using our MAT.

In Table 2, FairMOT(w/o ReID) + MAT means that FairMOT tracker completely abandons its ReID-based association and uses our MAT for motion-based prediction, reconnection, and association. Compared with this high-performance baseline (MOTA equals to 83.8), even without using its ReID embedding for appearance matching, our MAT can still bring a steady improvement on MOTA (**+2.7**), while the IDF1 only decreases by **0.1**.

Moreover, on the basis of FairMOT(w/o ReID) + MAT, if we further utilize the ReID embedding of FairMOT to filter out unreasonable (the threshold is set to 0.3) motion-based associations, then we have the FairMOT(w/ReID) + MAT. With the help of ReID module, FairMOT(w/ReID) + MAT outperforms FairMOT(w/o ReID) + MAT by **0.7** of IDF1, while the MOTA is only improved by **0.2**. Notice that including ReID branch requires additional computational resources and time.

Combining the results in Table 1 and Table 2, we can further draw the following conclusions. Such as, our MAT is indeed a plug-and-play solution, which can steadily bring about **3** points of MOTA boost and higher IDF1 to the baseline tracker. Besides, when the detector is powerful and frame rate is high enough, our MAT is sufficient to achieve a good performance (high MOTA) without ReID and other complex modules.

Memory & Speed. Together with our re-implemented Cascaded RCNN detector and the proposed MAT, our complete tracker usually consumes about 2.65 GB of physical memory, and 2 GB of graphics memory.

In addition, according to Fig. 1, our complete tracker consists of five components, namely the IML, 3DII, DRC, Detector, and Association. Refer to the speed calculation manner of FairMOT, the time taken to read and preprocess the images is not counted. Then the detailed speed of our complete tracker in terms of its different components is shown in Table 3. The overall speed of our complete tracker can reach 18.5 FPS using just one RTX2080Ti GPU. Among all these counted components, the Detector (our re-implemented Cascaded RCNN) and Association (KM algorithm) components are both very time-consuming. But strictly speaking, they are completely independent of our pure MAT, which just contains the IML, 3DII, and DRC. As shown in Table 3, the speed of our pure MAT can reach as high as 50.6 FPS. Thus, our MAT is proved to be a lightweight solution.

Table 2
Portability experiments of MAT based on the modification of FairMOT.

Tracker	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
FairMOT	83.8	81.9	**8136**	44,631	1659
FairMOT(w/o ReID) + MAT	86.5	81.8	16,047	**28,467**	1128
FairMOT(w/ReID) + MAT	**86.7**	**82.5**	14,211	29,559	**1065**

Table 3
Detailed speed of each component of our complete tracker.

Component	Hz/FPS
IML	92.5
3DII	685.9
DRC	133.8
Detector	37.8
Association	126.9
Pure MAT (IML + 3DII + DRC)	50.6
Ours (complete tracker)	18.5

6.3. Parameter sensitivity analysis

According to the experiments in Fig. 4, we demonstrate that, for better long-range tracking performance, the maximal possible reconnection length L_{max} in formula (2) can be up to 120 frames. Here we fix L_{max} as 120 and then test the performance sensitivity of another parameter α in formula (2), as shown in Fig. 6. Notice that the other experimental settings of this subsection are the same as the ablation study of Section 4.2. The results in Fig. 6 reveal that, to determine the dynamic length of reliable reconnection window for better MOTA performance, the intensity of camera motion should play the main role when compared with the estimated Kalman velocity. Nevertheless, directly removing $|V_{Box}|$ from the formula (2) and setting α as 1 is not the best option. Besides, our final MOTA performance is not very sensitive with the variation of parameter α. Therefore, as described in the implementation details of Section 4.1, α in formula (2) can be fixed as 0.95 for all the experiments.

In addition, as mentioned in Fig. 2 and Section 3.2, for cyclic pseudo-observation trajectory filling process, there is a parameter in Step 3 for backward IML updating, namely the backward tracklet length. The backward tracklet consists of several tracks after point **B**, which are used for training a new backward IML model. Here we test the performance sensitivity of this parameter, and the corresponding MOTA results of different backward tracklet lengths are shown in Fig. 7. It is verified that our final MOTA performance is quite insensitive to the backward tracklet length. Thus, according to the results of Fig. 7, we fix this parameter as 3 for all the experiments.

6.4. Evaluation on benchmarks

As mentioned before, our complete MAT mainly devotes to enhance the tracking part of a tracker, so it can easily cooperate with the public or private detector. Besides, it can be regarded as a near-online (denoted using **NO**) tracker.

Here, we evaluate the comprehensive performance of our MAT approach on the test sets of MOT16 and MOT17, using both the

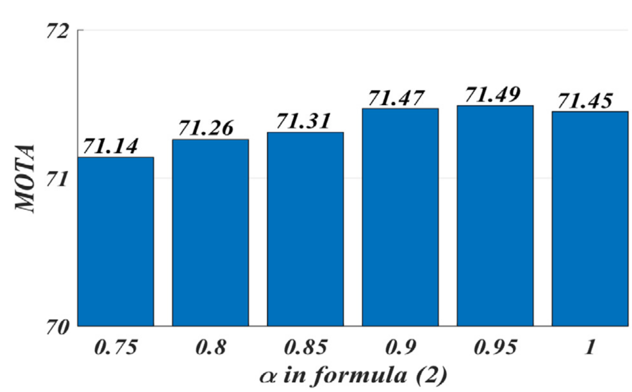

Fig. 6. Sensitivity analysis experiment of α in formula (2).

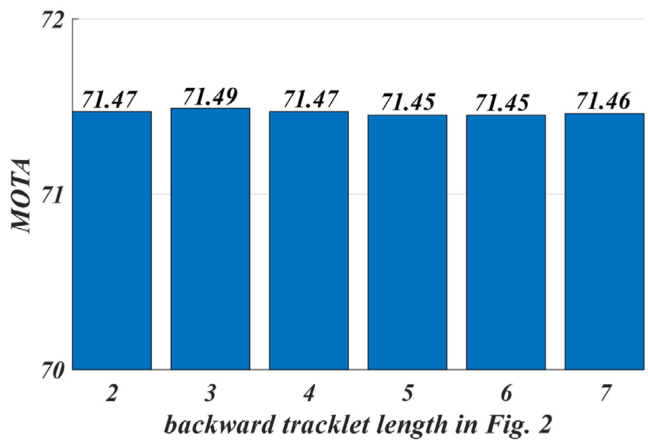

Fig. 7. Sensitivity analysis experiment of backward tracklet length in Fig. 2.

public and private detectors. We also compare MAT with other state-of-the-art (SOTA) public and private trackers, which have been officially published or peer reviewed with results on these benchmarks. The comparisons are shown in Table 4 and Table 5, where the SOTA offline, online, and near-online MOT methods are all included. Such as the offline trackers (denoted using **Off**) like MPNTracker [5], Lif_T [16], NOTA [45], JBNOT [46], and LPCMOT [47]. The online trackers (denoted using **O**) like Trackor++, DeepMOT [29], UMA [32], UnsupTrack [18], CenterTrack, JDE [26], FairMOT, CTracker, and Tube_TK [21]. And the near-online trackers (denoted using **NO**) like MHT-DAM [48], bLSTM-MTP [49], EDMT [50], and MLT [51].

As demonstrated in Table 4 and Table 5, our MAT significantly outperforms all these SOTA MOT methods by a large margin, especially in terms of MOTA and FN. Besides, for our MAT, the metrics of IDF1 and IDS are almost the best, no matter using the public or private detector. Additionally, our MAT has much lower time cost, as shown in the Hz column, compared to most of the trackers. It should be noted that, the main emphasis of this paper is the

Table 4
Comparisons of SOTA MOT methods on MOT16 test set.

Method	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓	Hz↑
Public Detector on MOT16						
MPNTracker(**Off**)	58.6	61.7	4949	70,252	354	6.5
Lif_T(**Off**)	61.3	64.7	4844	65,401	389	0.5
NOTA(**Off**)	49.8	55.3	7428	83,614	614	–
Trackor++(**O**)	54.4	52.5	3280	79,149	682	2.0
DeepMOT(**O**)	54.8	53.4	**2955**	78,765	645	–
UMA(**O**)	50.5	52.8	7587	81,924	685	5.0
UnsupTrack(**O**)	62.4	58.5	5909	61,981	588	1.9
MHT-DAM(**NO**)	45.8	46.1	6412	91,758	590	0.8
bLSTM-MTP(**NO**)	49.9	52.5	7111	83,676	579	**23.8**
EDMT(**NO**)	45.3	47.9	11,122	87,890	639	1.8
MLT(**NO**)	52.8	62.6	5362	80,444	**299**	–
Ours(NO)	**67.7**	**69.6**	6337	**52,234**	379	18.5
Private Detector on MOT16						
JDE(**O**)	64.4	55.8	–	–	1544	18.8
FairMOT(**O**)	68.7	**70.4**	–	–	953	25.9
CTracker(**O**)	67.6	57.2	**8934**	48,305	1897	**34.4**
Tube_TK(**O**)	66.9	62.2	11,544	47,502	1236	1.0
Ours(**NO**)	**70.5**	63.8	11,318	**41,592**	**928**	18.5

Table 5
Comparisons of SOTA MOT methods on MOT17 test set.

Method	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓	Hz↑
Public Detector on MOT17						
MPNTracker(**Off**)	58.8	61.7	17,413	213,594	1185	6.5
Lif_T(**Off**)	60.5	65.6	14,966	206,619	1189	0.5
JBNOT(**Off**)	52.6	50.8	31,572	232,659	3050	5.4
LPCMOT(**Off**)	59.0	66.8	23,102	206,948	1122	4.8
Trackor++(**O**)	53.5	52.3	12,201	248,047	2072	2.0
CenterTrack(**O**)	61.4	53.3	15,520	196,886	5326	–
DeepMOT(**O**)	53.7	53.8	**11,731**	247,447	1947	–
UMA(**O**)	53.1	54.4	22,893	239,534	2251	5.0
UnsupTrack(**O**)	61.7	58.1	16,872	197,632	1864	
MHT-DAM(**NO**)	50.7	47.2	22,875	252,889	2314	0.9
bLSTM-MTP(**NO**)	53.6	55.8	23,669	236,226	1845	**22.7**
EDMT(**NO**)	50.0	51.3	32,279	247,297	2264	0.6
MLT(**NO**)	54.8	62.9	19,118	234,303	**1077**	–
Ours(**NO**)	**67.1**	**69.2**	22,756	**161,547**	1279	18.5
Private Detector on MOT17						
FairMOT(**O**)	67.5	**69.8**	–	–	2868	25.9
CTracker(**O**)	66.6	57.4	**22,284**	160,491	5529	**34.4**
CenterTrack(**O**)	67.3	59.9	23,031	158,676	2898	–
Tube_TK(**O**)	63.0	58.6	27,060	177,483	4137	3.0
Ours(**NO**)	**69.5**	63.1	30,660	**138,741**	**2844**	18.5

motion-aware designs but not the optimization of detector. Hence our re-implemented detector may not be as powerful as other private detectors in terms of accuracy and speed. Even so, our MAT still wins the championship on the primary metric MOTA, which reflects the overall tracking performance.

To summarize, the reason why our proposed MAT can achieve superior tracking performance in MOTA and IDF1 is mainly due to the powerful motion compatibility of IML module, which ensures the continuity and purity of trajectories. Additionally, by further employing the DRC module, the long-range motion-based tracking and reconnection become true with dramatical decreases of FN and IDS. Furthermore, given a fast detector, our MAT can achieve the high tracking speed. Specifically, by imposing temporal-spatial constraints and filtering useless track-detection connections, our 3DII module can address another speed bottleneck for data association with nearly constant time cost.

6.5. Visualization results

First, to further highlight the motion-aware ability of our proposed IML and DRC modules, we present the visualization results of different trajectory filling strategies using SDP detector in Fig. 8. Just like the colors used in Fig. 2, the green boxes indicate the ground truth boxes, the blue boxes denote the inertia IML predictions, and the yellow boxes represent the filled tracking fragments by our DRC module. Frame 232 corresponds to the deactivated point **A** in Fig. 2, and frame 254 means the reconnection point **B**. The visualization comparisons in Fig. 8 support the discussions in Fig. 2 vividly and consistently, that the filled trajectory using our DRC module is smoother and closer to true trajectory with smaller offset. In contrary, if the pedestrian changes the walking direction slightly during being occluded, just like the target lady in Fig. 8, the inertia IML predictions become further and further away from ground truth boxes like the tracking from frame 232 to frame 252. Moreover, it may lead to a saltation when reconnected, such as the non-smooth transition of blue boxes from frame 252 to frame 254.

Moreover, as shown in Fig. 9, we select a typical occlusion scenario to compare our MAT with two related trackers qualitatively, namely the Tracktor++ and DeepSort. The tracking results of our MAT and Tracktor++ are based on the public SDP detector, and DeepSort uses its private detector. The red target pedestrian in the white dotted rectangle is occluded when passing by the streetlight. It can be seen from Fig. 9 that Tracktor++ and DeepSort are both unable to accurately track this man. They treat this target pedestrian as two different identities before and after being occluded, which will inevitably cause False Negatives and Identity Switches. In contrast, our MAT can address this problem effectively through the proposed IML and DRC modules, and maintain the trajectory accuracy and purity of the occluded object.

Other detailed qualitative visualization results are shown in Fig. 10. All the test sequences are tracked by our MAT using SDP

Fig. 8. Visualization comparisons of trajectory filling by inertia IML (blue) or DRC module (yellow) using SDP detector. The green boxes indicate the ground truth boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

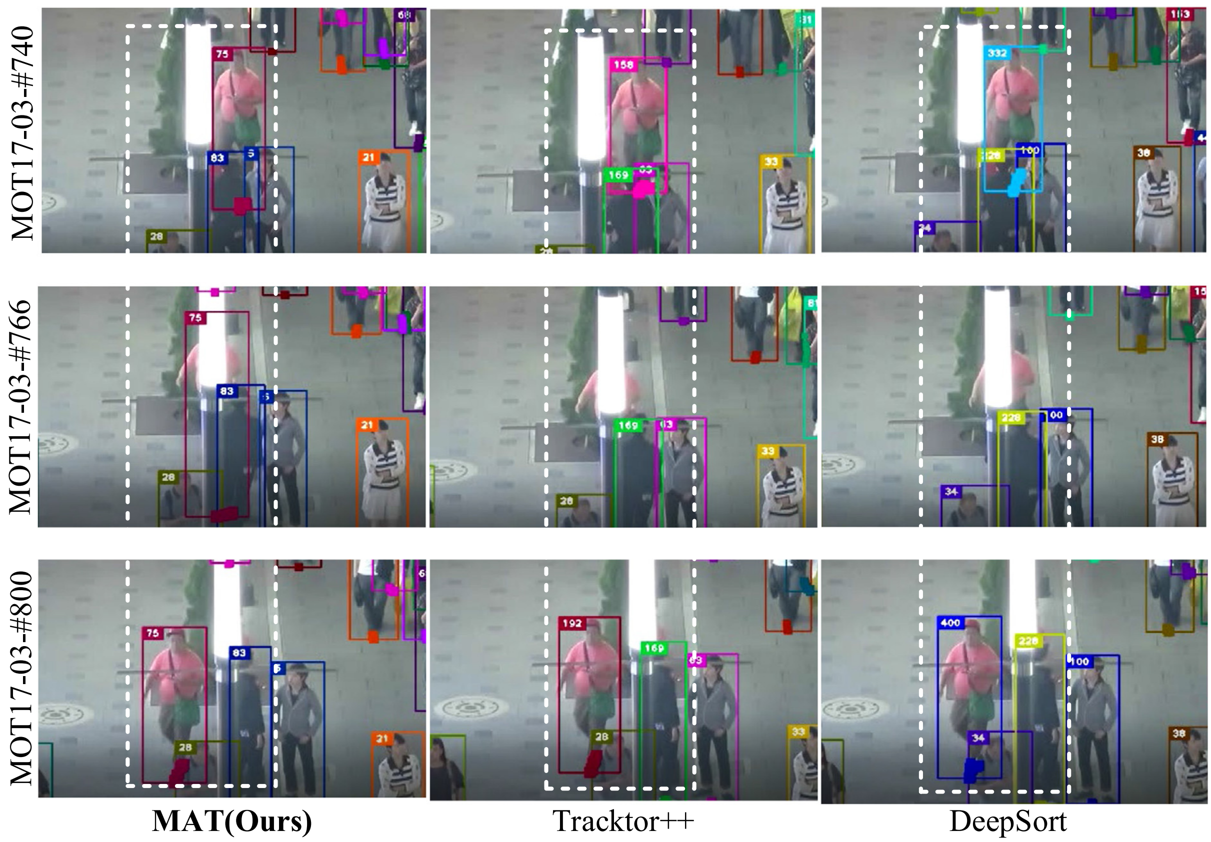

Fig. 9. Qualitative comparisons of our MAT with the Tracktor++ and DeepSort in a typical occlusion scenario. The white dotted rectangle indicates the tracking results of the red target pedestrian. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 10. Visualizations of tracking results of our MAT approach using SDP detector on MOT17 test set.

detections as observations. Intuitively, our proposed MAT can output accurate tracking boxes of pedestrians with consistent identities. Besides, our MAT is robust enough to the irregular camera motions (such as MOT17-06 and MOT17-14), crowded scenarios (such as MOT17-03), and various camera viewpoints (such as MOT17-03 and MOT17-07). Especially for MOT17-14 sequence, which is captured in a busy street by a fast-moving camera that is mounted on a bus. Our proposed MAT can still be able to track all the pedestrians steadily and continuously.

7. Conclusion

In this paper, we propose an enhanced MOT paradigm named as MAT, which mainly focuses on powerful motion-aware designs, such as motion-based prediction, reconnection, and association. Our MAT contains three key modules, namely the IML, DRC, and 3DII. In detail, the IML module is developed for blending various motion patterns compatibly. The DRC module is devised to determine reconnection window dynamically and fill trajectory smoothly. The 3DII module is presented for the fast temporal-spatial filtering of useless track-detection association connections. Extensive experiments are conducted on the widely used MOT16&17 challenging benchmarks, and all the results demonstrate the superiority of our MAT approach in terms of the SOTA accuracy and high efficiency. Besides, our MAT is a plug-and-play solution, its architecture is quite simple and has strong generality and extensibility, which can be easily embedded into any tracking-by-detection tracker or video detector.

CRediT authorship contribution statement

Shoudong Han: Conceptualization, Methodology, Writing – review & editing, Supervision. **Piao Huang:** Writing – original draft, Methodology, Software. **Hongwei Wang:** Visualization, Software. **En Yu:** Formal analysis, Software. **Donghaisheng Liu:** Investigation, Software. **Xiaofeng Pan:** Validation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61105006; Open Fund of Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education under Grant No. IPIC2019-01; and the China Scholarship Council under Grant No. 201906165066.

References

[1] Anton, M.; Laura, L.-T.; Ian, R.; Stefan, R.; and Konrad, S. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv: 1603.00831[cs.CV]. Ithaca, NY: Cornell University Library.*

[2] P. Bergmann, T. Meinhardt, L. Leal-Taixé, Tracking without bells and whistles, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 941–951.

[3] Bernardin, K., and Stiefelhagen, R. 2008. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Vide.* doi:10.1155/2008/246309.

[4] A. Bewley, Z.Y. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: *Proceedings of the IEEE International Conference on Image Processing*, 2016, pp. 3464–3468.

[5] G. Brasó, L. Leal-Taixé, Learning a neural solver for multiple object tracking, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6247–6257.

[6] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[7] X. Chen, H. Fang, T.Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C.L. Zitnick, Microsoft COCO: Common objects in context, in: *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[8] W. Choi, S. Savarese, Multiple target tracking in world coordinate with single, minimally calibrated camera, in: *Proceedings of the European Conference on Computer Vision*, 2010, pp. 553–567.

[9] P. Chu, H.B. Ling, Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6172–6181.

[10] Q. Chu, W. Ouyang, B. Liu, F. Zhu, N. Yu, DASOT: A unified framework integrating data association and single object tracking for online multi-object tracking, in: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 10672–10679.

[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[12] G.D. Evangelidis, E.Z. Psarakis, Parametric image alignment using enhanced correlation coefficient maximization, IEEE Trans. Pattern Anal. Mach. Intell. 30 (10) (2008) 1858–1865, https://doi.org/10.1109/TPAMI.2008.113.

[13] K. Fang, Y. Xiang, X.C. Li, S. Savarese, Recurrent autoregressive networks for online multi-object tracking, in: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 466–475.

[14] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645, https://doi.org/10.1109/TPAMI.2009.167.

[15] W.T. Feng, Z.H. Hu, W. Wu, J.J. Yan, W.L. Ouyang, Multi-object tracking with multiple cues and switcher-aware classification. *arXiv preprint arXiv:1901.06129 [cs.CV]. Ithaca, NY: Cornell University Library.*

[16] A. Hornakova, R. Henschel, B. Rosenhahn, P. Swoboda, Lifted disjoint paths with application in multiple object tracking. *arXiv preprint arXiv:2006.14550 [cs.CV]*, Cornell University Library, Ithaca, NY, 2020.

[17] L. Huang, X. Zhao, K. Huang, GlobalTrack: A simple and strong baseline for long-term tracking, in: *Proceedings of the The Thirty-Fourth Conference on Artificial Intelligence*, 2020, pp. 11037–11044.

[18] S. Karthik, A. Prabhu, V. Gandhi, Simple unsupervised multi-object tracking arXiv preprint arXiv:2006.02609 [cs.CV], Cornell University Library, Ithaca, NY, 2020.

[19] H.W. Kuhn, The Hungarian Method for the Assignment Problem, Springer, 2010.

[20] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, S. Savarese, Learning an image-based motion context for multiple people tracking, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3542–3549.

[21] B. Pang, Y.Z. Li, Y.F. Zhang, M.C. Li, C.W. Lu, TubeTK: Adopting tubes to track multi-object in a one-step training model, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6308–6318.

[22] J. Peng, T. Wang, W. Lin, J. Wang, J. See, S. Wen, E. Ding, TPM: Multiple object tracking with tracklet-plane matching, Pattern Recogn. 107 (2020) 107480, https://doi.org/10.1016/j.patcog.2020.107480.

[23] J.L. Peng, C.A. Wang, F.B. Wan, Y. Wu, Y.B. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Fu, Chained-Tracker: Chaining Paired Attentive Regression Results for end-to-end Joint Multiple-Object Detection and Tracking. *arXiv preprint arXiv:2007.14557v1[cs.CV]*, Cornell University Library, Ithaca, NY, 2020.

[24] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal, in: *Proceedings of the Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[25] G.A. Wang, Y.Z. Wang, H.T. Zhang, R.S. Gu, J.-N. Hwang, Exploit the connectivity: Multi-object tracking with trackletnet, in: *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 482–490.

[26] Z.D. Wang, L. Zheng, Y.X. Liu, S.J. Wang, Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605 [cs.CV]. Ithaca, NY: Cornell University Library*.

[27] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: *Proceedings of the IEEE International Conference on Image Processing*, 2017, pp. 3645–3649.

[28] Y. Xu, Z. Wang, Y. Yuan, G. Yu, SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines, in: *Proceedings of the The Thirty-Fourth Conference on Artificial Intelligence*, 2020, pp. 12549–12556.

[29] Y.H. Xu, A. Osep, Y.T. Ban, R. Horaud, L. Leal-Taixé, X. Alameda-Pineda, How to train your deep multi-object tracker, in: *Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6787–6796.

[30] B. Yang, C. Huang, R. Nevatia, Learning affinities and dependencies for multi-target tracking using a CRF model, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1233–1240.

[31] F. Yang, W. Choi, Y. Lin, Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2129–2137.

[32] J.B. Yin, W.G. Wang, Q.H. Meng, R.G. Yang, J.B. Shen, A unified object motion and affinity model for online multi-object tracking, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6768–6777.

[33] Zhang, Y. F.; Wang, C. Y.; Wang, X. G.; Zeng, W. J.; and Liu, W. Y. 2020. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888 [cs.CV]. Ithaca, NY: Cornell University Library.*
[34] X.Y. Zhou, V. Koltun, P. Krähenbühl, Tracking objects as points arXiv preprint arXiv:2004.01177 [cs.CV], Cornell University Library, Ithaca, NY, 2020.
[35] X.Y. Zhou, D.Q. Wang, P. Krähenbühl, Objects as points. arXiv preprint. arXiv:1904.07850 [cs.CV], Cornell University Library, Ithaca, NY, 2019.
[36] J. Zhu, H. Yang, N. Liu, M. Kim, W.J. Zhang, M.-H. Yang, Online multi-object tracking with dual matching attention networks, in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 366–382.
[37] J. Hong Yoon, C.R. Lee, M.H. Yang, K.J. Yoon, Online multi-object tracking via structural constraint event aggregation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1392–1400.
[38] S. Sharma, J.A. Ansari, J.K. Murthy, K.M. Krishna, Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking, in: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 3508–3515.
[39] J.H. Yoon, M.H. Yang, J. Lim, K.J. Yoon, Bayesian multi-object tracking using motion context from multiple objects, in: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 33–40.
[40] F. Saleh, S. Aliakbarian, H. Rezatofighi, M. Salzmann, S. Gould, Probabilistic tracklet scoring and inpainting for multiple object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14329–14339.
[41] B. Shuai, A. Berneshawi, X.Y. Li, D. Modolo, J. Tighe, SiamMOT: Siamese multi-object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12372–12382.
[42] Z.K. Zhou, N.N. Fan, K. Yang, H.P. Wang, Z.Y. He, Adaptive ensemble perception tracking, Neural Netw. 142 (2021) 316–328, https://doi.org/10.1016/j.neunet.2021.05.003.
[43] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, Z. He, Self-supervised deep correlation tracking, IEEE Trans. Image Process. 30 (2021) 976–985, https://doi.org/10.1109/TIP.8310.1109/TIP.2020.3037518.
[44] Z. He, X. Li, X. You, D. Tao, Y.Y. Tang, Connected component model for multi-object tracking, IEEE Trans. Image Process. 25 (8) (2016) 3698–3711, https://doi.org/10.1109/TIP.2016.2570553.
[45] L. Chen, H. Ai, R. Chen, Z. Zhuang, Aggregate tracklet appearance features for multi-object tracking, IEEE Signal Process Lett. 26 (11) (2019) 1613–1617, https://doi.org/10.1109/LSP.9710.1109/LSP.2019.2940922.
[46] R. Henschel, Y.Z. Zou, B. Rosenhahn, Multiple people tracking using body and joint detections, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 770–779.
[47] P. Dai, R.L. Weng, W. Choi, C.S. Zhang, Z.P. He, W. Ding, Learning a proposal classifier for multiple object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2443–2452.
[48] C. Kim, F.X. Li, A. Ciptadi, J.M. Rehg, Multiple hypothesis tracking revisited, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4696–4704.
[49] C. Kim, F.X. Li, M. Alotaibi, J.M. Rehg, Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9553–9562.
[50] J.H. Chen, H. Sheng, Y. Zhang, Z. Xiong, Enhancing detection model for multiple hypothesis tracking, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 18–27.
[51] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, Z. Xiong, Multiplex labeling graph for near-online tracking in crowded scenes, IEEE Internet Things J. 7 (9) (2020) 7892–7902, https://doi.org/10.1109/JIOT.2020.2996609.

Shoudong Han was born in Wuhan, China, on March 28, 1983. He received the B.S. degree, M.S. degree and Ph.D. degree in Control Science and Engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2005, 2007 and 2010, respectively. He is currently an Associate Professor in the National Key Laboratory of Science and Technology on Multispectral Information Processing, and School of Artificial Intelligence and Automation at HUST. His research interests include computer vision, deep / machine learning, graph representation learning, and pattern recognition. He has published a number of innovative works in top-tier journals such as TIP, PR, SP, PRL, and JVCIR.

Piao Huang was born in Xiaogan, China, on November 24, 1995. He received the B.S. degree in Mechatronic Engineering from the Huazhong Agricultural University, Wuhan, China, in 2017. He received the M.S. degree in Control Science and Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2020. His research interests focus on multi-object tracking, object detection, and person re-identification.

Hongwei Wang was born in Hubei, China, on August 10, 1998. He received the B.S. degree in Control Science and Engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. He is now pursuing the M.S. degree in the School of Artificial Intelligence and Automation at HUST. His research interests focus on multi-object tracking, motion prediction, object detection, and person re-identification.

En Yu was born in Linjiang, China, on July 5, 1998. He received the B.S. degree in Control Science and Engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. He is now pursuing the M.S. degree in the School of Artificial Intelligence and Automation at HUST. He devotes himself to the research of multi-object tracking, object detection, person re-identification, and self-supervised learning.

Donghaisheng Liu received the B.S. degree in Automation from Zhengzhou University, Zhengzhou, China, in 2019. He is currently pursuing the M.S. degree in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests focus on person re-identification, object detection, and image super-resolution.

Xiaofeng Pan was born in Anhui, China, on January 4, 1998. He received the B.S. degree in Control Science and Engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2020. He is currently pursuing the M.S. degree in the School of Artificial Intelligence and Automation at HUST. His research interests focus on object detection, instance segmentation, and pedestrian attribute recognition.