Hierarchical Clustering and Refinement for Generalized Multi-Camera Person Tracking

Zongyi Li[1 †] Runsheng Wang[1 †] He Li[1] Bohao Wei[1] Yuxuan Shi[1*]
Hefei Ling[1] Jiazhong Chen[1] Boyuan Liu[1] Zhongyang Li[1] Hanqing Zheng[1]
[1] Department of Computer Science and Technology, Huazhong University of Science and Technology.
{zongyili,wrsh,he_li,xavid,shiyx,lhefei,jzchen,leobryan,lzy123,zgwxzhq}@hust.edu.cn

Abstract

Multi-camera person tracking has gained significant attention in recent times, owing to its widespread application in surveillance scenarios. However, this task is challenging due to the variance viewpoints, heavy occlusion, and illumination changes. In order to tackle these challenges, we propose a novel Hierarchical Clustering and Refinement framework for Generalized Multi-Camera Person Tracking. Specifically, our framework comprises two main components: hierarchical clustering and hierarchical refinement. Compared with directly clustering tracklets among multiple cameras, our hierarchical clustering strategy can progressively assign tracklets to correct targets. Nevertheless, the clustering and tracking process would inevitably produce incorrect matchings. Therefore, a hierarchical refinement strategy is proposed to reduce these incorrect matches which includes: intra-camera tracklet level refinement, appearance refinement, spatial-temporal refinement, and face refinement. Extensive experiments show the effectiveness of our method, which achieves 92% *IDF1 in 2023 AI CITY CHALLENGE track1, ranking 5th on the leaderboard.*

1. Introduction

Recently, many researchers have been continuously exploring elaborated approaches for person retrieval among multiple cameras under various video surveillance scenarios to create more practical applications for person search. Compared with vanilla image-based person re-identification, Multi-Target, Multi-Camera Tracking (MTMCT) is a more practical yet challenging computer vision task. Apart from recognizing persons across cameras, it requires tracking multiple targets within every single camera, as well as cross-camera tracklets association. With the trajectories of multiple targets from multiple cameras, MTMCT provides a multi-view and more informative analysis of different targets and boosts the development and practicality of the public security system. Typically, the pipeline of MTMCT can be attributed to the following steps: 1) Detection of Pedestrians, 2) Single Camera Tracking, 3) Person ReID Feature Extraction, and 4) Cross-camera Association. Concretely, the first step localizes the positions of persons in each frame of the surveillance video. The second step conducts short-term tracking of persons detected from the first step using state-of-the-art Multiple Object Tracking (MOT) methods to obtain the short-term tracklets. The third step extracts the person re-identification (ReID) feature for each tracklet with deep person re-identification models. The final step utilizes the ReID features of tracklets obtained from the third step, and associates tracklets across cameras according to the similarities of extracted tracklet features.

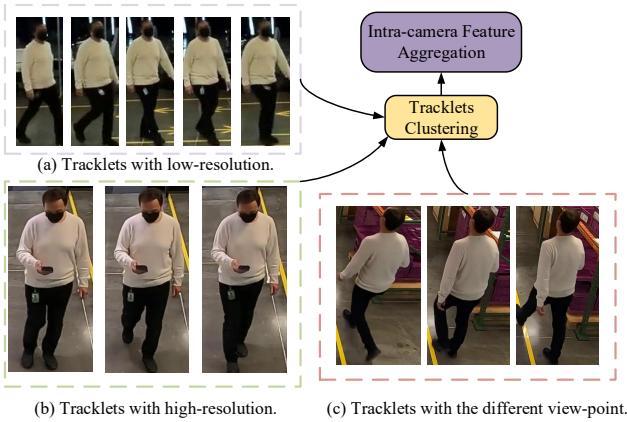

Figure 1. Clustering tracklets within a single camera. Tracklets of the same identity with different resolutions and different viewpoints can be aggregated by the clustering results.

However, there are several issues with the MTMCT task:

Firstly, MTMCT should be applied to diverse surveillance scenarios. Notably, with the development of the metaverse, tracking persons in virtual surveillance scenarios is

† Equal contribution. ∗ Corresponding author.

worth exploring. In the first track of AICity 2023, the MTMCT task under the diverse indoor surveillance scenarios, including the real surveillance scenarios and virtual surveillance scenarios, is considered.

Secondly, MOT within a single camera can only capture short-term tracklets, and associating targets with individual tracklets is not robust enough since it is susceptible to the variance of viewpoints, illumination, and resolution of input images. In real public places, such as offices and supermarkets, the same identity often appears and disappears repeatedly within the same camera. This indicates that one identity has multiple tracklets under the same camera. As shown in Fig. 1, the shown tracklets are of different resolutions and viewpoints. Intuitively, if we first aggregate the tracklets of the same target via intra-camera clustering, the aggregated features are more robust for the cross-camera association. Therefore, we propose **hierarchical clustering** as our clustering strategy. Since intra-camera tracklets clustering is easier than inter-camera tracklets, our method first conducts intra-camera clustering for intra-camera tracklets with extracted person ReID features to obtain more robust representations for all intra-camera targets. Subsequently, we take the intra-camera clustered features as the robust person ReID representations for corresponding targets and furtherly utilize the clustered features to associate cross-camera targets.

Finally, we observe the quality of frames in a tracklet cannot be assured, since the individual images suffer from occlusion and the variance of illumination. Moreover, there are many mis-clustered tracklets. Therefore, we propose **hierarchical refinement**, including the tracklet-level refinement and cluster-level refinement. For tracklet-level refinement, we filter out the low-quality frames which are dissimilar to the video-level tracklet features. For cluster-level refinement, we introduce face representations as another effective biometric clue for refining the clustering results. An example is shown in Fig. 2, the mis-clustered tracklet is clustered to c_1, since it has a similar ReID feature with c_1. As for the face representation, it differs greatly from c_1 and is similar to c_2. Specifically, we first pick up the tracklets which are possible to be mis-clustered by choosing the ones which are dissimilar from the corresponding clusters. Then, we associate the selected tracklets with face representations.

Our contributions can be attributed as following:

- We propose hierarchical clustering, which separately conducts intra-camera clustering and inter-camera clustering to associate identities across cameras.
- We propose hierarchical refinement, including tracklet-level refinement and cluster-level refinement. The two refinements respectively refine the tracklet features by filtering out low-quality frames, and boost the clustering results by taking advantage of face representations as another type of effective biometric clue.
- Our method achieves advanced performance and fifth place in the first track of AICity 2023.

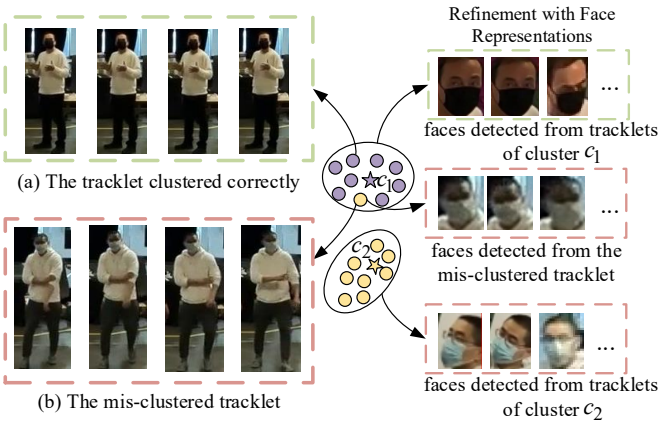

Figure 2. Cluster-level refinement with face representations. (a) shows a tracklet that is correctly clustered to cluster c_1. (b) shows a tracklet which is mis-clustered to c_1 (the ground truth cluster of the tracklet is c_2), since it has similar appearance feature with c_1. However, the faces of the mis-clustered tracklet differ extremely from the faces detected from c_1.

2. Related Work

2.1. Person Detection

Person detection is a critical task in computer vision that involves identifying and localizing individuals in images or videos. It has a wide range of applications, including video surveillance, autonomous driving, human-computer interaction, and more.

Generally, person detection algorithms can be divided into two main categories including one-stage detectors [33, 41] and two-stage detectors [2, 19, 42]. One-stage detectors, such as YOLO [41] and SSD [33], are known for their real-time performance and speed but may sacrifice some accuracy for speed. In contrast, two-stage detectors, such as Faster R-CNN [42], Mask R-CNN [19], and Cascade R-CNN [42], provide greater precision and flexibility but require more computational resources.

Owing to the success of transformer structures in natural language processing, transformer-based detectors such as DETR [4] and Swin Transformer [35] are booming recently. By utilizing vision transformers that treat an image as a sequence of patches, these detectors employ the self-attention mechanism to capture long-range dependencies. Consequently, they have demonstrated competitive performance on object detection benchmarks.

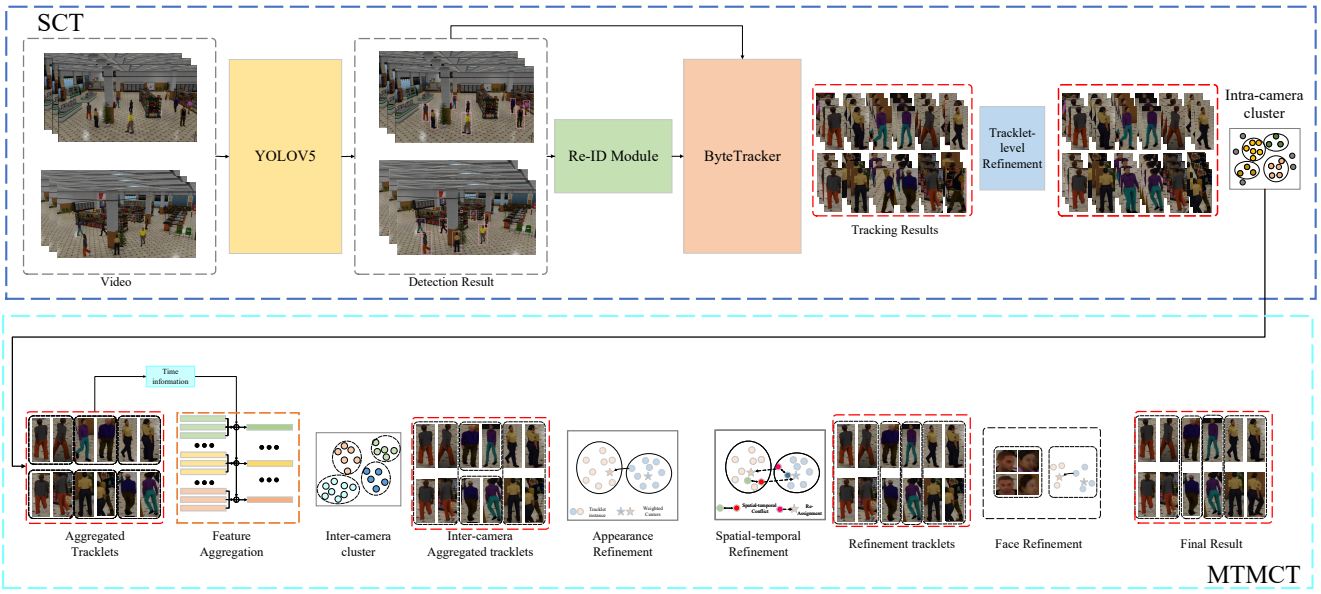

Figure 3. The pipeline of our hierarchical refinement for the MTMCT system. We first detect all persons using the YOLOv5 model and extract the person ReID feature using the ReID model. Then all bounding boxes and ReID features are sent to the ByteTracker to associate the pedestrian bounding boxes in the video. Hierarchical clustering is performed on intra and inter tracklets to merge single tracklets. And tracklet-level refinement and inter-camera tracklets refinement are performed to improve the clustering results.

2.2. Person Re-identification

Person re-identification (ReID) [21, 30, 34, 36, 55] is an important task in computer vision, particularly in surveillance systems, which aims to match pedestrians across different cameras. Over the past few years, remarkable progress has been made in Person ReID through advanced techniques such as self-supervised learning and transformer-based ReID.

Self-supervised learning techniques, such as contrastive learning, have emerged as successful approaches for person ReID representations learning without large amounts of labeled data [24]. These techniques have simplified the training of ReID models on large-scale datasets. Existing self-supervised learning methods can be classified into three categories. Firstly, generative self-supervised learning [8, 11, 25, 40] aims to generate synthetic samples, which are further involved to enlarge the training data and enhance the generalization performance of the ReID model. Secondly, contrastive self-supervised learning [5, 7, 9, 17, 18, 48] aims to train an encoder by drawing the embeddings of the same sample with distinct data augmentation closer while pushing embeddings of other samples away. Finally, adversarial self-supervised learning [12, 16, 27, 28] aims to generate fake samples by training a generator and distinguish them from genuine samples by a discriminator. Currently, contrastive self-supervised learning has established its dominance in computer vision.

CNN-based techniques have dominated the ReID community for several years. However, the popularity of pure-transformer models is on the rise. The TransReID model [21], for instance, was the first to effectively employ Vision Transformers for Person and Vehicle ReID, which achieves state-of-the-art results. Many other works try to utilize Transformers to aggregate features or information from CNN backbones. For example, [29, 44, 54] integrate Transformer layers into the CNN backbone to aggregate hierarchical features and align local features. Additionally, for video ReID, [34, 55] leverage Transformers to aggregate appearance features, spatial features, and temporal features in order to learn a discriminative representation for a person tracklet.

Combining the two methods above, TransReID-SSL [36] further investigates that DINO [6] algorithm with Transformer architecture obtains the best ReID performance among the existing self-supervised learning (SSL) methods and network architectures.

2.3. Single-camera Tracking

Single-camera tracking (SCT) is a subfield of computer vision that aims to track the movement of objects in a video sequence captured from a single camera [10]. There are currently two types of SCT algorithms. The first type follows the tracking-by-detection paradigm [1, 3, 13, 38, 51, 53, 56], while the second type called joint-detection-tracking com-

bines object detection with ReID in a single network [31, 45, 49, 57, 59].

Tracking-by-detection methods, such as SORT [1] and DeepSORT [51], first detect objects using deep detection models, and then obtain the trajectories of targets by data association between adjacent frames. Due to the improvement of object detection techniques [15, 19, 35, 41, 42], these methods have been dominant in the SCT task for years. Joint-detection-tracking methods, such as those incorporating appearance embedding or motion prediction into detection frameworks, achieve comparable performance with low computational costs. However, these methods face a challenge in optimizing the competition between different components, which ultimately constrains their tracking performance.

2.4. Multi-Camera People Tracking

Based on the results of the aforementioned tasks, the primary goal of multi-camera people tracking is to establish a series of tracking chains across different cameras. To enhance their pipeline, some works [22, 23, 26, 39] have incorporated external information about the camera setup. To prevent infeasible cross-camera transitions, [23, 32, 46] utilize scene topology, while camera adjacency is considered in [22, 32, 46]. In [22], the movement directions were used to determine the feasibility of camera transitions, and camera-specific regions are defined to identify the possibility of tracks appearing in multiple cameras. Furthermore, clustering approaches have been identified as effective for addressing this task in related literature [23, 26, 43, 47].

3. Method

3.1. Overview

Our proposed framework for Multi-Target Multi-Camera Tracking (MTMCT) is illustrated in Figure 3. The framework consists of four main components: Person Detection, Person Re-identification, Single-camera tracking, and Multiple-camera tracking. The overall process can be summarized as follows: (1). Using the person detector to obtain person bounding boxes from every camera view. (2). Extracting person ReID features from each bounding box by employing a pre-trained ReID model. (3) Utilizing the Single-camera tracking model to generate single-camera tracklets for each camera. (4) Clustering tracklet features to associate intra-camera tracklets. (5) Applying clustering methods for associating inter-camera tracklets. (6) Refining the inter-camera tracklets using appearance, spatial-temporal and face constraints.

3.2. Person Detection

Person detection is the initial and critical step in cross-camera tracking; therefore, utilizing a reliable detector is paramount. The person detection phase requires precise and non-missing pedestrian detection boxes. To effectively detect more pedestrians, we employ YOLOv5 as our detector. Overall, YOLOv5 is a highly accurate and efficient object detection algorithm that has been widely adopted in various applications.

In the competition scenario, both real and virtual pedestrian images are present. As the pre-trained YOLOv5 model has already achieved excellent person detection performance on the COCO dataset, we use it to detect persons in real-world scenes. For virtual scenarios, we train the YOLOv5 detector from the scratch using the virtual dataset provided by AICity2023. We ignore other categories and only detect pedestrians in the scene by applying NMS to remove duplicate detection boxes. With the YOLOv5 detector, we obtain the detection boxes and confidence scores of pedestrians in the corresponding videos.

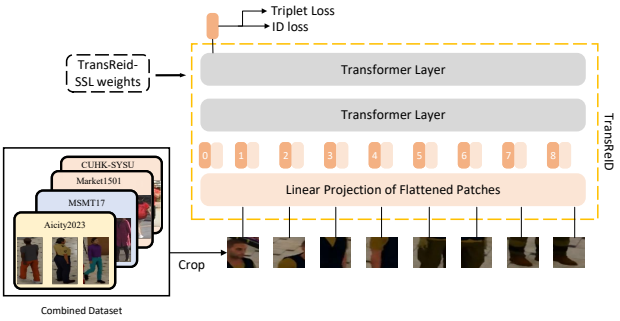

Figure 4. The pipeline of our ReID model. We initialize TransReID model by the TransReID-SSL pre-trained weights and fine-tune the ReID model on a combined dataset.

3.3. Person Re-identification

Our person re-identification model is based on the TransReID-SSL [36], which has been pre-trained on the LUperson dataset [14], and is known for its ability to extract robust and domain-invariant ReID features. As shown in Fig. 4, to further improve its performance, we train the TransReID [21] model on a combined dataset comprising Market-1501 [58], MSMT17 [50], CUHK-SYSU [52], and the AiCity2023 dataset. We initialize the model's weights using TransReID-SSL pre-trained model and fine-tune it with an input image of 256×128 size, with the Cross-Entropy loss and triplet loss. The Cross-Entropy loss function can be formulated as follows:

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right), \tag{1}$$

where y is the ID label for i-th image, and N is the number of images in the combined dataset. And the triplet loss can

be formulated as follows:

$$L_{tri} = \sum_{i=1}^{N} \max\left(m + d\left(f_i^a, f_i^n\right) - d\left(f_i^a, f_i^p\right), 0\right), \quad (2)$$

where d is the l_2 distance, f_i^p, f_i^n is the positive and negative samples and m is the margin of triplet loss.

3.4. Single-camera person tracking

For single-camera person tracking, we need to associate the detected bounding boxes in the video to obtain the corresponding tracklets. We use ByteTrack as the tracking algorithm, which can associate low-confidence detections by associating every detection box with a unique identity.

As shown in Fig. 3, we first extract person features using the ReID model, and then input the bounding box and ReID features into the tracking model. The tracker considers motion information and visual similarity to assign a tracklet ID to each detected box. With the tracking model, we can associate the pedestrian bounding boxes in the video to obtain tracklets.

Single-camera Tracklet-level Refinement Due to some crowded situations, people with different identities may be assigned to the same tracklet, which causes the ID switch problem. Therefore, we first calculate the intra-variance of the tracklet by calculating the distance between the feature of individual frames and the mean feature of all frames in a tracklet. If a tracklet contains different identities, its intra-variance would be high due to the appearance variance. Therefore if the intra-variance is greater than a certain threshold, we use the K-means algorithm to split the tracklet into two tracklets to reduce errors caused by single-camera trackers. Specifically, we set the threshold to 0.3.

Single-camera tracklet association Since AiCity2023 track1 is under the indoor setting, the same person may have multiple trajectories in a single camera. Therefore, we first roughly cluster these tracklets and merge these tracklets in the same clusters, considering them as the same person. To do this, we firstly obtain the tracklet feature by averaging the features of all frames in the tracklet. So the tracklet can be denoted as $trac = \{f, ti, to, c\}$, where f is the tracklet feature, ti and to is the time the person enters and exits the camera, c is the camera ID. Then the appearance distance among the tracklets can be denoted as $D_{appearance}$. Moreover, the Jaccard distance matrix is also calculated to combine neighbor information, which can be denoted as $D_{jaccard}$. Moreover, considering that tracklets within the same camera with overlapping time intervals cannot belong to the same cluster, we set the distance of these tracklets to 1.

$$D_{spatial}^{i,j} = \begin{cases} 1, & \{ti^i, to^i\} \cap \{ti^j, to^j\}! = 0 \\ 0 & \text{else} \end{cases}. \quad (3)$$

Therefore, the fusion distance matrix can be formulated as:

$$D = D_{appearance} + \alpha D_{jaccard} + \beta D_{spatial}, \quad (4)$$

where α and β are the weights parameters. After obtaining the fusion distance matrix, we perform DBSCAN algorithm to roughly cluster these tracklets and merge tracklets in one cluster.

3.5. Cross-camera person association

In this module, we will describe our multi-camera tracking framework, which takes tracklets aggregated from single-camera tracking as input. Single-camera person tracking has employed clustering to group intra-camera tracklets based on the fusion distance between them. Similarly, we use the K-means clustering algorithm to group these aggregated tracklets based on their aggregated features. Since directly averaging tracklet features does not take into account the length of each tracklet. Ideally, tracklets with longer spans should be given greater weight. Therefore, the merged intra-camera tracklets features are weighted by different tracklets time spans, which can be formulated as:

$$f_{sct}^k = \frac{1}{|\mathcal{I}_k|} \sum_{f_i \in \mathcal{I}_k} w_i f_i, \quad (5)$$

where $w_i = \frac{log(l_i)}{\sum_{j=0}^{n} log(l_j)}$ is the time weights for tracklets i, l_i is the number of images in tracklets i. This weighted method can pay more attention to longer tracklets. And the weighted sum features f_{sct} are used in the following inter-camera clustering process. After performing the clustering algorithm on intra-camera aggregated tracklets, tracklets belonging to the same cluster are assigned the same ID.

Moreover, clustering would inevitably lead to some incorrect ID assignments. To refine clustering results in the cross-camera person association process, we employ three Refinement methods: Appearance Refinement, Spatio-temporal information Refinement, and Face Refinement, to refine the cross-camera tracking results.

Appearance Refinement Since the clustering algorithm emphasizes overall similarity but neglects the length of each tracklets, where longer tracklets should be assigned bigger weights, we first calculate the center features for each cluster f_{mtmct}^k using the time-weighted sum by Eq 5. Next, we re-calculate the distance between each tracklet and the center of the cluster and reassign the identity of each tracklet to its nearest cluster. With this refinement, we can effectively improve the tracklets by taking their appearance and duration time into account.

Spatio-temporal Refinement Similar to the intra-camera person tracking method, we need to exclude abnormal tracklets in clustering results to obtain the final accurate MTMCT results since inter-camera clustering may also

cluster the tracklets in the same camera with overlapping time intervals. Specifically, after getting the inter-camera clustering result, we traverse all the clusters and reassign the tracklets with camera-time overlap to different clusters. For conflict pairs in the same cluster, only the closest tracklet to the cluster center will be remained, while other conflict tracklets are assigned to different clusters according to the similarity. which is shown in Fig. 3.

Face Refinement Since there are many low-resolution images in the video without faces, we do not directly incorporate the face representations when calculating similarity in clustering. Instead, we incorporate the face representations to refine the inter-camera clustering results. The face refinement involves two steps: First, we extract faces from all images using the MTCNN model. And the ArcFace model is used to extract the corresponding face features. Second, we calculate the average face features of tracklets in a cluster as the face representation of the cluster. Since face information is more informative, we directly assign the mis-clustered tracklets according to the face representations to a new cluster. Concretely, when face feature of a tracklet differs largely from that of the assigned cluster stemming from ReID features, we regard the tracklet as a mis-clustered one. Subsequently, when the face similarity between the tracklet and its nearest face cluster is larger than 0.8, we reassign the ID of this tracklet to the nearest face cluster.

With inter-camera clustering and hierarchical refinement strategies, all tracklets can be assigned to an ID label.

4. Experiments

4.1. Dataset

In this track, the MTMCT dataset contains real data and virtual synthetic data. This dataset has 1,491 minutes of videos and a total of 130 cameras. The video data are all in high resolution (1920x1080) at 30 FPS and are divided into 22 subsets, including 10 subsets for training, 5 subsets for validation, and 7 subsets for testing. Moreover, we also utilize three public person ReID datasets:Market-1501, MTMC17, CUHK-SYSU for ReID model training.

4.2. Evaluation Metric

We use IDF1, IDP and IDR as MTMCT evaluation metrics. The IDF1 score is a metric used to evaluate the performance of object detection models. It measures the ratio of correctly identified detections, taking into account both the ground truth and the false negative, true negative, and true positive counts. The IDF1 score is specifically derived from the counts of IDFN, IDTN, and IDTP and can be formulated as follows:

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN}. \quad (6)$$

4.3. Implementation Details

Our framework is implemented on RTX-3090 GPUs with 24G memory. For the person detection module, we utilize the YOLOv5-l model pre-trained on the COCO dataset to perform person detection in real scenarios. In virtual scenarios, we train the YOLOv5 model on the AIcity track1 dataset from the scratch. The IOU threshold for detection is set to 0.3, and the NMS threshold is set to 0.45. For the person ReID module, we train the TransReID model on the combined dataset which consists of Market, MSMT17, CUHK-SYSU, with the pre-trained TransReID model as the initialization weights. Additionally, we adopt ByteTrack for single-camera person tracking.

4.4. Quantitative Analysis

In this subsection, we report the ablative analysis of the proposed hierarchical clustering strategy, the hierarchical refinement, as well as different backbones for extracting person ReID features for all tracklets. Concretely, the ablative study of proposed clustering and refinement strategies is conducted on the test set of AICity 2023 track 1 dataset, while the ablative experiment of ReID backbones is conducted on the validation dataset divided by our-self.

The ablative results of proposed clustering and refinement strategies are shown in Tab. 1. It is impressive that the hierarchical clustering ("intra-camera cluster" together with "inter-camera" cluster in Tab. 1) promotes the performance by a large margin. Moreover, different refinement strategies are able to boost performance. Specifically, appearance, spatial-temporal, and tracklet-level refinement promote the performance by 2%, 1%, and 2%, respectively.

Ablative results with different ReID backbones are shown in Tab.2. Apparently, TransReID performs better than Resnet50-based models. Moreover, fine-tuning TransReID trained with a self-supervised scheme achieves the best result. Therefore, we choose TransReID-SSL as our backbone.

As shown in Fig. 3, our method achieves 0.921 IDF1 in the Track1 of AIcity2023 challenge, which ranks fifth place compared with other teams.

4.5. Visualization

The final MTMCT results are visualized in Fig5. Each colum of this figure represents trajectories in the different camera, while each row presents the trajectories in the same camera but at different time. For instance, the ID 17 person appear in different camera with various viewpoint and occlusion, yet our methods also can effectively match the their tracklets across these different cameras.

Furthermore, we present the results of inter-cluster analysis as illustrated in Fig. 6. The tracklets belonging to the same cluster have been assigned with the same IDs. In addition, our approach is capable of managing cases where there

Inter-camera Cluster	Intra-camera Cluster	Appearance Refinement	Spatial-temporal Refinement	Tracklet-level Refinement	Performance (ID-F1)
✓					0.77
✓	✓				0.82
✓	✓	✓			0.84
✓	✓	✓	✓		0.85
✓	✓	✓	✓	✓	0.87

Table 1. Ablation study for proposed clustering and refinement strategies.

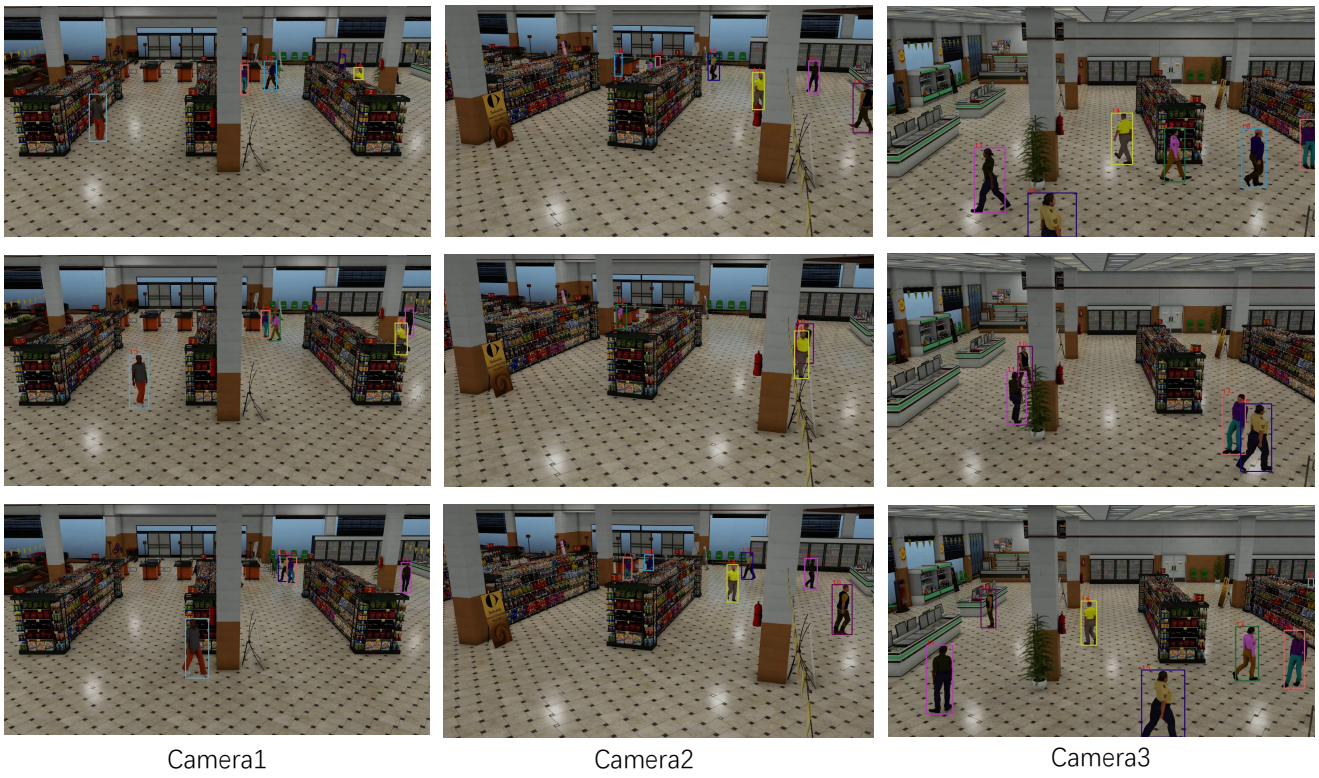

Figure 5. Visualization of final tracking results on AIcity2023 test set. The same ID people are marked with the same color in different cameras.

Backbones	Rank-1	Rank-5	Rank-10	mAP
Resnet50 [20]	0.74	0.76	0.76	0.68
Resnet50-IBN [37]	0.77	0.82	0.85	0.73
TransReID [21]	0.87	0.88	0.89	0.83
TransReID-SSL [36]	**0.91**	**0.96**	**0.97**	**0.88**

Table 2. Ablation study for different backbones.

is body overlap between distinct person images, as can be seen in cluster 2. With the assistance of the anti-occlusion ability of the TransReID model, we are able to match partial tracklets to the complete-body tracklets, as demonstrated in cluster 3 of Fig. 6.

Rank	TeamID	IDF1
1	6	0.9536
2	9	0.9417
3	41	0.9331
4	51	0.9284
5	**113 (ours)**	**0.9207**
6	133	0.9109
7	34	0.9104
8	82	0.8981
9	151	0.8676
10	38	0.8676

Table 3. Comparison with other teams on track1, and our teams take fifth place.

Figure 6. Visualization of final inter-clustering results on AIcity2023 test set. Different clusters represent different people.

5. Conclusions

In this paper, we propose an effective and novel MTMCT framework, consisting of person detection, single-camera multiple target tracking using MOT algorithms, ReID features extraction, hierarchical clustering, and hierarchical refinement. We demonstrate that hierarchical clustering, where the intra-camera and inter-camera clustering algorithms are conducted sequentially, is extremely beneficial for the indoor video surveillance scenario, where persons appear and disappear in a camera repeatedly. Moreover, we propose a number of refinement strategies, which mainly include tracklet-level refinement and cluster-level refinement. Notably, we use face representations as another type of clue to correct the mis-clustered tracklet, and furtherly boost the cross-camera association performance. We believe that MTMCT with other biometric clues, such as face and gait representations, is worth exploring in the future for the computer vision community.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 61972169, in part by China Postdoctoral Science Foundation 2022M711251, in part by the National key research and development program of China(2019QY(Y)0202, 2022YFB2601802), in part by the Major Scientific and Technological Project of Hubei Province (2022BAA046, 2022BAA042), in part by the Research Programme on Applied Fundamentals and Frontier Technologies of Wuhan(2020010601012182), in part by the Key Joint Projects of Enterprises(U22B2017), and the Knowledge Innovation Program of Wuhan-Basic Research.

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. 3, 4

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[3] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[10] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 3

[11] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 3

[12] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *Advances in neural information processing systems*, 32, 2019. 3

[13] Yunhao Du, Yang Song, Bo Yang, and Yanyun Zhao. Strongsort: Make deepsort great again. *arXiv preprint arXiv:2202.13514*, 2022. 3

[14] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, 2021. 4

[15] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[16] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015. 3

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[19] 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, 4

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[21] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 3, 4, 7

[22] Yuhang He, Jie Han, Wentao Yu, Xiaopeng Hong, Xing Wei, and Yihong Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 576–577, 2020. 4

[23] 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. 4

[24] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 3

[25] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 3

[26] Philipp Kohl, Andreas Specker, Arne Schumann, and Jurgen Beyerer. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1042–1043, 2020. 4

[27] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 577–593. Springer, 2016. 3

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[29] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. 3

[30] Zongyi Li, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Qian Wang, and Fengfan Zhou. Reliability exploration with self-ensemble learning for domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1527–1535, 2022. 3

[31] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 4

[32] Chong Liu, Yuqi Zhang, Hao Luo, Jiasheng Tang, Weihua Chen, Xianzhe Xu, Fan Wang, Hao Li, and Yi-Dong Shen. City-scale multi-camera vehicle tracking guided by crossroad zones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4137, 2021. 4

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 2

[34] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, Xuesheng Qian, and Xiaoyun Yang. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv preprint arXiv:2104.01745*, 2021. 3

[35] 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, 4

[36] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021. 3, 4, 7

[37] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 7

[38] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 3

[39] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 588–589, 2020. 4

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[41] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2, 4

[42] 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, 4

[43] 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. 4

[44] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *arXiv preprint arXiv:2107.05475*, 2021. 3

[45] Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. Siammot: Siamese multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12372–12382, 2021. 4

[46] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4173–4182, 2021. 4

[47] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017. 4

[48] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 3

[49] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020. 4

[50] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 4

[51] 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. 3, 4

[52] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017. 4

[53] 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. 3

[54] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021. 3

[55] Tianyu Zhang, Longhui Wei, Lingxi Xie, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*, 2021. 3

[56] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 1–21. Springer, 2022. 3

[57] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 4

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 4

[59] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 474–490. Springer, 2020. 4