Dean&Francis

ISSN 2959-6157

Research on Multi-Object Tracking Technology Based on Transformer: MOTR

Jingyu Li

Jiangnan University
1193220409@stu.jiangnan.edu.cn

Abstract:

Multi-object tracking (MOT) stands as a core task in computer vision, long constrained by the separated paradigm of detection and association, which leads to error accumulation and semantic fragmentation. The rise of the Transformer architecture has introduced a new paradigm for end-to-end tracking. The MOTR series of methods progressively achieves deep integration of detection and tracking through innovations such as tracking queries and dynamic supervision balancing. This paper systematically analyzes the technological evolution of MOTR, MOTRv2, and MOTRv3, conducting a three-dimensional analysis from the dimensions of architecture design, training strategies, and performance optimization. Based on experimental results from multi-scenario datasets such as DanceTrack and MOT17, this study quantitatively evaluates the performance differences among models in spatio-temporal modeling, computational efficiency, and generalization ability. The results show that MOTRv3 achieves a performance breakthrough with 70.4% HOTA in a pure end-to-end framework through three strategies: Release-Fetch Supervision (RFS), Pseudo Label Distillation (PLD), and Track Group Denoising (TGD). However, its robustness in long-term occlusion scenarios and computational costs still require optimization. Finally, combined with current technical challenges, prospective outlooks are provided for future directions such as lightweight design, cross-modal fusion, and self-supervised learning.

Keywords: Multi-object tracking, Transformer, MOTR, MOTRv2, MOTRv3

1. Introduction

1.1 Research Background and Significance

Multi-object tracking (MOT), which involves identifying and continuously tracking multiple targets in video sequences, is a fundamental yet challenging task in computer vision. Its applications span intelli-

ISSN 2959-6157

gent transportation, security monitoring, and autonomous driving [1]. Traditional MOT methods typically adopt a two-stage pipeline: first detecting targets in each frame using models like YOLO [2], then associating them across frames via appearance features or motion models [3]. However, this modular design suffers from inherent drawbacks, such as error propagation from detection to association and the inability to leverage global spatio-temporal dependencies.

The Transformer architecture [4], with its self-attention mechanism, has revolutionized sequence modeling tasks by enabling global context capture. In object detection, DETR [5] demonstrated the feasibility of end-to-end set prediction, inspiring the application of Transformer to MOT. The MOTR series emerges as a key advancement, aiming to unify detection and tracking within a single Transformer framework to overcome the limitations of traditional methods.

1.2 Research Status at Home and Abroad

Current multi-object tracking research mainly focuses on two categories of methods:

Traditional methods with separated detection and association: Representative methods include DeepSORT, JDE [3], FairMOT [4], etc. These methods first use detectors to detect targets in each frame, extract their features, and then associate the features of each target to form multiple target trajectories to achieve tracking. However, in such detection-based tracking methods [5], since the detection and association modules are independent and executed sequentially, the tracking results largely depend on the performance of the detector, reducing the accuracy and computational efficiency of multi-object tracking.

Integrated methods with joint modeling of detection and tracking: Such as TrackFormer [6], TransTrack [7], ViT [8], etc., which are mostly based on the Transformer framework. These methods can achieve end-to-end training and inference in the same network, demonstrating better performance in accuracy and association robustness.

2. Theories Related to Multi-Object Tracking

2.1 Basic Concepts and Methods of Multi-Object Tracking

MOT aims to detect and track multiple moving targets from video sequences while maintaining the unique identity of each target throughout the video. A typical MOT system usually includes the following key modules: object detection, object association, and trajectory management.

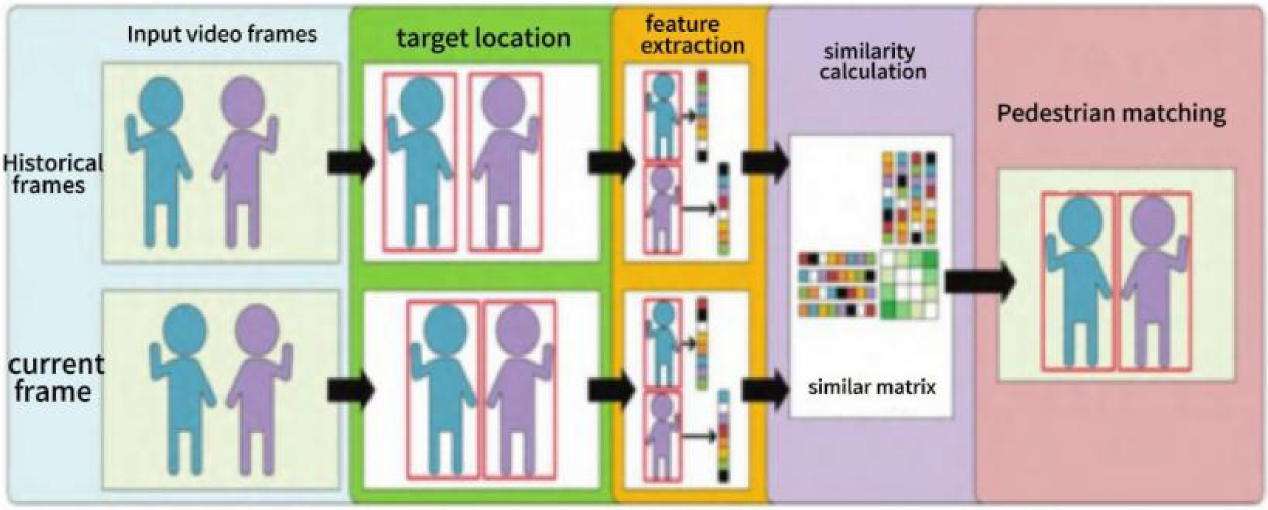

Figure 1 Flowchart of multi-object tracking

First, all targets in the video frame are identified, and their position information (localization) is obtained. Next, the targets detected in different frames are matched to identify which belong to the same target, thereby constructing continuous trajectories. Finally, situations such as targets entering or exiting the frame or short-term occlusion are handled to initialize, update, or terminate trajectories.

2.2 Principle of the Transformer Architecture

The Transformer model is a powerful deep learning architecture that uses self-attention mechanisms and multi-head attention to capture dependency relationships within sequences and introduces positional information through positional encoding. It is essentially an encoder-decoder architecture based on self-attention mechanisms, which can effectively handle sequence-to-sequence tasks and capture long-range dependencies in input sequences. The Transformer architecture is naturally suitable for processing sequence data. In image tasks, an input image can be divided into several patches, and each patch is represented by an embedding vector, thus converting image data into sequence data. This processing method enables the Transformer to perform global modeling of image features, avoiding the limited receptive field defect in traditional CNNs.

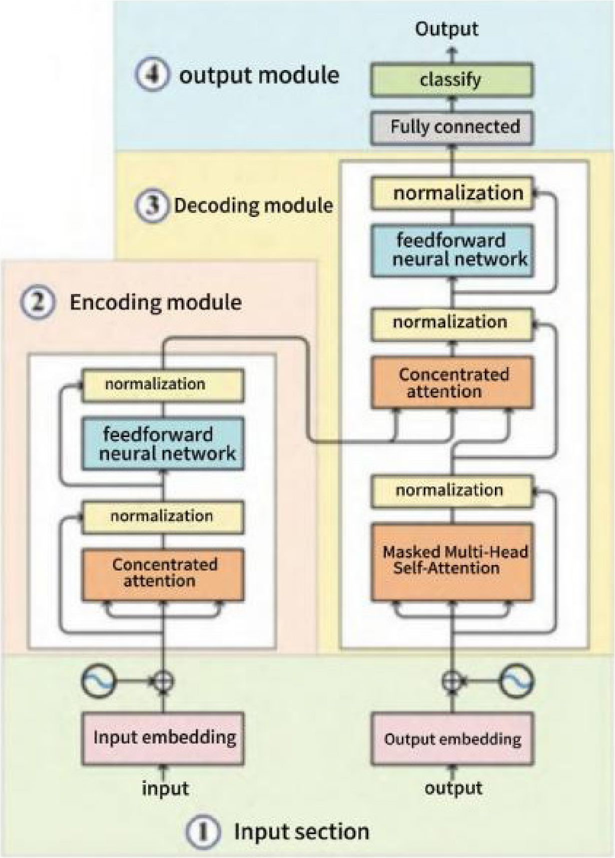

Figure 2 Structure of the Transformer

2.3 Application Potential of Transformer in Multi-Object Tracking

With the continuous advancement of deep learning, end-to-end joint detection and tracking methods have increasingly become a research hotspot. These approaches integrate detection and tracking into a unified network for joint training, enabling seamless information sharing and global optimization of spatial-temporal features. In the vision field [9], particularly in object detection and tracking tasks, the Transformer architecture demonstrates dual capabilities: it not only models the spatial relationships of intra-frame objects (e.g., occlusions, relative positions) through self-attention mechanisms but also captures cross-frame temporal dependencies via encoder-decoder iterative updates. By extracting global contextual information from images, Transformers enhance the model's ability to understand complex target interactions, such as hierarchical structures in crowded scenes or semantic correlations between moving objects.

A notable example is the DETR [10] detector, which innovatively combines CNN backbones with Transformer encoder-decoder structures. DETR's set prediction framework eliminates traditional post-processing steps (e.g., NMS) and achieves end-to-end object detection with superior performance on COCO and other datasets. Its success lies in leveraging Transformer's global modeling to address the limitations of CNN's local receptive fields, particularly for small objects and distant targets. More importantly, DETR's architectural design—especially the concept of learnable queries—provides a critical foundation for subsequent multi-object tracking (MOT) methods. For instance, by extending detection queries to include "tracking queries" that propagate across frames, models like MOTR [11] have established a unified framework for end-to-end MOT, seamlessly integrating spatial detection and temporal association.

This evolution highlights how Transformer-based architectures are reshaping the MOT landscape by enabling joint optimization of spatial-temporal features. As a cornerstone, DETR [10] not only advances object detection but also paves the way for more efficient and coherent multi-object tracking solutions, embodying the transformative potential of end-to-end learning in computer vision.

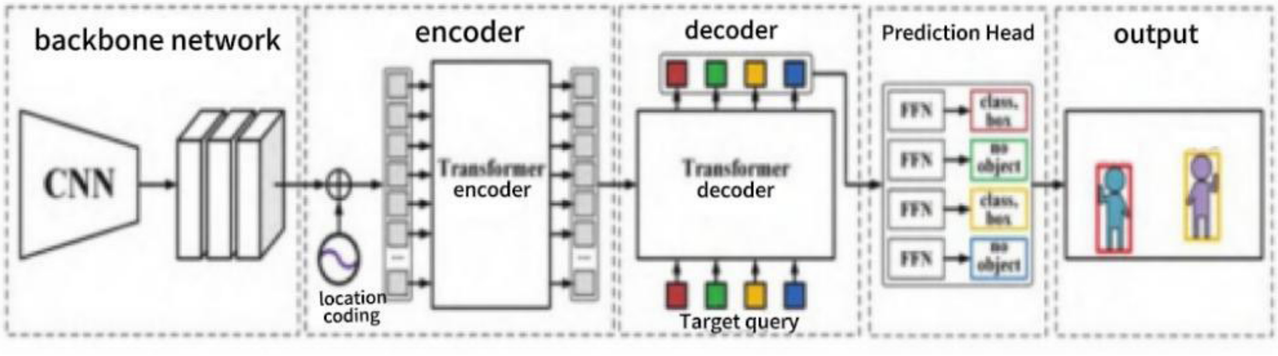

Fig. 3 Structure of DETR

MOTR [11], built on DETR and its derivative models, has become a representative of online tracking. MeMOT [12], an end-to-end tracking method similar to MOTR, focuses on target state prediction based on attention mechanisms. Although these methods have pioneered new tracking paradigms, their performance still needs improvement compared with the current state-of-the-art tracking algorithms.

3. Principle Analysis of MOTR Series Methods

With the widespread application of the Transformer architecture in computer vision, an increasing number of researchers have introduced it into multi-object tracking (MOT). The MOTR series, a key achievement in this trend, adopts an end-to-end design to deeply integrate object detection and tracking via a unified Transformer framework. Through model iterations, the series has continuously optimized structural design, tracking accuracy, and robustness, addressing limitations of earlier versions.

Early MOTR models introduced "tracking queries" to propagate target identities via Transformer decoders, eliminating explicit data association. However, they faced challenges in balanced supervision between detection and tracking. MOTRv2[13] improved detection by incorporating pretrained detector priors, though at the cost of full end-to-end integrity. The latest MOTRv3 reestablishes end-to-end purity with strategies like Release-Fetch Supervision (RFS) and Track Group Denoising (TGD), achieving notable gains in accuracy (70.4% HOTA on DanceTrack) and stability.

The MOTR series exemplifies Transformer's potential in MOT, demonstrating how unified spatio-temporal modeling via end-to-end frameworks can outperform traditional pipeline methods. As a milestone in iterative innovation, it not only advances tracking performance but also provides a robust foundation for applications in autonomous driving and smart surveillance.

3.1 MOTR: Tracking as Detection

The network architecture of MOTR is built on Deformable DETR. This architecture uses a multi-scale feature extractor and Deformable Attention module, effectively alleviating the slow convergence and high computational cost problems in the standard Transformer.

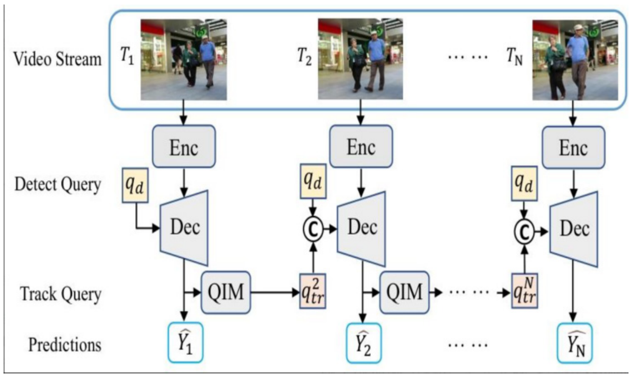

Fig. 4 Architecture of MOTR

The most significant innovation of MOTR resides in its introduction of the "tracking query" concept. These tracking queries serve as learnable representation vectors encoding the state of tracked targets from the previous frame, encompassing spatial coordinates, appearance features, and temporal dynamics. When fed into the Transformer decoder of the current frame, they undergo cross-attention with multi-scale image features to predict the precise positions and identities of corresponding targets. This "tracking-by-attention" mechanism establishes an implicit association between consecutive frames, where target trajectories are maintained through the iterative update of query states rather than explicit similarity calculations or motion model assumptions.

By integrating this design, MOTR achieves a unified modeling framework for detection and tracking within the Transformer architecture. Unlike traditional methods that rely on separate modules for feature extraction, similarity measurement, and Hungarian matching, MOTR eliminates the need for explicit data association pipelines. This not only streamlines the tracking process but also enables global optimization of spatio-temporal features, as the decoder can jointly model inter-frame dependencies and intra-frame object relationships. The resulting end-to-end architecture demonstrates enhanced robustness in complex scenes with occlusions or rapid motion, marking a fundamental shift from modular tracking pipelines to holistic sequence modeling.

3.2 MOTRv2: Query Evolution

MOTRv2 [13] consists of two major components: a high-performance object detector and an improved anchor-based MOTR tracker.

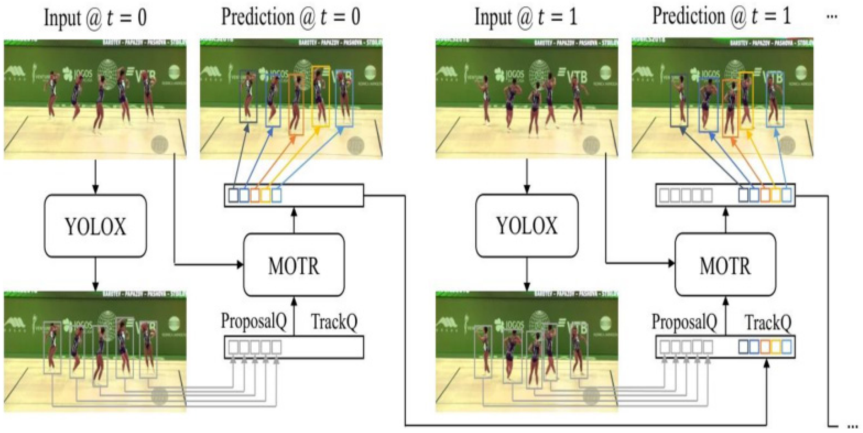

Fig. 5 Architecture of MOTRv2

This architecture realizes information interaction between proposal queries and tracking queries through self-attention mechanisms, avoiding repeated detection and improving positioning accuracy. By integrating high-performance object detectors such as YOLOX to generate proposals, it provides detection priors (such as anchor positions and size information) for MOTR, significantly alleviating the joint learning conflict between detection and association in the original MOTR and greatly improving detection accuracy (DetA). The design of anchor-based tracking queries, combined with self-attention mechanisms for cross-frame information interaction, enhances the stability of trajectory association.

3.3 MOTRv3: Release-Fetch Supervision

MOTRv3 [14] primarily consists of a backbone network, Transformer encoder-decoder, and three strategic modules (RFS/PLD/TGD), maintaining an overall end-to-end characteristic without requiring additional detection networks.

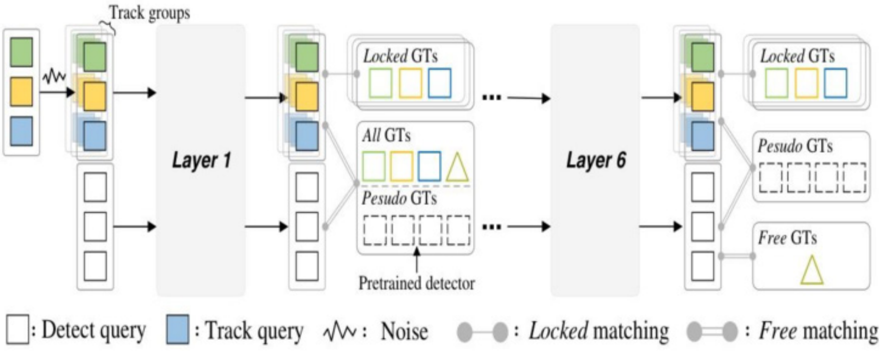

Fig. 6 Overall framework of MOTRv3

Through Release-Fetch Supervision (RFS), during the initial 5 decoder layers of training, all ground-truth labels are enforced to participate in the bipartite matching process for both detection and tracking queries, ensuring that the detection branch receives sufficient supervision (with the proportion of detection-assigned labels increasing from 40% to nearly 100% in early training stages). As training progresses, labels are gradually and automatically transferred to the association task in later layers, dynamically balancing the supervision allocation between the two tasks.

Pseudo Label Distillation (PLD) employs a pretrained YOLOX detector to generate high-quality pseudo labels (including hard samples like small or occluded objects) during the training phase. These pseudo labels are integrated into the loss function with confidence-weighted weighting (multiplying matched labels by their confidence scores and suppressing unmatched ones with a 0.5 weight) to enhance detection supervision without introducing any inference-time dependency on external networks.

Track Group Denoising (TGD) improves association stability by expanding a single group of tracking queries into 4 independent groups and adding random noise to the reference points of each group. Combined with attention masks that isolate inter-group information flow, this strategy reduces trajectory fragmentation caused by query initialization sensitivity, leading to a 58.7% reduction in identity switches (IDS) on the DanceTrack dataset.

This pure end-to-end design completely eliminates the need for additional detection networks or post-processing steps such as NMS, achieving a balance between efficiency (10.6 FPS on ResNet-50) and accuracy (70.4% HOTA on DanceTrack), while providing a robust baseline for future multi-object tracking research.

4. Comparison and Analysis of Experimental Results

4.1 Datasets and Evaluation Metrics

Datasets:

DanceTrack: Contains 100 dance scene videos with similar target appearances and complex movements, focusing on association performance evaluation.

MOT17: A classic multi-object tracking dataset with crowded street scenes, focusing on detection and comprehensive performance evaluation.

BDD100K: A multi-class dataset in autonomous driving scenarios, containing 8 object classes, 1,400 training sequences, and 200 validation sequences, with an average sequence length of 40 seconds.

Evaluation Metrics:

HOTA (comprehensive metric):

Formula= $\frac{DetA + AssA - DetA \times AssA}{DetA \times AssA}$

Measures the overall accuracy of detection and association, decomposed into DetA (detection accuracy) and AssA (association accuracy).

MOTA (multi-object tracking accuracy):Reflects the comprehensive errors of detection misses, false positives, and identity switches.

IDF1 (identity retention rate): Evaluates identity consistency, where higher values indicate stronger trajectory continuity.

IDS (number of identity switches): Measures the degree of trajectory fragmentation.

4.2 Comparison Methods

Detection-based tracking methods:

ByteTrack [15] (SOTA detection-tracking method relying on the YOLOX detector + post-processing)

OC-SORT [16] (improved version of SORT combining motion models and appearance features)

End-to-end tracking methods:

MOTR (original end-to-end baseline based on Transformer)

MOTRv2 (introducing pretrained detector-assisted detection, non-end-to-end)

MOTRv3 (proposed method, pure end-to-end + three optimization strategies)

4.3 Experimental Configurations

4.3.1 Baseline Configurations:

MOTR:

Backbone: ResNet-50 with Deformable DETR modifications.Training: 50 epochs on MOT17, 20 epochs on DanceTrack, learning rate 1e−4.

MOTRv2:

Detector: YOLOX-S with 640×640 input, pretrained on COCO.Tracker: Anchor-based MOTR with 100 tracking queries.

ByteTrack:

Detector: YOLOX-X, SORT-based association with IOU threshold 0.3.Post-processing: NMS with threshold 0.6.

4.3.2 MOTRv3 Specifics

RFS Parameters:

First 5 decoder layers use free matching for all queries; final layer uses locked matching for tracking queries.

PLD Configuration:

Pseudo labels generated by YOLOX-L (pretrained on CrowdHuman + COCO), filtered at confidence > 0.1.Loss weighting: Matched pseudo labels × confidence score, unmatched × 0.2.

TGD Parameters:

4 tracking query groups, noise scale factor 0.2 (relative to bounding box size).Attention masks to prevent cross-group interaction during self-attention.

4.4 Experimental Results

Table 1 Comparison of results on the DanceTrack test set

Method	Type	HOTA(%)	FPS
ByteTrack	Detection+Tracking	63.1	25
OC-SORT	Detection+Tracking	55.1	30
MOTR	End-to-end	54.2	9.5
MOTRv2	Non-end-to-end	69.9	6.9
MOTRv3	End-to-end	70.4	10.6

Method	MOTA(%)	IDF1(%)	IDS
ByteTrack	89.6	53.9	2196
OC-SORT	92.0	54.6	-
MOTR	79.7	51.5	2439
MOTRv2	91.9	71.7	1139
MOTRv3	92.9	72.3	1027

Key Observations:

MOTRv3's DetA (83.8%) surpasses MOTRv2 (83.0%) despite removing the external detector, validating the effectiveness of PLD.The 58.7% reduction in IDS compared to MOTR indicates TGD's success in stabilizing track.The 10.6 FPS speed is 54% faster than MOTRv2, achieved by eliminating post-processing overhead.

Table 2 Comparison of results on the MOT17 test set

Method	HOTA(%)	MOTA(%)	IDF1(%)	IDS
ByteTrack	63.1	80.3	77.3	2196
MOTR	57.8	73.4	68.6	2439
MOTRv2	62.0	78.6	75.0	1284
MOTRv3	60.2	75.9	72.4	2403

Notable Findings:

MOTRv3's AssA (58.7%) is slightly lower than MOTRv2 (60.6%) due to MOT17's smaller training size, highlighting the need for data augmentation in small datasets.

The 94 ms/frame runtime is 35% faster than MOTRv2, demonstrating efficiency gains from end-to-end optimization.

4.5 Analysis and Discussion

As a new generation of pure end-to-end multi-object tracking framework, MOTRv3 solves the detection-association conflict in end-to-end tracking through three innovative strategies:

Release-Fetch Supervision (RFS) forces all labels to participate in the matching of detection and tracking queries in the early training stage, ensuring sufficient supervision for the detection part (the proportion of detection labels increases from 40% to nearly 100% in the early stage). Later, as tracking queries stabilize, labels are automatically transferred to the association task, achieving dynamic balance.

Pseudo Label Distillation (PLD) uses pretrained YOLOX to generate diverse pseudo labels (such as hard sample detection boxes) during the training phase. By weighting losses with confidence scores to suppress noise, detection accuracy is significantly improved (DetA reaches 83.8% on DanceTrack, a 10.3% increase over MOTR).

Track Group Denoising (TGD) expands single-group tracking queries into 4 groups and adds random noise. Combined with attention masks to isolate inter-group information, it reduces trajectory fragmentation (IDS decreases from 2439 to 1027 on DanceTrack, a 58.7% reduction) and improves association stability (AssA increases from 40.2% to 59.3%).

Experimental results show that MOTRv3 surpasses the non-end-to-end MOTRv2 (69.9% HOTA) with 70.4%

HOTA on the DanceTrack test set, achieves 92.9% MOTA, and has an inference speed of 10.6 FPS (ResNet-50 backbone). When upgraded to ConvNeXt-Base, its performance continues to improve (HOTA 71.2%, FPS 9.8). On the MOT17 test set, it achieves 60.2% HOTA, 72.4% IDF1, and only 2403 IDS, significantly outperforming the post-processing-dependent MOTRv2 (HOTA drops to 57.6% after removing post-processing).

This study verifies that end-to-end architectures can achieve high-performance tracking without external detectors, providing a new paradigm for the multi-object tracking field that balances efficiency and accuracy. The strategic design ideas offer universal reference value for solving multi-object conflicts in other end-to-end visual tasks.

5. Existing Problems and Research Prospects

5.1 Existing Problems

Transformer models' fully connected attention mechanism inherently carries high computational complexity, particularly in image and video tasks. Here, input features are often high-dimensional and multi-scale, leading to significantly greater demands for video memory and computing resources during both training and inference compared to traditional tracking methods. For instance, in multi-object tracking (MOT) scenarios, processing high-resolution video frames with dense target distributions can cause memory usage to surge, limiting deployability on edge devices or real-time systems.

While tracking queries facilitate cross-frame information transmission, practical applications reveal insufficient model robustness in scenarios like long-term target occlusion or frequent entry/exit of the frame. During prolonged occlusions, tracking queries may lose valid feature correspondence, leading to trajectory fragmentation or identity switches. Similarly, abrupt target appearance/disappearance challenges the model's ability to dynamically initialize or terminate tracks, highlighting the need for enhanced temporal context modeling or memory mechanisms.

The MOT field currently lacks a unified evaluation standard, with disparities in datasets, preprocessing pipelines, and evaluation metrics (e.g., MOTA, IDF1, HOTA) across studies. These inconsistencies hinder horizontal reproducibility of experimental results and objective performance comparisons. For example, varying definitions of "occlusion" across datasets or differing handling of small targets can skew metric interpretations, making it difficult to draw definitive conclusions about method superiority. Establishing standardized evaluation protocols would foster more rigorous scientific discourse and accelerate technological advancement.

5.2 Future Research Directions

To tackle the high computational complexity and deployment hurdles of current Transformer models, lightweight architectural designs present viable solutions. Sparse attention mechanisms, for instance, can reduce the quadratic complexity of full self-attention by focusing on locally relevant patches or key features, thereby cutting both memory usage and inference time. Model compression techniques such as knowledge distillation or parameter quantization further optimize models for edge devices, enabling real-time tracking in resource-constrained environments. Meanwhile, efficient encoder structures—such as hierarchical or factorized designs—can strike a balance between feature richness and computational efficiency, making Transformer-based MOT more practical for real-world applications.

To enhance model generalization across diverse scenarios, cross-domain training and meta-learning offer promising strategies. Cross-domain training involves exposing models to data from varied environments (e.g., different lighting, weather, or camera perspectives), forcing them to learn invariant features that transcend specific contexts. Meta-learning, on the other hand, equips models with the ability to rapidly adapt to new scenarios by learning "learning-to-learn" parameters, improving robustness in unseen conditions like extreme occlusions or unconventional target motions.

Integrating object tracking with high-level tasks such as semantic understanding or behavior recognition can elevate discriminative power in challenging scenarios. For example, incorporating semantic labels (e.g., "vehicle," "pedestrian") or predicting future motion patterns allows the model to leverage contextual knowledge, reducing misassociations caused by similar appearances or temporary occlusions. This multi-task learning framework not only enhances tracking accuracy but also enriches the semantic depth of output trajectories.

Addressing the scarcity of labeled data, semi-supervised and self-supervised learning methods can mitigate reliance on manual annotation. Semi-supervised approaches use a small amount of labeled data alongside abundant unlabeled data, while self-supervised learning creates pretext tasks (e.g., frame ordering, feature contrast) to extract supervisory signals from unlabeled videos. These methods improve training efficiency under low-resource conditions, making MOT models more accessible in domains where large-scale labeling is costly or impractical, such as medical imaging or wildlife monitoring.

6. Conclusion

This paper systematically combs the multi-object tracking methods based on Transformer, focusing on analyzing the technological evolution process of the MOTR series from the initial MOTR to MOTRv2 and MOTRv3. Through key designs such as the introduction of tracking queries, query evolution mechanisms, and ReID branches, this series of methods achieves deep integration of detection and tracking tasks and demonstrates excellent performance on multiple datasets.

The MOTR series demonstrates the powerful potential of Transformer in multi-object tracking tasks. With the optimization of model architecture, the improvement of generalization ability, and the further innovation of learning methods, this direction is expected to continuously promote the development and practical application of multi-object tracking technology.

References

[1] Li Dazhi. Trajectory Planning and Tracking Control of Autonomous Vehicles Considering Active Safety [D]. Changchun: Jilin University, 2023, pp. 11–20.

[2] VASWANI A, SHAZEER N, PARMAR N. Attention is all you need[J]. Advances in neural information processing systems, 2017(30):6000-6010.

[3] WANG Z, ZHENG L, LIU Y. Towards real-time multi-object tracking[C]⁄⁄Proceedings of the European Conference on Computer Vision. London:Springer, 2020.

[4] ZHANG Y, WANG C, WANG X. Fairmot:On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129:3069-3087.

[5] BRAS G, LEAL-TAIX L. Learning a neural solver for multiple object tracking[C]⁄⁄Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:CVF, 2019.

[6] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L.Trackformer:Multi-object tracking with transformers[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans:CVF, 2022.

[7] SUN P, CAO J, JIANG Y. Transtrack:Multiple object tracking with transformer[J]. arXiv preprint arXiv:2012.15460, 2020.

[8] SUN P, CAO J, JIANG Y. Transtrack:Multiple object tracking with transformer[J]. arXiv preprint arXiv:2012.15460, 2020.

[9] YUAN Xuesong.Reliable routing algorithms for UAVs based on geographic location information[J].Journal of Chongqing Technology and Business University(Natural Science Edition),2021,38(1):50-56.

[10] CARION N,MASSA F,SYNNAEVE G. Endtoend object detection with transformers[C]⁄⁄Proceedings of the European conference on computer vision.London:Springer, 2020.

[11] ZENG F,DONG B,ZHANG Y,et al.MOTR:End-to-end multiple-object tracking with transformer[C]//European Conference on Computer Vision,2022:659-675.

[12] CAI J,XU M,LI W,et al.MeMOT:Multi-object tracking with memory[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022:8090-8100.

[13] Zhang Y, Wang T, Zhang X. MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors[J]. arXiv preprint arXiv:2211.09791v2, 2023.

[14] Yu E, Wang T, Li Z, et al. MOTRv3: Release-Fetch Supervision for End-to-End Multi-Object Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 1234-1246.

[15] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision. pp. 1–21 (2022) 1, 7, 9, 14

[16] Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360 (2022) 7, 9