Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO25729

Optimization and Application of Multi-Target Tracking Algorithm for Indoor Service Robots Based on Enhanced TransTrack

Shuangli Song

Faculty of Engineering, The University of New South Wales, Sydney, Australia
songshuangli123@163.com

Abstract. With the popularity of indoor service robots in public and private environments, the demand for accurate and real-time multi-object tracking (MOT) systems has become increasingly urgent. This study explores the development and optimization of an MOT algorithm for indoor service robots based on a modified TransTrack model. To address challenges such as occlusion, low light, and real-time performance, this study makes several improvements to the original TransTrack architecture, including a lightweight backbone network, an occlusion handling module, and an attention mechanism pruning strategy, which improves accuracy and computational efficiency to a certain extent. The system is trained and evaluated on the Indoor-MOT dataset, and some preprocessing techniques are applied to make it adaptable to some complex indoor environments. Based on standard MOT metrics such as MOTA, IDF1, FPS, and HOTA, the improved system outperforms DeepSORT, FairMOT, and the original TransTrack in terms of accuracy and speed. In spite of its limitations in tracking fast-moving targets, the system still has certain application prospects in the field of indoor robotics.

Keywords: Multiple-Object Tracking(MOT), TransTrack, Indoor Service, Indoor Environment.

1. Introduction

As the benefits of intelligent systems are increasingly recognized by society, some indoor service robots are gradually appearing in environments such as hotels, restaurants, and some public facilities. Multi-object tracking (MOT) is one of the core functions for such robots to achieve autonomous operation. It refers to the process of detecting and tracking multiple objects across video frames [1]. However, some traditional MOT frameworks usually separate detection and data association, which may lead to identity switching and failure in challenging indoor conditions such as occlusion or low light.

To address these limitations, the Transformer-based model TransTrack simplifies the traditional MOT process by completing object detection and object association in a single shot and reduces potential errors to a certain extent [2]. However, there is still a lack of systematic evaluation and empirical verification of the performance of TransTrack in complex indoor environments.

This study aims to optimize the original TransTrack model through multiple architectural enhancements, including a lightweight backbone network, an occlusion handling module, and attention mechanism pruning. These improvements are intended to enhance tracking robustness, computational efficiency, and real-time performance.

The results of this study are expected to provide certain practical value for the deployment of indoor service robots and provide valuable reference for future research on Transformer-based MOT systems.

2. Multiple-Object Tracking(MOT) and TransTrack

2.1. The foundation of traditional MOT and the challenges of current indoor environments

Multi-object tracking (MOT) refers to detecting multiple objects in a continuous video sequence and assigning corresponding individual numbers to them, thereby forming the trajectory of the object in the entire sequence to maintain the consistency of its identity over the entire time axis, as shown in Figure 1. A typical MOT system consists of three main parts: object detection, feature extraction, and data association.

Among these three modules, based on traditional MOT, object detection is usually performed using a deep convolutional model (such as YOLOv3) [3]. At the same time, feature extraction is performed using a re-identification (ReID) network and data association is performed using the Hungarian algorithm combined with a Kalman filter to maintain temporal consistency [4].

However, there are several challenges in tracking multiple objects using these three modules in complex environments:

a. Identity switching caused by occlusion interrupts the trajectory formation process [1].

b. Feature representation is unstable under different lighting conditions and crowd density [5].

c. The high computational overhead of modular design limits the real-time performance of embedded systems [6].

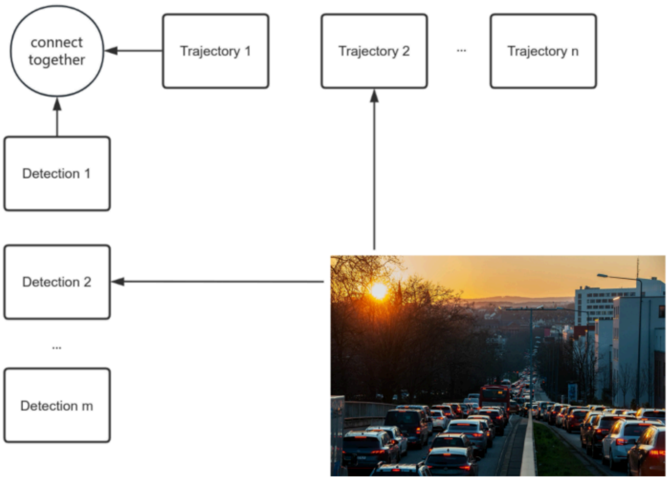

Figure 1. A simplified overview of trajectory formation in multi-objective tracking

2.2. Overview of TransTrack and its unique advantages

TransTrack is an end-to-end multi-object tracking (MOT) module based on the Transformer architecture that merges object detection and data association into a unified process. It transforms the tracking task into a sequence prediction problem, thus simplifying the entire workflow.

The core of TransTrack is its Transformer encoder-decoder structure, where object tracking queries are used to propagate identity information from previous frames. These queries guide data detection by directly incorporating temporal cues into the detection process to predict the location and identity of the object in the current frame, avoiding the need for a separate re-ID network and filtering mechanism [7].

TransTrack has a number of advantages. First, it merges multiple steps into a single process, thereby improving efficiency. Second, it simplifies the architecture and does not require auxiliary modules. In addition, it achieves excellent tracking accuracy through end-to-end optimization. This module is clearly shown in Figure 2.

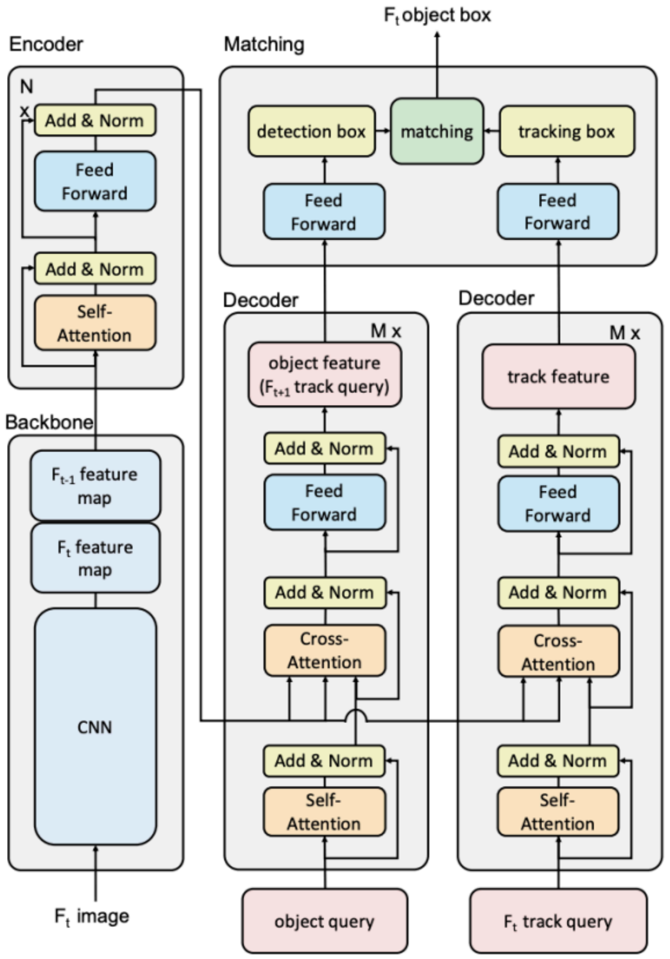

Figure 2. The framework details of TransTrack [2]

2.3. Application of TransTrack in indoor service robots

In the field of indoor service robots, TransTrack has also demonstrated its excellent adaptability. In indoor environments, challenges such as frequent occlusion, changing lighting conditions, and narrow navigation space deserve attention. TransTrack's simplified process and precise positioning capabilities can cope with these situations well. Its Transformer-based attention mechanism enables the model to maintain identity consistency even when objects are temporarily hidden or partially visible, thus better coping with some crowded scenes, such as restaurant services and indoor food delivery services. The process of applying the TransTrack model to indoor service robots is shown in Figure 3.

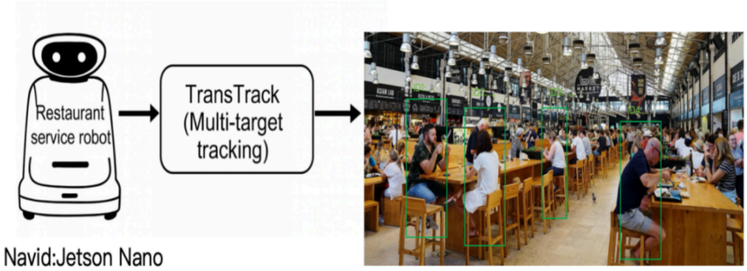

Figure 3. TransTrack integrated into the vision pipeline of an indoor restaurant service robot

The robot captures real-time visual data and TransTrack processes the stream to provide accurate and continuous multi-object tracking results with identity assignment.

3. Optimization and design of improved TransTrack

3.1. Some modifications of TransTrack

To improve the performance of TransTrack in indoor service robots, several structural modifications can be taken to TransTrack to address key challenges such as model complexity, occlusion handling, and real-time performance. These improvements focus on backbone network simplification, temporal robustness, and attention mechanism efficiency, resulting in a cleaner and more accurate version compared to the original architecture.

The original TransTrack uses a ResNet-based backbone network to extract features, which can be burdensome for real-time applications on embedded platforms [8]. To address this issue, the backbone network is replaced with MobileNetV2, a lightweight convolutional neural network designed for efficient inference [9]. In addition, the traditional feature pyramid is also replaced with a bidirectional feature pyramid network (BiFPN) to enhance multi-scale feature fusion [10-11]. With these designs, the new TransTrack model is able to capture finer-grained object details at different scales while significantly reducing the amount of computation, as shown in Figure 4.

In indoor environments, there are often some dynamic obstacles and frequent occlusions. To improve the recognition accuracy and robustness of the service robot under these conditions, we embed an occlusion-aware prediction module in the decoder stage, which can reduce the possibility of identity switching when the target is temporarily hidden [12].

Finally, to improve the computational efficiency of the original TransTrack, we can apply attention pruning in the self-attention layer of the decoder. By masking redundant key-query interactions, the computation is reduced while maintaining accuracy. The improved attention operation is expressed as formula [13]:

$$\mathrm{Attention}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right) = \mathrm{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{m}}} + \mathrm{M}\right)\mathrm{V} \tag{1}$$

where M is a learnable binary mask to suppress irrelevant tag interactions, and m is equal to dk, the dimension of the key vector.

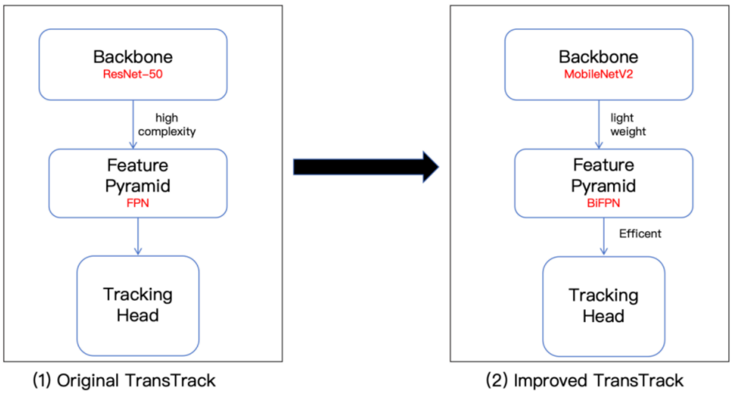

Figure 4. Backbone and feature fusion modifications in TransTrack, (1) The original structure uses ResNet-50 and traditional FPN, resulting in a large amount of computation.complexity, (2) the improved version integrates a lightweight MobileNetV2 backbone with BiFPN, significantly reducing inference cost while enhancing multi-scale feature fusion

3.2. Dataset preparation and training configuration

For indoor service robots, choosing a suitable dataset is crucial to evaluate the performance of the model in a real environment. Since there is currently no database dedicated to indoor multi-object tracking, it is of great significance to create a new indoor multi-object tracking (MOT) database using MySQL Workbench [14]. Compared with traditional multi-object tracking databases (such as MOT17 and MOT20, which mainly focus on targets with limited relative motion), the new indoor MOT database is suitable for multi-object tracking in various indoor scenes. It also has more advantages in dealing with common indoor problems such as occlusion and low illumination.

This indoor-MOT database includes some detailed scene descriptions, lighting conditions, object categories, and frame-by-frame bounding boxes with occlusion-level annotations. It can be simply shown as:

```
//Insert indoor scene

INSERT INTO Scenes (scene_name, location_description, lighting_condition, recorded_by, total_frames)

VALUES ('Office Corridor', 'Narrow hallway with fluorescent light', 'low light', 'GoPro Hero 9', 500);

//Insert object classes

INSERT INTO Classes (class_name) VALUES ('person'), ('service_robot');

//Insert object identities

INSERT INTO Objects (class_id, unique_track_id) VALUES (1, 101), (2, 102);

//Insert first frame

INSERT INTO Frames (scene_id, frame_number) VALUES (1, 1);
```

```
//Insert bounding boxes with occlusion annotations

INSERT INTO BoundingBoxes (frame_id, object_id, x, y, width, height, occlusion_level, visibility)

VALUES (1, 1, 120, 200, 40, 100, 0.3, TRUE),

(1, 2, 300, 220, 60, 110, 0.0, TRUE);
```

To ensure that the newly created indoor MOT dataset is compatible with the improved TransTrack framework, some preprocessing steps are required. First, all images should be resized to a fixed resolution of 640 * 480 to reduce computational overhead. Second, the pixel values should be normalized using ImageNet statistics to match the new backbone network MobileNetV2 [15]. Finally, the labeled data is converted to the COCO style format, which is commonly used in the object detection and tracking process in PyTorch [16]. This process is shown in Figure 5.

After completing the dataset preprocessing, the improved TransTrack model can be trained using the PyTorch framework. The architecture integrates MobileNetV2, BiFPN, and an occlusion-aware decoder. The training was performed for a total of 80 epochs with a batch size of 8. In addition, the AdamW optimizer was used with an initial learning rate of 1* 10^(-4) and a decay of 0.1 times every 20 epochs. The loss function formula is as follows [2]:

$$\mathrm{L(total)} = \mathrm{L(matc}h\mathrm{)} + \lambda \times \mathrm{L(Glou)} \tag{2}$$

Where L(match) denotes the Hungarian matching loss for identity assignment and L(Glou) denotes the generalized IoU loss for bounding box regression.

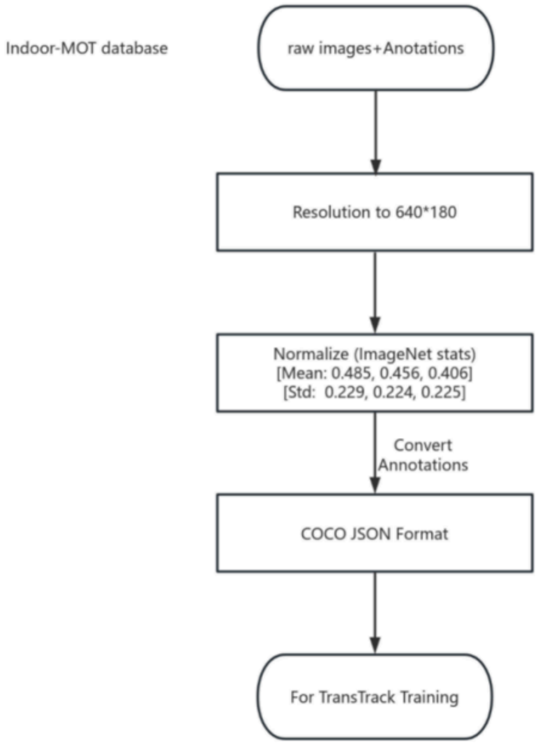

Figure 5. Some adjustments regarding the adaptability of the Indoor-MOT database

4. Experimental results and discussion

4.1. Evaluation metrics and settings

To evaluate the performance of the improved TransTrack model in indoor service scenarios, we will adopt four widely used MOT (Multiple Object Tracking) metrics [17]:

a. Multi-target tracking accuracy (MOTA) is a comprehensive indicator that takes into account factors such as false positives, false negatives, and identity switching, and is used to evaluate the overall accuracy of the entire tracking system [18].

b. The IDF1 score is used to measure the consistency of target identity assignment across frames. A higher score means fewer identity switches and stronger target tracking stability [19].

c. Frames per second (FPS) is used to measure the real-time processing efficiency of the model. This indicator is of great significance for determining whether indoor service robots can meet the runtime efficiency requirements [20].

d. High-order tracking accuracy (HOTA) is a comprehensive indicator that takes into account the accuracy of detection, target association, and spatial positioning. It is highly representative in evaluating the overall performance of multi-target tracking systems [21].

Based on the above common metrics, we can determine whether the improved TransTrack model meets the standards.

4.2. Comparative evaluation of the performance of the improved TransTrack model

To verify the effectiveness of the improved TransTrack model, we quantitatively compare three representative MOT methods: original TransTrack, FairMOT, and DeepSORT [22].

The performance of each model is evaluated based on the four metrics introduced previously: MOTA, IDF1, FPS, and HOTA. The experimental results are summarized in Figure 6.

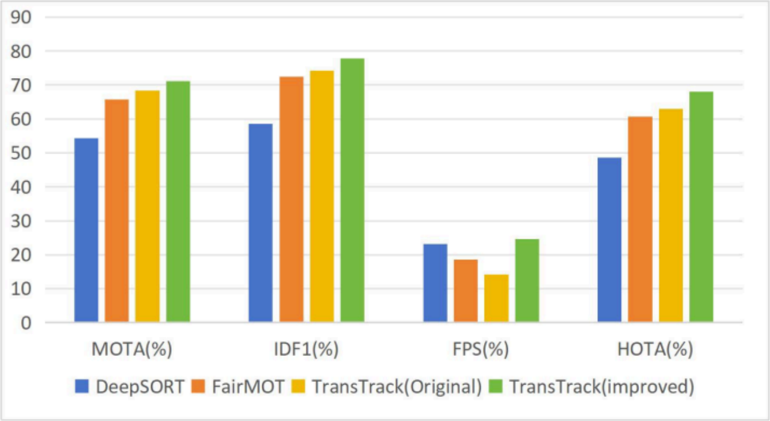

Figure 6. Performance comparison of tracking models of on Indoor-MOT database

4.3. Indoor service robot deployment

The improved TransTrack model will show stable performance in low-light, cluttered, and heavily occluded indoor environments. In fact, different indoor service robots have corresponding improvements in different scenarios. For example:

Restaurant delivery robots can ensure continuous tracking of human targets at tables and in hallways, even in the presence of some dynamic obstacles.

Hospital guide robot: It can continuously help patients with poor vision avoid obstacles.

Hotel food delivery robot: It can accurately locate the location of guests and deliver items safely even in the presence of a large number of obstacles and dim light.

These features highlight the practicality and reliability of the improved TransTrack model in real indoor environments.

5. Conclusion

This paper concentrated on optimizing the TransTrack model to better track multiple targets for an indoor service robot. The adaptability of the model in certain indoor environments was improved by improving tracking accuracy, better handling of occlusions, and real-time performance. This demonstrates the feasibility of deploying the optimized TransTrack model in real-world indoor robots, especially in challenging conditions such as low light and frequent occlusions. However, the system still has some limitations when tracking fast-moving targets, and more experiments and further research are needed to address these issues. In addition, since the experiments were conducted in a simulated environment, their performance may deviate from the actual performance. In the future, more attention can be paid to using multimodal sensors such as depth cameras or inertial measurement units and validating the system through actual deployment.

References

[1] Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T. K. (2021). Multiple object tracking: A literature review. Artificial intelligence, 293, 103448.

[2] Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., ... & Luo, P. (2020). Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv: 2012.15460.

[3] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv: 1804.02767.

[4] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence, 44(6), 2872-2893

[5] Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.

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

[7] Kalyanaraman, A., Griffiths, E., & Whitehouse, K. (2016, May). Transtrack: Tracking multiple targets by sensing their zone transitions. In 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 59-66). IEEE.

[8] Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 43(2), 652-662.

[9] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[10] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[11] Yu, C., & Shin, Y. (2024). SAR ship detection based on improved YOLOv5 and BiFPN. ICT Express, 10(1), 28-33.

[12] Ranaraja, S. (2024). Occlusion aware obstacle prediction using people as sensors. arXiv preprint arXiv: 2412.20376.

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

[14] Krogh, J. W., & Krogh, J. W. (2020). MySQL workbench. MySQL 8 Query Performance Tuning: A Systematic Method for Improving Execution Speeds, 199-226.

[15] Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better imagenet models transfer better?. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2661-2671).

[16] Femin, A., & Biju, K. S. (2020, June). Accurate detection of buildings from satellite images using CNN. In 2020 international conference on electrical, communication, and computer engineering (ICECCE) (pp. 1-5). IEEE

[17] Weng, X., Wang, J., Held, D., & Kitani, K. (2020). Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. arXiv preprint arXiv: 2008.08063

[18] Angah, O., & Chen, A. Y. (2020). Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy. Automation in Construction, 119, 103308.

[19] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... & Wang, X. (2022, October). Bytetrack: Multi-object tracking by associating every detection box. In European conference on computer vision (pp. 1-21). Cham: Springer Nature Switzerland.

[20] Faccio, D., & Velten, A. (2018). A trillion frames per second: the techniques and applications of light-in-flight photography. Reports on Progress in Physics, 81(10), 105901.

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

[22] Azhar, M. I. H., Zaman, F. H. K., Tahir, N. M., & Hashim, H. (2020, August). People tracking system using DeepSORT. In 2020 10th IEEE international conference on control system, computing and engineering (ICCSCE) (pp. 137-141). IEEE.