An Online Approach and Evaluation Method for Tracking People Across Cameras in Extremely Long Video Sequence

Cheng-Yen Yang[1*], Hsiang-Wei Huang[1*], Pyong-Kun Kim[2*], Zhongyu Jiang[1]
Kwang-Ju Kim[2], Chung-I Huang[3], Haiqing Du[4], Jenq-Neng Hwang[1]

1 UW Information Processing Lab 2 ETRI 3 NCHC 4 BUPT

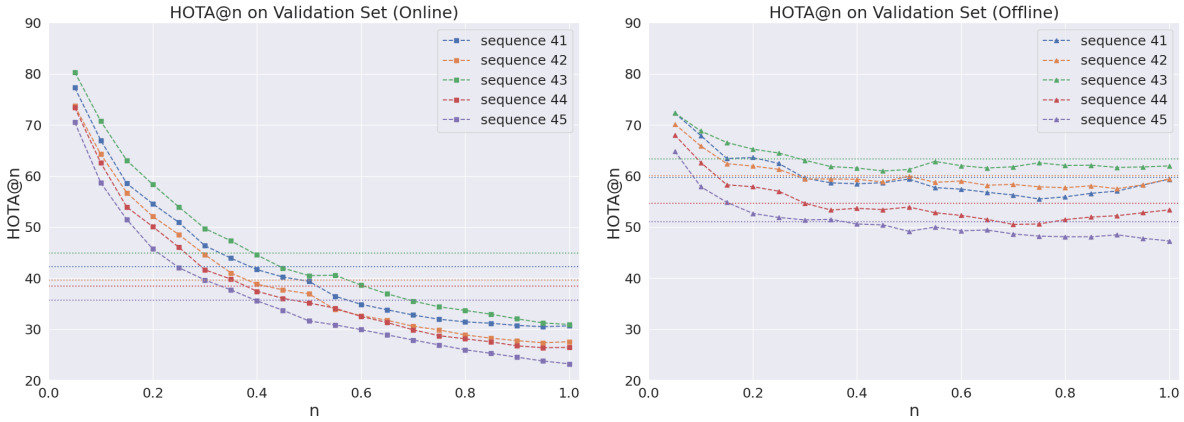

Figure 1. The HOTA@n curve on the AIC24 validation set is shown for various sequences using online method (left) and offline method (right). The online method struggles with re-identification issues lead to less accurate HOTA evaluations in extremely long sequences. The horizontal dashed lines indicate the mHOTA value for each sequence, with each having a length of 24000 frames.

Abstract

Multi-camera Multi-Object Tracking has drawn significant attention in recent years due to its critical role in surveillance, analytics, and related fields. Various challenges, including non-overlapping regions, varying occlusion conditions, and the need for cross-domain generalization in multi-camera tracking systems, remain unsolved in the field. We propose a novel online tracking framework that capitalizes on real-time camera calibration to achieve consistent multi-object tracking across camera networks. Our approach seamlessly integrates spatial and temporal association techniques, ensuring robust tracking even in long-duration videos. However, standard tracking evaluation metrics like CLEAR or HOTA fall short of accurately interpreting the performance of tracking over extended video sequences. Another contribution of this study is the proposal of a new evaluation metric, mHOTA, which provides a better assessment of tracking performance over prolonged periods. Our comprehensive experiments on the AIC24 Multi-Camera People Tracking dataset demonstrate the effectiveness and scalability of our method, along with the capability of the proposed evaluation metric. The code will be available at https://github.com/ipl-uw/mHOTA.

1. Introduction

With the pervasive surveillance in recent years, the ability to effectively track individuals across multiple camera feeds within extended video sequences is paramount for various applications ranging from security surveillance to retail analytics. However, the task becomes significantly challenging when dealing with extremely long video sequences, where traditional tracking methods may fail due to computational constraints or loss of tracking accuracy over time. In this paper, we introduce an innovative online approach and evaluation method tailored specifically for tracking people across cameras in such extremely long

* These authors are the core members of the challenge team: UW-ETRI and consider equal contribution to the work.

sequence scenarios.

Most of the existing multi-camera people tracking frameworks [9, 10, 12, 13] incorporate state-of-the-art single camera online tracking methods [1, 28] and an offline global link model to conduct tracking across cameras and achieve robust and accurate tracking under long video sequence. The online tracking method either used motion [5, 14, 15, 25, 26] or appearance [1, 8, 16, 23], while the offline global link model often leverages appearance and spatio-temporal information to conduct association.

However, although it achieves superior performance with an online tracking method and offline global link model, the area of online multi-camera tracking method is still under exploration. Online multi-camera tracking shares the ability to handle video in an online or even real-time manner, which usually represents a lower computational cost and also the potential for real-world application, where sometimes processing the input sequence without access to future information is needed. For this reason, we propose a fully online approach for multi-camera people tracking, aiming to facilitate research in this area and provide a more practical solution. Furthermore, we also proposed a comprehensive evaluation framework designed to assess the robustness and reliability of our tracking approach across diverse tracking settings. This evaluation framework encompasses metrics for the 3D tracking accuracy under different lengths of video sequence.

Our contributions can be summarized as follows:

- A new metric, mHOTA, is introduced to address the shortcomings of existing evaluation methods, with the goal of establishing a fair benchmarking standard that accurately assesses online and offline tracking methods on extremely long sequences.
- We present an online and real-time multi-camera tracking framework leveraging camera calibration for spatial and temporal association, optimized for extended video sequences.
- The approach is validated on a multi-camera tracking dataset, achieving fifth place in the 2024 AI City Challenge Track 1 under 3DHOTA.

2. Related Work

2.1. Single-Camera Multi-Object Tracking

Since the introduction of deep learning technology, the single-camera multi-object tracking field has made significant progress, mainly adhering to the detection-tracking paradigm [2, 6, 18], which detects the location of the object to be tracked in each image frame extracted from a video, and connects the objects between frames based on different association clue. With the significant advances in object detection, tracking by detection has been widely adopted as the dominant paradigm in the MOT field. Early studies [5, 24] applied the Kalman filter to predict the position in the next frame and used the motion feature as a way to conduct data association. As a follow-up study, [24, 27] focused on improving the association accuracy by extracting appearance feature information and used as an association clue. Recently, a lot of research has been proposed to improve performance by incorporating different modules or global link models into existing algorithms and boosting the tracking performance. ByteTrack [29] proved that not only high-confidence detection results but also low-confidence detection results can contribute to better tracking performance. Several works [8, 15] proposed to enhance the performance with an extra tracklet-level association after online tracking. Despite recent advancements in online tracking algorithms, the predominant focus remains on single-camera tracking. However, the significance of tracking individuals across multiple cameras cannot be overlooked, as it serves as a crucial application. To address this need, we have developed an online multi-camera people tracking algorithm designed to advance research in this area.

2.2. Multi-Camera Multi-Object Tracking

Following single-camera tracking research, there have been efforts to adapt them to multi-camera applications. Most recent methods for multi-camera multi-object tracking systems solve the problem with two separate stages. Usually, a single-camera online tracking stage within each camera and a global association stage with a global link model conducting cross camera association based on spatio-temporal information and tracklet's feature similarity. While a two-stage approach is common in most multi-camera tracking frameworks, the single-camera tracking stage in recent works [11–13] usually does not consider using global information from different cameras to conduct association. Furthermore, some previous works [7, 13, 21] indicated that the ID switch problem might also occur during the single-camera tracking stage, which might further harm the final performance of the multi-camera tracking system. To address this issue, it is usually beneficial to conduct online tracking with the assistance of global information. In this work, we propose using an object's world coordinates to conduct online tracking, which utilizes each object's location and motion features under the world coordinate for association. This approach enables us to leverage information from multiple camera sources and maintain the tracking process online, even with multiple cameras as inputs. When multiple cameras serve as input sources, the tracking process becomes more robust against various common challenges in multi-object tracking, including occlusion and missing detections.

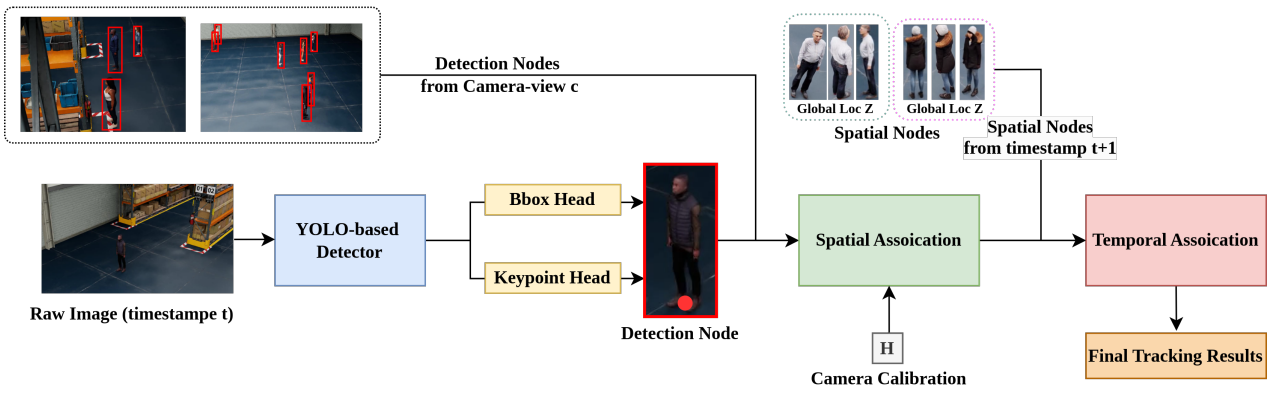

Figure 2. The pipeline of our online multi-camera people tracking framework. The extracted frames from all camera views are fed into the same **Detector** (Sec 3.1) to predict the detection nodes in image space. The footpoint locations are projected into world coordinates for the **Spatial Association** (Sec 3.2) to generate spatial nodes, which then undergo **Temporal Association** (Sec 3.3).

3. Method

3.1. Detection

In this section, we introduce an enhanced detection methodology leveraging the YOLO-based detector framework [17]. Specifically, we augment the detector architecture by incorporating an additional head dedicated to pose estimation, which is crucial for robust people tracking across multiple cameras. Let θ_D denote the detection head of the original YOLO-based detector, and θ_P represent the pose estimation head. Since the ground-truth world coordinates of each track are provided, we can utilize H_c, the ground-truth camera projection matrix, to reproject the corresponding foot point location in the image coordinates of each visible detection. We define the prediction bounding box and image coordinate pair as follows:

$$\mathbf{X}^{(c,t)} = \left\{ \left(x_0^{(c,t)}, k_0^{(c,t)}\right), \left(x_1^{(c,t)}, k_1^{(c,t)}\right), \cdots \right\} \quad (1)$$

where $x_i^{(c,t)} \in \mathbb{R}^5$ represents the bounding coordinate and confidence score under camera view $c \in \mathbf{C}$ at timestamp t, and $k_i^{(c,t)} \in \mathbb{R}^3$ is the corresponding keypoint prediction and confidence score in world coordinates for the i-th detection, which is projected from predicted image coordinates with H_c. The representation $\mathbf{X}$ is what we call **Detection Node** in the following content.

Our joint training objective is formulated as follows:

$$\mathcal{L}_{\text{total}}(\theta_D, \theta_P) = \lambda_{bbox} \cdot \left(\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{class}}\right) + \lambda_{pose} \cdot \mathcal{L}_{\text{pose}} \quad (2)$$

where $\mathcal{L}_{\text{loc}}$ denotes the bounding box regression loss, $\mathcal{L}_{\text{class}}$ represents the classification loss, and $\mathcal{L}_{\text{pose}}$ is the pose estimation loss. By jointly optimizing these components, our enhanced detector not only accurately localizes and classifies objects but also provides reliable pose estimates, thereby facilitating robust people tracking across diverse camera views in extremely long sequences.

Algorithm 1 Pseudo-code of Spatial Association

Require: Detection nodes, $\mathbf{X}^t$, Clustering threshold, T_c
Aspect ratio threshold, T_r
Initialize $\mathbf{Z}$ with detections
$\mathbf{Z}^t \leftarrow []$
for $x^{(c,t)}$ in $\mathbf{X}^t$ **do**
  if check_aspect$(x^{(c,t)}) > T_r$ **then**
    $\mathbf{Z}$.append$(\mathbf{z}(x^{(c,t)}, c))$
  end if
end for
# Merge all $\mathbf{z}$ not from the same view with distance $< T_c$
while nearest_distance$(\mathbf{Z}) < T_c$ **do**
  merge_nearest$(\mathbf{Z})$
end while

3.2. Spatial Association

To conduct an efficient and effective multi-view online MOT pipeline, we decouple the spatial and temporal association. For spatial association, as shown in Alg 1, we cluster the detection results, $\mathbf{X}^{(c,t)}$, across all the views based on their spatial locations, into **Spatial Nodes**, $\mathbf{Z}^t$,

$$\mathbf{Z}^t = \left\{ \mathbf{z}_0^t, \mathbf{z}_1^t, \cdots \right\} \quad (3)$$

where $\mathbf{z}_i^t$ represents a Spatial Node as it is a set of detection nodes across different cameras $c \in \mathbf{C}$ at timestamp t, clustered by world coordinates. All detection nodes, $\mathbf{X}^{(c,t)}$, are filtered by aspect ratio threshold, T_r, to remove potential noisy detection results.

It is important to highlight that the bounding box information at the image level is maintained in the spatial node, playing a crucial role in the tracking process. This data is essential for deriving aggregate world coordinates, facilitating the seamless merging of different camera perspectives. Moreover, it is valuable in post-processing, where it helps improve the accuracy of tracking results. Keeping this information ensures a more detailed understanding of the scene, which is critical for efficient decision-making in the tracking system.

3.3. Online Temporal Association

In the temporal association step, we employ standard tracking techniques under world coordinate systems, where the spatial node serves as the fundamental unit that we aim to associate across different timestamps.

Similar to multi-object tracking in traditional image space, the Kalman filter [4] can handle linear motion very well, especially since our tracking targets consist of humans who follow simple movement patterns around the environment in the Omniverse. The Kalman filter has four parameters, x, y, x', and y', denoting the position in the x and y directions and their respective velocities. These parameters are essential for modeling the state of a tracked object and updating its position and velocity estimates over time:

$$\dot{\mathbf{x}} = \left[x, y, x', y'\right]. \tag{4}$$

In addition to the Kalman filter, we incorporate standard association techniques to match detected objects across frames and cameras. This matching is based on a combination of spatial proximity and motion prediction. We employ a cost matrix that quantifies the likelihood of matches using $L1$ distance between existing tracks and new detections, optimizing the association problem using the Hungarian algorithm to minimize the overall cost.

3.4. ID Re-assignment

To improve the tracking performance, we further proposed an ID Re-assignment method aims to merge tracklets from the same identity after online tracking. After temporal association is finished, we can obtain multiple tracklet fragments from the whole sequence. However, several common tracking errors can happen, including 1. the same tracklet contained detections from multiple identities and were assigned with the same tracking ID, and 2. different tracklets contained the detections from the same identity but were assigned with different tracking IDs. These are usually caused by either the error during spatial and temporal association or the failure in ReID when the identity reappears in videos. To handle these error cases, we proposed an appearance-based ID Re-assignment method, which contains two stages including a tracklet splitting stage and a tracklet merging stage.

Tracklet Splitting. We first perform tracklet splitting to make tracklets into small fragments. For a tracklet with length N, the appearance feature extracted from a ReID model can be denoted as $F \in \mathbb{R}^{N\times D}$. To determine whether the tracklet contains features from multiple identities, we calculate the inner average pairwise cosine distance D_{inner} of all the features in each tracklet, which the average inner distance D_{inner} can be expressed as:

$$D_{inner} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{F_i \cdot F_j}{|F_i| \cdot |F_j|} \tag{5}$$

In this equation, F_i and F_j represent the appearance feature vectors of frames i and j in the tracklet, respectively. For those tracklets with D_{inner} bigger than threshold τ_{inner}, we further conduct hierarchical clustering to split the tracklets into more fragmented tracklets based on their appearance features. The number of clusters ranging from 2 to $k_{cluster}$, whenever the cluster results do not resulted in all the clustered tracklets having D_{inner} smaller than τ_{inner}, we increase the number of clusters by one and continue another round of clustering process.

Tracklet Merging. After tracklet splitting, we should obtain many fragmented tracklets that require further merging processing and assign the same fragment tracklets with a unified tracking ID. We conduct an average pooling of features for each tracklet and use hierarchical clustering to merge these tracklet fragments. Note that we do not merge any tracklets that have spatial and temporal overlap or larger distance than merging threshold τ_{merge}. We keep merging until no more tracklet pair can be further merged.

4. The mHOTA Evaluation Metrics

HOTA. Higher Order Tracking Accuracy [19] is designed to overcome limitations of previous CLEAR metrics [3] such as IDF1 and MOTA by considering higher-order associations and detection at the same time. A simple representation of HOTA in the multi-camera multi-object tracking setting is:

$$\text{HOTA}_\alpha = \frac{1}{N_C} \sum_{c\in C} \sqrt{DetA_\alpha \cdot AssA_\alpha} \tag{6}$$

where α represents the IoU threshold used to calculate DetA and AssA while N_C represents the number of views. As the final HOTA is computed by averaging over different thresholds.

3DHOTA. The organizer [20, 22] defines a modified version of HOTA called 3DHOTA that serves as a benchmarking metric for multi-camera multi-object tracking. Unlike the traditional HOTA, which averages the matching of predicted tracks with ground-truth tracks under each camera

Figure 3. Samples from the **AIC24 Multi-Camera People Tracking dataset**. This illustration showcases sequences across training, validation, and test sets, including warehouse, market, and hospital environments. The dataset is composed of synchronized videos from 8 to 16 distinct camera angles, each spanning over 13 minutes, highlighting challenges such as **non-overlapping regions**, **varying occlusion conditions**, and the need for **cross-domain generalization** in multi-camera tracking systems.

view, 3DHOTA is calculated directly in world coordinates, where tracks are represented as an x, y coordinate only. Instead of using Intersection over Union (IoU), they employ *L2* distance and a fixed threshold β to determine if the predictions qualify to match the ground truth. This approach restricts each track to predict only one world coordinate at each time t, ensuring consistency across all camera views.

However, a significant limitation of all existing evaluation metrics is their inability to adequately evaluate online methods in long video sequence datasets. This is because of the strategy used to determine the matching between predicted tracks and ground truth, which often favors brute force post-processing methods like setting maximum tracklet numbers.

mHOTA. To address this, we present mHOTA, which can serve as an evaluation metric for extremely long video sequences. Given a video sequence, we divide it into overlapping segments of length L_n where $L_n = \mathbf{n} \times L_{max}$ as $0 < \mathbf{n} \leq 1.0$ stands for length ratio and L_{max} stands for total length of the sequence s. The 3DHOTA@n can be defined as:

$$\text{3DHOTA@n} = \frac{1}{M} \sum_{i=0}^{M-1} \text{3DHOTA}(\text{s}_i^n) \tag{7}$$

$$\text{s}_i^n = s\left(\underbrace{\lfloor L_n/2 \rfloor \cdot i}_{start}, \underbrace{\lfloor L_n/2 \rfloor \cdot i + L_n}_{end} \right) \tag{8}$$

where we try to average the 3DHOTA values over a total of M overlapping sequences s_i^n. The $s(f_{start}, f_{end})$ stands for the slicing operation of the original s sequence.

This calculation ensures that 3DHOTA@n represents the average 3DHOTA score over the selected segments of length ratio n, taking into account the standard 3DHOTA calculation method across different thresholds.

Finally, we adapt the similar idea of mAP (mean average precision), defining the mHOTA as approximately the integral as summation that iterates over values of $\mathbf{n}$ from 0.05 to 1.00 in steps of 0.05:

$$\text{mHOTA} = \int_{0<n\leq 1} \text{3DHOTA@n} \tag{9}$$

$$\approx \frac{1}{19} \sum_{\substack{n=0.05 \\ n+=0.05}}^{1.00} \text{3DHOTA@n}. \tag{10}$$

5. Experiment

5.1. Datasets

The AIC24 Multi-camera People Tracking dataset, introduced by the AI City Challenge [22] this year, offers a comprehensive collection of data sourced from multiple camera feeds within a synthetic environment. These large-scale synthetic datasets were meticulously crafted using the NVIDIA Omniverse Platform, encompassing three distinct indoor scenarios. In total, the dataset comprises 90 scenes, each scene featuring approximately 16 cameras. Each camera provides synchronized high-resolution 1080p video feeds at 30 frames per second with a total length of around 10 minutes, enriched with detailed tracking annotations such as tracking IDs, bounding boxes, and world coordinates across camera views.

It's worth noting that the training and validation data only contain the warehouse scenario as in Figure 3. In contrast, the testing set presents a diverse array of scenarios including warehouse, hospital, and supermarket. This diversity introduces additional challenges in both detection and tracking tasks, stemming from the variations in object size, camera shooting angle, and environment settings across scenarios.

Sequence Name	@0.05			@0.2			@0.5			@0.05-1.00		
	HOTA	AssA	DetA	HOTA	AssA	DetA	HOTA	AssA	DetA	mHOTA	mAssA	mDetA
scene_041	77.28	69.21	86.52	54.57	34.64	86.45	39.39	18.12	86.09	42.22	22.67	86.23
scene_042	73.67	67.73	80.38	52.09	34.02	80.16	36.93	17.00	80.51	39.66	21.73	80.29
scene_043	80.32	75.43	85.71	58.34	39.94	85.66	40.51	19.25	85.41	44.88	25.77	85.57
scene_044	73.43	63.65	84.95	50.09	29.80	84.71	35.16	14.62	84.75	38.41	19.41	84.37
scene_045	70.50	60.14	82.91	45.73	25.31	82.91	31.64	12.10	82.83	35.72	17.25	82.97
Total	75.04	67.23	84.09	52.16	32.74	83.98	36.73	16.22	82.92	40.18	21.37	83.89

Table 1. The online result HOTA, AssA, DetA and its mHOTA on validation data of AIC24 dataset.

Type	HOTA@0.05	HOTA@0.2	HOTA@0.5	HOTA@1.0	mHOTA	mAssA	mDetA	mLocA
Online	75.04	52.16	36.73	27.79	40.18	21.37	83.89	95.97
Offline	69.92	60.59	57.03	56.58	58.07	49.48	68.81	95.30

Table 2. Comparison of the online and offline method of the tracking performance on validation data of AIC24 dataset.

Model	Type	Scene	HOTA	AssA	DetA	LocA
Baseline	*Online*	W	28.08	9.53	82.61	99.20
Baseline$^+$	*Offline*	W	79.41	78.34	80.51	99.25
Baseline	*Offline*	All	53.58	47.94	60.91	91.02
Baseline$^+$	*Offline*	All	**57.15**	**54.80**	**59.88**	**91.24**

Table 3. Overall tacking performance on test data of AIC24 dataset. Due to number of submission tries being limited, the first two rows of result (in gray color) are recompute based on the evaluation script provided by the organizer.

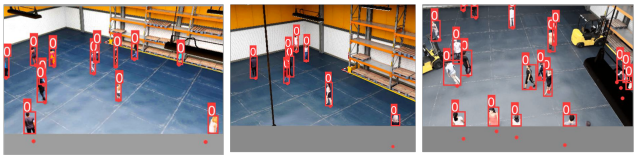

Figure 4. The visualization of the detection results on AIC24 testing dataset. The gray area are padding region to demonstratet the the keypoint predictions can land outside of the image.k

5.2. Implementation Details

Detector. With a frame rate of 30 FPS, the training and validation sets contain over 20 million high-resolution frames from 60 annotated sequences. To balance training time and detection performance, we train our YOLO-based detector at a sampling rate of 100 using all scenes in AIC24 and a sample rate of 250 using all hospital and market scenes in AIC23. The YOLOv8-x model was trained with an initial learning rate of 0.01 and a weight decay of 0.01 for 60 epochs. Keypoint annotations are in world coordinates, so we reproject them to image coordinates using the provided projection matrix. Note that some keypoints may extend beyond the bounding box or image due to occlusion. We retain these out-of-bound keypoints in the hope of improving both bounding box and keypoint predictions.

Spatial Association. We use hierarchical clustering for spatial association with clustering threshold, T_c, and aspect ratio threshold, T_r. More details and ablation studies are shown in Sec 3 and Table 4.

Temporal Association. We employ the default Kalman filter parameters from ByteTrack [28]. Considering the potential for long-term occlusion, we use a larger tracking buffer of 90 frames, which is equivalent to 3 seconds. The matching threshold is set to 2.5 (meter).

ID Re-assignment. For extracting the appearance feature, we trained an OSNet model [30] using the training procedure outlined in [13], employing ReID data sampled from this year's AI City Challenge training and validation sets [22]. The final sampled dataset comprises 41,757 training images, 20,919 testing images, and 21,210 query images. We set the tracklet re-assignment parameter k to 10, τ_{inner} to 0.3, and τ_{merge} to 0.15.

5.3. Experiment Results

Online Results. We implemented our online method on the first five scenes of the validation set from the 2024 AI City Challenge multi-camera people tracking dataset [22]. The tracking performance is shown in Table 2. As depicted in the table, tracking metrics related to association, including HOTA and AssA, drastically degrade as the length ratio of the sequence increases, indicating the drawback of online tracking under extremely long sequences. Under a length ratio of 0.05, online tracking maintains robust tracking accuracy with over 75% HOTA, while under a higher length ratio setting such as 0.5, the failure in long-term ReID leads to a significant drop in tracking performance. We also report the performance of our online method on the test set in Table 3.

Offline Results. To further improve performance, we incorporate offline tracklet merging as a post-processing method to boost the final performance. We compare the performance between the online and offline methods in Table

T_c	T_r	mHOTA	mAssA	mDetA	mLocA
1.0	0	34.09	16.51	79.92	95.00
2.5	0	38.05	19.57	82.75	94.06
1.0	1.4	38.78	20.16	83.56	95.83
2.5	1.4	40.18	21.37	83.89	95.97

Table 4. Online tracking performance of using different spatial association parameters on validation data of AIC24 dataset. The *default* setting stands for direction using $(x+0.5w, y+0.95h)$ of each box as the foot point coordinate in image space.

BBox	Keypoint	mHOTA	mAssA	mDetA	mLocA
prediction	*default*	32.53	15.76	76.15	90.50
prediction	*prediction*	40.18	21.37	83.89	95.97
GT	*default*	52.37	30.69	96.76	99.32
GT	*GT*	60.63	39.29	99.85	99.74

Table 5. Online tracking performance of using different bounding box and keypoint source on validation data of AIC24 dataset.

2. In a lower length ratio like 0.05, the online method achieves a higher HOTA compared to the offline method. This is mainly caused by incorrect merging in the offline post-processing, which might merge tracklets from different identities incorrectly, resulting in the removal of some parts of overlapped tracklets (with the same tracking ID under the same frame) during the evaluation stage. However, with a higher length ratio like 0.5 or 1.0, the offline method achieves better performance. This is attributed to the ability of offline post-processing to conduct long-term ReID. The performance of the offline method is reported in 3.

5.4. Ablation Studies

Different Spatial and Aspect Ratio Thresholds. As shown in Table 4, several different combinations of spatial and aspect ratio thresholds are tested. We achieve the best performance with $T_c = 2.5$ and $T_r = 1.4$. T_r is set to remove the noisy or partially detected human bounding boxes. The reason why larger T_c does not lead to clustering wrong instances is that we make sure in each spatial node, $\mathbf{z}_i$, all detection nodes are from different cameras.

Effectiveness of Keypoint Estimation. To evaluate the effectiveness of keypoint estimation, we compared the mHOTA from different sources of detected bounding boxes and spatial keypoint estimation. For bounding boxes, we try to utilize predicted bounding boxes from Yolov8 and ground truth bounding boxes. On the other hand, for keypoint estimation, we try predicted and default keypoints from the bottom of the bounding boxes. As shown in Table 5, with the help of more accurate estimated keypoints or ground truth keypoints, the mHOTA improves 7.65% with predicted bounding boxes and 8.26% with ground truth bounding boxes.

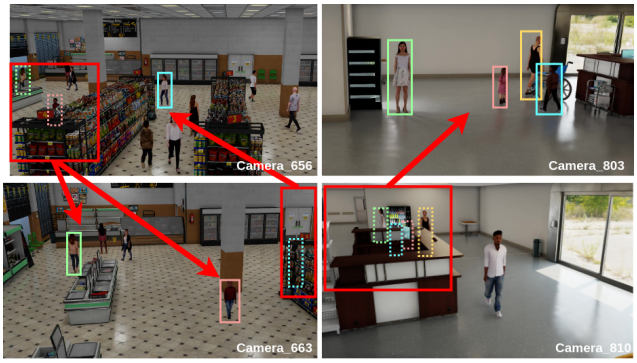

Figure 5. Example of the impact of occlusion on tracking in a multi-camera setup, where selective masking (region in red) and view optimization can enhance keypoint prediction and spatial association.

6. Discussion and Limitations

Exploring Online Re-Identification. In Tables 1, 2, and 3, it is evident that the performance degradation of the online method intensifies with the increase in frame count. Consequently, offline re-identification techniques, such as track ID reassignment, splitting, and merging, coupled with post-processing methods like interpolation, can surpass the online method in performance. Nevertheless, online tracking holds significant importance in real-world multi-camera multi-object tracking applications due to its due to its real-time processing capabilities and ability to provide immediate situational awareness, which is critical for dynamic decision-making and timely response in various scenarios.

A potential strategy to mitigate this issue is to execute offline tracking every n_k frame, allowing for accurate reconstruction of the track's appearance features (from different cameras), motion states, and historical trajectory. This approach could enhance future frame association, especially if temporal association involves matching these elements. However, this may result in a trade-off, as the method might not support real-time implementation.

Keypoint Reliability. The method proposed in this paper relies heavily on the accuracy of keypoint prediction. With supervised learning, it performs exceptionally well in the *warehouse* scene during testing, as indicated in Table 3, with a nearly perfect Localization Accuracy (LocA) highlighting the success of the spatial association step. However, when this method is adapted to unfamiliar settings such as *market* and *hospital*, which feature different backgrounds, camera angles, and distances to the target, there is a noticeable and significant drop in performance. To overcome this, one possible approach is to employ unsupervised domain adaptation techniques for detector training, aiming to reduce the domain disparity in such scenarios.

Camera-View Selection. In our experiments, we observed

that not all cameras should be treated equally. For instance, as shown in Figure 5, some tracks are significantly occluded by the aisle, leading to substantial errors in keypoint prediction in the image coordinates. These errors tend to amplify when reprojected back to world coordinates due to the increased distance from the camera. However, these same areas might be clearly captured by another camera. By selectively masking out the bounding boxes and keypoint predictions in the affected area, we often achieve improved results in spatial association. Determining which camera views to retain or which areas to mask out presents an intriguing challenge, with the potential for resolution through learning-based methods.

7. Conclusion

We introduce an online and real-time multi-camera people tracking framework that utilizes established camera calibration for spatial association in world coordinates, followed by temporal association. Additionally, we address the current shortcomings in metrics and the absence of adequate evaluation methods for online tracking in lengthy sequences. To address this gap, we propose the mHOTA metric, which offers a more comprehensive evaluation approach and hopes to establish it as a standard benchmark for assessing the effectiveness of online tracking methods in extended video sequences. Our method is thoroughly evaluated on a multi-camera people tracking dataset across different scenarios. Our proposed approach, along with post-processing, achieved a fifth-place ranking on the public test set of the 2024 AI City Challenge Track 1 in terms of 3DHOTA.

Acknowledgement

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean Government (Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry) under Grant 24ZD1120. We also want to acknowledge and thank National Center for High-performance Computing from Taiwan for providing the computing resources.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 2

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 4

[4] 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. 4

[5] 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. IEEE, 2016. 2

[6] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1515–1522. IEEE, 2009. 2

[7] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalized global graph model-based approach for multicamera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, 2016. 2

[8] 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

[9] Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In *CVPR workshops*, pages 416–424, 2019. 2

[10] Hung-Min Hsu, Yizhou Wang, and Jenq-Neng Hwang. Traffic-aware multi-camera tracking of vehicles based on reid and camera link model. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 964–972, 2020. 2

[11] Hung-Min Hsu, Yizhou Wang, Jiarui Cai, and Jenq-Neng Hwang. Multi-target multi-camera tracking of vehicles by graph auto-encoder and self-supervised camera link model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 489–499, 2022. 2

[12] Hsiang-Wei Huang, Cheng-Yen Yang, and Jenq-Neng Hwang. Multi-target multi-camera vehicle tracking using transformer-based camera link model and spatial-temporal information. *arXiv preprint arXiv:2301.07805*, 2023. 2

[13] 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, 6

[14] Hsiang-Wei Huang, Cheng-Yen Yang, Samartha Ramkumar, Chung-I Huang, Jenq-Neng Hwang, Pyong-Kun Kim, Kyoungoh Lee, and Kwangju Kim. Observation centric and central distance recovery for athlete tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 454–460, 2023. 2

[15] Hsiang-Wei Huang, Cheng-Yen Yang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Exploring learning-based motion models in multi-object tracking. *arXiv preprint arXiv:2403.10826*, 2024. 2

[16] Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. Iterative scale-up expansioniou and deep features association for multi-object tracking in sports. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 163–172, 2024. 2

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, 2023. 3

[18] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(7): 1409–1422, 2011. 2

[19] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian ´ Leibe. Hota: A higher order metric for evaluating multi-object tracking, 2021. International journal of computer vision, 129(2):548–578. 4

[20] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. 4

[21] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 343–356. Springer, 2012. 2

[22] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 4, 5, 6

[23] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. In 2017 IEEE international conference on image processing(ICIP), pages 3645–3649. IEEE. 2

[24] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[25] Cheng-Yen Yang, Alan Yu Shyang Tan, Melanie J. Underwood, Charlotte Bodie, Zhongyu Jiang, Steve George, Karl Warr, Jenq-Neng Hwang, and Emma Jones. Multi-object tracking by iteratively associating detections with uniform appearance for trawl-based fishing bycatch monitoring, 2023. 2

[26] Cheng-Yen Yang, Hsiang-Wei Huang, Zhongyu Jiang, Heng-Cheng Kuo, Jie Mei, Chung-I Huang, and Jenq-Neng Hwang. Sea you later: Metadata-guided long-term re-identification for uav-based multi-object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 805–812, 2024. 2

[27] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 36–42. Springer, 2016. 2

[28] 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 *European conference on computer vision*, pages 1–21. Springer, 2022. 2, 6

[29] 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. 2022. 2

[30] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification, 2019. Proceedings of the IEEE/CVF International Conference on Computer Vision. 6