A Robust Online Multi-Camera People Tracking System With Geometric Consistency and State-aware Re-ID Correction

Zhenyu Xie[1] Zelin Ni[1] Wenjie Yang[1] Yuang Zhang[1]
Yihang Chen[1,3] Yang Zhang[2] Xiao Ma[2]
[1]Shanghai Jiao Tong University [2]AI Lab, Lenovo Research [3]Monash University
{sp.sat,nzl5116190,13633491388,zyayoung,yhchen.ee}@sjtu.edu.cn
{zhangyang20,maxiao3}@lenovo.com

Abstract

Multi-camera multiple people tracking is a crucial technology for surveillance, crowd management, and social behavior analysis, enabling large-scale monitoring and comprehensive understanding of complex scenarios involving multiple individuals across different camera views. However, due to severe occlusion within the scene and significant variations in camera viewpoints, there are high demands for matching and correlating the same target among different cameras, especially in an online setting. To address this challenge, we propose a novel online multi-camera multiple people tracking system. This system integrates geometric-consistent constraints and appearance information of the targets, effectively improving tracking accuracy. Additionally, we design a state-aware Re-ID correction mechanism that adaptively leverages Re-ID features to correct mismatches among targets. This system has demonstrated good adaptability across various scenarios. Our proposed system is evaluated in track1 of the 2024 AI City Challenge [38], achieving a HOTA score of 67.2175% and securing the 2nd position on the leaderboard. The code will be available at: https://github.com/ZhenyuX1E/PoseTrack.

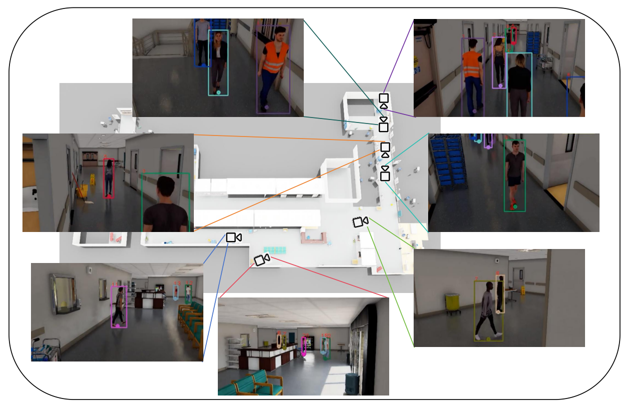


Figure 1. Illustration of the multi-camera multiple people tracking task. This task requires detecting and tracking each individual in occluded scenes, and assigning the same ID to the same object across different cameras.

1. Introduction

Multi-camera people tracking (MCPT) is a vital research area in computer vision aimed at accurately monitoring and tracing individuals across various camera views. By seamlessly integrating data from multiple cameras, MCPT transcends the limitations of traditional single-camera tracking methods, offering unparalleled accuracy and robustness. A traditional online MCPT pipeline encompasses critical stages such as pedestrian detection, re-identification (Re-ID) for extracting distinctive features, multi-camera matching for associating detection results between different cameras, and ID initialization or update for storing the trajectories up to the current moment.

This advanced technology finds applications across various sectors, with a growing demand for indoor people tracking in recent years. In storage, it optimizes inventory management and workflow efficiency. Supermarkets benefit from detailed customer behavior analysis for targeted marketing and improved security. In hospitals, MCPT enhances patient flow management and resource allocation, contributing to better care standards. With its ability to provide actionable insights and enhance surveillance capabilities, multi-camera multiple people tracking emerges as an indispensable tool across various domains, driving innovation and efficiency in today's interconnected world.

In recent years, a series of efforts have been made to enhance the MCPT task, leading to significant improvements in its applicability across various scenarios, as well as its robustness and accuracy. However, challenges such as occlusions in dense scenes and variations of the same individual across different camera perspectives pose higher demands on algorithms.

To solve these issues, we design an online MCPT system that integrates spatial information within and across cameras and state-aware appearance information of the targets. Specifically, when matching multi-view detections to tracked targets, we simultaneously consider 2D spatial affinity, 3D epipolar affinity, homography affinity and adaptive state-aware Re-ID affinity. Among these, the first three are designed to meet the geometric constraints within single-view and across different views. The latter helps to correct the problem of of ID switches during and after heavy occlusions. Furthermore, to avoid multiple fragments resulted from significant Re-ID differences for the same individual across different viewpoints, we specially design a Re-ID feature bank to store diverse Re-ID features corresponding to different poses and angles, enabling our system with a powerful online re-identification ability.

Our contributions can be summarized as follows:

- We design a robust online MCPT system. By incorporating geometric-consistent constraints and appearance information, effective multi-camera multi people tracking can be achieved in different scenarios.
- We propose a state-aware Re-ID correction mechanism to address the problem of ID switches during and after heavy occlusions, with a special-designed feature bank containing diverse Re-ID features corresponding to different poses and angles.
- We evaluated our system in track1 of the 2024 AI City Challenge that consists of many densely occluded scenes and achieved the second place in the leaderboard with a HOTA score of 67.2175.

The paper is organized as follows: In Section 2, we offer a comprehensive review of existing literature. Section 3 elucidates the methodology proposed in this paper. Following that, Section 4 delves into the intricate details of our implementation and the ensuing experimental outcomes. Finally, in Section 5, we engage in a discussion of the findings and draw conclusions from our research endeavor.

2. Related Works

2.1. Object Detection

Object detection is a crucial task in computer vision, aiming to accurately localize and classify objects in images and videos. In recent years, various approaches have been developed, categorized into anchor-based and anchor-free detection methods.

Anchor-based methods utilize predefined anchor boxes to predict object positions and categories. One prominent approach is Faster R-CNN [24], which introduces a region proposal network (RPN) to efficiently generate region of interest (ROI) proposals. This method significantly improves detection efficiency by separating region proposal generation from the subsequent classification and localization tasks. Another notable advancement is Mask R-CNN [9], an extension of Faster R-CNN that incorporates a parallel mask prediction head for instance segmentation.

In contrast, anchor-free detection methods do not rely on predefined anchor boxes. Instead, they directly predict object locations and categories without anchor-based constraints. You Only Look Once (YOLO) [23] and its follow-up work including YOLOv3 [22], YOLOv7 [35] and YOLOX [8] are pioneering anchor-free works which adopt a single-stage detection approach without anchor boxes, simplifying the detection pipeline while maintaining high precision. Transformer-based object detection methods such as DETR [2], Swin Transformer [17] and Vit-Det [14] are another anchor-free variants that utilize attention mechanisms to more accurately locate objects. Anchor-free methods bring faster running speed and can be applied to some real-time scenarios.

2.2. Re-Identification

Re-Identification (Re-ID) plays an important role in recognizing individuals across different scenes, cameras, or time instances. Advancements in deep learning have greatly improved Re-ID performance, with convolutional neural networks (CNNs) being commonly used to extract robust and discriminative features. During this time, three directions receive major attention, namely feature representation learning, metric learning and ranking optimization. To capture representative features, researchers explore global feature [36, 47], local feature [32, 42, 46], auxiliary feature with additional annotated information [29] or more augmented training samples [49] to construct feature vectors. In metric learning, researchers focus on the loss function used to guide the feature learning process, including identity loss [31, 49], verification loss [3, 48] and triplet loss [28, 39]. In the meanwhile, to improve the performance in the testing stage, the ranking order is optimized by similarity mining [43, 50], human interaction [16] and metric fusion [20].

2.3. Multi-Object Tracking

Multi-Object tracking typically adopts the track-by-detection paradigm where objects are first detected in each frame, and then linked across frames to form tracks. In the early stage, [1] is a representative work that utilizes Kalman filters to forecast the object location in the next frame and solves a assigning problem using Hungarian algorithm. Although SORT is simple and effective, it performs poorly in situations of occlusion and the disappearance of objects. Afterwards, various approaches are proposed to improve the performance in object association. [41] utilized additional appearance information to match objects. [34] proposed pixel-level tracking to achieve higher tracking accuracy. [44] associated every detection box to reduce missed detections and improve trajectory continuity.

2.4. Multi-camera people tracking

Multi-camera people tracking (MCPT) is a complex field within computer vision that focuses on identifying multiple individuals across different camera feeds. This task can be divided into offline and online tracking approaches, each with its distinct methodologies and challenges.

Offline tracking approach allows the system to access all the video frames and camera views at once, enabling comprehensive analysis and optimization. In previous works [6, 10, 40], graph-based methods are used to associate multiple image flows across cameras. Later, to facilitate intra-camera association, re-identification features are combined into models in [15, 19, 26, 33]. Furthermore, Some recent works [5, 11, 12, 45] adopt 3D pose estimation and camera parameters to acquire 3D human joints, which can be used as spatial information in the association step.

Different from offline tracking approach, online approach needs to form tracking trajectories using only the information available at the present time. [21] proposes a dynamic graph Model with link prediction to facilitate data association. [5] starts from the perspective of 3D pose estimation and iteratively updates 3D pose for each person in the process of multi-camera multi-people tracking. It is worth noting that online approach heavily rely on the performance of Re-ID models especially in the occlusion scenes.

3. Methods

In this section, we will first give an overview of our online tracking framework. Then the two main components of our framework will be detailed, i.e. multi-view matching with geometric consistency and state-aware Re-ID affinity and cross-view target initialization.

3.1. Overview

Generally, We store latest updated information of individuals as tracked targets for the ease of online tracking, and formulate the online tracking task as a problem associating newly detected human to tracked targets.

Specifically, for the i-th detection $D_{i,t,c}$ from the t-th frame of camera c, we estimate its 2D human body keypoints $\{x^k_{i,t,c}\}$ and its Re-ID feature $f_{i,t,c}$, where $k \in \{1, \cdots, N_K\}$, N_K the number of 2D human body keypoints and $f_{i,t,c} \in \mathbb{R}^{N_R}$ with N_R the length of feature vector. Thus, a detected human sample $D_{i,t,c}$ can be represented as $D_{i,t,c} = \{B_{i,t,c}, \{x^k_{i,t,c}\}, f_{i,t,c}\}$, where $B_{i,t,c} \in \mathbb{R}^4$ its bounding box and $x^k_{i,t,c} \in \mathbb{R}^2$ the 2D location of the k-th keypoint. In terms of target, we retain its last update 3D keypoints location $\{X^k_{i,t}\}$ in global coordinate, 2D keypoints $\{x^k_{i,t',c}\}$ and bounding boxes from each view $\{B_{i,t',c}\}$, where t' the last update time of the 2D keypoint or of the bounding box. Besides, a feature bank F_i containing differentiated Re-ID feature vectors of an single person is also maintained for each target. Therefore, a target at the t-th frame $T_{i,t}$ can be represented as a combination of the stored information, i.e. $T_{i,t} = \{\{X^k_{i,t}\}, \{x^k_{i,t',c}\}, \{B_{i,t',c}\}, F_{i,t}\}$. In the following paragraphs, we omit the index i of $D_{i,t,c}$ and $T_{i,t}$ for simplicity.

We also design 4 tracking states for targets: unconfirmed, confirmed, missing or deleted. Unconfirmed targets are initialized with only single-view detection and needs to be matched relentlessly for a certain number of frames to be transformed to a confirmed tracking state. If the target does not receive consecutive matches during this period, the target will be deleted to reduce the influence of false positive detection. Confirmed targets are targets being currently tracked, and can transfer to the missing state if the target does not receive any matches within a period of time.

3.2. Multi-view Matching with Geometric Consistency and State-aware Re-ID Affinity

Given frames from multi-view cameras, we iteratively associate detection samples to tracked targets view-by-view. The association problem can be equivalent to a weighted bipartite graph matching problem, with a cost matrix reflecting the weight of edges. The matching problem can be solved by Hungarian algorithm [13] view-by view once the cost matrix is determined. In our framework, we use an affinity matrix instead of a cost matrix, where the former can be considered as the negative of the latter. Therefore, how to modeling the affinity between detection samples and targets is a crucial component in multi-view matching.

In our framework, we specially design a combination of geometric consistency affinity and adaptive state-aware Re-ID affinity, which demonstrates a strong performance in scenes with crowd and heavy occlusions.

3.2.1 Geometric Consistency Affinity

In order to simultaneously consider matching detection samples and targets within single view as well as across different views, we construct the geometric consistency affinity matrix by incorporating single-view 2D spatial information and cross-view epipolar distance and homography distance.

2D Spatial Affinity In order to maintain the continuous consistency of multi-view tracking results within a single view, we follow previous methods[1, 44] by considering the position of the bounding box as one of the crucial criteria for tracking. We employ a Kalman Filter to predict bounding box locations in the next frame. Given a pair of target

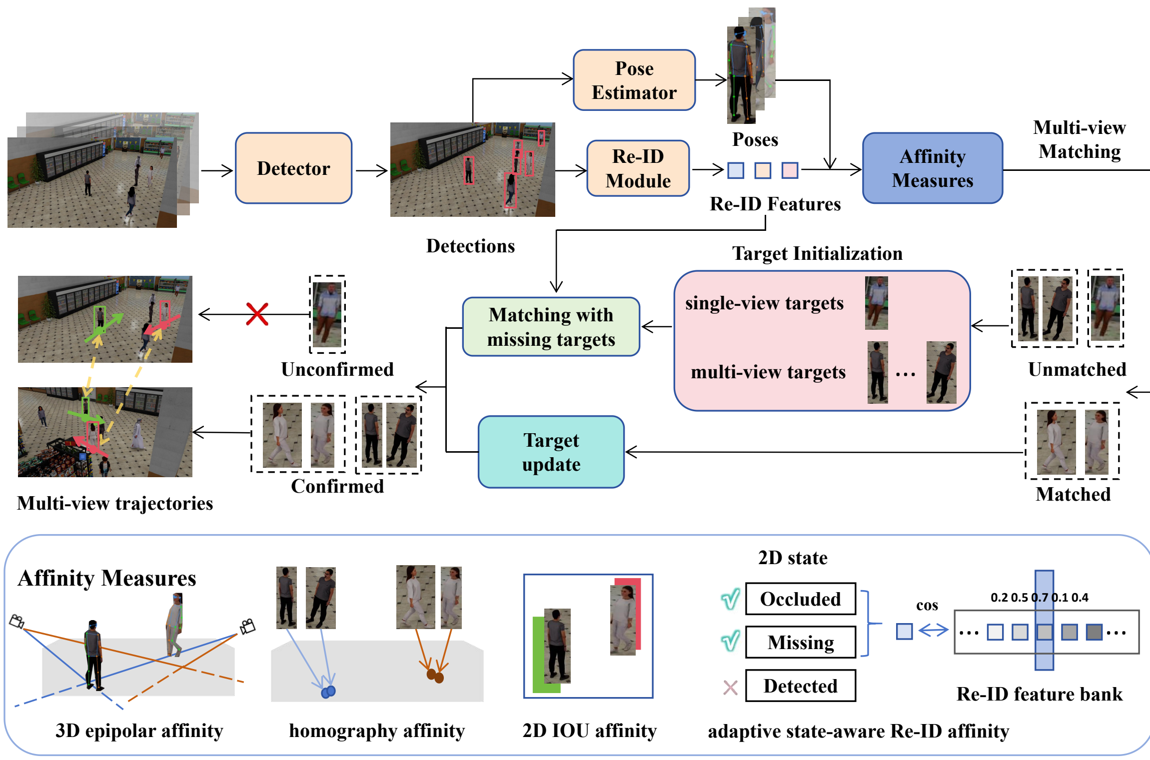


Figure 2. Illustration of the proposed online multi-camera people tracking system. Our system first performs pedestrian detection in each view. The detected bounding boxes are fed into a pose estimator and a Re-ID module to extract 2D keypoints and Re-ID features, which offers geometric and appearance information for multiple forms of affinity measures. Multi-view matching is then performed to associate multi-view detections with tracked targets based on the established affinity. The matched detections will update the state of the target, while the unmatched ones will initialize new targets under certain criteria. An additional missing targets matching procudure is implemented to re-identify re-emerged targets. Finally, the global coordinate of confirmed targets will be output.

and detection sample $(D_{t,c}, T_{t'})$, the Affinity between the detected bounding box and the predicted bounding box is defined using IOU metrics,

$$A_b\left(D_{t,c}, T_{t'}\right) = w_b(IOU(B_{D_{t,c}}, B^{KF}_{T_{t'},c}) - \alpha_b), \qquad (1)$$

where $B_{T_{t'},c}$ is the predicted bounding box by Kalman Filter from camera c of the target $T_{t'}$, N_c the number of cameras and α_b the threshold of 2D spatial affinity. Here we omit the i index of $D_{t,c}$ and $T_{t'}$ for simplicity. If there is no matched bounding box within a period of time, the affinity from camera c will be set to zero.

3D Epipolar Affinity To associate detection samples from different views to targets, we also introduce 3D epipolar distance into our affinity measurement. We back-project the detected 2D keypoint $x^k_{t,c}$ as a ray in 3D global coordinates,

$$\tilde{X}^k_t(\mu; x^k_{t,c}) = P^\dagger_c \tilde{x}^k_{t,c} + \mu \tilde{X}_c, \qquad (2)$$

where $\tilde{x}^k_{t,c}$ the homogeneous coordinate of $x^k_{t,c}$, $P^\dagger_c \in \mathbb{R}^{4\times 3}$ the pseudo-inverse of the camera projection matrix P_c. $\tilde{X}_c$ is the homogeneous coordinate of the camera center in global coordinates, which can be inferred using the following formula:

$$P_c \cdot \tilde{X}_c = \mathbf{0} \qquad (3)$$

The same operation is also performed to compute the ray $\tilde{X}^k_{t'}$ back-projected from the last updated 2D keypoint $x^k_{t',c'}$ of the target. The 3D epipolar affinity is defined as:

$$A_{epi}(D_{t,c}, T_{t'}) = \sum_{c' \neq c} \sum_{k=1}^{N_K} A_{epi}(x^k_{t,c}, x^k_{t',c'}), \qquad (4)$$

$$A_{epi}(x^k_{t,c}, x^k_{t',c'}) = w_{epi}\left(1 - \frac{d_l(X_t(\mu), X_{t'}(\mu))}{\alpha_{epi}}\right), \qquad (5)$$

where d_l denotes the line-to-line distance in 3D space and α_{epi} the threshold.

Homography Affinity Although the epipolar distance offers important information while associating individuals from different cameras, the distance can be inaccurate when two rays are nearly intersecting but the intersection point does not locate on any individual. Therefore, we further introduce homography distance to leverage global spatial information rather than intra-camera information. The function F_{bp} outputs the "bottom point" of a detection sample or a target from a certain camera. The "bottom point" represents the average position of the left and right ankle keypoints x_{la} and x_{ra} when their estimated confidence score is above the keypoint threshold α_{kp}, otherwise the midpoint of bottom line of the bounding box will be applied.

$$F_{bp}(D_{t,c}) = \begin{cases} \frac{x_{la}+x_{ra}}{2} & \text{if } c_{la}, c_{ra} > \alpha_{kp} \\ (\frac{x_1+x_2}{2}, y_2) & otherwise \end{cases}, \tag{6}$$

where (x_1, y_2) and (x_2, y_2) denote the coordinates of top-left vertex and bottom-right vertex of bounding box, respectively. Then, the "bottom points" can be reprojected into global coordinates when given the homography matrix $H_c \in \mathbb{R}^{3\times3}$of the ground plane,

$$F_g(D_{t,c}) = H_c \cdot F_{bp}(D_{t,c}). \tag{7}$$

We hereby perform homography transform on the bottom points of each detection sample and of each target from different views. The euclidean distance is calculated between the transformed points to measure the homography affinity,

$$A_h(D_{t,c}, T_{t'}) = \sum_{c' \neq c} w_h D_h(D_{t,c}, T_{t',c'}), \tag{8}$$

$$D_h(D_{t,c}, T_{t',c'}) = 1 - \frac{\|F_g(D_{t,c}), F_g(T_{t',c'})\|}{\alpha_h}. \tag{9}$$

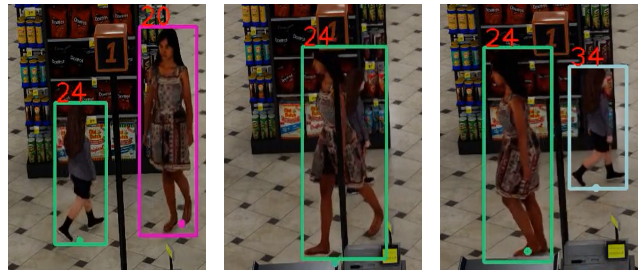
The overall geometric consistency affinity can be represented as a sum of the above-mentioned components,

$$A_{gc} = A_b + A_{epi} + A_h. \tag{10}$$

3.2.2 Adaptive State-aware Re-ID Affinity

Considering the scenario where two or multiple pedestrians are heavily occluded, the non-maximum suppression (NMS) might result in only one bounding box being detected, leading to potential ID switches among multiple pedestrians, which can be highly detrimental to data association. Therefore, we propose an adaptive state-aware Re-ID Affinity to incorporate Re-ID features as an auxiliary measure reducing ID switches during and after occlusions.

Here, we define a state variable called 2D state, which can be assigned with three different states: detected, occluded, or missing. The term "occluded" represents scenarios where the target suffers occlusion with other targets



(a) Before Occlusion (b) Occlusion (c) After Occlusion

Figure 3. Illustration of ID switch due to severe occlusion. Notably, the target initially tagged with ID 20 (purple) is erroneously matched to ID 24 (green) after occlusion. Besides, the target originally bearing ID 24 (green) is reinitialized to ID 34 (blue).

beyond a certain threshold. "Missing" indicates that a previously seen target has disappeared from the current camera view, while "detected" represents all other situations. Unlike the "tracking state" mentioned earlier, which represents the overall status of a target across all cameras, 2D state is a status measure recorded for each camera. For example, a target's tracking state is only marked as missing if it disappears from all views. In contrast, its 2D state will be marked as missing as soon as it disappears from the view of any camera.

For targets with at least one "occluded" or "missing" 2D state, we add a Re-ID-related metric as affinity. Specifically, for all detection samples and the above-mentioned tracked targets, we first calculate the Re-ID similarity between each pair, which is represented by the maximum value of the cosine similarities:

$$S(f_{D_{t,c}}, F_{t'}) = \max(\cos(f_{D_{t,c}}, F_{t'})), \tag{11}$$

where $f_{D_{t,c}}$ denotes the Re-ID feature vector from the detection sample $D_{t,c}$ and $F_{t'}$ represents the Re-ID feature bank of the tracked target $T_{t'}$, whose updating mechanism will be explained in 3.2.3. Then, the adaptive state-aware Re-ID affinity can be defined as

$$A_s = w_s(S(f_{D_{t,c}}, F_{t'}) - \alpha_s), \tag{12}$$

where w_s and α_s are the weight and the threshold of adaptive state-aware Re-ID affinity respectively. w_s is assigned with a strictly positive value only when the target has at least one "occluded" or "missing" 2D state, otherwise it will be assigned with a zero value.

As a result, the final affinity used in tracking can be a combination of geometric consistency and state-aware Re-ID affinity, i.e. $A_{final} = A_{gc} + A_s$.

3.2.3 Target Update

The information stored in targets will be updated once the matching results are determined based on the affinity

Figure 4. Examples of low intra-class similarity and high inter-class similarity. The blue box indicates low Re-ID similarity for the same target under different perspectives, and the red box represents high similarity for different targets due to similar clothing.

measures. The updating procedure can be primarily divided into single-view updates and multi-view joint updates. For single-view (or per-view) updates, the matched detection box and corresponding 2D keypoints are stored, and the Kalman filter is also updated based on the newly matched bounding box location. Multi-view joint updates primarily involve 3D keypoints update, correction with 3D keypoints, and feature bank update, which will be detailed in the following paragraphs.

3D keypoints update The update of 3D keypoints involves, for each trajectory, first iterating through all 2D keypoints in the visible views and selecting those above a certain threshold as reliable keypoints. Then, for all reliable keypoints, triangulation is performed to obain the 3D coordinates of the same keypoint by using the projection matrices of all valid views. the final output will be chosen from 3D keypoints, world coordinates after homography transform using 2D keypoints or using the mid point of bottom line of bounding box, following the exact priority order.

Feature Bank Update For persons re-entering the scene, we leverage Re-ID features to re-identify these targets. However, even with the current state-of-the-art Re-ID model [7], there are still difficulties when identifying the same person with various poses under different perspectives and different people with similar clothes, as illustrated in Figure 4. As a result, simply adopting a naive box-to-box Re-ID similarity calculation makes it difficult to identify a newly entering target.

To address the above problem, we design a Re-ID feature bank mechanism, where we retain diverse Re-ID features of each tracked target, hoping to correctly assign IDs under different angles. Our key insight is to ensure a Re-ID feature corresponding to a similar pose to the current detection is retained in the feature bank. One the one hand, Such design avoids the problem of mismatched different perspectives. On the other hand, the Re-ID similarity of the same target under close perspective will be higher than that of people wearing similar clothing, reducing the occurrence of incorrect allocation of IDs.

Algorithm 1 Re-ID Feature Bank Update

```
Initialize a fixed-size queue F as Re-ID feature bank
Define thresholds: α_k (keypoints), α_c (bbox confidence), α_b (IOU), α_cr (coverage rate), α_s (similarity)
for each bounding box B in tracking process do
    Let c_k be the confidence of upper body keypoints
    Let c_B be the confidence of bounding box B
    Let f_B be the Re-ID feature of bounding box B
    Let IOU(B, B'_t) be the intersection over union of B with a bounding box B'_t from tracked targets
    Let CR(B, B'_d) be the coverage rate of B with a detected bounding box B'_d
    Let S(f_B, f) be the cosine similarity between f_B and f
    if k > α_k and c > α_c and sum(IOU(B, B'_t) < α_b, ∀B'_t) < 2 and sum(CR(B, B'_d) < α_cr, ∀B'_d) < 2 then
        if ∀f ∈ F, S(f_B, f) < α_s then
            if size of Q is full then
                Dequeue the first element from Q
            end if
            Enqueue B to Q
        end if
    end if
end for
```

Specifically, we established a Re-ID feature queue with a fixed size. During the subsequent tracking process, if the detection sample of the same target under all viewpoints meets certain conditions, the corresponding Re-ID feature vector will be enqueued. The criteria requires that the detected person is easily recognizable and minimally occluded, demanding 1) the upper body to be visible, 2) the detection box's confidence level to exceed a certain threshold, and 3) the overlap with other boxes to be less than a certain threshold.

It is worth noting that in occlusion scenarios, the metric IOU cannot well reflect the occlusion degree of a bounding box, especially when small targets are occluded by large targets. Thus, we introduce coverage rate(CR) to help determine whether a detection sample is heavily occluded:

$$CR(B_{D_1}, B_{D_2}) = \frac{\text{Area}(B_{D_1} \cap B_{D_2})}{\text{Area}(B_{D_1})}, \tag{13}$$

where B_{D_1} and B_{D_2} represent the occluded bounding box and the occluding bounding box respectively.

Moreover, to retain discriminate Re-ID features, we require that the newly added Re-ID feature has a low maximum similarity with the existing features in the bank. Based on this, we constructed a Re-ID feature bank for each target. This mechanism is applied in subsequent missing track ID matching and the construction of adaptive state-aware Re-ID affinity. In the following steps, when calculating Re-ID similarity, we use the maximum similarity between the current Re-ID feature and the feature bank as the result.

3.3. Cross-view Target Initialization

For unmatched detections left from multi-view matching, we also try to associate them across different cameras. Similar to Equation 5 and Equation 8, we calculate the 3D epipolar affinity A_{epi} and homography affinity A_h, but between each pair of detection samples from different cameras, rather than between a detection sample and a target. The association affinity A_{ass} is defined as

$$A_{ass} = A_{epi} + A_h. \tag{14}$$

Inspired from [4, 5, 25], we formulate the association problem by a weighted graph partitioning problem with cycle-consistency. The problem can be solved via binary integer programming:

$$e^* = \underset{e}{argmax} \sum_{i,j} A_{ass}(i,j) e(i,j), \tag{15}$$

$$s.t. \quad e(i,j) \in 0, 1, \tag{16}$$

$$e(i,j) + e(j,k) <= 1 + e(i,k), \tag{17}$$

where $A_{ass}(i,j)$ denotes the association affinity between D_i and D_j, and $e(i,j)$ a binary variable indicating the existence of connection between D_i and D_j. The final results can be represented by a group of clusters, and we adopt different initialization strategies according to the size of clusters.

Single-view initialization When only one view is available for initiating a target, single-view initialization is performed. As the target only appears in one view, it is possible that it comes from a false positive detection. Thus we choose to set its tracking state as unconfirmed and needs several consecutive updates to transform into a confirmed target. Single-view initialization primarily initializes 2D information, including bounding boxes and 2D keypoints. Additionally, the corresponding Re-ID feature can also be involved in the Re-ID feature bank update.

Multi-view initialization When a target simultaneously appears in multiple views, multi-view initialization is performed. The tracking state for multi-view initialization is set directly to Confirmed. Compared to single-view initialization, multi-view initialization includes the initialization of 3D information. Specifically, in addition to initializing 2D information for each view similar to single-view initialization, 3D keypoints are also calculated and recorded through triangulation. Regarding the Re-ID feature bank, for each view that meets certain conditions, such as visible upper body keypoints, a bounding box confidence above a certain threshold, and minor occlusion, the decision to add the view's features to the feature bank is based on similarity, which is similar to the feature bank update process.

Matching with missing tracks For newly initialized targets, we need to re-identify whether they correspond to existing persons from missing targets, which avoids the situation where the trajectory of the same ID is split into multiple segments. Here, we use the Re-ID feature banks to determine if they corresponds to the same ID. Specifically, we calculate the similarity between all features in two feature banks, take the maximum value, and compare the value with the Re-ID similarity threshold α_s. If the similarity is above the threshold, the newly initialized target will be used to reactivate the missing targets, which means all the information will be transferred to the missing target. The latter will return to the confirmed tracking state and the former will be deleted. Otherwise, a new target will receive its new ID and advance to the matching pool.

4. Experiments

4.1. Dataset

The AIC24 Multi-Camera People Tracking (MCPT) dataset [38] consists of 90 multi-camera synthetic scenes, including common scenes such as storage areas, markets, and hospitals, generated by the NVIDIA Omniverse Platform. It comprises 40 scenes for training, 20 scenes for validation, and 30 scenes for testing. This iteration of the dataset marks a significant expansion in size, with the camera count growing from 129 to about 1,300, and the tracked individual count increasing from 156 to approximately 3,400. Additionally, 3D annotations and camera matrices are provided in the dataset. The videos are provided as high-resolution 1080p feeds, running at 30 frames per second, and come with tracking annotations that span across different camera views.

4.2. Evaluation Metrics

The Higher Order Tracking Accuracy (HOTA) based on 3D distance is used to rank the performance of each team on the leaderboard. HOTA, introduced by [18], was developed to rectify the issue where existing evaluation metrics disproportionately emphasized either detection or association. This metric balances the impact of detection, localization, and association. In the experiments, in addition to the HOTA value, we also provide scores for detection, localization, and association as references, marked as DetA, LocA and AssA respectively. Concretely, Detection Accuracy(DetA) assesses the correct identification of objects in an image or scene; Localization(LocA) measures the precision in pinpointing object locations; Association Accuracy (AssA) evaluates how accurately objects are tracked and identified over successive frames or views.

4.3. Implementation Details

Detection: Considering that the dataset predominantly consists of occlusion scenarios, in our experiments, we adopt the YOLOX model pretrained on the CrowdHuman dataset [27] by ByteTrack [44].

Re-identification: The network structure we adopted for Re-ID is MGN(R101) [37], whose weight is initialized with the pretrained weight from LUPerson [7]. To adapt to synthetic data, we finetune the model based on the training set and the validation set of AIC24 MCPT tracking dataset.

Pose Estimation: In pose estimation, we adopt the HRNet model [30] from the MMPose framework, known for its high-resolution networks that effectively maintain detailed spatial information throughout the model.

Parameter Selection: For the thresholds mentioned in Section 3, we empirically set the bounding box IOU threshold α_b as 0.5, the 3D epipolar distance threshold α_{epi} as 0.2, the homography distance α_h as 1.5(because the location of bottom points is often with significant noise) and the threshold of Re-ID affinity α_s as 0.5. Besides, we assign $(5, 1, 1, 5)$ to the affinity weights (w_b, w_{epi}, w_h, w_s). The bounding boxes and 2D keypoints are filtered with thresholds 0.3 and 0.7, respectively, to reduce false positive.

4.4. Experimental Results

Several Methods are evaluated in the test dataset of Track1 in the AI City Chanllenge 2024, as shown in Table 1. A series of improvements can be observed, mainly focusing on optimizing AssA, increasing from the original 30.2397% to the final 55.0560%. Specifically, compared to the baseline, the addition of Kalman filter improved the accuracy of predicted boxes, resulting in a gain of around 6 percentage points in association score. The incorporation of a feature bank optimized the matching of disappeared targets' IDs, reducing instances of trajectory splitting under the same ID and gaining around 2 percentage points in association. Furthermore, the addition of adaptive state-aware affinity for occluded and missing targets brought about a total increase of approximately 17 percentage points in association and around 13 percentage points in HOTA score. The final HOTA score reached 67.2175%.

Method	HOTA	DetA	LocA	AssA
Baseline	48.4516	80.2453	93.7445	30.2397
+ Kalman Filter	52.284	78.7058	93.474	36.4094
+Feature Bank	54.6255	82.1009	93.4924	38.1482
+Occluded Correction	64.1062	83.8852	93.8214	50.4193
+Missing Correction	**67.2175**	**84.0312**	**93.8221**	**55.0560**

Table 1. Performance comparison of different methods on the AIC24 dataset. The best results are highlighted in bold.

Our online multi-camera multi-people tracking method achieved a HOTA score of 67.2175 in the evaluation system of the AI City Challenge 2024 track 1, ranking second among all teams. The final leaderboard is shown in Table 2.

Ranking	Team ID	Team Name	HOTA
1	221	RIIPS	71.652
2	79	**SJTU-Lenovo(Ours)**	67.2175
3	142	FraunhoferIOSB	60.8792
4	40	NetsPresso	60.8233
5	8	UWIPL-ETRI	57.1445
6	50	ARV RETERIU	51.0556
7	5	SKKU Automation Lab	45.1575
8	124	STCHD	40.6202
9	162	Asilla	40.3361
10	21	TryThis	33.4879

Table 2. Leaderboard of Track 1 in the AI City Challenge 2024.

5. Conclusion

In this paper, we propose an online multi-camera multi-people tracking system that comprehensively considers the spatial and appearance information of the target. By introducing adaptive state-aware Re-ID affinity to correct the ID switch phenomenon under occlusion, our method significantly improves the accuracy of data association. When tested in different scenarios, this system demonstrates effectiveness and robustness. Our proposed system ranked second on the leaderboard of 2024 AI City Challenge Track1 in HOTA score.

References

[1] 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, 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8649–8658, 2018. 2

[4] Long Chen, Haizhou Ai, Rui Chen, and Zijie Zhuang. Aggregate tracklet appearance features for multi-object tracking. *IEEE Signal Processing Letters*, 26(11):1613–1617, 2019. 7

[5] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3279–3288, 2020. 3, 7

[6] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalized global graph model-based approach for multi-camera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, 2016. 3

[7] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 6, 8

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3650–3657, 2013. 3

[11] 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, et al. 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*, pages 5238–5248, 2023. 3

[12] Jeongho Kim, Wooksu Shin, Hancheol Park, and Jongwon Baek. Addressing the occlusion problem in multi-camera people tracking with human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5462–5468, 2023. 3

[13] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 2

[15] Zongyi Li, Runsheng Wang, He Li, Bohao Wei, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Boyuan Liu, Zhongyang Li, and Hanqing Zheng. Hierarchical clustering and refinement for generalized multi-camera person tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5519–5528, 2023. 3

[16] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang. Pop: Person re-identification post-rank optimisation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 441–448, 2013. 2

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[18] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 8

[19] Quang Qui-Vinh Nguyen, Huy Dinh-Anh Le, Truc Thi-Thanh Chau, Duc Trung Luu, Nhat Minh Chung, and Synh Viet-Uyen Ha. Multi-camera people tracking with mixture of realistic and synthetic knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5495–5505, 2023. 3

[20] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1846–1855, 2015. 2

[21] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13784–13793, 2021. 3

[22] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[25] Ergys Ristani and Carlo Tomasi. Tracking multiple people online and in real time. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 444–459. Springer, 2015. 7

[26] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 3

[27] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A bench-

mark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 8

[28] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1179–1188, 2018. 2

[29] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 475–491. Springer, 2016. 2

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 8

[31] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 3800–3808, 2017. 2

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 2

[33] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G Narasimhan. Self-supervised multi-view person association and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2794–2808, 2020. 3

[34] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7942–7951, 2019. 2

[35] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 2

[36] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1288–1296, 2016. 2

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 8

[38] 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. 1, 7

[39] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8042–8051, 2018. 2

[40] Longyin Wen, Zhen Lei, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision*, 122:313–333, 2017. 3

[41] 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

[42] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 2

[43] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, and Jun Chen. Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1239–1242, 2015. 2

[44] 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. 3, 8

[45] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2613–2626, 2022. 3

[46] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017. 2

[47] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 2

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017. 2

[49] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 2

[50] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 2