arXiv:2411.06197v2 [cs.CV] 28 Jun 2025

Tracking by Detection and Query: An Efficient End-to-End Framework for Multi-Object Tracking

Shukun Jia[a,b], Shiyu Hu[c], Yichao Cao[d], Feng Yang[e], Xin Lu[a,b], Xiaobo Lu[a,b,*]

[a]*School of Automation, Southeast University, Nanjing, 210096, China*
[b]*Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, 210096, China*
[c]*School of Physical & Mathematical Sciences, Nanyang Technological University, 639798, Singapore*
[d]*Big Data Institute, Central South University, Changsha, 410083, China*
[e]*School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China*

Abstract

Multi-object tracking (MOT) is dominated by two paradigms: tracking-by-detection (TBD) and tracking-by-query (TBQ). While TBD is decoupled and efficient, its fragmented association steps and heuristic matching pipelines often compromise robustness in complex scenarios. TBQ provides stronger semantic modeling through end-to-end learning, but suffers from high training cost and slow inference due to tight coupling between detection and association. To address these challenges, we propose TBDQ-Net, a unified tracking-by-detection-and-query (TBDQ) framework that effectively combines the strengths of both paradigms. Our method efficiently integrates pretrained, high-performance detectors with an MOT-tailored associator. The associator is lightweight and directly fetches information from the inference of detectors, enhancing the overall efficiency of the framework. The associator is also learnable, making it essential for fully end-to-end optimization, ensuring robust tracking capabilities. Specifically, the associator comprises two key modules: basic information interaction (BII) for comprehensive semantic interaction, and content-position alignment (CPA) for semantic and positional consistency. TBDQ-Net's effectiveness is extensively demonstrated on DanceTrack, SportsMOT and MOT20 benchmarks. As a structurally efficient and semantically robust tracking framework, it outperforms the leading TBD method by 6.0 IDF1 points on DanceTrack and achieves at least 37.5% faster inference than prominent TBQ methods.

Keywords: Multi-Object Tracking, Tracking by Detection, Tracking by Query

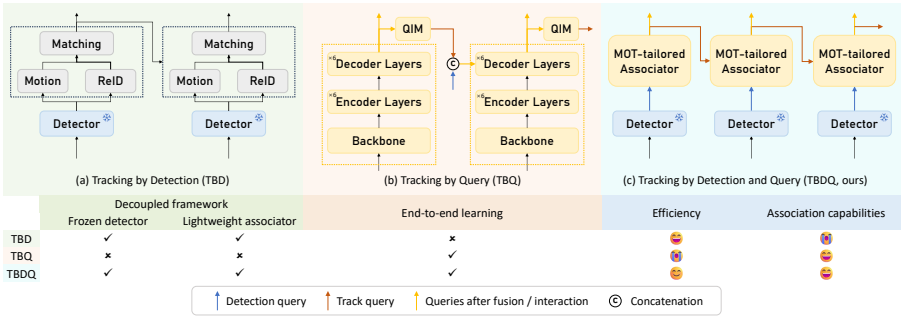


Figure 1: Comparison of mainstream paradigms. Tracking by detection and query (TBDQ) is a decoupled yet end-to-end framework, integrating a lightweight, learnable associator with a frozen, pretrained detector. *It outperforms conventional tracking-by-detection (TBD) approaches* in association performance and coherence, attributable to its learnable associator and end-to-end optimization. *Compared to tracking-by-query (TBQ) approaches,* it significantly boosts efficiency through a decoupled framework, which facilitates the efficient integration of pretrained detectors and a task-tailored associator.

1. Introduction

Multi-Object Tracking (MOT) is a pivotal task in applications involving scene understanding [1], autonomous driving [2], *etc*. Given a video with target classes, the goal is to recognize, localize and assign consistent identification numbers to objects over time. It essentially consists of two tasks: detection and association. Handling the relationship between them has been a central theme in the field. Accordingly, there are two dominant paradigms: the conventional tracking-by-detection (TBD) [3, 4, 5, 6] and the emerging tracking-by-query (TBQ) [7, 8, 9], which have complementary merits and demerits.

The TBD paradigm employs a decoupled framework, first detecting objects spatially and then associating them temporally. The modular framework facilitates the utilization of pretrained, high-performance detectors [10, 11] and lightweight association algorithms [3, 5], leading to outstanding practical efficiency. However, in the typical association process, appearance features and motion patterns are extracted independently and integrated heuristically, *resulting in a fragmented pipeline that lacks coherent modeling capabilities in complex scenarios*. Adopting a different philosophy, the TBQ

[1]* Corresponding author: Xiaobo Lu, email: xblu2013@126.com

paradigm employs an end-to-end framework to effectively address these limitations. It substantially enhances association capabilities through coherent utilization of tracking data. Nonetheless, TBQ methods typically couple detection and association, *meaning that their powerful association performance is intrinsically tied to joint training throughout the model.* This tight coupling often leads to task competition [7, 9] and substantially impedes overall efficiency, resulting in high training barriers and computational costs. In essence, despite their individual merits, these two mainstream paradigms possess complementary advantages: TBD excels in efficiency but struggles with comprehensive contextual understanding, while TBQ provides robust and expressive modeling capabilities, but at a significant computational cost. This clear structural conflict highlights a pressing demand for their unification to achieve both efficiency and robust association.

Recognizing this demand, recent works such as MOTIP [12] and MOTRv2 [13] have explored directions towards combining the merits of the two paradigms, achieving commendable progress. Specifically, MOTIP transforms the heuristic association of the TBD paradigm into a separate end-to-end ID prediction task; however, it still necessitates joint optimization of the detector and ID predictor for optimal performance, thereby failing to efficiently exploit off-the-shelf detection capabilities from pretrained models. Similarly, MOTRv2 decouples detection from the TBQ paradigm by leveraging a frozen, pretrained YOLOX for object detection. Yet, its association relies on an MOTR [7] model, whose capability still stems from the computationally expensive training of a full DETR-like architecture, including the backbone, encoder, and decoder. Consequently, despite their earnest efforts, these works offer limited improvements in overall efficiency.

In this work, we take a significant step towards unifying the strengths of both paradigms, introducing the tracking-by-detection-and-query (TBDQ) framework. Our model, TBDQ-Net, integrates a lightweight, MOT-tailored associator with a frozen, pretrained detector. Inspired by the TBD paradigm, the efficiency of TBDQ-Net intrinsically arises from two key aspects of the associator: its inherently lightweight design and compact structural integration. Specifically, the latter allows it to directly acquire necessary information from the detector's inference stream, fundamentally avoiding the cumbersome computations from DETR-like architectures that are prevalent in TBQ methods. More crucially, the associator is also learnable, which is essential for the fully

end-to-end framework to achieve robust association capabilities. Ultimately, TBDQ-Net sets a new standard by delivering unparalleled efficiency and robust tracking performance, effectively unifying the primary advantages of both paradigms. The structures and distinct features of each framework are visually illustrated in Figure 1. As the core component of TBDQ-Net, the proposed MOT-tailored associator consists of two key modules: the basic information interaction (BII) module, which facilitates semantic interaction between object queries and alleviates potential conflicts; and the content-position alignment (CPA) module, which aligns positional components with the updated semantic for information consistency.

TBDQ-Net is extensively evaluated in large-scale datasets, DanceTrack [14] and SportsMOT [15], as well as the extremely crowded benchmark MOT20 [16]. Compared to TBD methods, it acquires notably robust association capabilities in challenging scenarios. In DanceTrack, it surpasses the state-of-the-art TBD method Hybrid-SORT [17] by 3.7 points in the HOTA metric and 6.0 points in the IDF1 metric. Compared to TBQ methods, TBDQ-Net significantly improves the efficiency of both training and inference processes while maintaining the competitive tracking accuracy. Overall, the contributions are summarized as follows:

- We propose TBDQ-Net, a novel tracking-by-detection-and-query (TBDQ) framework, designed to resolve the inherent efficiency-accuracy trade-off in multi-object tracking (MOT). TBDQ-Net effectively unifies the decoupled framework of tracking-by-detection (TBD) and the end-to-end optimization of tracking-by-query (TBQ), achieving robust association with substantially higher efficiency.

- To enhance efficiency, we develop an efficient pipeline to leverage frozen, pre-trained detectors in a decoupled framework. This design allows detectors to not only yield high-quality detection results, but also provide immediate, crucial representations for subsequent association, effectively guaranteeing the framework's overall efficiency.

- To ensure robust association capabilities, we propose a learnable associator that seamlessly integrates within the end-to-end framework. This associator primarily

comprises the basic information interaction (BII) module for semantic interaction and the content-position alignment (CPA) module for information consistency.

- We demonstrate TBDQ-Net's effectiveness and practical utility through extensive experiments across challenging real-world scenarios, including dance, sports, and street scenes. It achieves a highly remarkable balance between tracking efficiency and accuracy, establishing a significant new benchmark in the field.

2. Related Work

Significant advancements in multi-object tracking (MOT) have been made through two primary paradigms: tracking-by-detection (TBD) and tracking-by-query (TBQ). This section provides a comprehensive review of these approaches.

2.1. Tracking by Detection

This paradigm has dominated the MOT field for years. SORT [3] performs location prediction using a Kalman Filter, with subsequent matching achieved through the Hungarian algorithm. Deep SORT [4] incorporates appearance features into the SORT pipeline to enhance re-identification (ReID) against misses and occlusions. OC-SORT [18] compensates for the linear motion assumption of SORT through an upgraded motion model. FairMOT [19] jointly learns detectors and appearance embeddings. ByteTrack [5] improves association by considering more low-score detections. FineTrack [20] aims to use diverse fine-grained representations. GTR [21], PuTR [22] and C-TWiX [23] upgrade heuristic association algorithms to learnable Transformer-based modules for enhanced association performance. SeaTrack [24] addresses nearshore MOT via a vision-based pipeline that is mainly based on motion information. Despite significant progress, these methods often suffer from complicated and severely fragmented designs, lacking coherent optimization.

2.2. Tracking by Query

These methods emerge in recent years and exhibit superior association robustness through end-to-end learning. They apply the query mechanism in MOT, where new

objects are identified by detection queries and existing ones are linked to track queries. TrackFormer [25] explores an end-to-end trainable framework but still requires operations such as non-maximum suppression (NMS) and ReID features. MOTR [7] achieves a complete end-to-end pipeline with techniques such as collective average loss and temporal aggregation network. After that, MeMOT [26] and MeMOTR [8] improve the utilization of temporal information. CO-MOT [9] and MOTRv3 [27] alleviate the conflict of MOTR by injecting more supervision into detection queries.

Apart from these methods, there are other works that endeavor to integrate the benefits of both TBD and TBQ paradigms. MOTIP [12] transforms the individual association of TBD into an end-to-end ID prediction task, building an ID predictor upon a pretrained detector. Nonetheless, it still requires joint training of the detector and ID predictor for optimal performance, failing to efficiently utilize off-the-shelf detection capabilities. MOTRv2 [13] uses predictions of the pretrained YOLOX [11] as prior knowledge, guiding the entire MOTR model to focus on association. However, its training still incorporates the backbone, encoder, decoder of DETR-like architectures, which is computationally intensive. In summary, albeit commendable progress, these approaches offer limited improvement in efficiency.

3. Methodology

3.1. Preliminary

The overall pipeline of TBDQ-Net is depicted in Figure 2. Given the frame I_t, detection results from a pretrained detector are encoded into object queries D_q^t, providing crucial detection information for downstream interaction. Concurrently, track queries T_q^t, representing existing objects, are propagated from the previous timestep. These two types of queries then undergo interaction within the learnable associator, where they engage in thorough semantic exchange and get their content components aligned with positional components. Subsequently, they are directly decoded into tracking results. *Overall, the efficient utilization of pretrained detectors and the proposed MOT-tailored learnable associator enable TBDQ-Net to acquire robust association*

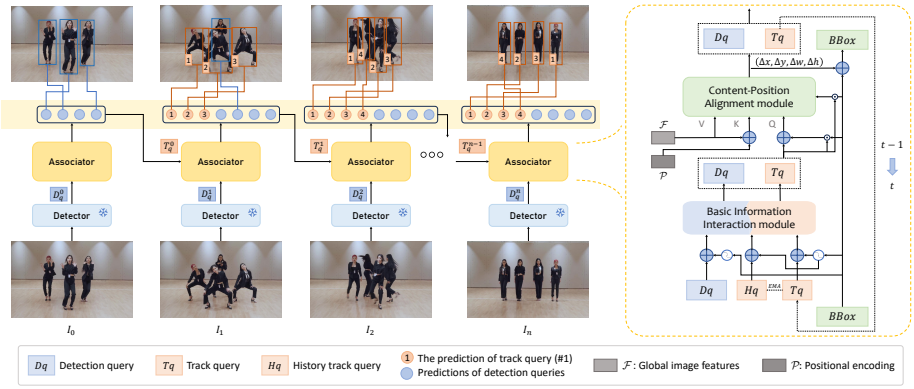


Figure 2: The pipeline of TBDQ-Net. Detection queries are derived from pretrained detectors, carrying detection information of each frame. Track queries are initially converted from decoded detection queries, then linked to fixed objects and propagated over time. For the first frame I_0, as track queries are not yet present, detection queries are immediately decoded into final predictions. For subsequent frames, the two types of object queries undergo interaction within the associator and are decoded into tracking results in an end-to-end manner. The dashed box on the right illustrates the structure of the proposed learnable associator. The basic information interaction (BII) module, depicted with two distinct colors, indicates that the modules used for updating detection queries and track queries possess individual parameter weights. Global image features $\mathcal{F}$ represent refined backbone features from pretrained detectors, while $\mathcal{P}$ indicate the 2D spatial positional encoding of these backbone features.

capabilities with remarkably higher efficiency. The following subsections elaborate on each key component.

3.2. Integration of Pretrained Detectors

Our approach efficiently integrates off-the-shelf pretrained detectors, a fundamental change from conventional TBQ models that typically incur computationally expensive joint training with their full DETR-like architectures. We achieve this by directly feeding the global image features $\mathcal{F}$ and detection queries D_q from pretrained detectors to the Transformer-based associator and decoder, which are easily accessible during the detector's inference process.

For CNN-based detectors, such as YOLOX, global image features $\mathcal{F}$ are derived by flattening the backbone features from the last three layers and mapping them via a multi-layer perceptron (MLP). Concurrently, detection queries D_q are formed by directly extracting features from the backbone based on the predicted detection boxes.

Conversely, Transformer-based detectors, which are inherently compatible with this framework, directly provide the necessary representations. Global image features $\mathcal{F}$ are obtained from the refinement features of their encoder, while detection queries D_q are derived from the output embeddings of their decoder.

Furthermore, inspired by Conditional-DETR [28] and DAB-DETR [29], we construct comprehensive object queries through distinct content and positional components. The former is introduced by aforementioned operations, while the latter is produced by the sinusoidal encoding of the predicted bounding boxes [29].

3.3. Basic Information Interaction Module

As a core component of the learnable associator, the basic information interaction (BII) module is designed to facilitate a comprehensive, bidirectional information flow between detection queries and track queries. This design serves a twofold objective: track queries are updated by detection queries carrying new detection information, ensuring their validity across time; simultaneously, detection queries related to already existing objects are suppressed by track queries, thereby mitigating potential conflicts between the two types of queries. Simply and intuitively, the BII module is achieved through the scaled dot-product attention that contains the variable Query (Q), Key (K) and Value (V):

$$\begin{aligned} O_1 &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V_1, \\ O_2 &= \text{norm}\left(O_1 + V_2\right), \\ O_3 &= \text{norm}\left(\text{FFN}\left(O_2\right) + O_2\right). \end{aligned} \tag{1}$$

Specifically, the standard vector V is replaced by two distinct vectors V_1 and V_2, where V_1 serves as the input to the attentive aggregation that updates the base information contained in V_2. d is the dimension of vectors Q, K, V_1, V_2. FFN means the standard feed-forward neural network [30]. On top of the BII module, the aforementioned bidirectional interaction is realized through specific configurations of the vectors Q, K,

V_1 and V_2. When updating detection queries,

$$\begin{aligned} Q &= \widetilde{D_q}, K = \text{concat}(\widetilde{D_q}, \widetilde{T_q}), \\ V_1 &= \text{concat}(D_q, N_q), V_2 = D_q. \end{aligned} \tag{2}$$

While for track queries,

$$\begin{aligned} Q &= \widetilde{T_q}, K = \text{concat}(\widetilde{D_q}, \widetilde{H_q}), \\ V_1 &= \text{concat}(D_q, H_q), V_2 = T_q. \end{aligned} \tag{3}$$

D_q, T_q and H_q denote the content components of detection queries, track queries and history track queries, respectively. $\widetilde{D_q}$, $\widetilde{T_q}$ and $\widetilde{H_q}$ stand for the queries including both the content and positional components:

$$\widetilde{E} = E + P_e(B), E \in \{D_q, T_q, H_q\}, \tag{4}$$

where E signifies the content component and $P_e(B)$ is the positional component. B denotes bounding boxes and $P_e(B)$ means the corresponding sinusoidal positional encoding [29]. $\widetilde{H_q}$ shares the same positional part with $\widetilde{T_q}$.

Though Equations (2) and (3) are both implemented by the BII module, they employ separate parameters to enhance their respective performance. Additionally, there are three vital points worth noting: ***(i)*** To reduce the impact of background noise and improve the interaction quality, we leverage the threshold τ_q to filter out low-quality detection queries before the bidirectional interaction. ***(ii)*** To alleviate potential conflict between detection queries and track queries, we suppress the former that exhibit high attention weights with the latter, as they tend to link the same object. Specifically, we utilize noisy queries N_q as a component of V_1 in Equation (2). To further enhance the learning capacity, N_q is constituted by M hard negative samples, *i.e.,* the filtered detection queries with the highest M confidence scores. M equals the number of track queries propagated from the previous moment. ***(iii)*** When updating track queries via Equation (3), we incorporate history track queries H_q to cope with missed objects. This allows track queries to be enhanced by H_q when their corresponding objects temporarily disappear. H_q are collected by the exponential moving average (EMA) method:

$$H_q^t = \begin{cases} w * T_q^t + (1 - w) * H_q^{t-1}, & \text{if } t > 0, \\ T_q^0, & \text{if } t = 0, \end{cases} \tag{5}$$

where w indicates the update weight of new information.

3.4. Content-Position Alignment Module

A temporal inconsistency arises when the content components of track queries are updated by the BII module, while their positional components retain information from the previous timestep. A similar misalignment also arises in detection queries following the semantic update within the BII module. To resolve these issues, the content-position alignment (CPA) module is introduced to align the positional components with their corresponding content components. The alignment process can be formulated as follows:

$$B' = f(E') + B, \tag{6}$$

where E' denotes the updated content components through the BII module, B signifies the initial positional components of queries, and B' represents the newly aligned ones. The function $f(\cdot)$ is a learnable transformation that generates a positional update derived from E'. In this work, we employ modulated cross-attention from DAB-DETR as the function $f(\cdot)$. This choice naturally facilitates the CPA module's implementation due to its inherent ability to simultaneously refine both positional information and object embeddings. Furthermore, to strengthen the impact of the CPA module, auxiliary losses are introduced during the training stage, which shares the same label assignment and calculation method as the final loss function.

3.5. Association Decoder

Through the BII and CPA modules, detection queries and track queries undergo thorough semantic exchange and information alignment across their content and positional components. Subsequently, these queries are directly decoded into predictions using a single Transformer decoder layer, ensuring the fully end-to-end nature of TBDQ-Net. In contrast to traditional TBD matching strategies that group objects into trajectories based on affinity matrices, this enables the model to understand the environment more holistically and acquire significantly stronger association capabilities for challenging motion patterns. Detailed demonstrations are provided in the experimental section.

Table 1: The number of frames (K) / sequences in training, validation and test subsets for each dataset. Density denotes the average number of boxes per frame.

Datasets	Training	Validation	Test	Total	Density
DanceTrack [14]	41.8 / 40	25.5 / 25	38.6 / 35	105.9 / 100	5.4
SportsMOT [15]	28.6 / 45	27.0 / 45	94.8 / 150	150.4 / 240	10.8
MOT20 [16]	8.9 / 4	–	4.5 / 4	13.4 / 8	123.2

3.6. Training and Inference

In the training stage, we adopt MOTR's approaches for label assignment and final loss calculation, which are widely employed in succeeding TBQ methods. During inference, tracked objects with confidence scores below τ_n are marked as inactive. These inactive objects will be removed from the tracklets if they persist for consecutive T frames. Please refer to the original paper [7] for more details.

4. Experiments

4.1. Datasets

TBDQ-Net is evaluated on two large-scale datasets, DanceTrack [14], SportsMOT [15], as well as MOT20 [16], a distinguished dataset for its highly crowded pedestrian scenes. DanceTrack is a widely adopted benchmark due to its abundant data and high-quality annotations. It presents significant challenges to the association task through similar appearances, diverse non-linear motion patterns, severe occlusions, and frequent clustering. SportsMOT focuses on tracking in sports scenes. In addition to similar yet distinguishable appearances, its challenge also lies in fast and variable-speed motions on courts. MOT20, a representative MOTChallenge benchmark, is specifically designed to evaluate tracking performance in highly congested street scenarios. Detailed statistics for these datasets are shown in Table 1.

4.2. Metrics

Tracking accuracies are measured on the widely used metrics: HOTA [31], DetA, AssA, MOTA [32] and IDF1 [33]. HOTA is a balanced metric, comprehensively evaluating tracking performance across both detection and association, where $HOTA =$

$\sqrt{DetA \cdot AssA}$. DetA and MOTA emphasize detection performance, while AssA and IDF1 reflect association capabilities. In MOT20, we follow the standard procedure to further report FP, FN and IDSW, where MOTA = 1 − (FP + FN + IDSW)/GT. FP, FN, IDSW and GT represent the number of predicted false positives, false negatives, ID switches and annotated ground truth boxes, respectively.

4.3. Implementation Details

General Details: The initial learning rate is 1.2×10^{-4} for all datasets. In DanceTrack, the training schedule is (6, 10, 12), meaning that the model is trained in 12 epochs, with the initial learning rate dropped by 10 at the 6th and 10th epoch, respectively. Due to the smaller number of training data in SportsMOT and MOT20, the training schedule expands to (12, 16, 20) and (20, 25, 30), in turn. For a fair comparison in each benchmark, TBDQ-Net is built upon the publicly released YOLOX, which is pretrained by ByteTrack [5] and widely utilized in subsequent research. Unless particularly specified, the image processing and sampling strategies are identical to those of CO-MOT [9] and MOTRv2 [13]. In all experiments, hyperparameters τ_q (Section 3.3), τ_n and T (Section 3.6) are set to 0.3, 0.5 and 20, respectively. *To focus on the essence of tracking and avoid artifacts from unrealistic, clumsy motion patterns [12], we did not utilize any extra data, e.g., crowdhuman [34] for joint training.*

Ablation Studies: The major ablation study is conducted in DanceTrack, due to the sufficient and high-quality data in training and validation subsets. For fast iteration, TBDQ-Net incorporates a pretrained DAB-DETR as the detector and is trained in lower resolutions. The shortest side ranges from 480 to 800 pixels with a step of 32 pixels, and the longest side is at most 1,333 pixels.

MOT20: Considering that the training data of MOT20 is 3 to 6 times fewer than that of the other two datasets, we employ the augmentation strategies of MOTR [7] to mitigate overfitting. Related configurations are separately investigated through ablation studies. As there is no validation subset in MOT20, we follow CenterTrack [10] to divide each sequence in the training subset into two parts: the first part is used for training, and the second part serves as the validation set.

Table 2: Efficiency comparison with TBQ methods. '#Parameters' refers to the total number of trainable parameters, which are expressed in millions (M). Training time corresponds to the model's training duration on DanceTrack, with the specific GPU resources clarified. Values are sourced from official publications if available. Otherwise, they are reproduced from the official implementations, which are marked by †. FPS (frames per second) of all methods is is tested *consistently on one NVIDIA 4090 GPU with FP32-precision*. For MOTRv2, values of training time and FPS are shown in gray to indicate they are derived from pre-computed offline detection results rather than online inference of YOLOX. Additionally, we report its practical FPS that incorporates the time of YOLOX's online inference. Similarly, we report the efficiency values of TBDQ-Net with the pretrained YOLOX for fair comparison.

TBQ Methods	#Parameters (M)↓	Training time↓	GPUs	FPS↑
MOTR [7]	43.9	60h	8×V100	14.8
MOTIP [12]	58.9	36h	8×4090	14.4
CO-MOT† [9]	41.7	125h	2×4090	14.9
MOTRv2† [13]	41.7	33h	2×4090	20.5
	41.7	–	–	16.0
TBDQ-Net (ours)	**8.5**	**22h**	2×4090	**22.0**

4.4. Comparison with Other Methods

Initially, we compare the *efficiency* of TBDQ-Net against representative TBQ methods, demonstrating its superiority in both training and inference stages. Subsequently, we evaluate the *accuracy* of multiple advanced methods across the DanceTrack, SportsMOT, and MOT20 benchmarks. These comparisons elucidate the distinctive characteristics of each method across different scenarios.

Comparison in Efficiency: As a fundamental advancement in efficiency, TBDQ-Net demonstrates a dramatic superiority over previous TBQ methods, as evidenced by Table 2. With a remarkably lower trainable parameter count of just 8.5M, TBDQ-Net substantially reduces model complexity compared to previous methods. *This improvement stems from the efficient integration of pretrained detectors and a specialized MOT-tailored associator, allowing TBDQ-Net to offload computationally intensive operations from cumbersome DETR-like architectures that are prevalent in TBQ methods.* Specifically, when compared to MOTIP, TBDQ-Net achieves a 52% increase in inference speed while simultaneously reducing the consumption of training resource by approximately 85%. Notably, against MOTRv2, another similar method that employs the pretrained YOLOX

Table 3: Comparison in DanceTrack. The upper part of the table presents tracking-by-detection (TBD) methods, while the lower part details tracking-by-query (TBQ) methods. Methods marked with * share the same pretrained detector YOLOX. The top-2 scores are displayed in **blue** and **green**, ***aiming to represent the competitive performance of TBDQ-Net.*** – indicates that data is not available.

Methods	Year	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
CenterTrack [10]	2020	41.8	78.1	22.6	86.8	35.7
FairMOT [19]	2021	39.7	66.7	23.8	82.2	40.8
ByteTrack* [5]	2022	47.7	71.0	32.1	89.6	53.9
FineTrack [20]	2023	52.7	72.4	38.5	89.9	59.8
OC-SORT* [18]	2023	55.1	80.3	38.3	92.0	54.6
PuTR* [22]	2024	55.8	–	–	91.9	58.2
ETTrack* [35]	2024	56.4	81.7	39.1	**92.2**	57.5
H-SORT* [17]	2024	65.7	–	–	91.8	67.4
MambaMOT* [36]	2025	56.1	80.8	39.0	90.3	54.9
C-TWiX* [23]	2025	62.1	81.8	47.2	91.4	63.6
MOTR [7]	2022	54.2	73.5	40.2	79.7	51.5
MeMOTR [8]	2023	68.5	80.5	58.4	89.9	71.2
CO-MOT [9]	2023	69.4	82.1	58.9	91.2	71.9
MOTRv2* [13]	2023	**69.9**	**83.0**	59.0	91.9	71.7
MOTIP [12]	2025	**70.0**	80.8	**60.8**	91.0	**75.1**
TBDQ-Net*	ours	69.4	80.8	**59.8**	89.5	**73.4**

detector, TBDQ-Net delivers a 37.5% faster practical inference while consuming at least 33% fewer training resources.

Comparison in DanceTrack: DanceTrack, with its aforementioned challenging factors, serves as an appropriate benchmark for the rigorous evaluation of target association. *Despite its significant efficiency improvements, TBDQ-Net achieves competitive tracking accuracy among advanced TBQ methods*, as shown in Table 3. Specifically, TBDQ-Net secures a top-3 score in the HOTA metric, highlighting its outstanding comprehensive capabilities in detection and association. When compared to MOTRv2, which utilizes the entire MOTR model for association, our proposed associator yields higher scores in AssA (+0.8) and IDF1 (+1.7), *underscoring the advantages of its more effectively MOT-tailored design in association.* For traditional TBD methods, their inherent weakness in association is amplified in such complex scenarios. TBDQ-Net effectively resolves this

Table 4: Comparison in SportsMOT. The *train* setting uses only the training subset, while *train+val* incorporates both training and validation subsets. ΔH quantifies the HOTA improvement from the *train* setting to the *train+val* setting, ***underscoring the model's ability to fully exploit larger data volumes***. The best score for each metric is marked in **blue**. All other notations are consistent with those presented in Table 3.

Methods	Year	*train*					*train+val*					ΔH
		HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑	
ByteTrack* [5]	2022	62.8	77.1	51.2	94.1	69.8	64.1	78.5	52.3	95.9	71.4	1.3
OC-SORT* [18]	2023	71.9	86.4	59.8	94.5	72.2	73.7	88.5	61.5	96.5	74.0	1.8
MixSORT* [15]	2023	–	–	–	–	–	74.1	88.5	62.0	96.5	74.4	–
ETTrack* [35]	2024	72.2	**86.9**	60.1	94.9	72.5	74.3	**88.8**	62.1	96.8	74.5	2.1
PuTR* [22]	2024	**73.0**	–	–	**95.1**	74.2	74.9	–	–	**97.0**	75.9	1.9
MambaMOT* [36]	2025	71.3	86.7	58.6	94.9	71.1	–	–	–	–	–	–
MeMOTR [8]	2023	70.0	83.1	59.1	91.5	71.4	–	–	–	–	–	–
MOTIP [12]	2025	71.9	83.4	62.0	92.9	**75.0**	75.2	86.5	65.4	96.1	**78.2**	3.3
TBDQ-Net*	ours	72.1	82.8	**62.8**	89.5	72.9	**75.5**	85.9	**66.5**	92.9	76.1	**3.4**

within the fully end-to-end framework that incorporates learnable association and query decoding mechanisms. Notably, TBDQ-Net achieves a 12.0 points improvement in AssA and at least a 6.0 points increase in IDF1, with only a slight trade-off in detection performance. Leveraging the same pretrained YOLOX weights, TBDQ-Net outperforms the best TBD method, Hybrid-SORT (H-SORT), by 3.7 points in HOTA.

Comparison in SportsMOT: This benchmark provides two evaluation settings: the general *train* setting, and the *train+val* setting that further incorporates the validation subset into training. As shown in Table 4, TBDQ-Net achieves competitive performance in HOTA and AssA under the *train* setting. Furthermore, when the volume of training data increases, TBDQ-Net gains the highest HOTA score. Crucially, TBDQ-Net demonstrates the largest HOTA improvement when transitioning from *train* to *train+val*, showcasing its robust data exploitation capabilities.

Comparison in MOT20: The severe crowd density and limited training data make MOT20 an exceptionally challenging benchmark, *particularly for query-based methods as they rely on sufficient learning data to model object relationships.* Consequently, few results for TBQ methods have been published. To bridge this gap and advance the research of query-based methods under such demanding conditions, we endeav-

Table 5: Comparison in MOT20. Methods with † are reproductions performed by our group. For MOTR, the number of queries was increased to 900 to better handle the crowded scenes and the training schedule is identical to ours, with all other settings adhering to the official configurations. MOTRv2 was reproduced using its official training configurations in the MOTChallenge benchmark. Scores marked in **blue** are the best ones among TBQ methods, ***serving to promote continued researches of query-based methods in this challenging scenario.*** All other notations are consistent with those presented in Table 3.

Methods	Year	HOTA↑	MOTA↑	IDF1↑	FP↓	FN↓	IDSW↓
FairMOT [19]	2021	54.6	61.8	67.3	103,440	88,901	5,243
CSTrack [37]	2022	54.0	66.6	68.6	25,404	144,358	3,196
PID-MOT [38]	2023	57.0	67.5	71.3	29,891	137,270	1,015
MSPNet [39]	2024	–	75.9	74.4	26,532	96,909	1,194
GLMB [40]	2024	54.2	67.7	67.3	29,597	134,534	2,911
PuTR* [22]	2024	61.1	74.8	74.1	–	–	–
H-SORT* [17]	2024	63.9	76.7	78.4	–	–	–
MOTR† [7]	2022	53.2	58.7	64.3	70,597	141,001	2,177
MeMOT [26]	2022	54.1	63.7	66.1	47,882	137,983	1,938
MOTRv2† [13]	2023	51.7	51.4	62.8	**10,111**	240,882	**604**
TBDQ-Net*	ours	**60.2**	**72.2**	**72.9**	30,950	**111,869**	1,200

ored to reproduce several state-of-the-art TBQ approaches. Our efforts confirmed that the representative methods MOTR and MOTRv2 have been successfully trained and evaluated on the official server, underscoring the inherent difficulties for TBQ methods. As shown in Table 5, TBDQ-Net achieves dominant advantages across the majority of metrics, demonstrating robust association capabilities even in this challenging scenario. Although MOTRv2 obtains the best scores in FP and IDSW, it comes at the cost of a significantly inferior FN score, leading to compromised balanced tracking performance.

Generally, traditional TBD methods exhibit superior tracking performance compared to TBQ methods. This can be attributed to two primary factors: ***(i)*** Despite the high pedestrian density, the motion patterns in the MOTChallenge benchmark tend to be linear and could be effectively resolved via intricate heuristic designs. ***(ii)*** The MOTChallenge benchmark continuously annotates and evaluates occluded, no-longer-visible pedestrians, intrinsically presenting significant hurdles to the fitting and generalization capabilities of TBQ methods. Ultimately, the issue of limited training data amplifies the

Table 6: Impact of each component within TBDQ-Net. 'self-attn' means the standard Transformer self-attention blocks.

BII	CPA	HOTA	DetA	AssA
✓		54.9	66.0	46.1
	✓	52.9	65.6	43.1
✓	✓	58.6	71.4	48.5
self-attn	✓	57.6	71.1	47.0

Table 7: Impact of different configurations for noisy detection queries (N_q), as defined in Equation (2).

Configurations	HOTA	DetA	AssA
track query	57.9	71.1	47.4
hard noisy query	58.6	71.4	48.5
easy noisy query	57.6	71.2	47.0
zeros	57.1	71.3	46.0
ones	58.1	71.1	47.8

two impacts, causing TBQ methods to perform inferior to TBD methods. *Through this work, we hope to stimulate more attentions to these challenges, enabling query-based methods to be competent in this domain.*

4.5. Ablation Studies

Analysis of TBDQ-Net Components: We assess the individual contributions of the proposed basic information interaction (BII) and content-position alignment (CPA) modules through an ablation study. When the BII module is excluded, detection queries and track queries are directly fed into the CPA module without any prior interaction. Conversely, when the CPA module is removed, the interacted object queries are immediately decoded into tracking results, lacking the crucial alignment between content and positional components. As evident from Table 6, optimal performance is attained when both modules are present, demonstrating the pivotal roles of both query interaction and content-position alignment in the overall framework. Furthermore, substituting the BII module with standard Transformer self-attention blocks results in a 1.0% drop in HOTA, suggesting the superior effectiveness of the BII module's specialized design for MOT.

Table 8: Impact of different information sources for updating track queries, specifically addressing the second term of K and V_1 in Equation (3).

Track query	Hist. track query	HOTA	DetA	AssA
✓	✓	57.3	71.6	46.2
✓		57.9	72.1	46.9
	✓	58.6	71.4	48.5

Table 9: Impact of different updating weights w in Equation (5).

Updating weight w	HOTA	DetA	AssA
0.9	58.3	72.0	47.5
0.8	58.1	71.3	47.7
0.7	58.6	71.4	48.5
0.6	58.1	71.1	47.9

Analysis of Interaction on Detection Queries: As elaborated in Section 3.3, one component of V_1 in Equation (2) incorporates noisy queries N_q to alleviate conflicts between detection queries and track queries. The impacts of various configurations for N_q are evaluated in Table 7. The first configuration uses general track queries instead of the noisy queries, resulting in degradations in HOTA (-0.7) and AssA (-1.1). Besides, alternative configurations for N_q are also explored, including easy noisy queries, an all-zero matrix, and an all-one matrix. These alternatives all yield suboptimal performance, failing to mitigate the inherent conflicts between detection queries and track queries as effectively as hard noisy queries. This is attributed to the effect of hard noisy queries, which not only disrupt potentially conflicting detection queries but also enhance the model's capacity to suppress challenging negative detections.

Analysis of Interaction on Track Queries: To disclose the impact of updating track queries with different information sources, we explore various configurations for K and V_1 in Equation (3). Detection queries, serving as the primary information source for track queries, are consistently included across all configurations. Therefore, we mainly analyze the other component of K and V_1. As shown in Table 8, suboptimal performance is achieved by either the combination of track queries and history track queries, or track queries alone. Because the former introduces redundancy in current information, while

Table 10: The performance of different values for hyperparameters τ_q, τ_n and T.

τ_q	HOTA	DetA	AssA	τ_n	HOTA	DetA	AssA	T	HOTA	DetA	AssA
0.2	57.8	71.8	46.9	0.4	57.5	70.9	47.0	10	57.3	72.1	45.9
0.3	58.6	71.4	48.5	0.5	58.6	71.4	48.5	15	57.7	71.7	46.8
0.4	58.0	71.3	47.5	0.6	58.2	70.7	48.2	20	58.6	71.4	48.5
0.5	57.4	71.9	46.2	0.7	55.6	67.6	46.1	25	57.8	71.2	47.3

Table 11: The performance of different augmentation strategies for clip length, random insert (p_i) and random dropout (p_d) in MOT20.

Clip length	HOTA	DetA	AssA	(p_i, p_d)	HOTA	DetA	AssA
6	61.4	63.3	59.8	(0.0, 0.0)	61.9	55.9	68.5
7	62.9	64.5	61.5	(0.1, 0.1)	72.3	72.5	72.2
8	70.1	70.3	70.0	(0.1, 0.2)	63.4	65.0	62.0
9	71.6	74.4	69.2	(0.2, 0.1)	71.6	74.4	69.2
10	64.7	65.5	64.1	(0.2, 0.2)	60.8	65.0	57.1

the latter suffers from a lack of crucial historical context, rendering the model less capable of robustly handling missed objects. In contrast, history track queries are able to strike an appropriate balance between current and historical information, yielding optimal performance. Furthermore, we investigate the optimal value for w in Equation (5), which controls the update weight of new information. The results in Table 9 align with the intuition that current information is more relevant for detection, while historical information enhances association. Ultimately, the highest HOTA score is achieved when w is set to 0.7.

Analysis of Other Hyperparameters: The impacts of different values for hyperparameters τ_q (Section 3.3), τ_n and T (Section 3.6) in DanceTrack are evaluated in Table 10. Although performance fluctuations are observed with different parameter settings, these variations are generally within an acceptable range. For SportsMOT and MOT20, while dataset-specific tuning of these hyperparameters may yield higher performance, we maintain consistency with the optimal values in DanceTrack for simplicity.

Ablation Studies in MOT20: As shown in Table 1, MOT20 is a markedly small dataset, posing severe challenges for the training of TBDQ-Net. To address this issue,

Figure 3: The visualization of tracking results in diverse challenging scenarios. Yellow stars signify the noteworthy individuals in each frame. Object IDs are discernible upon magnified viewing.

we increase the clip length in the clip-wise training fashion [7]. As shown in Table 11, extending the clip length substantially enhances tracking performance, as it enables the model to capture more coherent temporal cues from the limited training data, improving learning stability and fostering more robust object tracking. To augment the diversity of tracking cases, we further introduce random insertion of negative track instances with probability p_i and random dropout of positive track instances with probability p_d [7]. From Table 11, these operations effectively boost tracking performance by enriching constrained scenarios with more challenging cases.

4.6. Visualizations

Visualization of Tracking Results: The tracking results of TBDQ-Net are visualized in Figure 3. In DanceTrack, clustering and occlusion occur among dancers #246, #249, and #250; however, they are successfully tracked upon reappearance. A similar phenomenon is observed for dancer #247, who is heavily occluded by dancer #245, but is promptly re-identified when reappearing. In SportsMOT, TBDQ-Net effectively tracks players with rapid, irregular motions and visually similar appearances. In MOT20, TBDQ-Net consistently tracks pedestrian #6, despite spatial overlaps with surrounding individuals in extremely crowded environments. Similarly, it accurately differentiates be-

tween pedestrians #81 and #91. These results highlight the robust modeling capabilities of TBDQ-Net in complex scenarios.

Visualization of Attention Weights: To clarify the actual functionality of the BII module, we analyze its working mechanism by visualizing the attention weights. According to Equation (2), when updating detection queries, the attention weights can be represented as $W_d(\widetilde{D_q}, \widetilde{D_q})$ and $W_d(\widetilde{D_q}, \widetilde{T_q})$. The former corresponds to the self-attention weights among D_q, while the latter refers to the weights of noisy queries N_q, which serve to disrupt conflicting detection queries and mitigate competition between detection and track queries. Similarly, based on Equation (3), the attention weights can be divided into $W_t(\widetilde{T_q}, \widetilde{D_q})$ and $W_t(\widetilde{T_q}, \widetilde{H_q})$, through which the track queries are updated by detection queries and history track queries, respectively.

According to Figures 4b and 5b, gray cells are filled with large (top-2) attention weights, implying that these detection queries are effectively disrupted by the corresponding track queries, thereby preventing the direct competition between them. As shown in Figures 4c and 5c, track queries are primarily updated by their associated history track queries. Specifically, in Figure 4a, an inactive track query (#1) is reactivated. In this process, its detection query (#1) plays a more crucial role than general situations, which is depicted in Figure 4c. In Figure 5a, a new object appears (dancer #6). The attention weights in Figure 5b suggest that the detection query (#6) is not disrupted, so that its object can be identified as a new one.

Overall, the working process of the BII module aligns with its original motivation. *Although it may not strictly yield ideal results, such as when the values in the colored cells are not the largest in its row, the inherent robustness of the neural network ensures its effectiveness.* Furthermore, the optimized attention weights of the BII module are analogous to the affinity matrices that are used to associate detections with trajectories in TBD methods. *However, TBDQ-Net does not explicitly leverage these weights for straightforward association; it harnesses them to facilitate learnable interactions among object queries, which are subsequently decoded into tracking results, enabling the model to gain a more thorough understanding of environments and capture more robust tracking cues.*

(a) The case where a missed object (#1) is re-activated, corresponding to the green cell in subfigure (c).

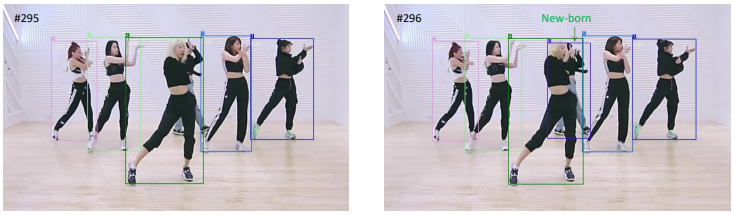
	0	1	2	3	4	5	0	1	2	3	4	5
0	0.154	0.0395	0.0524	0.0902	0.0425	0.0845	0.1436	0.0953	0.0637	0.0949	0.0522	0.0872
1	0.0675	0.1562	0.0714	0.0723	0.0682	0.0776	0.0659	0.1718	0.0613	0.0621	0.0564	0.0692
2	0.0645	0.0754	0.1468	0.0433	0.0386	0.1159	0.0767	0.148	0.1173	0.0432	0.029	0.1013
3	0.0938	0.0494	0.0314	0.1396	0.1126	0.0462	0.0936	0.104	0.0344	0.1302	0.1135	0.0511
4	0.049	0.0589	0.0237	0.1258	0.2102	0.0291	0.0571	0.1068	0.0239	0.1151	0.1667	0.0337
5	0.0931	0.0674	0.1002	0.0597	0.042	0.1117	0.0996	0.1319	0.0958	0.0601	0.0369	0.1014

(b) The illustration of $W_d(\widetilde{D_q}, \widetilde{T_q})$ (the orange column) and $W_d(\widetilde{D_q}, \widetilde{D_q})$ (the light blue column) when updating detection queries (the light blue row). The gray cells indicate that detection queries are disrupted by their corresponding track queries.

	0	1	2	3	4	5	0	1	2	3	4	5
0	0.1713	0.0636	0.076	0.1258	0.0817	0.1138	0.0912	0.0464	0.0559	0.0634	0.0495	0.0614
1	0.0965	0.2001	0.0864	0.0985	0.1219	0.0958	0.022	0.1167	0.0435	0.0316	0.0543	0.0327
2	0.103	0.0559	0.2781	0.0654	0.0502	0.1635	0.0396	0.0542	0.0838	0.0289	0.0223	0.0553
3	0.1301	0.0561	0.0463	0.1778	0.1878	0.0692	0.0606	0.0436	0.0371	0.0627	0.0855	0.0433
4	0.084	0.053	0.0395	0.1539	0.3227	0.055	0.0294	0.0434	0.0244	0.0523	0.1157	0.0266
5	0.1479	0.0572	0.174	0.0916	0.0656	0.1609	0.0517	0.0536	0.0691	0.0391	0.0344	0.0549

(c) The illustration of $W_t(\widetilde{T_q}, \widetilde{H_q})$ (the pink column) and $W_t(\widetilde{T_q}, \widetilde{D_q})$ (the light blue column) when updating track queries (the orange row). The yellow cells imply that track queries are mainly updated by their history track queries.

Figure 4: The working process on Frame #109. The attention weights of the BII module for updating detection queries (Equation (2)) and track queries (Equation (3)) are visualized in Figure 4b and Figure 4c, respectively. Rows represent the vector Q, while columns signify the vector K. Row and column indexes of tables are identical to the ID number of dancers in Figure 4a. For each query in Q, the largest weight value is marked in red, while the second and third highest are marked in light blue and green (optional).



(a) The case where a new-born object (#6) is recognized. Previously, the object was removed from the tracklets due to the long-time missing. When it re-appears, it is recognized as a new-born object, which is reflected by the green cell in subfigure (b).

	0	1	2	4	5	0	1	2	4	5	6
0	0.1391	0.0305	0.1151	0.1025	0.0482	0.142	0.0288	0.1103	0.1226	0.0545	0.1063
1	0.037	0.2362	0.0256	0.0438	0.1392	0.0339	0.1702	0.0213	0.0588	0.1089	0.1251
2	0.1408	0.0306	0.2121	0.0481	0.0357	0.1302	0.0242	0.1463	0.0834	0.0378	0.1108
4	0.0718	0.0231	0.0354	0.2773	0.0579	0.0913	0.0342	0.0467	0.207	0.0797	0.0757
5	0.0601	0.0965	0.0336	0.1169	0.134	0.0601	0.0998	0.0323	0.1126	0.1268	0.1273
6	0.0871	0.0893	0.069	0.0532	0.0963	0.077	0.0741	0.0579	0.0759	0.0881	0.2322

(b) The illustration of $W_d(\widetilde{D_q}, \widetilde{T_q})$ (the orange column) and $W_d(\widetilde{D_q}, \widetilde{D_q})$ (the light blue column) when updating detection queries (the light blue row). The green cell suggests that the detection query (dancer #6) is recognized as a new-born object.

	0	1	2	4	5	0	1	2	4	5	6
0	0.1889	0.0781	0.1617	0.1105	0.0993	0.0588	0.0415	0.0946	0.0558	0.0386	0.0722
1	0.0731	0.3521	0.0508	0.0509	0.1697	0.0269	0.1221	0.0279	0.0265	0.0382	0.0618
2	0.1798	0.0617	0.3015	0.0465	0.0735	0.0584	0.0259	0.1234	0.0346	0.029	0.0658
4	0.0763	0.0737	0.0484	0.2493	0.1287	0.0647	0.0507	0.0536	0.1482	0.0617	0.0447
5	0.0768	0.2169	0.0476	0.1298	0.1986	0.0356	0.0956	0.0372	0.0461	0.0427	0.0732

(c) The illustration of $W_t(\widetilde{T_q}, \widetilde{H_q})$ (the pink column) and $W_t(\widetilde{T_q}, \widetilde{D_q})$ (the light blue column) when updating track queries (the orange row).

Figure 5: The working process on Frame #296. Unless specifically clarified, it shares the same notations as Figure 4.

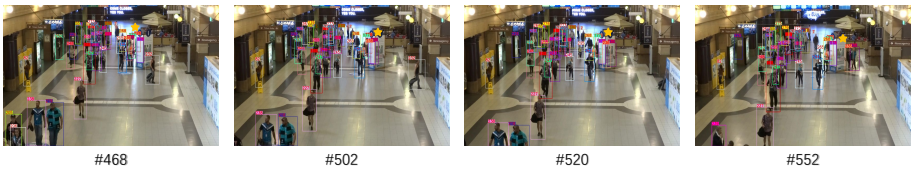


Figure 6: Visualization of the challenge in MOT20. The positions of invisible pedestrians #235 and #236 are interpolated through their trajectories before (frame #468) and after (frame #552) occlusions, leading to a more offline-like evaluation benchmark rather than an online one.

4.7. Limitation

TBDQ-Net demonstrates robust association capabilities in challenging scenarios. However, its detection performance still lags behind state-of-the-art methods, particularly TBD approaches. Therefore, further strengthening detection performance without compromising association accuracy remains a crucial direction for future research. Moreover, despite TBDQ-Net advancing query-based methods to competitive performance on MOT20, a discernible gap remains compared to state-of-the-art TBD approaches. This disparity suggests a promising avenue for more efficient data utilization strategies, such as properly integrating priors with trainable modules. Additionally, as depicted in Figure 6, the rigorous evaluation of occluded and invisible objects within the MOTChallenge benchmark poses a particular challenge for TBDQ-Net, as its object embedding can be distorted by occluders. This underscores the need for more robust query encoding techniques. Addressing these issues will substantially boost the competence of TBDQ-Net. We intend to explore these avenues in future research.

5. Conclusion

This work proposes TBDQ-Net, a novel tracking-by-detection-and-query (TBDQ) framework that seamlessly integrates the complementary strengths of the mainstream tracking-by-detection (TBD) and tracking-by-query (TBQ) paradigms. For the first time, TBDQ-Net significantly enhances the efficiency of a fully end-to-end multi-object tracking (MOT) framework, while maintaining robust tracking capabilities. This is enabled by its decoupled architecture that effectively integrates a compact MOT-tailored learnable associator with pretrained, high-performance detectors. Furthermore, this design proves instrumental in enabling query-based methods to demonstrate competence in highly crowded object tracking scenarios. We envision that TBDQ-Net, with its decoupled and end-to-end framework, alongside its outstanding balance in efficiency and accuracy, will foster advancements of MOT and broaden its accessibility to the wider computer vision community within the prevailing end-to-end era.

Acknowledgements

We are grateful for the support of the National Natural Science Foundation of China (No. 62271143), the Frontier Technologies R&D Program of Jiangsu (No.BF2024060), and the Big Data Computing Center of Southeast University.

References

[1] Y. Li, G. Yang, Z. Su, S. Li, J. Yang, L. He, Indoor scene multi-object tracking based on region search and memory buffer pool, Pattern Recognition 165 (2025) 111623.

[2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, H. Li, Planning-oriented autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

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

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

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

[6] T. Mandel, M. Jimenez, E. Risley, T. Nammoto, R. Williams, M. Panoff, M. Ballesteros, B. Suarez, Detection confidence driven multi-object tracking to recover reliable tracks from unreliable detections, Pattern Recognition 135 (2023) 109107.

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

[8] R. Gao, L. Wang, Memotr: Long-term memory-augmented transformer for multi-object tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9901–9910.

[9] F. Yan, W. Luo, Y. Zhong, Y. Gan, L. Ma, Bridging the gap between end-to-end and non-end-to-end multi-object tracking, arXiv preprint arXiv:2305.12724 (2023).

[10] X. Zhou, V. Koltun, P. Krähenbühl, Tracking objects as points, in: European conference on computer vision, Springer, 2020, pp. 474–490.

[11] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, arXiv preprint arXiv:2107.08430 (2021).

[12] R. Gao, J. Qi, L. Wang, Multiple object tracking as id prediction, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 27883–27893.

[13] Y. Zhang, T. Wang, X. 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, 2023, pp. 22056–22065.

[14] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, P. Luo, Dancetrack: Multi-object tracking in uniform appearance and diverse motion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20993–21002.

[15] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, L. Wang, Sportsmot: A large multi-object tracking dataset in multiple sports scenes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 9921–9931.

[16] P. Dendorfer, Mot20: A benchmark for multi object tracking in crowded scenes, arXiv preprint arXiv:2003.09003 (2020).

[17] M. Yang, G. Han, B. Yan, W. Zhang, J. Qi, H. Lu, D. Wang, Hybrid-sort: Weak cues matter for online multi-object tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 6504–6512.

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

[19] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, Fairmot: On the fairness of detection and re-identification in multiple object tracking, International journal of computer vision 129 (2021) 3069–3087.

[20] H. Ren, S. Han, H. Ding, Z. Zhang, H. Wang, F. Wang, Focus on details: Online multi-object tracking with diverse fine-grained representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 11289–11298.

[21] X. Zhou, T. Yin, V. Koltun, P. Krähenbühl, Global tracking transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8771–8780.

[22] C. Liu, H. Li, Z. Wang, R. Xu, Putr: A pure transformer for decoupled and online multi-object tracking, arXiv preprint arXiv:2405.14119 (2024).

[23] M. Miah, G.-A. Bilodeau, N. Saunier, Learning data association for multi-object tracking using only coordinates, Pattern Recognition 160 (2025) 111169.

[24] J. Ding, W. Li, M. Yang, Y. Zhao, L. Pei, A. Tian, Seatrack: Rethinking observation-centric sort for robust nearshore multiple object tracking, Pattern Recognition 159 (2025) 111091.

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

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

[27] E. Yu, T. Wang, Z. Li, Y. Zhang, X. Zhang, W. Tao, Motrv3: Release-fetch supervision for end-to-end multi-object tracking, arXiv preprint arXiv:2305.14298 (2023).

[28] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang, Conditional detr for fast training convergence, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3651–3660.

[29] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, DAB-DETR: Dynamic anchor boxes are better queries for DETR, in: International Conference on Learning Representations, 2022.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

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

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

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

[34] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, Crowdhuman: A benchmark for detecting human in a crowd, arXiv preprint arXiv:1805.00123 (2018).

[35] X. Han, N. Oishi, Y. Tian, E. Ucurum, R. Young, C. Chatwin, P. Birch, Ettrack: enhanced temporal motion predictor for multi-object tracking, Applied Intelligence 55 (1) (2025) 1–17.

[36] H.-W. Huang, C.-Y. Yang, W. Chai, Z. Jiang, J.-N. Hwang, Mambamot: State-space model as motion predictor for multi-object tracking, ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2024).

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

[38] W. Lv, N. Zhang, J. Zhang, D. Zeng, One-shot multiple object tracking with robust id preservation, IEEE Transactions on Circuits and Systems for Video Technology (2023).

[39] Y. Zheng, H. Qi, L. Li, S. Li, Y. Huang, C. He, D. Wang, Motion-guided and occlusion-aware multi-object tracking with hierarchical matching, Pattern Recognition 151 (2024) 110369.

[40] L. Van Ma, T. T. D. Nguyen, C. Shim, D. Y. Kim, N. Ha, M. Jeon, Visual multi-object tracking with re-identification and occlusion handling using labeled random finite sets, Pattern Recognition 156 (2024) 110785.