MoTRa: Motion-Aware Target Representation Learning for End-to-End Multi-Object Tracking

Yuanzhou Huang, **Songwei Pei***, **Shuhuai Wang**, **Bingfeng Liu**, **Qian Li**, **Shangguang Wang**
School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing, China
{huangyuanzhou, peisongwei, wangshuhuai, bfliu, li.qian, sgwang}@bupt.edu.cn

Abstract

Multi-object tracking (MOT) has long faced challenges with identity switches, especially for targets with low appearance discriminability and complex motion. Existing end-to-end trackers typically enhance robustness by modeling long-term temporal information across target-level representations, yet these representations remain insufficiently discriminative for spatially and visually similar targets. In this paper, we present MoTRa, a Motion-aware Target Representation learning framework for end-to-end MOT that enriches target-level representations with adaptive motion cues. As a result, each representation adaptively balances motion cues and appearance-related content cues under varying conditions, enhancing its discriminability for tracking. To achieve this, we propose a feature fusion and alignment module that extracts motion cues and adaptively fuses them with content features into target representations. To further regularize the dynamically fused representations, we propose a target-specific contrastive learning strategy that promotes intra-trajectory consistency and inter-target separability. Experimental results demonstrate the effectiveness of our method, achieving competitive performance on multiple benchmarks.

1 Introduction

Multi-object tracking (MOT) aims to localize all relevant targets in each video frame and assign consistent identities (IDs) over time, enabling applications such as pedestrian analytics [Zhang *et al.*, 2024], autonomous driving [Li *et al.*, 2024], and situational awareness [Liu *et al.*, 2025]. However, achieving reliable identity association remains challenging due to highly similar targets and frequent occlusions.

Most existing MOT methods follow the tracking-by-detection paradigm [Shim *et al.*, 2025], typically coupling an off-the-shelf detector with motion- or appearance-based association strategies. While these approaches achieve impressive performance, their decoupled design and separate optimization make it difficult to jointly model unified target representations and motion dynamics, thereby limiting long-term tracking robustness in challenging scenarios [Zeng *et al.*, 2022]. Recently, end-to-end MOT methods have emerged. By unifying detection and association within Transformer-based architectures, these methods can better learn target representations and capture long-term temporal dependencies for robust tracking. Most prior methods adopt the tracking-by-propagation paradigm [Zeng *et al.*, 2025], where tracked targets are represented by autoregressive track queries propagated across frames to localize their counterparts in the current frame. In contrast, recent work such as MOTIP [Gao *et al.*, 2025] formulates MOT as an in-context ID prediction task, further simplifying the tracking pipeline by binding ID information to each target representation, enabling direct ID decoding for data association. Overall, these advances in end-to-end methods demonstrate that better modeling and use of target representations has strong potential for robust tracking in complex scenarios. Nevertheless, current end-to-end methods still struggle to produce sufficiently discriminative target representations, particularly in challenging situations involving severe occlusions and complex interactions.

Building upon DETR-like object detection frameworks

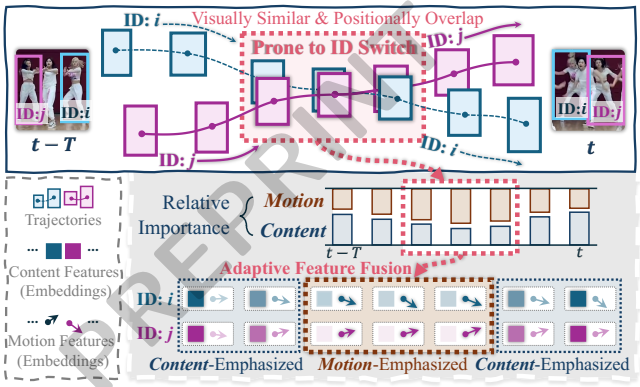

Figure 1: Illustration of Motion-Aware Target Representation. Unlike most end-to-end MOT methods that directly reuse detection output embeddings as target-level features, MoTRa augments these embeddings by adaptively fusing motion cues with content cues that capture appearance-related information. This enables the model to focus on the most discriminative cues under different tracking conditions and better distinguish targets in challenging scenarios.

*Corresponding author.

[Carion *et al.*, 2020], recent end-to-end MOT methods typically adopt DETR output embeddings as target-level representations for association, since these embeddings inherently encode positional and appearance information. However, as illustrated in Figure 1, when targets experience complex interactions and occlusions, their output embeddings can become highly similar due to visual and positional similarity, increasing the risk of ID switches and compromising tracking robustness. To address this problem, recent solutions enhance these representations with temporal information [Gao and Wang, 2023], yet the target-level representations remain detection-oriented and limited in discriminability, thereby constraining overall tracking performance. Although introducing additional tracking cues is a natural way to strengthen target representations, naively fusing heterogeneous cues can dilute discriminative signals and hinder reliable association. This leads to a pivotal question: *Is it possible to adaptively fuse multiple tracking cues so that the model can emphasize the most discriminative cues under varying tracking conditions?*

Motivated by the above perspective, we propose MoTRa, an end-to-end MOT method towards learning more discriminative target representations for robust tracking, as shown in Figure 1. Since DETR output embeddings implicitly encode distinct cues, explicitly enhancing specific cues under different tracking conditions becomes challenging. We therefore decouple the output embeddings into content embeddings and positional information, and further augment the latter with motion trends to form motion embeddings. Here, the content embedding encodes semantic information of the target's appearance and identity, whereas the motion embedding captures inter-frame variations in spatial position. Afterwards, we introduce a motion-aware feature fusion and alignment module (MFA), in which an adaptive gating mechanism fuses motion and content embeddings to emphasize discriminative cues under different tracking conditions. To alleviate the feature inconsistency induced by such dynamic fusion, we further design a target-specific contrastive learning strategy (TSCL), which promotes intra-trajectory consistency and inter-target separability for more stable and discriminative representations. Overall, MoTRa preserves the simplicity of end-to-end architectures while explicitly leveraging readily available motion and content cues to learn tracking-oriented target representations, thereby enhancing target discriminability and achieving impressive performance on DanceTrack [Sun *et al.*, 2022], SportsMOT [Cui *et al.*, 2023], and BFT [Zheng *et al.*, 2024]. Extensive ablation studies validate the effectiveness of the proposed modules. The major contributions of this work are as follows:

- We propose MoTRa, an end-to-end MOT framework that learns discriminative, tracking-oriented target representations for robust tracking in complex scenarios.
- We design an MFA module for adaptive fusion of motion and content cues under varying tracking conditions, and introduce a TSCL strategy to enhance intra-trajectory consistency and inter-target separability.
- Extensive experiments demonstrate that MoTRa enhances tracking performance and achieves competitive results on DanceTrack, SportsMOT, and BFT.

2 Related Works

Tracking-by-Detection. Most existing MOT methods follow the tracking-by-detection (TbD) paradigm, which consists of a detection stage followed by an independent association stage. In the association stage, early TbD methods [Bewley *et al.*, 2016] primarily leverage motion cues through position prediction and IoU-based matching [Yang *et al.*, 2023]. Later works [Wojke *et al.*, 2017] further refine this by incorporating appearance cues with ReID models for similarity matching. To improve matching robustness, recent approaches refine this process by integrating motion and appearance [Seidenschwarz *et al.*, 2023] and by incorporating additional tracking cues [Yang *et al.*, 2024]. However, these attempts still rely on hand-crafted, similarity-based matching heuristics instead of learning a unified target representation, making it difficult to exploit temporal context for robust tracking, especially under high motion complexity and low appearance discriminability [Qin *et al.*, 2024].

End-to-End Tracking. Most prior methods adopt the Tracking-by-Propagation (TbP) framework [Zeng *et al.*, 2022], which extends DETR-like detectors [Zhu *et al.*, 2020] to tracking by using learnable detect queries and propagated track queries to identify new and existing targets, respectively. Specifically, track queries are derived from previous-frame output embeddings [Meinhardt *et al.*, 2022], which are optimized for detection and encode target positional and appearance information. To improve long-term tracking, follow-up works enrich track queries with long-term motion patterns via temporal aggregation modules [Gao and Wang, 2023]. However, TbP still suffers from conflicts caused by the unified processing of the two query types [Huang *et al.*, 2025b]. The recent ID-prediction framework [Gao *et al.*, 2025] alleviates this by decoupling detection and association, where a DETR-like detector first detects objects and obtains their output embeddings, which are then combined with learnable ID tokens for ID prediction. Compared with TbP, this design simplifies the pipeline and allows longer-term historical information to be exploited for association. In this paper, MoTRa follows the ID-prediction paradigm, but learns more discriminative, tracking-oriented representations for association. Rather than directly reusing detection embeddings, it explicitly integrates motion and content cues and emphasizes the most informative cues according to each target's evolving tracking condition, thereby improving tracking performance in complex scenarios.

3 Methods

Most previous Transformer-based end-to-end MOT methods leverage detection output embeddings to represent targets and perform association [Huang *et al.*, 2025b; Gao *et al.*, 2025]. In contrast, our aim is to learn tracking-oriented representations by explicitly integrating motion and content cues, enabling the model to adaptively emphasize discriminative cues based on each target's evolving tracking condition, thereby improving performance in scenarios with complex motion and occlusion. To this end, we present MoTRa, as shown in Figure 2.

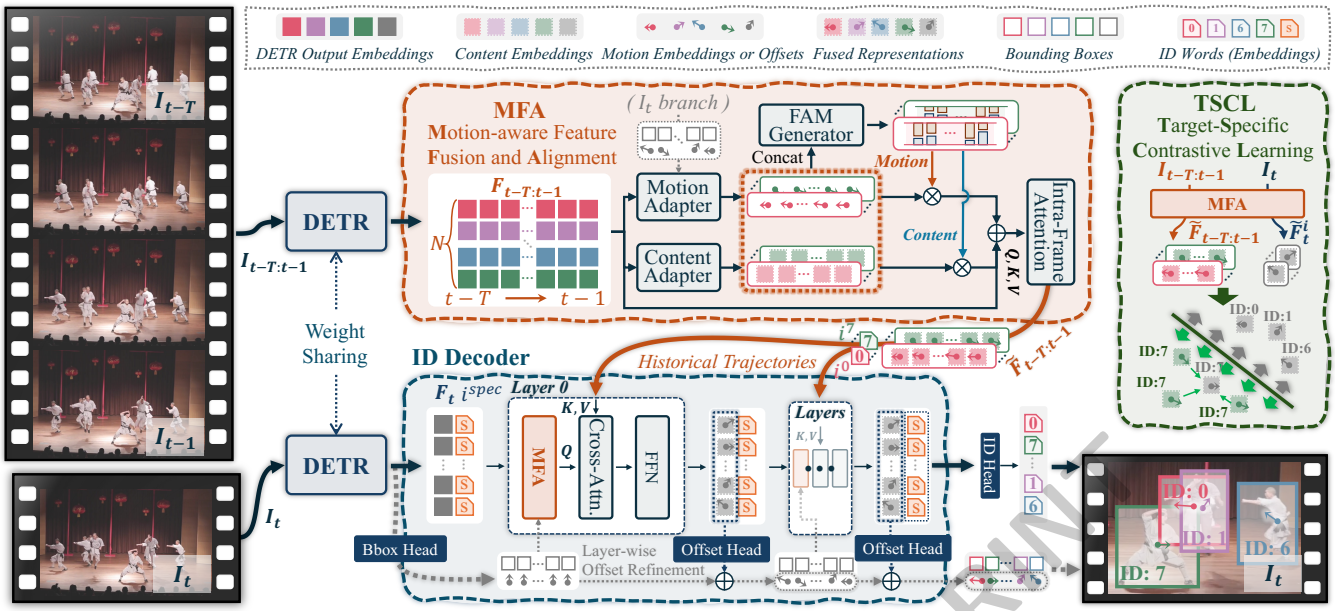

Figure 2: Overview of MoTRa. Unlike methods that directly use DETR output embeddings as target-level features for tracking, MoTRa learns motion-aware, tracking-oriented representations through two key components: an MFA module that adaptively fuses motion and content cues based on each target's state, and a TSCL strategy that enforces temporal consistency while maintaining discriminability across targets.

3.1 Overview

We implement MoTRa based on an ID prediction framework [Gao *et al.*, 2025], which detects targets and predicts their tracking identities within a single model, thereby enabling end-to-end optimization. Given a sequence of video frames $I_{t-T:t}$, a DETR-like detector is employed to detect targets in each frame and extract their output embeddings $F_{t-T:t}$. For association, tracked historical targets $\mathcal{T}_{t-T:t-1}$ are represented by combining their target-level features $F_{t-T:t-1}$ with correlated learnable ID tokens $\mathcal{I} = \{i^1, \ldots, i^K\}$ as in-context prompts, while newly detected targets in the current frame $\mathcal{D}_t$ are represented by F_t and a special ID token i^{spec}. Finally, an ID decoder facilitates interactions between $\mathcal{D}_t$ and $\mathcal{T}_{t-T:t-1}$, allowing the model to predict the ID labels of $\mathcal{D}_t$ from their ID token i^{spec} via an ID prediction head. However, as $F_{t-T:t}$ are optimized for detection, these features cannot reliably distinguish targets under complex occlusions.

In this work, rather than relying on detection-oriented $F_{t-T:t}$, we focus on learning more discriminative, tracking-oriented features $\widetilde{F}_{t-T:t}$ for target representation in both $\mathcal{T}_{t-T:t-1}$ and $\mathcal{D}_t$, thereby enhancing association performance. Specifically, we introduce MFA to adaptively integrate additional motion cues and obtain $\widetilde{F}_{t-T:t}$ (Section 3.2). In addition, we design a TSCL strategy to ensure the temporal consistency and discriminability of $\widetilde{F}_{t-T:t}$ (Section 3.3).

3.2 Motion-aware Feature Fusion and Alignment

In real-world tracking scenarios, cue reliability varies across tracking conditions. Positional information is often sufficient when targets are well separated, whereas appearance cues are effective when visual differences are clear. However, when targets experience severe occlusion, the relative reliability of positional and appearance cues becomes uncertain, and static positional cues can be less discriminative than motion cues. Motivated by this, we introduce motion cues into representation learning and propose MFA to derive target-level representations $\widetilde{F}_{t-T:t}$ that adaptively emphasize discriminative cues under varying tracking conditions.

Motion Cues. Compared with positional cues, motion provides more effective discrimination under occlusion. Unlike cues such as pose, which typically require additional models, motion features are easily derived and can be seamlessly integrated into end-to-end frameworks. Therefore, we exploit motion as an additional tracking cue to enhance association performance. Given the k-th target at frame j $(t - T \leq j \leq t)$, we formulate its motion cue as $m_j^k = (x_j^k, y_j^k, w_j^k, h_j^k, \Delta x_j^k, \Delta y_j^k, \Delta w_j^k, \Delta h_j^k) \in \mathbb{R}^8$, where Δ denotes the offsets from frame j to $j+1$, and and all terms are normalized by the image resolution. Specifically, the bounding box coordinates of the target are obtained from its corresponding DETR output embedding F_j^k using the Bbox head. For targets in the historical trajectories $\mathcal{T}_{t-T:t-1}$, the motion offsets are computed as coordinate differences between consecutive frames. Since there is no future frame for historical targets at frame $t-1$, we duplicate the offsets at frame $t-2$ to pad those at frame $t-1$. For targets from the new detections in $\mathcal{D}_t$, since the ID decoder enables interactions between $\mathcal{D}_t$ and $\mathcal{T}_{t-T:t-1}$, we directly predict the motion offsets of these targets from the ID decoder output embeddings using dedicated offset heads, thereby compensating for their unknown motion offsets. Inspired by [Zhu *et al.*, 2020], we refine the predicted offsets layer by layer, enabling the decoder to progressively adjust the predicted offsets across layers.

Feature Fusion and Alignment. Although motion cues are incorporated, they may not reliably distinguish targets under all tracking conditions. To adaptively exploit complementary information, inspired by gating mechanisms [Mu and Lin, 2025], we employ a generator that predicts a Feature Assignment Matrix (FAM) to dynamically balance motion and content cues. Given the DETR output embeddings for all N_j targets in the j-th frame ($t - T \leq j \leq t$), denoted as $F_j \in \mathbb{R}^{N_j \times C}$, we first extract both content and motion embeddings for these targets. Specifically, content embeddings $F'_j \in \mathbb{R}^{N_j \times C}$ are obtained by refining F_j with a feed-forward network (FFN) as a content adapter that emphasizes semantic information relevant to appearance and identity. In parallel, the motion cues $M_j \in \mathbb{R}^{N_j \times 8}$, which encode inter-frame variations in spatial position, are projected into the same C-dimensional space via a dedicated FFN motion adapter, yielding motion embeddings $M'_j \in \mathbb{R}^{N_j \times C}$. Then, by concatenating F'_j and M'_j, we feed them into the FAM generator, which consists of an FFN followed by a softmax operation, to predict the feature assignment matrix:

$$A_j = \mathrm{SoftMax}\left(\mathrm{FFN}\left(\mathrm{Cat}(F'_j, M'_j)\right)\right), \tag{1}$$

where $A_j = [A_j^{\text{cont}}, A_j^{\text{moti}}] \in \mathbb{R}^{N_j \times 2}$ stores the normalized weights assigned to the content and motion embeddings, respectively. Finally, we adaptively fuse the motion and content branches, while keeping a residual connection from F_j:

$$F_j^{fuse} = F_j + A_j^{\text{cont}} \odot F'_j + A_j^{\text{moti}} \odot M'_j, \tag{2}$$

where $F_j^{fuse} \in \mathbb{R}^{N_j \times C}$ and $\odot$ denotes element-wise multiplication. After the fusion stage, an intra-frame attention module is applied to further refine the fused features F_j^{fuse} by explicitly modeling interactions among targets within the same frame, which enhances contextual alignment and inter-target separability, yielding the final tracking-oriented features $\widetilde{F}_j$ for representing targets.

MFA for Current-Frame Detection Refinement. Accurate identity prediction requires effective interactions between historical trajectories and current detections in the ID decoder. To this end, we apply MFA before each cross-attention layer to refine target-level representations for both historical trajectories and current detections. As MFA already contains intra-frame attention, which serves the same role as self-attention, we replace the original self-attention module with MFA to refine current-frame features. Together with the offset refinement in MFA, this design allows the decoder to progressively update motion cues and adaptively adjust the fusion patterns between motion and content features.

3.3 Target-Specific Contrastive Learning

While MFA enables feature fusion adaptive to each target's evolving tracking condition, such dynamic fusion may cause the representations of the same target to become temporally inconsistent and reduce the separability between different targets. To address this, we propose a TSCL strategy to supervise the MFA-enhanced target representations, encouraging them to be temporally coherent and identity-discriminative.

Following the current-to-history interaction in the ID decoder cross-attention, TSCL constructs contrastive samples over current-history target pairs, making the supervision consistent with the ID prediction process. For the MFA-enhanced representations, we denote the representation of target identity k at historical frame j, where $t - T \leq j \leq t - 1$, as $\widetilde{F}_j^k$, and its current-frame representation from the l-th decoder layer as $\widetilde{F}_{t,l}^k$. TSCL constructs contrastive pairs as $(\widetilde{F}_j^k, \widetilde{F}_{t,l}^k)$, treating same-identity pairs as positives and different-identity pairs as negatives. This encourages $\widetilde{F}_j^k$ to be closer to $\widetilde{F}_{t,l}^k$ while separating it from current-frame representations of other identities in layer l. Accordingly, the InfoNCE loss [He *et al.*, 2020] is formulated as:

$$L_{\text{NCE}}^{k,j,l} = -\log \frac{\exp(\widetilde{F}_j^k \cdot \widetilde{F}_{t,l}^k)/\tau}{\sum_{k'=0}^{N_t} \exp(\widetilde{F}_j^k \cdot \widetilde{F}_{t,l}^{k'}))/\tau}, \tag{3}$$

where $\tau \in (0, 1]$ is a temperature hyper-parameter and N_t is the total number of targets in the frame t. To fully leverage historical information and enforce consistency across time and decoder depth, we extend this loss by aggregating over all historical frames and all decoder layers:

$$L_{cl} = \frac{1}{N_a} \sum_{k=1}^{N_a} \frac{1}{T} \sum_{j=t-T}^{T} \frac{1}{N_l} \sum_{l=1}^{N_l} L_{\text{NCE}}^{k,j,l}, \tag{4}$$

where N_a is the number of targets that appear in both the historical frames and the current frame t, and N_l is the total number of decoder layers.

3.4 Training

In this paper, we introduce additional losses to learn more robust motion-aware representations. Specifically, we adopt a standard L1 loss for offset regression [Zhang *et al.*, 2025], denoted as $\mathcal{L}_{off}$, and use Eq. 4 for contrastive learning. The tracking loss is defined as:

$$\mathcal{L}_{trk} = \lambda_{id}\mathcal{L}_{id} + \lambda_{off}\mathcal{L}_{off} + \lambda_{cl}\mathcal{L}_{cl}, \tag{5}$$

where $\mathcal{L}_{id}$ is a cross-entropy loss for ID classification [Gao *et al.*, 2025], and each λ is a hyper-parameter for the corresponding loss term. In addition, a detection loss is adopted to supervise the DETR-like detection module [Zhu *et al.*, 2020]:

$$\mathcal{L}_{det} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{giou}\mathcal{L}_{giou}, \tag{6}$$

where $\mathcal{L}_{cls}$ is used for object classification, and $\mathcal{L}_{L_1}$ and $\mathcal{L}_{giou}$ are used for bounding box regression. Finally, we train MoTRa with a unified loss $\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{trk}$.

4 Experiments

4.1 Datasets and Metrics

We evaluate our method on the DanceTrack [Sun *et al.*, 2022], SportsMOT [Cui *et al.*, 2023], and BFT [Zheng *et al.*, 2024] datasets, which feature high appearance similarity and complex motion. We adopt Higher Order Tracking Accuracy (HOTA), Detection Accuracy (DetA), and Association Accuracy (AssA) from the HOTA metric family [Luiten *et al.*, 2021], as well as MOTA [Bernardin and Stiefelhagen, 2008] and IDF1 [Ristani *et al.*, 2016]. IDF1 and AssA mainly reflect association quality, while MOTA and DetA focus on detection performance. HOTA integrates DetA and AssA to provide a balanced assessment of both detection and association.

Dataset	Methods	E2E	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
DanceTrack [Sun *et al.*, 2022]	Hybrid-SORT* [Yang *et al.*, 2024]		62.2	63.0	91.6	/	/
	DfTrack* [Huang *et al.*, 2025a]		65.5	68.2	92.7	52.4	82.1
	TrackTrack* [Shim *et al.*, 2025]		66.5	67.8	**93.6**	52.9	/
	MeMOTR [Gao and Wang, 2023]	✓	68.5	71.2	89.9	58.4	80.5
	MOTRv2† [Zhang *et al.*, 2023]	✓	69.9	71.7	91.9	59.0	83.0
	MOTRv3† [Yu *et al.*, 2023]	✓	70.4	72.3	92.9	59.3	**83.8**
	SambaMOTR [Segu *et al.*, 2024]	✓	67.2	70.5	88.1	57.5	78.8
	TGFormer* [Zeng *et al.*, 2025]	✓	69.1	71.9	91.3	58.3	81.9
	MOTIP [Gao *et al.*, 2025]	✓	69.6	74.7	90.6	60.4	80.4
	MoTRa (ours)	✓	**71.1**	**76.1**	90.9	**62.7**	80.8
SportsMOT [Cui *et al.*, 2023]	MeMOTR [Gao and Wang, 2023]	✓	68.8	69.9	90.2	57.8	82.0
	SambaMOTR [Segu *et al.*, 2024]	✓	69.8	71.9	90.3	59.4	82.2
	MOTIP [Gao *et al.*, 2025]	✓	72.6	77.1	92.4	63.2	83.5
	MoTRa (ours)	✓	**73.8**	**77.9**	**93.0**	**64.8**	**84.1**
BFT [Zheng *et al.*, 2024]	NetTrack* [Zheng *et al.*, 2024]		68.7	77.0	**78.9**	65.2	**72.6**
	SambaMOTR [Segu *et al.*, 2024]	✓	69.6	81.9	72.0	73.6	66.0
	MOTIP [Gao *et al.*, 2025]	✓	70.5	79.3	77.1	71.8	69.6
	MoTRa (ours)	✓	**71.3**	**82.9**	77.8	**72.7**	70.3

Table 1: Comparison with state-of-the-art methods on the DanceTrack, SportsMOT, and BFT test sets. Methods marked with $*$ adopt a stronger detector (*e.g.*, YOLOX-X [Ge *et al.*, 2021], DAB-DETR [Liu *et al.*, 2022]), while those marked with † are trained with additional data [Shao *et al.*, 2018]. E2E denotes end-to-end MOT methods. The best result is highlighted in bold.

4.2 Implementation Details

MoTRa is built upon the MOTIP baseline [Gao *et al.*, 2025]. For a fair comparison, we use the same Deformable DETR [Zhu *et al.*, 2020] as the DETR-like detector and keep the pre-trained weights and parameter settings identical to MOTIP. Specifically, we store up to $K = 50$ ID tokens for labeling target identities, based on the maximum number of targets in a frame. The maximum length of historical frames T is set to 29, 59, and 19 for DanceTrack [Sun *et al.*, 2022], SportsMOT [Cui *et al.*, 2023], and BFT [Zheng *et al.*, 2024], respectively. During training, we set λ_{cls}, λ_{L1}, λ_{giou}, and λ_{id} to 2.0, 5.0, 2.0, and 1.0, respectively. We additionally set $\lambda_{off} = 5.0$, identical to λ_{L1}, and $\lambda_{cl} = 1.0$ for representation learning. In TSCL, we set the temperature parameter τ to 0.02. We use the AdamW optimizer with an initial learning rate of 0.0001 and a batch size of 1, and multiply the learning rate by 0.1 at scheduled epochs. For DanceTrack, training lasts 10 epochs, with the learning rate reduced at epochs 5 and 9. For SportsMOT, training lasts 13 epochs, with reductions at epochs 8 and 12. For BFT, training lasts 22 epochs, with reductions at epochs 16 and 20. During inference, we set the thresholds for DETR output filtering, newborn objects, and ID prediction to 0.3, 0.6, and 0.2, respectively. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.3 Comparison with State-of-the-art Methods

We compare MoTRa with previous methods in Table 1. For Transformer-based end-to-end methods, different DETR-like detectors [Liu *et al.*, 2022] and extra detection datasets [Shao *et al.*, 2018] can significantly affect detection performance and robustness [Gao and Wang, 2023; Segu *et al.*, 2024]. Following MOTIP [Gao *et al.*, 2025], we adopt Deformable DETR [Zhu *et al.*, 2020] and train MoTRa without extra detection datasets for a fair comparison. Despite these constraints, MoTRa still achieves superior association performance, particularly in terms of HOTA, IDF1, and AssA.

DanceTrack and SportsMOT. DanceTrack [Sun *et al.*, 2022] presents group dancer tracking scenarios characterized by uniform appearance and diverse motion, while SportsMOT [Cui *et al.*, 2023] focuses on athlete tracking with high-speed movements and complex interactions. Both datasets pose significant challenges for association, especially in distinguishing targets under occlusion. Therefore, we select DanceTrack and SportsMOT as our primary benchmarks for comparative evaluation. For non-end-to-end methods such as Hybrid-SORT [Yang *et al.*, 2024], which leverage powerful detectors [Ge *et al.*, 2021] and incorporate additional tracking cues to enhance performance, the reliance on manual design makes it challenging to learn unified feature representations for targets, resulting in association performance that is still significantly outperformed by end-to-end methods. For end-to-end methods, MOTIP [Gao *et al.*, 2025] simplifies MOT as an ID prediction task and represents targets with additional identity tokens, achieving superior performance compared to previous tracking-by-propagation approaches such as MeMOTR [Gao and Wang, 2023]. Building on the ID prediction framework, MoTRa further enhances target representations by adaptively integrating motion cues and employing a TSCL training strategy. This yields improvements over the baseline, particularly in HOTA, IDF1, and AssA, demonstrating the effectiveness of our method.

BFT. Unlike human tracking, BFT [Zheng *et al.*, 2024] focuses on open-world, highly dynamic bird tracking, which also poses significant challenges in distinguishing targets due to low discriminability. Compared to the MOTIP baseline, MoTRa still demonstrates improvements in overall performance, with a particularly notable gain of +3.6 IDF1, further demonstrating the generalization ability of our approach.

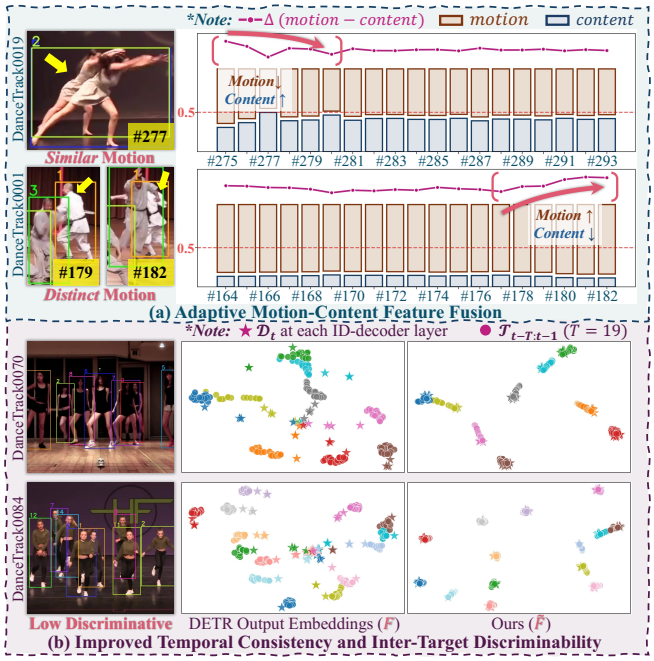

Figure 3: Visualization of motion-aware target representations. The points are visualized with t-SNE, and points with the same color correspond to the same identity.

4.4 Visualization Analysis

We provide a visual analysis of MoTRa's effectiveness in representing targets in challenging scenarios on DanceTrack [Sun *et al.*, 2022], as shown in Figure 3. In Figure 3(a), we visualize how the fusion of motion and content features adapts across frames and scenarios. Since targets across scenarios naturally rely differently on motion and content cues, we focus on temporal changes in the fusion weights (curves) rather than their absolute magnitudes. For the same target, the visualization shows that changing tracking conditions drive our model to rebalance content and motion. Besides, targets with different occlusion patterns also induce different fusion strategies. In DanceTrack0019, the occluded dancer gradually exhibits poses and motions similar to the occluder, leading to a decreased motion weight at frame #277, whereas in DanceTrack0001, larger motion and position differences result in the opposite trend.

In Figure 3(b), we illustrate the effectiveness of the proposed TSCL strategy. Despite the distinct fusion patterns of motion and content features across frames, representations of the same targets in both history frames (circle markers) and the current frame (star markers) remain well clustered, demonstrating strong temporal consistency. Moreover, targets with different identities are more clearly separated than in the original DETR output embeddings, further highlighting the improved discriminability of our learned representations.

4.5 Ablation Study

We conduct ablation experiments on the DanceTrack validation set, using MOTIP [Gao *et al.*, 2025] as the baseline. The hyperparameter T is fixed to 19 during ablation.

Row	Extra Cue	Gating	TSCL	HOTA↑	IDF1↑	MOTA↑
#1	N/A (Baseline)			57.8	60.8	80.8
#2	Position	✓		58.9	62.1	80.5
#3	Position	✓	✓	59.7	64.0	81.4
#4	Motion	✓		59.8	63.3	80.1
#5	Motion*	✓		60.5	64.4	81.1
#6	Motion*		✓	59.9	64.2	80.5
#7	Motion*	✓	✓	61.2	65.8	81.1

Table 2: Ablation study of motion-aware target representation learning. Motion* denotes the use of L_{off} for motion offset prediction.

Target Representation Learning. Compared with previous end-to-end MOT methods [Gao *et al.*, 2025], MoTRa introduces additional motion cues into target representation, leverages a gating mechanism to adaptively fuse them, and adopts a contrastive learning strategy to enhance representation discriminability. We therefore conduct an overall ablation to validate these components, as shown in Table 2.

For additional tracking cues, position and motion can be derived from detection results without extra models. The position cue encodes absolute coordinates, and the motion cue captures their temporal changes. Although the original DETR output embeddings F already contain positional information, explicitly decoupling and adaptively fusing these cues improves tracking, yielding +1.1 HOTA (Row #2 vs. Row #1). On top of position cues, incorporating motion offsets as motion cues brings an additional +0.9 HOTA (Row #4 vs. Row #2), confirming the benefit of motion information. Since no future frame is available to compute motion offsets for new detections, we predict their motion offsets to compensate for this missing information, further improving their representations for association (Row #5 vs. Row #4).

For different fusion strategies, we adopt a gating mechanism that allows the model to adaptively fuse multiple cues (Row #7), rather than simply adding them (Row #6). This design enables target representations to emphasize different cues under varying conditions (Figure 3(a)). The results show that the adaptive fusion strategy achieves better performance, bringing a gain of +1.3 HOTA and further validating the effectiveness of our gating-based fusion.

For the contrastive learning strategy, we adopt TSCL during training to enforce temporal consistency for the same targets and enhance discriminability between different targets. As shown in Table 2, applying TSCL further improves tracking performance for both position (Row #3 vs. Row #2) and motion (Row #7 vs. Row #5) modeling. These improvements, together with the visualization in Figure 3(b), further demonstrate the effectiveness of the TSCL strategy.

Architectural Design. To facilitate more effective information interaction between historical trajectories and detections in the current frame, we apply MFA to both historical targets $\mathcal{T}$ and new detections $\mathcal{D}$. We therefore conduct ablations on these structural designs, as shown in Table 3. Notably, since TSCL cannot be adopted when MFA is applied only to $\mathcal{T}$, we disable TSCL in all ablations in Table 3 for a fair comparison.

Building on the baseline model, we first apply MFA only

Target Set	History	Current	Layers	HOTA↑	IDF1↑	MOTA↑
N/A (Baseline)				57.8	60.8	80.8
$\mathcal{T}$	✓			59.4	63.3	80.1
$\mathcal{T}$ & $\mathcal{D}$	✓	✓	Single	59.6	63.9	79.8
$\mathcal{T}$ & $\mathcal{D}^*$	✓	✓	Multi	60.2	64.2	80.7
$\mathcal{T}$ & $\mathcal{D}$	✓	✓	Multi	60.5	64.4	81.1

Table 3: Ablation study for applying MFA to different target sets. Iterative offset refinement across layers is not used for D^*.

FAM predict	Residual	HOTA↑	IDF1↑	MOTA↑
F	✓	58.3	62.2	80.4
ADD$(F' + M')$	✓	60.0	63.2	81.7
Cat(F', M')		59.7	63.1	80.1
Cat(F', M')	✓	61.2	65.8	81.1

Table 4: Ablation study of gating and fusion strategies in MFA.

Attention	Placement	HOTA↑	IDF1↑	MOTA↑
$\mathcal{D}$	After Fusion	60.5	64.0	80.6
$\mathcal{D}$ & $\mathcal{T}$	Before Fusion	60.3	63.7	81.5
$\mathcal{D}$ & $\mathcal{T}$	After Fusion	61.2	65.8	81.1

Table 5: Ablation study of intra-frame attention in MFA.

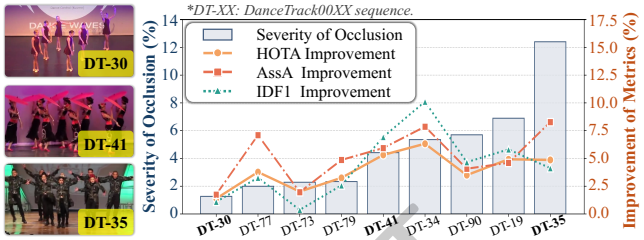

Figure 4: Performance improvements over the baseline under different occlusion severities.

to $\mathcal{T}$. Despite the semantic misalignment between $\mathcal{T}$ and $\mathcal{D}$, incorporating motion cues already helps distinguish different trajectories, yielding a gain of +1.6 HOTA. To further align $\mathcal{T}$ and $\mathcal{D}$, we then apply a single MFA module to $\mathcal{D}$ directly after detection. However, at this stage it is difficult to predict motion offsets for $\mathcal{D}$, leading to ineffective fusion and only negligible gains. To obtain predicted motion offsets and update fusion strategies at each decoder layer, we instead apply MFA across layers, which brings notable improvements over using MFA only on $\mathcal{T}$. Finally, inspired by the bounding box refinement in [Zhu *et al.*, 2020], we refine motion offsets iteratively across layers rather than predicting them independently, yielding a slight additional improvement.

Gating and Fusion Strategies. To adaptively fuse distinct cues, we predict an assignment matrix (FAM) conditioned on the concatenation of the content and motion embeddings, F' and M'. As shown in Table 4, directly using the DETR output embeddings F for FAM prediction brings only negligible gains over the baseline. Although F already encodes appearance and positional information, it lacks temporal motion cues and is therefore suboptimal for learning effective FAMs. In contrast, incorporating motion via ADD or Cat yields clear improvements in tracking performance, with the concatenation design performing best. We further introduce a residual connection during feature fusion to stabilize optimization, which proves crucial for obtaining notable gains.

Intra-frame Attention. The self-attention mechanism in the ID decoder enables information exchange among detected targets within the current frame, which facilitates identity discrimination for association [Gao *et al.*, 2025]. Motivated by this, we extend self-attention to historical targets as an intra-frame attention module and integrate it into our MFA. As shown in Table 5, applying the attention to historical targets after feature fusion yields a gain of +0.7 HOTA, indicating that historical targets within the same frame also benefit from information exchange for better feature alignment. We further investigate the placement of the attention relative to feature fusion. Placing attention before fusion allows information exchange only among the raw DETR output embeddings, which lack temporal motion cues. In contrast, placing attention after fusion operates on already fused representations, helping align their semantic information and better enhancing discriminability, thereby achieving better performance.

Effectiveness under occlusion. We further evaluate association performance under different occlusion levels, as shown in Figure 4. Occlusion severity for each video is defined as the ratio between severe overlaps (IoU > 0.6) and all overlaps (IoU > 0.1) [Shim *et al.*, 2025]. MoTRa shows overall gains in IDF1 and AssA over the baseline, especially under higher occlusion severity, indicating improved robustness.

Inference Speed and Parameters. Compared to the baseline, which runs at 27.1 FPS with 59M parameters, our method achieves 25.8 FPS with 63M parameters on the DanceTrack validation set. Despite the extra computational cost, our model remains feasible for practical applications.

4.6 Discussion

Although we introduce a motion-aware target representation learning method, its detection gain remains limited since ID prediction is decoupled from detection. Building on motion cues, we plan to more effectively exploit predicted motion to assist detection in subsequent frames, which we believe can further improve overall tracking performance.

5 Conclusion

This paper presents MoTRa, a motion-aware target representation learning framework that enhances detection-oriented features with motion-aware tracking cues to improve target discriminability through an MFA module and a TSCL strategy. MFA adaptively fuses motion cues with appearance-related content cues to emphasize the most informative cues under varying tracking conditions, while TSCL regularizes the fused representations by promoting intra-trajectory consistency and inter-target separability. Experiments demonstrate the effectiveness of MoTRa.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work was supported by the Beijing Municipal Science and Technology Program (Z251100003625011), the Beijing Natural Science Foundation (L252018), the NSFC through grant No.62402054, the China Postdoctoral Science Foundation through grant 2024M760279, and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under grant BX20250390.

References

[Bernardin and Stiefelhagen, 2008] 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.

[Bewley *et al.*, 2016] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing*, pages 3464–3468. IEEE, 2016.

[Carion *et al.*, 2020] 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.

[Cui *et al.*, 2023] 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.

[Gao and Wang, 2023] 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.

[Gao *et al.*, 2025] Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27883–27893, 2025.

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

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[Huang *et al.*, 2025a] Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao Wei. Dftrack: Deconfused data association framework for multi-object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):10367–10381, 2025.

[Huang *et al.*, 2025b] Yuanzhou Huang, Songwei Pei, and Rui Zeng. Dqformer: Transformer with decoupled query augmentations for end-to-end multi-object tracking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(6):1–23, 2025.

[Li *et al.*, 2024] Haijun Li, Zhuye Xu, Changxi Ma, and Xiao Tang. Multi-object tracking algorithm for unmanned vehicle autonomous driving scene based on online spatiotemporal feature correlation. *IEEE Access*, 12:116489–116497, 2024.

[Liu *et al.*, 2022] 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.

[Liu *et al.*, 2025] Yue Liu, Shijie Zhang, Huayi Li, and Chao Zhang. Satsort: Online multi-spacecraft visual tracking algorithm. *Acta Astronautica*, 232:103–113, 2025.

[Luiten *et al.*, 2021] 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.

[Meinhardt *et al.*, 2022] 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.

[Mu and Lin, 2025] Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

[Qin *et al.*, 2024] Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19004, 2024.

[Ristani *et al.*, 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing.

[Segu *et al.*, 2024] 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.

[Seidenschwarz *et al.*, 2023] Jenny Seidenschwarz, Guillem Brasó, Victor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13813–13823, 2023.

[Shao *et al.*, 2018] 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.

[Shim *et al.*, 2025] Kyujin Shim, Kangwook Ko, Yujin Yang, and Changick Kim. Focusing on tracks for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11687–11696, 2025.

[Sun *et al.*, 2022] 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.

[Wojke *et al.*, 2017] 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*, pages 3645–3649. IEEE, 2017.

[Yang *et al.*, 2023] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4799–4808, 2023.

[Yang *et al.*, 2024] 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*, volume 38, pages 6504–6512, 2024.

[Yu *et al.*, 2023] 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.

[Zeng *et al.*, 2022] 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.

[Zeng *et al.*, 2025] Rui Zeng, Yuanzhou Huang, and Songwei Pei. Tgformer: Transformer with track query group for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9824–9832, 2025.

[Zhang *et al.*, 2023] 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.

[Zhang *et al.*, 2024] Yang Zhang, Xiaoyan Cui, and Yu Bai. Research on an improved motr-gru multi-target tracking and behavior analysis method. In *2024 4th International Symposium on Artificial Intelligence and Intelligent Manufacturing*, pages 623–626, 2024.

[Zhang *et al.*, 2025] Bozhou Zhang, Nan Song, Xin Jin, and Li Zhang. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6854–6863, 2025.

[Zheng *et al.*, 2024] Guangze Zheng, Shijie Lin, Haobo Zuo, Changhong Fu, and Jia Pan. Nettrack: Tracking highly dynamic objects with a net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19145–19155, 2024.

[Zhu *et al.*, 2020] 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.