

History-Aware Transformation of ReID Features for Multiple Object Tracking

Ruopeng Gao¹, Yuyao Wang¹, Chunxu Liu¹, and Limin Wang^{1,2,✉}

¹ Nanjing University, Nanjing, China

² Shanghai AI Laboratory, Shanghai, China

<https://github.com/MCG-NJU/HATReID-MOT>

{ruopenggao, wayfareryy}@gmail.com chunxu.liu@smail.nju.edu.cn

lmwang@nju.edu.cn

Abstract. In Multiple Object Tracking (MOT), Re-identification (ReID) features are widely employed as a powerful cue for object association. However, they are often wielded as a one-size-fits-all hammer, applied uniformly across all videos through simple similarity metrics. We argue that this overlooks a fundamental truth: MOT is not a general retrieval problem, but a context-specific task of discriminating targets within a single video. To this end, we advocate for the adjustment of visual features based on the context specific to each video sequence for better adaptation. In this paper, we propose a history-aware feature transformation method that dynamically crafts a more discriminative subspace tailored to each video’s unique sample distribution. Specifically, we treat the historical features of established trajectories as context and employ a tailored Fisher Linear Discriminant (FLD) to project the raw ReID features into a sequence-specific representation space. Extensive experiments demonstrate that our training-free method dramatically enhances the discriminative power of features from diverse ReID backbones, resulting in marked and consistent gains in tracking accuracy. Our findings provide compelling evidence that MOT inherently favors context-specific representation over the direct application of generic ReID features. We hope our work inspires the community to move beyond the naive application of ReID features and towards a deeper exploration of their purposeful customization for MOT.

Keywords: Multi-object tracking · MOT · Tracking

1 Introduction

Multiple Object Tracking (MOT) is a fundamental computer vision task that aims to detect objects and maintain their identities across video frames. Its primary goal is to generate a distinct trajectory for each target by associating its corresponding detections over time [30]. As a critical component for understanding dynamic scenes, MOT serves as an essential prerequisite for a wide range

✉ : Corresponding author.

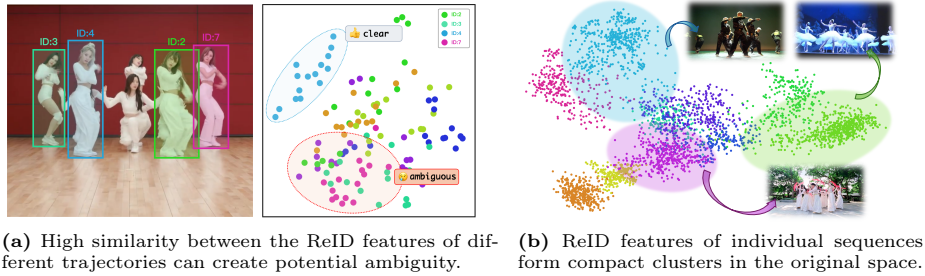


Fig. 1: Visualization of ReID features in the original representation space [33].

of downstream applications, such as autonomous driving [21], human behavior analysis [9], trajectory forecasting [23], and public surveillance.

The tracking-by-detection paradigm [3, 5, 60] has long been the dominant and most widely adopted approach in the field of multiple object tracking. According to the task definition, it decouples the complex tracking problem into two sequential subtasks: first, an object detector localizes all targets within each frame, and second, an association algorithm links these detections across frames to form individual trajectories. As the former step is well-addressed by powerful detectors [17, 48], the crux of this paradigm lies in the association stage. To solve this association problem, most methods [5, 12, 61] model existing trajectories with discriminative cues and then allocate identities by minimizing the matching cost.

Given that distinct targets often exhibit unique visual characteristics, appearance has emerged as a powerful and prevalent discriminative feature for trajectory modeling. In practice, visual features are typically extracted using off-the-shelf Re-Identification (ReID) models [33], and a cost matrix is then formulated by the cosine distance. Despite its demonstrated success [1, 53, 57, 61], a latent contradiction persists within this paradigm. According to the definition, the goal of a general ReID model is to learn a universal feature representation capable of distinguishing any given identity from a large, open set. In contrast, the challenge within MOT is to discriminate only between the limited set of targets appearing in a specific video, which constitutes a more nuanced, expert-level requirement. As illustrated in Fig. 1a, targets within the same video sequence always exhibit a high degree of similarity, making them difficult to distinguish in a generic, globally-trained feature space [33]. Furthermore, this intra-sequence similarity causes their representations to cluster within a confined subspace of the original space, leading to redundancy, as shown in Fig. 1b. Based on the foregoing observations, a natural question arises: *can we seek a specialized representation subspace for MOT, one that is more focused on distinguishing identities within the constrained set of a given sequence?*

In this paper, we first confirm the significant influence of representation discriminability on tracking performance. Accordingly, we posit that an ideal representation space should be structured for each sequence to minimize intra-

trajectory distances while maximizing inter-trajectory distances. Such a configuration would effectively increase the separability between positive and negative pairs, thereby enhancing the tracking ability. Coincidentally, this principle is conceptually analogous to the objective of Fisher Linear Discriminant (FLD) [14], where trajectories are viewed as different categories. Since tracking is an online, continuous inference process, the established historical context can be regarded as a dynamic approximation of the data distribution for the current sequence. Building on these discussions, in practice, we feed the existing trajectories as conditional input into the FLD algorithm. By computing the closed-form solution, we derive a projection matrix that maps the original features onto a more discriminative subspace tailored for separating different trajectories. The experiment results reveal that this simple yet effective feature transformation substantially enhances discriminability and boosts overall tracking performance. Nevertheless, we revisit this process and argue that MOT possesses some task-specific demands that should not be overlooked. Firstly, since a target’s features gradually change over time in online tracking, we use a temporally weighted average to construct the trajectory’s center, rather than the naïve averaging. This makes the resulting representation better suited for the similarity assessment required at the current timestep. Secondly, historical trajectories are not always reliable due to occlusions and tracking errors. Moreover, the tracker needs to handle newborn objects, which are not considered in the transformed space. These challenges underscore the importance of retaining the original feature space with its strong robustness and generalization capabilities. To this end, we combine the similarity scores from both the general and specialized representations, thereby leveraging the complementary strengths of each.

To clearly validate the impact of our ReID feature transformation, our experiments are primarily conducted on trackers that rely solely on appearance [27,33]. This approach minimizes the complex designs and potential interference introduced by other tracking cues [3,52]. In practice, we build a ReID-based tracker upon the most widely-used ReID model [33] in the MOT community [35,57] and validate the effectiveness of our components. Relying solely on the ReID cue, our method achieves significant performance improvements. Remarkably, in some scenarios [10], our algorithm substantially outperforms methods that combine multiple clues [1,10,35], establishing a new state-of-the-art result. This finding strongly indicates that the full potential of appearance information has been underestimated in past research. We also confirm the generalization capability of our proposed method by applying it to [27] with various visual encoders [19,25,29,63], observing stable performance boosts across every case. Additionally, we conduct experiments on several hybrid-based methods [5,57]. The results demonstrate that our approach can be seamlessly integrated into these advanced trackers, achieving state-of-the-art performance.

To sum up, our main contributions include:

- Following our analysis in Sec. 3.2, we equip Fisher Linear Discriminant with historical tracklet supervision to transform ReID features, enhancing their discriminability.

- To address the practical needs of MOT task, we propose two customized components, *temporally-weighted trajectory centroid* (Sec. 4.2) and *knowledge integration* (Sec. 4.3), which further improve our tracking performance.
- To prove the effectiveness of our method, we conduct extensive experiments on ReID-based methods, demonstrating consistent performance gains across diverse scenarios (Tab. 1, Tab. 2 and Tab. 3). We also validate its versatility by seamlessly integrating it into hybrid-based trackers, pushing their state-of-the-art performance even further.

2 Related Work

Tracking-by-Detection methods decouple the multiple object tracking (MOT) task into two sub-tasks: object detection and data association. While a minority of studies [24] have explored customized detection methods, the vast majority of research [31, 37, 45, 60, 64] has focused on the design of the target association algorithm. In this process, researchers model trajectories and measure affinities based on diverse cues. For visual appearance, most methods [1, 35, 36, 57] directly utilize off-the-shelf ReID models [33] to extract features. While some approaches [41, 51, 61] employ custom-designed extractors, they still adhere to the fundamental principles and supervision methods of traditional ReID methods [33, 44]. Regarding location information, the most classic method [3, 60] is to use the Kalman filter [52] for linear estimation of the motion. To handle non-linear dynamics [10, 47] and other complex cases, recent methods have introduced many tailored rules [5, 13, 57, 58] or adopted learnable modules for motion prediction [12, 22, 34, 35, 42, 54]. Many approaches [13, 35, 36, 53, 57] also fuse the two aforementioned cues together to fully leverage their respective advantages. Furthermore, some other methods introduce even more information modalities, such as Bird’s-Eye-View (BEV) perspectives [12] and depth information [1, 37, 50]. Most relevant to our work are several studies that aim to customize the ReID branch of the MOT task: [20] seeks to mitigate the mismatch between its global temporal training and local temporal inference, [8] performs group-wise similarity calculation to address the long-tail distribution problem, [28] helps newborn targets acquire more robust representations, [6] sharpens the distinction in similarity, [49] employs a condition-aware strategy to enhance the matching and updating procedures for individual trajectories, and [18] utilizes segmentation guidance to improve object-level ReID feature extraction. Although these methods aim to improve the discriminative capability of visual representations, they overlook the fact that tracking only needs to distinguish objects within a finite set, rather than from a potentially unbounded identity space as in conventional ReID. This observation serves as the key motivation of our work, as it makes it possible to use tracking histories to guide the optimization of the feature space. To the best of our knowledge, this is the first work to provide an intuitive explanation, a systematic analysis, and a practical solution to this issue.

End-to-End MOT models are emerging forces, bypassing hand-crafted algorithms [5, 60] to formulate multi-object tracking in an end-to-end man-

ner [15, 59]. A typical form is to expand DETR [7, 65] into MOT tasks, representing different trajectories through the propagation of track queries [38, 59]. Follow-up methods incorporated temporal information [4, 16, 46] and mitigated the imbalance of supervision signals [55, 62], leading to better tracking performance. Nevertheless, end-to-end methods still face the challenges of high computational costs and a strong need for training data, which will require future research. Although beyond our main focus, this paradigm may also benefit from our findings, as discussed in the conclusion.

3 Preliminary

3.1 ReID-based Tracker

The tracking-by-detection paradigm [3, 5, 60] treats multiple object tracking as a two-step process. First, an object detector $\mathcal{D}$ is employed to localize all targets in a given frame I_t . Subsequently, these detections are associated with established trajectories based on a cost matrix or used to initialize new tracks. Following our discussion in Sec. 1, we simplify our experimental scope by concentrating on trackers that use only appearance cues for data association. Given an object bounding box, $\mathbf{b}_{t,i}$, in the t -th frame, a feature extraction network Φ is applied to obtain the corresponding visual feature $\mathbf{f}_{t,i}$, often referred to as a re-identification (ReID) feature. It is used to represent the appearance of each detection and to construct the feature of each trajectory. In practice, while numerous methods [36, 53, 57] for trajectory modeling exist, we adopt the widely-used Exponential Moving Average (EMA) update strategy due to its proven efficiency and effectiveness, as formulated below:

$$\hat{\mathbf{f}}_{t,\tau_j} = \lambda \mathbf{f}_{t,\tau_j} + (1 - \lambda) \hat{\mathbf{f}}_{t-1,\tau_j}, \quad (1)$$

where $\hat{\mathbf{f}}_{t-1,\tau_j}$ represents the appearance feature of track τ_j aggregated up to timestep $t - 1$, $\mathbf{f}_{t,\tau_j}$ is the ReID feature obtained from the extractor Φ at the current frame I_t , and λ is a momentum coefficient, typically set to a small value close to 0, that controls the update ratio.

Once the aforementioned features are prepared, we compute the matching cost for each detection-trajectory pair using a similarity metric. A common practice is to use the cosine similarity, which is calculated as follows:

$$\text{Cost}(t, i, \tau_j) = 1 - \text{Sim}(t, i, \tau_j) = 1 - \frac{\mathbf{f}_{t,i} \cdot \hat{\mathbf{f}}_{t-1,\tau_j}}{\|\mathbf{f}_{t,i}\| \|\hat{\mathbf{f}}_{t-1,\tau_j}\|}. \quad (2)$$

Accordingly, a cost matrix is constructed for the current frame based on all potential assignments. The Hungarian algorithm is then employed to find the globally optimal matching solution. Following this, the features of the matched tracks are updated according to Eq. (1), in preparation for the next time step.

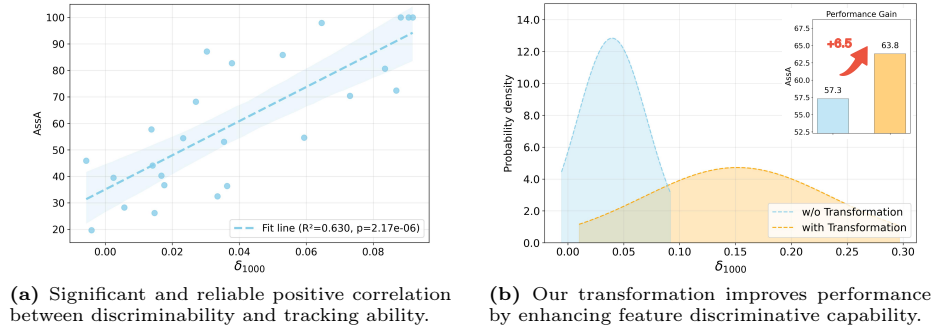


Fig. 2: Correlation between ReID feature discriminability δ_{1000} and tracking accuracy AssA on DanceTrack [47]. Similar analysis on [10] can be found in Appendix Fig.5.

3.2 Discriminative Capability Analysis

As stated in Sec. 3.1, since the tracker relies solely on appearance features for discrimination, it is intuitive to assume that the discriminative power of the ReID features is directly correlated with the tracking performance.

To validate this hypothesis, it is necessary to quantify the discriminative capability of the representation space. Since the tracking process relies on cosine similarity for affinity measurement, as shown in Eq. (2), we also adopt it as the cornerstone for evaluating the discriminability. To be more specific, we measure the discriminative power for the i -th detection in frame t using a score, $\delta(t, i)$. This score is defined as the similarity margin between the detection’s positive track and its most confusing negative track. Furthermore, since tracking failures are minority events within a given sequence, we focus on the most challenging cases. Therefore, for each video, we select the 1000 worst scores and compute their average. This metric, termed δ_{1000} , is used to quantify the discriminative ability of the ReID representations for a specific sequence (details in Appendix A.1). Accordingly, we conduct a statistical analysis on the representative dataset DanceTrack [47], as shown in Fig. 2a. The results reveal a significant and reliable positive correlation between the discriminative capability (δ_{1000}) of the ReID features and the object association accuracy (AssA [32]). This conclusion provides a clear motivation for our work: to boost tracking performance by explicitly enhancing the discriminability of the representation space, described in Sec. 4.

3.3 Fisher Linear Discriminant

Fisher Linear Discriminant (FLD) [14], also widely known as Linear Discriminant Analysis (LDA), is a classic supervised method used for both dimensionality reduction and classification. The core principle is to find a linear transformation that projects high-dimensional data onto a lower-dimensional space where the classes are maximally separated. In other words, the projection pulls the means of different classes far apart while keeping the data within each class tightly

clustered. Mathematically, this is achieved by defining a within-class scatter matrix, $\mathbf{S}_W$, and a between-class scatter matrix, $\mathbf{S}_B$. Given a set of N feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} = \mathbf{X} \in \mathbb{R}^{N \times d}$, each feature $\mathbf{x}$ is associated with one of C classes, the scatter matrices can be formulated as:

$$\mathbf{S}_W = \sum_{c=1}^C \sum_{\mathbf{x} \in \mathbf{X}_c} (\mathbf{x} - \bar{\mathbf{x}}_c)(\mathbf{x} - \bar{\mathbf{x}}_c)^T, \quad (3)$$

$$\mathbf{S}_B = \sum_{c=1}^C N_c (\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^T, \quad (4)$$

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i, \quad \bar{\mathbf{x}}_c = \frac{1}{N_c} \sum_{\mathbf{x} \in \mathbf{X}_c} \mathbf{x}, \quad (5)$$

where $\mathbf{X}_c$ represents the subset of $\mathbf{X}$ pertaining to class c . The optimal projection matrix, $\mathbf{W} \in \mathbb{R}^{d \times d'}$, is found by maximizing the Fisher criterion [14], which is the ratio of the between-class scatter to the within-class scatter in the projected space:

$$J(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}. \quad (6)$$

By applying the projection matrix $\mathbf{W}$ derived above, each feature $\mathbf{x}$ is converted into a new d' -dimensional vector with enhanced discriminability, where $d' = \min(C - 1, d)$.

4 Method

Based on the analysis in Sec. 3.2 and the result shown in Fig. 2a, a clear positive correlation exists between the discriminative capability of the ReID features and the final tracking performance. Therefore, in this section, our primary goal is to find a more discriminative representation space for distinguishing between trajectories. To this end, we mainly employ Fisher Linear Discriminant (FLD) [14] along with several customized techniques, which are detailed in Sec. 4.1 and Sec. 4.2 - Sec. 4.3, respectively. The overall illustration is shown in Fig. 3.

4.1 History-Aware Transformation for ReID Features

As discussed in Sec. 1, current multi-object tracking (MOT) methods [35, 36, 57] largely adopt ReID features directly from traditional re-identification methods [33, 44]. Since these models are required to distinguish between a vast number of open-set identities, the features they produce are, by design, as general as possible. In contrast, the multiple object tracking task only requires recognizing a closed set of identities within a single video. This creates a dilemma where the generality of traditional ReID features becomes a liability, as they lack the

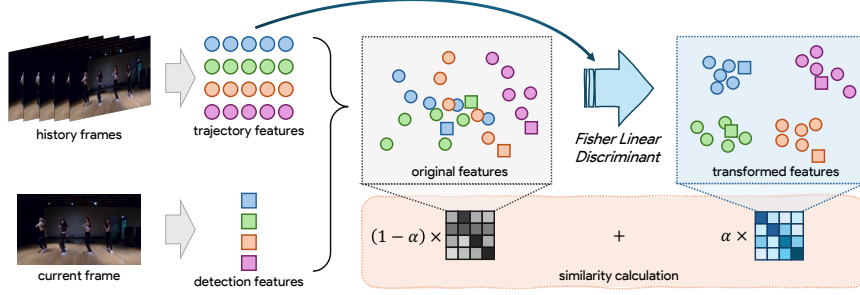


Fig. 3: Overview of our pipeline. We use different colors to indicate different identities (trajectories). In the original space, some overly similar targets cannot be well distinguished, leading to issues in the matching process. Therefore, we treat the trajectory features as conditions and apply a tailored *Fisher Linear Discriminant* to seek a better subspace for distinguishing different trajectories. Finally, both original and transformed features are used to calculate the similarity matrix, balancing generalization and specialization.

specificity needed to differentiate between these similar targets, as illustrated in Fig. 1. Therefore, we are motivated to seek a more specialized representation space to address the aforementioned challenges. Intuitively, this space should pull features belonging to the same trajectory closer, while pushing features from different trajectories further apart. This idea coincides perfectly with the objective of Fisher Linear Discriminant (FLD) [14] in its mathematical formulation, provided that we treat each *trajectory* as a *class* in the original framework. Specifically, by replacing the feature vector $\mathbf{x}$ in Eq. (3) - Eq. (5) with our ReID features $\mathbf{f}$, and substituting the number of classes C with the number of tracks N_τ , we can obtain the projection matrix $\mathbf{W}$ for tracking by maximizing the objective in Eq. (6).

However, FLD is a supervised method, which means it requires corresponding labels in addition to the feature vectors. This core prerequisite is unfulfilled in a standard tracking process. Therefore, we propose a history-aware dynamic labeling scheme to compensate for this absence. Practically, since tracking is an online process, at each timestep t , the historical track assignments from previous frames can serve as the supervisory signals for FLD. Although potential tracking errors exist, we believe the overall statistical signal remains reliable. Furthermore, since a target’s appearance gradually evolves during tracking, we only consider its T most recent features for each trajectory. This choice ensures both efficiency and effectiveness.

4.2 Temporally-Weighted Trajectory Centroid

Following the statement in Sec. 4.1, a naive implementation would be to average all T features $\{\mathbf{f}_{t-T, \tau_j}, \dots, \mathbf{f}_{t-2, \tau_j}, \mathbf{f}_{t-1, \tau_j}\}$ of the τ_j -th trajectory to serve as its mean feature center. According to the definition of FLD [14] and Eq. (4),

these feature centroids determine the distribution centers of the vectors after projection. Although this approach yields notable improvements, we still point out that it overlooks the temporal characteristics inherent in the tracking task. In online tracking, a target’s appearance evolves continuously over time. Even within the same trajectory, features that are closer temporally tend to have higher similarity. Hence, for identity allocation at the current moment, more recent ReID features should intuitively play a more significant role. In practice, we apply a temporal weighting to the mean calculation in Eq. (5):

$$\bar{\mathbf{f}} = \frac{1}{N_\tau} \sum_{j=1}^{N_\tau} \bar{\mathbf{f}}_{\tau_j}, \quad \bar{\mathbf{f}}_{\tau_j} = \frac{1}{\sum \lambda_{t'}} \sum_{t'=t-T}^{t-1} \lambda_{t'} \mathbf{f}_{t', \tau_j}, \quad \lambda_{t'} = (\lambda_0)^{t-t'}, \quad (7)$$

where λ_0 is a temporal decay coefficient with a value between 0 and 1. Using these temporal-weighted trajectory centroids in the calculation of Eq. (4) makes the final projection more attuned to the current temporal context, benefiting the similarity measurement at the time step t .

4.3 Knowledge Integration

Although we have found a more discriminative space conditioned on historical trajectories with the methods in Sec. 4.1 and Sec. 4.2, it still has some imperfections. First, the historical tracking results may contain errors, which can lead to a biased or suboptimal projection matrix. Second, because the transformed space is built only from the features of existing trajectories, it may not be robust enough for handling newborn targets. Therefore, we revisit the original representation space. Although it is not optimized for a given scenario, it offers more robust generalization capabilities, especially when facing unseen targets. This motivates our proposal to integrate it with the specialized subspace for a trade-off. Due to the disparate dimensionalities of these two spaces, our integration strategy operates on the similarity matrices rather than the vectors themselves. It can be formulated as follows:

$$\text{Cost}^*(t, i, \tau_j) = 1 - \text{Sim}^*(t, i, \tau_j) = 1 - [\alpha \cdot \text{Sim}'(t, i, \tau_j) + (1 - \alpha) \cdot \text{Sim}(t, i, \tau_j)], \quad (8)$$

where $\text{Sim}'(\cdot)$ is the similarity computed using the transformed ReID features, and α is a balancing coefficient. The Hungarian algorithm then finds the optimal assignment using the complete cost matrix constructed from the fused $\text{Cost}^*(\cdot)$. See Fig. 3 for an overview of this pipeline.

5 Experiments

5.1 Datasets and Metrics

Datasets. We select DanceTrack [47] and SportsMOT [10] as our primary experimental benchmarks because they both present a key challenge: targets within a

single video often exhibit a high degree of visual similarity. Specifically, DanceTrack features group dance scenarios, while SportsMOT includes three types of team sports. We also evaluate our approach on the TAO [11] dataset to demonstrate its effectiveness in diverse and general tracking cases. In addition, we report compact MOT17 [39] results in 4 and detailed results in Appendix B.1.

Metrics. On traditional MOT benchmarks [10, 39, 47], we select the Higher Order Tracking Accuracy (HOTA) [32] as the primary metric, especially its Association Accuracy (AssA) component. We also include MOTA [2] and IDF1 [43] in some experiments. To better evaluate the multi-category tracking problem, we employ the Tracking Every Thing Accuracy (TETA) [26] for TAO [11].

5.2 Implementation Details

To more clearly illustrate the improvements brought by our method, we focus our experiments on pure ReID-based trackers, as discussed in Sec. 3.1. Due to the lack of such publicly available trackers in the community, we construct a new tracker by combining the widely-used YOLOX [17] detector with the FastReID [33] model. For a fair comparison, we use the off-the-shelf weights from [5, 35, 57] for all network modules without any additional fine-tuning. To ensure the baseline performs at its full potential, we conduct an exhaustive grid search to identify the optimal hyperparameter configuration for each specific benchmark. This fortified baseline, denoted as *FastReID-MOT*, ensures that any subsequent performance gains are strictly attributable to our proposed methodology. To further evaluate the potential of our approach, we select several representative hybrid-based trackers [5, 57, 60] and incorporate our method into their ReID branches to boost performance. Regarding MASA [27], we also directly adopt the official model weights without any modifications, covering a diverse range of backbones to guarantee the breadth of our evaluation. For notation, we add the prefix **HAT-** to methods that use our **H**istory-**A**ware **T**ransformation approach.

5.3 State-of-the-art Comparison

FastReID-MOT. We compare our method (HAT-FastReID-MOT) against the baseline (FastReID-MOT) on DanceTrack [47] and SportsMOT [10] in Tab. 1 and Tab. 2. † indicates that hyperparameters are fine-tuned on the corresponding dataset to maximize performance; otherwise, the default settings from our ablation study are used, as stated in Sec. 5.4. Our approach yields substantial performance gains over the baseline. On the challenging DanceTrack dataset, our appearance-only method even achieves results comparable to several recent hybrid and motion-based trackers [35, 56]. Even more impressively, on the SportsMOT benchmark, our ReID-only tracker (HAT-FastReID-MOT) sets a new state-of-the-art. It significantly surpasses existing hybrid-based trackers [35] and end-to-end models [15, 46]. This result both vindicates our approach and highlights the need to reconsider the true potential of ReID features.

Hybrid-based Tracker. To further validate the effectiveness of our method, we inserted it into several recent well-known trackers [5, 57, 60]. The results

Table 1: Comparison with state-of-the-art methods on the DanceTrack test set.

Methods	HOTA	DetA	AssA
<u><i>motion-based:</i></u>			
ByteTrack [60]	47.7	71.0	32.1
OC-SORT [5]	55.1	80.3	38.3
C-BIoU [56]	60.6	81.3	45.4
<u><i>ReID-based:</i></u>			
QDTrack [40]	54.2	80.1	36.8
FastReID-MOT (baseline)	50.6	81.1	31.6
HAT-FastReID-MOT	58.6	81.3	42.3
HAT-FastReID-MOT†	61.2	81.6	46.0
<u><i>hybrid-based:</i></u>			
FairMOT [61]	39.7	66.7	23.8
DeepSORT [53]	45.6	71.0	29.7
StrongSORT [13]	55.6	80.7	38.6
DiffMOT [35]	62.3	82.5	47.2
Hybrid-SORT-ReID [57]	65.7	—	—
ByteTrack-ReID	52.4	71.0	38.7
HAT-ByteTrack-ReID	56.1	71.4	44.2
OC-SORT-ReID	60.8	81.0	45.7
HAT-OC-SORT-ReID	64.6	81.5	51.3
HAT-Hybrid-SORT-ReID	66.9	81.5	55.0
<u><i>end-to-end:</i></u>			
MeMOTR [16]	63.4	77.0	52.3
CO-MOT [55]	65.3	80.1	53.5
SambaMOTR [46]	67.2	78.8	57.5
MOTIP [15]	69.6	80.4	60.4

Table 2: Performance on the SportsMOT test set. Results marked with * denote joint training involving the validation set of SportsMOT.

Methods	HOTA	DetA	AssA
<u><i>motion-based:</i></u>			
ByteTrack [60]	62.8	77.1	51.2
OC-SORT [5]	71.9	86.4	59.8
ByteTrack* [60]	64.1	78.5	52.3
OC-SORT* [5]	73.7	88.5	61.5
<u><i>ReID-based:</i></u>			
QDTrack [40]	60.4	77.5	47.2
FastReID-MOT (baseline)	67.3	86.8	52.3
HAT-FastReID-MOT	78.1	87.3	69.9
HAT-FastReID-MOT†	78.9	87.4	71.3
HAT-FastReID-MOT†*	80.8	89.4	73.1
<u><i>hybrid-based:</i></u>			
BoT-SORT [1]	68.7	84.4	55.9
DiffMOT [35]	72.1	86.0	60.5
ByteTrack-ReID	65.1	76.8	55.1
HAT-ByteTrack-ReID	72.4	77.3	67.8
OC-SORT-ReID	74.1	86.8	63.3
HAT-OC-SORT-ReID	81.2	87.2	75.6
HAT-OC-SORT-ReID*	82.4	89.3	76.1
<u><i>end-to-end:</i></u>			
TrackFormer [38]	63.3	66.0	61.1
MeMOTR [16]	68.8	82.0	57.8
MOTIP [15]	72.6	83.5	63.2

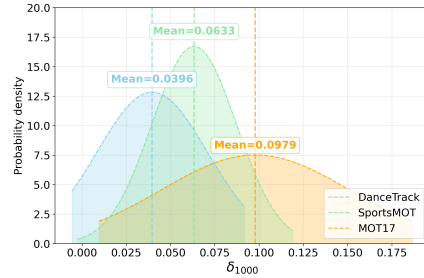
in Tab. 1 and Tab. 2 show that our method can consistently bring significant improvements when applied to hybrid-based trackers. While many prior studies [16, 46, 59] suggest that end-to-end methods possess an inherent advantage on DanceTrack due to their flexibility, our approach surpasses several recent E2E models [46, 55], demonstrating significant potential. A more detailed discussion regarding E2E methods is provided in Appendix C. The smaller performance gains on hybrid-based methods can be attributed to both performance saturation and the inherent design of these trackers, which often prioritizes motion and thus limits the impact of our appearance enhancements. In addition, the complex design of collaborative trackers also limits the improvement brought by a single module and makes it challenging to coordinate different modules. To further exploit the capability of the proposed approach in hybrid-based methods, we believe that it may be necessary to redesign the fusion algorithm of the tracker

Table 3: Evaluating our method by applying it to MASA [27]. All models are trained on a large-scale image segmentation dataset [25] with different visual backbones.

Methods	DanceTrack test		SportsMOT test		TAO val	
	HOTA	AssA	HOTA	AssA	TETA	AssocA
<i>MASA [27]:</i>						
MASA-R50	50.8	31.6	71.6	58.9	45.8	42.7
MASA-Detic	50.6	31.5	72.2	60.1	46.5	44.5
MASA-G-DINO	50.4	31.2	72.8	61.0	46.8	45.0
MASA-SAM-B	49.4	29.9	71.9	59.5	46.2	43.7
<i>Ours (HAT):</i>						
MASA-R50	54.3 (+3.5)	36.1 (+4.5)	73.7 (+2.1)	62.4 (+3.5)	46.4 (+0.6)	44.4 (+1.7)
MASA-Detic	54.3 (+3.7)	36.2 (+5.7)	74.5 (+2.3)	63.7 (+3.6)	47.2 (+0.7)	46.4 (+1.9)
MASA-G-DINO	53.9 (+3.5)	35.7 (+4.5)	74.7 (+1.9)	64.1 (+3.1)	47.5 (+0.7)	46.7 (+1.7)
MASA-SAM-B	52.1 (+2.7)	33.4 (+3.5)	73.4 (+1.5)	61.9 (+2.4)	46.9 (+0.7)	45.6 (+1.9)

Table 4: Performance comparison with state-of-the-art methods on MOT17 [39]. The best and second-best results are denoted in **bold** and underline, respectively.

Methods	HOTA	DetA	AssA	IDF1
<i>reid-based:</i>				
QDTrack [40]	53.9	55.6	52.7	66.3
FastReID-MOT (baseline)	<u>61.5</u>	<u>63.4</u>	<u>60.0</u>	<u>73.5</u>
HAT-FastReID-MOT [†]	63.5	64.0	63.2	77.5
<i>hybrid-based:</i>				
FairMOT [61]	59.3	60.9	58.0	72.3
DeepSORT [53]	61.2	63.1	59.7	74.5
DiffMOT [35]	64.5	64.7	64.6	79.3
OC-SORT-ReID	64.1	<u>64.4</u>	64.0	79.0
HAT-OC-SORT-ReID	<u>64.2</u>	<u>64.4</u>	<u>64.1</u>	<u>79.2</u>

**Fig. 4:** Comparison of ReID separability in the original feature space on DanceTrack [47], SportsMOT [10], and MOT17 [39] based on δ_{1000} .

itself and conduct extensive engineering-level parameter tuning, which is beyond the scope of this work.

MASA. To investigate the generalization of our method across different ReID representation spaces, we conduct experiments on MASA [27]. This framework is ideal as it includes a variety of visual backbones [19, 25, 29, 63] and is pre-trained on a general-purpose segmentation dataset [25]. To complement our evaluations on specialized datasets [10, 47], we further employ the TAO [11] benchmark. With its vast category diversity, TAO serves as a rigorous testbed to assess whether our approach maintains consistent reliability in open-world environments. Tab. 3 shows that our approach brings consistent and significant boosts across all tested visual backbones. However, the performance gains are less pronounced compared to those observed in Tab. 1 and Tab. 2. We attribute this to the fact that our method is designed to refine and distill an existing feature

Table 5: Comparison of transformation selections. *Oracle* and *YOLOX* denote the sources of the detection results, while d and d' indicate the original and projected feature dimension, respectively. N_{obj} and N_{id} are the total number of historical samples and trajectories, respectively. If $d' > d$, the target dimension will be set to d .

#	$\mathcal{D}$	Method	d'	DanceTrack val		SportsMOT val	
				HOTA	AssA	HOTA	AssA
# 1	Oracle	–	d	74.9	57.3	86.2	74.7
# 2		PCA	$N_{obj} - 1$	75.1 (+0.2)	57.6 (+0.3)	85.6 (-0.6)	73.9 (-0.8)
# 3		PCA	$N_{id} - 1$	56.3 (-18.6)	32.5 (-24.8)	69.4 (-16.8)	49.0 (-25.7)
# 4		FLD	$N_{id} - 1$	79.0 (+4.1)	63.8 (+6.5)	92.2 (+6.0)	85.4 (+10.7)
# 5	YOLOX	–	d	51.1	33.4	73.7	61.5
# 6		PCA	$N_{obj} - 1$	45.3 (-5.8)	26.7 (-6.7)	70.6 (-3.1)	56.4 (-5.1)
# 7		PCA	$N_{id} - 1$	43.5 (-7.6)	24.6 (-8.8)	61.3 (-12.4)	42.7 (-18.8)
# 8		FLD	$N_{id} - 1$	57.7 (+6.6)	42.7 (+9.3)	81.1 (+7.4)	74.1 (+12.6)

space. Since MASA is not explicitly trained on specific datasets, its initial representations may possess limited discriminative power in challenging scenarios, providing a sub-optimal foundation for our distillation process. Furthermore, the TAO [11] dataset encompasses a vast array of heterogeneous object categories with minimal inter-class similarity, which naturally constrains the headroom for our algorithm. Nevertheless, the consistent improvements achieved across such diverse scenarios underscore the robust generalization of our method.

MOT17. We provide a compact report of tracking performance on MOT17 [39] in Tab. 4. Unlike the strong performance observed on DanceTrack and SportsMOT, our method achieves only moderate results on MOT17. Even on our self-constructed FastReID-MOT baseline, it brings an improvement of only 2.0 HOTA. As shown in Fig. 4, we visualize the difference in discriminability within the original representation space across different datasets. MOT17 exhibits the highest appearance-level discriminability, likely due to the substantially different clothing of pedestrians. This property substantially limits the room for improvement of our proposed method. This may also explain the marginal improvement on TAO [11] in Tab. 3, which is a practical consideration in real-world scenarios.

5.4 Ablation Study

We verify the effectiveness of each component in this section, using the ReID-based tracker from Sec. 3.1 and Sec. 5.2 as the baseline. For all experiments, except those in Tab. 5, we use the detections from the public YOLOX model [5,17]. We perform a step-by-step inquiry in Tab. 5 to Tab. 8, with details provided in Appendix A.5. The gray cells denote the final default hyperparameter settings. **History-Aware Transformation.** As shown in Tab. 5, applying the FLD-based ReID feature transformation, as described in Sec. 4.1, significantly improves tracking performance over the baseline (# 1 and # 5). Following the correlation analysis in Sec. 3.2, we visualize the change in the discriminative ability

Table 6: History Length T

T	HOTA	AssA	MOTA	IDF1
10	54.3	37.8	86.7	53.2
20	56.2	40.4	86.5	55.2
40	57.0	41.4	86.5	56.6
60	57.7	42.7	86.3	56.7
80	56.3	40.5	86.2	55.4
∞	55.3	39.2	85.1	52.2

Table 7: Temporal Decay Coefficient λ_0

λ_0	HOTA	AssA	MOTA	IDF1
1.00	57.7	42.7	86.3	56.7
0.95	58.5	42.5	86.8	58.2
0.90	59.3	44.8	86.9	59.8
0.80	59.1	44.7	87.0	59.6
0.60	58.1	43.2	87.0	57.9
0.40	57.8	42.8	86.9	57.1

Table 8: Balancing Coefficient α .

α	HOTA	AssA	MOTA	IDF1
1.0	59.3	44.8	86.9	59.8
0.9	60.6	46.8	87.1	61.7
0.8	59.3	45.0	87.2	60.4
0.6	59.1	44.7	87.6	60.6
0.4	58.7	43.9	87.9	60.1

Table 9: Comparison of different fitting strategies for FLD.

FLD Strategy	DanceTrack		SportsMOT	
	HOTA	AssA	HOTA	AssA
–	51.1	33.4	73.7	61.5
Supervised	51.4	34.0	77.6	68.0
History-Aware	57.7	42.7	81.1	74.1

δ_{1000} of ReID features under an oracle detection setting, as shown in Fig. 2b. This serves as clear evidence that our history-aware transformation boosts the separability of visual representations, thereby improving tracking capabilities. For comparison, we evaluate a PCA-based transformation in Tab. 5, but it resulted in a performance drop. This is because Principal Component Analysis (PCA) is designed to maximize global variance and is oblivious to trajectory labels, which we believe are essential for finding an optimal space for tracking.

Supervised vs. History-Aware. While our method achieves significant improvements over the baseline in Tab. 5, these results alone do not fully substantiate our underlying motivation. A notable concern is whether this improvement stems from the introduction of FLD or the utilization of historical information as guidance, as we claimed. To verify this, we fit an FLD using the ground-truth labels and features from the training set (denoted as *supervised*) to serve as a competitor against our proposed approach. Experiments in Tab. 9 demonstrate that leveraging vast amounts of labeled data from different videos does not necessarily enable the FLD to identify a more appropriate projection space. This observation aligns with our argument in Fig. 1 and Sec. 1, namely that the optimal subspace varies significantly across different sequences, simultaneously reinforcing that our underlying motivation holds enduring significance beyond the current implementation. This insight may inspire future research to employ more advanced algorithmic designs to achieve this objective, rather than being limited to the use of FLD.

History Length T . As stated at the end of Sec. 4.1, we only consider ReID features from the T most recent frames. Although using a too-short temporal length T decreases the credibility of the reference samples, it still provides a

notable enhancement compared to the baseline tracker (54.3 *vs.* 51.1 HOTA), as shown in Tab. 6. Conversely, an excessively large T incorporates outdated features, making the distribution less representative and degrading performance.

Temporally-Weighted Trajectory Centroid. Following our discussion in Sec. 4.2 about the varying temporal importance of features, we introduce the coefficient λ_0 to weight them accordingly. Tab. 7 demonstrates that using the temporal-weighted trajectory centroid can significantly enhance tracking performance. However, it is essential to note that excessively small values of λ_0 may lead to an overreliance on recent samples, resulting in a decline in robustness.

Knowledge Integration. In Tab. 8, we investigate various fusion coefficients α to balance robustness and specialization. These results indicate that this involves a trade-off, prompting us to choose 0.9 as our default setting. In addition, this supports the concept outlined in Sec. 4.3, valuing the complementarity of those two spaces can boost the reliability of ReID features.

6 Conclusion

In this paper, we challenge a long-standing practice in multiple object tracking: *the direct adoption of appearance matching strategies from the re-identification task, an approach we argue is fundamentally inappropriate for tracking*. We contend that visual representations in MOT should be tailored to discriminate among the finite set in a given video sequence, as opposed to the open-set challenge. Rather than the conventional approach of optimizing over training data, this should be an optimization specifically tailored to the current sequence that has been absent from previous methodologies. To this end, we proposed an approach that leverages the tracking history to guide an adaptive transformation of the feature space, thereby boosting its discriminability. Comprehensive experiments validate the effectiveness and versatility of our proposed approach and establish the new state-of-the-art performance. These results provide compelling evidence that the potential of ReID features in MOT has been significantly underestimated, while simultaneously suggesting that our motivation holds enduring value for future explorations beyond this specific implementation. Therefore, we hope our findings spur a wave of research into this crucial problem, whether in the form of new training-free components or as guiding principles for developing learnable modules.

Acknowledgements

This work is supported by the Basic Research Program of Jiangsu (No. BK20250009), the Fundamental Research Funds for the Central Universities (No.020214380140), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Aharon, N., Orfaig, R., Bobrovsky, B.: Bot-sort: Robust associations multi-pedestrian tracking. *CoRR* **abs/2206.14651** (2022)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.* **2008** (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F.T., Upcroft, B.: Simple online and realtime tracking. In: *ICIP*. pp. 3464–3468. IEEE (2016)
4. Cai, J., Xu, M., Li, W., Xiong, Y., Xia, W., Tu, Z., Soatto, S.: Memot: Multi-object tracking with memory. In: *CVPR*. pp. 8080–8090. IEEE (2022)
5. Cao, J., Weng, X., Khiroudkar, R., Pang, J., Kitani, K.: Observation-centric SORT: rethinking SORT for robust multi-object tracking. *CoRR* **abs/2203.14360** (2022)
6. Cao, X., Zheng, Y., Yao, Y., Qin, H., Cao, X., Guo, S.: TOPIC: A parallel association paradigm for multi-object tracking under complex motions and diverse scenes. *IEEE Trans. Image Process.* **34**, 743–758 (2025)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *ECCV* (1). *Lecture Notes in Computer Science*, vol. 12346, pp. 213–229. Springer (2020)
8. Chen, S., Yu, E., Li, J., Tao, W.: Delving into the trajectory long-tail distribution for multi-object tracking. In: *CVPR*. pp. 19341–19351. IEEE (2024)
9. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: *ECCV* (4). *Lecture Notes in Computer Science*, vol. 7575, pp. 215–230. Springer (2012)
10. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In: *ICCV* (2023)
11. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A large-scale benchmark for tracking any object. In: *ECCV* (5). *Lecture Notes in Computer Science*, vol. 12350, pp. 436–454. Springer (2020)
12. Dendorfer, P., Yugay, V., Osep, A., Leal-Taixé, L.: Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? In: *NeurIPS* (2022)
13. Du, Y., Zhao, Z., Song, Y., Zhao, Y., Su, F., Gong, T., Meng, H.: Strongsort: Make deepsort great again. *IEEE Trans. Multim.* **25**, 8725–8737 (2023)
14. Fisher, R.A.: The statistical utilization of multiple measurements. *Annals of eugenics* **8**(4), 376–386 (1938)
15. Gao, R., Qi, J., Wang, L.: Multiple object tracking as id prediction. In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. pp. 27883–27893 (June 2025)
16. Gao, R., Wang, L.: MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9901–9910 (October 2023)
17. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. *CoRR* **abs/2107.08430** (2021)
18. Guo, W., Wang, T., Gao, J., Zhang, T., Xu, C.: Occlusion-aware multi-object tracking via joint diffusion motion prediction and appearance purification. *IEEE Transactions on Circuits and Systems for Video Technology* (2026)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778. IEEE Computer Society (2016)
20. Hou, Y., Wang, Z., Wang, S., Zheng, L.: Adaptive affinity for associations in multi-target multi-camera tracking. *IEEE Trans. Image Process.* **31**, 612–622 (2022)

21. Hu, Y., Hu, L., Kong, Q., Fan, B.: A survey on end-to-end perception and prediction for autonomous driving. *Mach. Intell. Res.* **22**(6), 999–1030 (2025)
22. Huang, H., Yang, C., Chai, W., Jiang, Z., Hwang, J.: Exploring learning-based motion models in multi-object tracking. *CoRR* **abs/2403.10826** (2024)
23. Kesa, O., Styles, O., Sanchez, V.: Multiple object tracking and forecasting: Jointly predicting current and future object locations. In: *WACV (Workshops)*. pp. 560–569. IEEE (2022)
24. Khurana, T., Dave, A., Ramanan, D.: Detecting invisible people. In: *ICCV*. pp. 3154–3164. IEEE (2021)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: *ICCV*. pp. 3992–4003. IEEE (2023)
26. Li, S., Danelljan, M., Ding, H., Huang, T.E., Yu, F.: Tracking every thing in the wild. In: *ECCV (22)*. *Lecture Notes in Computer Science*, vol. 13682, pp. 498–515. Springer (2022)
27. Li, S., Ke, L., Danelljan, M., Piccinelli, L., Segù, M., Gool, L.V., Yu, F.: Matching anything by segmenting anything. In: *CVPR*. pp. 18963–18973. IEEE (2024)
28. Li, Y., Zhou, S., Qin, Z., Wang, L., Wang, J., Zheng, N.: Single-shot and multi-shot feature learning for multi-object tracking. *IEEE Trans. Multim.* **26**, 9515–9526 (2024)
29. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: *ECCV (47)*. *Lecture Notes in Computer Science*, vol. 15105, pp. 38–55. Springer (2024)
30. Liu, Y., Wang, Z., Zhou, X., Zheng, L.: A study of using synthetic data for effective association knowledge learning. *Mach. Intell. Res.* **20**(2), 194–206 (2023)
31. Liu, Z., Wang, X., Wang, C., Liu, W., Bai, X.: Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *CoRR* **abs/2306.05238** (2023)
32. Luiten, J., Osep, A., Dendorfer, P., Torr, P.H.S., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **129**(2), 548–578 (2021)
33. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: *CVPR Workshops*. pp. 1487–1495. Computer Vision Foundation / IEEE (2019)
34. Luo, R., Song, Z., Ma, L., Wei, J., Yang, W., Yang, M.: Diffusiontrack: Diffusion model for multi-object tracking. In: *AAAI*. pp. 3991–3999. AAAI Press (2024)
35. Lv, W., Huang, Y., Zhang, N., Lin, R., Han, M., Zeng, D.: Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In: *CVPR*. pp. 19321–19330. IEEE (2024)
36. Maggolino, G., Ahmad, A., Cao, J., Kitani, K.: Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813* (2023)
37. Mancusi, G., Panariello, A., Porrello, A., Fabbri, M., Calderara, S., Cucchiara, R.: Trackflow: Multi-object tracking with normalizing flows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9531–9543 (2023)
38. Meinhardt, T., Kirillov, A., Leal-Taixé, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: *CVPR*. pp. 8834–8844. IEEE (2022)
39. Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. *CoRR* **abs/1603.00831** (2016)

40. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: CVPR. pp. 164–173. Computer Vision Foundation / IEEE (2021)
41. Plaen, P.D., Marinello, N., Proesmans, M., Tuytelaars, T., Gool, L.V.: Contrastive learning for multi-object tracking with transformers. In: WACV. pp. 6853–6863. IEEE (2024)
42. Qin, Z., Zhou, S., Wang, L., Duan, J., Hua, G., Tang, W.: Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In: CVPR. pp. 17939–17948. IEEE (2023)
43. Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshops (2). Lecture Notes in Computer Science, vol. 9914, pp. 17–35 (2016)
44. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: CVPR. pp. 6036–6046. Computer Vision Foundation / IEEE Computer Society (2018)
45. Saraceni, L., Motoi, I.M., Nardi, D., Ciarfuglia, T.A.: Agrisort: A simple online real-time tracking-by-detection framework for robotics in precision agriculture. In: ICRA. pp. 2675–2682. IEEE (2024)
46. Segù, M., Piccinelli, L., Li, S., Yang, Y., Schiele, B., Gool, L.V.: Samba: Synchronized set-of-sequences modeling for multiple object tracking. *CoRR* **abs/2410.01806** (2024)
47. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: CVPR. pp. 20961–20970. IEEE (2022)
48. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS). pp. 1–6. IEEE (2024)
49. Wang, Y., Li, R., Zhang, D., Li, M., Cao, J., Zheng, Z.: Catrack: Condition-aware multi-object tracking with temporally enhanced appearance features. *Knowl. Based Syst.* **308**, 112760 (2025)
50. Wang, Y., Zhang, D., Li, R., Zheng, Z., Li, M.: PD-SORT: occlusion-robust multi-object tracking using pseudo-depth cues. *IEEE Trans. Consumer Electron.* **71**(1), 165–177 (2025)
51. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: ECCV (11). Lecture Notes in Computer Science, vol. 12356, pp. 107–122. Springer (2020)
52. Welch, G., Bishop, G., et al.: An introduction to the kalman filter. Tech. rep., University of North Carolina at Chapel Hill (1995)
53. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP. pp. 3645–3649. IEEE (2017)
54. Xiao, C., Cao, Q., Luo, Z., Lan, L.: Mambatrack: A simple baseline for multiple object tracking with state space model. In: ACM Multimedia. pp. 4082–4091. ACM (2024)
55. Yan, F., Luo, W., Zhong, Y., Gan, Y., Ma, L.: CO-MOT: boosting end-to-end transformer-based multi-object tracking via coopetition label assignment and shadow sets. In: ICLR. OpenReview.net (2025)
56. Yang, F., Odashima, S., Masui, S., Jiang, S.: Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In: WACV. pp. 4788–4797. IEEE (2023)

57. Yang, M., Han, G., Yan, B., Zhang, W., Qi, J., Lu, H., Wang, D.: Hybrid-sort: Weak cues matter for online multi-object tracking. CoRR **abs/2308.00783** (2023)
58. Yi, K., Luo, K., Luo, X., Huang, J., Wu, H., Hu, R., Hao, W.: Ucmctrack: Multi-object tracking with uniform camera motion compensation. In: AAAI. pp. 6702–6710. AAAI Press (2024)
59. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: end-to-end multiple-object tracking with transformer. In: ECCV (27). Lecture Notes in Computer Science, vol. 13687, pp. 659–675. Springer (2022)
60. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV (22). Lecture Notes in Computer Science, vol. 13682, pp. 1–21. Springer (2022)
61. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **129**(11), 3069–3087 (2021)
62. Zhang, Y., Wang, T., Zhang, X.: Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors (2022)
63. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (9). Lecture Notes in Computer Science, vol. 13669, pp. 350–368. Springer (2022)
64. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: ECCV (4). Lecture Notes in Computer Science, vol. 12349, pp. 474–490. Springer (2020)
65. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR. OpenReview.net (2021)