Multi-Object Tracking in Large-Scale Object Scenarios Using TrackFormer

1st Yinsong Wang
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
yinsong@chalmers.se

2nd Dian Wang
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden
dianw@chalmers.se

***Abstract*—This paper explores the use of transfer learning to adapt a multi-object tracking model, pre-trained on the MOT20 dataset for full-body tracking, to a new task of head tracking using the HT21 dataset. The aim is to improve tracking accuracy in high-density scenes where full-body detection is prone to failure due to crowding and occlusion. By fine-tuning the MOT20 model with selected subsets of HT21, we successfully shifted the model's focus from body to head tracking, leading to a significant increase in detection performance, particularly in crowded environments. However, the new model demonstrated limitations in handling head occlusion and consistent identity tracking after occlusions. Additionally, issues such as poor detection of stationary individuals persisted. Overall, the transfer learning approach proved effective, demonstrating its ability to adapt existing models for new tasks with improved accuracy, while also revealing areas for further refinement. Code available at GitHub.**

***Index Terms*—Multi-object tracking, Transformer, CNN, TrackFormer, Tracking-by-attention**

I. Introduction

In various fields, the *attention* mechanism has proven effective for tracking, as it enables the modeling of dependencies irrespective of their distance within input or output sequences [2]. Building on this concept, *transformer* architecture has been introduced to address multi-object tracking (MOT) challenges in diverse ways, particularly in the MOT challenge event.

The goal of MOT is to first recognize objects and then track their trajectories throughout a video sequence. With the development of advanced detection algorithms in the computer vision community, many researchers have adopted the tracking-by-detection paradigm: first detecting objects in individual frames and then associating these detections across the video sequence.

In this work, we utilize *TrackFormer*, an effective combination of convolutional neural networks (CNNs) and transformer architecture in crowded scenes. TrackFormer introduces the tracking-by-attention paradigm, which not only applies attention mechanisms to data association but also jointly performs tracking and detection. In this specific scenario, HeadHunter [3], a head-tracking neural network, detects 36 heads whereas Faster-RCNN [4], a full body detection neural network, can detect only 23 pedestrians out of 37 present in a crowded scene. In a nutshell, using the powerful ability from TrackFormer in head-tracking scenes could have a great effect.

II. Related Work

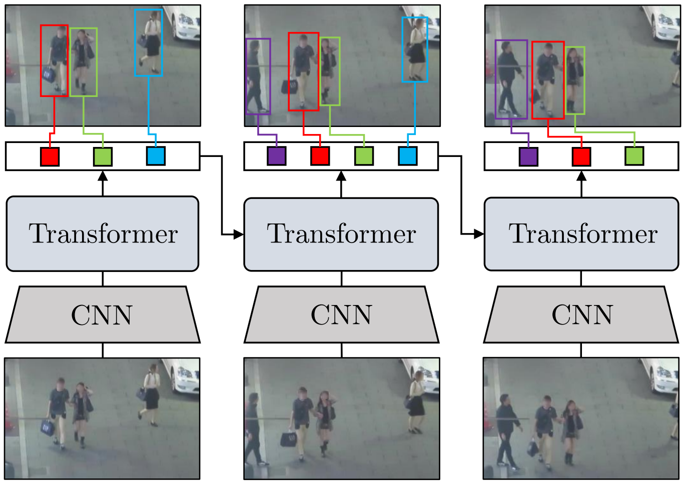


Fig. 1: Trackformer jointly tackles object detection and *track-by-attention* with Transformer. [1]

As an end-to-end trainable Transformer encoder-decoder architecture, **TrackFormer** utilizes a CNN to extract frame-level features, which are then fed into the encoder. The decoder transforms queries into bounding boxes associated with object identities. Each query represents an object and tracks it across both spatial and temporal sequences in an autoregressive manner. When a new object is detected by a static object query, that query is incorporated into subsequent tracking queries for future frames. At each frame, the encoder-decoder applies self-attention to compute output embeddings, generating both bounding boxes and identities, as well as updated tracking queries for the next frame.

A. Track by attention

The model operates in four key steps:

- Extracting frame-level features using a CNN backbone, in this case, ResNet-50.
- Encoding these frame features with self-attention in the Transformer encoder.

- Decoding queries using self-attention and encoder-decoder attention in the Transformer decoder.
- Mapping queries to bounding boxes and class predictions via multilayer perceptrons (MLPs).

The output embeddings accumulate bounding box and class information over multiple decoding layers, the output embeddings are initialed by 2 different queries: i) static object queries, allowing the model to track from the current frame; ii) autoregressive track queries, which are responsible for tracking objects across frames.

With these properties, the decoder can simultaneously generate both detection and tracking in a unified manner, known as the tracking-by-attention paradigm.

B. Track queries

The concept of *track query* is introduced in the decoder to enable frame-to-frame track generation. Track queries follow objects throughout a sequence, preserving their identity information while adapting to their changing positions in an autoregressive manner.

In figure 2, we provide a visual depiction of the track query concept. At frame $t = 0$, initial detections generate new track queries that follow their corresponding objects through subsequent frames. To achieve this, N_{object} object queries (in white) are decoded into output embeddings for potential track initialization. Each valid object detection $b0_0, b1_0, \ldots$, with a classification score above the threshold σ_{object} (i.e., output embeddings that do not predict the background class), initializes a new track query embedding.

Since not all objects appear in the first frame, the track identities $K_{t=0} = \{0, 1, \ldots\}$ only represent a subset of all potential tracks K. During the decoding step at any frame $t > 0$, track queries generate additional output embeddings corresponding to distinct identities (shown in different colors). The combined set of N_{object} and N_{track} output embeddings is initialized by learned object queries and temporally adapted track queries, respectively.

III. Method

A. Data pre-processing

In this project, two different datasets are utilized for the multi-object tracking: MOT20 and HT21. These two datasets serve distinct purposes - MOT20 is a full-body tracking dataset, while HT21 focuses on head tracking, particularly in crowded scenes.

The data is structured using the COCO (Common Objects in Context) format, which is widely adopted in computer vision tasks. In COCO format, annotations are stored in a JSON file, where each object instance is described by bounding boxes, categories, and segmentation masks. The bounding boxes are encoded using four parameters: $[x, y, width, height]$, where x and y represent the coordinates of the top-left corner, and the width and height define the size of the box. To convert the HT21 dataset into this format, a data generation function processes the unzipped images, producing COCO-formatted data that will be used later in the transfer learning and evaluation stages.

B. Transfer learning

Given the distinct focus of the datasets (full-body tracking in MOT20 and head tracking in HT21), the project adopts a transfer learning approach to fine-tune a pre-trained model trained on MOT20 using the HT21 dataset. Transfer learning allows the model to leverage previously learned representations - such as feature extraction layers focused on person detection - and adapt them for the new task of head tracking.

The pre-trained MOT20 model tracked full bodies but was less effective for heads, particularly in dense environments. To address this, we used HT21-02, HT21-03, and HT21-04 subsets to retrain the model for head tracking. Cross-validation with the HT21-01 subset ensured robust evaluation and reduced over-fitting risks.

Training ran for 10 epochs, balancing performance with time and efficiency. The fine-tuned model was preserved for subsequent tasks, such as evaluation on unseen data and the generation of evaluation metrics.

C. Performance and Evaluation

After completing the transfer learning process, the performance of the fine-tuned model is evaluated to assess its effectiveness in head tracking. Since the ground truth annotations for the test set in the MOTChallenge datasets are not publicly available, the evaluation is first carried out using the validation set from HT21-01. We then employ the official TrackEval tool, which is widely used for multi-object tracking evaluation, to measure the model's performance against key metrics.

The primary evaluation metrics used in the project are as follows:

- MOTA (Multiple Object Tracking Accuracy) [5]: This metric assesses overall tracking accuracy by considering false positives, false negatives, and identity switches. Higher values indicate better performance.
- MOTP (Multiple Object Tracking Precision) [5]: This score measures the accuracy of the predicted object positions compared to the ground truth, focusing on how well the model predicts the location of tracked objects.
- IDF1 (Identification F1 Score) [6]: This score evaluates the accuracy of object re-identification across frames. It is particularly important in tracking tasks where object identities must be preserved.
- HOTA (Higher Order Tracking Accuracy) [7]: This score provides a balanced view of tracking performance by combining detection accuracy and association accuracy, focusing on both object detection and identity preservation.
- MTR (Mostly Tracked): The percentage of ground-truth trajectories that are successfully tracked for more than 80% of their duration.
- MLR (Mostly Lost): The percentage of ground-truth trajectories that are tracked for less than 20% of their duration.

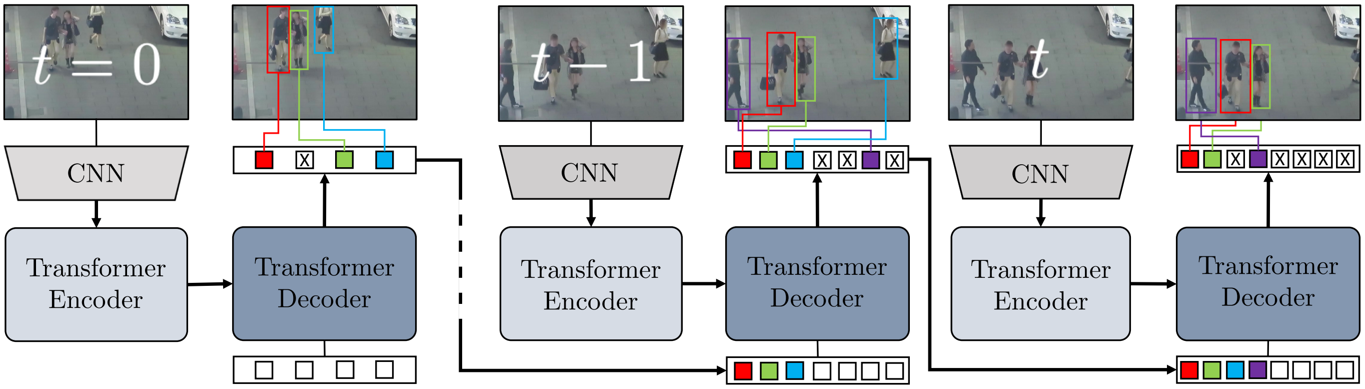


Fig. 2: **TrackFormer** casts multi-object tracking as a set prediction problem performing joint detection and **tracking-by-attention**. The architecture consists of a CNN for image feature extraction, a Transformer encoder for image feature encoding, and a Transformer decoder that applies self- and encoder-decoder attention to produce output embeddings with bounding box and class information. At frame t = 0, the decoder transforms Nobject object queries (white) to output embeddings either initializing new autoregressive **track queries** or predicting the background class (crossed). On subsequent frames, the decoder processes the joint set of Nobject + Ntrack queries to follow or remove (blue) existing tracks as well as initialize new tracks (purple). [1]

Following this initial evaluation of the HT21-01 validation set, we proceed to test the model on completely unseen datasets, specifically HT21-12 and HT21-13. These datasets have never been encountered by the model during training, providing a true test of its generalization capabilities. In this analysis, both models are evaluated on the same datasets, and keyframes are manually selected to visually compare the tracking accuracy between full-body tracking and head tracking. This comparison allows us to observe how the fine-tuned model performs under unseen crowded scenes.

IV. Results

Based on the evaluation methods described above, both quantitative metrics and qualitative visual results will be demonstrated below to show how the transfer learning approach improves tracking accuracy.

A. Evaluation metrics

The below table demonstrates the evaluation metrics for the HT21-01 dataset:

HT21-01	**MOTA**	**MOTP**	**IDF1**	**HOTA**	**MTR**	**MLR**
Result(%)	66.77	71.109	67.692	49.46	45.57	16.456

TABLE I: *Evaluation metrics for dataset HT21-01*

The model's MOTA (66.77%) indicates solid overall tracking performance, accounting for false positives, false negatives, and identity switches. This score suggests that the model is reliable at detecting and tracking multiple objects in crowded scenes. MOTP (71.11%), which measures the precision of object localization, shows that the model accurately predicts object positions, with bounding boxes closely aligning with the ground truth. Meanwhile, the IDF1 score (67.69%) highlights the model's effectiveness at maintaining consistent object identities across frames, though there is still room for improvement in identity preservation.

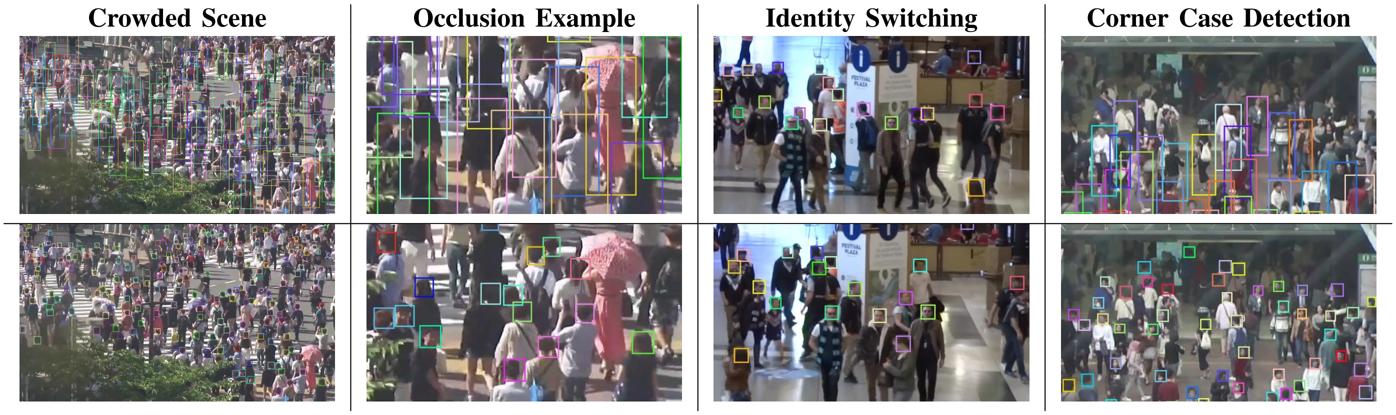
The HOTA score (49.46%) reflects a balanced ability to detect objects and maintain their associations over time, but the score also suggests that the model could improve in maintaining track continuity in more challenging scenarios. Looking at trajectory metrics, MTR (45.57%) indicates that nearly half of the objects were tracked successfully for most of their duration, while MLR (16.46%) shows that a relatively small portion of objects were mostly lost. These results suggest that while the model handles tracking well, identity switches and track fragmentation could be optimized further to enhance long-term tracking performance.

B. Visual Results

To further evaluate the impact of transfer learning on head tracking, we compared the results of the original MOT20 pretrained model and the fine-tuned model on the HT21-13 and HT21-12 datasets using four key image sequences.

Crowded Scene: In a densely populated intersection where the total number of individuals was 216, the transfer learning model demonstrated a significant improvement in detection accuracy. The original model was able to detect only 126 individuals, whereas the fine-tuned model detected 196 individuals. This represents a substantial increase in detection accuracy, highlighting the benefits of adapting the model for head tracking in high-density scenarios. The head tracking boxes make the image much clearer and easy to see as well.

Occlusion Example: In this comparison, the frame captures a woman in a red dress standing with her back to the camera, holding an umbrella. Interestingly, the original model trained for full-body tracking successfully detected the person despite the partial occlusion, while the transfer learning model, focused on head tracking, failed to detect her due to the occluded



Crowded Scene

Occlusion Example

Identity Switching

Corner Case Detection

Fig. 3: Performance Comparison of Detection Modes in a Crowded Scene
Crowded Scene: In a crowded scene with 216 people in total, the full-body detection mode successfully detected 126 individuals, whereas the head tracking mode detected 196 individuals. **Occlusion Example:** A notable case is an individual wearing a pink dress and holding a pink umbrella. This person was detected by full-body detection but missed by the head detection mode. **Identity Switching:** A person in a white T-shirt initially had a pink detection mark before passing an information board. Upon reappearing, the detection system assigned a green detection mark, indicating the model incorrectly identified them as a different person. **Corner Case Detection:** Pedestrians near the corners were rarely detected by the full-body detection mode. Although head detection improved performance in these areas, there remains significant room for further enhancement in detecting corner cases.

head. This suggests that full-body tracking may still be more effective in specific cases where the head is not visible.

Identity Switching: In a sequence of close consecutive frames, a person passing behind an obstruction was misidentified by the fine-tuned model after reappearing, treating the individual as a new object. This identity switch caused minor tracking errors, indicating that the head tracking model may struggle with consistent re-identification when occlusions occur, leading to occasional tracking fragmentation.

Corner Case Detection: In both models, there was a consistent issue in detecting individuals standing still in a corner of the frame. Neither the original model nor the transfer learning model performed well in detecting stationary people in this particular scenario, suggesting that motion-based features or low visibility conditions may impact detection performance.

V. Conclusions

In this project, we demonstrated that the TrackFormer model can effectively adapt to diverse use cases through transfer learning. By fine-tuning a MOT20 pre-trained model with the HT21 dataset, we successfully shifted its focus from full-body tracking to head tracking, resulting in substantial performance improvements in high-density environments. The enhanced detection accuracy in crowded scenes showcases the model's versatility and ability to handle specific challenges associated with head tracking.

However, this transition introduces certain limitations. The model's ability to detect individuals during partial head occlusion, which was better managed by the original full-body tracking model, declined noticeably. Additionally, challenges such as detecting stationary individuals in corners persisted in the fine-tuned version, and occasional identity switches occurred when individuals reappeared after passing behind obstacles.

Despite these limitations, the results indicate that the TrackFormer model, enhanced through transfer learning, can effectively meet the demands of head tracking tasks. This adaptability highlights the model's potential for broader applications and its capability to excel in specialized contexts while maintaining robust performance.

References

[1] T. Meinhardt, A. Kirillov, L. Leal-Taixé, and C. Feichtenhofer, ”TrackFormer: Multi-Object Tracking with Transformers,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, June 2022, pp. 8844–8854.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ”Attention is All You Need,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[3] R. Sundararaman, C. D. A. Braga, E. Marchand, and J. Pettré, ”Tracking pedestrian heads in dense crowd,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[5] Bernardin, K. & Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *Image and Video Processing, 2008(1):1-10, 2008.*

[6] Ristani, E., Solera, F., Zou, R., Cucchiara, R. & Tomasi, C. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *In ECCV workshop on Benchmarking Multi-Target Tracking, 2016.*

[7] Jonathon Luiten, A.O. & Leibe, B. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *International Journal of Computer Vision, 2020.*