Research and Analysis of Multi object Tracking Combined with Transformer Method

Junhao Yu

School of International Education, Jiangsu University of Technology, Jiangsu, China
2022705107@smail.jsut.edu.cn

Abstract. Multi-object tracking (MOT) is widely used in intelligent transportation, public safety, and autonomous driving. Traditional MOT algorithms, relying on local feature modeling and heuristic association rules, have reached their performance limits. This leads to significant performance degradation in complex scenarios such as object occlusion, identity switching (IDSW), and scale variations. In recent years, the Transformer architecture, with its global self-attention mechanism and long-term dependency modeling capabilities, has established a new paradigm for MOT, achieving end-to-end collaborative optimization of object detection and identity mapping. This paper provides a comprehensive overview of Transformer-based MOT research. First, it describes the task definition and key evaluation metrics, and compares traditional MOT methods, CNN-based methods, and Transformer-based solutions. Then, it analyzes the core innovations in three key areas: feature extraction, decoder design, and data association. The paper focuses on techniques such as bi-branch feature fusion, spatial location constraints, linear attention optimization, and multi-scale object modeling. Furthermore, this paper utilizes benchmark datasets such as MOT17, MOT20, KITTI, and VisDrone-MOT2019 to compare representative Transformer-based algorithms in terms of tracking accuracy (MOTA, IDF1, HOTA) and real-time performance (frame rate). Finally, this paper summarizes the unresolved challenges and future research directions in current research. The aim of this work is to lay a systematic foundation for the theoretical research and practical applications of Transformer-based MOT.

Keywords: Multi-object tracking; Transformer; Feature fusion; Data association; Occlusion handling

1 Introduction

Multi-object tracking (MOT) has become one of the most important research directions in computer vision. Its goal is to continuously detect and identify multiple objects in a video sequence while maintaining the object's identity. This task plays a crucial role in intelligent transportation systems, public safety, autonomous driving, and video surveillance systems. However, complex environmental conditions such as occlusion, lighting variations, and changes at different scales pose significant challenges to the

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_90

performance of MOT. Traditional MOT methods typically rely on local feature modeling and heuristic motion models, which limits their robustness and scalability in real-world scenarios.

With the success of deep learning, convolutional neural networks (CNNs) have demonstrated powerful feature extraction capabilities. However, CNNs inherently focus on local receptive fields, making it difficult to capture long-range dependencies and global contextual relationships. To overcome this limitation, the Transformer architecture, originally developed for Natural Language Processing (NLP), has been applied to computer vision due to its self-attention mechanism and global dependency modeling capabilities. This evolution has given rise to a new research paradigm—Transformer-based motion tracking (MOT), which enables the co-optimization of detection and association within a single, unified framework.

Recent studies have highlighted the potential of Transformer-based architectures to improve the accuracy and robustness of Motion Detection and Recognition (MOT). Feng et al. proposed the CNA DeepSORT algorithm, which combines convolutional features with channel neighborhood attention mechanisms, effectively improving identity consistency and occlusion handling capabilities [1]. Zhang et al. designed a CNN-Transformer hybrid model (CTMOT), which integrates local CNN and global Transformer features through a bidirectional bridging module, achieving state-of-the-art performance on the MOT17 and KITTI datasets while maintaining real-time performance [2]. Li et al. developed a Transformer-based multi-task branching algorithm (TMTB) for wide-angle UAV scenarios, which improves multi-scale feature rendering and motion detection [3]. Wu et al. proposed the CrossUnihead framework, which achieves feature decoupling and task interaction in the detection, classification, and re-identification branches [4]. Furthermore, Yang et al. proposed a Transformer-based dual-source motion model that integrates target and camera motion information to improve tracking accuracy in dynamic environments [5]. These representative studies collectively demonstrate that the Transformer architecture possesses powerful global context modeling capabilities and supports multi-task collaborative learning, thereby significantly improving detection accuracy and tracking consistency.

Therefore, this paper provides a comprehensive overview of Transformer-based multi-object tracking (MOT) algorithms. First, the development history of MOT is discussed, and the technical differences between traditional methods, CNN-based methods, and Transformer-based methods are compared. Subsequently, design innovations and optimization strategies in feature extraction, decoder architecture, and data association modules are analyzed. Finally, this paper uses benchmark datasets such as MOT17, MOT20, KITTI, and VisDrone-MOT2019 to compare representative algorithms in terms of accuracy (MOTA, IDF1, HOTA) and real-time efficiency (FPS). Furthermore, this paper discusses current challenges and explores future research directions, including developing resource-efficient Transformer models, multimodal fusion, and end-to-end collaborative optimization

2 Overview of Related Technologies

2.1 Core Principles of Transformer

Based on a self-attention mechanism, Transformer models global feature dependencies by mapping and interacting with queries, keys, and values. Unlike convolutional neural networks (CNNs), which are limited by local receptive fields, the Transformer can capture a wider range of dependencies and has higher parallel computing capabilities.

IIn the field of computer vision, Transformer uses structural modifications such as Patch Embedding and Positional Encoding to serialize 2D images into feature sequences that can be processed by attention mechanisms, thereby achieving global modeling in the spatial dimension.

To address the characteristics of vision tasks, various improved architectures have emerged, such as DeiT, Swin Transformer, and Deformable Transformer, which have been optimized in terms of hierarchical feature representation and spatial adaptability.

2.2 Basic Framework of Visual Tracking

Visual target tracking tasks can be divided into two categories: Long-Term Tracking (LTT) and Multi-Object Tracking (MOT).

Long-term tracking (LTT) primarily deals with complex scenarios such as the long-term disappearance and deformation of targets. Its core processes include feature extraction, target modeling, template updates, and position prediction.

The main objective of multi-target tracking (MOT) is to achieve robust target recognition and persistent target identity preservation simultaneously in continuous video sequences. Its core components include multi-target feature extraction, data association, and trajectory management to achieve an optimal balance between tracking accuracy and real-time capability [6]. Transformer-based MOT algorithms typically implement this process through an end-to-end paradigm. Their basic framework usually consists of a feature extraction backbone, a transformer-encoder-decoder structure, and a multi-task prediction head. It utilizes a global self-awareness mechanism for the joint processing of recognition and association, thereby significantly simplifying the complex post-processing and heuristic association steps of traditional methods.

2.3 Evaluation Metrics

The Multi-Object Tracking (MOT) research community has established a standardized evaluation system to comprehensively measure algorithms across key dimensions: detection quality, identity consistency, association accuracy, and running efficiency. This review relies on four core metrics for evaluation and analysis: MOTA, IDF1, HOTA, and FPS.

MOTA is a core metric for measuring overall tracking accuracy, considering three main error types: false negatives (FN), false positives (FP), and identity switching (ID). It reflects the cumulative quality of detection and association and is a primary standard for measuring the overall capability of an algorithm.

IDF1 measures an algorithm's ability to maintain identity consistency throughout the entire tracking sequence. It is calculated as the harmonic mean of Identity Precision (IDP) and Identity Recall (IDR), with particular emphasis on maintaining identity

consistency in complex situations such as occlusion, crowded scenes, or overlapping targets.

HOTA breaks down tracking performance into detection accuracy (DetA) and association accuracy (AssA). It uses the geometric mean of these two components to evaluate performance, thus providing a more balanced and interpretable tracking quality metric that covers detection, association, and localization.

FPS is used to evaluate the operational efficiency and real-time performance of a tracker. It represents the number of video frames processed per second, and is therefore particularly important for real-world scenarios with extremely high real-time performance requirements, such as drone monitoring and autonomous driving.

3 Application of Transformer in Long-Term Single-Object Tracking

The challenge of long-term tracking (LTT) lies in how to achieve reliable re-identification and template maintenance after the target undergoes long-term occlusion, deformation, or temporary disappearance. Traditional correlation filters or Siamese networks (SiamRPN, SiamRCNN) rely on fixed template matching, making it difficult to adapt to spatiotemporal changes in the target. In contrast, the Transformer, with its global attention mechanism and spatiotemporal modeling capabilities, offers a new approach to long-term tracking.

3.1 Transformer Feature Extraction and Dynamic Template Update Mechanism

The Transformer's self-attention mechanism can simultaneously capture long-range dependency information between the target and the background, thus achieving higher feature extraction for semantic consistency. The TT-DTU algorithm is a typical Transformer-based long-term tracking framework [7]. This method is based on the STARK framework and introduces a dynamic template update mechanism. It evaluates the template quality through a scoring prediction head and adaptively selects reliable templates, thereby maintaining template stability in deformation and occlusion scenarios.

The scoring prediction head consists of a deep cross-correlation and asymmetric hybrid attention mechanism. It interactively models the appearance and position information of the current frame through learnable scoring labels and outputs a confidence score through a Sigmoid activation function. The system only performs an update when the template confidence exceeds a threshold (τ=0.5), thus effectively avoiding background contamination. Experiments show that this mechanism reduces the risk of template drift while maintaining tracking robustness.

3.2 Performance Verification

TT-DTU achieved superior performance on multiple public datasets: on the LaSOT dataset, its S-score reached 68.1%, and on the VOT2021-LT dataset, its F-score reached 70.3%, representing improvements of approximately 1.0% and 0.8% over STARK, respectively. On the TrackingNet dataset, its NP-score reached 88.0%, and on the

UAV123 dataset, its P-score reached 90.2%, both of which were state-of-the-art results at the time (Table 1). These results validate the robustness and generalization ability of the Transformer-based template update strategy in complex scenarios.

Table 1. Performance Comparison Table of Long-Term Single-Object Tracking Methods

Method	LaSOT (S-score)	VOT2021-LT (F-score)	TrackingNet (NP-score)	UAV123 (P-score)
STARK-ST101	67.1%	69.5%	86.9%	89.3%
LTMU	60.5%	69.1%	–	–
SiamRCNN	65.3%	–	85.4%	83.4%
TT-DTU [7]	**68.1%**	**70.3%**	**88.0%**	**90.2%**

4 Progress in Technologies of Transformer-based Multi-Object Tracking

The core challenge of multi-object tracking (MOT) lies in balancing detection accuracy, object identity consistency, and real-time performance in complex dynamic scenes. The traditional "detection-association" separation paradigm easily leads to error accumulation between modules, resulting in frequent object identity switching and object loss. Transformers, with their global modeling capabilities and flexible structural scalability, offer a new solution for multi-object tracking. In recent years, researchers have made significant progress in applying Transformers to multi-object tracking, mainly focusing on three areas: architectural improvement, feature fusion, and data association optimization.

4.1 Tracker Architecture Improvements

In terms of architecture design, researchers Li Hao and Li Jia proposed a Transformer-based multi-task branch multi-object tracking algorithm (TMTB) for multi-scale target detection and tracking in wide-view drone scenarios [3]. This method enhances the features of targets at different scales through a multi-branch structure: the small target branch uses 1×1 convolution to enhance local details, while the large target branch captures global context information through a window attention mechanism. It also introduces a motion compensation module to fuse camera motion features, thereby improving robustness in dynamic scenarios. On the VisDrone-MOT2019 dataset, the algorithm achieves a MOTA of 48.3%, which is 5.9% higher than JDE, with an 8.2% increase in small-target recall rate and a 75% improvement in running speed, realizing efficient and real-time tracking in complex drone scenarios. As shown in Fig.1, the wide-view tracking framework illustrates the multi-task feature fusion and prediction pipeline of TMTB.

To address the feature interference problem caused by the coupling of detection and tracking tasks in traditional single-decoder structures, researchers Wang Li et al.

proposed a dual-decoder-based Transformer multi-object tracking method [6]. The model relies on the collaborative work of a detection decoder and a tracking decoder with clear division of labor: the detection decoder is responsible for target detection and appearance feature extraction, while the tracking decoder predicts the target's position in the current frame based on the Track Query information from the previous frame. It realizes the matching of results from the two decoders through the Hungarian algorithm. This architecture effectively reduces conflicts between detection and tracking, improving the continuity and stability of tracking. As shown in Figure 2, the dual-decoder structure illustrates the division of labor between the detection and tracking branches.

To further improve computational efficiency, Yang Chen et al. proposed the FLSTrack model based on a dual-decoder structure [8], introducing Focused Linear Attention which reduces the computational complexity of attention from $O(N^2)$ to $O(N)$. This method achieves real-time performance of 30.1 FPS on the MOT17 dataset, with a HOTA index of 66.2%, becoming a representative Transformer architecture that balances accuracy and speed.

Researchers Wu Fang and Zhang Yan proposed the CrossUnihead structure (UniTracker), which achieves unified modeling of detection, classification, and ReID tasks through Transformers [4]. This model improves cross-task information interaction capability via a Multi-Head Feature Sharing Mechanism, significantly reducing the ID Switch rate and verifying the potential of Transformers in end-to-end multi-task collaborative tracking.As shown in Fig.3, the CrossUnihead architecture illustrates the unified modeling of detection, classification, and ReID tasks.

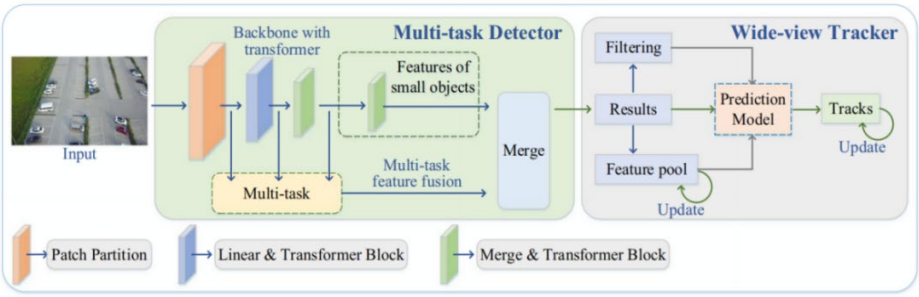

Fig. 1 Framework for tracking models [3]

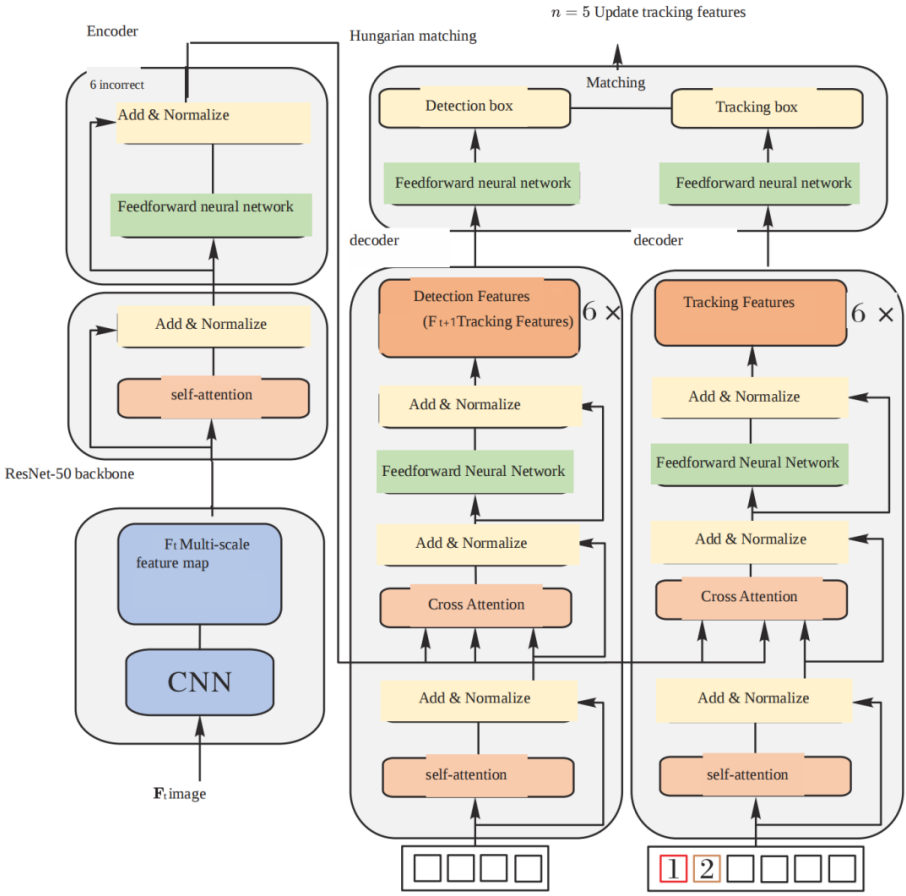

Fig. 2. Dual-decoder structure details [6]

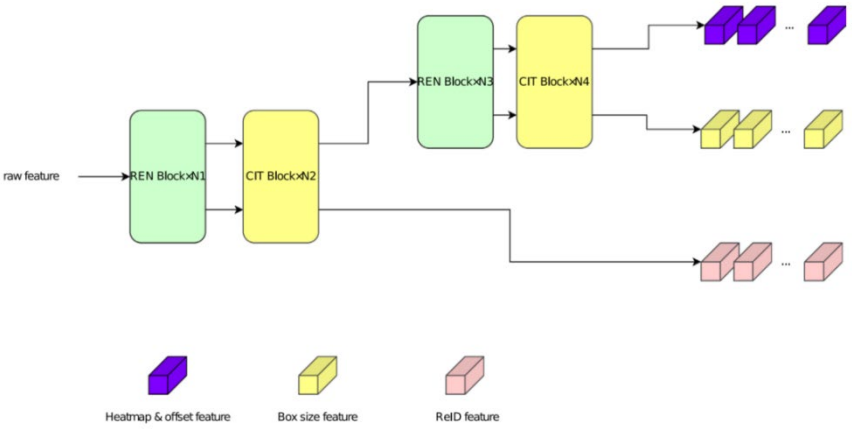

Fig. 3. Overview of transformer-based CrossUnihead [4]

4.2 Feature Fusion and Multi-Scale Modeling

In terms of feature fusion, researchers Zhang Yingjun et al. proposed the CTMOT algorithm with a CNN-Transformer hybrid structure, achieving complementarity between local and global features through parallel CNN and Transformer backbone networks [2]. The CNN branch adopts ShuffleNet V2 as its backbone to extract local edge and texture information of targets; the Transformer branch is based on the Swin Transformer architecture to capture long-range dependencies and global contextual relationships. The two branches realize feature interaction via a Two-way Bridge Module (TBM), organically integrating the global semantic focus of Transformers with the local detail perception of CNNs. Experimental results show that this method achieves a MOTA of 92.36% on the KITTI dataset and 88.57% on the UA-DETRAC dataset, with significantly improved small-target recognition performance.

The CNA-DeepSORT algorithm proposed by researchers Feng Kun et al. combines convolutional features with Channel–Neighborhood Attention [1], enhancing feature representation capability and identity preservation performance in occlusion scenarios. This method achieves higher ID consistency in complex scenarios, verifying the importance of multi-layer feature fusion for multi-object tracking tasks.

In addition, researchers Chen Xi et al. proposed a Scale-aware Transformer, which achieves cross-scale feature fusion by dynamically adjusting Patch size and introducing a scale attention mechanism [9]. The model adopts 8×8 Patches for small targets to preserve detailed features, and incorporates a scale weight factor into self-attention to enhance the saliency of small targets. On the MOT20 dataset, the ID Switch rate decreases by 22.4% and the small-target recall rate increases by 15.7%, significantly improving the performance of Transformers in multi-scale scenarios. As shown in Fig.4, the Scale-aware module illustrates the cross-scale feature fusion mechanism.

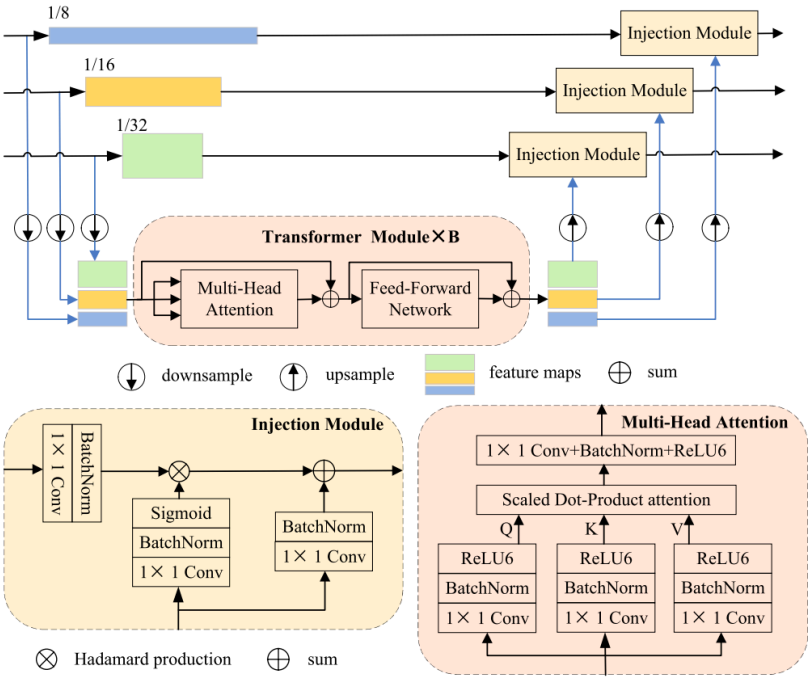

Fig. 4. Scale-aware transformer module [9]

4.3 Data Association and Motion Model Optimization

In terms of data association, Wu Yue et al. proposed an end-to-end multi-target tracking algorithm that combines Transformer and spatial position constraints [10]. This method uses the spatial constraint mechanism of reference points to narrow down the range of candidate targets and combines IoU matching to complete identity association, effectively reducing the complexity of the additional ReID network. Yang Jing et al. proposed a dual-source motion model, which models both target motion information and camera motion vectors simultaneously and captures temporal dependencies through Transformer, thereby significantly improving the association accuracy in dynamic scenes [5]. These studies demonstrate the great potential of Transformer in the data association stage of multi-target tracking (MOT).

5 Key Algorithm Performance Comparison and Analysis

To further verify the effectiveness of the proposed algorithm, we compared it with several representative methods on standard benchmark datasets such as MOT17, focusing on core performance metrics such as MOTA, IDF1, HOTA, IDs, and FPS. As shown in Table 2, the comparison results show that different optimization strategies exhibit different advantages across various performance dimensions.

In terms of overall tracking accuracy, FLSTrack [8], Scale-aware Transformer [9], and CTMOT [2] demonstrated strong competitiveness. FLSTrack [8], with its dual-decoder architecture and focused linear attention mechanism, achieved high scores on both the MOTA (79.8%) and HOTA (66.2%) datasets. Scale-aware Transformer [9] (MOTA 77.5%) and CTMOT [2] (MOTA 76.4%) also showed stable performance, indicating that the multi-scale modeling and feature fusion strategies discussed in Section 4.2 can effectively improve tracking accuracy. In terms of performance, CTMOT [2] and FLSTrack [8] achieved 35.0 FPS and 30.1 FPS respectively, meeting the requirements of real-time applications. CTMOT [2] benefits from its CNN-Transformer hybrid backbone network, while FLSTrack [8] reduces the computational complexity of attention from $O(N^2)$ to $O(N)$ by focusing on a linear attention mechanism, thus achieving high accuracy without sacrificing inference speed. In contrast, most other algorithms run at speeds between 15 and 22 FPS, making them less suitable for applications requiring strict real-time processing.

In terms of identity consistency, CNA-DeepSORT [1] (543 IDs) and TKTR [10] (639 IDs) achieved significantly lower identity switching counts. CNA-DeepSORT [1] enhances the discriminative power of appearance features by combining the attention mechanism with DeepSORT's ReID embedding. In contrast, TKTR [10] achieves similar improvements without relying on the ReID network. It utilizes Transformer-based spatial reference point constraints to mitigate identity fragmentation caused by occlusion or target interaction. FLSTrack [8] (1023 IDs) also exhibits a relatively low number of ID switching, reflecting the association stability brought about by its dual-decoder design.

TMTB [3] and CTMOT [2] demonstrate strong generalization capabilities across various scene types. TMTB [3], designed specifically for wide-angle UAV environments, employs a multi-task branch architecture and achieves 48.3% MOTA on the VisDrone dataset, significantly outperforming baseline methods. CTMOT [2] achieves 92.36% MOTA on the KITTI dataset, demonstrating the adaptability of its hybrid backbone network to various scenes (e.g., vehicle-mounted perspective).

Overall, Transformer-based MOT algorithms show great potential, but their respective technical advantages differ. FLSTrack [8] achieves a balanced trade-off between accuracy and speed; CTMOT [2] performs well in real-time performance and scene generalization; while TKTR [10] and CNA-DeepSORT [1] provide effective strategies for reducing identity switching and improving association stability.

Table 2. Performance Comparison of Representative Transformer-based Multi-Object Tracking Algorithms

Algorithm (Ref.)	**Dataset**	**MOTA (↑)**	**IDF1 (↑)**	**HOTA (↑)**	**IDs (↓)**	**FPS (↑)**
CTMOT [2]	MOT17 /KITTI	76.4% /92.36%	-	-	2317/-	35.0/-

FLSTrack [8]	MOT17	79.8%	78.9%	66.2%	1023	30.1
Scale-aware [9]	MOT17	77.5%	74.9%	-	2235	19.5
UniTracker [4]	MOT17	71.1%	70.0%	-	3573	20.3
TKTR [10]	MOT17	68.2%	63.6%	-	639	-
Dual-Decoder [6]	MOT17	65.4%	64.8%	-	2952	-
Two-Source [5]	MOT17	57.6%	56.1%	-	2636	-
CNA-DeepSORT [1]	MOT16	65.1%	-	-	543	16.0
TMTB [3]	VisDrone	48.3%	-	-	-	21.0

6 Challenges and Future Outlook

Despite significant progress in global modeling and association of Transformer-based multi-object tracking (MOT) algorithms, this technology still faces a series of technical challenges in complex scenes and real-time applications. These challenges also point to key directions for future research.

The current key challenges mainly focus on three aspects. First, there are bottlenecks in computational complexity and real-time performance. The core self-attention mechanism of Transformer has a computational complexity ($O(N^2)$), resulting in the consumption of a large amount of resources in high-resolution videos and long sequence scenes. Although methods like FLSTrack have introduced optimization measures such as linear attention, how to significantly reduce latency and improve inference speed (FPS) while maintaining or exceeding the current level of accuracy remains a key issue limiting the widespread application of this technology in real-time systems (such as autonomous driving and surveillance) [8]. Second, there is the issue of robustness to multi-scale and small targets. The dependence of Transformer on image patch segmentation may lead to the dilution or loss of detailed information of extremely small targets. Although existing methods such as Scale-aware [9] and TMTB [3] have attempted to improve the algorithm through multi-scale modeling or scene-specific optimization, designing a Transformer encoder structure that can adaptively,

efficiently, and accurately represent objects of different scales and densities remains a core challenge for improving the generalization ability of the algorithm. Furthermore, maintaining the target identity under long-term occlusion is also a challenge. Recovering the correct target ID under severe and prolonged occlusion is very difficult, requiring long-term memory of the target's appearance and motion patterns. Current methods, such as TKTR [10] and CNA-DeepSORT [1], have reduced the number of target ID switching, but Transformer needs to model temporal dependencies more deeply to stably maintain the target identity.

The future development directions should mainly focus on the following directions: Developing efficient and lightweight Transformer architectures is crucial. To meet the demands of higher frame rates, it is necessary to explore low-complexity mechanisms such as sparse attention or kernel attention, and to achieve model miniaturization through techniques such as knowledge distillation. In addition, deep fusion modeling of multimodal information is key to enhancing the robustness of the model in harsh environments. This requires combining visual data with heterogeneous data sources such as LiDAR depth information. Future models must be designed with flexible cross-modal attention mechanisms to effectively align and fuse this data. The ultimate goal is to achieve fully end-to-end spatiotemporal joint modeling, establishing a unified Transformer framework capable of simultaneously learning and predicting spatial detection information and temporal trajectory sequences, thereby fundamentally solving the separation problem between detection, association, and motion prediction.

7 Conclusion

This paper provides a comprehensive review of Transformer-based multi-object tracking (MOT) algorithms in recent years. The Transformer architecture, leveraging its self-attention mechanism to provide global modeling and long-range dependency capabilities, introduces a novel end-to-end solution to the MOT field. This paradigm effectively alleviates the limitations of local features and heuristic associations inherent in traditional methods, thereby reducing performance bottlenecks.

This paper delves into three main research directions: tracker architecture improvement, multi-scale feature fusion, and data association optimization. In terms of architecture, dual-decoder and multi-task branching structures have been widely adopted. At the feature level, methods such as CTMOT and Scale-aware Transformer successfully combine the local characteristics of CNNs with the global characteristics of Transformers. Regarding data association, TKTR demonstrates significant stability in preserving target identity by introducing positional constraints. Performance comparisons with representative algorithms such as FLSTrack validate the advantages of lightweight attention mechanisms in balancing tracking accuracy and real-time performance.

Despite the immense potential of Transformer-based multi-object tracking (MOT) algorithms, their high computational complexity and insufficient robustness in extremely small target scenarios remain major obstacles to their widespread application. Future research will focus on developing more efficient Transformer

architectures, leveraging multimodal features for deep fusion, and achieving true end-to-end spatiotemporal joint modeling, thereby breaking through the limits of multi-object tracking technology in terms of accuracy and real-time performance.

References

1. Feng, K., Huo, W., Xu, W., Li, M., Li, T.: CNA-DeepSORT algorithm for multi-target tracking, Multimedia Tools and Applications, vol. 83, pp. 4731–4755 (2024)
2. Zhang. Y., Bai. X., Xie. B.: CNN-Transformer feature fusion multi-object tracking algorithm, Computer Engineering and Applications, vol. 60, no. 2, pp. 180–189 (2024)
3. Li, H., Li, J.: TMTB: Transformer based multi-task branching multi-object tracking algorithm for wide-view scenes, Multimedia Tools and Applications, vol. 83, pp. 41015–41032 (2024)
4. Wu, F., Zhang, Y.: UniTracker: transformer-based CrossUnihead for multi-object tracking, Journal of Real-Time Image Processing, vol. 21, p. 133 (2024)
5. Yang, J., Ge, H., Su, S., Liu, G.: Transformer-based two-source motion model for multi-object tracking, Applied Intelligence, vol. 52, pp. 9967–9979 (2022)
6. Wang, L., Xuan, S., Qin, X., Li, Z.: Multi-object tracking method based on dual-decoder Transformer, Journal of Computer Applications, vol. 43, no. 6, pp. 1919-1929 (2023)
7. Zhang, H., Peng, X., Wang, X.: Long-term tracking with transformer and template update, EURASIP Journal on Advances in Signal Processing, vol. 2022, p. 124 (2022)
8. Zu, D., Duan, X., Kong, G., Long, H.: FLSTrack: focused linear attention swin-transformer network with dual-branch decoder for end-to-end multi-object tracking, Signal, Image and Video Processing, vol. 19, p. 25 (2025)
9. Xiang, X., Zhou, X., Wang, X., Zhai, M., El Saddik, A.: Multi-object tracking with scale-aware transformer and enhanced association strategy, Multimedia Systems, vol. 31, p. 122 (2025)
10. Wu, Y., Luo, J., Zhang, P., Ren, Y.: End-to-end multi-object tracking with location constraint using Transformer, Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), vol. 35, no. 3, pp. 563-570 (2023)