ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240543

MC-SORT: A Motion Correction-Based Framework for Long-Term Multiple Object Tracking

Xiangyu Li[a], Yunchuan Qin[a,*], Ruihui Li[a,*], Guanghua Tan[a], Zhuo Tang[a] and Kenli Li[b,a]

[a]College of Computer Science and Electronic Engineering, Hunan University
[b]State Key Laboratory of Advanced Design and Manufacturing Technology for Vehicle, Hunan University

Abstract. Long-term occlusion is one of the most formidable challenges in Multi-Object Tracking (MOT). The motion models of existing SORT-based trackers are unreliable in estimating the motion states of long-term occluded targets. This is mainly because as the occlusion period increases, the increases speed of estimation errors in the motion model increases faster. In practical applications, we believe that the estimation error of the tracker during long-term occlusion is mainly concentrated in the estimation error of the motion model on the velocity of the occluded target. In this work, we have demonstrated that in the long-term occlusion period, appropriately correcting the estimated values of the motion model on the target motion velocity and fully utilizing the temporal and attribute information of the target's historical trajectory as calculation indicators of correlation are beneficial for improving the robustness of the tracker in long-term occlusion. We refer to our proposed motion correction-based framework as MC-SORT, which mainly consists of a Momentum Compensation Module (MCM) and a Backtracking Re-association (BRA) module. The former can correct the estimated value of the target's motion state during long-term occlusion, the latter uses the temporal and attribute information of the target's historical trajectory during long-term occlusion as correlation indicators to measure the degree of correlation between the target and trajectory. Our proposed MC-SORT has the characteristics of simplicity, online, real-time, and plug-and-play, particularly improving the robustness of the tracker in long-term occlusion. The extensive experimental results on the MOT17 and MOT20 datasets demonstrate the robustness and superiority of our framework.

1 Introduction

Multi-object Tracking (MOT) represents a classic challenge in the field of computer vision, its goal is to reliably track the trajectory of each object within a continuous video stream. MOT serves as a fundamental task for various complex real-world applications, including autonomous driving[8], intelligent search and rescue[7], and intelligent supervision[25], among others. With the continuous development of object detection technology and the proposal of real-time online tracking paradigm SORT[4], significant advancements have been made in the MOT domain. Nevertheless, many technical challenges remain, with occlusion being one of the most critical issues.

As shown in Figure 1 (a), occlusion can roughly be divided into two main categories for analysis: mutual occlusion between pedestrians and occlusion caused by obstacles. Most short-term occlusions are due to the first type, while the second type, especially involving large obstacles, often leads to long-term occlusion. Current SORT-based trackers perform well in short-term occlusion but have a systemic limitation for long-term occlusion, which we will discuss in detail in Section 3.1. Our goal is to develop a motion model-based MOT method that is robust to long-term occlusion. We have analyzed existing SORT-based trackers and found that most of them exhibit systematic shortcomings in handling long-term occlusion. As shown in Figure 1 (b), when the target reappears after a long-term occlusion, the error between the estimated value and the actual value of the target far exceeds the tracking ability of SORT-based trackers. We analyze that this may be caused by the accumulation of estimation errors in the motion model during long-term occlusion, and we will explain the specific analysis process in detail in Section 3.1. Although current SORT-based trackers have excellent performance in dealing with unobstructed or short-term occluded targets, there is an

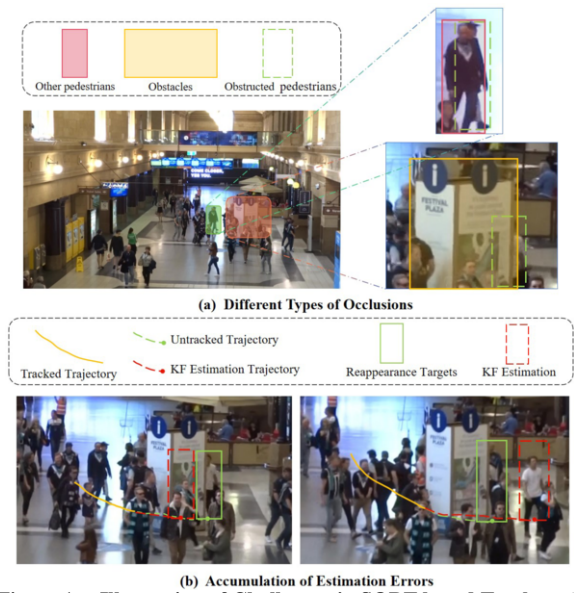

Figure 1. Illustration of Challenges in SORT-based Trackers.(a) describing the occurrence of different occlusions. The green area triggers short-term occlusion caused by mutual occlusion between pedestrians. At the same time, the tangerine area has triggered long-term occlusion caused by a large obstacle. (b) illustrates the accumulation of estimation error during the long-term occlusion. These two figures respectively illustrate two situations where the motion model estimates the target velocity lower than the actual target velocity and the motion model estimates the target velocity higher than the actual target velocity. The estimation errors after long-term occlusion far exceed the association ability of SORT-based trackers, ultimately leading to tracking failure.

* Corresponding Author.

urgent need to address their limitations in long-term occlusion.

In this work, we realized a significant limitation of SORT-based trackers in dealing with long-term occlusion. The error between the estimated and true values of the target trajectory during long-term occlusion increases with the growth of the occlusion period. When the target reappears, the huge errors will ultimately lead to a complete loss of the target. Although the observation-centered approach proposed by OC-SORT aims to alleviate the process noise accumulation in motion models caused by occlusion, it does not consider the temporal information of target motion and the attitude information of the target, with some limitations in dealing with long-term occlusion.

To alleviate the limitations of SORT-based trackers under long-term occlusion, we propose a hypothesis that in long-term occlusion, the motion model's estimation of the direction consistency of the target trajectory is more reliable than its estimation of the target velocity. Based on this assumption, we propose a plug-and-play framework based on motion correction, which mainly consists of two parts. Firstly, for targets entering the long-term occlusion, we compensate for their motion estimation during the occlusion period, reduce the error between their estimation and the actual values, alleviate the accumulation of estimation errors during the long-term occlusion, and correct the target's motion trajectory during the occlusion period. We call this part the Momentum Compensation Module (MCM), which mainly corrects and compensates for the target's motion during the long-term occlusion. Next, we re-associate the target that has reappeared after long-term occlusion with the existing trajectories of long-term occlusion targets. We incorporate the attitude information (location, shape, and physical attributes of targets) and temporal information (the generation time of each estimation) of the target's historical trajectory into the correlation matrix. We call this part the Backtracking Re-association (BRA), which mainly uses the motion-corrected target's historical trajectory to calculate the correlation between different trajectories and the reappeared target.

The proposed method is called MC-SORT, which has the characteristics of simplicity, online, real-time, and plug-and-play, and it can significantly improve the robustness over long-term occlusion. Our contributions are summarized as the following:

- We recognize a critical limitation of SORT-based trackers under long-term occlusion. Based on experience and analysis, to address this limitation, we propose a hypothesis that under long-term occlusion, the estimation of target trajectory direction consistency by the motion model is more reliable than the estimation of target velocity.
- We propose a motion correction-based framework, MC-SORT, to alleviate the limitations of SORT-based trackers under long-term occlusion. It has the characteristics of simplicity, real-time, online, and plug-and-play, and without any additional learning and training.
- We have designed a Momentum Compensation Module (MCM) and a Backtracking Re-association (BRA) module, which can improve the association ability of the tracker under long-term occlusion. They effectively improve the performance of existing SORT-based trackers under long-term occlusion and perform well on multiple datasets.

2 Related Work

2.1 Tracking-by-Detection

The Tracking-by-Detection paradigm typically consists of three parts: an object detector[29, 13, 12], a motion estimation model[17], and an association module. In this paradigm, the motion model first estimates each retained trajectory, while the object detector detects the current frame result, and then inputs the estimated value and detection value into the correlation module to generate the tracking result of the current frame. Various methods have been applied to enhance this paradigm, achieving seemingly gratifying progress. SORT[4], pioneering the use of the Kalman Filter (KF)[17] as a motion model, estimates future states of objects, marking a significant leap forward. DeepSORT[34] proposes a cascaded matching strategy based on SORT[4], it classifies trajectories of different qualities into different priorities and matches them at different levels, the shorter the occlusion period, the higher the priority. Additionally, other approaches[27, 33, 34, 36] seek to harness the appearance attributes of different objects to enhance match precision, incorporate the extracted appearance features into the associated cost matrix. ByteTrack[40] uses low confidence detection results for re-association, which can fully utilize detection information as much as possible. It is different from other trackers that filter low-confidence detection. Bot-SORT[1] fine-tunes the Kalman Filter (KF)[17] parameters and introduces a Global Motion Compensation (GMC) technique, aiming to refine the estimation of motion. OC-SORT[6] adopts an observation-centered tracking method instead of an estimation-centered approach to reduce the accumulation of motion model errors. Building upon OC-SORT[6], Deep OC-SORT[22] utilizes Camera Motion Compensation (CMC) and dynamic update strategies, further enhancing match accuracy. Adopting this Tracking-by-Detection paradigm, we propose a framework focused on motion correction, particularly aimed at tackling long-term occlusions.

2.2 Motion Models

Following the Kalman Filter's (KF)[17] adoption as the foundational motion model in multi-object tracking by SORT[4], this approach has gained popularity and widespread implementation in numerous studies[34, 40, 39, 15]. Several investigations focus on refining detection results to improve the precision of motion model forecasts. Study [2] addresses detection biases induced by camera vibrations through camera motion compensation (CMC). Bot-SORT[1] employs a global motion compensation (GMC) technique to tackle detection biases spanning various frames. GIAOTracker[11] introduces the NSA Kalman Filter (NSA-KF) based on the Unscented Kalman Filter (UKF), factoring in detection scores in target motion estimations. CIWT[26] attempts to employ the Extended Kalman Filter (EKF) as the motion model for forecasting the trajectories of nonlinearly moving targets.

2.3 Occlusion

Occlusion of targets presents a significant challenge in the field of multi-object tracking. Some approaches[14, 36] strive to categorize targets depending on their occlusion status. Study [41] employs head detection to mitigate target disappearance within crowds, effectively reducing the occurrence of occlusions and enhancing tracking efficacy. Quo Vadis[10] proposes a method combining homography estimation with depth maps to generate a pseudo-3D scene, aimed at accurately predicting the movements of occluded targets and enhancing motion forecast precision. Some methods[5, 35] leverage a memory bank to bolster long-term tracking precision. This memory bank is capable of diminishing the uncertainties in embedding

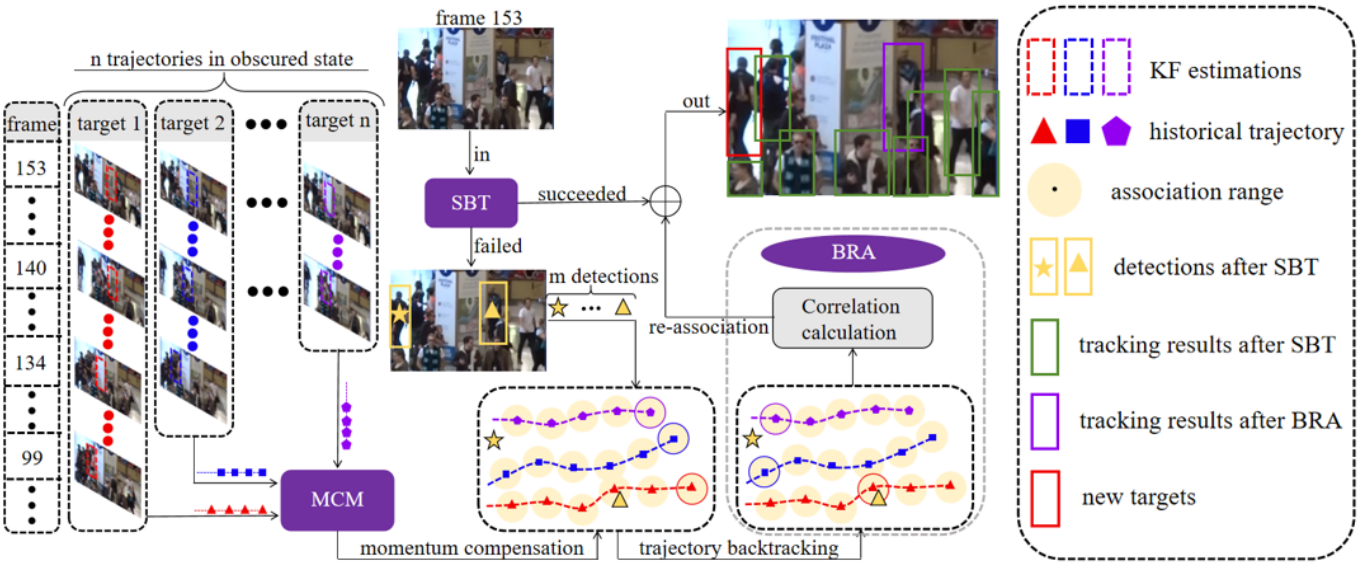

Figure 2. The pipeline of our proposed MC-SORT. SBT represents the SORT-based tracker, MCM represents our proposed **Momentum Compensation Module**, and BRA is the **Backtracking Re-association** module. When the occluded targets reappear at frame 153, the first attempt is to track them by SBT, and at the same time, the MCM is used to compensate for the momentum of n trajectories in the occluded state to obtain n historical trajectories that have already been momentum-compensated. The m targets that were failed tracked by SBT and the n historical trajectories that passed through MCM will proceed to the next step. Next, input the n historical trajectory and m suspicious detections from the previous step into the BRA module to calculate their correlation and generate a re-association result. Finally, add the generated re-association result to the successful tracking result by SBT to obtain the final tracking result.

or appearance features due to occlusions, thereby aiding in the reduction of data association inaccuracies stemming from occlusions. MotionTrack[28] proposed a learnable interaction module capable of conducting correlation analyses on historical trajectory features to re-identify occluded targets. Some others[31, 32, 16] utilize the Re-ID technique to enhance data association precision in occlusion scenarios. Our research mainly focuses on the accumulation of errors between the motion model estimation and the true values when the target is obscured for a long time. Specifically, we modify the existing motion estimation through the historical trajectory of the target to reduce the accumulation of errors.

3 Methodology

In this section, we first analyze the limitations of existing SORT-based trackers and introduce our proposed method MC-SORT from their limitations. Next, we will provide a detailed introduction to the two important components of MC-SORT, the Momentum Compensation Module (MCM) and the Backtracking Re-association (BRA) module, in the following text. It should be noted that our method is based on the limitations of SORT-based trackers and is an effective supplement to existing SORT-based trackers. It does not require additional learning and training and can be plug-and-play.

3.1 Limitation of SORT-based methods

The SORT-based trackers mainly consist of two parts: motion model (Kalman Filter) and association algorithm (Hungarian Algorithm). This type of tracker first uses a motion model to estimate the trajectory estimation value for the next frame based on the existing trajectory set, then uses IoU or appearance features as association evaluation indicators between the detection set and the trajectory estimation set to generate the corresponding correlation matrix, and finally forms the tracking result through association algorithms. This type of tracker is particularly effective for scenes with little or no occlusion, especially if all targets, in reality, can be accurately detected, then its tracking results can be completely consistent with the actual target motion situation.

However, in practical applications, due to the imaging principle of images, completely occluded targets can be considered as completely lost data in the image. Therefore, current image-based detection algorithms are unable to accurately detect such occluded targets, making occlusion a major challenge in the field of multi-target tracking. To solve the occlusion problem, SORT-based trackers are based on the assumption that the target's displacement in a short period is a uniform linear motion. Therefore, when the target is occluded in a short period, it is feasible to predict the short-term occluded trajectory of the target through the linear motion prediction model Kalman Filter, because the error between the estimated value of the Kalman Filter and the true value of the target is within the ability range of the association algorithm. OC-SORT has demonstrated the sensitivity of the Kalman Filter to state noise and the amplification of errors by occlusion time. Therefore, we have noticed a limitation of the SORT-based trackers. When long-term occlusion occurs, due to the accumulation of motion errors by the Kalman Filter, the estimated trajectory may have a huge deviation from the true value. That is, the motion estimation value of the Kalman Filter is unreliable when long-term occlusion occurs. Based on this analysis, we have reconsidered the observation-centered update strategy proposed by OC-SORT and its proposed OCM module which relies on consistency in the direction of movement. Combining their performance on the MOT and DanceTrack datasets, we propose a hypothesis that for most long-term occluded targets, because their destination is generally known and clear, their motion direction will rarely change significantly during the occlusion period. Therefore, we believe that the accumulation of motion errors by the Kalman Filter is mainly reflected in the target's motion velocity rather than the target's motion direction. That is, in long-term occlusion, we can trust the Kalman Filter's upward estimation and reduce its dependence on motion velocity estimation.

To verify our hypothesis, we selected MOT17 and MOT20 as

the analysis datasets. For details, please refer to Section I in Appendix[18]. Thanks to the visible parameter of their true values, we can easily simulate different situations of occlusion. Through data analysis and calculation, we can preliminarily prove our hypothesis from the statistical results. Based on this, we identified a limitation of SORT-based trackers, where the estimation of target velocity by the motion model is not reliable during long-term occlusion, severely suppressing the performance of SORT-based trackers in long-term occlusion scenarios. Based on our hypothesis, we propose a plug-and-play framework MC-SORT as a supplement to existing SORT-based trackers, which mainly consists of two modules: the MCM module and the BRA module. We will provide a detailed explanation of this framework in the following chapters.

3.2 *Overview of MC-SORT*

MC-SORT mainly executes two steps for long-term occlusion targets:

- **Step 1: Motion State Compensation.** Compensate for the motion state of each target trajectory that has entered a long-term occlusion.
- **Step 2: Re-association Reappearing Targets.** Backtracking the historical trajectory of targets that have reappeared after long-term occlusion and re-association them to the trajectory.

Each trajectory that has undergone long-term occlusions and reappearance will go through several stages: being tracked period, short-term occlusion period, long-term occlusion period, and reappearance period after long-term occlusion. For the two stages of being tracked and short-term occlusion, we believe that the current SORT-based trackers can adapt well, so our method mainly focuses on improving the long-term occlusion period and the reappearance period after long-term occlusion.

Motion State Compensation. In Section 3.1, we proposed a hypothesis that the error accumulation of the Kalman Filter under long-term occlusion is mainly concentrated on the motion velocity. Based on this assumption, we consider a scenario in which the estimated velocity of the Kalman Filter is much lower than the true target motion velocity when the target is in a long-term occlusion. This scenario can lead to a large error between the estimated value of the Kalman filter and the true value of the target, exceeding the limit of the association algorithm, resulting in tracking failure. SORT generates correlation matrices based on IoU evaluation metrics as the cost matrix for the association algorithm. Therefore, when there is no overlap between the motion estimation value and the true value, association failure will inevitably occur during the association process, leading to tracking failure. Although OC-SORT proposes the OCM module from an observational perspective that incorporates the direction consistency of tracks in the cost matrix for the association, for long-term occluded targets, the generated cost matrix completely loses temporal and target attribute information, so there are still some limitations. As shown in Figure 3, the impact of temporal information and attribute information is demonstrated. When the target enters a long-term occlusion, according to the method proposed by OC-SORT, only considering the influence of the object's motion direction, it can be concluded that the correlation between the reappeared target and $target_2$ is higher, because the motion direction estimation of $target_2$ is more closely related to the reappeared target than the estimation of $target_1$, which is completely opposite to the actual situation. In fact, the real situation is that $target_1$ has a higher correlation with the reappeared target because the attribute and temporal

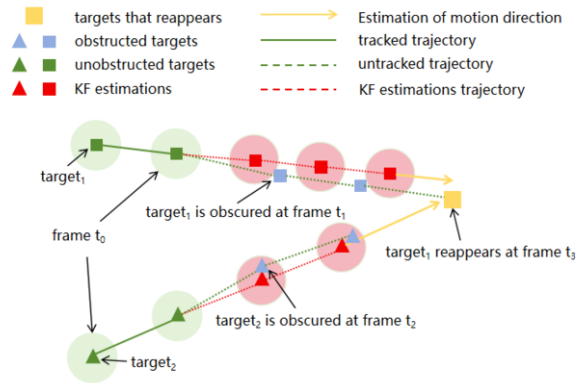

Figure 3. Impact of Temporal and Attribute information. Different shapes represent the attribute of different targets ("▲" represent $target_1$, "■" represent $target_2$), and the circle area with shadow indicates the range that an estimate can be associated with the truth value. Different colors of the same target represent different states of the target (green represents the tracked state, light blue represents obscured targets actual value, yellow represents reappearing targets, and red represents obscured targets estimation). t_n represents different moments, where $target_1$ is obscured at frame t_1 while $target_2$ is not obscured at this time. $target_2$ is obscured at frame t_2, at which point both $target_1$ and $target_2$ are in obscured state. At frame t_3, $target_1$ reappeared, while $target_2$ was still in occlusion.

characteristics of $target_1$ are closer to the reappeared target. Even from the perspective of movement direction, the difference between the correlation of $target_1$ and the correlation of $target_2$ is limited. For this reason, we propose a momentum compensation module as a supplement to the motion model under long-term occlusion, compensating for the motion state during long-term occlusion. Compared with the method proposed by OC-SORT, we incorporate the temporal and attribute information of the target into the calculation of the correlation matrix in the association algorithm, fully considering the impact of temporal and attribute information.

Re-association Reappearing Targets. Due to the significant deviation in the estimation of motion velocity by the motion model mentioned earlier, there is a situation where the estimated velocity of the motion model exceeds the true motion velocity of the target when it reappears after long-term occlusion. To address this issue, we extended the original trajectory information from the temporal dimension, preserving both the temporal and attribute information of the target during long-term occlusion. All the retained information is called the historical trajectory set, abbreviated as the historical trajectory. When a target that has been obscured for a long time reappears, the historical trajectory of the target is used as input for the association algorithm to calculate the correlation between the estimated trajectory and the reappeared target, and ultimately re-associate the reappeared target.

3.3 *Momentum Compensation Module*

In practice, when the target is in a long-term occlusion period, the error between the estimated motion value of the target and its true value will increase with the increase of occlusion time, which is called error accumulation.

Assuming the limit range of the association algorithm part in the SORT-based trackers is δ. When the error $E(i, t_0)$ between the motion estimation value $\hat{S}_i(t_0)$ and the true value $S_i(t_0)$ of the $target_i$ at frame $t_0 \in (t_{start}, t_{end})$ is exceeding δ, it indicates that the $target_i$ failed to track at frame t_0. Therefore, it can be easily concluded that at frame $t \in (t_0, t_{end})$, the error $E(i, t)$ between the

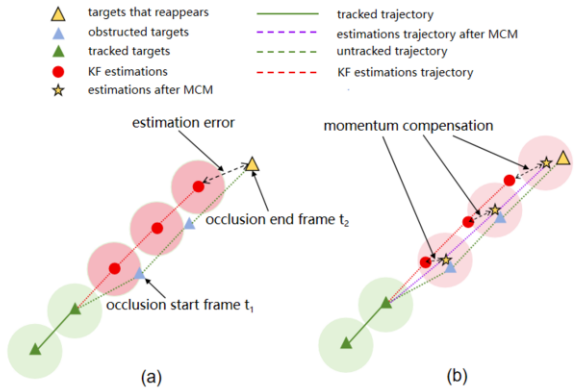

Figure 4. Illustration of how MCM works during long-term occlusion. (a) describes the working process of KF when occlusion occurs. The target enters the occluded state starting from frame t_1 and reappears at frame t_2. The red dot represents the KF estimation during the occlusion and the circle area with shadow indicates the range that an estimate can be associated with the truth value. The accumulation of errors from frame t_1 to frame t_2 resulted in the estimation error between the reappeared target and the original trajectory exceeding the association range, leading to tracking failure. (b) describes the scenario of using the MCM module. The red dot represents the original KF estimation and the yellow star represents the KF estimation through MCM. After entering the occluded state, the MCM compensates the momentum of the historical trajectory to suppress the accumulation of estimation errors, so that the estimation error between the target estimation value at frame t_2 and the true value is within the association range, and the target can be re-associated with the original trajectory.

estimated value $\hat{S}_i(t)$ and the true value $S_i(t)$ will always exceed the limit range of the association algorithm, due to the principle of error accumulation. Therefore, when the target reappears, it will inevitably fail to associate with the original trajectory, resulting in the permanent disappearance of $target_i$. The calculation of $E(\cdot)$ is as follows:

$$E(i,t) = S_i(t) - \hat{S}_i(t) \quad , t \in (t_{start}, t_{end}) \tag{1}$$

where $S_i(t)$ represents the motion state of the $target_i$ at frame t, $\hat{S}_i(t)$ represents the motion estimation state of the $target_i$ at frame t, t_{start} represents the start frame of the long-term occlusion period, t_{end} represents the end frame of the long-term occlusion period.

In order to solve the problem of the estimated and true values exceeding the limit range of the association algorithm during the long-term occlusion period, we propose a Momentum Compensation Module (MCM) to alleviate this problem. As mentioned earlier, we believe that the error between the estimated value and the true value is mainly reflected in the difference in the target velocity. The main function of the MCM is to compensate for the error between the estimated value and the true value, weaken its error accumulation effect, and reduce the error value to within the limit range δ of the association algorithm. We consider that the obscured time is positively correlated with the error value, and the specific calculation of the momentum compensation value $M_c(i,t)$ for $target_i$ at the frame t is as follows:

$$M_c(i,t) = \Delta\hat{S}_i(t) \times T(t - t_{start}) \quad , t \in (t_{start}, t_{end}) \tag{2}$$

where the $T(\cdot)$ means the temporal correlation function, it is positively correlated with the obscured time of targets, it dynamically adjusts the motion correction by the occlusion time to suppress the increase of estimation error caused by the increase of occlusion time. The $\Delta\hat{S}_i(t)$ means the relative estimated value of the motion model. The specific calculations for $\Delta\hat{S}_i(t)$ and $T(\cdot)$ are as follows:

$$\Delta\hat{S}_i(t) = \hat{S}_i(t) - \hat{S}_i(t-1) \tag{3}$$

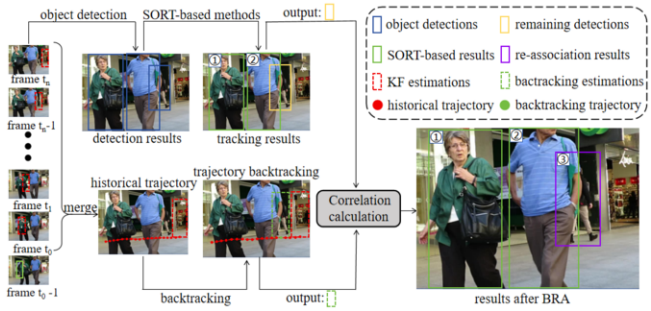

Figure 5. Illustration of the BRA module. t_0 is the starting frame for $target_i$ to enter long-term occlusion, and t_n is the ending frame. The BRA module acts on frame t_n where the target reappears after long-term occlusion. When the occluded target reappears, the historical trajectory of $target_i$ is backtracking to frames $t_n - 4$, and the temporal and attribute information of the backtracking is used to calculate the correlation between $target_i$ and the reappeared target. Finally, the trajectory of $target_i$ is re-associated with the reappeared target through the BRA module.

$$T(t - t_{start}) = \sigma \times \sqrt{Exp(\frac{t - t_{start}}{t_{end} - t_{start}})} \tag{4}$$

where σ represents temporal correlation weight. It should be noted that we only use the MCM module for momentum compensation on target trajectories that are in the long-term occlusion period, and retain all momentum-compensated motion states and temporal information of the target during the long-term occlusion period as its historical trajectory, denoted as $H^{traj} = \{\hat{S}(t_{start}) + M_c(t_{start}), \cdots, \hat{S}(t_{end}) + M_c(t_{end})\}$.

3.4 Backtracking Re-association

When the target reappears after long-term occlusion, due to the accumulation of errors, it may occur that the motion model's estimated value of the target's current frame exceeds its actual position, leading to the failure of target association. Although we have compensated for the momentum of the target's motion estimation during occlusion, we still need to consider the occurrence of this situation, especially the possibility of excessive compensation. For this reason, we propose a Backtracking Re-association (BRA) module to solve the above issue.

Specifically, based on the assumption we proposed earlier, if the estimated value in the direction of the target motion vector is reliable, even if the estimated value when the target reappears exceeds its actual value, its reappearance position will inevitably be associated with an estimated value in the historical trajectory during its long-term occlusion period. At the same time, we consider the impact of temporal information, as the increase in occlusion time amplifies the error in the estimated values of the target motion model. Therefore, estimates with larger temporal values in historical trajectories have lower confidence levels.

We use the historical trajectory of long-term occlusion targets as the basic input unit of BRA. Taking n long-term occluded target trajectories $\mathcal{H} = \{H_1^{traj}, \cdots, H_n^{traj}\}$ and m newly emerged target detection results $\mathcal{D} \in \mathbb{R}^{m \times 4}$ as examples. For individual trajectory $H_i^{traj} \in \mathcal{H}$ and detection result $D_j \in \mathcal{D}$, the correlation score calculation is as follows:

$$C_{br}(H_i^{traj}, D_j) = \max_t \left(\Phi_i(t) \times C_{IoU}(H_i^{traj}(t), D_j)\right) \tag{5}$$

where the $H_i^{traj}(t)$ represents the motion estimation value of the historical trajectory set of $target_i$ after passing through MCM at the

Table 1. Comparison with the state-of-the-art methods under the "private detector" protocol on the MOT17 test set. ↑ means higher is better, ↓ means lower is better. The best results for each metric are **bolded**.

Tracker	Venue	IDF1↑	HOTA↑	MOTA↑	AssA↑	FP↓	FN↓	IDs↓	Frag↓
CSTrack[19]	TIP'22	72.6	59.3	74.9	57.9	23847	114303	3567	7668
RelationTrack[37]	TMM'22	74.7	61	73.8	61.5	27999	118623	1374	2166
TrackFormer[23]	CVPR'22	68	-	74.1	-	34602	108777	2829	-
MeMOT[5]	CVPR'22	69	56.9	72.5	55.2	37221	115248	2724	-
MTrack[36]	CVPR'22	73.5	-	72.1	-	53361	101844	2028	-
MOTR[38]	ECCV'22	68.6	57.8	73.4	55.7	-	-	2439	-
ByteTrack[40]	ECCV'22	77.3	63.1	80.3	62.0	25491	83721	2196	2277
QuoVadis[10]	NeurIPS'22	77.7	63.1	80.3	62.1	25491	83721	2103	2277
P3AFormer[42]	ECCV'22	78.1	-	81.2	-	17281	86861	1893	-
OC-SORT[6]	CVPR'23	77.5	63.2	78	63.2	**15100**	108000	1950	2040
MotionTrack[28]	CVPR'23	80.1	65.1	81.1	65.1	23802	81660	1140	**1605**
UTM[35]	CVPR'23	78.7	64	**81.8**	62.5	25077	**76298**	1431	1889
MC-SORT (ours)	-	**80.9**	**65.2**	80.6	**65.7**	22605	85494	**1125**	1794

Table 2. Comparison with the state-of-the-art methods under the "private detector" protocol on the MOT20 test set. ↑ means higher is better, ↓ means lower is better. The best results for each metric are **bolded**.

Tracker	Venue	IDF1↑	HOTA↑	MOTA↑	AssA↑	FP↓	FN↓	IDs↓	Frag↓
CSTrack[19]	TIP'22	68.6	54	66.6	50.0	25404	144358	3196	7632
RelationTrack[37]	TMM'22	70.5	56.5	67.2	56.4	61134	104597	4243	8236
MeMOT[5]	CVPR'22	66.1	54.1	63.7	55.0	47882	137982	1938	-
MTrack[36]	CVPR'22	69.2	-	63.5	-	96123	86964	6031	-
ByteTrack[40]	ECCV'22	75.2	61.3	77.8	59.6	26249	87594	1223	1460
QuoVadis[10]	NeurIPS'22	75.7	61.5	77.8	59.9	26249	87594	1187	1460
P3AFormer[42]	ECCV'22	76.4	-	78.1	-	25413	86510	1332	-
OC-SORT[6]	CVPR'23	75.9	62.1	75.5	62.0	**18000**	108000	**913**	**1198**
MotionTrack[28]	CVPR'23	76.5	62.8	78	61.8	28629	84152	1165	1321
UTM[35]	CVPR'23	76.9	62.5	**78.2**	61.4	29964	**81516**	1228	1342
MC-SORT (ours)	-	**77.7**	**63.6**	77.9	**63.3**	23573	89394	1233	1463

frame t, $C_{IoU}(H_i^{traj}(t), D_j)$ calculates the IoU (Intersection over Union) between $H_i^{traj}(t)$ and D_j. The $\max_t(\cdot)$ indicates the maximum is calculated along the direction of the temporal dimension t of the $H_i^{traj}(t)$. The $\Phi(\cdot)$ means temporal confidence function, it is a function used to describe the credibility of estimated values. For $target_i$, the calculation of $\Phi_i(t)$ is as follows:

$$\log(\Phi_i(t)) = \frac{t - t_{start}}{t_{end} - t_{start}} \times \log(\lambda) \quad , t \in (t_{start}, t_{end}) \tag{6}$$

where t_{start} represents the starting time when $target_i$ enters long-term occlusion, t_{end} represents the time when detection D_j appears, which is the time when the $target_i$ reappears, and λ represents the temporal confidence weight.

4 Experiments

4.1 Experimental Setup

Datasets. To verify the robustness of our proposed MC-SORT for long-term occlusion, we selected MOT17[24] and MOT20[9] as experimental datasets, due to their dense targets and numerous obstacles. We evaluate our MC-SORT on MOT17[24] and MOT20[9] under the "private detection" protocol.

Metrics. We use HOTA[21] as the main indicator because it maintains a better balance between detection and association accuracy, which can evaluate the performance of the tracker from an overall perspective. We also emphasize AssA[21] and IDF1[30] to evaluate the performance of associations, as they better reflect the accuracy of the association. Other metrics we report, such as MOTA[3], are highly related to detection performance.

Implementation Details. We implemented our MC-SORT in PyTorch and performed all experiments on one NVIDIA A100. In order to fairly evaluate our framework, we have inserted our designed module into existing methods while ensuring the integrity of the original algorithm. To ensure fairness, the universal YOLOX[13] detector is employed, utilizing weights trained by ByteTrack[40] on both MOT17 and MOT20 datasets. To verify the generality and effectiveness of our method, our framework does not require any additional training. For the tracking process, we use the same strategy as ByteTrack[40], using a dual-layer data association strategy. The default high and low thresholds are 0.6 and 0.1, respectively. After unsuccessful backtracking by the History Backtracking algorithm module, the threshold for newly generated trajectories needs to be set to 0.7. The temporal correlation weight in MCM is set to 0.025, and the temporal confidence weight in BRA is set to 0.6.

4.2 Benchmark Results

MOT17. As shown in Table 1, our MC-SORT framework outperforms state-of-the-art methods in most key metrics. Especially, the indicators IDF1, HOTA, AssA, and IDs all rank first, which proves that our proposed framework has reached the most advanced level in association accuracy, and we have also reached the current advanced level in other indicators, demonstrating the robustness of our algorithm. Our framework focuses on solving the problem of long-term occlusion, which enables the tracker to have stronger capabilities when facing long-term occlusion. Therefore, it can generate more accurate association results, which can also be reflected in the indicators reflecting association ability (IDF1, AssA).

MOT20. As shown in Table 2, our MC-SORT framework still achieved state-of-the-art results on the MOT20 dataset under the "private detector" protocol. It should be pointed out that UTM ranks first in MOTA metrics because of its complex structures and mechanisms, and it is only slightly ahead of our results in MOTA. Our framework does not require additional learning and training, and is a plug and play simple framework. In this case, we are only 0.3 lower than UTM

Table 3. Ablation studies of the **Momentum Compensation Module**(**MCM**) and the **Backtracking Re-association**(**BRA**) on the MOT17 validation set. The best results for each metric are **bolded**.

Setting	HOTA↑	IDF1↑	MOTA↑	AssA↑
Baseline	69.19	82.03	78.64	71.50
Baseline+BRA	69.38	82.35	78.84	71.75
Baseline+MCM	69.54	82.60	78.85	72.22
Baseline+BRA+MCM	**69.67**	**82.83**	**78.85**	**72.48**

Table 4. Results of applying MC-SORT to 4 different state-of-the-art trackers on the MOT17 validation set. "†" means that the tracker uses the Re ID module. The best results for each metric are **bolded**.

Method	w/MC-SORT	HOTA(↑)	IDF1(↑)	MOTA(↑)
ByteTrack[40]		67.88	79.77	77.66
	✓	**68.48**	**81.03**	**78.03**
BoT-SORT[1]		69.14	81.62	78.44
	✓	**69.50**	**82.30**	**78.91**
BoT-SORT†[1]		68.73	81.50	78.48
	✓	**69.67**	**82.83**	**78.85**
SparsTrack[20]		69.00	81.97	77.87
	✓	**69.23**	**82.13**	**78.00**
OC-SORT[6]		66.37	77.93	**74.54**
	✓	**66.74**	**78.69**	74.42

in the MOTA indicator, but HOTA exceeds it by 1.1, IDF1 exceeds it by 0.8, and AssA exceeds it by 1.9, which reflects our association ability far superior to UTM. Our method ranks first in HOTA, IDF1, and AssA metrics, fully demonstrating the advantages of our method in terms of association ability.

4.3 Ablation Study

Effect of Each Component. Table 3 lists the contributions of the different modules we proposed on the MOT17 validation set. Due to the plug-and-play characteristic of the modules we propose, we can directly add and remove corresponding modules to achieve ablation effects. The results demonstrate the effectiveness of our proposed module in MC-SORT. The results indicate that both the MCM module and the BRA module alone can improve the performance of the baseline, especially in terms of the improvement in association ability, which is significantly intuitive. When the MCM and BRA modules work together, the baseline improvement reaches its maximum, indicating that our proposed modules have a positive interaction between themselves. The overall performance of the ablation study indicates that our proposed method is effective in improving the association ability of the tracker.

Applications on other trackers. We have selected several state-of-the-art SORT-based trackers to demonstrate the generality and robustness of our proposed MC-SORT framework. Due to the plug-and-play characteristic of our proposed method, we can directly apply our approach to these trackers. The experimental results indicate that our method has varying degrees of improvement on different SORT-based trackers, which proves the generality and robustness of our proposed method.

Extra Evaluation of Crowd and Occlusion. Our approach is mainly proposed to address the challenge of long-term occlusion. To verify the effectiveness of our proposed method under long-term occlusion, inspired by the crowdMOT dataset proposed in MotionTrack[28], we transformed the MOT17 validation set into a subset with more frequent long-term occlusion occurrences. Thanks to the visibility score in the MOT dataset, we consider the true value of the visibility score below ρ as the occluded value. When a target has a continuous Δt frame visibility score below ρ throughout the entire video stream, we consider it as a track with long-term occlusion. We extract all the targets that suit the conditions in the MOT17 validation set as a new validation set to demonstrate the effectiveness of our proposed method under long-term occlusion. We evaluated the performance of the validation sets formed by the two visibility thresholds ρ and several Δt. We chose HOTA, IDF1, and AssA as our evaluation metrics, which better reflect the tracking ability of the tracker. As shown in Table 5, the experimental results demonstrate the superiority and robustness of our proposed method for long-term occlusion, indirectly verifying the correctness of the theory proposed in Section 3.1.

Table 5. Evaluation of HOTA, IDF1, and AssA for crowd and occlusion cases on the MOT17 validation set. Increases in metrics are marked in green.

ρ	Metrics	Setting	$\geq$ **20**	$\geq$ **40**	$\geq$ **60**	$\geq$ **80**
$\leq$ **0.25**	HOTA(↑)	Baseline	41.68	33.95	31.30	25.29
		Ours	42.58	34.86	32.06	26.44
		Improvement	+0.89	+0.92	+0.76	+1.14
	IDF1(↑)	Baseline	46.31	34.68	30.04	21.42
		Ours	47.43	35.57	30.77	22.42
		Improvement	+1.12	+0.89	+0.72	+1.00
	AssA(↑)	Baseline	54.12	51.42	51.78	47.94
		Ours	56.13	53.93	54.20	52.19
		Improvement	+2.01	+2.52	+2.43	+4.25
$\leq$ **0.15**	HOTA(↑)	Baseline	32.92	25.45	20.95	14.79
		Ours	34.08	26.57	21.51	15.46
		Improvement	+1.16	+1.1	+0.56	+0.67
	IDF1(↑)	Baseline	34.53	22.96	17.04	9.97
		Ours	36.00	23.97	17.53	10.29
		Improvement	+1.47	+1.01	+0.50	+0.32
	AssA(↑)	Baseline	46.82	46.23	42.17	35.59
		Ours	49.87	50.16	44.25	38.78
		Improvement	+3.05	+3.93	+2.07	+3.19

5 Conclusions

We analyzed existing SORT-based trackers and recognized a limitation of them. Due to long-term occlusion, the error between the estimated value and the actual value of a motion model often exceeds the limit range of the association algorithm, leading to tracking failure. To address this issue, we propose a hypothesis based on experiments and experience that the error of the tracker's motion model during long-term occlusion is mainly reflected in its estimation of the motion velocity. Based on this assumption, we propose a motion correction-based framework MC-SORT. It can improve the robustness of existing SORT-based trackers under long-term occlusion while maintaining its simplicity, online, and real-time characteristics. Our experiments on different datasets have achieved state-of-the-art results, especially for multi-object tracking under long-term occlusion, where our method has significant gain effects. It is worth noting that our method does not require additional training and learning, and can be plug-and-play, which is an effective supplement to existing SORT-based trackers under long-term occlusion.

6 Future Work

Our MC-SORT mainly revises human-led motion estimation, which mainly faces multi-object tracking of traffic scenes. For non-traffic scenarios, such as tracking wild creatures such as bees, our method still has limitations (mainly due to the unpredictability of target motion estimation in these scenarios), and we will study the application of our method in non-traffic scenarios in the future.

Acknowledgements

The work is supported by the Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Major Project (2021ZD40300), National Natural Science Foundation of China (No.62202151), and Research Funds of State Key Laboratory of Advanced Design and Manufacturing Technology for Vehicle.

References

[1] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[2] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 941–951, 2019.

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[4] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[5] J. Cai, M. Xu, W. Li, Y. Xiong, W. Xia, Z. Tu, and S. Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8100, 2022.

[6] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9686–9696, 2023.

[7] J. L. Casper, M. Micire, and R. R. Murphy. Issues in intelligent robots for search and rescue. In *Unmanned ground vehicle technology II*, volume 4024, pages 292–302. SPIE, 2000.

[8] C. Chen, A. Seff, A. Kornhauser, and J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE international conference on computer vision*, pages 2722–2730, 2015.

[9] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[10] P. Dendorfer, V. Yugay, A. Osep, and L. Leal-Taixé. Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? *Advances in Neural Information Processing Systems*, 35:15657–15671, 2022.

[11] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong. Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 2809–2819, 2021.

[12] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[13] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[14] S. Guo, J. Wang, X. Wang, and D. Tao. Online multiple object tracking with cross-task synergy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8136–8145, 2021.

[15] S. Han, P. Huang, H. Wang, E. Yu, D. Liu, and X. Pan. Mat: Motion-aware multi-object tracking. *Neurocomputing*, 476:75–86, 2022.

[16] Y. Huang, F. Zhu, Z. Zeng, X. Qiu, Y. Shen, and J. Wu. Sqe: a self quality evaluation metric for parameters optimization in multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8306–8314, 2020.

[17] R. E. Kalman. A new approach to linear filtering and prediction problems. 1960.

[18] X. Li. Appendix for "MC-SORT: A Motion Correction-based Framework for Long-Term Multiple Object Tracking", Aug. 2024. URL https://doi.org/10.5281/zenodo.13284090.

[19] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022.

[20] Z. Liu, X. Wang, C. Wang, W. Liu, and X. Bai. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *arXiv preprint arXiv:2306.05238*, 2023.

[21] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021.

[22] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3025–3029. IEEE, 2023.

[23] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022.

[24] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[25] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[26] A. Osep, W. Mehner, M. Mathias, and B. Leibe. Combined image-and world-space tracking in traffic scenes. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1988–1995. IEEE, 2017.

[27] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021.

[28] Z. Qin, S. Zhou, L. Wang, J. Duan, G. Hua, and W. Tang. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17939–17948, 2023.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[30] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[31] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13813–13823, 2023.

[32] B. Shuai, A. G. Berneshawi, D. Modolo, and J. Tighe. Multi-object tracking with siamese track-rcnn. *arXiv preprint arXiv:2004.07786*, 2020.

[33] Q. Wang, Y. Zheng, P. Pan, and Y. Xu. Multiple object tracking with correlation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3876–3886, 2021.

[34] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[35] S. You, H. Yao, B.-k. Bao, and C. Xu. Utm: A unified multiple object tracking model with identity-aware feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21876–21886, 2023.

[36] E. Yu, Z. Li, and S. Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8843, 2022.

[37] E. Yu, Z. Li, S. Han, and H. Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 2022.

[38] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022.

[39] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.

[40] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[41] Y. Zhang, H. Chen, Z. Lai, Z. Zhang, and D. Yuan. Handling heavy occlusion in dense crowd tracking by focusing on the heads. In *Australasian Joint Conference on Artificial Intelligence*, pages 79–90. Springer, 2023.

[42] Z. Zhao, Z. Wu, Y. Zhuang, B. Li, and J. Jia. Tracking objects as pixel-wise distributions. In *European Conference on Computer Vision*, pages 76–94. Springer, 2022.