This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/MMUL.2019.2954405,
IEEE Transactions on Neural Networks and Learning Systems

Logical Topology Inference via CPGCN Joint Optimizing with Pedestrian Re-id

Keyang Cheng, Member, IEEE, Qing Liu, Member, IEEE, Rabia Tahir, Member, IEEE Liangmin Wang, Member, IEEE and Maozhen Li, Member, IEEE

Abstract—With the rise of artificial intelligence, deep learning has become the main research method of pedestrian recognition re-identification(re-id). However, most of the existing researches usually just determine the retrieval order based on the geographical location of cameras, which ignore the spatio-temporal logic characteristics of pedestrian flow. Furthermore, most of these methods rely on common object detection to detect and match pedestrians directly, which will separate the logical connection between videos from different cameras. In this research, a novel pedestrian re-identification model assisted by logical topological inference is proposed, which includes:(1) A joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference which makes the multi-camera logical topology provide the retrieval order and the confidence for re-identification. And meanwhile, the results of pedestrian re-identification as a feedback modify logical topological inference. (2) A dynamic spatio-temporal information driving logical topology inference method via conditional probability graph convolution(CPGCN) with forest-based transition activation mechanism(RF-TAM) is proposed, which focuses on the pedestrian's walking direction at different moments. (3) A pedestrian group cluster graph convolution network(GC-GCN) is designed to measure the correlation between embedded pedestrian features. Some experimental analysis and real scene experiments on datasets CUHK-SYSU, PRW, SLP and UJS-reID indicate that the designed model can achieve a better logical topology inference with accuracy of 87.3%, and achieve the top-1 accuracy of 77.4% and the mAP accuracy of 74.3% for pedestrian re-identification.

Index Terms—pedestrian re-identification, graph convolution network, logical topology inference.

I. Introduction

PEDESTRIAN re-identification(re-ID) is a research focusing on computer vision technology. It aims to establish identity correspondences across different cameras. Although many researchers have made explorations in the field of pedestrian re-ID and got some achievements, it is still very hard to do re-identify pedestrian in a large-scale surveillance system. Because there are serious external factors such as occlusion and illumination change which will reduce the robustness of pedestrian features.

Keyang Cheng was with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013 China e-mail: kycheng@ujs.edu.cn.

Qing Liu was with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013 China.

Rabia Tahir was with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013 China.

Liangmin Wang was with the Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, 212013 China.

Maozhen Li was with the Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, U.K.

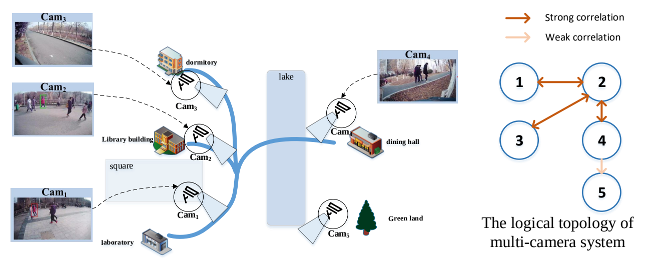

Fig. 1. The multi-cameras logical topology refers to the inherent pedestrian logical relationship among multi-cameras. The task of pedestrian re-identification is to search a specific pedestrian in non-overlapping cameras. The logical topology information between multi-cameras can effectively assist pedestrian re-identification.

For the pedestrian re-identification problem, it is usually considered as a metric learning task [3]. In some methods, the pedestrian in the pedestrian gallery are simply matched, and the error of the target pedestrian in the complex multi-camera scene is ignored. In other words, the logical topology information between multi-camera network is ignored. In fact, as shown in Fig. 1, the logical topology information between multi-cameras can effectively assist pedestrian re-ID. The multi-camera logical topology is a representation of inherent pedestrian-based spatial-temporal correlation between multi-cameras and the task of pedestrian re-identification is to search a selected pedestrian in variable cameras.

To discover the spatial-temporal correlation between cameras, many researchers have proposed effective methods to predict the logical topology between multi-cameras. In the early years, some methods [24], [26] inferred the multi-camera topology via simple occurrence correlation between special events of people. Such methods need some prior knowledge which is related to the camera entering and exiting point, and make too much false matching. In recent years, some novel approaches [30], [38] inferring topology based on pedestrian appearance have been proposed. This type of methods have greatly improved the re-identification accuracy. Furthermore, some methods [8], [9] apply the multi-camera topology to pedestrian re-identification to optimize the retrieval order of the multi-camera. As a result, the efficiency of pedestrian re-identification is also improved.

Although the above work has make great efforts to address the mentioned problems, these research works totally have weaknesses as below: 1)Most of the work separate the logical topology inference of the non-overlapping cameras from re-identification. Although these methods obviously improve the results of re-identification in same ways, the effect of such methods will be greatly reduced in multi-camera scene. 2) Most of the existing camera logical topology inference methods model correlation based on the video sub-area where the pedestrian appeared or based on the pedestrian appearance. They ignore people's behavior tends changing dynamically over time throughout the day. The logical topology of the camera in different time periods should change dynamically.

To address these deficiencies, an novel pedestrian re-identification model assisted by logical topological inference is proposed in this paper. The relationship of the multi-camera videos is considered in the proposed model to provide retrieval order and confidence for pedestrian re-identification, not just based on the distance between the cameras. And then the results of pedestrian re-identification will update the score of the logical topological inference. And the group cluster graph convolution network(GC-GCN) is provided to measure the distance of a cluster of pedestrian features, and can obviously improve the accurate of pedestrian re-identification. In general, it can provide more efficient and accurate pedestrian re-identification directly to the surveillance video environment.

The main contributions of this work are as below:

1) A joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference is designed. The multi-camera logical topology provides the retrieval order and the confidence of pedestrian re-identification. Meanwhile, the results of re-identification as a feedback will modify logical topological inference. The logical topological structure of the camera system can be discovered precisely by the joint optimization mechanism, so that the results of pedestrian re-identification will be re-ranked according to the multi-camera logical topology, thereby the accuracy of pedestrian re-identification can be improved.
2) A dynamic spatio-temporal information driving logical topology inference method via conditional probability graph convolution(CPGCN) is proposed, which focuses on four view directions of one video at different times. A forest-based transition activation mechanism(RF-TAM) is proposed to measure the inherent spatial-temporal and causal relationships within and across non-overlapping multi-cameras.
3) A pedestrian group cluster graph convolution network(GC-GCN) is designed to measure the correlation between embedded pedestrian features in multi-camera system. In generally speaking, pedestrians tend to walk in groups among different cameras. Most members of a group appearing in the video of a camera will appear in the video of another logically related camera with a high probability. Therefore, a GC-GCN is designed to model this process, the group pedestrian matching is used to assist pedestrian re-identification with single pedestrian.

There is a simple summary of the relationship between the three contributions: First, the first item proposes a joint optimization mechanism to jointly learn pedestrian re-identification and logical topological inference, so that the two can promote each other. Secondly, within the joint optimization mechanism, the logical topology inference method driven by the dynamic spatiotemporal information described in the second item and the pedestrian re-identification method based on the GC-GCN described in the third item are proposed. That is, the second and third items are the two sub-components of the joint optimization mechanism in the first item about logical topological inference and pedestrian re-identification.

The structure of this paper is described as below: The research motivation and innovation of this paper are introduced in the first section(Sec. I), The existing research results in this research field are introduced in the second section(Sec. II), the proposed model of logical topology inference is described in the third section(Sec. III). The settings and the results of the experiments and the analyses of the proposed method are shown in the fourth section(Sec. IV). The last section(Sec. V) is a conclusion of this paper.

II. Related Work

A. Pedestrian re-identification

As a research hotspot in the field of artificial intelligence, pedestrian re-identification has been proposed for many years, and has numerous research results. Traditional pedestrian re-identification research treat the re-identification as a metric learning problem. Many researchers have proposed various pedestrian identification methods such as [6], [19], [21], [22], [32], [42], [45] based on metric learning. Most methods combine pedestrian attribute classification with ID classification to measure the distance of pedestrian features. Miraj et al. [1] use hypothesis transfer learning to measure the metric between multi-cameras. The main idea of this method is transfer the prior knowledge from the existing models and just using the original models and limited annotation dataset. Kaiwei et al. [49] propose a fully unsupervised method: HCT, which only uses the unannotated labels. They regard the samples with different labels as a cluster, and then combine a fixed number of clusters according to the cluster distance. Then, after all the clusters are merged, the pseudo label will be reset. Finally, the model is optimized by the triplet loss. The experiment shows that their unsupervised method reaches a good perform on the mission of re-identification.

In these few years, there are some researchers have begun to consider combining detection and re-identification

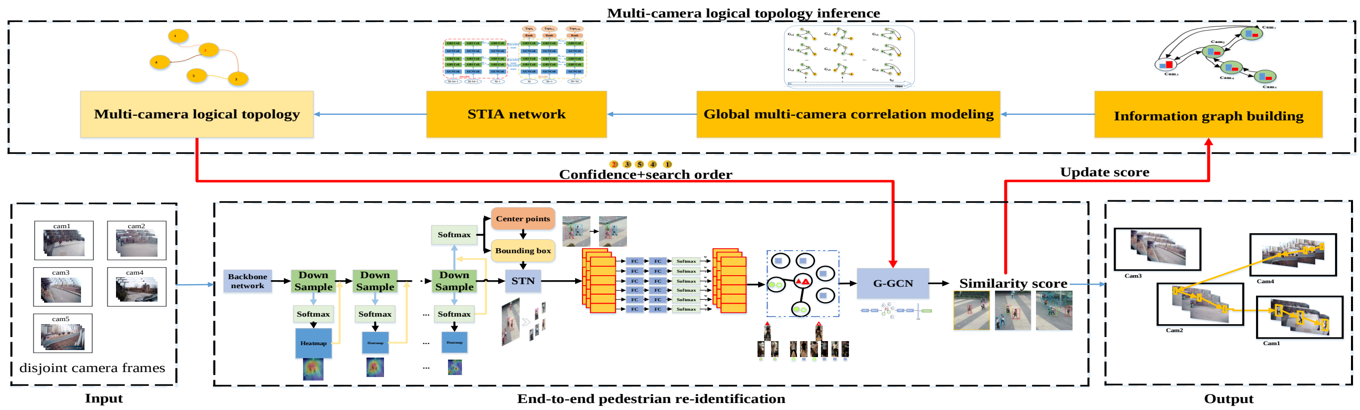

Fig. 2. A diagram illustrating the approach for the joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference.

into a whole identification problem named as pedestrian search. During these works, Xiao et al. [44] design the original person search network. Different from the conventional ID classifier, they design a pedestrian feature matrix to save the ID of each batch in the training process. Munjal et al. [28] propose a new query-based person search network which extracts the global information from the dataset and then outputs a query-based context re-identification score. Chen et al. [4] propose a feasible pedestrian re-identification model, which uses two separate CNN streams to extract foreground information and features patches, respectively. The method not only extracts robust features for each person ID but also considers the information complementarity of the background. Yao et al. [47] introduce an OR similarity indicator, which includes an objectness branch and a exclusion branch. The objectness branch can slow down the impact of noisy features, and improve the accuracy of person search in a way of ranking samples.

Although the existing methods have made great progress in accuracy, the multi-camera logical topology is not applyed to the re-ID task. Therefore, the existing methods are not suitable to the complex distributed multi-camera system directly.

B. Multi-camera logical topology inference

To discover the spatio-temporal correlations of multi-cameras, many researchers have tried to build multi-camera topology and camera coordinate. Some researches [2], [31], [34]–[37] directly define the logical topology of multi-camera. But in fact, in most cases, the logical topology of multi-camera is unknown. Thus, it is necessary to design a method to obtain the multi-camera logical topology. Many methods for estimating the logical topology of non-overlapping multi-camera system have been proposed. According to different research strategies, these methods can be sorted into two categories: one is unsupervised method of inherent pattern matching through a specified area in camera view and another is relying on calibrated cameras and inter-camera object tracking such as person tracking.

1) Inherent pattern matching with specified regions of camera view: Loy et al. [24] propose an unsupervised method which divides the camera view into multiple sub-areas, and then calculates the similarity within the corresponding sub-active area in view of the two cameras. His method successfully solves the problem that the camera needs to be calibrated, and achieves the purpose of calculating correlation by matching the patterns of multiple dynamic regions in camera view. However, this mode ignores the walking directions of pedestrians, and the resulting logical topology is the absolute positional relationship between the cameras. Li et al. [17] divide the camera view into some sub-areas and try to find the co-occurrence of sub-areas. This method achieves good results to a certain extent, but it is not suitable for distributed multi-camera system. Therefore, this method can not model the transition delay between cameras.

2) Relying on calibrated cameras and inter-camera person tracking: Javed et al. [34] firstly propose a model using the object tracking methods within the camera views to obtain correlations. Their method can determine the correlation between cameras through pedestrian trajectories, but this treat is easily affected by occlusion, camera orientation, and dynamic appearance of clothing. The problem of person tracking is still unsolved. Nam et al. [29] introduce a model to estimate multi-camera logical topology which based on the results of object tracking. Although their method is taken into account the change in orientation of pedestrians during walking, they still rely on the color model to identify and match pedestrians. Some proposed methods [15], [40] calculate correlation by building the transition time which is detected in different camera views. Makris et al. [27] define and observe the activity of person and build the multi-camera logical topology according to the activity between cameras. Their method can effectively avoid solving the correspondence problem. However, their method requires camera calibration in advance and is not suitable for

complex and large surveillance systems. In recent years, more advanced methods have been designed. Cho et al. [9] joint the pedestrian re-identification and multi-camera logical topology into a model for training. In their method, the random forest is applied to re-identify pedestrian in an unknown multi-camera system. However, the disadvantage is that the logical topology inferred by their method can not be updated dynamicly. As we all know, people's walking directions tends to change dynamically over time throughout the day. Therefore, the logical topology of the camera in different time periods should also change dynamically.

III. Method

In this section, the proposed joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference will be introduced in detail. In the method, the multi-camera logical topology provides the retrieval order and the confidence of pedestrian re-identification. Moreover, the results of pedestrian re-identification as a feedback will modify logical topological inference. After several iterative trainings, the inferred logical topology tends to be a stable state. It can provide accurate retrieval order and confidence for pedestrian re-identification. The Fig. 2 is the framework of the proposed mechanism. In order to describe the operation process of the whole model in details, we use an abstract formula 1 to express the mutual promotion between the pedestrian re-identification model and the logical topological structure inference model.

$$\begin{aligned} &\rho_{C_A,C_B}(target) = \\ &[argmin(||X^i_{C_A} - Y^j_{C_B}||)] * P(C_B|Par_G(C_B)) \end{aligned} \tag{1}$$

where $\rho_{C_A,C_B}(target)$ refers to the similarity of the pedestrian $target$ between cameras C_A and C_B. Assume that camera C_A is the camera where the target pedestrian first appears, and C_B is any camera in the multi-camera system. The formula $argmin||X^i_{C_A} - Y^j_{C_B}||$ is mainly solved by the deep learning model GC-GCN(Sec.III-B). Xi and Yj represents the $i-th$ pedestrian feature and the $j-th$ pedestrian feature in cameras C_A and C_B, respectively. $P(C_B|Par_G(C_B))$ represents the weight of camera C_B in the multi-camera system, and the weight is obtained through logical topology inference(Sec.III-A3).

A. Dynamic logical topological inference

We fully consider the pedestrian's walking direction and build the spatiotemporal relationship between cameras. Then, the logical topology between the cameras is inferred based on the spatiotemporal and causal relationship, so as to optimize pedestrian re-identification. In this subsection, a dynamic logical topological inference approach is proposed. During the approach, a random forest is utilized to establish the camera-to-camera transition distribution and a random forest-based transition activation mechanism(RF-TAM) is proposed to active the neighbor nodes with a probability. Then a spatiotemporal information aggregation(STIA) model is proposed to infer the dynamic logical topology. An diagram of our approach is given in Fig. 3.

Pedestrian walking trajectories usually pass through multiple cameras, so it is reasonable to count the occurrence of pedestrians among multiple cameras and establish correlations. The designed end-to-end pedestrian re-identification framework is applied to detect pedestrians. In order to model the walking directions of pedestrians well, we establish observation points as shown in Fig. 4 for the camera's field of view: defining the observation points O_1, O_2, O_3, O_4 along the directions of top, right, bottom and left of the view field. During pedestrian detection and re-identification, the entry and exit points of pedestrians at the sight observation points will be recorded. The pedestrian similarity distribution is supposed that is established between the two cameras Cam_A and Cam_B. Cam_A is marked as the source camera and Cam_B is marked as the target camera. The extracted pedestrian features are stored in the *s_gallery* and *t_gallery* respectively. *s_gallery* stores the pedestrian features of the source camera Cam_A, and *t_gallery* stores the pedestrian features of the target camera Cam_B.

1) Local camera-to-camera transition distribution establishment: Although there are many re-identification models improving the re-identification performance, in the complicated surveillance system, factors such as illumination and occlusion will do harm to the re-identification. When handling the task of a large number of pedestrians in surveillance system, we have tested several widely used classifiers such as convolutional neural networks(CNN), graph convolutional networks(GCN) and random forest(RF). Because of the strong interpretability of decision tree and random forest structure, we can clearly understand the optimization process of the model, so as to better adjust the performance of the model. The RF can minimize the recognition error, which caused by external factors on rough pedestrian re-identification task. Therefore the RF is employed in our application finally.

A set of pedestrian features in a camera captured by a object detection model is denoted as:

$$F^{C_A} = \left\{ (f^{C_A}_{i,j}, y_i) | 1 \le i \le N^{C_A}, 1 \le j \le M^{C_A}_i \right\} \tag{2}$$

where y_i is the annotation of pedestrian i, $M^{C_A}_i$ denotes the number of features of person i in camera C_A, and N^{C_A} is the total number of pedesstrians. Moreover, a random forest is trained according to the set F^{C_A}.

For a person j in the video captured by camera C_B, the transition distribution can be estimated by aggregating outputs as:

$$P^{C_A}(y|v^{C_B}_{j,l}) = \frac{1}{N} \sum_{t=1}^{N} P^{C_A}_t(y|v^{v_B}_{j,l}) \tag{3}$$

where P^{C_A} represents the probability of the decision tree, N represents the number of decision trees,v denotes the

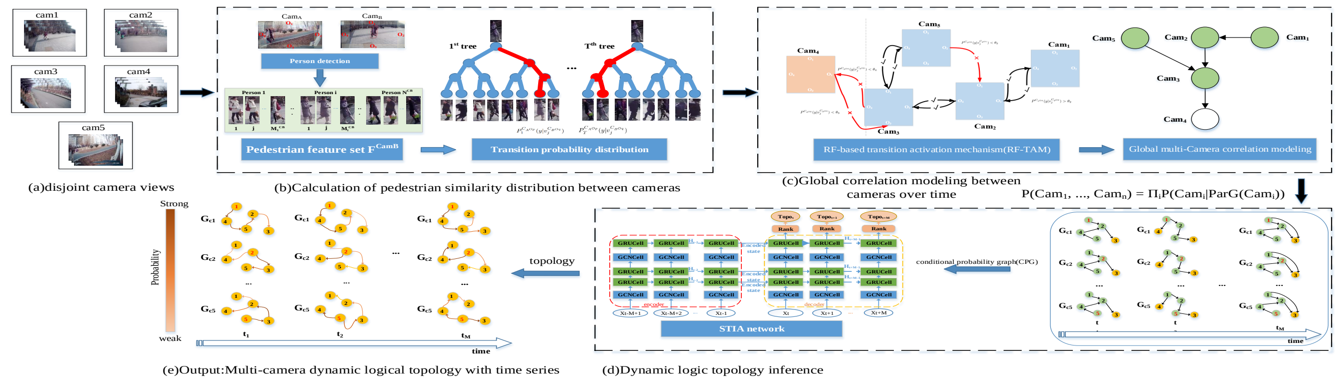

Fig. 3. An overview of the dynamic logical topological inference approach.

(a)Pedestrian enters camera view from observation point O_2 and exits camera view from observation point O_1 (b)Pedestrian enters camera view from observation point O_4 and exits camera view from observation point O_2

Fig. 4. Establish observation points in different directions for the camera's field of view.

features of pedestrian appearance, and $v_j^{C_A}$ denotes the k-th appearance of pedestrian i in camera C_A , respectively.

To achieve a multiple target re-identification result, we expand this case: the observation points are marked, multiple features of person j are extracted and the results are averaged as:

$$P^{C_{AO_p}}(y|v_j^{C_{BO_q}}) = \frac{1}{M_j^{C_{AO_p}}} \sum_{l=1}^{M_j^{C_{AO_p}}} P^{C_{AO_p}}(y|v_{j,l}^{C_{BO_q}}) \quad (4)$$

where $M_j^{C_A}$ is the number of features of the person j. C_{AO_p} and C_{BO_q} denote the observation point p in camera A and the observation point q in camera B, respectively. The whole transition establishment process is shown in Fig. 3(b).

2) RF-based node transition activation mechanism: In fact, not all the transition distribution is effective due to some missing and false matches. In this case, a RF-based transition activation mechanism(RF-TAM) is proposed. As shown in Fig. 3(c), the multi-camera topology can be formalized as a graph structure. Actually, it can be defined as $G(V, E)$, among then V denotes the cameras and E denotes the inter-camera transition distribution. In the graph, each camera node v has two status: active or inactive. An active node set is defined as S_0. Initially, S_0 only contains the camera node in which the pedestrian appear firstly. For each neighbor node B, it has a threshold θ_B defined as:

$$\theta_B = \frac{D_{in}}{D_{in} + D_{out}} \quad (5)$$

where D_{in} and D_{out} denote the in-degree and out-degree of node B respectively.

In the update process, each node in S_0 has an opportunity to activate neighbor nodes which in inactive status. The neighbor node B will be activated if the following conditions are met:

$$P^{C_{AO_p}}(y|v_j^{C_{BO_q}}) > \theta_B \quad (6)$$

The iterative update strategy as shown in the following steps:

- Step 1: Given the initial set of active nodes S_0, when node A is activated at time t, it has a chance to affect its neighbor node B. The condition for successfully activating the neighbor node is $P^{C_{AO_p}}(y|v_j^{C_{BO_q}}) > \theta_B$.
- Step 2: If B has multiple neighbor nodes that are all newly activated nodes, then these nodes will try to activate node B. If node A successfully activates node B, then at time $t+1$, node B will become active and be added to the set of active node set S_0.
- Step 3: At time $t+1$, the activated node B will have an impact on other neighboring nodes, that is, it will try to activate other neighboring nodes, and repeat the above process of Step 1 and Step 2.

All activation processes are independent. When no node can be activated, the activation process ends. Each node has only one chance to activate its own neighbor nodes. When the activation process is over, a final active nodes set S_0 is obtained. The overall process of RF-based transition activation mechanism(RF-TAM) is shown in algorithm 1. The computational complexity of RF-based transition activation mechanism(RF-TAM) is at the $O(n^2)$ level. RF-TAM is essentially an influence propagation maximization model. The fault tolerance of the model can be increased by maximizing the correlation of the node to its neighbor nodes.

3) Global multi-camera topology inference: In a multi-camera system, the state of one camera is not only affected by neighboring camera nodes, but also related to the state of the previous cameras. Therefore, conditional probability is used to model the global correlation of the camera and the joint distribution in the graph is solved by Bayesian

Algorithm 1: Framework of RF-based transition activation mechanism(RF-TAM)

Input: N:Camera node collection
Output: $G(N, W, A)$:a set of information graph with influence maximization

for $i \in N$ do
  for $j \in neighbour\_node(N_i)$ do
    Statistical pedestrian distribution:
    $P_i(X) = N_{\mu_u, \sum_u}(X)$,
    $P_j(Y) = N_{\mu_v, \sum_v}(Y)$;
    calculate information: $W_{i,j}(t, O_m, O_n)$;
  end
  Build information graph: $G(N, W)$;
  k=0,j=0,a set of activated nodes A;
  assign threshold $\theta_j = \frac{D_{in}}{D_{in}+D_{out}}$;
  while k=0 or $(A_k \neq A_k, k \geq 1)$ do
    $A_{k+1} = A_k$; inactive $= N - A_k$ for all $j \in inactive$ do
      if $\sum_{l \quad connected \quad toj, l \in A_i} W_{l,j} \geq \theta_j$ then
        activate j; $A_{k+1} = A_{k+1} \bigcup \{j\}$;
      end
    end
    i=i+1;
  end
  $G(N, W, A)$ append $G_i(N_i, W_i, A_i)$
end
Return $G(N, W, A)$

network(BN). The BN means a joint distribution via the chain rule for Bayesian networks:

$$BN = P(Cam_1, ..., Cam_n) = \Pi_i P(Cam_i | Par_G(Cam_i)) \tag{7}$$

where $Par(Cam_i)$ denotes the parent nodes of node i, and $Par(Cam_i) \in S_0$. For a specific node i, its conditional probability can be expressed as:

$$\begin{aligned} P(Cam_i | Par_G(Cam_i)) &= \frac{P(Cam_i \cup Par_G(Cam_i))}{P(Par_G(Cam_i))} \\ &= \frac{P(Par_G(Cam_i)|Cam_i) * P(Cam_i)}{P(Par_G(Cam_i))} \\ &= \frac{P(Par_G(Cam_i)|Cam_i) * P(Cam_i)}{\sum_{j=1}^{n} P(Cam_j) * P(P(Par_G(Cam_i)|Cam_j)} \end{aligned} \tag{8}$$

At last, the spatio-temporal information aggregation(STIA) model is proposed to infer the dynamic logical topology. The overall structure of the inference model can be shown in Fig. 3(d). The entire model is an encoder-decoder structure. For the encoder part, the time series of the first M moments are encoded as the initial hidden state in the decoder. And the state of the first M moments and the state of the current moment jointly predict the logical topology order of the future M moments in the decoder part.

In GCNCell, a Conditional Probability Graph Convolution Network(CPGCN) structure is used to aggregate spatial features. $G(C, A, CPG)$ represents this CPGCN structure. C represents the set of camera nodes in the multi-camera network, and $C = \{cam_1, cam_2, ..., cam_N\}$ while N denotes the number of camera nodes. A represents the adjacency matrix in the camera system. A is a binary matrix and $A \in R^{N*N}$. $A_{ij} = 1$ when A_{ij} is a connection between node i and j, otherwise $A_{ij} = 0$. CPG is a matrix with $N*N$ items, which represents the conditional probability of the corresponding camera in the information graph. The graph convolution process is described as follows:

$$F = GCN(A, CPG) = \sigma(\tilde{D}^{-\frac{1}{2}} * \tilde{A} * \tilde{D}^{-\frac{1}{2}} * CPG * W) \tag{9}$$

where $\tilde{A}$ denotes adjacency matrix with self circulation, which is represented as $\tilde{A} = A + I$, $\tilde{D}^{-\frac{1}{2}}$ denotes degree matrix, and CPG denotes the conditional probability. The whole hidden layer structure is expressed as:

$$h_v^{l+1} = ReLU(b^{(l)} + \frac{1}{N_v} \sum_{u \in N_v} h_u^{(l)} CPG^{(l)}) \tag{10}$$

B. Group correlation graph model for pedestrian re-identification

During this section, a backbone model is applied to extract low-dimensional features and then a designed multi-scale heatmap attention mechanism(MHAM) is used to improve the coordinate box of the pedestrians. The center point of features and coordinate box of the pedestrian are iteratively optimized in training process. Moreover, the pedestrian group cluster graph convolution netwotk(GC-GCN) is designed to calculate the distance of features. Fig. 5 shows the pedestrian detection and re-identification framework.

1) Multi-scale heatmap attention and offset optimization of pedestrian detection: As a backbone, the ResNet is applied to extract low-dimensional pedestrian features. Then a multi-scale heatmap attention mechanism(MHAM) is designed to calculate the weight on low-dimensional features and thus enhance its robustness. In an uncropped image, except for foreground information such as pedestrians, there are many noisy backgrounds with rich texture information. These noise information will have a great impact on the model. As a measure to solve this problem, the MHAM will add weights to the low-dimensional features and find the hot area of the pedestrian in the features. Finally, the weighted features will be sent into the detect head to do detect pedestrains.

Traditional object detection usually performs poorly on datasets with only pedestrians, mainly due to mismatches caused by similar appearances of pedestrians and inaccurate coordinates obtained through regression. In order to solve the problem of imbalanced classification within class, the detection head part of MHAM relies on the regression coordinates box as well as uses the center point of the pedestrian feature as an auxiliary recognition. The bounding box coordinate is calculated by optimizing the offset between the center of the features and the detection coordinate.

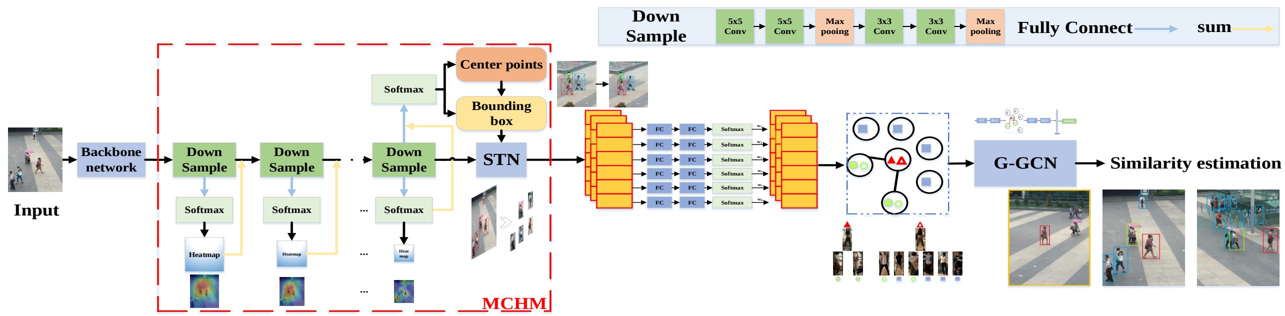

Fig. 5. An overview of the end-to-end pedestrian re-identification framework.

For a image, each pedestrian bounding box is defined as $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$. The pedestrian center point(c_x^i, c_y^i) is assigned as $c_x^i = \frac{x_1^i + x_2^i}{2}$ and $c_y^i = \frac{y_1^i + y_2^i}{2}$, respectively. (x_1^i, y_1^i) and (x_2^i, y_2^i) are the top left point and the down right point of the pedestrian coordinate, respectively. The position of the bounding box can be calculated by $(\tilde{c_x^i}, \tilde{c_y^i}) = (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$. Meanwhile, the heatmap of the position (x, y) can be defined as $M_{xy} = \sum_{i=1}^{N} exp^{-\frac{(x-\tilde{c_x^i})^2 + (y-\tilde{c_y^i})^2}{2\sigma_c^2}}$. Among them, N and σ_c is the number of pedestrians and the standard deviation, respectively. The model is trained with focal loss and in a form of pixel-wise regression. The training focal loss can be shown as below:

$$L_h = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{M}_{xy})^{\alpha} lg(\hat{M}_{xy}) & if \quad M_{xy} = 1;\\ (1-\hat{M}_{xy})^{\beta}(\hat{M}_{xy})^{\alpha} lg(1-\hat{M}_{xy}) & otherwise\end{cases} \tag{11}$$

where $\hat{M}$ denotes the heatmap of the image, and α, β are the parameters.

The size of the bounding box is defined as $\hat{S} \in R^{W*H*2}$ and the offset between bounding box and center point is defined as $\hat{O} = R^{W*H*2}$. Each ground truth of the image is assumed as $b^i = (x_1^i, y_1^i, x_2^i, y_2^i)$, then the size of the ground truth can be calculated by $s^i = (x_2^i - x_1^i, y_2^i - y_1^i)$. Furthermore, the ground truth offset can be obtained by $o^i = (\frac{c_x^i}{4}, \frac{c_y^i}{4}) - (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$. The output size and offset of the bounding box are defined as $\hat{S}^i$ and $\hat{O}^i$, respectively. Then l_1 loss is enforced for the two outputs:

$$L_{box} = \sum_{i=1}^{N} ||o^i - \hat{o}^i||_1 + ||s^i - \hat{s}^i||_1 \tag{12}$$

The clipped pedestrians can be obtained by a spatial transformer networks(STN) with the MHAM and the pedestrian head detection. Usually in the real world, when multiple pedestrians wear similar clothing, the appearance of a single pedestrian is very similar to other pedestrians, which has a great impact on pedestrian re-identification. So in this model, a group of pedestrian features is clustered and applied to calculate pedestrian similarity by the multi-block features. A positive feature pairs means that the pedestrians who appear on both query library and search gallery. In the task of re-identification, the distance between two features is used to judge whether they belong to the same ID or not. x_i^r, x_j^r are the $r-th$ block of pedestrian feature i and j. Finally, as shown in Fig. 5, the final similarity $dist(i,j)$ between pedestrian features can be defined as a weighted average of the similarities of different body parts as the below formula:

$$dist(i,j) = \sum_{r=1}^{R} w_r * d(x_i^r, x_j^r) \tag{13}$$

where d is the distance between x_i^r and x_j^r,usually the Euclidean distance is applied as the distance. R represents the number of body part and in our model the number is six. w_r is denoted as the optimized weight of the $r-th$ feature part of the pedestrian.

The features of different pedestrian body parts often have different contribute to pedestrian re-identification. This is mainly because the proportions of body parts are different and they are easily influenced by environmental factors such as occlusion and illumination. Thus, the model will output the weights w_r by a classifier which after the fully connected layers. For a pair of person ID(i,j), the training annotation y will be set to 1 if these two samples are the same pedestrian, otherwise $y = -1$. the model is trained and optimized according to the following formula:

$$L_{ID} = \begin{cases} 1 - dist(i,j) & y = 1\\ max(0, dist(i,j) + \beta) & y = -1\end{cases} \tag{14}$$

The formula takes a gap parameter β between positive samples and negative samples to enhance the discrminativeness of the pedestrian features.

2) Pedestrian group cluster graph convolution for pedestrian re-identification: For a pair of images$A$ and B, it is captured from two non-overlapping cameras. Based on daily experience, if a group of pedestrians appear in two images, the target pedestrian in the crowd will also appear in both images with a high probability. According to this assumption, the similarity of the crowd will be used to assist in the task of re-identification.

Assuming that group pedestrian features are defined as $(A_i, B_i), i \in \{1, .., K\}$, each group has K pedestrians. As shown in Fig. 6, the K groups of pedestrian feature pairs and the remaining single pedestrian features are

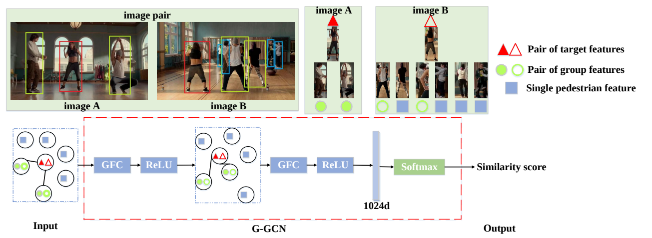

Fig. 6. The pipeline of the proposed model GC-GCN of re-identification.

abstracted into nodes, and the relationship between the nodes is formed to be a graph. The graph is defined as $G(V, E)$, where parameter V is pedestrian features, and E represents the relationship between the cameras. Each node in the graph is assigned with a pair of pedestrian features (X_{A_j}, X_{B_j}), and $j \in 0, ..., K$. In order to effectively spread and aggregate the information between graph nodes, the data in each node is calculated in the form of graph convolution. Assuming that the input of the graph convolution is $X \in R^{N*2d}$, where $N = K + 1$ and d denotes the pedestrian feature dimension. The parameter A is the adjacency matrix of graph convolution. $A_{i,j} = 1$ if the feature pairs are belong to the same people, otherwise $A_{i,j} = 0$. To simplify the processing of the model, the adjacency matrix A is normalized and can be regard as a feature stack of $\{A_1, ..., A_T\}$. Each A_t will be optimized symmetrically by the following forluma: $A_t = \Lambda_t^{-\frac{1}{2}} * \hat{A}_t * \Lambda_t^{-\frac{1}{2}}$, where $\hat{A}_t = A_t + I$ and Λ_t is the optimized degree matrix of $\hat{A}_t$. $\hat{A}$, Λ are the column of $\hat{A}_t$ and Λ_t,respectively. To maintain the structure of the pedestrian group features, a pedestrian group cluster graph convolution network(GC-GCN) is provided to aggregate node information and update the weights of nodes. As shown in Fig. 6, the propagation process of the GC-GCN is as below formula:

$$GCN(V^h, A)^{h+1} = RuLU(\Lambda^{-\frac{1}{2}} * \hat{A} * \Lambda^{-\frac{1}{2}} * V^{(h)} * W^{(h)}) \tag{15}$$

where $V^{(h)}$ denotes the output of the $h-th$ hidden layer features, $W^{(h)}$ denotes the optimizable weights and RuLU is the activation function applied in our model. A classifier is used at the end of the model for output

3) Iterative update strategy for joint pedestrian re-identification and topology inference training: For the entire model, the final result of re-identification is not solely dependent on the output of the GC-GCN model, but is affected by both GC-GCN and the logical topology of the multi-camera. For the output of GC-GCN model, the features will be reranked again according to the search order provided by the logical topology of multi-camera and then as a result of the re-identification.

In the training process of GCNCell, there have an initial logical topolog in the multi-camera system. The initial logical topology provides the initial weight W in $GCN(A, CPG)$. However, the initial weight is not accurate, because it is obtained by the rough pedestrian re-identification results. During the iterative training process of the strategy, the results of pedestrian re-identification will be re-ranked according to the logical topology, and pedestrian re-identification as a feedback will also update the features of logical topology inference model. Moreover, to comprehensively understand the process of the re-identification and logical topology inference, an optimized iterative update strategy can be expressed as below:

- Step 1: The end-to-end pedestrian re-identification framework is trained 10 epochs firstly. At this stage, the re-ranking of the topology search order will not be performed.
- Step 2: In the next 10 epochs, the pedestrian features will be sent to the random forest model and the similarity socre of every pair of pedestrian in pairs camera will be sent to the GCNCell as the initial weight W.
- Step 3: In each epoch in Step 2, there are 60 mini-batches for the STIA training. In each mini-batch, the weight of GCNCell will be updated iteratively. After iteration training with 60 mini-batches, the STIA model will be able to output a more reliable camera logical topology and the logical topology will be added into the GC-GCN model.
- Step 4: Repeat the above Step 2 and Step 3 until the camera logical topology converges or all training batches are completed.

When the iterative training process is over, the multi-camera logical topology can be inferred and the results of re-identification can be obtained well, moreover the results of re-identification can to be helpful the multi-camera logical topology inference. The overall process of iterative training is shown in algorithm 2. It can be seen from the pseudo-code structure that the computational complexity of our joint optimization mechanism is at the $O(n^2)$ level. The total loss function of the joint optimization mechanism can be expressed by the formula 16, where L_{STIA} is a common cross-entropy loss function and α, β, γ is assigned as 0.2, 0.2, 0.6, respectively.

$$L_{total} = \alpha L_{box} + \beta L_{ID} + \gamma L_{STIA} \tag{16}$$

IV. Experimental Analysis and Discussion

In this part, firstly, the dataset employed and the implementation of our approach will be introduced in details. Then some ablation studies and comparative experiments which include some quantitative and qualitative analysis of the method will be performed.

A. Datasets

1) SLP [8]: The SLP is a fully labeled large-scale pedestrian re-identification dataset with logical topology information of the multi-camera. The dataset contains a total of 2632 pedestrians and each pedestrian is fully labeled. There are a total of nine cameras in the dataset and the logical topology correlation between the cameras is also fully labeled.

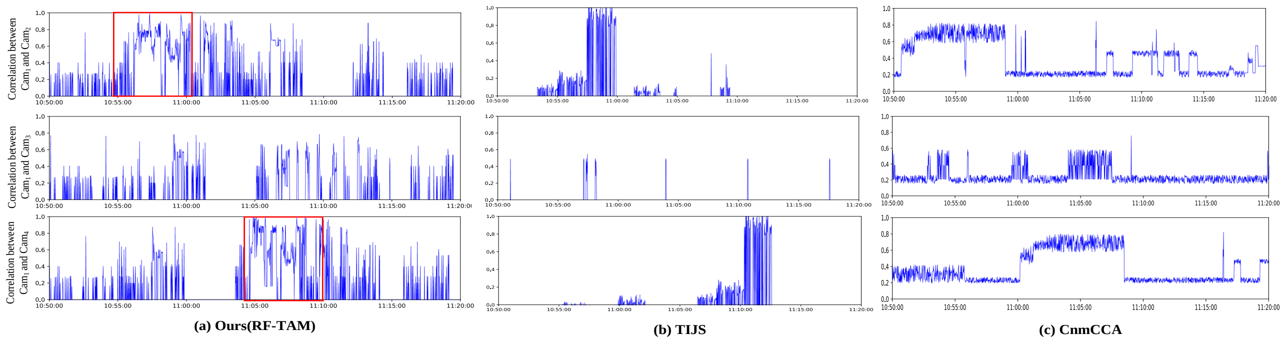

Fig. 7. Correlation analysis and comparison of inferred transition distributions among multiple cameras.

Algorithm 2: Process of joint pedestrian re-identification and logical topology inference training

---

```
Input: Video sequence data
Output: Multi-camera logical topology and
        pedestrian re-identification results
for Joint optimization mechanism training for 60
  epoches do
    if First 10 epoches then
        The end-to-end pedestrian re-identification
          framework in training.
    end
    else
        for 60 mini batches do
            The STIA in training.
        end
        The whole joint optimization mechanism is
          trained. The weight of GCNCell will be
          updated and multi-camera logical topology
          is inferred.
    end
end
Return Multi-camera logical topology and
  pedestrian re-identification results.
```

---

2) CUHK-SYSU [44]: This dataset is also a pedestrian re-identification dataset, and this dataset is more suitable for person search task. The images in the dataset are all with uncropped camera views. There are 8432 fully labeled pedestrian IDs and 96143 pedestrian bounding boxes in the dataset. The camera viewpoint, illumination, and occlusion are different among cameras in the dataset, and it is very close to a real surveillance system.
3) PRW [52]: Similar with the CUHK-SYSU, the PRW is also a person search dataset. It can be considered as an extension of the existing pedestrian re-identification dataset Market1501 [51]. Market1501 provides pedestrian bounding box information of each image.
4) Real scene dataset(UJS-reID): A UJS-reID dataset is collected in campus with non-overlapping camera views. The video is captured from multi-cameras with a frame rate of 15 FPS and is an enclosed area consisting of a laboratory, a student dormitory building, a cafeteria, and a library building. This is a typical scene on campus. With different times(8.30am.-9.30am., 10.30am.-11.30pm. and 4.30pm.-5.30pm.), the walking trajectory of students changes regularly. The physical topology of the scene is shown in Fig.1.

B. Qualitative and quantitative analysis of dynamic logical topology inference

1) Local camera-to-camera transition distribution establishment: The camera-to-camera transition distribution is measured by a random forest and then optimized by RF-TAM. To verify the effectiveness of RF-TAM. We compared RF-TAM with some advanced correlation analysis methods: TIJS [18] and CnmCCA [48]. The Fig. 7 illustrate that the designed model RF-TAM method can easily model the transition between two cameras, and match the correlation patterns between different camera pairs well. Although the TIJS [18] method can also calculate the transition, the internal pattern between the cameras are ignored, resulting in a poor final modeling effect. The CnmCCA [48] method can calculate the correlation between a single pair of cameras, but it cannot achieve correlation pattern matching between multiple pairs of cameras.

2) Global multi-camera logical topology inference: In a large-scale surveillance system, there is a causal relationship between each camera. The CPGCN is proposed to model the global correlation between all cameras and SITA network is used to infer the dynamic logical topology. The STIA network is implemented by the pyTorch deep learning framework and trained on two GPUs:Tesla P100*2. The initial learning rate of the model is 0.01 and it will reduced by 10 times every 10 epochs. The total number of training batches is 60 epochs.

To better verify this part, some ablation studies also performed in the CPGCN and STIA network. In Table I, the global correlation is modeled by the CPGCN and the commom GCN model, respectively. The CPGCN takes the conditional probability map as the input which contains the causal relationship between cameras while the common

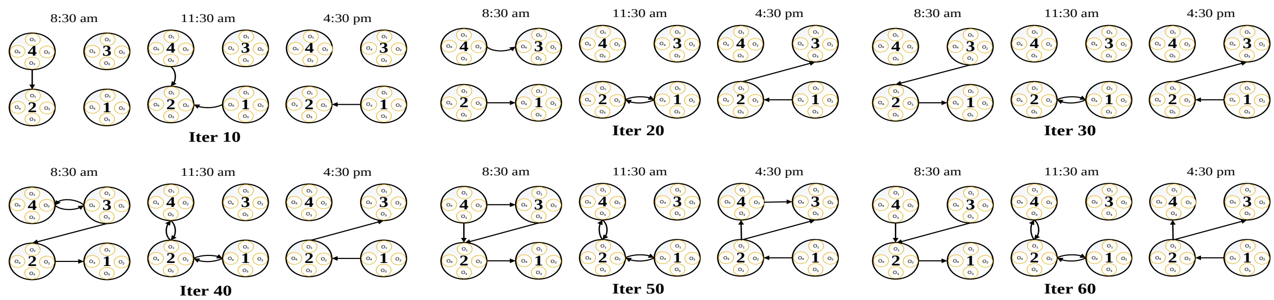

Fig. 8. During the iterative training process, the logical logical topology in non-overlapping multi-camera network changes dynamically. With the increase of iterative training, the logical topology structure sequence inferred is more closer to the real topology structure.

TABLE I
Ablation experiments of the CPGCN on real scene data.

Models	8.30am.	11.30am.	4.30pm.
LSTM+CPGCN	82.3	83	81.6
GLU+CPGCN	85.5	79	83
LSTM+GCN	68.5	772	65
GLU+GCN	79.5	83	82
GRU+GCN	81.5	76	73.4
GRU+CPGCN	88.5	87	85

TABLE II
A comparison of the cut distance between the inferred logical topology by various methods and the real logical topology.

	8:30am		11:30am		4:30pm		average
methods	edges	dist	edges	dist	edges	dist	dist
Actual	5	-	4	-	2	-	-
Distance-based	3	0.125	1	0.1875	3	0.0625	0.125
ODPR	2	0.5	2	0.125	1	0.0625	0.229
ours	3	0.125	4	0.0	3	0.0625	0.06

GCN just takes the adjacency matrix as the input. In addition, in order to eliminate the interference of other factors, LSTM and GRU modules are also added to the ablation experiment. It can be found from the experiment that the CPGCN including causality has a much higher accuracy than the common GCN model. Meanwhile, the GRU + CPGCN achieves the highest accuracy on the datasets of experiments.

The STIA network is applied to infer the dynamic logical topology inference and the CPG is employed as the input of STIA. In order to fully explain and demonstrate the performance and effect of the STIA model, we visualize the logical topology changes during the training process. The real time topology is visualized every ten epochs. The iterative training results can be shown in Fig. 8. From the experimental results, it can be known that as the training batches increases, the logical topology inferred is getting closer to the real one. Meanwhile, some more comparison experiments are conducted on the datasets SLP and UJS-reID as shown in TABLE. IV. Some methods [7], [11], [12], [33], [39], [53] are applied to the comparative experiments on dataset UJS-reID and SLP. The final inferred logical topology is shown in Fig. 10. To better measure the similarity between every two logical topologies, the similarity measurement method of isomorphic graphs is used to measure the similarity of two topological graphs. The similarity is in the form of cut distance [23], which is shown as:

$$dist(G_1, G_2) = max(\frac{|e_{G_1}(U,W) - e_{G_2}(U,W)|}{|V|^2}) \quad (17)$$

where G_1 and G_2 represent the two logical topology graphs respectively. G_1 and G_2 have the same node set V. U,W are any two subsets of the camera set V and $U, W \in V$. e_G is the number of edges between U and W in G. It is worth mentioning that the cut distance is more accurate for dense graphs. Therefore, this indicator can effectively measure the similarity between the logical topological structures of large multi-cameras. TABLE. II records the cut distance(the dist column of the table) between various methods and the real logical topology. The results show that the logical topology inferred by our method can fit the actual dynamic logical topology in different time. That means the proposed method in this paper is superior to existing methods. Furthermore, in order to describe the performance of the logical topology structure itself, we define the normalized cue distance accuracy to measure the accuracy performance of the logical topology structure:

$$ACC = 1 - \frac{dist(G_1, G_2) - dist_{min}(G_1, G_2)}{dist_{max}(G_1, G_2) - dist_{min}(G_1, G_2)} \quad (18)$$

where $dist(G_1, G_2)$ represents the cut distance between logical topology G_1 and G_2, $dist_{max}(G_1, G_2)$ and $dist_{min}(G_1, G_2)$ represent the minimum and maximum cut distance in the logical topology structure sample respectively. The Fig. 9 shows the interrelationship between logical topological inference model and re-ID model during the training process.

As can be found from the Fig. 11(a), a curve of accuracy is drawn to show the performance of the inferred multi-camera logical topology. Besides, the accuracy of the proposed pedestrian re-identification framework combained with multi-camera logical topology is presented in Fig. 11(b).

The TABLE III shows the performance of several methods in time cost. Among them, our method consumes

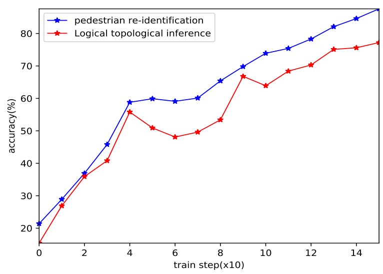

Fig. 9. Interrelationship between topological inference model and re-ID model.

TABLE III
A comparison of the time performance which retrieve the target for the first time in a camera.

	times(seconds)			
methods	8:30am	11:30am	4:30pm	average
Distance-based(error 25%) [7]	22	19	25	22.0
Distance-based(error 50%) [7]	25	26	28	26.3
ODPR [12]	32	28	30	30.0
ours	18	14	15	15.7

the least time under the condition of obtaining the same recognition results. The main reason is that we search the cameras according to the cameras order which provided by the logical topology. In this way, it can greatly reduce the retrieval time of empty cameras(the cameras without target pedestrian), thus the method can decrease the retrieval time of the unified multi-camera surveillance system. Moreover, we select the results of the 10th, 20th, 40th, and 60th epochs in the training process, and calculate the confusion matrix based on the recognition results. This can more intuitively explore the accuracy changes during the training process. The confusion matrix is shown in Fig. 12. The figure presents the errors with both the pedestrian re-ID predictions and the labeled person IDs. The re-ID predictions is obtained by the joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference model.From the Fig. 12, we can find that, with the increase of training iterations, the accuracy of pedestrian re-identification is getting higher and higher, and the predicted pedestrian ID is getting closer to the ground truth ID. When the difference between the re-ID results and the ground truth ID is the smallest, the accuracy of pedestrian re-identification reaches the highest, and the model converges, which means that the problem is well addressed by our joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference model.

C. Comparative and ablation experiments for pedestrian re-ID with GC-GCN

The MHAM is mainly used to obtain the region of interest of pedestrians in the non-cropped image, and increase the weight of the regional features, so as to improve the robustness of the features. In order to verify the effectiveness of the MHAM, a series of relevant comparative experiments as well as ablation experiments are performed. In some common object detection models, the bounding boxes are proposed by anchors, which can be called as anchor-based object detection. In addition, there are some other one stage object detection methods called anchor-free methods. In order to objectively reflect the ef-

TABLE IV
The comparative re-ID test results on datasets SLP and UJS-reID.

	SLP		UJS-reID	
methods	mAP(%)	R1(%)	mAP(%)	R1(%)
Db [7]	58.9	65.8	75.9	82.3
ODPR [12]	43.5	49.6	56.8	63.8
PCB [39]	47.3	48.1	54.1	60.0
IDE [53]	33.5	49.2	46.7	53.9
TriNet [11]	39.5	45.8	53.3	60.2
AWTL [33]	59.5	53.3	66.9	69.7
no topology	56.3	65.7	68.7	76.3
ours	63.4	68.5	78.0	85.1

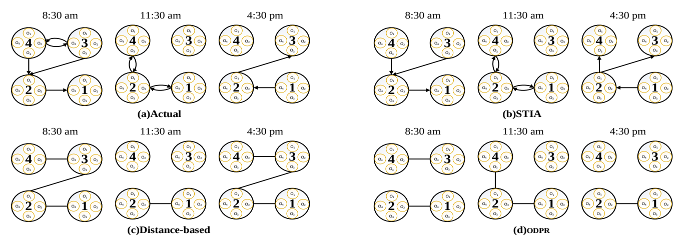

Fig. 10. Comparison of dynamic logical topological structures on real scene data inferred from different models.

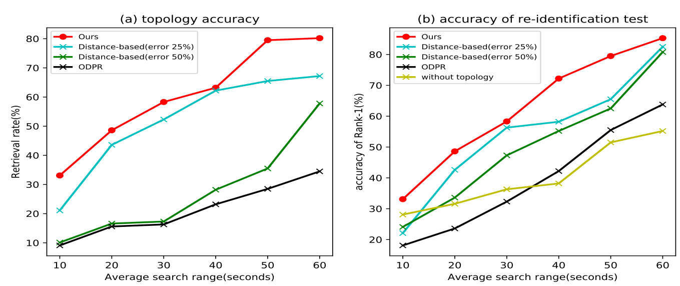

Fig. 11. The experimental results of logical topology and pedestrian re-identification accuracy change in multi-camera environment.

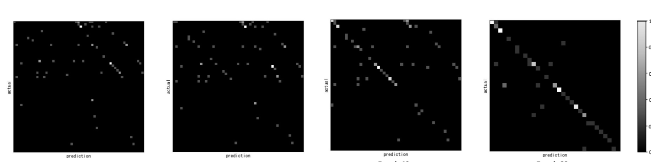

Fig. 12. The effectiveness of our proposed joint optimization mechanism of pedestrian re-identification and multi-camera logical topology inference. The confusion matrix is in the form of 93 $*$ 93 grids. Each grid indicates a person ID. Totally summing up 93 IDs, which approximates the number of person IDs in the full training set of the real scene dataset.

TABLE V
Comparative experiments of accuracy between the proposed MHAM and traditional object detection models.

Models	Accuracy(%)		
	AP	AP_{50}	AP_{60}
Faster-RCNN [25]	27.0	47.1	37.1
RGB-DCNN [54]	36.4	60.0	39.1
FB-SSD [20]	44.9	63.5	47.0
CornerNet [16]	47.3	63.7	53.7
CenterNet [10]	52.5	64.8	56.5
MHAM(ours)	57.1	70.3	57.4

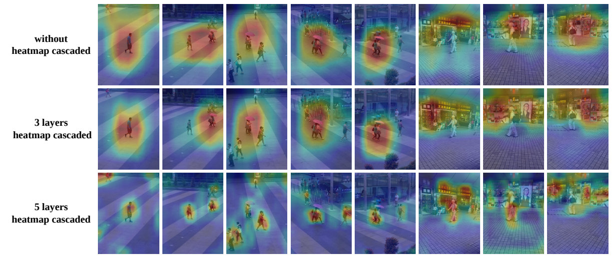

Fig. 13. Visualization of the effectiveness of the proposed method MHAM. As the number of heatmap layers added, the model focuses on the pedestrian area more concentrated. That is to say, the less background information the feature contains, the better it is for pedestrian detection.

fectiveness of the model, the proposed MHAM is compared with the above-mentioned two types of one-stage and two-stage common object detection methods respectively. The comparison experiments are conducted on the datasets CUHK-SYSU. Moreover, RGB-DCNN [54], Faster-RCNN [25] and FB-SSD [20] are adopted as the two stage object detection method, while CornerNet [16] and CenterNet [10] are adopted as the one stage method. As shown in Table. V, the experimental performance of the proposed MHAM is better than that of the common object detection method. Because the common object detection model are usually employed for a variety of objects. But for the scene with only single pedestrian object, its performance will suffer. That is to say, the common object detection model can not well distinguish the gap within the classification. The MHAM can effectively improve the discrimination of intra class gaps.

As for the ablation studies, the performance of MHAM is the most worthy of in-depth study. We decompose MHAM into several structures with different depths and conduct experiments separately. The 3-layer and 5-layer MHAM are regarded as the shallow attention structure and the deep attention structure, respectively. In addition, as a comparison, we also eliminate the MHAM structure and directly measure the accuracy of the original model structure. The Fig. 13 shows the performance of the ablation studies.As shown in the first row of the figure, the pedestrian area that the model focuses on is very rough. The area of interest contains too much background information, and it is difficult to get accurate location information of pedestrian in the whole image. This results in a particularly large deviation of the bounding box during pedestrian detection. After adding MHAM, the pedestrian attention area is significantly more concentrated, which indicates that the robustness of the model to extract features is significantly enhanced. Moreover, compared with the shallow attention structure(3-layer), the deep attention structure(5-layer) can more accurately focus on the pedestrians. In other words, compared with the original model structure, the MHAM structure can effectively improve the robustness of pedestrian features.

In traditional machine learning, pedestrian features are usually extracted manually, and then mathematical distance calculation formulas such as Euclidean distance are used to calculate similarity to determine whether they are the same pedestrian. However, the performance of manual features directly affects the distance calculation, and also directly affects the pedestrian re-identification. As deep learning method, the proposed GC-GCN comprehensively considers the similarity of group pedestrian features in the form of graph convolution, and outputs the similarity based on a group of pedestrians, thereby improving the accuracy of pedestrian re-identification. In this subsection, firstly, we make a comparison between some traditional manual feature extraction methods such as DSIFT [50], LOMO [13] and some related methods such as IAN [43] and Dis-GCN [14] which also using deep learning. The specific experimental data are shown in Fig. 14. It can be found from the results that the accuracy of the deep learning model is generally higher than the artificial feature extraction method. More quantitative experiment results are shown in Table VI. It is worth mentioning that due to the different collection scenarios of the CUHK-SYSU and PRW datasets, the pedestrians' dresses and postures in the obtained data are different. In addition, because the camera's shooting angle and light angle are different. The feature distribution of the data set is different, which has a greater impact on the robustness of the model to extract features, and ultimately leads to a greater difference in the accuracy of pedestrian re-recognition in different datasets. Finally, as described in Table VII, the proposed model have compared with the latest research methods of pedestrian re-recognition. We have compared with the SOTA method from the perspectives of the number of network layers, the amount of parameters, and the accuracy of the model. It can be concluded from the experimental results that although the accuracy of our model is not the best one, our model is better than other models with the same number of network layers. In addition, in the case of comparable accuracy, the number of parameters of our model is much reduced compared to other methods. It means that our model is more suitable for edge devices.

In the following part, the MHAM and GC-GCN are regard as a whole framework and some more ablation studies are performed to explore the performance of the framework. As shown in the Table VIII, different CNN backbone models and different feature extractor

TABLE VI
Comparative experiments of accuracy between the proposed pedestrian detection method and the traditional pedestrian feature extraction methods

		Models						
Accuracy(%)	Datasets	DSIFT [50]+Euclidean	DSIFT [50]+KISSME	LOMO [13]+XQDA	IAN(Res34) [43]	IAN(Res50) [43]	Dis-GCN [14]	MHAM(ours)
mAP	CUHK-SYSU	33.7	46.9	66.3	72.8	74.9	14.8	65.1
Rank-1		37.9	54.2	73.3	77.5	79.1	81.1	89.1
Rank-5		16.2	61.1	79.9	85.1	86.8	89.6	94.9
Rank-10		57.4	79.0	88.8	93.1	96.6	94.1	96.7
mAP	PRW	17.6	18.5	21.1	23.5	36.1	41.6	58.2
Rank-1		24.1	26.1	24.2	50.1	57.4	55.9	73.1
Rank-5		33.1	31.1	35.1	60.5	64.9	64.5	79.9
Rank-10		41.1	41.1	44.0	74.8	76.1	71.4	87.5

TABLE VII
Multi-dimensional comparative analysis results with some State-of-the-art methods.

Models	dataset	layers(estimated)	parameters(estimated, Mb)	mAP(%)	Rank-1(%)
BUFF [46]	PRW	100+	10M	44.4	82.4
TCTS [41]		100+	10M	46.8	87.5
NAE+ [5]		50+	5M	44.0	81.1
Ours		50+	2M	62.4	79.1

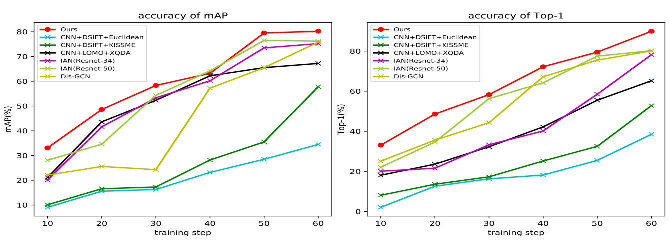

Fig. 14. Comparative experiments of different pedestrian re-identification methods.

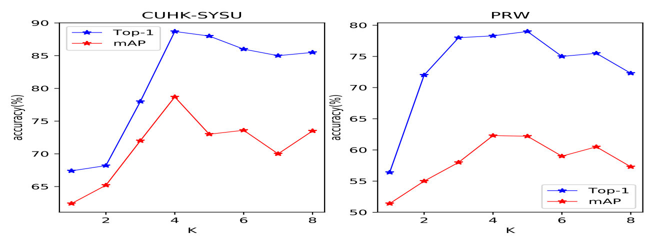

Fig. 15. Influence of different group number K on pedestrian re-identification accuracy.

TABLE VIII
Ablation studies of vary backbones and distance measurement methods on the re-ID dataset of CUHK-SYSU.

Models	distance	mAP(%)	Rank-1(%)
Res34+MHAM	GC-GCN	73	78.3
Res34+MHAM	Euclidean	58.1	63
Res34+MHAM	Cosine	56.3	59.8
Res50+MHAM	GC-GCN	78.2	88.7
Res50+MHAM	Euclidean	68.4	71.6
Res50+MHAM	Cosine	65.3	70.9

Fig. 16. Visualize the re-identification results of different numbers of pedestrian groups. The red bounding box in the middle is the selected pedestrian to be identified, the yellow bounding box is the pedestrians that appear in pairs around the target pedestrian, and the blue bounding box is the pedestrian that appears for the first time.

including GC-GCN and mathematical distance calculation formula such as Euclidean distance and consine distance are employed in the ablation studies. Moreover, we test the accuracy of pedestrian re-identification with different group K. Curves in Fig. 15 and Fig. 16 show that the value of K has a certain impact on the accuracy, and at the peak, it can be found that the proposed GC-GCN model can significantly improve the effect of pedestrian re-identification. This is reasonable, because in daily life, people usually walk in groups of four or five pedestrians, rarely more than five people in a group. In other words the proposed framework can give a more accuracy person search result in crowded scenes.

D. Real scene application

This section mainly describes the experimental results of the proposed model STIA in a real surveillance environment. We conducted a pedestrian search experiment on a set of surveillance video data on campus. We collect actual video data through multiple cameras. The dataset UJS-reID provided by this paper is captured at school by five non-overlapping video cameras with a frame rate of 15 FPS. The scene of the dataset is an enclosed area consisting of a laboratory, a student dormitory building,

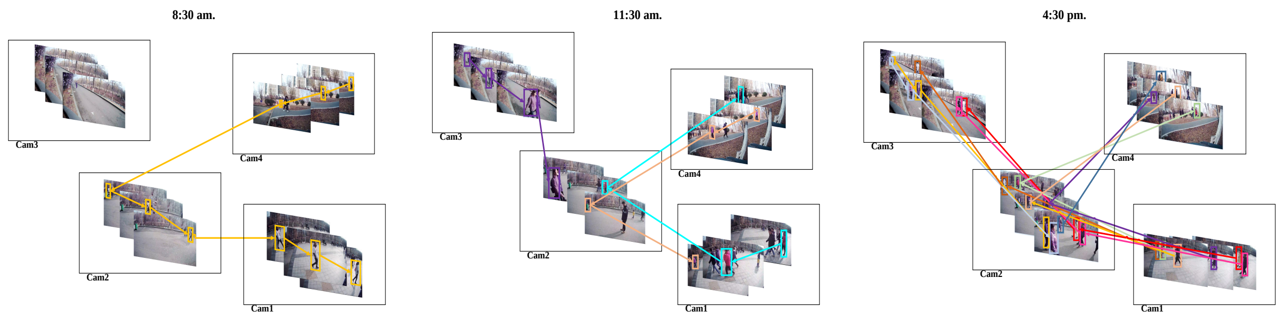

Fig. 17. Pedestrian walking trajectories on the real scene video at different times.

a cafeteria, and a library building. This is a typical scene on campus. At different times(8.30am.-9.30am., 10.30am.-11.30pm. and 4.30pm.-5.30pm.), the walking trajectory of students will change regularly. In this scene, in 8.30am.-9.30am., most students leave the cafeteria area to the teaching building area. In 10.30am.-11.30pm., many students walk from the teaching building to the cafeteria area. While in 4.30pm.-5.30pm., most students leave the teaching building area to the canteen area or the dormitory area, and almost no students walk from the dormitory area to the teaching building area. This is a typical application scenario on campus, and the logical topology between cameras that captured video in these areas also changes dynamically over time. This change is not only a change in the correlation between cameras, but also a change in the causality between two cameras with the same correlation. Based on this, we try to capture this logical structure and causality between multiple cameras, and use the logical topology to promote the optimization of the search sequence for pedestrian re-identification and the final recognition confidence. From the Fig. 17, it can be shown that the proposed method build a logical topology as $Cam_4 - Cam_2 - Cam_1$ in 8.30am.-9.30am, $Cam_1 - Cam_2 - Cam_4 - Cam_3$ in 10.30am.-11.30pm. and $Cam_1 - Cam_2 - Cam_3 - Cam_4$ in 4.30pm.-5.30pm, and the results of re-identification are improved significantly.

V. Conclusion

In this paper, we focus on the temporal and spatial relationship of pedestrians in video frames from different camera. And an novel pedestrian re-identification model assisted by logical topological inference is proposed. The multi-camera logical topology provides the retrieval order and the confidence of pedestrian re-identification. Meanwhile, the results of pedestrian re-identification as a feedback will modify logical topological inference. A dynamic spatio-temporal information driving logical topology inference method via conditional probability graph convolution is proposed. A time-delayed Jensen-Shannon divergence model is proposed to model causality in spatio and temporal within and across camera views. For two overlapping cameras, there is a time delay error between pedestrians passing through multiple cameras. And a pedestrian group cluster graph convolution network(GC-GCN) is provided to measure the distance of group features in multi-camera system. According to the determined logical topology information, when pedestrians walk between cameras which is logically associated, there will be a groups across cameras synchronously. Therefore, a GC-GCN is designed to model this process, so as to make full use of the group matching to enhance the re-ID of single pedestrian.

Acknowledgment

This research is supported by National Natural Science Foundation of China (61972183) and the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

References

[1] S. M. Ahmed, A. R. Lejbolle, R. Panda, and A. K. Roy-Chowdhury. Camera on-boarding for person re-identification using hypothesis transfer learning. In CVPR 2020: Computer Vision and Pattern Recognition, pages 12144–12153, 2020.

[2] Y. Cai and G. Medioni. Exploring context information for inter-camera multiple target tracking. In IEEE Winter Conference on Applications of Computer Vision, pages 761–768. IEEE, 2014.

[3] Z. Chang, Q. Zhou, M. Yu, S. Zheng, H. Yang, and T. P. Wu. Distribution context aware loss for person re-identification. In 2019 IEEE Visual Communications and Image Processing (VCIP), pages 1–4, 2019.

[4] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person search via a mask-guided two-stream cnn model. In Proceedings of the European Conference on Computer Vision (ECCV), pages 764–781, 2018.

[5] D. Chen, S. Zhang, J. Yang, and B. Schiele. Norm-aware embedding for efficient person search. pages 12615–12624, 2020.

[6] X. Chen, H. Sui, J. Fang, W. Feng, and M. Zhou. Vehicle re-identification using distance-based global and partial multi-regional feature learning. IEEE Transactions on Intelligent Transportation Systems, 22:1276–1286, 2021.

[7] Y. Cho and K. Yoon. Distance-based camera network topology inference for person re-identification. Pattern Recognition Letters, 125:220–227, 2019.

[8] Y.-J. Cho, S.-A. Kim, J.-H. Park, K. Lee, and K.-J. Yoon. Joint person re-identification and camera network topology inference in multiple cameras. Computer Vision and Image Understanding, 180:34–46, 2019.

[9] Y.-J. Cho, J.-H. Park, S.-A. Kim, K. Lee, and K.-J. Yoon. Unified framework for automated person re-identification and camera network topology inference in camera networks. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 2601–2607. IEEE, 2017.

[10] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. arXiv: Computer Vision and Pattern Recognition, 2019.

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. ArXiv, abs/1703.07737, 2017.
[12] N. Jiang, S. Bai, Y. Xu, C. Xing, Z. Zhou, and W. Wu. Online inter-camera trajectory association exploiting person re-identification and camera topology. In Proceedings of the 26th ACM international conference on Multimedia, pages 1457–1465, 2018.
[13] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. pages 2288–2295, 2012.
[14] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. C. H. Lee, B. Glocker, and D. Rueckert. Distance metric learning using graph convolutional networks: Application to functional brain networks. pages 469–477, 2017.
[15] P. Kumar and K. Dogancay. Analysis of brightness transfer function for matching targets across networked cameras. In DICTA, pages 250–255, 2011.
[16] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. pages 765–781, 2018.
[17] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In European conference on computer vision, pages 383–395. Springer, 2008.
[18] J. Li, P. Shang, and X. Zhang. Time series irreversibility analysis using jensen–shannon divergence calculated by permutation pattern. Nonlinear Dynamics, 96(4):2637–2652, 2019.
[19] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang. Hierarchical temporal modeling with mutual distance matching for video based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 31:503–511, 2021.
[20] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. IEEE Transactions on Circuits and Systems for Video Technology, pages 1–1, 2019.
[21] S. Lin, C. Li, and A. Kot. Multi-domain adversarial feature generalization for person re-identification. IEEE Transactions on Image Processing, 30:1596–1607, 2021.
[22] Q. Liu, K. Cheng, and B. Wu. Person search via anchor-free detection and part-based group feature similarity estimation. pages 242–254, 2020.
[23] L. Lovász. Large networks and graph limits, volume 60. American Mathematical Soc., 2012.
[24] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. International Journal of Computer Vision, 90(1):106–129, 2010.
[25] X. Mai, H. Zhang, X. Jia, and M. Q. H. Meng. Faster r-cnn with classifier fusion for automatic detection of small fruits. IEEE Transactions on Automation Science and Engineering, pages 1–15, 2020.
[26] D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In In Advanced Video and Signal Based Surveillance, pages 183–188. IEEE, 2003.
[27] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition., volume 2, pages II–II. IEEE, 2004.
[28] B. Munjal, S. Amin, F. Tombari, and F. Galasso. Query-guided end-to-end person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 811–820. IEEE, 2019.
[29] Y. Nam, S. Rho, and J. H. Park. Inference topology of distributed camera networks with multiple cameras. Multimedia tools and applications, 67(1):289–309, 2013.
[30] C. Niu and E. Grimson. Recovering non-overlapping network topology using far-field vehicle tracking data. In 18th International Conference on Pattern Recognition (ICPR'06), volume 4, pages 944–949, 2006.
[31] A. Rahimi, B. Dunagan, and T. Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 1, pages 187–194. IEEE, 2004.
[32] L. Ren, J. Lu, J. Feng, and J. Zhou. Uniform and variational deep learning for rgb-d object recognition and person re-identification. IEEE Transactions on Image Processing, 28(10):4970–4983, 2019.
[33] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6036–6046, 2018.
[34] M. Shah, O. Javed, K. Shafique, and Z. Rasheed. Tracking across multiple cameras with disjoint views, Nov. 11 2008. US Patent 7,450,735.
[35] X. Shu, G. J. Qi, J. Tang, and J. Wang. Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation. In the 23rd ACM international conference, 2015.
[36] X. Shu, J. Tang, G. J. Qi, W. Liu, and J. Yang. Hierarchical long short-term concurrent memory for human interaction recognition. 2018.
[37] X. Shu, L. Zhang, G. J. Qi, W. Liu, and J. Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99):1–1, 2021.
[38] C. Stauffer. Learning to track objects through unobserved regions. In 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05), volume 2, pages 96–102. IEEE, 2005.
[39] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. ArXiv, abs/1711.09349, 2018.
[40] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In Proceedings of the IEEE International Conference on Computer Vision., volume 2, pages 1842–1849. IEEE, 2005.
[41] C. Wang, B. Ma, H. Chang, S. Shan, and X. Chen. Tcts: A task-consistent two-stage framework for person search. pages 11952–11961, 2020.
[42] W. Wu, D. Tao, H. Li, Z. Yang, and J. Cheng. Deep features for person re-identification on metric learning. Pattern Recognit., 110:107424, 2021.
[43] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, and J. Feng. Ian: The individual aggregation network for person search. Pattern Recognition, 87:332–340, 2019.
[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3415–3424. IEEE, 2017.
[45] N. B. Yan Y, Zhang Q. Learning context graph for person search. pages 2158–2167, 2019.
[46] W. Yang, D. Li, X. Chen, and K. Huang. Bottom-up foreground-aware feature fusion for person search. pages 3404–3412, 2020.
[47] H. Yao and C. Xu. Joint person objectness and repulsion for person search. IEEE Transactions on Image Processing, 30:685–696, 2021.
[48] Y. Yuan, X. Shen, Y. Li, B. Li, J. Gou, J. Qiang, X. Zhang, and Q. Sun. Composite nonlinear multiset canonical correlation analysis for multiview feature learning and recognition. Concurrency and Computation: Practice and Experience, page e5476, 2019.
[49] K. Zeng, M. Ning, Y. Wang, and Y. Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In CVPR 2020: Computer Vision and Pattern Recognition, pages 13657–13665, 2020.
[50] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. pages 3586–3593, 2013.
[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. 2015 IEEE International Conference on Computer Vision (ICCV), pages 1116–1124, 2015.
[52] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1367–1376, 2017.
[53] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14:1 – 20, 2018.
[54] X. Zhu, C. Chen, B. Zheng, X. Yang, H. Gan, C. Zheng, A. Yang, L. Mao, and Y. Xue. Automatic recognition of lactating sow postures by refined two-stream rgb-d faster r-cnn. Biosystems Engineering, 189:116–132, 2020.