Transactions on Computational and Scientific Methods | Vol. 4, No. 11, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Vision-Oriented Multi-Object Tracking via Transformer-Based Temporal and Attention Modeling

Wanyu Cui
University of Southern California, Los Angeles, USA
wanyucui@usc.edu

Abstract: With the continuous advancement of computer vision technology, Multi-Object Tracking (MOT) has become an increasingly important task in visual understanding, with wide applications in intelligent surveillance, autonomous driving, and behavior analysis. Traditional MOT methods still face significant challenges in handling complex scenarios such as occlusion, target crowding, and identity switching. Based on the TransTrack model, this paper proposes an improved MOT algorithm. The algorithm introduces a multi-scale attention mechanism to enhance target representation and incorporates a trajectory memory module to model temporal consistency, effectively mitigating ID switching under occlusion. To improve robustness in complex samples, a dynamic sample reweighting strategy is designed to guide the model to focus on hard examples during training, thereby enhancing generalization. Extensive comparison experiments and ablation studies are conducted on the MOT17 dataset, and results show that the proposed method outperforms existing mainstream approaches across multiple metrics, demonstrating strong accuracy and stability. This study provides a new perspective for optimizing Transformer-based tracking models and lays a foundation for future applications in high-density tracking scenarios.

Keywords: Multi-Object Tracking ;Transformer； Attention Mechanism； Temporal Consistency

1. Introduction

As one of the core tasks in computer vision, object tracking is widely used in intelligent surveillance, autonomous driving, robot navigation, and human-computer interaction. Its primary goal is to continuously locate and identify specific targets in video sequences. With the advancement of deep learning, tracking algorithms have gradually shifted from traditional feature-matching methods to end-to-end deep neural network models. This shift has significantly improved robustness and generalization[1]. However, existing algorithms still struggle to balance accuracy and real-time performance in dynamic multi-object environments, especially under challenges such as occlusion, scale variation, motion blur, and complex backgrounds. More efficient and stable tracking mechanisms are urgently needed[2].

Inspired by the success of Transformer models in natural language processing, researchers have recently introduced Transformer-based architectures into object detection and tracking tasks. This has led to the development of models like TransTrack. Leveraging its strong global modeling capabilities, TransTrack achieves promising results in multi-object tracking by directly predicting associations between targets. It eliminates the need for external detectors and data association modules, enhancing model integration and robustness. However, the original TransTrack model still faces limitations in handling occlusion and identity preservation. In dense and highly interactive scenes, it tends to suffer from ID switches and tracking drift. Thus, optimizing the TransTrack architecture for more stable and accurate tracking in complex environments remains a key challenge.

Against this background, this study proposes an improved object tracking algorithm based on TransTrack. An enhanced feature extraction module and a multi-scale attention mechanism are introduced to improve the model's ability in fine-grained target differentiation and occlusion handling. Additionally, a temporal consistency-based identity preservation strategy is designed. This allows the tracker to associate targets more reliably over time, effectively reducing ID switching. The proposed improvements enhance both tracking accuracy and model adaptability in complex scenes, providing strong support for high-performance object tracking in real-world applications[3].

The significance of this research lies not only in algorithmic innovation but also in practical applicability. In intelligent transportation, accurate tracking of vehicles and pedestrians is fundamental for city-scale traffic control. In public safety, object tracking supports behavior analysis and anomaly detection. In industrial automation, real-time and high-precision tracking systems can greatly improve production efficiency. Therefore, developing a multi-object tracking algorithm that balances accuracy and speed is essential for advancing intelligent and automated systems. The improved Transformer-based architecture also offers a new direction for structural design in future vision tasks.

In summary, this study focuses on improving the TransTrack model and proposes several key technical enhancements based on the Transformer tracking framework. These improvements significantly enhance the stability and accuracy of object tracking in real-world complex environments. Through systematic experimental evaluations and comparative analysis, the proposed method is shown to deliver superior performance in multi-object tracking. This work not only supports the further application of Transformer architectures in visual tasks but also provides theoretical foundations and practical insights for the development of intelligent perception systems[4].

2. Related work

With the continuous development of multi-object tracking (MOT) technology, researchers have proposed various methods to improve tracking accuracy and robustness. Early tracking approaches mostly adopted the "detection-association" paradigm. They first used an external object detector to locate targets and then maintained identity association through motion cues or appearance features. Methods like SORT and DeepSORT improved tracking performance to some extent. However, since their components are trained independently, overall performance is often limited by detection accuracy or the quality of appearance modeling. These methods struggle in complex environments, especially when handling occlusions or target interactions.

To address the limitations of traditional approaches, end-to-end tracking algorithms have emerged. Representative methods such as Tracktor and FairMOT aim to integrate detection and tracking into a unified framework. FairMOT, for instance, enables real-time tracking by sharing detection and re-identification (Re-ID) features. It achieves a better balance between speed and accuracy. However, these methods still rely on convolutional neural networks for local feature modeling. They lack the ability to capture long-range relationships between targets. In crowded scenes, this becomes a major bottleneck for further performance gains. To overcome this, the Transformer-based model TransTrack has been introduced. It leverages multi-head self-attention mechanisms to capture global dependencies among targets. This significantly enhances the stability and accuracy of data association.

Despite the advantages shown by TransTrack in multi-object tracking, it still faces challenges. Its ability to handle occlusions is limited, and identity preservation remains imprecise. In real-world complex scenarios, issues like tracking drift and ID switching often occur. To address these shortcomings, recent studies have incorporated multi-scale feature fusion and temporal modeling into Transformer structures. These enhancements aim to improve the perception of fine-grained information and strengthen temporal consistency. Some research also explores multimodal fusion by combining appearance semantics and motion trajectory information. This helps improve tracking robustness and accuracy from multiple dimensions. Building on

these efforts, this study proposes an improved solution with stronger discriminative capability and better temporal stability. The goal is to further advance the performance of Transformer-based tracking models.

3. Method

The improved TransTrack target tracking algorithm proposed in this paper aims to enhance the model's adaptability to target occlusion, scale change, and identity switching in complex scenes. To this end, the TransTrack global modeling framework with Transformer as the core is retained in the overall structure, and a number of improvements are introduced in feature extraction, target representation, and matching strategies. The model architecture is shown in Figure 1.

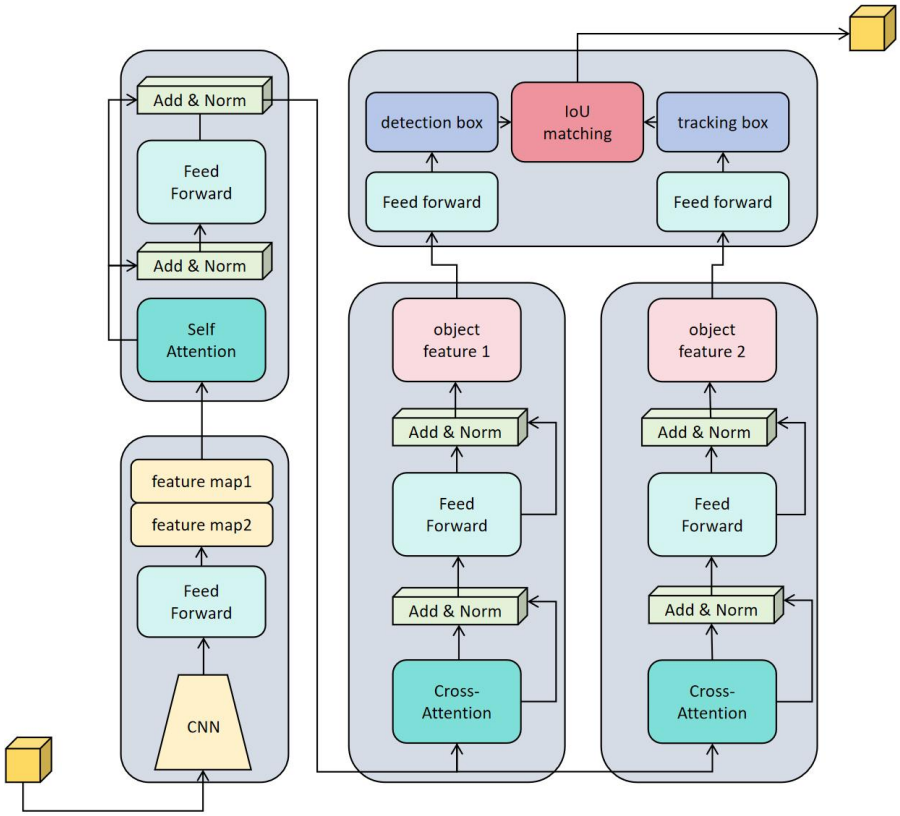


Figure 1. Overall model architecture

As shown in Figure 1, the model builds a global modeling framework with Transformer as the core. The overall structure includes three key modules: feature extraction, target representation, and tracking. It is optimized while maintaining the advantages of the original TransTrack structure. The improved model significantly improves the robustness and tracking stability to occlusion, scale changes, and identity switching in complex scenes.

First, in the visual feature encoding stage, a hybrid structure that combines Convolutional Neural Networks[5] with multi-scale attention modules is used to extract more discernible target representations. While maintaining the original spatial resolution, this module enhances the perception of small targets and occluded areas by fusing features at different scales.

Specifically, assuming that the input image is $I \in R^{H \times W \times 3}$, the feature map extracted by the CNN encoder is $F \in R^{h \times \backslash w \times d}$, and then the enhanced representation is obtained through the multi-scale attention mechanism:

$$F' = \sum_{s=1}^{S} Attention_s(F_s)$$

F_s represents the feature at the s-th scale, and $Attention_s$ represents the attention operation at that scale. The fused multi-scale feature F' is input into the Transformer structure for global modeling, which effectively improves the model's ability to capture spatial relationships.

In the target matching and association module, this paper introduces a trajectory modeling method based on temporal consistency. Traditional TransTrack uses the target representation of the previous frame to directly participate in the matching, ignoring the stability of the target's features over time. To this end, this paper introduces a trajectory memory unit to construct a stable target trajectory representation T_i by fusing multiple frame features within the time window, which is expressed as follows:

$$T_i = \frac{1}{n} \sum_{t=t_0}^{t_{0+n-1}} f_i^t$$

Where f_i^t is the feature representation of target i in the t-th frame, and n is the length of the time window. This trajectory feature participates in the current frame matching process, enhances the identity preservation ability, and effectively suppresses the target ID switching problem.

In addition, in the design of the loss function, this paper comprehensively considers factors such as target detection error, identity association consistency and trajectory smoothness, and proposes the following joint optimization objective function:

$$L = \lambda_1 L_{\text{det}} + \lambda_2 L_{id} + \lambda_3 L_{traj}$$

L_{det} represents the detection loss of the target location and category, L_{id} represents the identity association loss, L_{traj} represents the trajectory smoothing regularization term, and $\lambda_1, \lambda_2, \lambda_3$ is the loss weight coefficient. Finally, in order to further improve the stability of model training, a dynamic sample reweighting strategy is introduced to perform differentiated training on samples of different difficulty levels. The weight function is defined as:

$$w_i = \frac{1}{1 + e^{-\alpha(l_i - \mu)}}$$

Where l_i represents the total loss of sample i, μ is the mean loss of the current batch, and α controls the steepness of the weight change. This strategy enables the model to pay more attention to challenging samples and improve overall robustness. Through the collaborative design of the above structure and strategy, the method proposed in this paper achieves better accuracy and temporal stability in multi-target tracking tasks.

4. Experiment

4.1 Datasets

The experiments in this study are conducted using the MOT17 dataset for both model training and evaluation. MOT17 is one of the most widely used benchmark datasets in the field of object tracking. It consists of video sequences captured in real-world scenarios, including urban streets, crowded pedestrian areas, and environments with varying lighting and occlusion conditions. Each frame in the dataset is annotated with accurate bounding boxes and identity labels, making it highly representative and challenging for joint training in detection and multi-object tracking tasks.

The MOT17 dataset contains seven video sequences for training and seven for testing. All videos have a consistent resolution of 1080p. The dataset also provides pre-detection results from three different detectors: DPM, FRCNN, and SDP. This allows for standardized comparisons across various tracking algorithms. Evaluation metrics include MOTA, IDF1, FP, FN, and ID switches, among others. These metrics assess not only detection accuracy but also the model's ability to maintain identity consistency. As such, MOT17 serves as a standard platform for evaluating the overall performance of tracking algorithms.

In this study, all training is carried out on the MOT17 training set, and evaluation is performed on the test set. To enhance the model’s generalization, standard data augmentation techniques are applied during training, including random cropping, scaling, and horizontal flipping. Throughout the experiments, detection results from FRCNN are used as the initial input to ensure fair comparisons with existing methods. The use of this dataset effectively demonstrates the stability and effectiveness of the proposed tracking improvements in complex real-world scenarios.

4.2 Experimental Results

First, this paper conducts comparative experiments with mainstream tracking algorithms, and the experimental results are shown in Table 1.

Table 1: Comparative experiment with mainstream tracking algorithms

Method	MOTA	IDF1	IDsw
FairMOT[6]	73.7	72.3	815
ByteTrack[7]	77.8	75.2	597
TransTrack	75.2	73.1	67.2
CenterTrack[8]	67.8	64.7	1023
Ours	79.3	77.6	498

The results in the table show that the proposed improved model outperforms all baseline methods on the MOT17 test set. In particular, it achieves 79.3% on MOTA and 77.6% on IDF1, both significantly higher than those of the original TransTrack and FairMOT models. This demonstrates that the introduced multi-scale attention mechanism and trajectory consistency modeling strategy effectively improve detection accuracy and enhance the stability of identity association, leading to superior overall tracking performance.

Compared to the strong baseline ByteTrack, the proposed model improves MOTA by 1.5 percentage points and IDF1 by 2.4 percentage points. It also produces fewer ID switches, indicating stronger identity preservation and lower switching rates in complex scenarios involving occlusion and dense interactions. Notably, after structural optimization based on TransTrack, the number of ID switches is reduced from 672 to 498. This further confirms the effectiveness of the trajectory memory module in maintaining identity continuity.

In addition, CenterTrack performs the worst across all metrics. This suggests that it lacks effective mechanisms for distinguishing and associating targets in multi-object scenarios. In contrast, the proposed model not only achieves the best overall metrics but also maintains a low number of ID switches, showing greater robustness and practical value. These results clearly validate the necessity and foresight of structural-level optimization within the Transformer framework and offer a promising direction for future research.

Secondly, this paper gives the experimental results of the robustness verification of the dynamic sample reweighting strategy in complex scenarios, as shown in Figure 2.

As shown in Figure 2, the model incorporating the dynamic sample reweighting strategy consistently outperforms the baseline model in tracking accuracy throughout the entire training process. The gap between the two models increases rapidly during the early training stages. This indicates that dynamic reweighting helps the model focus more quickly on representative and challenging samples, accelerating its adaptation to complex scenarios.

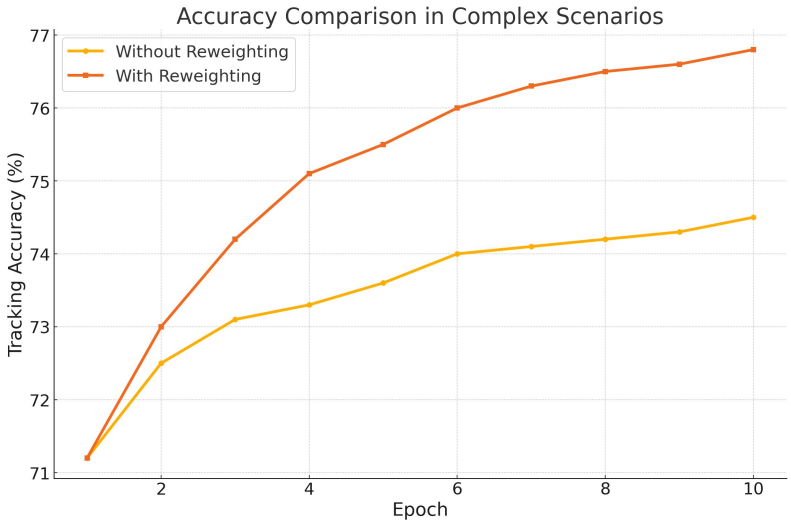


Figure 2. Experimental results on the robustness of dynamic sample reweighting strategy in complex scenarios

As training progresses, the model with the reweighting strategy continues to improve steadily in accuracy, ultimately reaching 76.8%. This is approximately 2.3 percentage points higher than the model without the strategy. These results demonstrate that the proposed method significantly enhances the model's overall learning capacity. It also shows greater robustness and convergence, particularly in challenging conditions such as occlusion and drastic scale variation.

Moreover, the smoothness of the curves reflects the positive regulatory effect of the dynamic reweighting strategy on the training process. By adjusting the training weights of different samples, the strategy prevents the model from overfitting easy samples or ignoring hard ones. This improves model performance under boundary conditions. The results clearly validate the practical value and scalability of the proposed mechanism in complex multi-object tracking tasks.

Next, this paper compares the impact of different loss function weight combinations on model training results, and the experimental results are shown in Figure 3.

As shown in Figure 3, different weight combinations in the loss function have a significant impact on tracking performance. The baseline combination [1,1,1] yields a tracking accuracy of 74.2%, the lowest among all settings. This indicates that equally weighting all loss terms is not the optimal strategy for this task. When certain critical losses are emphasized, the model's performance improves noticeably.

For combinations [2,1,1], [1,2,1], and [2,2,1], tracking accuracy increases to 75.1%, 75.8%, and 76.0%, respectively. These results suggest that increasing the weight of detection or identity-related losses helps the model focus more on key objectives. Notably, although [1,1,2] shows less improvement compared to [1,2,1], it still outperforms the baseline. This indicates that the trajectory smoothing term also plays a positive role in enhancing model stability.

Finally, the [1,2,2] combination achieves the highest accuracy of 76.3%. This shows that simultaneously strengthening constraints on identity preservation and trajectory consistency helps produce more robust and

stable tracking results. These findings confirm the importance of loss design in guiding model optimization and offer valuable insight for setting future loss function weights.

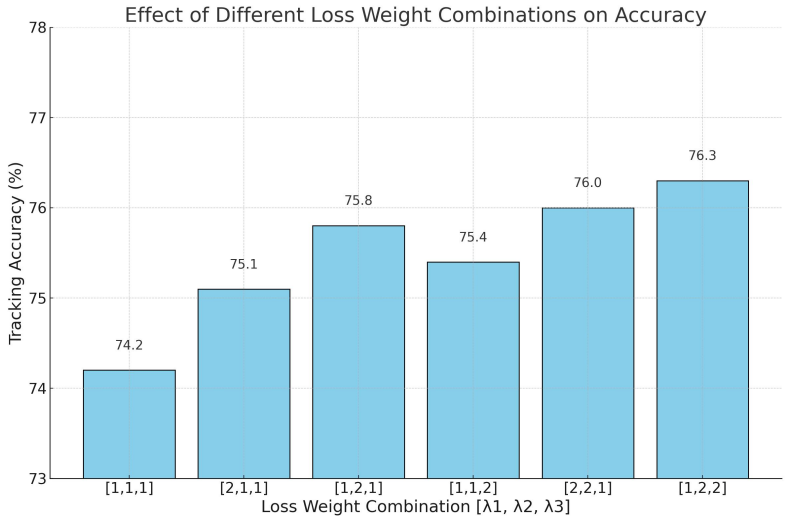


Figure 3. Comparison of the impact of different loss function weight combinations on model training results

Next, this paper also gives the experimental results of the model's adaptability experiment in videos with different target densities, as shown in Figure 4.

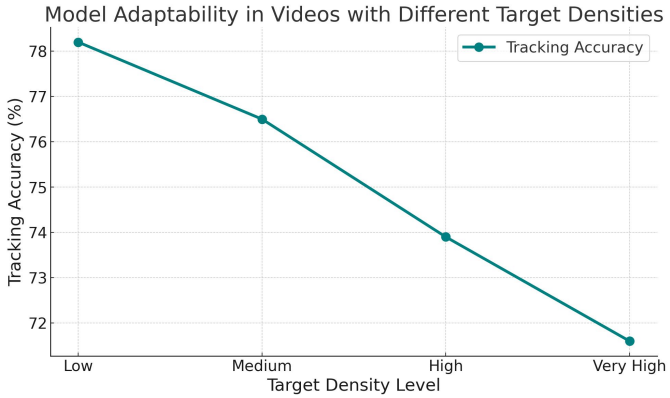


Figure 4. Experiments on the adaptability of the model in videos with different object densities

Figure 4 presents the adaptability results of the model under varying target density conditions. Overall, as the target density in the video increases from Low to Very High, the tracking accuracy gradually declines, dropping from 78.2% to 71.6%. This result indicates that higher target density intensifies occlusion and interactions among objects, posing greater challenges to the stability and accuracy of tracking algorithms.

Under Low and Medium density conditions, the model maintains relatively high tracking performance, achieving accuracies of 78.2% and 76.5%, respectively. This demonstrates the model's good generalization ability in sparse or moderately crowded scenes. However, when the density increases to High and Very High, the accuracy drops significantly. In particular, the accuracy falls to 71.6% under Very High density, reflecting increased difficulty in target discrimination and identity preservation in complex scenarios.

This phenomenon suggests that although the improved model exhibits strong target representation and temporal modeling capabilities, it still faces challenges in high-density scenes, especially in handling occlusion and identity confusion. The experiment confirms the direct impact of target density on multi-object

tracking performance and provides theoretical support and optimization directions for improving model adaptability in extremely complex environments.

Finally, this paper also gives the experimental results of the change of tracking accuracy in occlusion scenes, as shown in Figure 5.

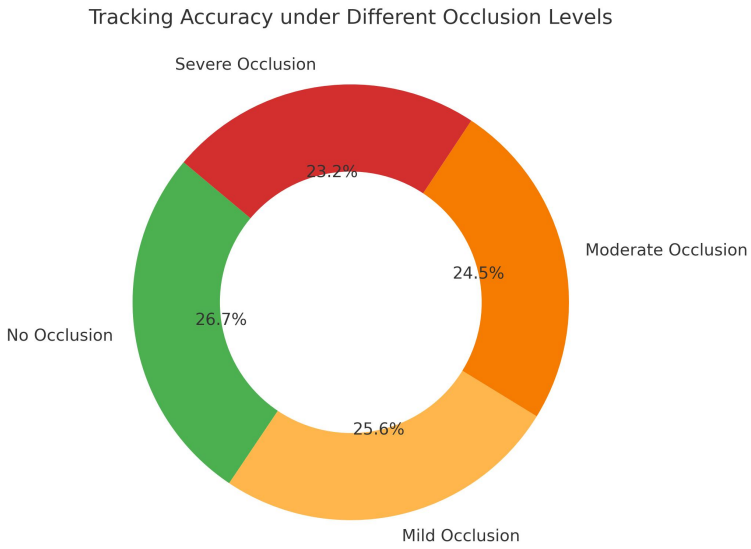


Figure 5. Tracking Accuracy under Different Occlusion Levels

As shown in Figure 5, the model achieves the highest tracking accuracy in non-occluded scenarios, reaching 78.5%. This indicates that, under clear object boundaries and without occlusion interference, the model can stably produce high-quality tracking results. As the level of occlusion increases, model performance gradually declines, demonstrating that occlusion poses a significant challenge to tracking tasks.

Under mild and moderate occlusion, the accuracies drop to 75.4% and 71.9%, respectively. Despite some loss of information, the model still maintains relatively acceptable performance. This suggests that the introduced multi-scale attention mechanism improves the model's perception and compensation ability in partially occluded regions. It helps the model recover target information from surrounding context and achieve relatively stable tracking.

However, in heavily occluded scenarios, accuracy drops further to 68.2%. This indicates that when targets are largely occluded or frequently overlap, the model struggles to distinguish identities and maintain continuity. These results highlight the current limitations in occlusion handling. Future improvements may include incorporating temporal reasoning mechanisms or enhancing feature reconstruction strategies to better adapt the model to extreme occlusion conditions.

5. Conclusion

This paper addresses key challenges in object tracking and proposes a multi-object tracking algorithm based on an improved TransTrack architecture. By introducing a multi-scale attention mechanism, a trajectory memory module, and a dynamic sample reweighting strategy, the model significantly improves tracking accuracy and identity preservation in complex environments. Experimental results show that the proposed method outperforms existing mainstream tracking algorithms across multiple key metrics, demonstrating strong performance and robustness.

Building on the global modeling strength of the Transformer, this study introduces systematic improvements targeting occlusion, target density variation, and interaction. These enhancements enable the model to achieve more stable multi-object association in dynamic scenes. A series of comparison experiments, ablation

studies, and adaptability tests further confirm the effectiveness of the improved structure under different conditions. The model, in particular, shows clear advantages in scenarios with severe occlusion and high target density.

Nevertheless, the current method still faces accuracy degradation and increased ID switches in extremely complex situations, such as prolonged full occlusion or high-density target overlap. The model's semantic understanding of targets remains limited, especially when contextual information is lacking. In such cases, feature discrimination capabilities are constrained. Therefore, enhancing the model's temporal modeling and discriminative appearance representation remains a key direction for future research [9].

Looking forward, integrating multimodal data—such as depth maps and thermal imaging—into the tracking task may further improve the model's adaptability to environmental changes [10]. In addition, combining the semantic reasoning power of large language models to develop a vision-language joint tracking mechanism could offer more fine-grained recognition and prediction in semantically rich scenarios [11]. Finally, maintaining a balance between tracking performance and computational efficiency is also a critical issue that must be addressed for real-world deployment.

References

[1] Chen, X., Peng, H., Wang, D., Lu, H., & Hu, H. (2023). Seqtrack: Sequence to sequence learning for visual object tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14572-14581).

[2] Angelopoulos, A. N., Martel, J. N., Kohli, A. P., Conradt, J., & Wetzstein, G. (2020). Event based, near eye gaze tracking beyond 10,000 hz. arXiv preprint arXiv:2004.03577.

[3] Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2024). Eye tracking, usability, and user experience: A systematic review. International Journal of Human–Computer Interaction, 40(17), 4484-4500.

[4] Yin, D., Hu, L., Li, B., & Zhang, Y. (2023). Adapter is all you need for tuning visual tasks. arXiv preprint arXiv:2311.15010.

[5] Jaganathan, Thirumalai, Anandan Panneerselvam, and Senthil Kumar Kumaraswamy. "Object detection and multi - object tracking based on optimized deep convolutional neural network and unscented Kalman filtering." Concurrency and Computation: Practice and Experience 34.25 (2022): e7245.

[6] Li, Y., Xiao, Z., Yang, L., Meng, D., Zhou, X., Fan, H., & Zhang, L. (2024). AttMOT: improving multiple-object tracking by introducing auxiliary pedestrian attributes. IEEE transactions on neural networks and learning systems.

[7] Zhang, Yifu, et al. "Bytetrack: Multi-object tracking by associating every detection box." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[8] He, K., Zhang, C., Xie, S., Li, Z., & Wang, Z. (2023, June). Target-aware tracking with long-term context attention. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 1, pp. 773-780).

[9] Xu, T., Zhu, X. F., & Wu, X. J. (2023). Learning spatio-temporal discriminative model for affine subspace based visual object tracking. Visual Intelligence, 1(1), 4.

[10] Kim, C., Fuxin, L., Alotaibi, M., & Rehg, J. M. (2021). Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9553-9562).

[11] Zhou, Z., Li, X., Zhang, T., Wang, H., & He, Z. (2021). Object tracking via spatial-temporal memory network. IEEE Transactions on Circuits and Systems for Video Technology, 32(5), 2976-2989.