Sciences Methods and Technologies International Journal (SciMeTech)

ISSN: 3085-5284, journal homepage: www.scimetech.com

Hybrid Multi-Person Tracking Framework for Dense and Dynamic Environments.

Oussama Lachihab, My Ahmed El Kiram, Latifa Er-rajy.

Laboratory of Computer Science and Smart Systems, Faculty of Science Semlalia, Cady Ayyad University, Marrakech MOROCCO.

How to cite:

Oussama Lachihab, My Ahmed El Kiram, Latifa Er-rajy, "Hybrid Multi-Person Tracking Framework for Dense and Dynamic Environments", Sciences Methods and Technologies International Journal (SciMeTech), Vol 2, Issue 2, p 186-191

Keywords:

Multi-Object Tracking
Yolov8s
Face Detection
Person Tracking
Appearance Features
Kalman Filter
DINOv2

Abstract

Multi-object tracking in crowded scenes remains challenging due to occlusions, appearance ambiguity, and unreliable detections. We propose a tracking framework that leverages both body and face cues to improve identity consistency. Our method uses YOLOv8s to detect persons and faces, followed by a spatial association step to link them into person-level observations. We then employ DINOv2 to extract feature embeddings from detected regions, which are used within a tracking pipeline that combines motion and appearance information. The tracker adopts a cascade matching strategy to associate detections across frames and handle challenging cases such as occlusions and missed detections. Experimental results on a sequence of MOT17 dataset demonstrate that incorporating complementary cues can help maintain identity consistency, while also revealing limitations in dense scenarios.

I. Introduction

Recent advances in artificial intelligence have dramatically expanded the deployment of video-based surveillance systems in smart cities[1], public transportation hubs[2], airports[3], shopping malls[4], and large-scale outdoor events[5]. In these environments, AI-powered cameras continuously ingest multi-camera streams to monitor crowd flows, detect unauthorized access, and flag potential security incidents in real time. Underpinning many of these systems is the task of multi-person tracking (MPT)[6], which enables authorities to maintain persistent identities across space and time, trace trajectories, and reconstruct interaction patterns among individuals. Robust multi-object tracking (MOT) is therefore a core computer vision problem and a critical enabler of proactive security operations, including anomaly detection, threat localization, and post-incident forensics.

Despite progress, tracking multiple persons in dense, dynamic environments remains highly challenging. In crowded scenes, individuals frequently overlap, move rapidly, and execute non-rigid motions, leading to severe occlusions, short-term disappearances, and frequent identity switches. At the same time, large variations in scale, pose, lighting, and camera viewpoint complicate appearance-based re-identification, while abrupt changes in crowd density require the system to adapt association strategies on the fly. These factors make it difficult to maintain stable tracklets over long sequences, especially when the active field-of-view includes hundreds of individuals in a single frame. Current benchmark studies on datasets such as CrowdTrack[7] and DanceTrack[8] have shown that even state-of-the-art trackers exhibit significant performance degradation under such conditions, highlighting a persistent gap between controlled benchmarks and real-world urban scenes. A wealth of existing MOT methods[9] has been developed to address these challenges, including detector-centric pipelines such as DeepSORT[10], association-aware frameworks like ByteTrack[11], and end-to-end transformer-based approaches such as MOTR[12] and TrackFormer[13]. While these methods have achieved strong performance on standard benchmarks such as MOT17, MOT20[14], and DanceTrack, they often struggle with highly crowded scenes and varying crowd densities. To address these limitations, we propose an adaptive transformer-based multi-person tracking framework for dense and dynamic environments.

First, we demonstrate that a frozen DINOv2-small[15] backbone, pretrained via self-supervised self-distillation with no Re-ID supervision, yields sufficiently discriminative instance-level representations for multi-person tracking and reidentification without domain-specific fine-tuning, establishing self-supervised vision foundation models as a viable zeroannotation alternative to supervised Re-ID embedders in tracking pipelines.

Second, a density-aware adaptive tracking regime in which a per-frame scene density estimator drives continuous interpolation of matching parameters and selectively activates a three-round cascade association strategy, eliminating both hard mode-switching instability and unnecessary computation across varying crowd conditions.

II. Methods

a. Overview:

Our framework performs multi-object tracking by integrating detection, representation learning, and adaptive data association into a unified pipeline (Fig. 1). Given an input frame, the system first detects faces and bodies using YOLOv8s[16]. Detected instances are spatially associated to form person-level observations, which are then encoded into discriminative embeddings using DINOv2 and a transformer-based fusion module (Fig. 2). Finally, an adaptive cascade tracker performs multi-stage data association using motion, geometry, and appearance cues.

A key design principle of our approach is that body detections provide the primary identity signal, while face information is treated as an optional cue that enhances discriminability when reliable.

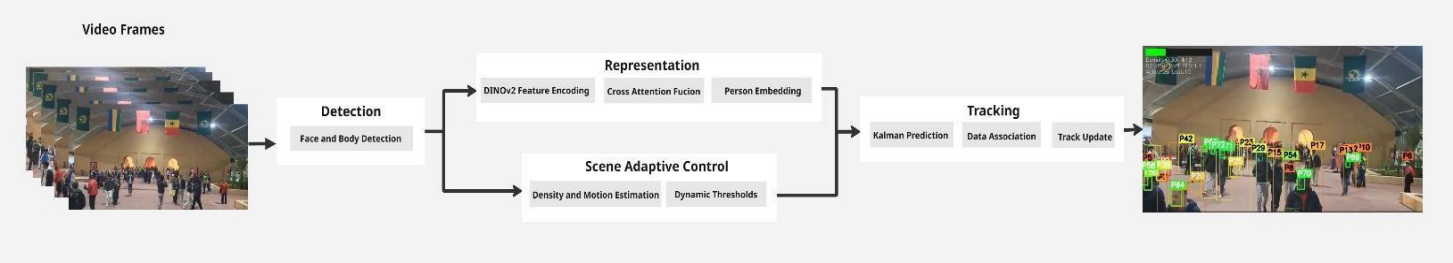

Fig. 1. Overview of the proposed multi-object tracking framework using YOLOv8 detections, DINOv2-based embeddings, cross-attention fusion, and adaptive cascade tracking.

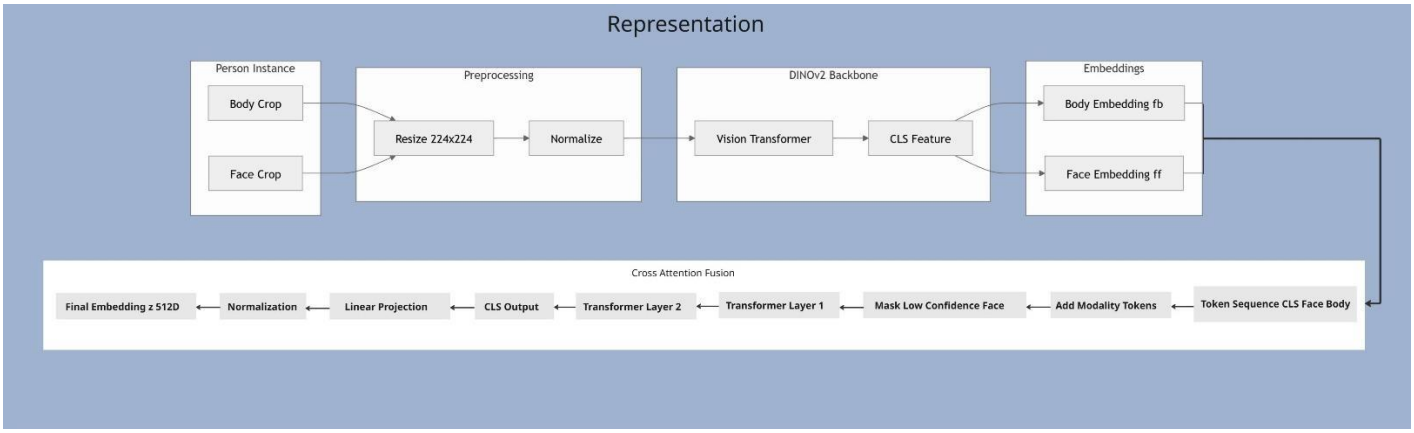

Fig. 2. Detailed representation module where face and body embeddings are fused via cross-attention with confidence-aware masking to produce a robust person embedding.

b. Detection:

We employ YOLOv8s to jointly detect faces and full-body bounding boxes. To improve robustness under challenging conditions, detections are divided into high-confidence **Dh** and low-confidence **Dl** sets using two thresholds.

c. Person-Level Representation :

1. Crop Extraction and Preprocessing.

For each person instance, we extract face and body image crops from the input frame. Each crop is resized to 224×224 and normalized using standard ImageNet statistics before feature extraction.

2. Feature Encoding with DINOv2.

We use DINOv2 as a frozen backbone to extract modality-specific embeddings. For each crop, we obtain the representation from the CLS token of the final transformer layer:

$f_b, f_f \in R^{384}$

where f_b and f_f denote body and face embeddings, respectively. All embeddings are L2-normalized to facilitate cosine similarity computation.

3. Cross-Attention Fusion.

To combine face and body features, we propose a lightweight transformer-based fusion module (Fig. 2). Instead of relying on fixed-weight averaging, we construct a token sequence:

[CLS, f_b, f_f]

Each token is augmented with a learned modality embedding to encode whether it originates from the face or body stream. To handle unreliable face detections, we introduce a confidence-based masking mechanism that suppresses the face token when its detection confidence falls below a predefined threshold. The sequence is processed using a two-layer transformer encoder:

H=Transformer([CLS, f_b, f_f])

The final person embedding is obtained from the CLS token:

$z = Norm(Proj(H_{CLS})) \in R^{512}$

This formulation enables adaptive fusion, where the model dynamically adjusts the contribution of each modality based on its reliability. In particular, the system degrades gracefully to body-only representations when face information is absent or noisy.

d. Adaptive Cascade Tracking:

1. Motion Model.

Each tracked object is modeled using a Kalman filter with state:

$(c_x, c_y, w, h, v_x, v_y, v_w, v_h)$

which predicts the object's motion across frames.

2. Multi-Stage Data Association

We perform data association using a three-stage cascade:

- **Round 1 (Primary matching):** high-confidence detections are matched to active tracks using a weighted combination of IoU and embedding similarity.
- **Round 2 (Geometric recovery):** low-confidence detections are matched to unmatched active tracks using IoU only.
- **Round 3 (Re-identification):** remaining high-confidence detections are matched to previously lost tracks using appearance similarity.

Matching is solved using the Hungarian algorithm.

3. Scene-Adaptive Control

To handle diverse environments, we introduce a scene-aware controller that estimates object density, and motion variation from the current frame. These signals dynamically adjust the balance between IoU and embedding costs, matching thresholds, activation of cascade stages and Kalman process noise. This enables the tracker to adapt between sparse and highly crowded scenes without retraining.

4. Track Update and Lifecycle

Each track maintains an appearance embedding updated via exponential moving average:

$\boldsymbol{z_t = \alpha z_{t-1} + (1 - \alpha)\hat{z}_t}$

Tracks transition between three states: active (currently tracked), lost (temporarily unmatched), terminated (removed after prolonged absence). This strategy ensures stable identity preservation over time.

III. Results and discussion

a. Experimental setup :

We evaluate the proposed tracking framework on the MOT17 benchmark, using the MOT17-02-DPM sequence. The sequence contains 600 frames at a resolution of 1920×1080 and 30 FPS, with 62 annotated identities. Ground-truth annotations were filtered to remove distractor classes, resulting in 18,581 valid annotations. Tracking performance is evaluated using standard MOT metrics, including HOTA, MOTA, and IDF1. Runtime performance is also reported with a detailed stage-wise breakdown.

b. Quantitative Results:

Table 1. Quantitative tracking performance on the MOT17 MOT17-02-DPM sequence. We report standard multi-object tracking metrics, including HOTA, DetA, AssA, MOTA, and IDF1. The results reflect the performance of the proposed framework under partial temporal coverage, highlighting the impact of detection and association quality on overall tracking accuracy.

Metric	Value
HOTA	8.48%
DetA	25.22%
AssA	2.85%
MOTA	-1.23%
IDF1	24.51%
ID Switches	366
FPS	4.8

The proposed method achieves moderate detection accuracy (DetA = 25.22%) but struggles in association (AssA = 2.85%), leading to a relatively low HOTA score. The IDF1 score of 24.51% indicates that the learned representation captures some identity information, but identity consistency across frames remains limited. The negative MOTA score is primarily due to missed detections and identity switches, which are amplified by incomplete frame coverage (see Section 4.3).

c. Qualitative results:

Fig. 3. Qualitative tracking results of the proposed framework on challenging crowded scenes. (a–d) show representative frames illustrating multi-person tracking under occlusion, close interactions, and motion. The tracker maintains consistent identities across frames (e.g., P81) despite partial occlusions and viewpoint changes, while some identity switches and missed associations occur in dense regions.

As shown in Fig. 3, the tracker is able to maintain consistent identities across frames despite occlusions, close interactions, and viewpoint changes, demonstrating the effectiveness of the proposed representation and association strategy. In particular, identities such as P81 remain stable over time even in moderately crowded scenarios. However, failure cases are also observed, including occasional identity switches and the presence of floating bounding boxes, where detections are temporarily assigned without proper spatial alignment to visible targets. These artifacts highlight limitations in detection reliability and association under challenging conditions, suggesting room for improvement in both localization and temporal consistency.

d. Coverage analysis:

A key limitation observed in the current implementation is partial temporal coverage. Predictions span frames 1–52, whereas ground truth covers the full range (1–600), resulting in only 9% coverage. This significantly impacts all tracking metrics, particularly MOTA, due to large numbers of false negatives, HOTA, due to missing temporal associations and IDF1, due to fragmented trajectories.

This suggests that the evaluation is currently dominated by early-frame performance, and improving full-sequence processing is expected to yield substantial gains.

e. Runtime analysis:

The system operates at 4.8 FPS, with computation dominated by feature extraction and detection. `

Table 2. Stage-wise runtime analysis of the proposed tracking framework. We report the mean processing time, variance, and relative computational cost of each module. The embedding stage (DINOv2 + fusion) dominates the runtime, while tracking and data association introduce negligible overhead.

Stage	Mean Time (ms)	% Time	FPS
Embedding (DINOv2 + Fusion)	115.63	85.9%	8.6
Detection (YOLOv8)	13.10	9.7%	78.0
Tracking	2.51	1.9%	398.2
Matching	2.43	1.8%	412.3
Kalman Filter	0.67	0.5%	1491.1
Scene Analysis	0.26	0.2%	3919.5

The embedding stage (DINOv2 + fusion) is the primary computational bottleneck, accounting for over 85% of total runtime. In contrast, tracking and association are highly efficient, contributing less than 3% combined.

This indicates that the system is representation-bound rather than tracking-bound, suggesting that optimization efforts should focus on feature extraction efficiency, batching and GPU utilization and lightweight embedding alternatives.

IV. Discussion:

The results highlight several important insights:

1. Representation Quality vs. Association Gap: while the learned embeddings achieve a reasonable IDF1 score, the low association accuracy (AssA) indicates that identity matching remains a bottleneck.
2. Impact of Partial Coverage: the current evaluation is heavily affected by incomplete frame processing. Extending inference to the full sequence is expected to significantly improve all metrics.
3. Computational Trade-offs: the use of DINOv2 provides strong features but introduces substantial computational overhead, limiting real-time performance.
4. Strength of Modular Design: despite low overall scores, the system demonstrates efficient tracking infrastructure, flexible handling of missing modalities and adaptability to scene conditions.

These properties provide a strong foundation for further improvements.

V. Conclusion

We presented an adaptive multi-object tracking framework that integrates detection, representation learning, and hierarchical data association into a unified pipeline. The proposed approach leverages YOLOv8s for joint face and body detection, and leverage DINOv2 to extract robust modality-specific features. A lightweight cross-attention fusion module enables adaptive integration of face and body cues, allowing the system to remain effective even when face detections are unreliable or missing.

In addition, we introduced an adaptive cascade tracking strategy that combines motion, geometric, and appearance signals through a multi-stage association process. A scene-aware controller dynamically adjusts tracking parameters based on crowd density and motion, improving robustness across varying conditions.

Experimental results on a sequence of MOT17 benchmark highlight the effectiveness of the proposed design, particularly in terms of modularity and flexibility. While the current performance is limited by partial temporal coverage and computational overhead from feature extraction, the framework demonstrates a foundation for handling challenging tracking scenarios with incomplete or noisy observations.

Future work will focus on improving temporal coverage, enhancing association accuracy, and optimizing the representation stage to achieve real-time performance. Additionally, exploring lightweight embedding models or end-toend training strategies could further improve both efficiency and tracking robustness.

References

Add references from googlescholar in APA format

[1] Sharma, H., & Kanwal, N. (2025). Video surveillance in smart cities: current status, challenges & future directions. *Multimedia Tools and Applications*, *84*(16), 15787-15832.

[2] Quintana-Ramirez, I., Sequeira, L., & Ruiz-Mas, J. (2021). An edge-cloud approach for video surveillance in public transport vehicles. *IEEE Latin America Transactions*, *19*(10), 1763-1771.

[3] Sujkowski, M., Kozuba, J., Uchroński, P., Banaś, A., Pulit, P., & Gryżewska, L. (2023). Artificial intelligence systems for supporting video surveillance operators at international airport. *Transportation Research Procedia*, *74*, 1284-1291.

[4] Arroyo, R., Yebes, J. J., Bergasa, L. M., Daza, I. G., & Almazán, J. (2015). Expert video-surveillance system for realtime detection of suspicious behaviors in shopping malls. *Expert systems with Applications*, *42*(21), 7991-8005.

[5] Song, X., Sun, L., Lei, J., Tao, D., Yuan, G., & Song, M. (2016). Event-based large scale surveillance video summarization. *Neurocomputing*, *187*, 66-74.

[6] Stadler, D., & Beyerer, J. (2023). An improved association pipeline for multi-person tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3170-3179).

[7] Fu, T., Chen, Y., Chen, Z., Zhao, M., Li, B., & Xue, X. (2025). CrowdTrack: A Benchmark for Difficult Multiple Pedestrian Tracking in Real Scenarios. *arXiv preprint arXiv:2507.02479*.

[8] Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., & Luo, P. (2022). Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2099321002).

[9] Luo, W., Zhao, X., & Kim, T. K. (2014). Multiple object tracking: A review. *arXiv preprint arXiv:1409.7618*, *1*(1), 1. [10] Veeramani, B., Raymond, J. W., & Chanda, P. (2018). DeepSort: deep convolutional networks for sorting haploid maize seeds. *BMC bioinformatics*, *19*(Suppl 9), 289.

[11] 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.

[12] Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., & Wei, Y. (2022, October). Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision* (pp. 659-675). Cham: Springer Nature Switzerland. [13] Meinhardt, T., Kirillov, A., Leal-Taixe, L., & Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8844-8854).

[14] Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., ... & Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

[15] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[16] Lachihab, O., El Kiram, A., & Errajy, L. (2026). Fine-Tuning YOLOv8s for Unified Human and Face Detection in Crowded Environments. *Engineering, Technology & Applied Science Research*, *16*(1), 32348-32356.