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

Yinsong Wang, Dian Wang

Chalmers University of Technology
Electrical Engineering Department

CHALMERS UNIVERSITY OF TECHNOLOGY

Introduction

Figure 1: Full-body detection vs Head-tracking in a crowded scene

The primary goal of Multi-Object Tracking(MOT) is to recognize objects and track their trajectories throughout a video sequence, often using the tracking-by-detection paradigm: first detecting objects in individual frames, and then associating these detections across frames. In this work, we use TrackFormer[1], which combines integrated CNNs with transformer architecture to track objects in crowded scenes. TrackFormer applies the tracking-by-attention paradigm, enabling simultaneous tracking and detection through attention-based data association. For instance, in figure 1, head-tracking mode detects 196 heads, while full body detection detects only 126 pedestrians out of 216 present[2]. This highlights TrackFormer's potential in crowded scenes to improve performance in challenging tracking tasks.

Method & Results

- **Data Pre-processing**
  - HT21 Dataset Preparation
  - COCO Format Generation
- **Transfer Learning**
  - Training Set: HT21-02, HT21-03, HT21-04
  - Cross Validation Set: HT21-01
  - Pre-trained Model: MOT20_checkpoint_epoch_50.pth
  - Epoch: 10
- **Performance and Evaluation**
  - Apply the evaluation metrics: MOTA[2], MOTP[2], IDF1[3], HOTA[4], MTR, MLR.

Results:

Occlusion Example | **Identity Switching** | **Corner Case Detection**

Figure 2: Performance Comparison of Detection Modes in a Crowded Scene

Architecture

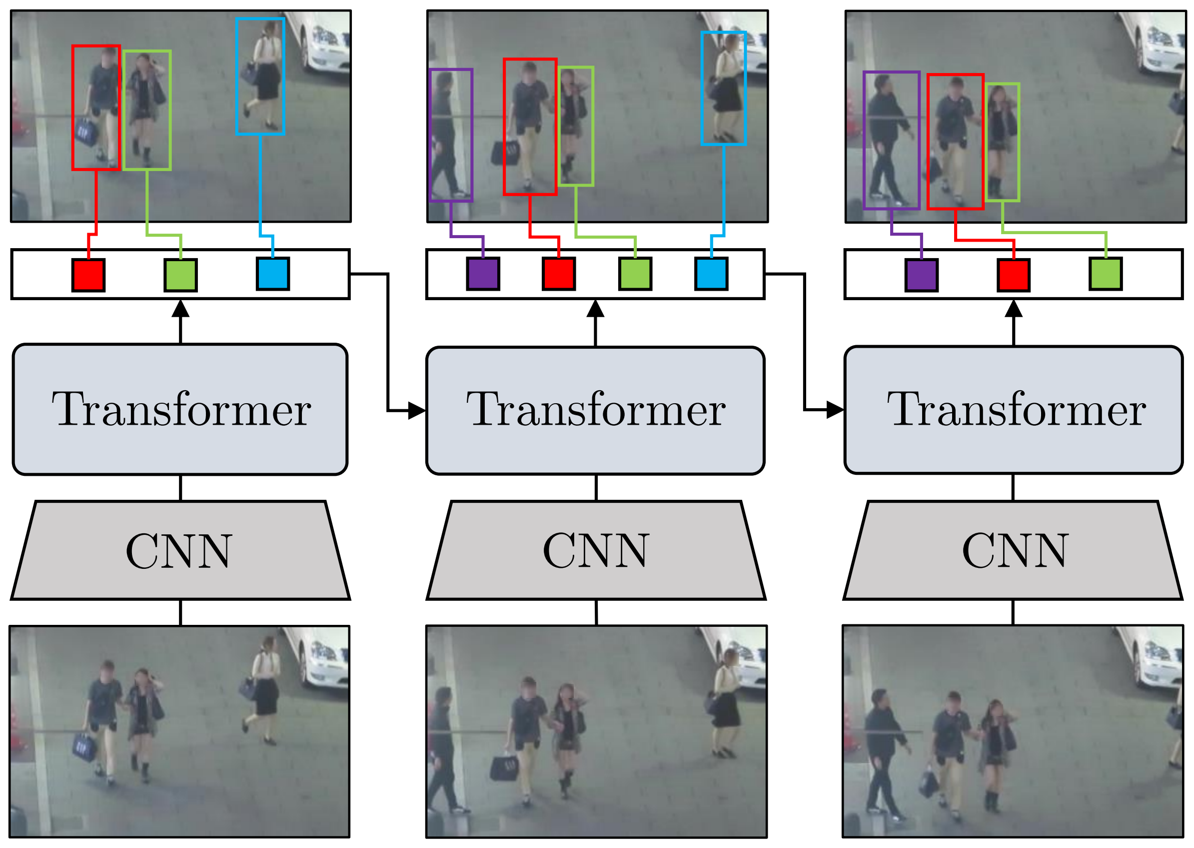

Figure 3: **TrackFormer** jointly tackles object detection and *track-by-attention* with Transformer.[1]

As an end-to-end trainable Transformer encoder-decoder architecture, **TrackFormer** integrates CNN with Transformer.

- It utilizes a **CNN** to extract frame-level features, which are then passed 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.
  - Generates both **bounding boxes** and **identities**.
  - Updates **tracking queries** for the next frame.

Conclusion

This project demonstrated substantial performance improvements in high-density environments, confirming the ability of transfer learning to adapt the TrackFormer model and enhance detection accuracy in crowded scenes by shifting its focus from full-body tracking to head tracking.

Limitations:

- The model's ability to detect individuals with partial head occlusion declined compared to the original full-body tracking model.
- Challenges in detecting stationary individuals in corners persisted in the fine-tuned version.
- Occasional identity switches occurred, especially when individuals reappeared after passing behind obstacles.

Overall, this project highlights the power of TrackFormer in enhancing model performance for large-scale object scenarios.

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] 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.

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

[4] 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.*

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

yinsong@student.chalmers.se dianw@student.chalmers.se