LA-MOTR: End-to-End Multi-Object Tracking by Learnable Association

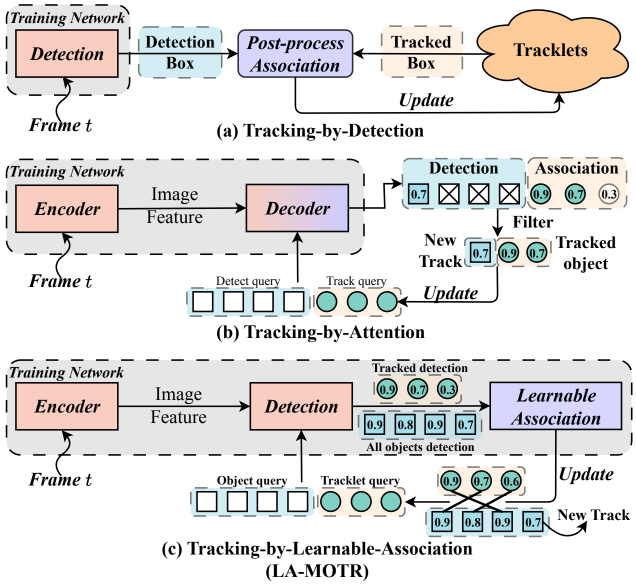
Peng Wang, Yongcai Wang*, Hualong Cao, Wang Chen, Deying Li
School of Information, Renmin University of China
{peng.wang, ycw, caohualong, chenwang, deyingli}@ruc.edu.cn

Abstract

*This paper proposes **LA-MOTR**, a novel Tracking-by-Learnable-Association framework that resolves the competing optimization objectives between detection and association in end-to-end Tracking-by-Attention (TbA) Multi-Object Tracking. Current TbA methods employ shared decoders for simultaneous object detection and tracklet association, often resulting in task interference and suboptimal accuracy. By contrast, our end-to-end framework decouples these tasks into two specialized modules: Separated Object-Tracklet Detection (SOTD) and Spatial-Guided Learnable Association (SGLA). This decoupled design offers flexibility and explainability. In particular, SOTD independently detects new objects and existing tracklets in each frame, while SGLA associates them via Spatial-Weighted Learnable Attention module guided by relative spatial cues. Temporal coherence is further maintained through Tracklet Updates Module. The learnable association mechanism resolves the inherent suboptimal association issues in decoupled frameworks, avoiding the task interference commonly observed in joint approaches. Evaluations on DanceTrack, MOT17, and SportMOT datasets demonstrate state-of-the-art performance. Extensive ablation studies validate the effectiveness of the designed modules. Code is available at* `https://github.com/PenK1nG/LA-MOTR`.

1. Introduction

Multi-Object Tracking (MOT) [36, 47] involves associating targets across a continuous video sequence to maintain consistent IDs for each object. As a fundamental task in computer vision, MOT has been crucial in applications ranging from early video surveillance systems [10, 12, 22] to modern autonomous driving technologies [6, 57]. Traditional MOT frameworks, such as Tracking-by-Detection (TbD) [4, 29, 50, 53, 60], employ a sequential pipeline: first detecting objects per-frame using established detectors [15, 20, 38, 63], then associating [5, 26, 39, 53] these detections with existing tracklets, as illustrated in Figure 1(a).



(a) Tracking-by-Detection

(b) Tracking-by-Attention

(c) Tracking-by-Learnable-Association (LA-MOTR)

Figure 1. **Comparison of Different MOT Frameworks**. (a) consists of a training-based detector and a post-processing association. (b) utilizes a decoder to simultaneously detect new targets and associate tracklets. (c) the proposed framework **LA-MOTR** introduces a learnable association mechanism to enhance the end-to-end pipeline by decoupling the detection and association.

However, decoupled frameworks find it challenging to simultaneously optimize detection and association [58, 60]. With advancements in deep learning [9, 21, 48], recent studies [43, 46, 58] have attempted to use an end-to-end pipeline to simultaneously address the object detection and the tracklet association. Among these, the Tracking-by-Attention (TbA) framework, which utilizes Transformer-based models for end-to-end tracking, has achieved the best performance. TbA [58] extends the DETR [9, 31, 65] framework to use a single decoder based on queries to perform both the detection the association tasks, as illustrated in Figure 1(b). This approach of directly decoding new target detections and matching them with historical trajectories through the network performs exceptionally well in scenarios with complex target movements [11, 47]. However, when applied

[*] Corresponding author.

to crowded environments such as MOT17 [36], the performance significantly declines. This decline is primarily due to the difficulty in detecting new targets, which subsequently reduces tracking accuracy. Studies [18, 54, 56, 62] indicate that the number of new targets supervising the detection queries is significantly smaller than the number of the tracklets supervising the track queries. To mitigate this issue, these methods adjust the supervision strength [54, 56] across different query types.

We argue that a more fundamental issue in TbA arises from the shared use of a single attention decoder network for both the detect queries and the track queries, introducing functional ambiguity. Detect queries utilize the attention decoder to identify new objects, while track queries employ the same decoder to associate current detections with existing tracklets. This dual usage blurs the decoder's functionality, reducing its effectiveness for both detecting new targets and associating them with historical track queries. Consequently, this increases the risk of tracking failures, as shown in Figure 6(a). Additionally, end-to-end networks often prioritize track queries due to their stronger supervisory signals [56, 62], which detracts the detection of new targets.

To resolve the competing optimization objectives between detection and association, we introduce **LA-MOTR**, a novel *Tracking-by-Learnable-Association* framework for end-to-end MOT. LA-MOTR separately detects all objects and historical tracklets within the current frame and associates them using an online learnable association module to update the tracklets. The updated tracklets in the current frame then guide tracklet detection in the next frame. The procedure is illustrated in Figure 1(c). In particular, for each frame, we utilize the tracklet queries propagated across frames alongside the blank object queries to separately detect the targets of historical tracklets and all objects in the current frame. Next, we propose a *Spatial-Guided Learnable Association* (SGLA) module that leverages relative spatial cues from all objects and tracklets to guide feature interaction and compute association scores. Specifically, in SGLA, we employ Spatial-Weighted Attention to fuse object features with tracklet features, providing mutual information that enables more accurate detection and matching of blurred or occluded objects to their corresponding trajectories. Finally, we use the tracklet features obtained from current frame as the tracklet queries for the next frame. The key contributions can be summarized as following:

- We propose LA-MOTR, a novel Tracking-by-Learnable-Association framework that effectively resolves competing optimization, enabling more flexible and interpretable detection and association in end-to-end MOT.
- We present the Separated Object-Tracklet Detection module, Spatial-Guided Learnable Association module, and Tracklet Update module to facilitate LA-MOTR.
- Our approach achieves state-of-the-art performance among end-to-end MOT methods on challenging datasets, including DanceTrack, MOT17, and SportsMOT. Experimental results and comprehensive ablation studies validate its effectiveness in synchronizing detection accuracy and tracking consistency.

2. Related Work

2.1. Tracking-by-Detection

Tracking-by-Detection (TbD) [1, 4, 14, 16, 29] first detects objects in each frame and then associates these detections with existing trajectories. In pedestrian tracking scenarios using fixed cameras [12, 28, 36], targets generally exhibit linear motion and maintain consistent appearance features. Consequently, association in TbD primarily relies on motion and appearance information. SORT [4] utilizes a Kalman filter [26] to predict object positions based on historical trajectories and computes the Intersection over Union (IoU) with detected targets. Deep SORT [53] enhances SORT by incorporating a feature extraction network and employing cosine similarity for matching. To balance speed and performance, JDE [51] and FairMOT [60] integrate detection and feature extraction into a unified network.

In more complex scenarios [11, 47, 52], the assumptions of linear target motion and consistent appearance features make association challenging. ByteTrack [61] addresses this issue by reusing low-confidence detections to enhance association robustness. OC-SORT [8] reduces error accumulation in the Kalman filter during occlusion periods by computing a virtual trajectory. Hybrid-SORT [55] incorporates confidence and height state as additional cues in the association process. DroneMOT [49] decomposes the drone's motion and integrates it with the target's motion to construct the association matrix.

However, these two-stage TbD methods are unable to simultaneously optimize detection and association, often requiring customized association strategies and manual parameter tuning. In contrast, our LA-MOTR introduces a learnable association module within an end-to-end framework, enabling the network to learn association relationships across diverse scenarios.

2.2. Tracking-by-Attention

Tracking-by-Attention (TbA) employs an end-to-end framework, eliminating the need for manual parameter tuning across diverse scenarios. Inspired by DETR [9, 65], TbA utilizes learnable queries to detect objects and historical trajectory queries to monitor them in the current frame. In contrast to TransTrack [46] and TrackFormer [35], which rely on Intersection over Union (IoU) or Non-Maximum Suppression (NMS) for matching final trajectories, MOTR [58] uses track queries to associate tracklets and detects newly appearing objects using de-

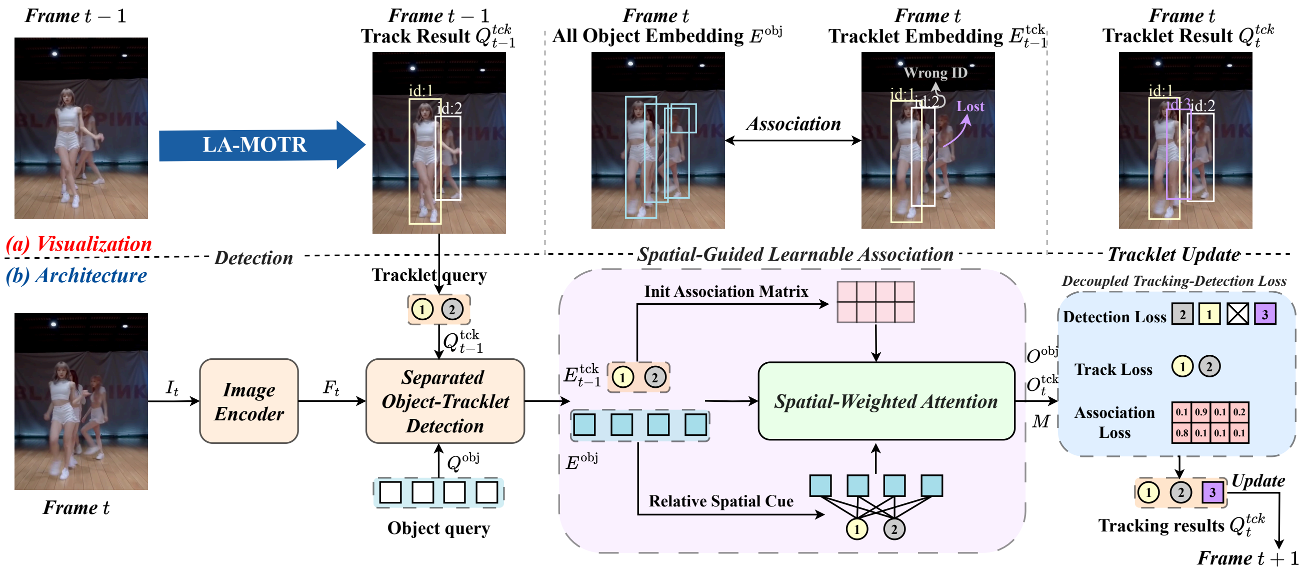


Figure 2. The overall **architecture** of LA-MOTR along with a **visualization** of its corresponding processes. Initially, the image is processed by the image encoder. Concurrently, the tracklet query and blank object query are input into the **Separated Object-Tracklet Detection**, which outputs the tracklet and all objects detection embeddings for the current frame. Then, relative spatial cues between objects and tracklets guide their feature interactions within the **Spatial-Guided Learnable Association** module to create an association matrix linking objects to tracklets. Finally, the tracklet query for time $t+1$ is generated, completing the tracking cycle.

tect queries. This approach implicitly integrates detection and trajectory association within a single attention decoder. Overall, this learning-based, end-to-end methodology effectively manages tracking across various scenarios.

The complexity of an integrated attention decoder poses challenges in balancing new object detection with existing trajectory tracking. To address this, MOTRv2 [62] incorporates a YOLOx [20] detector to generate queries for all objects, reducing conflicts between detection and association tasks. Building on this, MOTRv3 [56] enhances detection and association supervision by correcting label assignment imbalances during training. Similarly, CoMOT [54] resolves label assignment disparities, enabling a tracking system that operates without periodic renewal or re-initialization. LAID [25] follows the TbD pipeline by using a frozen detector and adding an association network to specifically optimize association and minimize competition. MOTIP [19] redefines the association task as the ID prediction identification. Additionally, some methods leverage extended temporal information to improve performance. MeMOT [7] utilizes a tracking memory bank, while MeMOTR [18] employs long-term memory through a customized memory-attention layer to maintain stable IDs.

Despite these advancements, these methods remain within the DETR framework, embedding association within the same decoder that handles both historical information and new trajectory detection. In contrast, LA-MOTR decouples the entire process into separate network modules, thereby preserving the robust functionalities of object detection and specialized association components.

3. Method

We propose **LA-MOTR**, a novel end-to-end framework with learnable association. As illustrated in Figure 2, our method divides the tracking process into three specialized components: Separated Object-Tracklet Detection (Section 3.1), Spatial-Guided Learnable Association (Section 3.2), and Tracklet Update (Section 3.4). At each frame t, the captured image $I_t \in \mathbb{R}^{W \times H \times 3}$ is processed to extract image features F_t. These features, along with blank object queries Q^{obj} and previous tracklet queries Q_{t-1}^{tck}, are input into a decoder to generate corresponding detection embeddings E^{obj} and E_t^{tck}. Subsequently, the Separated Object-Tracklet Detection component utilizes spatial cues from E^{obj} and E_t^{tck} to construct Spatial-Guided Attention, guiding embedding features interactions. This process generates the outputs O^{obj}, O_t^{tck}, and an association matrix M. Finally, the tracklet query output O_t^{tck} serves as the detection for existing tracklets in the current frame, while the object query output O^{obj} contains the detection results for all targets (see Section 3.3). The association matrix M, which matches the object and tracklet query outputs, is used to update the tracklet queries Q_t^{tck} for the next frame and to identify newly appeared targets in the current frame.

3.1. Separated Object-Tracklet Detection

In the Separated Object-Tracklet Detection module, the image I_t at the t-th frame is first processed by a ResNet-

50 [21] backbone and a transformer encoder [31] to extract the image features F_t. These features serve as keys and values in a cross-attention mechanism. The tracklet queries Q^{tck}_{t-1} from the previous frame are concatenated with the object queries Q^{obj} to form the input queries. This architecture produces two types of detection embeddings: E^{tck}_t for tracklets and E^{obj} for objects. In contrast to end-to-end models like the decoder in MOTR [58], which supervise object queries only with newly detected objects, our approachs object queries include detections for all objects in the current frame to facilitate better feature interaction and association, as illustrated in Figure 2(a).

Separating tracklet queries and object queries to obtain detection embeddings for existing tracklets and all targets in the current frame forms the foundation of the new framework. This method effectively captures contextual information and the relationships between the two types of outputs, providing reliable embeddings for subsequent learnable association. These embeddings are crucial for accurate detection and robust tracking over time.

3.2. Spatial-Guided Learnable Association

After obtaining the detection embeddings for historical tracklets and all objects, we perform implicit association to acquire the interaction embeddings and association relationships. To achieve this, we propose the Spatial-Guided Learnable Association (SGLA), a learned association module based on the Edge-Augmented Graph Transformer[13, 24], to obtain the association matrix. SGLA leverages Relative-Spatial Cues between tracklets and objects to enhance the association matrix while simultaneously generating weights for Spatial-Weighted Attention. The Spatial-Weighted Attention within SGLA comprises a multi-layer structure, where the association matrix $M_{(l)}$ from each layer l is used as input for the subsequent layer. In the initial layer, $O^{\text{obj}}_{(0)}$ and $O^{\text{tck}}_{(0)}$ represent the object detection embedding $E^{\text{obj}} \in \mathbb{R}^{N_d \times d}$ and the tracklet detection embedding $E^{\text{tck}}_t \in \mathbb{R}^{N_t \times d}$ from the Separated Object-Tracklet Detection Module, respectively. Additionally, the association matrix $M_{(0)} \in \mathbb{R}^{N_d \times N_t \times d}$ is initialized as a zero matrix.

Relative-Spatial Cue. For each layer l, the object embeddings $O^{\text{obj}}_{(l)}$ and tracklet embeddings $O^{\text{tck}}_{(l)}$ are processed through a detection head to generate bounding boxes $B^{\text{obj}}_{(l)}$ and $B^{\text{tck}}_{(l)}$, as illustrated in Figure 3. This design enables direct extraction of spatial cues within both object and historical tracklet embeddings. Notably, the head is shared with the one used in Section 3.3.

Subsequently, the box differences $B^{diff}_{(l)}$ are calculated by determining the pairwise differences between $B^{\text{obj}}_{(l)}$ and $B^{\text{tck}}_{(l)}$. The aggregated box differences $B^{diff}_{(l)} \in \mathbb{R}^{N_d \times N_t \times 4}$ are embedded using an MLP, generating the relative deviation encoding $E^{diff}_{(l)} \in \mathbb{R}^{N_d \times N_t \times d}$, which encodes the relative spatial relationships between the boxes.

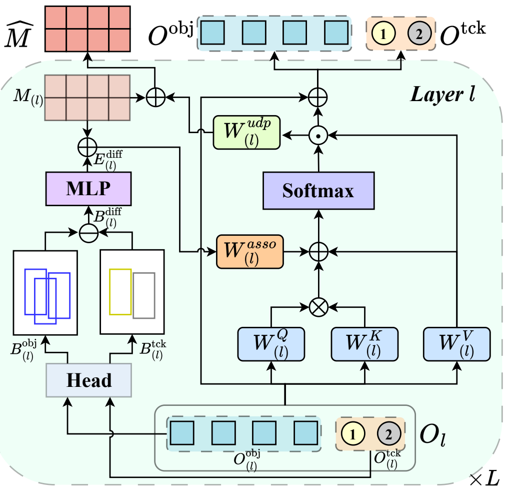


Figure 3. The Spatial-Weighted Attention consists of L layers. Each layer receives outputs from the preceding layer, $O^{\text{obj}}_{(l)}$, $O^{\text{tck}}_{(l)}$, and $M_{(l)}$. This design ensures relative spatial cues within each layer can effectively guide the association matrix $M_{(l)}$.

Spatial-Weighted Attention. In Figure 3, we improve the self-attention mechanism by incorporating relative spatial cues between objects and tracklets as weights to guide the association matrix M. For each layer l, the object output $O^{\text{obj}}_{(l)}$ and the tracklet output $O^{\text{tck}}_{(l)}$ are concatenated into a unified representation $O_{(l)} \in \mathbb{R}^{(N_d+N_t)\times d}$. We add $E^{\text{diff}}_{(l)}$ to the association matrix $M_{(l)}$ and integrate it into the self-attention computation to produce the attention matrix $A_{(l)}$, defined as

$$A_{(l)} = \text{softmax}\left(\frac{(O_{(l)}W^{Q}_{(l)})(O_{(l)}W^{K}_{(l)})^{\top}}{\sqrt{d_k}} + M_{(l)}W^{\text{asso}}_{(l)}\right).$$

This integration utilizes the spatial cue to strengthen contextual relationships between objects and tracklets. The computed attention matrix $A_{(l)}$ is applied to the value matrix and combined with $O_{(l)}$ through a residual connection, producing the output for the next layer: $O_{(l+1)} = O_{(l)} + A_{(l)}(O_{(l)}W^{V}_{(l)})$. Furthermore, we dynamically update the association matrix as $M_{(l+1)} = M_{(l)} + A_{(l)}W^{\text{upd}}_{(l)}$, enabling iterative refinement of the associations between objects and tracklets. This closed-loop design ensures spatial and temporal consistency across layers.

The Spatial-Guided Learnable Association design exploits spatial consistency across consecutive frames by utilizing relative spatial positions as weighting factors. This promotes bidirectional information exchange between object and tracklet embeddings via the Spatial-Weighted Attention module. As shown in Figure 2(a), tracklet embeddings acquire additional features from object embeddings for targets lost in prior frames, enhancing robustness to temporary occlusions and reappearances. Simultaneously, ob-

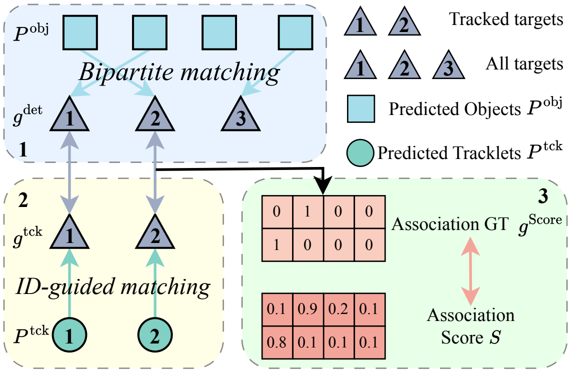


Figure 4. **Details of Decoupled Tracking-Detection Loss.** Steps 1, 2, and 3 respectively describe the matching processes for calculating the detection loss, tracking loss, and association loss.

ject queries benefit from the temporal information in tracklet queries, improving detection accuracy as illustrated in Figure 6(a).

3.3. Decoupled Tracking-Detection Loss

After obtaining the object output O^{obj}, the track output O_t^{tck}, and the association matrix $\hat{M}$, the detection head first produces the predicted object detections P^{obj} and the predicted tracklets P^{tck}. A sigmoid function is then applied to the association matrix $\hat{M}$ to calculate the association scores S for all tracklet-detection pairs.

As illustrated in Figure 4, objects P^{obj} are matched with existing targets in the current frame using the Hungarian algorithm [27], establishing the detection ground truth g^{det}. Similarly, predicted tracklets P^{tck} are matched with historical tracklets by their identifiers to determine the tracklet ground truth g^{tck}. If a target matches both a detection and a track, the association score is set to 1; otherwise, it is set to 0, resulting in the ground truth association score g^{score}. Unmatched detections are then initialized as new trajectories. During inference, detected objects and tracklets are directly matched using association scores S. Objects with scores above 0.5 are initialized as new trajectories, defining the trajectories for the current frame. Detailed inference procedures are provided in the supplementary materials.

For a training sequence comprising T frames, the total loss is decomposed into detection loss $\mathcal{L}_{\text{det}}$, track loss $\mathcal{L}_{\text{tck}}$, and association loss $\mathcal{L}_{\text{asso}}$. Both $\mathcal{L}_{\text{det}}$ and $\mathcal{L}_{\text{tck}}$ consist of classification loss $\mathcal{L}_{\text{cls}}$ (Focal Loss [41]), regression loss $\mathcal{L}_{\text{reg}}$ (L1 loss [37]), and bounding box loss $\mathcal{L}_{\text{box}}$ (GIoU loss [38]). The association loss $\mathcal{L}_{\text{asso}}$ is computed using binary cross-entropy [42] on the two score matrices. The overall loss for the training sequence is formulated as

$$\mathcal{L} = \sum_{t=1}^{T} (\lambda_{\text{det}} \mathcal{L}_{\text{det}}^{t} + \lambda_{\text{tck}} \mathcal{L}_{\text{tck}}^{t}) + \sum_{t=2}^{T} \lambda_{\text{asso}} \mathcal{L}_{\text{asso}}^{t} \qquad (1)$$

3.4. Tracklet Update

In previous Tracking-by-Attention methods, the tracklet query for the next frame, Q_{t+1}^{tck}, is typically updated by introducing various learnable networks between frames or across multiple frames. This strategy aims to ensure continuity in the target feature representation and maintain long-term memory. In contrast, our method derives the tracklet features through Spatial-Guided Learnable Association, effectively integrating them with the feature representations of all objects in the current frame. Consequently, we only need to ensure the continuity of inter-frame features, as the features of all current tracklet can be fused with all objects from the association process. Therefore, we utilize only the object outputs O^{obj} and tracklet outputs O_t^{tck} from the current frame to generate the tracklet queries for the next frame:

$$Q_{t+1}^{\text{tck}} = w \cdot O_t^{\text{tck}} + (1 - w) \cdot O^{\text{obj}} \qquad (2)$$

As illustrated in Figure 6(a), the tracklet query results demonstrate high accuracy, enabling effective object detection even in crowded or motion-blurred scenarios. This method not only reduces model complexity but also maintains tracking continuity without adding additional learnable parameters, as shown in Table 6.

4. Experiments

4.1. Datasets and Metrics

Datasets. We evaluate our method primarily on the Dancetrack [47], SportsMOT [11], MOT17 [36]. Dancetrack [47] focuses on dance performances, featuring over 105k annotated frames. The rapid movements of dancers with similar appearances significantly increase the difficulty of the tracking task. SportsMOT [11] consists of approximately 150K annotated frames capturing various sports events, providing a rich set of scenarios with fast-moving athletes and high-speed cameras. MOT17 [36] is a widely used benchmark in multi-object tracking characterized by crowded scenes and challenging occlusions.

Evaluation Metrics. We assess the performance of multi-object tracking systems using Higher Order Tracking Accuracy (HOTA) [33], which evaluates detection accuracy, association accuracy, and temporal consistency. Additionally, we follow standard evaluation protocols by employing Multiple-Object Tracking Accuracy (MOTA) [3], Identity Switches (IDS) [2], and Identity F1 Score (IDF1) [40].

4.2. Implementation Details

LA-MOTR employs Dab-Deformable-DETR [31] with a ResNet50 [21] backbone. The model is initialized using the official pre-trained weights from the COCO [30] dataset. In addition, we apply data augmentation techniques such as random scaling and cropping. Specifically, we resize input images so that the shorter side is 800 pixels and the longer side does not exceed 1536 pixels.

As discussed in Section 3.4, training track query updates over the long term is unnecessary. Instead, we train LA-MOTR using fixed three-frame clips as batches across the three datasets. Within each clip, input frames are sampled at random intervals ranging from 1 to 10. For the DanceTrack dataset [47], we train for 18 epochs, reducing the learning rate by a factor of ten at the 12th epoch. Similarly, training on SportsMOT [11] is conducted over 28 epochs, with the learning rate decreased at the 18th epoch. Following MeMOTR [18], we combine the CrowdHuman [44] validation set with the MOT17 [36] training data to create an augmented training set, which helps prevent overfitting due to limited data. We train the model on this set for 130 epochs, reducing the learning rate at the 120th epoch.

All experiments were performed using PyTorch on four NVIDIA 3090 GPUs. The batch size was set to one per GPU, with each batch comprising a video clip. We employed the AdamW [32] optimizer with an initial learning rate of 2.0×10^{-4}. During training, tracked targets with detection scores below the threshold $\tau_{\text{det}} = 0.5$ and IoU below $\tau_{\text{IoU}} = 0.5$ were excluded. The weight parameter w in the tracklet query update is set to 1, 0.9, and 0.5 for the DanceTrack, SportsMOT, and MOT17 datasets, respectively. The loss coefficients are set as follows: $\lambda_{\text{cls}} = 2$, $\lambda_{\text{reg}} = 5$, $\lambda_{\text{box}} = 2$, and $\lambda_{\text{asso}} = 10$. As shown in Table 7, LA-MOTR maintained network parameters and speed despite the inclusion of the learnable association module compared to other end-to-end methods.

4.3. Comparison with State-of-the-art Methods

We compare LA-MOTR with major end-to-end online methods, including TransTrack [46], MOTR [58], MeMOTR [18], Co-MOT [54], MOTRv2 [62], MOTRv3 [56], MOTIP [19], and LAID [25]. Additionally, we evaluate LA-MOTR against state-of-the-art traditional two-stage online methods such as QDTrack [17], FairMOT [60], CenterTrack [64], ByteTrack [61], and OC-SORT [8].

Dancetrack dataset. Target movements in Dancetrack [47] are nonlinear and irregular, frequently experiencing occlusions and blurring. Consequently, associating targets with tracklets presents a significant challenge. As shown in Table 1, we trained LA-MOTR on the DanceTrack training set and compared it with state-of-the-art methods on the DanceTrack test set [47]. LA-MOTR demonstrates competitive performance across various metrics. Compared to OC-SORT, LA-MOTR improves HOTA, DetA, and AssA by 17.0%, 4.1%, and 20.7%, respectively. These results highlight the effectiveness of LA-MOTR as an end-to-end object tracking approach in complex scenarios. Furthermore, when compared to other end-to-end methods such as MeMOTR, LA-MOTR exhibits significant advantages, including a 1.8% increase in HOTA. Additionally, we implemented LA-MOTR with the standard Deformable-DETR backbone, achieving a HOTA score of 66.5%.

Method	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
Two-Stages Methods					
FairMOT[60]	39.7	82.2	40.8	23.8	66.7
CenterTrack[64]	41.8	86.8	35.7	22.6	78.1
ByteTrack[61]	47.7	89.6	53.9	32.1	71.0
QDTrack[17]	54.2	87.7	50.4	36.8	80.1
OC-SORT[8]	55.1	92.0	54.6	38.3	80.3
End-to-End Methods					
TransTrack[46]	45.5	88.4	45.2	27.5	75.9
MOTR*[58]	54.2	79.7	51.5	40.2	73.5
MeMOTR*[18]	63.4	85.4	65.5	52.3	77.0
MeMOTR[18]	68.5	89.9	71.2	58.4	80.5
CO-MOT*†[54]	69.4	91.2	71.9	58.9	82.1
MOTRv2*†[62]	69.9	91.9	71.7	59.0	<u>83.0</u>
MOTRv3*†[56]	<u>70.4</u>	<u>92.9</u>	72.3	59.3	83.8
MOTIP*[19]	67.5	90.3	72.2	57.6	79.4
MOTIP[19]	70.0	91.0	**75.1**	**60.8**	80.8
LAID[25]	69.6	89.9	<u>73.5</u>	<u>59.9</u>	81.8
LA-MOTR*	66.5	92.5	69.7	53.9	80.4
LA-MOTR	**71.1**	**93.5**	71.8	59.0	**84.4**

Table 1. **Tracking results on the DanceTrack test set.** Results for existing methods are sourced from DanceTrack. * indicates the use of the standard Deformable-DETR as the backbone, and † denotes the use of an additional data. The best performance is presented in **bold**, and the second-best performance is <u>underlined</u>.

Our proposed Spatial-Guided Learnable Association effectively integrates relative spatial cues, achieving association performance with 93.5% MOTA and 59.0% AssA. Importantly, unlike some methods [56, 62] that utilize an additional detector, LA-MOTR achieves state-of-the-art performance in DetA by reaching 84.4%, demonstrating its effectiveness without requiring auxiliary detection.

MOT17 dataset. The MOT17 dataset serves as a widely recognized benchmark for pedestrian detection and tracking, characterized by densely packed targets and linear motion patterns. As previously discussed, TbA methods face challenges in effectively balancing detection and tracking. In contrast, TbD approaches employ more robust detectors and accurately predict target positions, resulting in improved performance. LA-MOTR addresses these challenges by decoupling detection and tracking tasks. This architecture enables the model to leverage advanced detection capabilities while maintaining effective tracking performance. On the MOT17 benchmarks, LA-MOTR achieves a 57.4% HOTA and 75.3% MOTA when trained only on the MOT17 training set, surpassing multiple state-of-the-art end-to-end MOT approaches, as demonstrated in Table 2.

However, Transformer-based methods often overfit the MOT17 dataset due to its limited training set of approxi-

Method	Add. Data	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
Two-Stages Methods						
FairMOT[60]	CH,CP,ETHZ	59.3	73.7	72.3	58.0	60.9
CenterTrack[64]	CH	52.2	67.8	64.7	51.0	53.8
ByteTrack[61]	CH,CP,ETHZ	63.1	80.3	77.3	62.0	64.5
QDTrack[17]	CH	63.5	77.5	78.7	64.5	62.6
OC-SORT[8]	CH,CP,ETHZ	63.2	78.0	77.5	63.4	63.2
End-to-End Methods						
TransTrack[46]	CH	54.1	74.5	63.9	47.9	61.6
MOTR*[58]	CH	57.8	73.4	68.6	55.7	60.3
Co-MOT*[54]	✗	60.1	72.6	72.7	60.6	59.5
MOTRv2*[62]	CH	57.6	70.1	70.3	57.5	58.1
MOTRv2*[62]	CH	62.0	78.6	75.0	60.6	63.8
MeMOTR*[18]	CH	58.8	72.8	71.5	58.4	59.6
MOTRv3*[56]	CH	60.2	75.9	72.4	58.7	62.1
MOTIP[19]	CH	59.2	75.5	71.2	56.9	62.0
LA-MOTR*	✗	53.8	72.1	67.4	60.5	53.9
LA-MOTR*	CH	60.8	79.8	72.2	63.3	59.2
LA-MOTR	✗	57.4	75.3	70.1	61.7	58.0
LA-MOTR	CH	**62.6**	**80.7**	**73.8**	**63.9**	**62.9**

Table 2. **Tracking results on the MOT17 test set.** Additional training datasets include CrowdHuman[44] (CH), CityPersons[59] (CP), and ETHZ[45]. The best and second performance among the End-to-End methods is marked in **bold** and underline. Gray color indicates method with offline post-processing, which are excluded from comparison. * indicates the use of the standard Deformable-DETR as the backbone.

mately 5K frames [18]. To mitigate this, we incorporated the CrowdHuman validation set (about 4.37K frames) into the training process, resulting in a 62.6% increase in the HOTA score. This improvement suggests that the small size of MOT17 contributes to lower performance compared to TbD approaches. Nevertheless, our results demonstrate that LA-MOTR competes effectively with advanced TbD methods that use sophisticated detectors, highlighting its potential for robust multiple object tracking.

SportsMOT dataset. To further demonstrate the association capabilities of LA-MOTR, we evaluate our method on SportsMOT, where the targets exhibit highly similar appearances and move simultaneously with the camera. Notably, all models are trained only on the SportsMOT training set. The results demonstrate that our method achieves impressive performance with a HOTA score of 72.4%, particularly excelling in the AssA metric with a score of 61.8%.

Method	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
Two-Stages Methods					
FairMOT[60]	49.3	86.4	53.5	34.7	70.2
CenterTrack[64]	62.7	90.8	60.0	48.0	82.1
ByteTrack[61]	62.8	94.1	69.8	51.2	77.1
OC-SORT[8]	71.9	94.5	72.2	59.8	86.4
QDTrack [17]	60.4	90.1	62.3	47.2	77.5
DiffMOT[34]	72.1	94.5	72.8	60.5	86.0
Deep-IOU[23]	74.1	95.1	75.0	63.1	87.2
End-to-End Methods					
TransTrack*[46]	68.9	92.6	71.5	57.5	82.7
MeMOTR*[18]	68.8	90.2	69.9	57.8	82.0
MeMOTR[18]	70.0	91.5	71.4	59.1	83.1
SambaMOTR*[43]	69.8	90.3	71.9	59.4	82.2
MOTIP[19]	71.9	92.9	**75.0**	62.0	83.4
LAID[25]	71.7	89.2	72.7	**62.4**	82.5
LA-MOTR*	69.5	93.5	70.0	57.9	84.4
LA-MOTR	**72.4**	**95.6**	73.3	61.8	**86.6**

Table 3. **Tracking results on the SportsMOT test set.** The best and second performance among the End-to-End methods is marked in **bold** and underline. * indicates the use of the standard Deformable-DETR as the backbone Results for existing methods are sourced from SportsMOT.

4.4. Ablation Study

This section analyzes key components of our pipeline: the association method, tracklet query updates, and training clip length. We perform ablation experiments on the DanceTrack [47] dataset, which includes challenging scenarios such as severe target motion blur, occlusions, and difficult object associations. The model is trained on the training set and evaluated on the validation set.

Association Methods. Table 4 presents an ablation study comparing various association methods following object and tracklet detections. The methods are categorized into hand-crafted and end-to-end approaches, each retrained to ensure a fair comparison. Hand-crafted methods combine FairMOT's[60] detection and classification heads with LA-MOTR's SOTD to obtain bounding boxes and appearance features. For association, ByteTrack's[61] strategy is employed, utilizing IoU, appearance similarity, or their combination.

Association	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
Hand-Craft					
IoU	49.864	73.729	50.522	37.642	66.458
Feature	46.977	70.211	47.210	34.509	64.336
Combination	57.920	81.970	58.866	45.076	73.397
End-to-End					
Single Decoder	60.220	86.075	61.648	47.509	75.994
SGLA **w/o** RSC	62.753	88.237	64.651	49.084	78.992
SGLA	**64.474**	**90.537**	**66.956**	**51.913**	**80.366**

Table 4. **Ablation Experiments on Association Methods.** IoU refers to matching based on bounding box IoU, Feature refers to matching based on feature representations.

q	k, v	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
E^{obj}	E_t^{tck}	63.896	89.374	66.483	50.397	**80.672**
E_t^{tck}	E^{obj}	64.008	90.280	67.194	51.779	80.201
concat		**64.474**	**90.537**	**66.956**	**51.913**	80.366

Table 5. **Ablation Experiments on attention mechanism in SGLA.** We evaluate using E^{obj} and E_t^{tck} as queries, keys, and values in cross-attention, and their concatenation for self-attention.

End-to-end methods include using MOTR with a single

Query Update	HOTA↑	MOTA↑	IDF1↑	AssA↑	DetA↑
TAN[58]	62.740	88.994	64.608	51.694	79.376
TIM[18]	**64.589**	**91.074**	66.910	**52.159**	79.831
LA-MOTR	64.474	90.537	**66.956**	51.913	**80.366**

Table 6. **Ablation Experiments on Tracklet Query Update.** TAN and TIM are learning-based methods for updating tracklet queries.

decoder for detection and matching, as well as our proposed Spatial-Guided Learnable Association (GLA). Additionally, we assess the impact of removing Relative Spatial Cues (RSC). The results indicate that end-to-end association methods outperform hand-crafted approaches. Compared to single-decoder models, our decoupled method, LA-MOTR, achieves a HOTA improvement of 64.474%. Furthermore, incorporating RSC enhances the interaction between objects and tracklets, resulting in an additional 1.721% increase in HOTA.

Attention Mechanism in SGLA. We perform ablation experiments on the Spatial-Guided Learnable Association (SGLA) attention mechanism, as shown in Table 5. Performance was evaluated using E^{obj} and E_t^{tck} in various combinations of queries, keys, and values within cross-attention modules. Additionally, we replaced cross-attention with concatenated E^{obj} and E_t^{tck} in self-attention modules to assess their combined effects. The first row results show that integrating E^{obj} significantly improves DetA, while the second row indicates that E_t^{tck} substantially enhances association. These findings demonstrate that E^{obj} and E_t^{tck} represent the model's detection and association capabilities, respectively. This validates our choice of using self-attention for comprehensive interaction on both embeddings, achieving a HOTA score of 64.474%.

Tracklet Query Update Method. We evaluated various track query update strategies through ablation experiments (Table 6). TAN [58] with a learned update achieved a HOTA of 62.740%, while TIM [18], which integrates long-term memory features, attained 64.589%. Our proposed method (Section 3.4) directly combines tracklet and object outputs for the next frame without additional learning mechanisms, resulting in a HOTA of 64.474% and a higher DetA of 80.366%. These results demonstrate that LA-MOTR's association transformer effectively facilitates information interaction among current frame targets and efficiently performs feature association without relying on long-term strategies.

Tracklet Query Update Weight. We conducted an ablation study on the weight parameter w used in the query update process (Section 3.4). We sampled 10 uniformly distributed values of w between 0 and 1, trained all models on three datasets in Figure 5. The results indicate that the optimal w is dataset-dependent. For datasets with complex target motion, such as DanceTrack and SportsMOT, the best

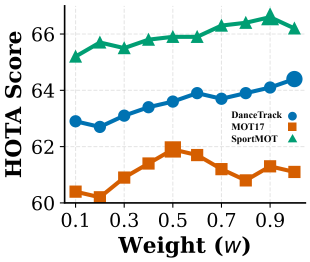


Figure 5. Ablation study on tracklet update weight w.

Method	Params	FPS
MOTR[58]	43.91M	13.47
MeMOTR[18]	50.13M	12.33
LA-MOTR	50.29M	11.96

Table 7. Network Parameters and Inference Frame Per Second of End-to-End MOT Methods.

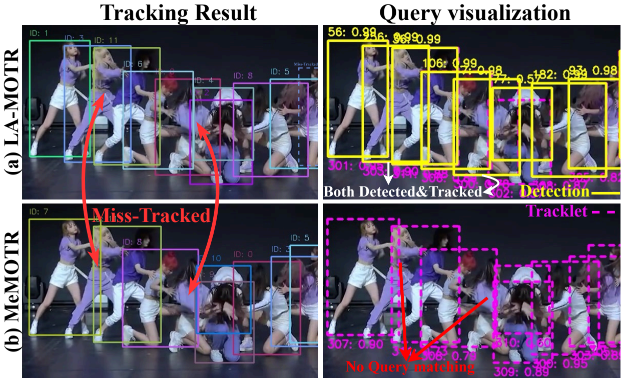


Figure 6. **Visualization of LA-MOTR and MeMOTR under occlusion.** MeMOTR loses two trajectories and fails to detect them through query decoding, whereas LA-MOTR successfully tracks all trajectories with corresponding object and tracklet queries.

w values are 1 and 0.9, respectively. In contrast, for linear motion datasets like MOT17, an optimal w of 0.5 was identified. This suggests that for more nonlinear target motions, the model relies more on historical tracklet features, whereas for linear motion, the current frame object features provide sufficient information to predict the next frame's tracklets.

5. Conclusion

We introduce LA-MOTR, a pioneering end-to-end framework that integrates a learnable association module to separately manage object and tracklet queries. This separation effectively alleviates the functional ambiguities present in existing Tracking-by-Attention approaches. Comprehensive experiments conducted on challenging datasets demonstrate that our method outperforms current end-to-end multi-object tracking techniques, achieving state-of-the-art performance.

Limitations and Future Work. Ablation studies indicate that the long-term memory component in track query updates enhances MOT accuracy. In future work, we will explore learnable associations across additional dimensions and develop improved query update mechanisms to more effectively preserve essential temporal information.

Acknowledgments

Dr. Li is supported in part by the National Natural Science Foundation of China Grant No. 12071478. Dr. Wang is supported in part by the National Natural Science Foundation of China Grant No. 61972404, Public Computing Cloud, Renmin University of China, and the Blockchain Lab, School of Information, Renmin University of China.

References

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

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 1, 2

[5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017. 1

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[7] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8100, 2022. 3

[8] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9686–9696, 2023. 2, 6, 7

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2

[10] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 215–230. Springer, 2012. 1

[11] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9921–9931, 2023. 1, 2, 5, 6

[12] P Dendorfer. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2

[13] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023. 4

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

[15] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 1

[16] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[17] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6, 7

[18] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9901–9910, 2023. 2, 3, 6, 7, 8

[19] Ruopeng Gao, Yijun Zhang, and Limin Wang. Multiple object tracking as id prediction. *arXiv preprint arXiv:2403.16848*, 2024. 3, 6, 7

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

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 4, 5

[22] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004. 1

[23] Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 163–172, 2024. 7

[24] Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 655–665, 2022. 4

[25] Shukun Jia, Yichao Cao, Feng Yang, Xin Lu, and Xiaobo Lu. Multi-object tracking by detection and query: an efficient end-to-end manner. *arXiv preprint arXiv:2411.06197*, 2024. 3, 6, 7

[26] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 1, 2

[27] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[28] L Leal-Taixé. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2

[29] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 1, 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[31] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 1, 4, 5

[32] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[33] 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. *International journal of computer vision*, 129:548–578, 2021. 5

[34] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19321–19330, 2024. 7

[35] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[36] Anton Milan. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 5, 6

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 5

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1, 5

[39] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 1

[40] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 5

[41] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 5

[42] Usha Ruby and Vamsidhar Yendapalli. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, 9(10), 2020. 5

[43] Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Bernt Schiele, and Luc Van Gool. Samba: Synchronized set-of-sequences modeling for multiple object tracking. *arXiv preprint arXiv:2410.01806*, 2024. 1, 7

[44] 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. 6, 7

[45] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008. 7

[46] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 1, 2, 6, 7

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 1, 2, 5, 6, 7

[48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[49] Peng Wang, Yongcai Wang, and Deying Li. Dronemot: Drone-based multi-object tracking considering detection difficulties and simultaneous moving of drones and objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7397–7404. IEEE, 2024. 2

[50] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5740–5748, 2024. 1

[51] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European conference on computer vision*, pages 107–122. Springer, 2020. 2

[52] Longyin Wen, Pengfei Zhu, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Jiayu Zheng, Tao Peng, Xinyao Wang,

Yue Zhang, et al. Visdrone-mot2019: The vision meets drone multiple object tracking challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[53] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2

[54] Feng Yan, Weixin Luo, Yujie Zhong, Yiyang Gan, and Lin Ma. Bridging the gap between end-to-end and non-end-to-end multi-object tracking. *arXiv preprint arXiv:2305.12724*, 2023. 2, 3, 6, 7

[55] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6504–6512, 2024. 2

[56] En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023. 2, 3, 6, 7

[57] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1

[58] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 1, 2, 4, 6, 7, 8

[59] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 7

[60] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021. 1, 2, 6, 7

[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 *European conference on computer vision*, pages 1–21. Springer, 2022. 2, 6, 7

[62] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 2, 3, 6, 7

[63] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1

[64] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 6, 7

[65] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2