Global Trajectory Helps Person Retrieval in a Camera Network

Xin Zhang, Xiaohua Xie, Member, IEEE, Jianhuang Lai, Senior Member, IEEE Wei-Shi Zheng, Member, IEEE

Abstract—We are concerned about retrieving a query person from the videos taken by a non-overlapping camera network. Existing methods often rely on pure visual matching or consider temporal constraint, but ignore the spatial information of the camera network. To address this problem, we propose a framework of person retrieval based on cross-camera trajectory generation which integrates both temporal and spatial information. To obtain the pedestrian trajectories, we propose a new cross-camera spatio-temporal model that integrates the walking habits of pedestrians and the path layout between cameras, forming a joint probability distribution. Such a spatio-temporal model among a camera network can be specified using sparsely sampled pedestrian data. Based on the spatio-temporal model, the cross-camera trajectories of a specific pedestrian can be extracted by the conditional random field model, and further optimized by the restricted nonnegative matrix factorization. Finally, a trajectory re-ranking technology is proposed to improve the person retrieval results. To verify the effectiveness of our approach, we build the first dataset of cross-camera pedestrian trajectories over an actual monitoring scenario, namely the Person Trajectory Dataset. Extensive experiments have verified the effectiveness and robustness of the proposed method.

Index Terms—Person retrieval, Trajectory Generation, Person re-id, Conditional Random Field

I. INTRODUCTION

In this paper, we consider the issue of retrieving a query person from the videos taken by a camera network over an area such as a living community or campus. Such a practical task is often associated with trajectory inference and has critical public security and commercial customer analysis applications. It is closely related to person re-identification and person tracking technologies.

For person retrieval, existing methods mainly rely on visual data such as single images [1]–[5] or videos [6]–[13]. These methods focus on extracting visual features of persons robust to illumination variation [14], resolution variation [15], [15] and occlusion [16]–[18], and study how to measure the similarity of these features. Yet, it is difficult to achieve accurate pedestrian matching by relying on pure visual information. For a camera network, besides vision cues, a large amount of non-visual information can be utilized, such as the geographical coordinates of the camera, the shooting time of each video frame, the roads of the concerned area, and walking habits of pedestrians among this camera network. Obviously, in a practical scene, mining non-visual cues may be beneficial to construct spatio-temporal constraints for enhancing pedestrian retrieval. This inspires us to exploit pedestrian trajectory for pedestrian retrieval because pedestrian trajectory contains not only the information relevant to the visual similarity between pedestrian images but also the spatio-temporal cues of the camera network. In addition, the obtained pedestrian trajectories can support other applications.

The frame images, video tracklets, and cross-camera trajectory of a pedestrian form hierarchical clusterings and can be regarded as the low-order representation to the high-order representation of the pedestrian progressively. Compared to a single frame image, the clusters generally provide richer information. In other words, using person tracklets or person trajectories can enhance the person retrieval. However, how to effectively obtaining the person trajectory is still an open question. The existing pedestrian detection and intra-camera pedestrian tracking technologies can get the set of person tracklets under each camera well. In our method, we further infer the cross-camera trajectories of a pedestrian by combining the tracklet matching, conditional random field, and non-negative matrix factorization technologies. After that, we develop a trajectory re-ranking technology to optimize the person retrieval results. Note that in our method the query may be a pedestrian tracklet or only a single pedestrian image, without the need of inputting spatio-temporal information.

A key to cross-camera pedestrian analysis is to model the empirical spatio-temporal relationship of the camera network. However, spatio-temporal modeling is challenging. Firstly, the pedestrian trajectories are often varying. For example, there may be multiple paths between a pair of cameras, and the walking time of pedestrians for the same path may also be different. Secondly, there is often a lack of data for establishing the spatio-temporal model. Currently, there are mainly three data sets can support the research of exploiting spatio-temporal relationship between cameras for person retrieval, i.e., Market1501 [19], DukeMTMC [20], and Campus4k [21]. However, these datasets only contain the images or video frames with their shooting time but no additional information such as the distance between cameras. Based on these databases, some researchers have attempted to use temporal constraints to improve the accuracy of person association [22]–[24]. These methods mainly consider the distribution of time interval for the same person appearing in a pair of cameras. In practice, it is difficult to collect enough samples for the distribution statistics of the time interval between each camera pair. To
overcome this deficiency, we propose a new spatio-temporal model by integrating the walking habits of pedestrians and the path distance between cameras, resulting in a joint probability distribution of temporal and spatial information. Here, the walking habits are mainly related to the walking speed and the tendency of path selection between cameras. This model utilizes not only visual cues but also non-visual cues, including the coordinates of the cameras and the path map of the area of interest, which are both easy to obtain in the actual scene. With the assistance of non-visual cues, the complete spatio-temporal relationships among the whole camera network can be modeled by the sparse sampling of pedestrian visual data. Based on the spatio-temporal model, we develop a conditional random field model to infer the cross-camera spatio-temporal adjacency graph for specific pedestrian. Finally, we propose to optimize the pedestrian trajectory from the spatio-temporal adjacency graph by the restricted nonnegative matrix factorization (25).

To verify the effectiveness of the proposed method, we collected a cross-camera pedestrian trajectory dataset in an actual residential area for the experiment. The experiment shows that the inference of pedestrian trajectories promotes the person retrieval accuracy. At the same time, experiments show that our method can also be combined with other Bayesian-based spatio-temporal methods to get better results.

The contributions of this paper are as follows:

- A novel framework of person retrieval based on cross-camera trajectories is presented. The pedestrian trajectories which imply both temporal and spatial information of the camera network help to promote the person retrieval accuracy, and further support other applications.
- A spatial-temporal model based on the joint probability distribution is designed for supporting the trajectory generation in a camera network, which embeds the walking habits of pedestrians and the path distances between cameras and can be constructed by the sparse sampling of pedestrian data.
- A spatio-temporal conditional random field method is developed to infer the pedestrian trajectories. The restricted nonnegative matrix factorization is employed to optimized the inference result. A trajectory re-ranking technology is proposed to improve the person retrieval results.
- The first cross-camera person trajectory dataset from the actual scene is collected. An evaluation method for person trajectory retrieval is also introduced. The dataset will be released when the paper is accepted.

The remainder of this paper is organized as follows. The related works are reviewed in Section II. Our dataset and the evaluation method of trajectory generation are introduced in Section III. The proposed method is introduced in Section IV. Experimental results are shown in Section V. Conclusion is in Section VI.

II. RELATED WORK

This section reviews related works about spatio-temporal modeling, trajectory generation, and re-ranking, which are all critical parts of our proposed framework.

A. Spatio-Temporal Model for Pedestrian Matching

To improve the accuracy of pedestrian matching, some researchers have proposed combining spatio-temporal information with visual features. For this purpose, how to build the spatio-temporal model for a camera network is essential. In 2018, Lv et al. (23) proposed using spatio-temporal constraints in their unsupervised person re-id method. They make a strong assumption that a gallery person always appears in the time window \( (t - \Delta t, t + \Delta t) \) when given a query person image at the time \( t \). In 2020, Xie et al. (21) proposed an association method to integrate both spatio-temporal probability and visual probability into a joint probability through an unsupervised way, which is to optimize person retrieval results in the prediction phase. In 2016, Huang et al. (22) proposed using Weibull distribution to model the distribution of cameras pair and to use it in the sorting stage after searching. In 2019, our group (24) proposed a fast Histogram-Parzen (HP) method to fit the spatio-temporal relationship between cameras. Above mentioned methods all focus on the relationship between two cameras while ignoring the overall relationship among a camera network. Therefore, these methods require sufficient data between each camera pair to establish a complete spatio-temporal relationship, which is difficult to meet in practice. In this paper, we model the spatio-temporal relationship of the camera network as a whole, which embeds the path distances between cameras. Based on this model, we can complete the construction of the spatio-temporal relationship between all camera pairs with a sparse sampling of pedestrian spatio-temporal trajectory. In addition, the existing person retrieval methods using spatio-temporal information are mainly based on the Bayesian strategy, which requires knowing the spatio-temporal information of the query in advance, which limits the use scenario of this kind of method. The clustering and scattering mechanism proposed in this paper can optimize the retrieval under the condition of unknown query spatio-temporal information and is applicable to wider scenarios.

B. Pedestrian Trajectory Generation

Some existing video analysis tasks have considered the generation of cross-camera pedestrian trajectories. For example, Multi-Target Multi-Camera Tracking (MTMCT) aims to determine the positions of every person at all times from video streams taken by multiple cameras (25). According to the different application scenarios, it can be divided into centralized and distributed tracking approaches. The centralized approach is wildly used in overlapping camera views tracking systems. It aims to reduce the effect of occlusions and noisy observations in a crowded environment or in the monitoring of small areas (27), (28). The distributed approach is generally used in applications which are dealing with non-overlapping camera views that are designed to monitor large areas most of the time (29). In this kind of MTMCT method, they usually assume that the camera’s field of view is often non-overlapping, and there is no explicit geometric relationship between cameras. Javed et al. (30) proposed a method that calculates the similarity of the appearance and that of spatio-temporal data between every two cameras to judge who this person is. The difference
between the above work and ours is that MTMCT focused on the association, but ours focused on both association and retrieval collaboration. Furthermore, existing MTMCT studies have limitations, such as small intervals between cameras, short duration in a video, and less tracking targets. Our work exceeds these limits.

C. Re-ranking for Person Retrieval

Re-ranking means using high confidence samples to reorder the initial search results, which has been studied in the generic instance retrieval [31]–[34], [35]. Some researchers have also designed reordering methods for person Re-id task [36]–[40]. Li et al. [40] proposed a re-ranking method by analyzing the relative information and direct information of near neighbors of each pair of images. Garcia et al. [38] proposed an unsupervised re-ranking method by jointly considering the content and context information in the ranking list. Zhong et al. [31] proposed a k-reciprocal encoding method to re-ranking the re-ID results. These re-ranking methods only use the visual information of persons while ignoring the temporal or spatial information. In 2019, our group [24] proposed a spatio-temporal two-stream re-ranking method to reorder the list of the retrieval result. However, this method requires that the query’s spatio-temporal information be known, limiting its application scenarios. In this paper, we develop a novel trajectory re-ranking technology to optimize the person retrieval. The proposed trajectory re-ranking method does not need the spatio-temporal information of the query.

III. PERSON TRAJECTORY DATASET

Thanks to many public Re-id datasets [19]–[21], [42], great progress has been made in person retrieval. However, existing Re-id datasets do not provide labels of the cross camera trajectory of persons. This brings a great challenge to the research of spatio-temporal relationship. To break this dilemma, we collect a dataset called Person Trajectory Dataset (PTD), containing complete image sequences from non-overlapping outdoor cameras in a residential area, labeled full person trajectories.

This PTD is collected from a camera network of nine cameras from 8 a.m. to 8 p.m. The spatial distribution of the cameras for this dataset is shown in Fig. 1. The person image sequence acquired is then preliminarily labeled by FPN target detection method [43] and DeepSort tracking method [44], and then the annotation results are corrected manually. The images of any person that only appears under one camera are also eliminated to ensure the persons in PTD have appeared under at least two different cameras.

The dataset can be divided into two parts: visual data and spatio-temporal data. The visual data includes 17,996 images of 662 individuals, of which 5,003 images of 197 individuals are used as the testing set, and the remaining 12,993 images of 465 people are used as the training set.

Fig. 1: The spatial distribution of the cameras in the Person Trajectory Dataset. For each camera, the satellite enlarged image and the camera view of the corresponding cameras are displayed. The following numbers, such as SQ0921, indicate the camera index.

For spatio-temporal data, it also consists of a training set and testing set, within which the data are derived from PTD’s training set and testing set, respectively. Fig. 2 shows the amount of data between each camera pair in spatio-temporal data. From Fig. 2, we can see that it is tough to establish the spatio-temporal relationship model of each camera pair when there is limited data. First, there is an imbalance in the amount of data between different camera pairs. For example, there are 228 samples in camera pair No. 3, while there are only 12 samples in camera pair No. 6. Second, the number of positive and negative samples is also unbalanced across the camera pair. In spatio-temporal data, a positive example refers to the time difference in a specific person’s trajectory under each camera. A negative example refers to the time difference between the different trajectories of the same person and other people under all cameras. The sample imbalance between camera pairs poses a great challenge to the spatio-temporal modeling between cameras. Because of imbalance in spatio-temporal data, only ten camera pairs are used in training, and their distribution is shown in Fig. 3. In Fig. 3, the distance represents the length of the shortest path between the corresponding camera pairs, and the time difference means the absolute value of the time past between a person’s appearances in their corresponding camera pair.

We would like to compare the PTD with the existing MTMCT datasets. There are mainly three public MTMCT benchmark datasets, PET09S2L1 [45], CAMPUS [46], and EPFL [47]. In PETS, seven cameras are used to film about ten targets entering and walking through a footpath. In CAMPUS, they collected about 15 persons in the four video subsets. The amounts of persons in these MTMCT datasets are too small to be used to evaluate person retrieval tasks. Compared with these MTMCT datasets, the PTD data comes from the surveillance
camera of the real scene, with more extensive coverage and more pedestrians recorded.

IV. PROPOSED METHOD

A. Overview

In this paper, we mainly solve three sub-problems. The first one is how to build the spatio-temporal model under a camera network. The second one is how to use the specific spatio-temporal model to extract the possible trajectories of different persons. The third one is how to retrieve the trajectories according to the query. The spatio-temporal relationship between cameras can be defined as a function \( \psi(I_{c,t}, I_{c',t'}) \), where \( I_{c,t} \) represents the person image shot by the \( c \)-th camera at time \( t \). \( \psi \) is a function that outputs the probability of \( I_{c,t} \) and \( I_{c',t'} \) belonging to the same trajectory. We need to specify the spatio-temporal model \( \psi \) by training samples between different cameras and predict spatio-temporal relationship of testing samples. For the trajectory generation model, the input includes the query image and a gallery set, and the output is a set of galleries. Each trajectory should satisfy two constraints: similarity constraint and spatio-temporal constraint. Similarity constraint means that the query and the images of the generated trajectory should be similar. The spatio-temporal constraint means that the generated trajectory should satisfy the spatio-temporal relationship between cameras. Note that we allow the camera nodes in the gallery to appear in multiple possible trajectories, as long as these trajectories can meet the similarity and spatio-temporal constraints. Then, the goal of the person retrieval with trajectory generation is to retrieve all possible trajectories of the query person. The overall framework of our method is shown in Fig. 4. In short, our method mainly includes the following steps:

1) Constructing a cross-camera spatio-temporal model.
2) Obtaining the query-relevant candidate tracklets from the gallery set according to pure visual cues.
3) Using the spatio-temporal model to obtain the spatio-temporal probability between different tracklets in the candidate set.
4) Using the conditional random field method to obtain a spatio-temporal consistent graph model.
5) Using restricted non-negative matrix factorization to infer the trajectories in the optimized graph model.
6) Using trajectory re-ranking technology to enhance pedestrian retrieval.

B. Spatio-Temporal Modeling

We propose a novel spatio-temporal model based on the path distances between camera pairs. We assume that the camera network in an area satisfies the spatio-temporal relation \( \psi(d_{ij}, \Delta t) \). The \( \psi(d_{ij}, \Delta t) \) is a two-dimensional probability distribution function, where \( d_{ij} \) is the path distance between camera \( c_i \) and \( c_j \), and \( \Delta t \) is the time length for persons to walk from camera \( c_i \) to \( c_j \). When the path distance between two cameras and the corresponding time length are input into the model, it returns a probability to indicate the likelihood that the person is moving continuously from one camera to another. Because we often do not have enough data for all camera pairs to compute the spatio-temporal probability distribution, we choose to establish a joint probability model of the camera network. Firstly, we train the spatio-temporal model using the camera pairs with enough data, and then get the spatio-temporal probability of all camera pairs by interpolation. We use Multi-layer Perceptron (MLP) to fit the spatio-temporal probability model \( \psi \). MLP is smooth and continuous, which satisfies the assumption that spatio-temporal model is a two-dimensional continuous probability distribution. Our model consists of three network layers: input, hidden, and output. The input layer includes the path distance \( d \) and time length \( \Delta t \). The output layer returns whether the input value meets the time constraint. Assuming that the probability of the final output is \( \hat{y}_i \) and the label is \( y_i \), we train the model with the cross entropy loss function, which is expressed by the mathematical formula as follows:

\[
Loss = \sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{1}
\]

where \( n \) is number of mini-batch.

The Fig. 5 illustrates the spatio-temporal fitting on PTD using different models, including DecisionTree, Adaboost, and MLP. As shown, the probability distribution obtained by MLP is more smooth, which is more in line with the actual situation of pedestrians walking. In the experiment section, we will verify that MLP indeed works better for the final pedestrian retrieval.

C. Extraction of Candidate Tracklet Set

We use visual similarity to obtain the candidate tracklet set. Hierarchical clustering(HC) is a common clustering method, which does not need to specify the number of cluster categories. We represent a set of tracklets under one camera by \( T^c = \{o^c_1, o^c_2, ..., o^c_{n^c}\}, c = 1, 2, ..., N \), where \( c \) represents the camera index and \( n^c \) represents the number of tracklets in camera \( c \). We use feature extraction model, such as Resnet50, to extract features for each tracklet in the gallery, which can be expressed as follows:

\[
f^c_i = \Omega(o^c_i) = \frac{\sum_{j=1}^{k} F(I^c_{ij})}{k} \tag{2}
\]
Fig. 3: Time distribution under different camera pairs in Spatio-Temporal Dataset. (a) The positive sample distribution in training set. (b) The positive sample distribution in testing set.

Hierarchical Clustering
Tracklet Candidate Set
Spatio-Temporal Association
Original Graph
Optimized Graph
Trajectories

Fig. 4: The framework of our method for trajectory generation.

Hierarchical
Clustering
Tracklet
Candidate Set
CRF
RNMF
Spatio-Temporal
Association
Original Graph
Optimized Graph
Trajectories

Fig. 5: Visualization of spatio-temporal fitting by different models. (a) DecisionTree. (b) Adaboost. (c) MLP. Their AUC scores are 0.934, 0.986, and 0.987, respectively, on the test set of PTD.

where $\Omega$ is the average function and $F$ is the image feature extract function. $f_{ij}^c$ represents a specific image in tracklet $i$ with index $j$ in camera $c$. By the feature extraction model, we can get the feature set of the gallery. For a single camera $c$, the features of $i$-th tracklet is extracted and averaged to get the corresponding feature vector $f_c^i$. Then, the hierarchical clustering algorithm is carried on the set $\omega = \{f_1^1, f_2^1, \ldots, f_{nk}^c\}$ to get the clustering result $\hat{\omega} = \{\omega_1, \omega_2, \ldots, \omega_m\}$, where $\omega_m = \{f_p^q | o_p^q \in \omega_m\}$ represents the $m$-th cluster nodes.

D. Cross-camera Trajectory Generation

In this section, we introduce our trajectory generation module, which can be divided into two stages, adjacency graph generation and trajectory optimization.

1) Adjacency graph generation: At this stage, we aims to get the spatiotemporal consistency between the members in the candidate set as much as possible. We take the candidate set generated by HC as the input. First, we define that each element $f_q^p$ in the candidate set $\omega_q$ is a vertex $v_q$ in the graph and all the vertices constitute the vertex set $V = \{v_1, v_2, \ldots, v_n\}$. Each vertex $v_i$ has two attributes $c_i$ and $t_i$, which indicate
the camera they belong to and the time they appear. Then, we combine the vertices in pairs, adding directed edges between each pair of vertices. We define the \( e_{ij} \) as the edge between \( v_i \) and \( v_j \). The direction of \( e_{ij} \) is determined by the time of \( v_i \) and \( v_j \). When the time \( t_i \) is prior to \( t_j \), the direction is from \( v_i \) to \( v_j \), otherwise from \( v_j \) to \( v_i \). All edges \( e_{ij} \) forms the set \( E \). Then we define a probability function \( Q(e_{ij}) \) whose inputs are edges \( e_{ij} \) in \( E \). The \( Q \) function can be expressed by the following formula:

\[
Q(e_{ij}) = \psi(d_{ij}, \Delta t_{ij})
\]  

(3)

where \( \psi \) is the spatio-temporal model introduced in Section [V-B]. The \( d_{ij} \) is the distance between \( e_i \) and \( e_j \), and \( \Delta t_{ij} \) is the time difference \( |t_i - t_j| \). After that, then we apply the conditional random field to obtain a spatial-temporal consistent trajectory model. The spatial-temporal association graph obtained from the previous step is defined as the graph \( G = \{V, E\} \), the graph obtained after conditional random field processing is represented by \( \hat{G} = \{V, \hat{E}\} \) and the total probability model of spatial-temporal consistency graph of the trajectory is defined as \( \Pi(V, E) \). Therefore, the method of obtaining spatial-temporal consistent graph can be expressed by the following formula:

\[
\hat{E} = \arg \max_E \Pi(V, E)
\]  

(4)

The total probability \( \Pi(V, E) \) describes the probability that the constructed graph satisfies spatial-temporal consistency under the condition that the edges are \( E \) and the nodes are \( V \). When the total probability \( \Pi(V, E) \) obtains the maximum value, the corresponding \( E \) is the solution \( \hat{E} \). Because the total probability formula is difficult to calculate, we solve the formula \( (4) \) by approximate inference. We assume that the probability of each pair of nodes satisfying spatial-temporal constraints is independent of each other, then the total probability formula can be expressed as:

\[
\Pi(V, E) \approx \prod_i \Lambda(e_{ij})
\]  

(5)

where \( \Lambda(e_{ij}) \) represents the probability of whether the edge satisfies spatial-temporal consistency, \( e_{ij} \) represents vertex \( i \) and vertex \( j \). We use the mean field method to solve the \( \Lambda(e_{ij}) \). By using the mean field method, the \( \Lambda(e_{ij}) \) can be calculated iteratively by the following formula:

\[
\Lambda^{T+1}(e_{ij}) = \frac{e^{\left[\rho_1(\Lambda^T(e_{ij}) - t_1) + \rho_2(\Upsilon(v_i, v_j) - t_2)\right]}}{Z}
\]  

(6)

where \( \rho_1, \rho_2, t_1, t_2 \) are hyper-parameters, \( T \) represents the number of current iteration. For the first iteration, we initialize \( \Lambda^0(e_{ij}) = Q(e_{ij}) \). \( Z \) is the normalization term, and it can be calculated by the following formula:

\[
Z = e^{\left[\rho_1(1-t_1) + \rho_2(1-t_2)\right]}
\]  

(7)

\( \Upsilon(v_i, v_j) \) is the spatial-temporal consistent probability of vertex \( i \) and vertex \( j \), it can be calculated by following formula:

\[
\Upsilon(v_i, v_j) = \frac{\sum_k \Lambda^T(e_{ik}) \Lambda^T(e_{jk})}{\sqrt{\sum_k \Lambda^T(e_{ik})^2} \sqrt{\sum_k \Lambda^T(e_{jk})^2}}
\]  

(8)

After \( T \) iterations, we can get the \( \Lambda^T(E) \). \( \Lambda^T(E) \) is the set of all \( \Lambda^T(e_{ij}) \). Finally, we set a threshold \( \kappa \) and remove all the edges less than \( \kappa \) in \( \Lambda^T(E) \), and the remaining edges are set \( \hat{E} \).

2) Trajectory optimization: We model the problem of extracting pedestrian trajectory from graph \( \hat{G}(V, E) \) as the following mathematical problem: Given graph \( \hat{G}(V, E) \), we need to find one or more subsets of it, which satisfy the following rules:

1) Each subset represents a cross camera trajectory. The nodes in the track are sorted according to time, and the time before and after the nodes should meet a certain spatio-temporal relationship.
2) Each subset includes at least one node.
3) The union of all subsets is equal to \( V \).
4) A node from \( V \) could appear in multiple subsets and be shared by multiple possible trajectories.

Rules 1, 2 and 3 are easy to understand. Here we focus on rule 4: In practice, nodes should belong to only one trajectory, not multiple trajectories. The reason why we allow nodes to belong to multiple trajectories here is because sometimes it is more effective that retain all possible trajectories than forcing a solution without duplicate nodes. Moreover, the solution without duplicate nodes is included in the solution with duplicate nodes. In the following experiments, we will compare a variety of methods, including the method of obtaining repeated node solution and the method without repeated node solution to illustrate this situation. We choose restricted non-negative matrix factorization (RNMF) \( (25) \) to get the final pedestrian trajectory. Using RNMF has two advantages in this task. One is that it does not need to know the number of clusters in advance. The other is that compared with other algorithms, its performance is the best. First, we represent graph \( \hat{G}(V, E) \) as adjacency matrix \( \hat{S} \), its size is \( N \times K \). \( N \) is the number of nodes in set \( V \), and the element \( S_{ij} \) is weight of \( \hat{e}_{ij} \) from \( \hat{E} \). Given the adjacency matrix \( \hat{S} \), we need to find an assignment matrix \( \hat{A} \) to satisfy the following constraints:

\[
\hat{A} = \arg \min_A ||S - AA^T||
\]  

\[
s.t. A \in \{0, 1\}^{N \times K}
\]  

(9)

\[
A1_{K} = 1_{N}
\]  

(10)

where \( \hat{A} \) is the assignment matrix, which is a binary matrix and its size is \( N \times K \). \( K \) is the number of possible trajectories in \( S \). To obtain the value of \( K \), we calculate the eigenvalues of \( S \) and define the absolute values of all eigenvalues as \( \varsigma_1, \varsigma_2, ..., \varsigma_n \). We define a threshold \( \bar{\varsigma} \) and get the value of \( K \) through the following formula:

\[
K = \sum_i 1(\varsigma_i > \bar{\varsigma})
\]  

(10)

where \( 1(\varsigma_i > \bar{\varsigma}) \) is a function that when \( \varsigma_i > \bar{\varsigma} \) is true, it is 1, otherwise it is 0. The \( A \) indicates which trajectories each node should be assigned to. The \( 1_N \) and \( 1_K \) are defined as the column vector of \( N \times 1 \) and \( K \times 1 \) with value of 1. Under the condition of binary, equation \( (9) \) is difficult to solve, so
where we convert equation (9) into the following formula to solve it:
\[ A^* = \arg\min_{A' \geq 0} ||S - A'A^T|| + \tau ||A_{1k} - 1_N|| \quad (11) \]

where \( A' \) is the real version of \( A \), \( \tau \) is the hyper-parameter, and \( ||A_{1k} - 1_N|| \) is the regularization term. We solve this equation iteratively according to (25):
\[ A'^n = A' \odot \sqrt{(SA' + 2\tau 1_1K^T) \odot (4A'A^T A' + 2\tau 1_K 1_K^T)} \quad (12) \]

where \( \odot \) represents the division of matrix elements and \( \odot \) represents the multiplication of matrix elements. After obtaining the solution \( A'^n \), we traverse all rows of \( A'^n \) and select the column of the maximum value of each row as its corresponding trajectory.

E. Analysis of Conditional Random Field

In this section, we analyze the conditional random field module. The main question we are concerned about is the conditions under which \( \Lambda(e_{ij}) \) will increase and the conditions under which \( \Lambda(e_{ij}) \) will decrease in the CRF module. We define the spatio-temporal consistent gain function as follows:
\[ D(e_{ij}, T) = \Lambda^{T+1}(e_{ij}) - \Lambda^T(e_{ij} = 1) \quad (13) \]

Now consider the first iteration, bring formula (6) into formula (13), we can get the following formula:
\[ D(e_{ij}) = \exp(ln(ZQ(e_{ij}))) \times \left( e^{\rho_1(Q(e_{ij}) - t_1)} + \rho_2(T(v_i, e_j) - t_2) - ln(ZQ(e_{ij})) \right) - 1 \quad (14) \]

It is easy to see from formula (14) that the condition of \( D > 0 \) is equivalent to the following inequality:
\[ \rho_1(Q(e_{ij}) - t_1) + \rho_2(T(v_i, e_j) - t_2) - ln(ZQ(e_{ij})) > 0 \quad (15) \]

Bring formula (8) into formula (15), we can get the following inequality:
\[ \rho_1 e^{\rho_1(Q(e_{ij}) - t_1)} + \rho_2 e^{\rho_2(T(v_i, e_j) - t_2)} - ln(ZQ(e_{ij})) > \ln(Z) + \rho_1 t_1 + \rho_2 t_2 \quad (16) \]

In formula (16), we take its approximate expression for \( \frac{\rho_1}{\sqrt{\sum_k Q(e_{ik})^2}} \) to make it independent of \( Q(e_{ij}) \):
\[ \rho_1 e^{\rho_1 Q(e_{ij})} - \rho_2 e^{\rho_2 Q(e_{jk})} - \ln(Z) > \rho_1 t_1 + \rho_2 t_2 \quad (17) \]

In formula (17), we take their approximate expression for \( \rho_1 e^{\rho_1 Q(e_{ij})} \) and \( \rho_2 e^{\rho_2 Q(e_{jk})} \):
F. Trajectory Re-ranking for Person Retrieval

In the previous stage, we have generated multiple possible trajectories for each query. However, there may still be some problems for pedestrian retrieval by directly using these trajectories as results. On the one hand, each trajectory has multiple nodes. When there is an error in trajectory generation, these nodes may compose multiple persons of different identities. On the other hand, multiple trajectories may share one or more nodes, so persons between different trajectories may overlap. In order to reduce the above two problems, we use a flattening operation to enhance the person retrieval results. Firstly, we assume that \( p_1, p_2, p_3, \ldots, p_n \) is the result of person trajectory retrieval. Among them, the order of \( p_i \) is sorted by similarity with the query, with high similarity in the front and low similarity in the back. At the same time, all nodes in each trajectory are sorted by their similarity with the query. Then we connect all the trajectories to get a new sequence and traverse the new sequence from front to back. If the current node has not appeared before, it will be retained. Otherwise, it will be deleted.

V. EXPERIMENT

The experiments focus on investigating seven aspects: 1) Comparison of performance between our method with the traditional person retrieval methods; 2) Ablation study for trajectory generation; 3) Effectiveness of trajectory re-ranking; 4) Effectiveness of strategy of applying the spatio-temporal model; 5) Spatio-temporal robustness of the proposed method; 6) The generalization ability of the proposed method.

A. Evaluation Criteria

Besides evaluating the person retrieval or re-identification, the purpose of PTD is used for evaluating the spatio-temporal modeling and the trajectory generation. For the spatio-temporal model evaluation, the receiver operating characteristic curve (ROC), and the Area Under Curve (AUC) are used. For the person retrieval evaluation, we use three indicators to evaluate: Cumulative Matching Characteristics (CMC), mean Average Precision (mAP), and Trajectory Rank Score (TRS). For CMC and mAP, we use the standard settings, and for AUC, we use the standard settings on them to conduct experiments. Trajectory rank score (TRS) is proposed in this paper to measure the performance of trajectory generation algorithm in person retrieval. TRS is defined as follows. Suppose the input query feature vector is \( f_{in} \). In the gallery, the person trajectory with the same identity as query vector can be represented by \( S_g \), where \( S_g = \{S_{g1}, S_{g2}, \ldots, S_{gn}\} \) and each \( S_{gi} \) represents one trajectory. Each trajectory is composed of multiple tracklets, which can be expressed as \( S_{g1} = \{T_{g11}, T_{g12}, \ldots\} \), where \( T_{g11} \) represents a tracklet that belongs to \( S_{g1} \). We use the method proposed in Section IV to generate the person trajectory and sort them by the similarity with query. It is assumed that the output result can be expressed in \( S_{out} \), where \( S_{out} = \{S_{o1}, S_{o2}, \ldots, S_{on}\} \). Let \( T \) be an element in the candidate set, and define \( n(T) \) to represent the times of node \( T \) in all predicted trajectories, then the weight of node \( T \) is:

\[
w(T) = \frac{1}{N(T)}
\]  

(20)

Then we define the intersection score of the ground truth and the prediction as follows:

\[
|S_{gi} \cap S_{oj}| = \sum w(T_k)\bar{w}_{ik}
\]  

(21)

where \( \bar{w}_{ij} \) indicates that the identity of \( S_{gi} \) is the same as the identity of \( T_k \). Then we define union score:

\[
|S_{gi} \cup S_{oj}| = \sum \bar{w}_{ik}
\]  

(22)

Finally, we define an intersection ratio operation \( \sigma \) to calculate the similarity between the two trajectories:

\[
\sigma(S_{i1}, S_{j1}) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}
\]  

(23)

\[
TRS = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{1}{r_j} \sigma(S_{gi}, S_{oj})
\]  

(24)

where \( r_j \) represents the order of similarity of the \( S_{oj} \). Obviously, a higher TRS indicates a better method.

B. Experiment Setting

Our experiment mainly uses our data set Person Trajectory Dataset. We chose Resnet-50 as our backbone, from which we removed the last two layers to get the feature extraction model. First of all, we conduct supervised training on the training set of PTD, with the loss function of triplet loss, a learning rate of 0.0002, and a decay rate of 0.0005. The Adaptive Moment Estimation (ADAM) method is used to optimize the model. We randomly selected 64 identities in each mini-batch and four images for each identity. A total of 256 images are sent to the network for training, and we train 50 epochs in total, with the learning rate decreased to one-tenth in 20 and 40 periods. To train the spatio-temporal model, we use a multi-layer perceptron composed of a three-layer neural network to train. The input layer inputs the time difference and distance. The middle layer contains 100 leaf nodes. The output layer produces the probability of the data belonging to a particular trajectory using a specific set of inputs.

C. Person Retrieval Performance

In this section, we experiment with person retrieval strategies using different information. This experiment will verify that using more additional information, such as association under a single camera and cross camera trajectory information, can significantly improve the performance of pedestrian retrieval. According to the different information used in the retrieval process, we divide the person retrieval into three types, image-based retrieval, video tracklet-based person retrieval, and trajectory-based person retrieval. When query and gallery are both images, it is called image based method. When one of query or gallery is video, we call it video based method. When there is cross-camera trajectory in the gallery, we call it trajectory-based method. The experimental results are shown in Table IV. For the image-based method, we apply the most commonly used configuration in the Re-id task to the experiment. For the video based-person method, we use the image-based model to extract features from multiple frames.
TABLE I: Results of person retrieval using different matching patterns. Here “I”, “V”, and “T” indicate the “Image”, “Video”, and “Trajectory”, respectively. “Manual” and “Auto” indicate using manual or automatic methods for the trajectory generation, respectively. “*” indicates that the method utilizes the spatio-temporal information of query.

| Matching Pattern | mAP  | Rank-1 | Rank-5 | Rank-10 |
|------------------|------|--------|--------|---------|
| I-to-I           | 64.7 | 76.2   | 85.2   | 89.7    |
| V-to-V           | 85.6 | 84.5   | 94.8   | 96.9    |
| V-to-I           | 74.1 | 84.5   | 91.8   | 95.4    |
| I-to-V           | 78.2 | 76.2   | 89.4   | 93.5    |
| I-to-T(Manual)   | 83.7 | 79.5   | 92.0   | 95.7    |
| V-to-T(Manual)   | 89.9 | 86.8   | 96.4   | 98.5    |
| V-to-T(Auto)     | 89.1 | 90.2   | 94.8   | 95.9    |
| V-to-T(Auto)*    | 92.4 | 94.3   | 97.9   | 99.0    |

D. Ablation Experiment for Trajectory Generation

In this section, we investigate the different modules in trajectory generation. For the optimization stage of the trajectory generation module, we performed ablation experiments on the CRF module. For the trajectory optimization module, in addition to the NRMF module used in this paper, we compare its alternative Graph Search (GS) algorithm. The GS is a greedy algorithm. Considering the output result of the candidate set generation module graph \( G(V, E) \), we first define the start and end points of a trajectory. The starting point should have zero in-degree, while the ending point should have zero out-degree. By traversing the nodes in \( G \), we can get a set \( s \) of all vertices with in-degree 0 and we put them to set \( V_{\text{start}} \). Using the same method, the nodes with out-degree 0 are put to form the terminal set \( V_{\text{end}} \). Then we use the depth first search algorithm (DFS) to get final trajectories. We use \( p_1, p_2, ... , p_n \) to represent these trajectories and use \( f_{p_1}, f_{p_2}, ... , f_{p_n} \) to represent their features. Finally, we use the query feature \( f_{in} \) to compare with these trajectory features again to get their similarity. The final sorted result is the final result of person trajectory retrieval.

In terms of experimental parameter setting, for HC, we set the hyper parameter threshold to 0.07. When the conditional random field module is used in the experiment, we set the number of iterations \( T \) as 1, \( \rho_1 \) as 1.69, \( \rho_2 \) as 1.5, \( t_1 \) as 0, \( t_2 \) as 0, and \( \iota \) as 0.82, \( \kappa \) as 0.9. When we use the RNMF method to solve the trajectory, we set the number of iterations of the parameter to 200, \( \varepsilon \) to 0.5, and the \( \tau \) to 4.

In terms of quantitative comparison criteria, we select four indicators to compare the performance of different methods: TRS, mAP, Rank and ANOPN. The ANOPN represents the Average Number of Occurrences Per Node. It represents the sum of the number of nodes of all generated trajectories divided by the number of nodes in the gallery. The results are shown in Table [I] and Table [III]. In Table [I], we compare the results of conditional random field module with different spatio-temporal models. When using MLP spatio-temporal model, the conditional random field method can achieve the best results. The effect of AdaBoost and DecisionTree spatio-temporal model is not obvious after adding conditional random field, and the effect of MLP is obviously improved after adding conditional random field. In Table [III], we compared the cooperation degree between conditional random field and RNMF under the condition of using MLP spatio-temporal model. The baseline of our experiment is the result obtained in the V-to-V matching pattern without any post-processing. M0 adds clustering method on the baseline, does not use any spatio-temporal model, and directly takes the clustering results as the trajectory. Compared with baseline, M0 increased by 0.7% and 1.6% on mAP and Rank-1 respectively. This shows that the method of using the results of simple clustering as trajectories is effective. Compared with M0, M1 has better results after adding the graph search method. This indicates that there are incorrect clustering results in the results of clustering, and incorrect clustering results can be effectively excluded by adding the graph search method with spatio-temporal model. However, this also introduces an additional
null
strategy and the existing Bayesian strategy are not mutually exclusive but promote each other.

Next, we visualize our strategy. The visualization results are shown in Fig. 8. We selected an example from the test set to visualize the effectiveness of our method. It shows the results of visualization of query and its most similar top 10 in the gallery using T-SNE. It can be seen that the nearest sample to query is the wrong sample ID 424, and the correct sample should be ID 162. When we use the strategy proposed, we will establish the pedestrian trajectory in the gallery. It can be seen that the nearest sample ID 424 belongs to the trajectory with sample ID 138, and their center is farther away from the query than the sample ID 162, so the sample ID 162 becomes the top 1.

G. Spatio-Temporal Robustness

In this section, we study the influence of spatio-temporal noise on the method in this paper. Firstly, we analyze the source of spatio-temporal noise. We conclude that spatio-temporal noise mainly comes from the following ways: 1) the walking speed of different pedestrians is varying; 2) There are different paths between a pair of cameras; 3) We cannot judge whether a pedestrian is moving forward or doing something else when she is not in the camera. In order to study the influence of the above conditions on the method in this paper, we add different levels of random uniform noise from 10 minutes to 60 minutes at 10-minute intervals to the spatio-temporal data of the test set. The experimental results are obtained by averaging ten times in each experiment. The experimental results are shown in Fig. 9. It can be seen that after adding 60 minutes of random uniform noise, mAP only decreases by 0.6% and Rank-1 only decreases by 0.8%. Therefore, our method is robust to spatio-temporal noise in the camera network.

H. Generalization

Finally, we are concerned about whether the trajectory generation method and trajectory re-ranking method proposed in this paper can be applied to other feature extraction models besides Resnet-50. We use more person Re-id models for feature extraction, including the supervised ones, PCB [4] and MGN [5], and the unsupervised one, CAP [50]. The experimental results are shown in Table VI. As shown, the proposed trajectory retrieval strategy is effective on all feature extraction models. Compared with the Image-to-Image mode using only visual information, the trajectory retrieval method using our clustering and scattering strategy can greatly improve the retrieval performance. In Video-to-Video mode, without using feature re-ranking, the proposed trajectory retrieval strategy can improve the mAP by 1.4% and Rank-1 by 1.5% on the CAP model. In the supervised model Resnet50, MGN and PCB, the mAP has increased by 2.2%, 1.8%, and 1.2%, respectively. When using feature re-ranking, MGN achieves the best performance. Integrating trajectory re-ranking method to feature re-ranking method of MGN model can improve Rank-1 by 0.1 points. When using CAP, Resnet50, PCB model, the mAP has increased by 1.4%, 3.9%, 1.9%, and Rank-1 has increased by 1.5%, 4.2%, 2.1%, respectively.

We also draw the critical curves of different visual feature extraction models, as shown in Fig. 10. As can be seen from Fig. 10, the larger the positive gain area of the model with good


TABLE V: Comparison With Other Spatio-Temporal Strategies

| Matching Pattern | Method | Spatio-Temporal Strategy | Spatio-Temporal Model | Query Spatio -Temporal Information | mAP↑ | Rank-1↑ | TRS↑ |
|------------------|--------|--------------------------|----------------------|-----------------------------------|------|--------|------|
| V-to-V           | Baseline [22] | -            | Bayesian             | No                                | 56.1 | 67.7   | -    |
|                  | Wang et al. [24] | -            | Bayesian             | Yes                               | 75.8 | 86.4   | -    |
|                  | Wang et al. [24] | -            | Bayesian             | Yes                               | 64.7 | 76.2   | -    |
|                  | Xie et al. [27] | -            | Bayesian             | Yes                               | 59.3 | 73.0   | -    |
| V-to-V           | Ours Cluster+Scattering | MLP | Bayesian             | No                                | 81.6 | 92.8   | -    |
|                  | Wang et al. [24] | +Ours Cluster+Scattering | Bayesian             | No                                | 85.6 | 94.5   | -    |
| V-to-V           | ✓ ✓ | CAP [50] | 83.0 | 81.4 | 0.77 |
| V-to-V           | ✓ ✓ | MGN [5] | 95.4 | 94.3 | 0.90 |
| V-to-T           | ✓ ✓ | CAP [50] | 83.0 | 81.4 | 0.77 |
| V-to-T           | ✓ ✓ | MGN [5] | 95.4 | 94.3 | 0.90 |
| V-to-T           | ✓ ✓ | Resnet50 [48] | 89.5 | 89.7 | 0.83 |
| V-to-T           | ✓ ✓ | PCB [1] | 85.3 | 85.1 | 0.78 |

TABLE VI: Person retrieval results using different feature extraction models. Here “V” and “T” indicate video and trajectory, respectively. The “FR” and “TR” indicate feature re-ranking and trajectory re-ranking, respectively.

| Matching pattern | FR | TR | Feature Model | mAP↑ | Rank-1↑ | TRS↑ |
|------------------|----|----|---------------|------|--------|------|
| I-to-I           | -  | -  | CAP [50]      | 56.1 | 67.7   | -    |
| V-to-V           | -  | -  | MGN [5]       | 75.8 | 86.4   | -    |
| V-to-V           | +  | -  | Resnet50 [48] | 64.7 | 76.2   | -    |
| V-to-V           | ✓  | -  | PCB [1]       | 59.3 | 73.0   | -    |
| V-to-V           | ✓  | ✓  | CAP [50]      | 81.6 | 92.8   | -    |
| V-to-V           | ✓  | ✓  | MGN [5]       | 88.9 | 88.7   | -    |
| V-to-T           | ✓  | ✓  | CAP [50]      | 83.0 | 81.4   | 0.77 |
| V-to-T           | ✓  | ✓  | MGN [5]       | 94.8 | 94.3   | 0.90 |
| V-to-T           | ✓  | ✓  | Resnet50 [48] | 87.8 | 89.2   | 0.82 |
| V-to-T           | ✓  | ✓  | PCB [1]       | 84.9 | 84.5   | 0.78 |

visual feature extraction performance, the smaller the positive gain area of the model with poor visual feature extraction. When the visual feature extraction is accurate, there are fewer errors in the generation candidate set. The CRF module will relax the spatio-temporal establishment range of the same trajectory and make the nodes in the candidate set belong to the same trajectory as much as possible. When the visual feature extraction model is poor, there are more errors in the candidate set. The CRF module should use a more strict positive gain region to reduce the errors in the generated trajectory.

VI. CONCLUSION

Towards practical application scenarios, we propose a person retrieval framework based on cross-camera trajectory. In this framework, both visual features and non-visual cues of the camera network are exploited. Specifically, we propose a pedestrian trajectory generation method, which can find the potential scattering patterns in the clustering results of visual feature space and then improve the accuracy of pedestrian retrieval results. In the trajectory generation method, we propose a cross-camera spatio-temporal model that integrates the walking habits of pedestrians and the path distribution between cameras and develop a scheme for trajectory generation based on the spatio-temporal model. We also propose a CRF module to optimize the spatio-temporal probability graph. Then we obtain the specific pedestrian trajectory information by solving the restricted non-negative matrix factorization method. Furthermore, a trajectory re-ranking method is used to enhance the person retrieval accuracy. To verify the proposed method, we build a cross-camera person trajectory dataset from a real scenario.

REFERENCES

[1] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, “Style normalization and restoration for generalizable person re-identification,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3143–3152.
[2] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, “Beyond human parts: Dual part-aligned representations for person re-identification,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3642–3651.
[3] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in European Conference on Computer Vision, 2018.
[5] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” ACM Multimedia, 2018.
[6] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by video ranking,” in European Conference on Computer Vision, 2014, pp. 688–703.
[7] N. McLaughlin, J. M. del Rincon, and P. Miller, “Recurrent convolutional network for video-based person re-identification,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1325–1334.
[8] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, “See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6776–6785.
[9] J. You, A. Wu, X. Li, and W.-S. Zheng, “Top-push video-based person re-identification,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1345–1353.
[10] L. Wu, Y. Wang, J. Gao, and X. Li, “Where-and-when to look: Deep siamese attention networks for video-based person re-identification,” IEEE Transactions on Multimedia, vol. 21, no. 6, pp. 1412–1424, 2019.
[11] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, “Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1169–1178.
[12] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “Vrsc: Occlusion-free video person re-identification,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7183–7192.
[13] W. Ruan, C. Liang, Y. Yu, Z. Wang, W. Liu, J. Chen, and J. Ma, “Correlation discrepancy insight network for video re-identification,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 16, no. 4, pp. 1–21, 2020.
[14] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, “Illumination-adaptive person re-identification,” IEEE Transactions on Multimedia, vol. 22, no. 12, pp. 3064–3074, 2020.

[15] Z. Cheng, Q. Dong, S. Gong, and X. Zhu, “Inter-task association critic for cross-resolution person re-identification,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2605–2615.

[16] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, “High-order information matters: Learning relation and topology for occluded person re-identification,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6449–6458.

[17] P. Gao, J. Wang, H. Lu, and Z. Liu, “Pose-guided visible part matching for occluded person reid,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11 744–11 752.

[18] C. Zhao, X. Lv, S. Dou, S. Zhang, J. Wu, and L. Wang, “Incremental generative occlusion adversarial suppression network for person reid,” IEEE Transactions on Image Processing, vol. 30, pp. 4212–4224, 2021.

[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116–1124.

[20] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” CoRR, vol. abs/1609.01775, 2016. [Online]. Available: http://arxiv.org/abs/1609.01775

[21] Q. Xie, W. Zhou, G.-J. Qi, Q. Tian, and H. Li, “Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization,” IEEE Transactions on Multimedia, pp. 1–1, 2020.

[22] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang, “Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations,” in MMM 2016 Proceedings, Part I, of the 22nd International Conference on MultiMedia Modeling - Volume 9526, 2016, pp. 174–186.

[23] J. Lv, W. Chen, Q. Li, and C. Yang, “Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7948–7956.

[24] G. Wang, J. Lai, P. Huang, and X. Xie, “Spatial-temporal person re-identification,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 8933–8940, 2019.

[25] Y. He, X. Wei, X. Hong, W. Shi, and Y. Gong, “Multi-target multi-camera tracking by tracklet-to-target assignment,” IEEE Transactions on Image Processing, vol. 29, pp. 5191–5205, 2020.

[26] Y. Hou, L. Zheng, Z. Wang, and S. Wang, “Locality aware appearance metric for multi-target multi-camera tracking,” 2018.

[27] M. C. Liem and D. M. Gavrila, “Joint multi-person detection and tracking from overlapping cameras,” Computer Vision and Image Understanding, vol. 128, pp. 36–50, 2014.

[28] G. Nie, A. Liu, Y. Su, H. Luan, Z. Yang, L. Cao, and R. Ji, “Single/cross-camera multiple-person tracking by graph matching,” Neurocomputing, vol. 139, pp. 220–232, 2014.

[29] R. Iguerainaiis, D. Merad, K. Aziz, and P. Drap, “People tracking in multi-camera systems: a review,” Multimedia Tools and Applications, vol. 78, no. 8, pp. 10773–10793, 2019.

[30] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, “Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views,” Computer Vision and Image Understanding, vol. 109, no. 2, pp. 146–162, 2008.

[31] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3652–3661.

[32] O. Chun, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, “Total recall: Automatic query expansion with a generative feature model for object retrieval,” in 2007 IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.

[33] H. Jegou, H. Harzallah, and C. Schmid, “A contextual dissimilarity measure for accurate and efficient image search,” in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[34] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. van Gool, “Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors,” in CVPR 2011, 2011, pp. 777–784.

[35] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, “Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3013–3020.

[36] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, “Query-adaptive late fusion for image search and person re-identification,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1741–1750.

[37] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, “Deep ranking for person re-identification via joint representation learning,” IEEE Transactions on Image Processing, vol. 25, no. 5, pp. 2353–2367, 2016.

[38] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, “Person re-identification ranking optimisation by discriminant context information analysis,” in 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1305–1313.

[39] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, “Discriminant context information analysis for post-ranking person re-identification,” IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1650–1665, 2017.

[40] W. Li, Y. Wu, M. Mukunoki, and M. Minoh, “Common-near-neighbor analysis for person re-identification,” in 2012 19th IEEE International Conference on Image Processing, 2012, pp. 1621–1624.

[41] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen, “Person re-identification with content and context re-ranking,” Multimedia Tools and Applications, vol. 74, no. 17, pp. 6989–7014, 2015.

[42] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “Mars: A video benchmark for large-scale person re-identification,” in European Conference on Computer Vision, 2016, pp. 868–884.

[43] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944.

[44] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3464–3468.

[45] J. Ferrynam and A. Shahrokni, “Pets2009: Dataset and challenge,” Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on, pp. 1–6, 2010.

[46] J. Xu, X. Liu, Y. Liu, and S.-C. Zhu, “Multi-view people tracking via hierarchical trajectory composition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4250–4265.

[47] F. Fleuret, J. Berczal, R. Lengagne, and P. Fua, “Multicamera people tracking with a probabilistic occupancy map,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 267–282, 2008.

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[49] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.

[50] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, “Camera-aware proxies for unsupervised person re-identification,” in AAAI, 2020, pp. 2764–2772.