Focusing on Tracks for Online Multi-Object Tracking

Kyujin Shim Kangwook Ko Yujin Yang Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{kjshim1028, kokangook623, ujin.y, changick}@kaist.ac.kr

Abstract

Multi-object tracking (MOT) is a critical task in computer vision, requiring the accurate identification and continuous tracking of multiple objects across video frames. However, current state-of-the-art methods mainly rely on a global optimization technique and multi-stage cascade association strategy, and those approaches often overlook the specific characteristics of assignment task in MOT and useful detection results that may represent occluded objects. To address these challenges, we propose a novel Track-Focused Online Multi-Object Tracker (TrackTrack) with two key strategies: Track-Perspective-Based Association (TPA) and Track-Aware Initialization (TAI). The TPA strategy associates each track with the most suitable detection result by choosing the one with the minimum distance from all available detection results in a track-perspective manner. On the other hand, TAI precludes the generation of spurious tracks in the track-aware aspect by suppressing track initialization of detection results that heavily overlap with current active tracks and more confident detection results. Extensive experiments on MOT17, MOT20, and DanceTrack demonstrate that our TrackTrack outperforms current state-of-the-art trackers, offering improved robustness and accuracy across diverse and challenging tracking scenarios.

1. introduction

Multi-object tracking (MOT) is a fundamental task in computer vision that plays a crucial role in various applications [10, 20, 24, 26, 38, 41, 60]. However, MOT still faces diverse challenges due to the dynamic and unpredictable nature of real-world environments, including occlusions, similar appearances between different objects, and varying densities and motions of tracking targets. Many state-of-the-art methods [14, 56, 57, 61] follow the tracking-by-detection (TBD) framework, the dominant approach in MOT, to solve these problems. In this framework, trackers first detect objects in each frame and adequately link them to form complete trajectories across a video. More specifically, the detection results are compared with tracks, which are tracking results until the previous frame, and associated with the same objects using various distance measurements such as Intersection over Union (IoU) [8, 29, 61] and cosine distances of appearance features [1, 31, 36]. With the recent development of object detection techniques [18, 50], most approaches mainly focus on improving data association by building multiple association stages [33, 61] and introducing additional metrics, such as pseudo-depth [29] or moving direction [8, 56], while solving the matching between tracks and detection results with the Hungarian algorithm [34].

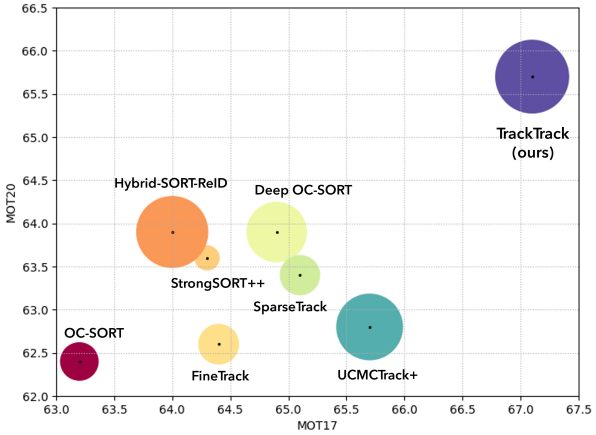
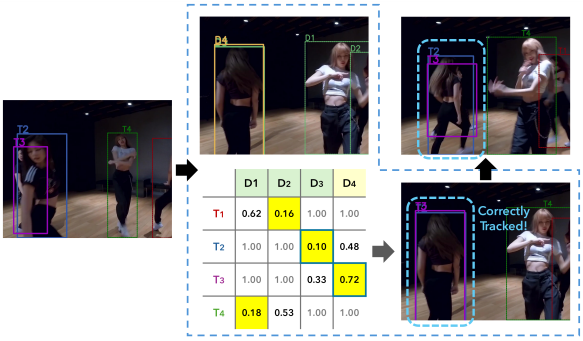


Figure 1. Comparisons of HOTA scores on the test set of MOT17, MOT20, and DanceTrack with our TrackTrack and other state-of-the-art methods. The radius of each circle means its HOTA score on DanceTrack. Our TrackTrack shows the highest and most consistently outperforming results on all datasets, demonstrating strong robustness and adaptability in every tracking scenario.

The Hungarian algorithm [34] is widely recognized for finding the minimum global cost matching between two independent sets, making it highly suitable for tasks requiring the lowest total costs, such as allocating workers to tasks in a factory. However, applying this algorithm to MOT during data association between tracks and detection results presents additional challenges. In MOT, tracks represent actual objects from previous frames, whereas detection results are object candidates of the current frame. Unlike matching between independent sets, as in job allocation to workers,

(a) Result of Hungarian matching



(b) Result of our Track-Perspective-Based Association (TPA)

Figure 2. Visualization of performance enhancement of our TPA compared to typical Hungarian matching. The symbol "T" means track, and "D" means detection result. The detection results with green boxes are high-confidence detection results, and the detection results with yellow boxes are deleted high-confidence detection results during NMS. Our proposed TPA successfully maintains track ID through accurate association, even in challenging cases with severe overlapping and fast movements.

tracks and detection results are highly interdependent, as both represent the same underlying objects, and there is a correct answer for each assignment. Consequently, framing the data association problem in MOT as selecting the optimal match for each track would be more appropriate than solely minimizing global costs, particularly in scenarios involving occlusions, which critically impact tracking performance, as illustrated in Fig. 2.

On the other hand, many conventional methods [8, 14, 33, 56, 61] attempt to incorporate low-confidence detection results, which may represent partially visible objects, by employing a multi-stage matching cascade. However, in this cascade, it first matches high-confidence detection results with tracks and subsequently matches low-confidence detection results only with the remaining tracks. Thus, it leads to suboptimal utilization of low-confidence detection results that can be better options for certain tracks, lowering tracking performance in scenes with occlusion.

To address these issues, we propose a Track-Perspective-Based Association (TPA) method, which thoroughly utilizes all available detection results, including high-confidence and low-confidence detection results and even high-confidence detection results that are discarded during non-maximum suppression (NMS), while focusing on the local perspective of each track. Additionally, we introduce Track-Aware Initialization (TAI), which selectively initializes new tracks only with the most feasible detected boxes among the redundant candidates. More specifically, our TPA compares every pair of tracks and detection results simultaneously through a single joint association stage to improve the utilization of every candidate. Then, it iteratively associates tracks and detection results that present the minimum distance while prioritizing local matching precision, ensuring each track is matched with the most appropriate detection result in each track-perspective manner. Meanwhile, TAI excludes detection results that significantly overlap with active tracks and other more confident detection results from the set of track initialization candidates, preventing spurious tracks and contributing to a more reliable tracking process. Together, these methods form the basis of our novel tracker, Track-Focused Online Multi-Object Tracker (TrackTrack), designed to enhance robustness and accuracy in challenging tracking scenarios with improved data association and initialization process. Extensive experiments on MOT17 [12], MOT20 [11], and DanceTrack [48] demonstrate that our TrackTrack consistently outperforms state-of-the-art trackers, highlighting its robustness and effectiveness in diverse scenarios, as in Fig. 1.

The main contributions of our work are:

- We introduce a novel Track-Focused Online Multi-Object Tracker (TrackTrack) with two main components of Track-Perspective-Based Association (TPA) and Track-Aware Initialization (TAI).
- We propose TPA that prioritizes local matching accuracy, enhancing robustness during data association by ensuring each track is merged with the most suitable detection result while considering every pairwise distance with all available detection results through a single-stage joint association.
- We present TAI that leverages active tracks to prevent the creation of redundant tracks during the track initialization process and improves overall tracking stability.
- We demonstrate the effectiveness and robustness of our TrackTrack through extensive experiments on MOT17, MOT20, and DanceTrack. It consistently achieves state-of-the-art performance across all datasets compared to previous cutting-edge methods.

2. Related Work

2.1. Tracking-by-detection

The tracking-by-detection (TBD) paradigm, which separates the tracking task into two distinct subproblems of

object detection and data association, is the predominant approach in MOT. With the recent advancements in high-performance object detectors, such as Faster R-CNN [37], Feature Pyramid Networks (FPN) [28], YOLOX [18], and YOLOv7 [50], most methods focus on developing the association aspect, primarily through accurate affinity or distance estimation between detected objects. In this context, various techniques are proposed, including learnable neural solver [5] and probabilistic tracklet scoring [40]. More specifically, Xu *et al.* [54] introduce Spatial-Temporal Relation Networks that predict distances between detected objects by aggregating their appearance, location, and topology information. Similarly, Sun *et al.* [49] propose a Deep Affinity Network to estimate pairwise affinities between detected results of distinct frames in an end-to-end fashion.

Motion model-based trackers [22, 31, 47, 51], which are straightforward and highly effective MOT algorithms with state-of-the-art performance, are also actively suggested. They typically predict the subsequent locations of each previous track using a motion model, usually the Kalman Filter [23], and associate the tracks with current detection results using their pairwise distances and the Hungarian algorithm [34]. Building on this standard tracking pipeline, various techniques are presented for better data association. For example, some improve their Kalman filter-based motion model by considering confidence scores of detection results [13, 22] or the camera motion of each video [57]. Also, diverse kinds of distance metrics using Re-ID features [36, 51], pseudo-depth [29], moving direction [8], and detection confidence [56] are newly suggested.

2.2. Data Association

Data association, which targets accurate merging between detected objects and existing tracks, is an essential part of MOT systems. Beyond the limitations of early methods [2, 16, 35] that often require high computational costs or excessive prior knowledge, such as the number of targets, diverse data association schemes are proposed in contemporary multi-object trackers. For example, SORT [4] suggests simple matching between detected results and previous tracks based on their pairwise spatial distances, and DeepSORT [52] presents a matching cascade that offers priority to more recent tracks while treating newborn tracks through a separate stage. ByteTrack [61] introduces a multi-step association process that first matches high-confidence detection results to existing tracks and subsequently associates low-confidence detection results with remaining tracks to reduce lost cases of tracks. In LG-Track [33], Meng *et al.* further divide detected boxes into four categories based on their localization and classification confidence scores and then sequentially match them to tracks through four individual stages of association. Most recently, Huang *et al.* [21] propose decomposed associating strategies to refine its original association results by decomposing the conventional matching problem into several subproblems. However, most methods, including the aforementioned trackers, still perform data association through a global optimization and multi-stage matching scheme. In contrast, our tracker better solves the association task with our Track-Perspective-Based Association (TPA) approach, which prioritizes local matching precision and fully incorporates all feasible detection results through a single joint association scheme. Also, during TPA, detection results discarded by NMS are employed, which are neglected in most trackers. Simultaneously, our Track-Aware Initialization (TAI) enhances the track initialization process, contributing to overall tracking stability.

2.3. Considering Every Detection Results

Prior to the advent of deep learning-based methods, the strategy of leveraging all detection proposals in multi-object tracking is investigated in early methods. For instance, Wu and Nevatia [53] suggest a method for detecting multiple partially occluded humans in single images by employing a Bayesian combination of edgelet part detectors, which models individuals as assemblies of body parts. Leibe *et al.* [25] propose a coupled detection and trajectory estimation framework, addressing multi-object tracking as a joint optimization of detection and trajectory estimation. A generative model is also utilized to estimate probabilistic occupancy maps of the ground plane for a multi-people tracking system [15]. On the other hand, Breitenstein *et al.* [6, 7] introduce an online tracking-by-detection approach that utilizes a detector confidence particle filter and unreliable information sources for robust tracking. These methodologies underscore the efficacy of incorporating all detection results, complemented by partial likelihoods, to achieve robust multi-object tracking in challenging environments.

3. Method

In this section, we introduce our proposed tracker, Track-Focused Multi-Object Tracker (TrackTrack). Unlike most methods that treat the association between tracks and detection results as a global cost optimization problem, we focus on track and fully utilize all the redundant detection results, which are object candidates, by Track-Perspective-Based Association (TPA) and Track-Aware Initialization (TAI). In the following subsections, we describe the details of TPA and TAI, followed by the overall tracking pipeline.

3.1. Track-Perspective-Based Association (TPA)

Most TBD-based trackers adopt the Hungarian algorithm to associate the tracks and detection results, mainly utilizing high-confidence detection results. While there are some methods that utilize low-confidence detection results

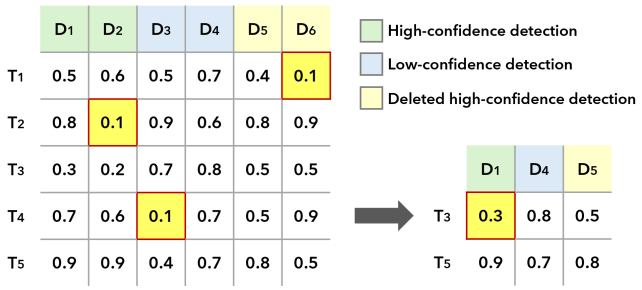


Figure 3. An overview of our TPA. First, we calculate a pairwise cost matrix between tracks and a complete set of detection results, which includes high-confidence, low-confidence, and high-confidence detection results omitted during NMS. We then identify the matchable pairs whose distances are minimum in both their corresponding rows and columns of the cost matrix, and we associate the discovered pairs and exclude their rows and columns from the matrix. Again, we search for the next set of pairs with minimum distances from the remaining matrix in the same manner and continue this association scheme until all remaining costs exceed the matching threshold.

[46, 61], these approaches are limited in that the use of low-confidence detection results is restricted by matching cascades, or they perform matching through global cost optimization, overlooking the interdependent relationship between tracks and detection results. We propose a Track-Perspective-Based Association (TPA) strategy, which maximizes the use of all detection results, which are object candidates, to provide optimal matching for each track.

Let $\mathcal{T} = \{T_1, T_2, \cdots, T_n\}$ represent the set of existing tracks and $\mathcal{D} = \{d_1, d_2, \cdots, d_m\}$ denote the set of detected results in a given frame. In our approach, we utilize three types of detection results to form the aggregated set $\mathcal{D}$: high-confidence detection results after NMS ($\mathcal{D}_{\text{high}}$), low-confidence detection results after NMS ($\mathcal{D}_{\text{low}}$), and highly confident but deleted detection results ($\mathcal{D}_{\text{del}}$) that are obtained by gathering results with high-confidence detection scores but deleted during NMS by $\mathcal{D}_{\text{high}}$.

The distance C_{ij} between the i^{th} track T_i and the j^{th} detection result d_j is calculated as

$$C_{ij} = \begin{cases} c(T_i, d_j), & d_j \in \mathcal{D}_{\text{high}} \\ c(T_i, d_j) + \tau_p, & d_j \in \mathcal{D}_{\text{low}} \\ c(T_i, d_j) + \tau_q, & d_j \in \mathcal{D}_{\text{del}} \end{cases} \quad (1)$$

where τ_p and τ_q are penalty terms for the low-confidence and discarded detection results, respectively. The distance function $c(T_i, d_j)$ between track T_i and detection result d_j is composed with HMIoU, cosine, confidence, and angular distances similar to the previous work [56].

With the computed cost matrix $\mathbf{C} = (C_{ij})$, we associate the tracks with the proper detection results in a track-perspective manner by selecting the best match for each track among the multiple detection results. More specifically, for the set of tracks $\mathcal{T}$ and detection results $\mathcal{D}$, our TPA algorithm iteratively selects appropriate pairs of (T_i, d_j) where each d_j shows the minimum cost among all detection results for a given track T_i, and each T_i shows the minimum cost among all tracks also for a given detection result d_j. The matched pairs $\mathcal{M}$ are identified as follows:

$$\mathcal{M} = \{(T_i, d_j) | T_i = \underset{T_l \in \mathcal{T}}{\operatorname{argmin}} C_{lj}, d_j = \underset{d_k \in \mathcal{D}}{\operatorname{argmin}} C_{ik}, C_{ij} < \tau_m\}, \quad (2)$$

where τ_m is a matching threshold. After determining the set of matched pairs $\mathcal{M}$, we remove the matched tracks and detection results from the original sets $\mathcal{T}$ and $\mathcal{D}$ to form new sets $\mathcal{T}'$ and $\mathcal{D}'$:

$$\mathcal{T}' = \mathcal{T} \setminus \{T_i \mid (T_i, d) \in \mathcal{M}\}, \quad \text{and} \quad (3)$$

$$\mathcal{D}' = \mathcal{D} \setminus \{d_j \mid (T, d_j) \in \mathcal{M}\}. \quad (4)$$

We then iteratively apply the same association procedure using the updated sets $\mathcal{T}'$ and $\mathcal{D}'$ until there are no matchable pairs with distance values lower than the threshold. Also, at each iteration, the matching threshold τ_m is decreased by the reduction term r to gradually tighten the condition of proper matching. The graphical illustration of our TPA is depicted in Fig. 3.

This minimum distance-based matching process can ensure that each track is associated with the most proper detection result unless the target objects of the tracks are totally occluded or disappeared and cannot be detected. By focusing on each track in this manner and utilizing all the probable detection results comprehensively, our TPA method not only improves the accuracy of track associations but also enhances the robustness of the tracking process, particularly in scenarios where traditional methods may struggle due to high detection noise or occlusions.

3.2. Track-Aware Initialization

Traditionally, in TBD-based trackers, high-confidence detection results that are not associated with existing tracks and exceed a certain confidence score threshold are initialized as new tracks to handle newly appearing objects in videos. In this work, we propose a Track-Aware Initialization (TAI) strategy that overcomes the limitations of this conventional approach by leveraging information from the active tracks, as presented in Fig. 4. Specifically, we treat the final locations of matched tracks during TPA as undeletable anchors or detection results with a confidence score of 1. We then apply non-maximum suppression to a combined set of these predefined anchors and the unmatched high-confidence detection results that are not associated during TPA. Finally, the remaining detection results

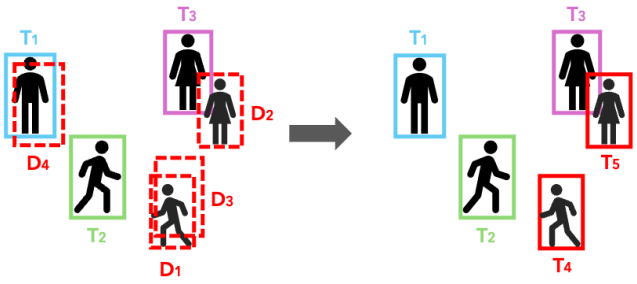


Figure 4. An overview of our Track-Aware Initialization. First, we set the last known positions of the matched tracks as undeletable anchors or detection results with confidence scores of 1. Then, we perform non-maximum suppression on the remaining high-confidence detection results after the association together with the predefined anchors so that the detection results that have significant overlaps with any tracks or more confident detection results can be discarded from initialization candidates. Only the left detection results after these procedures become our new tracks.

after the NMS process are initialized as our new tracks. This track-aware procedure results in more accurate track initialization by reducing the number of spurious tracks, as it discards detection results that significantly overlap with current active tracks and other confident detection results from initialization candidates. Consequently, our TAI enhances overall tracking stability and reliability, particularly in challenging scenarios with high detection noise and occlusions, by improving the quality of track initialization. Note that our TAI can prioritize the certainty of object existence by lowering the IoU threshold during NMS and can also adapt to denser crowds by increasing the threshold. It can be flexibly adjusted to be suited for each application, as algorithms need to be adjustable to some degree.

3.3. Tracking Pipeline

In our TrackTrack, we first detect objects of interest through a detection model and predict the current locations of the existing tracks by using the NSA Kalman Filter [23] as a motion model. The tracks and a total set of detection results, which contains all the high-confidence, low-confidence, and discarded high-confidence detection results, are compared, and we associate each track with the proper match through TPA. The remaining high-confidence detection results are then associated with recently initialized tracks that have not yet accumulated sufficient temporal information after initialization, again using an assignment algorithm similar to the TPA strategy. Specifically, we define tracks with less than three frames as the recently initialized tracks that are often noises or false positives, such as reflection. This separate assignment step is for blocking potential disruption of the unreliable one or two-frame-length tracks during the association with the confirmed tracks (tracked three frames or more). Also, it prioritizes the confirmed tracks to associate with the detection results and prevents noisy short-tracking results. Note that our methods do not need additional training to improve the association stage of the tracker.

4. Experiments

4.1. Datasets

In this study, we employed the MOT17 [12], MOT20 [11], and DanceTrack [48] datasets for our experiments. MOT17 features diverse crowd scenarios with both static and moving cameras. MOT20 focuses on more complex environments with high-density crowds, presenting additional challenges for tracking algorithms. DanceTrack addresses the intricacies of tracking objects that appear similar and exhibit nonlinear motion, particularly in the context of group dance. It emphasizes the importance of motion analysis over visual cues, setting it apart from MOT17 and MOT20.

4.2. Metrics

We evaluated the tracking performance using various metrics, including CLEAR [3], IDF1 [39], HOTA [30], and their relatives. Representatively, MOTA, which is included in the CLEAR metric group, measures three types of errors made by the tracker: false positives, missed detections, and identity switches. IDF1 focuses on identification accuracy, assessing how well the tracker maintains correct identity labels across frames. HOTA is known as the most balanced and comprehensive score that simultaneously evaluates the accuracy of detection, association, and localization in a unified measurement while better aligning with human visual evaluation. In our experiments, we considered HOTA to be the primary metric for assessing the overall effectiveness of multi-object trackers following the previous works [21, 51, 56, 57].

4.3. Implementation Details

As an object detector, we adopted the YOLOX-x model [18], similar to the previous works [8, 56, 61]. From a COCO [27] pre-trained weight, the model was trained for 80 epochs with each target dataset and additional CrowdHuman [43] and Widerperson [59] datasets while following the default training configuration of the prior work [61]. For our feature extractor, we utilized the SBS50 model of the FastReID [19] framework similar to the earlier methods [1, 21, 31]. Each extractor was trained through the basic settings configured by the framework with its respective target dataset. During the tracking, we used the detection results after applying NMS with an IoU threshold of 0.80. $\mathcal{D}_{\text{del}}$ was derived by removing $\mathcal{D}_{\text{high}}$ and low-confidence detection results from the set of detected results obtained after applying NMS with an IoU threshold of 0.95. NSA Kalman Filter [13], camera motion compensation [1], and post-processing techniques [14, 58] were utilized following the previous

MOT17				
Tracker	HOTA↑	IDF1↑	MOTA↑	AssA↑
Offline-Based				
SUSHI [9]	66.5	83.1	81.1	67.8
CoNo-Link [17]	67.1	83.7	82.7	67.8
Online-Based				
Hybrid-SORT-ReID [56]	64.0	78.7	79.9	63.5
FineTrack [36]	64.3	79.5	80.0	64.5
StongSORT++ [14]	64.4	79.5	79.6	64.4
Deep OC-SORT [31]	64.9	80.6	79.4	65.9
DeconfuseTrack [21]	64.9	80.6	80.4	65.1
SparseTrack [29]	65.1	80.1	81.0	65.1
DATrack [32]	65.4	80.4	81.4	65.4
CMTrack [44]	65.5	81.5	80.7	66.1
AdapTrack [45]	65.7	82.3	79.9	66.9
UCMCTrack+ [57]	65.7	81.0	80.6	66.4
PIA [47]	66.0	81.1	**82.2**	65.8
ImprAsso [46]	66.4	82.1	**82.2**	66.6
TrackTrack (Ours)	**67.1**	**83.1**	81.8	**68.2**

Table 1. Comparison results on the MOT17 test set with state-of-the-art methods.

MOT20				
Tracker	HOTA↑	IDF1↑	MOTA↑	AssA↑
Offline-Based				
SUSHI [9]	64.3	79.8	74.3	67.5
CoNo-Link [17]	65.9	81.8	77.5	68.0
Online-Based				
StrongSORT++ [14]	62.6	77.0	73.8	64.0
UCMCTrack+ [57]	62.8	77.4	75.6	63.5
DeconfuseTrack [21]	63.3	77.6	78.1	62.7
DATrack [32]	63.4	77.4	77.8	62.9
SparseTrack [29]	63.4	77.3	78.2	62.8
FineTrack [36]	63.6	79.0	77.9	63.8
Deep OC-SORT [31]	63.9	79.2	75.6	65.7
Hybrid-SORT-ReID [56]	63.9	78.4	76.7	64.5
ImprAsso [46]	64.6	78.8	**78.6**	64.6
PIA [47]	64.7	79.0	78.5	64.9
CMTrack [44]	64.8	79.9	76.2	66.7
AdapTrack [45]	65.0	80.7	75.0	**67.8**
TrackTrack (Ours)	**65.7**	**80.9**	78.0	67.3

Table 2. Comparison results on the MOT20 test set with state-of-the-art methods.

methods [1, 8, 14, 44, 45], and we re-tracked objects that were missed for up to two seconds. All the following experiments were performed with an NVIDIA GeForce RTX 3090 GPU and Intel(R) Core(TM) i7-11700K @ 3.60GHz CPU. Detailed hyper-parameter settings and their ablative studies are depicted in the supplementary material.

4.4. Experimental Results

In this section, we introduce the quantitative results of our tracker on MOT17, MOT20, and DanceTrack. As shown in Tables 1, 2, and 3, our TrackTrack achieves superior tracking performance compared to other state-of-the-art trackers in all the MOT17, MOT20, and DanceTrack datasets. In the MOT17 test set, TrackTrack exhibits the highest HOTA score of 67.1, surpassing all other online-based trackers and showing highly comparable or even better performance than the offline-based trackers, indicating a most accurate and robust tracking performance. Furthermore, our method also excels in IDF1 and AssA metrics with scores of 83.1 and 68.2, respectively. It suggests the particular effectiveness of our tracker in association and maintaining identity consistency through the novel association and initialization strategies. In the case of the MOT20 test set, our tracker continues to outperform and achieves the highest scores in both HOTA and IDF1 among online trackers. Also, it shows comparable and better performance than the state-of-the-art offline-based methods. These results demonstrate its strong tracking capability under highly crowded conditions. Finally, as in Table 3, our approach highly surpasses all other

DanceTrack				
Tracker	HOTA↑	IDF1↑	MOTA↑	AssA↑
Offline-Based				
SUSHI [9]	63.3	63.4	88.7	50.1
CoNo-Link [17]	63.8	64.1	89.7	50.7
Online-Based				
FineTrack [36]	52.7	59.8	89.9	38.5
OC-SORT [8]	55.1	54.6	92.0	38.3
SparseTrack [29]	55.5	58.3	91.3	39.1
StrongSORT++ [14]	55.6	55.2	91.1	38.6
GHOST [42]	56.7	57.7	91.3	39.8
CBIoU [55]	60.6	61.6	91.6	45.4
Deep OC-SORT [31]	61.3	61.5	92.3	45.8
CMTrack [44]	61.8	63.3	92.5	46.4
UCMCTrack+ [57]	63.6	65.0	88.9	51.3
Hybrid-SORT-ReID [56]	65.7	67.4	91.8	-
TrackTrack (Ours)	**66.5**	**67.8**	**93.6**	**52.9**

Table 3. Comparison results on the DanceTrack test set with state-of-the-art methods.

methods, including offline-based techniques, on the DanceTrack test set at every metric, underscoring its robustness against nonlinear motion and similar appearance of target objects. One more notable strength of our proposed solution is its consistent outperformance across all evaluated datasets, highlighting the robustness and adaptability of our proposed tracker in a wide range of tracking scenarios. In contrast, instability in tracking performance across datasets

		MOT17-val		DanceTrack-val	
TPA	TAI	HOTA↑	AssA↑	HOTA↑	AssA↑
		67.3	69.6	61.9	47.7
✓		68.5	72.0	62.9	49.1
	✓	67.4	69.6	62.6	48.7
✓	✓	**69.1**	**72.7**	**63.3**	**49.7**

Table 4. An ablative study for our proposed strategies. TPA and TAI denote Track-Perspective-Based Association and Track-Aware Initialization, respectively.

can be readily observed in other state-of-the-art methods.

4.5. Ablation Studies

In this section, we performed ablation studies to comprehensively examine our tracker and the proposed strategies. First, we analyzed the impact of applying each TPA and TAI in the tracker. Then, we investigated the influence of the proposed assignment algorithm, joint association scheme, and usage of the deleted detection results $\mathcal{D}_{\text{del}}$ during TPA.

4.5.1. Component Ablation

Table 4 shows the performance contributions of each proposed strategy, specifically Track-Perspective-Based Association (TPA) and Track-Aware Initialization (TAI), with the validation sets from MOT17 and DanceTrack. As in the table, the inclusion of each strategy remarkably improves both HOTA and AssA (Association Accuracy) scores. More specifically, the baseline tracker, which follows the typical Hungarian algorithm-based multi-stage association process and the threshold-based track initialization scheme of the prior works [8, 56, 61], achieves a HOTA score of 67.3 on MOT17 and a HOTA score of 61.9 on DanceTrack. However, after introducing TPA on the baseline, the HOTA scores for each dataset are improved to 68.5 and 62.9, respectively, and after presenting TAI, they are enhanced to 67.4 and 62.6, respectively. Furthermore, the combined application of TPA and TAI resulted in the highest performance, with the HOTA scores reaching 69.1 and 63.3. These results demonstrate that both TPA and TAI positively contribute to the tracking performance, and each presented strategy offers distinct advantages in enhancing the accuracy and robustness of the tracker.

4.5.2. Assignment Method

In Table 5, we can confirm the performance comparison of our proposed assignment method against the traditional Hungarian algorithm using the MOT17 and DanceTrack validation sets. The result clearly demonstrates the superiority of our method. By applying our assignment algorithm, 1.0%p and 2.2%p of each HOTA and AssA score are increased on the MOT17 validation set. Moreover, the improvement is even more pronounced on the DanceTrack validation set, where our method leads to 1.7%p and 2.6%p of enhancement on each metric compared to the Hungarian algorithm. These findings highlight the effectiveness of our assignment strategy that considers local matching precision when the total set of detection results are compared to earlier tracks during association.

	MOT17-val		DanceTrack-val	
Assignment	HOTA↑	AssA↑	HOTA↑	AssA↑
Hungarian	68.1	70.5	61.6	47.1
Ours	**69.1**	**72.7**	**63.3**	**49.7**

Table 5. Comparisons between using the Hungarian algorithm and our assignment method that considers local matching precision.

	MOT17-val		DanceTrack-val	
Association	HOTA↑	AssA↑	HOTA↑	AssA↑
Multi-Stage	68.3	71.5	63.2	**49.8**
Joint (Ours)	**69.1**	**72.7**	**63.3**	49.7

Table 6. Comparisons between the multi-stage cascade association of each $\mathcal{D}_{\text{high}}$, $\mathcal{D}_{\text{low}}$, and $\mathcal{D}_{\text{del}}$ similar to the previous works [8, 56, 61] and joint association of all detection results as we proposed.

	MOT17-val		DanceTrack-val	
Use $\mathcal{D}_{\text{del}}$	HOTA↑	AssA↑	HOTA↑	AssA↑
✗	68.6	71.9	62.1	47.6
✓	**69.1**	**72.7**	**63.3**	**49.7**

Table 7. An ablative study for the influence of utilizing the deleted detection results $\mathcal{D}_{\text{del}}$ in our tracking process.

4.5.3. Association Stage

An ablative result to demonstrate the effectiveness of our joint association scheme within TPA is presented in Table 6. The result indicates that the joint scheme outperforms the multi-stage association strategy, which matches each detection result $\mathcal{D}_{\text{high}}$, $\mathcal{D}_{\text{low}}$, and $\mathcal{D}_{\text{del}}$ through separate stages similar to the most existing works [8, 56, 61]. For example, on the MOT17 validation set, the joint approach achieves a HOTA score of 69.1 and an AssA score of 72.7, compared to 68.3 and 71.5 of the multi-stage scheme. These results demonstrate that jointly matching all detection results with tracks, as proposed in our TrackTrack, leads to more accurate tracking results compared to the conventional multi-stage association process.

4.5.4. Using Deleted Detection Results

In Table 7, an evaluation result for the impact of utilizing the deleted detection results $\mathcal{D}_{\text{del}}$ is shown. The result reveals that incorporating $\mathcal{D}_{\text{del}}$ enhances tracking performance in all cases. Notably, HOTA and AssA scores are improved from 68.6 to 69.1 and from 71.9 to 72.7, respectively, on MOT17

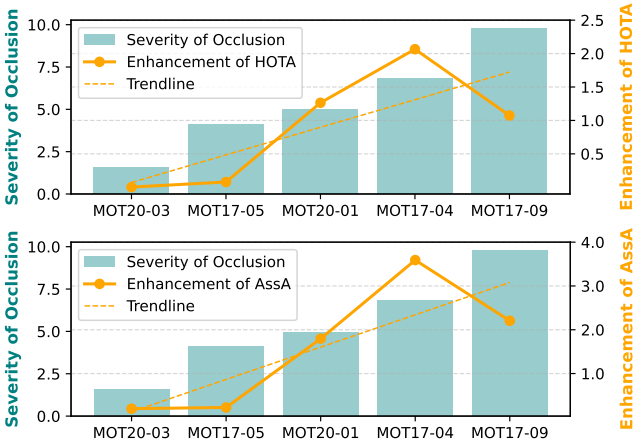


Figure 5. Performance improvement in our TrackTrack compared to the baseline based on the severity of occlusion. The enhancement of HOTA and AssA tends to increase with the occlusion severity, indicating that TrackTrack becomes increasingly effective in more challenging conditions involving severe occlusions.

and increased from 62.1 to 63.3 and from 47.6 to 49.7, respectively, on DanceTrack after including $\mathcal{D}_{\text{del}}$ in our tracking procedure. These findings suggest that leveraging the deleted detection results contributes to better performance by proposing broader matchable candidates for tracks.

4.6. Effectiveness Under Occlusion Conditions

The performance enhancement by our TrackTrack compared to the baseline in various scenarios with different levels of occlusion is shown in Fig. 5. The severity of occlusion is calculated as the ratio of the total number of severe overlaps (IoU $>$ 0.6) to the total number of overlaps (IoU $>$ 0.1) for each video. We can confirm that TrackTrack outperforms the baseline method more effectively as the severity of occlusion increases in both HOTA and AssA metrics that measure overall tracking performance and quality of association, respectively. This highlights the robustness of our strategies in scenarios where accurate tracking is most challenging, ultimately reducing tracking failure and aiding precise association in the presence of substantial occlusions.

4.7. Computational Cost

The analysis of the computational costs of our proposed schemes in TrackTrack is presented in Table 8. Specifically, when using the multi-stage association strategy [8, 56, 61] with the Hungarian algorithm [34], the system achieves 155.44 FPS on MOT17 and 394.84 FPS on DanceTrack. Then, the speeds are improved to 157.96 FPS and 400.13 FPS on MOT17 and DanceTrack, respectively, after transitioning to our iterative assignment strategy while still utilizing the multi-stage scheme. Finally, the most substantial performance gain is observed when both our assignment method and joint association strategy are applied, achieving the highest frame rates of 161.52 FPS on MOT17 and 408.37 FPS on DanceTrack, which are 3.9% and 3.4% improved results from the beginning baseline. These results indicate that our tracker not only enhances tracking accuracy but also reduces computation time, making it a more efficient alternative for multi-object tracking applications. Note that, for more precise comparison, we averaged the results from five times of execution for each case and excluded the post-processing techniques [14, 58] during the measurements. Also, as our assignment method is implemented in straightforward Python code for convenience, it has significant room for more computational optimization, while the implementation of the Hungarian algorithm is already highly optimized through the predefined library [1].

		MOT17-val	**DanceTrack-val**
Assignment	Association	FPS↑	FPS↑
Hungarian	Multi-Stage	155.44	394.84
Ours	Multi-Stage	157.96	400.13
Ours	Joint (Ours)	**161.52**	**408.37**

Table 8. Comparisons of computational costs between adopting the Hungarian algorithm and our proposed assignment method and using the multi-stage association strategy and our joint scheme in TrackTrack. Each value represents the frame per second (FPS) of the corresponding case.

5. Conclusion

In this work, we have presented TrackTrack, a novel approach to online multi-object tracking that focuses on improving data association and track initialization through our Track-Perspective-Based Association (TPA) and Track-Aware Initialization (TAI) techniques. By addressing the limitations of the traditional tracking-by-detection framework, particularly the drawbacks of global cost minimization and the suboptimal use of low-confidence detection results, TrackTrack effectively enhances tracking robustness, particularly in challenging scenarios involving occlusions and dense object interactions. TPA ensures that each track is matched with the most appropriate detection result by iteratively comparing detection candidates from a local perspective, while TAI prevents the creation of spurious tracks by carefully selecting reliable detection results for new track initialization. Extensive experimental results demonstrate that TrackTrack consistently outperforms state-of-the-art trackers on MOT17, MOT20, and DanceTrack datasets, showing superior accuracy and robustness.

[1] https://pypi.org/project/lapjv/

Acknowledgments

This research was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea(NRF) funded by the Ministry of Science & ICT and Korea Customs Service (NRF-2021M3I1A1097906).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv:2206.14651*, 2022. 1, 5, 6

[2] Yaakov Bar-Shalom, Thomas E. Fortmann, and Peter G. Cable. Tracking and data association. *The Journal of the Acoustical Society of America*, 87(2):918–919, 1990. 3

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP JIVP*, pages 1–10, 2008. 5

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468, 2016. 3

[5] Guillem Braso and Laura Leal-Taixe. Learning a neural solver for multiple object tracking. In *CVPR*, 2020. 3

[6] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, pages 1515–1522, 2009. 3

[7] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE TPAMI*, 33(9):1820–1833, 2011. 3

[8] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 1, 2, 3, 5, 6, 7, 8

[9] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *CVPR*, pages 22877–22887, 2023. 6

[10] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, pages 9921–9931, 2023. 1

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi-object tracking in crowded scenes. *arXiv:2003.09003*, 2020. 2, 5

[12] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 129:845–881, 2021. 2, 5

[13] Yunhao Du, Junfeng Wan, Yanyun Zhao, Binyu Zhang, Zhihang Tong, and Junhao Dong. Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021. In *ICCV*, 2021. 3, 5

[14] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE TMM*, 25:8725–8737, 2023. 1, 2, 5, 6, 8

[15] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE TPAMI*, 30(2):267–282, 2008. 3

[16] T. Fortmann, Y. Bar-Shalom, and M. Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 8(3):173–184, 1983. 3

[17] Yan Gao, Haojun Xu, Jie Li, Nannan Wang, and Xinbo Gao. Multi-scene generalized trajectory global graph solver with composite nodes for multiple object tracking. *AAAI*, 38(3): 1842–1850, 2024. 6

[18] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021. 1, 3, 5

[19] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv:2006.02631*, 2020. 5

[20] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *ICCV*, pages 5390–5399, 2019. 1

[21] Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao Wei. Deconfusetrack: Dealing with confusion for multi-object tracking. In *CVPR*, pages 19290–19299, 2024. 3, 5, 6

[22] Hyeonchul Jung, Seokjun Kang, Takgen Kim, and HyeongKi Kim. Conftrack: Kalman filter-based multi-person tracking by utilizing confidence score of detection box. In *WACV*, pages 6583–6592, 2024. 3

[23] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960. 3, 5

[24] Long Lan, Xinchao Wang, Gang Hua, and Dacheng Tao. Semi-online multi-people tracking by re-identification. *IJCV*, 128:1937–1955, 2020. 1

[25] Bastian Leibe, Konrad Schindler, and Luc Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, pages 1–8, 2007. 3

[26] Rui Li, Baopeng Zhang, Jun Liu, Wei Liu, Jian Zhao, and Zhu Teng. Heterogeneous diversity driven active learning for multi-object tracking. In *ICCV*, pages 9932–9941, 2023. 1

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, page 740–755, 2014. 5

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[29] Zelin Liu, Xinggang Wang, Cheng Wang, Wenyu Liu, and Xiang Bai. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *arXiv:2306.05238*, 2023. 1, 3, 6

[30] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, pages 548–578, 2020. 5

[31] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *ICIP*, 2023. 1, 3, 5, 6

[32] Ting Meng, Chunyun Fu, Mingguang Huang, Xiyang Wang, Jiawei He, Tao Huang, and Wankai Shi. Localization-guided track: A deep association multi-object tracking framework based on localization confidence of detections. *arXiv:2309.09765*, 2023. 6

[33] Ting Meng, Chunyun Fu, Mingguang Huang, Xiyang Wang, Jiawei He, Tao Huang, and Wankai Shi. Localization-guided track: A deep association multi-object tracking framework based on localization confidence of detections. *arXiv:2309.09765*, 2023. 1, 2, 3

[34] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5:32–38, 1957. 1, 3, 8

[35] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979. 3

[36] Hao Ren, Shoudong Han, Huilin Ding, Ziwen Zhang, Hongwei Wang, and Faquan Wang. Focus on details: Online multi-object tracking with diverse fine-grained representation. In *CVPR*, pages 11289–11298, 2023. 1, 3, 6

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39:1137–1149, 2017. 3

[38] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *CVPR*, pages 6036–6046, 2018. 1

[39] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, pages 17–35, 2016. 5

[40] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *CVPR*, pages 14329–14339, 2021. 3

[41] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *CVPR*, pages 6951–6960, 2017. 1

[42] Jenny Seidenschwarz, Guillem Brasó, Víctor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *CVPR*, pages 13813–13823, 2023. 6

[43] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint*, arXiv:1805.00123, 2018. 5

[44] Kyujin Shim, Jubi Hwang, Kangwook Ko, and Changick Kim. A confidence-aware matching strategy for generalized multi-object tracking. In *ICIP*, 2024. 6

[45] Kyujin Shim, Kangwook Ko, Jubi Hwang, and Changick Kim. Adaptrack: Adaptive thresholding-based matching for multi-object tracking. In *ICIP*, 2024. 6

[46] Daniel Stadler and Jürgen Beyerer. An improved association pipeline for multi-person tracking. In *CVPRW*, pages 3170–3179, 2023. 4, 6

[47] Daniel Stadler and Jürgen Beyerer. Past information aggregation for multi-person tracking. In *ICIP*, pages 321–325, 2023. 3, 6

[48] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20993–21002, 2022. 2, 5

[49] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE TPAMI*, 43:104–119, 2021. 3

[50] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pages 7464–7475, 2023. 1, 3

[51] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung-Hin So, and Xin Li. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. *AAAI*, 38(6):5740–5748, 2024. 3, 5

[52] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649, 2017. 3

[53] Bo Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, pages 90–97, 2005. 3

[54] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *ICCV*, pages 3988–3998, 2019. 3

[55] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *WACV*, pages 4799–4808, 2023. 6

[56] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. *AAAI*, 38(7):6504–6512, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[57] Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu, and Wei Hao. Ucmctrack: Multi-object tracking with uniform camera motion compensation. *AAAI*, 38(7): 6702–6710, 2024. 1, 3, 5, 6

[58] Kai Zeng, Yujie You, Tao Shen, Qingwang Wang, Zhimin Tao, Zhifeng Wang, and Quanjun Liu. Nct:noise-control multi-object tracking. *Complex & Intelligent Systems*, 9(4): 4331–4347, 2023. 5, 8

[59] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE TMM*, 22(2): 380–393, 2020. 5

[60] Wenwei Zhang, Hui Zhou, Shuyang Sun, Zhe Wang, Jianping Shi, and Chen Change Loy. Robust multi-modality multi-object tracking. In *ICCV*, pages 2365–2374, 2019. 1

[61] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 1, 2, 3, 4, 5, 7, 8