Robust Thermal Pedestrian Multi-Object Tracking through Three-Stage Trajectory Refinement

Anonymous CVPR submission

Paper ID 11

Abstract

Thermal multi-object tracking (MOT) is essential for surveillance and safety-critical perception in low-visibility environments, yet it remains challenging due to weak appearance cues, low contrast, frequent occlusions, and unstable object boundaries. We present a three-stage framework for thermal pedestrian MOT on the TP-MOT benchmark, consisting of thermal pedestrian detection, online tracking, and offline trajectory refinement. Our main contribution is a lightweight post-processing module that treats identity continuity as a first-class objective by stitching fragmented tracklets using temporal continuity, spatial proximity, motion consistency, and border-aware constraints. Unlike approaches that rely on heavy re-identification models or costly global optimization, the proposed refinement stage improves identity preservation while allowing the online tracker to remain simple and efficient. The full system further provides a unified pipeline for inference, visualization, evaluation, and submission generation, enabling reproducible benchmarking across tracker variants. Experiments on the PBVS Thermal MOT benchmark show that the proposed refinement stage consistently reduces fragmentation and improves continuity over raw tracking outputs, yielding a robust and practical real-time solution for thermal pedestrian surveillance.

1. Introduction

Multi-object tracking (MOT) in thermal imagery has become increasingly important for surveillance, public safety, intelligent transportation, and other vision systems that must operate reliably under poor illumination, adverse weather, or low-visibility conditions. Unlike RGB cameras, thermal sensors capture infrared radiation and therefore remain effective when color and texture cues are severely degraded. This property makes thermal imaging especially attractive for pedestrian monitoring in nighttime and safety-critical environments. At the same time, thermal pedestrian tracking remains difficult because objects often appear with weak boundaries, limited appearance variation, and highly ambiguous local structure, all of which make detection and identity preservation substantially more challenging than in conventional RGB settings.

Despite steady progress in modern MOT, thermal-based tracking continues to present several unresolved difficulties. First, the low-information nature of thermal imagery reduces the discriminative power of appearance features, making it harder to distinguish nearby pedestrians or recover identities after occlusion. Second, crowded scenes, missed detections, and abrupt motion changes frequently fragment trajectories and induce identity switches. Third, many high-performing tracking systems depend on heavy re-identification modules or complex motion models, which can improve robustness but often increase computational cost and deployment complexity. These limitations are particularly restrictive when the goal is to build a practical system that is both accurate and efficient for real-time thermal pedestrian tracking.

To address these challenges, this codebase presents a practical and extensible framework for thermal pedestrian multiple-object tracking, developed around the PBVS 2025 challenge setting. The repository adopts a modular two-stage pipeline that separates object detection from object association, allowing each stage to be configured and tuned independently. In its default formulation, the system combines a thermal-adapted YOLOv8 detector with a lightweight SORT-based tracking stage, emphasizing strong detection quality, stable association, and minimal runtime overhead. Rather than relying primarily on computationally expensive re-identification, the framework improves identity continuity through carefully tuned association parameters, an online ID-switch postprocessor, and an additional offline tracklet-stitching procedure that merges short fragmented trajectories when spatial and temporal evidence is consistent. This design yields a codebase that is both engineering-oriented and research-friendly, supporting reproducible experimentation, submission generation, and straightforward extension to alternative tracker backends.

CVPR #11

In brief, the main contributions are as follows:

- We present a modular thermal pedestrian MOT method that integrates detection, tracking, identity-repair within a unified pipeline.
- We introduce complementary identity-continuity mechanisms, including online ID-switch correction and offline tracklet stitching, to reduce fragmentation caused by occlusion, missed detections, and complex motion.
- We provide an extensible experimental framework with configurable tracker backend and standardized scripts, making the repository suitable for both practical deployment and future thermal MOT research.

2. Related Work

Modern multi-object tracking (MOT) systems are largely organized around the *tracking-by-detection* paradigm, where an object detector first produces per-frame observations and a tracker subsequently enforces temporal identity consistency. Over the past decade, a diverse family of association strategies, motion models, and auxiliary cues such as appearance or segmentation have been proposed to improve identity stability under occlusion and detection noise. This section reviews the most relevant strands of prior work and situates our system within this landscape.

2.1. Tracking-by-Detection

A foundational approach in modern MOT is SORT [3], which combines Kalman filtering with Hungarian assignment based on geometric overlap . Despite its simplicity, SORT remains widely used due to its efficiency and modular design, making it a natural baseline for detector-centered tracking systems. Subsequent work has focused on strengthening data association while retaining the same basic structure. ByteTrack [17] demonstrated that discarding low-confidence detections prematurely can significantly degrade identity continuity; instead, these detections can be recovered in a secondary matching stage to reduce track fragmentation under occlusion. BoT-SORT [1] further enhances this paradigm by integrating improved motion modeling, global motion compensation, and optional appearance embeddings to improve robustness in complex scenes. In parallel, OC-SORT [4] revisits the Kalman update formulation by emphasizing observation-centric updates, improving tracking stability under non-linear motion and short-term occlusions.

2.2. Identity Recovery and Tracklet Stitching

To address ID switch, where a tracker incorrectly assigns different tracking IDs to the same target, previous works have proposed identity recovery methods or tracklet stitching models that use different sources of information such as appearance, spatial, and temporal cues to associate fragmented tracklets and reassign correct tracking IDs after the online tracking process. These models typically rely on motion features or appearance features of tracklets for tracklet level association. For example, Translink [18] incorporates a CNN and a temporal attention network to extract and encode the appearance features of a tracklet and formulates the merging of tracklet pairs as a binary classification task. AFLink [7] relies only on spatial and temporal information. Some methods [6, 8, 16] adopt feature clustering techniques to merge tracklets and improve performance in multi-camera tracking scenarios. MambaMOT [9] proposes a motion model that acts as a motion predictor and extracts tracklet motion features for tracklet association.

3. Method

The released inference wrapper file instantiates a detector–tracker–repair pipeline for single-class thermal pedestrian multiple-object tracking. In the provided validation configuration, the system processes six thermal sequences (`seq2`, `seq17`, `seq22`, `seq47`, `seq54`, and `seq66`) sampled at 10 FPS with image resolution 960×1280. The overall design is intentionally modular: detections are first generated and cached, trajectories are then estimated online, and identity continuity is subsequently refined by two offline post-processing stages. This decomposition follows the tracking-by-detection philosophy of SORT [3] while adapting it to the fragmentation patterns that are common in thermal imagery [15].

Thermal pedestrian detection. For each frame I_t, we apply a YOLOv8 detector [10] in a single-class setting (*person*). The detector operates on resized inputs of 1920×1920, uses batched inference with batch size 10, and is configured with a very low confidence threshold (10^{-4}) and non-maximum suppression threshold 0.75 in order to preserve recall. Let

$$\mathcal{D}_t = \{(b_i^t, s_i^t)\}_{i=1}^{N_t}$$

denote the set of normalized bounding boxes and confidence scores predicted at time t. Instead of directly coupling detector and tracker in memory, the pipeline stores $\mathcal{D}_t$ as per-frame YOLO text files. This choice improves reproducibility, allows detector/tracker ablations without rerunning the full pipeline, and enables direct inspection of detector failure modes.

Online tracking. Before tracking, the stored detections are converted from normalized YOLO coordinates to image-space (x_1, y_1, x_2, y_2) boxes and filtered with a stricter threshold $s_i^t \geq 0.5$. Online association is then performed with SORT [3], which combines a constant-velocity Kalman predictor with Hungarian matching over bounding-box overlap. Given predicted track boxes $\hat{\mathcal{T}}_t$ and detections

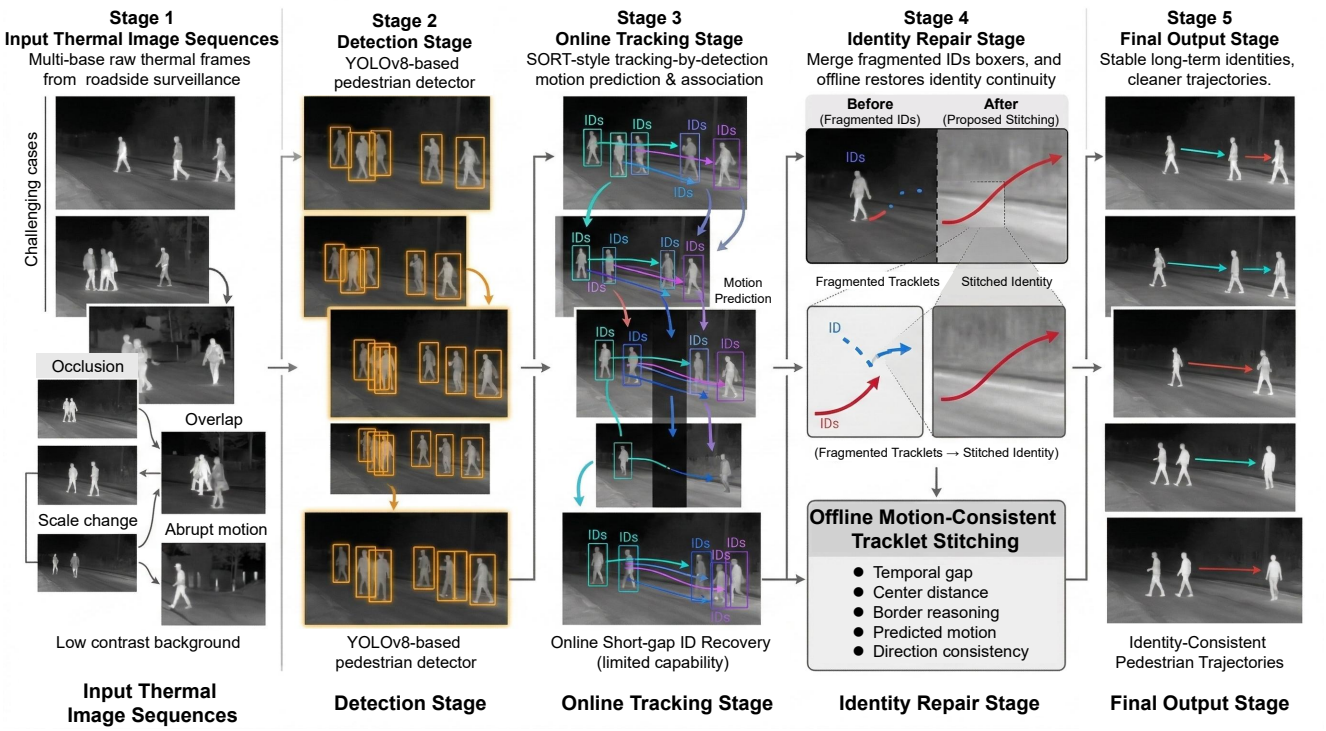

Figure 2. Overview of the proposed thermal pedestrian multi-object tracking pipeline. The framework consists of five stages. Stage 1 shows challenging thermal surveillance inputs with occlusion, overlap, scale changes, abrupt motion, and low contrast. Stage 2 detects pedestrians using a YOLOv8-based detector. Stage 3 performs online tracking-by-detection with to produce initial trajectories. Stage 4 introduces the proposed identity repair module, which stitches fragmented tracklets using temporal gaps, spatial proximity, motion prediction, border reasoning, and direction consistency. Stage 5 outputs stable identity-consistent pedestrian trajectories.

$\mathcal{D}_t$, the association stage maximizes pairwise IoU, or equivalently minimizes

$$C_{ij} = 1 - \mathrm{IoU}(\hat{b}_j^t, b_i^t).$$

In the released configuration, tracks are retained for up to 40 missed frames, and the IoU gate is set to 0.001, which indicates a deliberately permissive association policy. This setting is useful in thermal scenes where shape deformation, low contrast, and partial occlusion can make frame-to-frame overlap unstable even when identity continuity is still visually plausible.

Online short-gap identity repair. A key implementation detail is that identity correction is not deferred entirely to the end of the sequence. Immediately after each SORT update, we apply an online remapping module that keeps a memory of recently lost output identities. If a newborn raw track u appears within $\Delta t \leq 7$ frames of a recently lost track v, and if their centers satisfy

$$\|c_u - c_v\|_2 \leq 60,$$

the newborn trajectory inherits the previous output identity rather than emitting a new ID. Because the benchmark is single-class, the optional label-consistency check is effectively always satisfied. This step is particularly effective for thermal videos, where brief detector interruptions often arise from local temperature ambiguity, partial truncation, or low signal-to-noise regions.

Offline motion-consistent stitching. The first offline post-processing stage explicitly targets short illegal disappear/reappear events. We treat a tracklet ending away from the image border as an implausible disappearance and a tracklet starting away from the border as an implausible appearance. For an old tracklet a and a new tracklet b, with temporal gap

$$\Delta t = t_b^{\mathrm{start}} - t_a^{\mathrm{end}},$$

we estimate endpoint velocity from a 3-point temporal window and extrapolate the previous trajectory as

$$\tilde{c}_a = c_a^{\mathrm{end}} + v_a \Delta t.$$

We merge b into a only when all of the following hold: 1) both endpoints lie outside a 60-pixel border band, 2) $1 \leq \Delta t \leq 30$, 3) the predicted displacement satisfies

$\|\tilde{c}_a - c_b^{\text{start}}\|_2 \leq 80$ pixels, and 4) if the motion magnitude is informative (speed ≥ 0.25 pixels/frame), the angle between the two velocity vectors is at most $45°$ and the speed ratio is at most 3.0. This stage is intentionally conservative: it only repairs fragments that are both temporally short and kinematically consistent.

Learned relinking of longer fragments. A second offline stage addresses longer-range fragmentation using a fixed logistic relinking function. For each candidate predecessor–successor pair, we construct a 48-dimensional feature vector that encodes temporal gap, Euclidean and axis-wise displacement, tracklet lengths, box-size ratios, border distances, edge indicators, global velocity consistency, and multi-window extrapolation errors computed over temporal windows $\{2, 3, 5, 10, 20\}$. A candidate pair is considered only if the gap is at most 60 frames, the endpoint distance is at most 120 pixels, and both tracklets contain at least two observations; moreover, the predecessor must not terminate near a 25-pixel image border. The final relinking probability is

$$p = \sigma(w^\top z),$$

where z is the normalized feature vector and $\sigma(\cdot)$ is the sigmoid function. We accept a merge only when $p \geq 0.95$ and the candidate is mutually optimal for both the old and new tracklet. This mutual-best constraint makes the relinking pass high-precision and prevents cascade errors that would otherwise amplify early mistakes.

Output formatting and reproducibility. After post-processing, each visible target is written in MOT-style format as

$$(f, \text{id}, x, y, w, h, s, \ell, -1, -1),$$

where f is the 1-indexed frame number, id is the repaired identity, (x, y, w, h) is the image-space bounding box, s is the confidence score, and ℓ is the class label. The script optionally renders qualitative overlays during detection, tracking, and post-processing, and finally collects one `seq*_thermal.txt` file per sequence into a dated submission archive.

Overall, the released system can be viewed as a high-recall thermal detector, a lightweight online motion tracker, and a two-stage identity-repair backend. This asymmetry is well suited to thermal MOT: the detector maximizes candidate coverage, the online tracker preserves short-term temporal continuity, and the post-processing stages recover long-range identity consistency under missed detections, brief occlusions, and mid-frame fragmentation.

4. Experiments

4.1. Dataset Splits

The repository configuration tracks six validation sequences: `seq2`, `seq17`, `seq22`, `seq47`, `seq54`, `seq66`. These correspond to challenge-style sequences where public annotations are not distributed in this workspace.

4.2. Evaluation Metrics

We evaluate tracking performance using standard CLEAR MOT and identity-aware metrics that are widely used in multi-object tracking benchmarks and implemented in MOTChallenge-style evaluation toolchains [2, 12]. In particular, we report MOTA, MOTP, IDF1, IDP, IDR, Recall, and Precision. These metrics jointly measure localization quality, and temporal identity consistency.

Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives aggregated over all frames. Let $IDSW$ denote the number of identity switches, and GT denote the total number of ground-truth objects across all frames. For identity-aware evaluation, $IDTP$, $IDFP$, and $IDFN$ denote identity true positives, identity false positives, and identity false negatives, respectively.

MOTA. Multi-Object Tracking Accuracy summarizes the dominant tracking errors into a single score [2]:

$$\text{MOTA} = 1 - \frac{FN + FP + IDSW}{GT}. \tag{1}$$

A higher MOTA indicates fewer missed detections, fewer false alarms, and fewer identity switches. Because it penalizes all three error sources simultaneously, MOTA is commonly used as an overall measure of tracking performance.

MOTP. Multi-Object Tracking Precision measures localization quality on matched object–hypothesis pairs [2]. In overlap-based evaluation, it reflects the spatial agreement between matched predictions and ground-truth objects across all true-positive associations. In many benchmarks, including PBVS, MOTP is reported in a lower-is-better localization-error form.

IDP, IDR, and IDF1. Identity metrics evaluate whether detections are assigned the correct trajectory identity over time [13]. Using identity counts,

$$\text{IDP} = \frac{IDTP}{IDTP + IDFP}, \tag{2}$$

$$\text{IDR} = \frac{IDTP}{IDTP + IDFN}, \tag{3}$$

Table 1. PBVS website leaderboard results.

#	Participant	MOTA	MOTP	IDF1	IDP	IDR	Recall	Precision
1	**Ours**	0.99	0.13	0.86	0.86	0.86	1.00	1.00
2	wqetwet	0.97	0.12	0.97	0.98	0.96	0.98	1.00
3	SKKU-AutoLab	0.97	0.14	0.94	0.95	0.93	0.98	1.00
4	wwww123	0.85	0.12	0.86	0.91	0.82	0.89	0.98
5	spcke	0.82	0.14	0.86	0.90	0.83	0.88	0.94

$$\text{IDF1} = \frac{2\,IDTP}{2\,IDTP + IDFP + IDFN} \tag{4}$$

IDP measures identity precision, IDR measures identity recall, and IDF1 is their harmonic mean. These metrics are particularly sensitive to trajectory fragmentation and identity switches, making them informative for evaluating identity preservation over time.

Recall and Precision. Detection Recall and Precision are defined as

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{6}$$

Recall measures the proportion of ground-truth objects that are successfully detected, while Precision measures the proportion of predicted detections that correspond to true objects. These metrics describe detection coverage and false-alarm behavior within the tracking pipeline [12].

4.3. Experimental Setup

Experiments are conducted on the PBVS Thermal MOT dataset, which contains roadside thermal surveillance sequences captured under nighttime conditions. The dataset exhibits several challenges common in thermal tracking scenarios, including low contrast, partial occlusion, overlapping pedestrians, scale variation, and abrupt motion. All experiments follow the official challenge evaluation protocol. The detector is a YOLOv8 model trained for single-class pedestrian detection, which we adopted the weight from [15]. Inference is performed on resized 1920×1920 frames. Detection results are stored and reused across tracker variants to ensure fair comparison. The tracking stage is evaluated using several widely used MOT association strategies implemented within the same pipeline, including SORT [3], ByteTrack [17], BoT-SORT [1], OC-SORT [4], BoostTrack [14], DiffMOT [11], and a segmentation-assisted variant based on SAM3 [5].

Table 2. Leaderboard ranking based on the weighted score.

Rank	Team	Weighted Score
1	wqetwet	0.7575
2	SKKU-AutoLab	0.7550
3	**Ours**	0.7425
4	wwww123	0.6700
5	spcke	0.6600

Table 3. Extension of YOLOv8 detector TP-MOT evaluation server results reported by [15].

Method	MOTA ↑	MOTP ↓	IDF1 ↑
ByteTrack	0.9173	0.1367	0.7659
BoT-SORT	0.9174	0.1368	0.7605
BoostTrack	0.8654	0.1555	0.7545
DiffMOT	0.8630	0.1611	0.7812
OC-SORT	0.9071	0.1236	0.5685
SAM3	0.9003	0.2104	0.6073
SORT	**0.9844**	**0.1263**	**0.8130**
SORT + Stitching	**0.9853**	**0.1262**	**0.8545**

4.4. Comparison with Existing Trackers

Table 3 compares several representative tracking methods under the same detection inputs. Among the tested trackers, SORT provides the strongest baseline performance in this thermal benchmark.

More complex trackers such as ByteTrack and BoT-SORT achieve competitive detection recall but show lower identity consistency in this dataset. DiffMOT and BoostTrack provide stronger motion modeling but remain sensitive to missed detections and fragmentation under severe occlusion. These observations highlight an important practical insight: in thermal pedestrian tracking, identity continuity is often dominated by fragmentation recovery rather than by increasing tracker complexity alone.

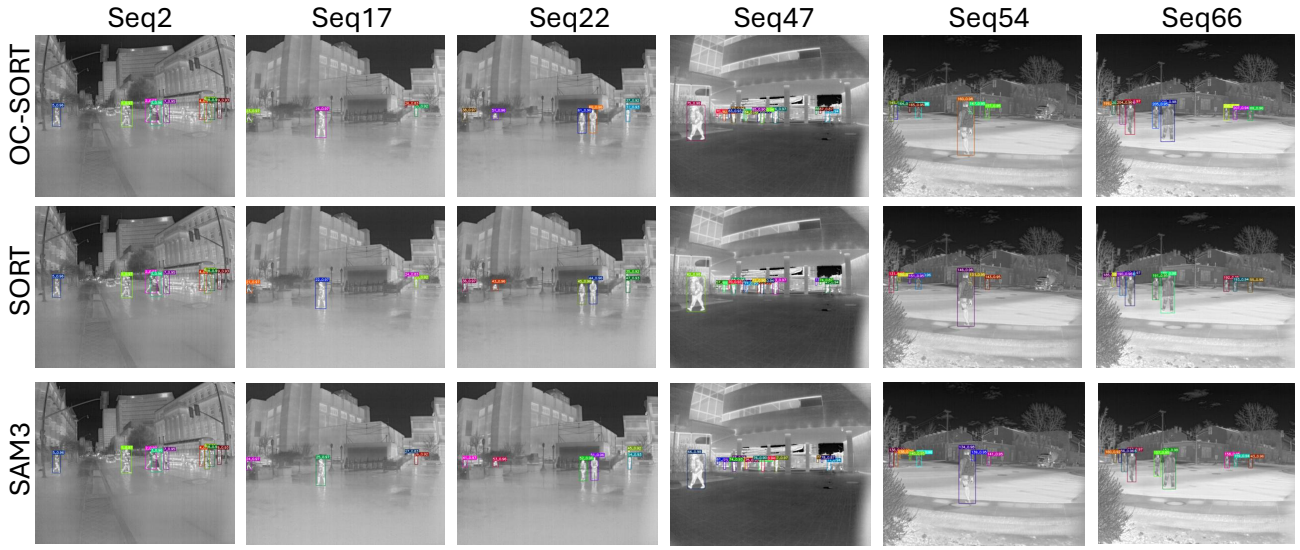

Figure 3. Qualitative comparison of OC-SORT, SORT, SAM3, and the proposed trajectory-refinement method on representative PBVS thermal MOT sequences. The proposed approach yields more stable identity assignments and smoother trajectories under occlusion and motion changes.

4.5. Ablation on Tracklet Stitching

The offline stitching module further reduces trajectory fragmentation by reconnecting short tracklets that satisfy temporal and motion-consistency constraints. Compared with the baseline pipeline, this stage reduces the total number of fragmented tracks and increases average trajectory length. After introducing the proposed identity-repair module, the extended *SORT + Stitching* configuration achieves the best overall performance with a MOTA of 0.9853 and IDF1 of 0.8545. Compared with the vanilla SORT baseline, the proposed refinement stage improves IDF1 by over 4% while maintaining nearly identical localization accuracy.

4.6. Qualitative Results

Figure 3 visualizes representative tracking results across several challenging sequences. Compared with other trackers, the proposed framework produces smoother trajectories and more stable identity assignments when pedestrians undergo temporary occlusion or cross paths. In particular, the proposed border-aware stitching strategy prevents spurious identity births in the interior of the image while still allowing legitimate entries and exits near frame boundaries. This constraint proves especially effective in roadside thermal surveillance scenarios, where pedestrians frequently move through the field of view.

5. Conclusion

This work presents a practical framework for thermal pedestrian multi object tracking that prioritizes robustness, efficiency, and modular design. By separating detection and association into configurable components, the system allows flexible experimentation while maintaining stable real time performance. The integration of online identity repair and offline tracklet stitching further improves identity continuity in challenging thermal scenarios involving occlusion, missed detections, and ambiguous appearance cues. Together, these design choices provide a reliable baseline and an extensible platform for future research in thermal based multi object tracking.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking, 2022. 2, 5

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. In *EURASIP Journal on Image and Video Processing*, 2008. 4

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2, 5

[4] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-Centric SORT: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, 2023. 2, 5

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. 5

[6] Riu Cherdchusakulchai, Sasin Phimsiri, Visarut Trairattanapa, Suchat Tungjitnob, Wasu Kudisthalert, Pornprom Kiawjak, Ek Thamwiwatthana, Phawat Borisuitsawat, Teepakorn Tosawadi, Pakcheera Choppradi, et al. Online multi-camera people tracking with spatial-temporal mechanism and anchor-feature hierarchical clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7198–7207, 2024. 2

[7] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 2

[8] Hsiang-Wei Huang, Cheng-Yen Yang, Zhongyu Jiang, Pyong-Kun Kim, Kyoungoh Lee, Kwangju Kim, Samartha Ramkumar, Chaitanya Mullapudi, In-Su Jang, Chung-I Huang, and Jenq-Neng Hwang. Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency id re-assignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5239–5249, 2023. 2

[9] Hsiang-Wei Huang, Cheng-Yen Yang, Wenhao Chai, Zhongyu Jiang, and Jeng-Neng Hwang. Mambamot: State-space model as motion predictor for multi-object tracking. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[10] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 2024. 2

[11] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19321–19330, 2024. 5

[12] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. In *ECCV Workshops*, 2016. 4, 5

[13] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016. 4

[14] Vukasin Stanojevic and Branimir Todorovic. Boosttrack++: Revisiting the tracking-by-detection paradigm for multiple object tracking, 2024. 5

[15] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Chi Dai Tran, Quoc Pham-Nam Ho, Huy-Hung Nguyen, and Jae Wook Jeon. A novel tuning method for real-time multiple-object tracking utilizing thermal sensor with complexity motion pattern, 2025. 2, 5

[16] Cheng-Yen Yang, Hsiang-Wei Huang, Pyong-Kun Kim, Zhongyu Jiang, Kwang-Ju Kim, Chung-I Huang, Haiqing Du, and Jenq-Neng Hwang. An online approach and evaluation method for tracking people across cameras in extremely long video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7037–7045, 2024. 2

[17] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 2, 5

[18] Yanting Zhang, Shuanghong Wang, Yuxuan Fan, Gaoang Wang, and Cairong Yan. Translink: Transformer-based embedding for tracklets' global link. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2