RESEARCH-ARTICLE

Graph based Spatial-temporal Fusion for Multi-modal Person Re-identification

YAOBIN ZHANG, South China University of Technology, Guangzhou, Guangdong, China

JIANMING LV, South China University of Technology, Guangzhou, Guangdong, China

CHEN LIU, Institute of Computing Technology Chinese Academy of Sciences, Beijing, Beijing, China

HONGMIN CAI, South China University of Technology, Guangzhou, Guangdong, China

Published: 26 October 2023

MM '23: The 31st ACM International Conference on Multimedia
October 29 - November 3, 2023
Ottawa ON, Canada

Conference Sponsors:
SIGMM

Graph based Spatial-temporal Fusion for Multi-modal Person Re-identification

Yaobin Zhang
202121045205@mail.scut.edu.cn
South China University of Technology
Guangzhou, China

Jianming Lv*
jmlv@scut.edu.cn
South China University of Technology
Guangzhou, China

Chen Liu
liuchen@ncut.edu.cn
Institute of Computing Technology, CAS
Beijing, China

Hongmin Cai*
hmcai@scut.edu.cn
South China University of Technology
Guangzhou, China

ABSTRACT

As a challenging task, unsupervised person re-identification (Re-ID) aims to optimize the pedestrian matching model based on the unlabeled image frames from surveillance videos. Recently, the fusion with the spatio-temporal clues of pedestrians have been proven effective to improve the performance of classification. However, most of these methods adopt some hard combination approaches by multiplying the visual scores with the spatio-temporal scores, which are sensitive to the noise caused by imprecise estimation of the spatio-temporal patterns in unlabeled datasets and limit the advantage of the fusion model. In this paper, we propose a Graph based Spatio-Temporal Fusion model for high-performance multi-modal person Re-ID, namely G-Fusion, to mitigate the impact of noise. In particular, we construct a graph of pedestrian images by selecting neighboring nodes based on the visual information and the transition time between cameras. Then we use a randomly initialized two-layer GraphSAGE model to obtain the multi-modal affinity matrix between images, and deploy the distillation learning to optimize the visual model by learning the affinity between the nodes. Finally, a graph-based multi-modal re-ranking method is deployed to make the decision in the testing phase for precise person Re-ID. Comprehensive experiments are conducted on two large-scale Re-ID datasets, and the results show that our method achieves a significant improvement of the performance while combined with SOTA unsupervised person Re-ID methods. Specifically, the mAP scores can reach 92.2%, and 80.4% on the Market-1501, and MSMT17 datasets respectively.

CCS CONCEPTS

• **Information systems** → *Specialized information retrieval.*

*Corresponding Author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613757

KEYWORDS

Unsupervised Person re-ID, Spatio-temporal, Graph, Re-ranking

ACM Reference Format:
Yaobin Zhang, Jianming Lv, Chen Liu, and Hongmin Cai. 2023. Graph based Spatial-temporal Fusion for Multi-modal Person Re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613757

1 INTRODUCTION

Person re-ID aims to capture the images containing the same person across different camera views [26, 41]. Recently, supervised person Re-ID algorithms [14, 29, 40] have made significant progress, but they rely on large and expensive annotated datasets. Consequently, an increasing number of researchers are turning to unsupervised Re-ID [5, 23, 27], which is more applicable to real-world scenarios with tremendous unlabeled image frames from surveillance videos.

Most existing unsupervised person Re-ID methods adopt pseudo label techniques [5, 18, 35], which apply clustering methods (such as DBSCAN [4]) on unlabeled data in the target domain to generate pseudo labels, resulting in significant performance gains. However, these methods heavily rely on carefully configured clustering hyper-parameters and require fine-tuning to achieve desirable results. Additionally, some methods use Generative Adversarial Networks (GAN) to reduce the domain shift caused by the diverse visual styles in different datasets [32, 42, 43].

Besides above pure visual Re-ID models, some multi-modal methods have been proposed to utilize the spatio-temporal pattern of pedestrians to further enhance the Re-ID performance. In particular, TFusion [21] adopted the Bayesian fusion to combine the spatio-temporal transition probabilities and visual scores, and used the fusion scores for evaluation. STCP [19] multiplied the spatio-temporal probability scores with visual scores to obtain the fusion scores for clustering and similarity judgement during the testing phase. JVTC [13] calculated the temporal consistency scores based on the time interval distribution between two cameras, and multiplied the scores with visual scores for model optimization.

All of above multi-modal methods conduct some hard combination strategies by multiplying the visual scores and the spatio-temporal scores, which are sensitive to the noise caused by the imprecise estimation of spatio-temporal patterns. Without the ground-truth labels of pedestrians, most of the methods estimate the spatio-temporal transition probabilities by analyzing the time interval of

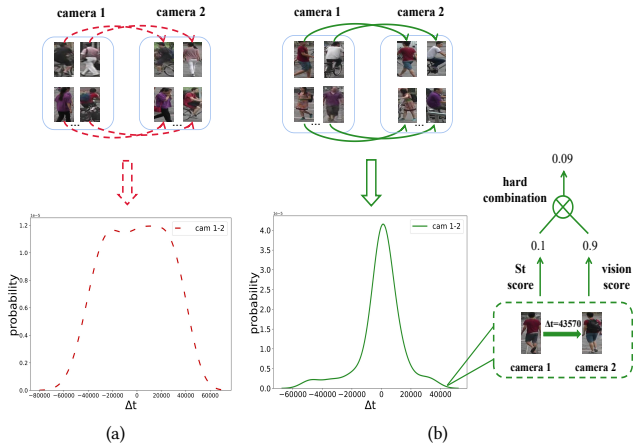

Figure 1: The spatio-temporal transition probability distribution is constructed on the Market-1501 dataset [39] using cross-camera image pairs of different types. (a) The distribution is constructed from visually similar image pairs. (b) The distribution is constructed from ground-truth person IDs.

the cross-camera images pairs containing the persons with similar appearance, which is judged by the visual model. For the scenarios with many visually similar negative samples, the model may learn incorrect cross-camera transition patterns, which result in wrong spatio-temporal scores as shown in Fig. 1. In particular, Fig. 1(a) shows the probability distribution of person transition from camera 1 to camera 2 based on the top 10 visually similar negative samples among all images in the Market-1501 dataset [39]. In contrast, Fig. 1 (b) depicts the ground-truth probability distribution of person transition. A significant difference can be observed between above two distributions, which may bring serious noise to the multiplication of visual scores and spatio-temporal scores.

Additionally, due to environmental factors and individual characteristics of person, transition between cameras is irregular for different persons. Therefore, the images of the persons with atypical transition time may receive much lower spatio-temporal scores. As shown in Fig. 1 (b), although the visual similarity of the positive image pair reaches 0.9, the model is highly likely to misidentify them as different persons after using the fusion of hard combination due to their low spatio-temporal transition probability of only 0.1. The hard combination with the noisy spatio-temporal information limits the advantage of the multi-modal fusion. In particular, the recently proposed pure visual unsupervised Re-ID methods [2, 30, 37] outperform the traditional multi-modal methods [13, 19, 21] with a large gap.

Based on the aforementioned analysis, we propose a Graph based Spatial-Temporal Fusion method, namely G-Fusion, which integrates spatial-temporal information softly to achieve significant improvement of Re-ID performance. By limiting the scope of the spatial-temporal information and only using it as a means of selecting neighbors in the graph, where a graph based aggregation method is conducted to learn structural affinities between images, the impact of the spatial-temporal noise is greatly reduced. Specifically, in the training phase, we select the cross-camera image pairs that are most likely to contain the same person based on visual similarity measurement and record their time intervals as a transition set for each pair of cameras. A graph of images is built by connecting the intra-camera image pairs with high visual similarity and the visually similar inter-camera image pairs which have the transition time within the fluctuation range of the records in the corresponding transition set. Then we use a randomly initialized two-layer GraphSAGE model [8] to obtain the affinity matrix between images, and utilize the distillation learning to make the model learn the relationships between nodes based on the graph. In the testing phase, a graph based multi-modal re-ranking method is proposed to make the final Re-ID decision.

The main contributions of this paper are as follows:

(1) We propose a Graph based Spatio-Temporal Fusion model for unsupervised person Re-ID, namely G-Fusion, which integrates the spatial-temporal information softly to reduce the impact of spatial-temporal noise and achieve significant improvement of performance.

(2) We apply a randomly initialized two-layer GraphSAGE model to obtain the multi-modal affinity matrix between images, and deploy the distillation learning to optimize the visual model by learning the affinity between the nodes. Meanwhile a graph-based multi-modal re-ranking method is proposed to enhance the precision of person Re-ID.

(3) We conduct comprehensive experiments on two commonly used Re-ID benchmark datasets, and the results show that G-Fusion can achieve superior performance with good scalability. More importantly, G-Fusion can be easily combined with other SOTA methods to significantly improve the overall performance.

2 RELATED WORK

This section discusses relevant literatures including unsupervised person Re-ID, graph-based person Re-ID, spatio-temporal fusion for supervised person Re-ID, and spatio-temporal fusion for unsupervised person Re-ID.

2.1 Unsupervised Person Re-ID

Most state-of-the-art unsupervised person Re-ID methods adopt pseudo-labeling techniques, which cluster and fine-tune on the target domain, and iteratively update the pseudo-labels and model parameters based on unlabeled data. In particular, MMT [6] proposed a synchronous average teaching framework for pseudo-label optimization, i.e., using more robust soft labels to perform online optimization on pseudo-labels. SpCL [7] treated each cluster and outlier as a single class, and then performs contrastive learning from instances to centroids. CAP [30] handled larger intra-personal differences caused by different cameras by generating proxies with camera awareness. ICE [2] improved previous class-level contrastive learning methods using pair-wise similarity scores between instances and used similarity scores as soft pseudo-labels to enhance consistency between the original view and the augmented views.

2.2 Spatio-Temporal Fusion for Person Re-ID

Due to the limitations of visual modality, many researchers have turned their attention to supervised multi-modal person Re-ID methods by fusing ground-truth spatio-temporal information. In

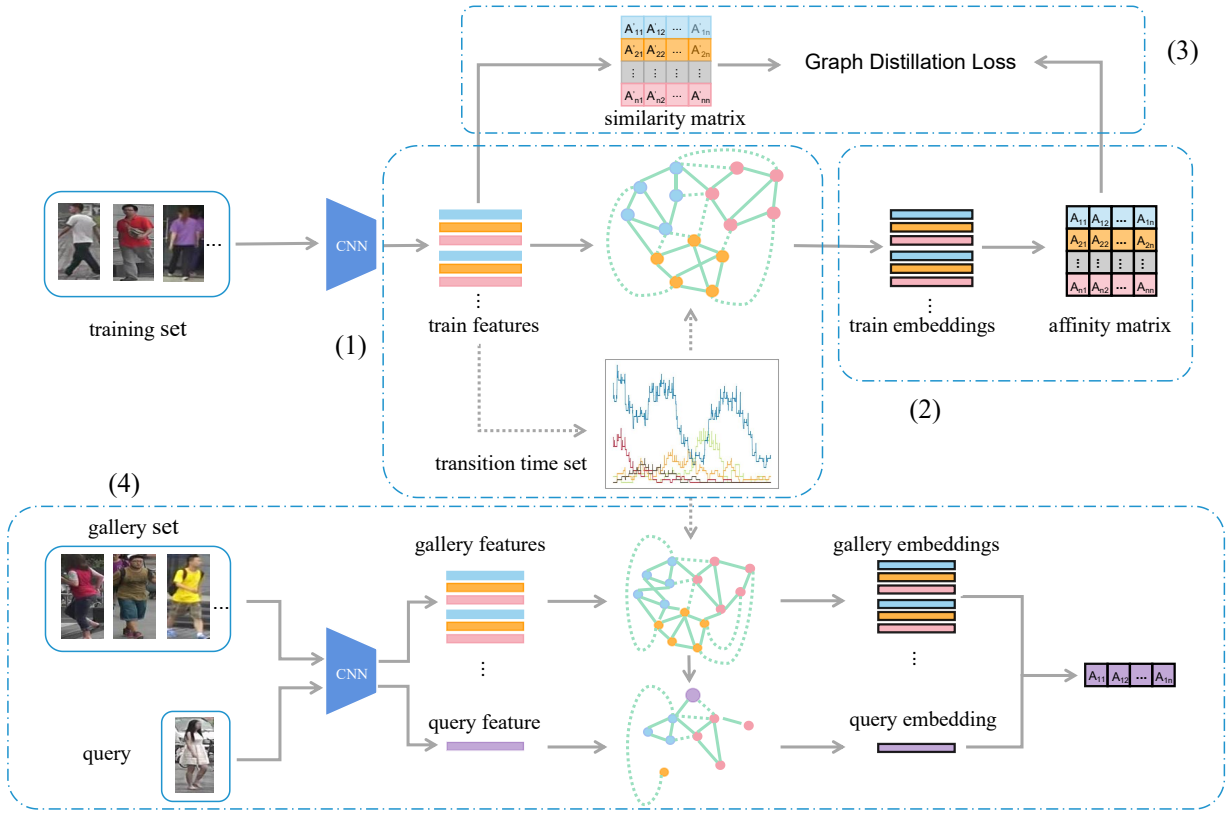

Figure 2: The Graph-based Spatio-Temporal Fusion framework, which consists of four main stages: (1) using spatio-temporal information to assist in selecting neighboring nodes and constructing a graph.(2) propagating node information and aggregating neighboring nodes to obtain the embedding of each node.(3) distilling the affinity matrix through learning to enable the model to learn the mutual relationships between nodes.(4) applying a graph based multi-modal re-ranking to reorder the test results.

particular, Wang et al. [28] proposed a novel dual-stream spatio-temporal framework for person Re-ID, which explored both visual semantic information and temporal information, and introduced a joint similarity measure using Logistic smoothing to combine the two types of information. Shu et al. [25] contributed a new large-scale spatio-temporal person Re-ID dataset (LaST) and proposed an effective baseline on this basis. Ren et al. [24] proposed an instance-level method to model the temporal and spatial patterns of person separately, and utilized the instance-level status information of each person to provide personalized predictions.

Above supervised multi-modal methods require real-world spatio-temporal data. However, in realistic scenarios, it is impossible to obtain the IDs for all person images, thereby preventing the acquisition of ground-truth spatio-temporal information. As a result, many researchers turned to the research of unsupervised multi-modal person Re-ID methods. In particular, Lv et al. [21] used Bayesian fusion to evaluate performance by combining visual and temporal modalities. Sridhar Raj S et al. [22] proposed a deep non-annotated method based on spatio-temporal correlation rules, which used visual features to incrementally cluster unlabelled person images and fine-tuned clustering by mining spatio-temporal correlation rules. Li et al. [13] calculated temporal consistency based on the time interval distribution between two cameras, fused it with visual scores to obtain a joint similarity score, and used this to optimize the model's clustering process.

2.3 Graph based Person Re-ID

As a widely used data structure, graph has been applied in various scenarios of person Re-ID. BMLC [36] proposed a multi-label prediction method based on the graph structure perspective, which predicted multiple labels by considering pairwise similarity and the distribution of neighboring nodes for each node. MGH [34] modeled the cross-camera heterogeneous data correlation using camera-perceiving hyper-edges, constructing a hyper-graph for feature learning and label refinement. HGA [38] proposed a coarse-to-fine heterogeneous graph alignment method by representing unlabeled data as heterogeneous graphs for each camera to find cross-camera person matching. HGO [17] proposed a heterogeneous graph driven optimization scheme, constructing a heterogeneous graph on the target domain and optimizing the model using heterogeneous affinity learning methods.

3 GRAPH BASED SPATIO-TEMPORAL FUSION

Fig.2 illustrates the overall framework of our proposed Graph-based Spatio-Temporal Fusion model, namely G-Fusion, which consists of

four main parts. Firstly, in the stage of **Spatio-temporal Fusion based Graph Construction**, we obtain persons' cross-camera transition time set by utilizing the visual similarity measurement, and further filter neighboring nodes based on whether their transition times fall in the fluctuation range of the corresponding transition set. Then, in the stage of **Graph based Node Information Propagation Stage**, we aggregate the neighbors of each node through a two-layer GraphSAGE [8] model to obtain node embeddings and calculate the affinity matrix between nodes. Subsequently, in the stage of **Distillation Learning of Affinity Matrix**, we apply the obtained affinity matrix to distill the model. Finally, in the stage of **Graph based Multi-modal Re-ranking**, we construct a graph of the gallery set and query set using the same construction strategy to obtain the node embeddings of queried images, which are used for the re-ranking of matching results. The details of each stage will be presented in the following sections.

3.1 Spatio-temporal Fusion based Graph Construction

As shown in Fig.1, the lack of ground-truth labels and the irregularity of personal spatio-temporal transition often result in unreliable spatio-temporal scores. The hard combination method by directly multiplying the visual scores and the spatio-temporal scores may even make the performance of the fusion model worse than the pure visual model in some special cases as shown in Fig. 1 (b). The key is to reduce the noise caused by the fusion of spatio-temporal information. Based on this analysis, we propose a soft fusion method of constructing a graph by connecting the nodes with higher spatio-temporal correlation as well as higher visual similarity. The representation of the graph is as follows:

$$\begin{aligned} G^t &=< V^t, E^t > \\ V^t &= \left\{ V_i^t \mid 0 \le i < N_t \right\} \\ E^t &= E_v^t \cup E_{st}^t \end{aligned} \tag{1}$$

Here, $V^t = \left\{ V_i^t \mid 0 \le i < N_t \right\}$ represents the unlabeled images from the training set of the target domain, where N_t is the total number of images. For each image $V_i^t \in V^t$, its visual feature vector is generated by the initialized visual model and denoted as v_i^t. The affinity between any pair of images $V_i^t, V_j^t \in V^t$ is calculated as follows:

$$S_{ij}^t = \exp\left(- \left\| v_i^t - v_j^t \right\|^2 / 2\alpha^2 \right) \tag{2}$$

The adoption of the Gaussian kernel function is aimed at enhancing the non-linear discriminative capacity of the affinity measurement.

E^t in Eq. (1) denotes the edge set of graph G^t, where E_v^t represents the edge set constructed from visual information, and E_{st}^t represents the edge set constructed based on spatio-temporal information. Specifically, for E_v^t, we sort the images from the training set V^t based on the affinity with the image V_i^t, and then select the top k_s images to construct E_v^t. We assume that for any image V_i^t, there is a high probability that its top k_s images belong to the same ID as V_i^t. Therefore, we can use these top k_s images to construct the set of transition time for cross-camera person tracking. The set of transition time $T(a, b)$ from any camera a to another camera b can be represented in the following form:

$$T(a,b) = \left\{ t_i - t_j \mid S_{ij}^t \in \left\{ S_i^t \right\}_{k_s}, C(V_i^t) = a, C(V_j^t) = b \right\} \tag{3}$$

where t_i represents the frame number of image V_i^t captured by camera a, and t_j represents the frame number of image V_j^t captured by camera b. $C(\cdot)$ indicates the camera ID of the image. $S_{ij}^t \in \left\{ S_i^t \right\}_{k_s}$ indicates the affinity between image V_i^t and image V_j^t is within the top k_s ranking of the affinity scores of image V_i^t.

Based on the transition set $T(a, b)$, the edge set E_{st}^t is constructed as follows. For any image V_i^t from camera a, an edge is built with another image V_j^t from camera b, if the following condition is satisfied:

$$\left[\Delta t_{ij} - \delta, \Delta t_{ij} + \delta \right] \cap T(a,b) \neq \varnothing \wedge S_{ij}^t \in \left\{ S_i^t \right\}_{k_l} \tag{4}$$

where $\Delta t_{ij} = t_i - t_j$ indicates the time interval between the two images. δ and k_l are pre-defined positive constant. $S_{ij}^t \in \left\{ S_i^t \right\}_{k_l}$ indicates the affinity between image V_i^t and image V_j^t is within the top k_l ranking of the affinity scores of image V_i^t. Eq. (4) defines the condition combining both spatio-temporal constraint and visual similarity, which makes sure only the images with relatively high similarity and correlated spatio-temporal pattern can be selected to build the connection in E_{st}^t .

3.2 Graph based Node Information Propagation

After constructing a graph that integrates spatio-temporal information, in order to better explore the deep relationships between images, we propose a node information propagation algorithm based on GraphSAGE [8]. In particular, from section 3.1, we can obtain the adjacency matrix representation A^t of the graph as follows:

$$A_{ij}^t = \begin{cases} S_{ij}^t, & E_{ij}^t \in E^t \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where the affinity S_{ij}^t is obtained from Eq (2) to represent the weight of an edge. Given that the adjacency matrix A^t is typically symmetrical, we can introduce a symmetric adjacency matrix $(A^t)^*$ as:

$$\left(A^t\right)^* = \frac{1}{2}\left(A^t + \left(A^t\right)^T \right) \tag{6}$$

Meanwhile, we adopt a two-layer GraphSAGE model to explore the structural relationship between images. Specifically, we obtain the embeddings of nodes by using the GCN-based aggregation functions to integrate the representation of all neighboring nodes at each layer in the following form:

$$\left(v_i^t\right)^k = \sigma\left(W^k \cdot \left(\left(D^t\right)^{-1} \left(A^t\right)^* \right)_{ij} \left(v_j^t\right)^{k-1} \right), \quad \forall j \in \mathbb{N}(i) \cup i \tag{7}$$

where $\left(v_i^t\right)^k$ denotes the embedding for the k^{th} layer obtained by aggregating the features of the $(k-1)^{th}$ layer's neighbors of node V_i^t. $\sigma(\cdot)$ is the ReLu activation function. Meanwhile, W^k represents the projection matrix for the kth layer. D^t is the degree matrix of A^t, which is normalized by using row normalization. $\mathbb{N}(i)$ denotes the neighborhood of node i. Moreover, we use the aggregation function in the form of GCN.

We observe that for any node V_i^t, it is difficult to find effective positive and negative samples to optimize the projection matrix W following the optimization procedure of traditional graph embedding. Additionally, the adjacency matrix already reflects the relationship of two nodes. Therefore, in Eq.(7), we use a randomly initialized projection matrix for aggregation. This not only speeds up the training process but also enables the exploration of deep connections between nodes.

Then, we calculate the affinity matrix M in the form of a Gaussian kernel function for the embeddings of the nodes in the last layer obtained after aggregation. Each element of M is defined as follows:

$$M_{ij} = \exp\left(-\left\|\left(v_i^t\right)^n - \left(v_j^t\right)^n\right\|^2 / 2\alpha^2\right) \tag{8}$$

Here $\left(v_i^t\right)^n$ represents the embedding of image V_i^t obtained by aggregating its neighbours through the n-layer GraphSAGE model. The higher the value of M_{ij}, the greater the probability that images V_i^t and V_j^t belong to the same person.

3.3 Distillation Learning of Affinity Matrix

In order to distill the knowledge embedded in the graph structure to optimize the visual model, we propose a graph based loss function inspired by the heterogeneous affinity learning method H-GO [17]. On the basis of taking into account the diversity of cameras, we adopted the cross-entropy between the affinity distribution and the similarity predicted by the model as the following loss function:

$$\mathcal{L} = -\sum_{i=1}^{N_t}\sum_{c=1}^{N_c}\sum_{\substack{j\in\mathbb{N}(i)\\ C_j=c}} \mathbb{M}_{ij}^{(c)} \log\left(\mathcal{P}^{(c)}\left(\mathcal{F}\left(V_j^t,\theta\right) \mid \mathcal{F}\left(V_i^t,\theta\right)\right)\right) \tag{9}$$

where c denotes the ID of a camera and N_c represents the total number of cameras, $C_j = c$ indicates that image V_j^t is captured by camera c. $\mathbb{N}(i)$ indicates the neighbor set of V_j^t. Eq.(9) groups nodes based on camera ID and independently calculates the affinity within each group. Specifically, the calculation formula for the affinity measure $\mathbb{M}^{(c)}$ related to grouping is as follows:

$$\mathbb{M}_{ij}^{(c)} = \begin{cases} \frac{M_{ij}}{\max\limits_{k\in\mathbb{N}(i),C_k=c}(M_{ik})}, & j \neq i \wedge j \in \mathbb{N}(i) \\ 0, & j \neq i \wedge j \notin \mathbb{N}(i) \\ 1, & j = i \end{cases} \tag{10}$$

Furthermore, $\mathcal{P}^{(c)}$ defines the possibility of the model predicting that image V_i^t and the image V_j^t that is captured by camera c belong to the same person:

$$\mathcal{P}^{(c)}\left(\mathcal{F}\left(V_j^t,\theta\right) \mid \mathcal{F}\left(V_i^t,\theta\right)\right) = \frac{\exp\left(\mathcal{F}\left(V_j^t,\theta\right)\cdot\mathcal{F}\left(V_i^t,\theta\right)/\tau\right)}{\sum\limits_{\substack{k\in\mathbb{N}(i)\\ C_k=c}}\exp\left(\mathcal{F}\left(V_k^t,\theta\right)\cdot\mathcal{F}\left(V_i^t,\theta\right)/\tau\right)} \tag{11}$$

However, the cost of optimizing model parameters by directly minimizing $\mathcal{L}$ is very expensive because we need to recalculate the feature vectors of all images in each iteration. Therefore, we employ a memory bank to store the feature vectors of all images. When calculating similarity, we retrieve the feature vector corresponding to image V_j^t from the memory bank and calculate the similarity with the feature vector generated by the model. Specifically, we define an approximate loss function for $\mathcal{L}$:

$$\mathcal{L} = -\sum_{i=1}^{N_t}\sum_{c=1}^{N_c}\sum_{\substack{j\in\mathbb{N}(i)\\ C_j=c}} \mathbb{M}_{ij}^{(c)} \log\left(\mathcal{P}^{(c)}\left(\mathbb{I}_j \mid \mathcal{F}\left(V_i^t,\theta\right)\right)\right) \tag{12}$$

The feature vector of image V_j^t obtained from the memory bank $\mathbb{I}$ is denoted as $\mathbb{I}_j$, and $\mathcal{P}^{(c)}\left(\mathbb{I}_j \mid \mathcal{F}\left(V_i^t,\theta\right)\right)$ is defined as follows:

$$\mathcal{P}^{(c)}\left(\mathbb{I}_j \mid \mathcal{F}\left(V_i^t,\theta\right)\right) = \frac{\exp\left(\mathbb{I}_j\cdot\mathcal{F}\left(V_i^t,\theta\right)/\tau\right)}{\sum\limits_{\substack{k\in\mathbb{N}(i)\\ C_k=c}}\exp\left(\mathbb{I}_k\cdot\mathcal{F}\left(V_i^t,\theta\right)/\tau\right)} \tag{13}$$

Regarding the feature vectors in memory bank $\mathbb{I}$, we adopt a momentum-based update strategy. In each iteration, the feature vector v_i^t generated from each mini-batch is used to update the corresponding entry $\mathbb{I}_i$ in memory bank $\mathbb{I}$:

$$\mathbb{I}_i = m\mathbb{I}_i + (1-m)v_i^t \tag{14}$$

Where $m \in [0,1]$ is the momentum coefficient of updating the feature vector. In the experiment, m was set to 0.01.

3.4 Graph based Multi-modal Re-ranking

During the testing phase, similar to the approach in sections 3.1 and 3.2, we construct a graph on the gallery set. The feature vectors $\left\{v_i^q\right\}_{i=1}^{N_q}$ and $\left\{v_j^g\right\}_{i=1}^{N_g}$ are extracted from the query set and the gallery set, respectively, using a well-trained backbone. Next, a similar approach as in Eq.(2) is used to obtain the affinity matrix S_{ij}^g between gallery-set images using a Gaussian kernel function.

Following the composition method similar to section 3.1, for any gallery set image V_i^g, all other images in the gallery set are sorted based on the affinity and the top k_s ones are selected as its neighbors. Meanwhile, based on the transition set $T(a,b)$ obtained in section 3.1, any pair of cross-camera images in the gallery set are also connected if the multi-modal constraint of Eq (4) is satisfied.

Then, following the method in section 3.2, we utilize a two-layer GraphSAGE model to obtain embeddings for the nodes in the graph:

$$\left(v_i^g\right)^n = \sigma\left(W^n\cdot\left(\left(D^g\right)^{-1}\left(A^g\right)^*\right)_{ij}\left(v_j^g\right)^{n-1}\right), \forall j \in \mathbb{N}(i)\cup i \tag{15}$$

where n denotes the number of layers in GraphSAGE. $\left(v_i^g\right)^n$ represents the embedding of node V_i^g at the n^{th} layer output by the model. A^g and D^g are defined similarly to section 3.2, where A^g is the adjacency matrix of the graph constructed from the gallery set and D^g is its corresponding degree matrix. $\left(v_j^g\right)^{n-1}$ is the feature vector of the neighboring nodes of node V_i^g at the $(n-1)^{th}$ layer.

For the images in the query set, they are treated as new nodes added to the graph. The judgment of the neighbor relationship is similar to section 3.1. Firstly, using the Gaussian kernel function to calculate the affinity S_{ij}^{qg} between the query set and the gallery set. Secondly, the top k_s gallery-set images for each query image V_i^q are obtained by sorting according to affinity, and they are treated as neighbors. Finally, based on spatio-temporal information, we select

neighboring nodes from the top k_l gallery images of the query image V_i^q according to Eq (4).

Then, similar to Eq.(15), the embedding of the query image is obtained through neighbor aggregation and random projection:

$$\left(v_i^q\right)^n = \sigma\left(W^n \cdot \left(\left(D^{qg}\right)^{-1}\left(A^{qg}\right)^*\right)_{ij}\left(v_j^g\right)^{n-1}\right), \forall j \in \mathbb{N}(i) \cup i \tag{16}$$

In the end, the similarity between the generated embeddings $\left(v_i^q\right)^n$ and $\left(v_i^g\right)^n$ of the query set and gallery set is calculated to obtain the final ranking results.

4 EXPERIMENTS AND RESULTS

In this section, we first introduce the datasets and experimental settings in Sec. 4.1 and Sec. 4.2, respectively, followed by the presentation of results in Sec. 4.3. Subsequently, we conduct ablation study in Sec. 4.4 and provide an analysis of parameter sensitivity in Sec. 4.5. Finally, the qualitative results are presented in Sec. 4.6.

4.1 Datasets

We will evaluate the Graph based Spatial-temporal Fusion method, namely G-Fusion, on two widely recognized large-scale datasets: Market-1501, and MSMT17. The detailed descriptions of these datasets are as follows:

Market-1501 [39]. Market-1501 is a dataset that includes a total of 15,01 identities captured in 32,217 images from six cameras. Each person is captured by at least two cameras. Its training set consists of a total of 12,936 images of 751 identities, while the test set comprises 19,281 images of 750 identities. Market-1501 provides sequence ID labels and partial frame labels for each image, which can be used as time information, and camera labels as spatial information.

MSMT17 [32]. MSMT17 is currently the largest image-based person Re-ID dataset, consisting of a total of 126,441 images of 4,101 identities captured by 15 cameras. The training set contains 32,621 images of 1,041 identities, and the testing set contains 93,820 images of 3,060 identities. Additionally, MSMT17 provides temporal information such as the time period label and local frame label for each image, as well as spatial information in the form of camera labels.

4.2 Configuration of Models

ResNet-50 [10] is employed as the feature extractor $\mathcal{F}$ to transform the input samples into 2048-dimensional feature vectors.

During the supervised training phase in the source domain, we pre-trained our model on ImageNet [3]. We chose the Adam optimizer and set the mini-batch size to 64. We trained the backbone using the conventional identity loss and hard-batch triplet loss with the learning rate of 0.00035 for a total of 80 epochs.

During the unsupervised training phase in the target domain, we resized images to 256×128 using data augmentation of random cropping, flipping, and color jittering. We use Adam optimizer with a mini-batch size of 64 and a learning rate of 0.00035 for a total of 8 epochs for training the model. The hyper-parameter α in Eq.(2) and Eq.(8) were set to 1, while the temperature parameter in Eq.(13) was set to 0.05 and the momentum coefficient in Eq.(14) was set to 0.01. Besides, δ in Eq.(4) was set to 25.

Furthermore, for graph construction during both the training and testing stages, we set k_s and k_l to 3 and 60, respectively. All experiments were conducted on two GTX 1080Ti GPUs with 80 CPU cores and 128GB of memory.

4.3 Comparison Results

The results of our four cross-domain experiments are presented in Tables 1, where the model is pre-trained on the labeled source dataset and transferred to the unlabeled target dataset. As shown in Table 1, our method outperforms most existing methods in the cross-domain experiments on Market-1501 and MSMT17. Specifically, after re-ranking, we achieved a top-1 accuracy of 94.4% and an mAP of 85.9% on Market-1501, which are respectively 1.3% and 22.4% higher than the unsupervised spatial-temporal fusion method STGAL[33]. Our method also surpasses most purely visual methods and approaches transformer-based method. Furthermore, on the MSMT17 dataset, we achieved a top-1 accuracy of 72.7% and an mAP of 52.1% after re-ranking, which are respectively 24.1% and 27% higher than the best unsupervised person Re-ID method JVTC+ [13] that fused spatio-temporal information, and even surpassing the best-performing model TransReID-SSL[20] by 1.0% in terms of mAP.

Additionally, we validated the effectiveness of our method in combination with other pure visual methods. Table 1 displays that the performance of models significantly surpasses the original methods when our method is incorporated. For instance, MMT+G-Fusion* outperforms MMT [6] by 12.6% and 42.7% in mAP on the Market-1501 and MSMT17 datasets, respectively. Moreover, TransReID-SSL+G-Fusion* achieved state-of-the-art results on both Market-1501 and MSMT17 datasets. Furthermore, it can be observed from the table that there is a significant improvement of performance after re-ranking. This is because after processing the gallery set of the target-domain, the model can more easily identify images in the gallery set related to the query image, thus considerably enhancing the performance of the model.

4.4 Ablation Study

The most crucial parts of our method include integrating spatial-temporal information in training and testing stage when constructing the graph and re-ranking based on graph. In order to analyze the contribution of these components clearly, we conducted ablation studies on two datasets and show the results in Table 2. Specifically, Model 0 indicates the basic model only using vision information to construct the graph in the training phase and not conducting post-processing method in the testing phase. Model 1 indicates the model adding spatio-temporal information to construct the graph in the training phase. Model 2 represents using spatio-temporal information to construct the graph in the training phase and conducting reranking in the testing phase. Model 3 indicates the model adding re-ranking in the testing stage. Model 4 represents using re-ranking and spatio-temporal information in the testing phase. Model 5 is the full model equipped with all components. As shown in Table 2, the effectiveness of solely using visual information in the training and testing stages is significantly lower than that of models that incorporate spatio-temporal information. Moreover, if spatio-temporal information is not used in either the training or

Table 1: Comparison with state-of-the-art spatio-temporal fusion and pure visual methods on Market-1501 and MSMT17. (*) indicates the re-ranking is implemented in the testing stage.

Type	Method	MSMT17→Market-1501				Market-1501→MSMT17			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
unsupervised spatio-temporal fusion methods	RASTF(USL) [16]	25.7	64.3	-	-	-	-	-	-
	TAUDL(USL) [15]	41.2	63.7	-	-	-	-	-	-
	TCUL [1]	44.5	72.1	86.3	-	-	-	-	-
	STAR-DAC(USL) [22]	33.9	67.0	80.6	84.9	17.66	41.51	55.12	64.86
	JVTC+ [13]	-	-	-	-	25.1	48.6	65.3	68.2
	STGAL(USL) [33]	63.5	93.1	-	-	-	-	-	-
unsupervised pure visual methods	H-GO [17]	65.2	88.4	94.4	95.9	9.3	25.1	35.9	41.2
	MMT [6]	76.6	90.3	96.9	98.1	22.7	50.2	64.0	69.9
	SpCL [7]	75.5	89.7	96.0	97.5	21.3	45.8	57.6	63.2
	P^2LR [9]	-	-	-	-	29.0	58.8	71.2	76.0
	AWB [31]	79.4	92.6	97.1	98.2	29.0	57.3	70.7	75.9
	CAP [30]	75.9	90.4	95.6	97.2	31.7	63.5	75.1	78.9
	CIDAM(USL) [11]	78.4	90.9	-	-	34.3	65.9	-	-
	CCL [12]	-	-	-	-	35.8	65.8	-	-
	ICE [2]	82.6	93.9	97.5	98.4	38.8	69.4	80.2	84.2
	PAT [37]	-	-	-	-	41.5	67.8	74.5	78.9
	TransReID-SSL [20]	89.5	95.4	98.0	98.8	51.1	74.3	83.6	87.1
Ours	G-Fusion	75.4	90.1	95.7	97.2	26.6	54.0	67.1	72.4
	G-Fusion*	85.9	94.4	96.9	97.4	52.1	72.7	78.7	80.5
	MMT+G-Fusion*	89.2	95.8	97.9	98.3	65.4	82.5	86.5	87.7
	SpCL+G-Fusion*	90.0	95.7	98.0	98.2	65.5	82.0	85.9	86.9
	CAP+G-Fusion*	87.9	95.3	97.3	97.7	66.5	82.8	87.1	88.0
	ICE+G-Fusion*	91.0	96.8	98.6	98.8	73.6	87.6	90.4	90.9
	TransReID-SSL+G-Fusion*	**92.2**	**96.6**	**98.7**	**99.0**	**80.4**	**89.5**	**92.3**	**92.8**

Table 2: Ablation study of our method with different configurations of key components.ST_{train} and ST_{test} means integrating spatial-temporal information when constructing the graph in the training stage and in the testing stage.

Model	Components				MSMT17→Market-1501				Market-1501→MSMT17			
	Baseline	ST_{train}	re-ranking	ST_{test}	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
0	✓	✗	✗	✗	60.4	81.4	90.8	93.4	9.3	23.2	33.9	39.5
1	✓	✓	✗	✗	75.4	90.1	95.7	97.2	26.6	54.0	67.1	72.4
2	✓	✓	✓	✗	76.7	89.8	94.2	95.4	25.3	51.2	62.5	66.9
3	✓	✗	✓	✗	60.3	81.0	87.9	90.2	8.5	21.3	29.1	32.8
4	✓	✗	✓	✓	74.4	89.0	93.1	94.4	21.8	44.9	51.6	54.1
5	✓	✓	✓	✓	**85.9**	**94.4**	**96.9**	**97.4**	**52.1**	**72.7**	**78.7**	**80.5**

testing phase, the mAP on the Market-1501 dataset is decreased by 25.5% and the Rank-1 is decreased by 13.0% compared to the complete model. These results indicate that spatio-temporal information plays a vital role in both the training and testing stages of our method.

4.5 Parameter Sensitivity Analysis

We analyzed the sensitivity of the hyper-parameters, k_s and k_l, used in our method during the training and testing phases, and the results are shown in Fig.3 and Fig.4. The parameter k_s represents the number of neighbors based on visual information in graph construction, while k_l represents the upper limit of the number of neighbors based on spatio-temporal information. As shown in Fig.3, during the training phase, the model achieved optimal performance when k_s ranged from 2 to 3 and k_l ranged from 60 to 100. A significant increase in model performance was observed as k_s increased from 1 to 2. This is because when k_s is equal to 1, each node's neighborhood only includes itself, making it impossible to establish relationships with other nodes. However, when k_s increased from 2 to 10, there was a noticeable decrease in model performance on the MSMT17 dataset, because this could lead to considering negative samples that are visually similar as neighbors, introducing many noisy images that can affect model performance. As depicted in Fig.4, during the testing phase, the model performance exhibits a similar variation trend to that observed during the training phase as k_s and k_l undergo changes. The optimal performance of the model

is achieved when k_s ranges from 2 to 3, and k_l ranges from 60 to 100. In our experiments, we set k_s to 3 and k_l to 60 for both the training and testing phases.

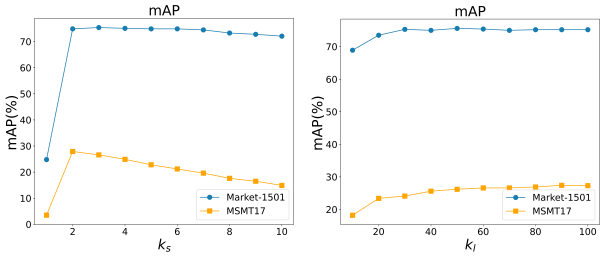

Figure 3: Parameters analysis of k_s and k_l on Market-1501 and MSMT17 in the training stage.

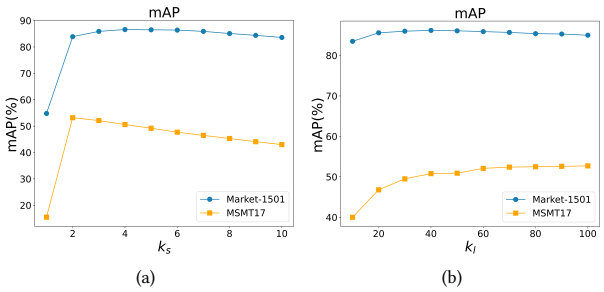

Figure 4: Parameters analysis of k_s and k_l on Market-1501 and MSMT17 in the testing stage.

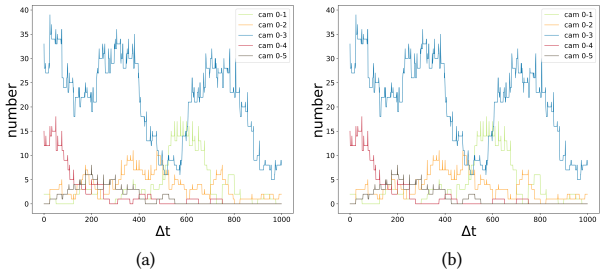

Figure 5: The transition time set in the Market-1501 dataset extracted by our model and the ground truth. (a) Extracted by our model. (b) Extracted by the ground truth.

4.6 Qualitative Results

In order to better demonstrate the effectiveness of the fusion of spatio-temporal information in our method, we present in Fig.5 the quantity statistics of person transition time extracted by our model and the ground truth. By comparing Fig.5(a) and Fig.5(b), it can be seen that the spatio-temporal information extracted by our method is generally close to the ground truth. Moreover, the time periods where the number of person transitions is 0 are also consistent with the actual person transition situations. Therefore, the method of using the proximity between the transition times of any two cross-camera images and the transition time set to assist visual information judgment on whether they belong to the same person is effective.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel Graph based Spatio-temporal Fusion model, for unsupervised person Re-ID, namely G-Fusion. We adopt a soft fusion strategy by leveraging person transition time to assist in visual information filtering for neighbor nodes selection, thus reducing the noise caused by spatio-temporal information fusion. We use GraphSAGE-based Node Information Propagation method to obtain the embeddings of nodes and employ the distillation learning of affinity matrix to guide model optimization. Finally, during the testing phase, we apply the graph based multi-modal Re-ranking to further improve the performance. Comprehensive experiments based on two large-scale Re-ID datasets show that G-Fusion can significant improve the performance of SOTA unsupervised person Re-ID methods.

In future work, we will extend the proposed method to streaming scenarios to achieve dynamic graph expansion for handling continuously incoming streaming data. Moreover, we will explore the possibility of deploying the method on a real camera network. This will enable us to tackle real-world challenges in the field of person Re-ID.

ACKNOWLEDGMENTS

This work was supported by the Science and Technology Program of Guangzhou, China (2023B03J1388) and the Key-Area Research and Development Program of Guangzhou City (202206030009).

REFERENCES

[1] Usman Ali, Bayram Bayramli, and Hongtao Lu. 2019. Temporal Continuity Based Unsupervised Learning for Person Re-Identification. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part V 26*. Springer, 770–778.

[2] Hao Chen, Benoit Lagadec, and Francois Bremond. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14960–14969.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.

[5] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *proceedings of the IEEE/CVF international conference on computer vision*. 6112–6121.

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526* (2020).

[7] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems* 33 (2020), 11309–11321.

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

[9] Jian Han, Ya-Li Li, and Shengjin Wang. 2022. Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 790–798.

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

[11] Pingting Hong, Dayan Wu, Bo Li, and Weipinng Wang. 2022. Camera-specific Informative Data Augmentation Module for Unbalanced Person Re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*. 501–510.

[12] Takashi Isobe, Dong Li, Lu Tian, Weihua Chen, Yi Shan, and Shengjin Wang. 2021. Towards Discriminative Representation Learning for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8526–8536.

[13] Jianing Li and Shiliang Zhang. 2020. Joint visual and temporal consistency for unsupervised domain adaptive person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 483–499.

[14] Jianing Li, Shiliang Zhang, Qi Tian, Meng Wang, and Wen Gao. 2019. Pose-guided representation learning for person re-identification. *IEEE transactions on pattern analysis and machine intelligence* 44, 2 (2019), 622–635.

[15] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2018. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*. 737–753.

[16] Wei Li, Meibin Qi, Ning Yang, Guowu Zhou, and Yubing Yang. 2020. Unsupervised Spatial-Temporal Model Based on Region Alignment for Person Re-identification. In *Journal of Physics: Conference Series*, Vol. 1518. IOP Publishing, 012025.

[17] Shaochuan Lin, Jianming Lv, Zhenguo Yang, Qing Li, and Wei-Shi Zheng. 2022. Heterogeneous graph driven unsupervised domain adaptation of person re-identification. *Neurocomputing* 471 (2022), 1–11.

[18] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. 2019. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8738–8745.

[19] Sen Ling, Hua Yang, Chuang Liu, Lin Chen, and Hongtian Zhao. 2022. Spatial-Temporal Constrained Pseudo-labeling for Unsupervised Person Re-identification via GCN Inference. In *Digital TV and Wireless Multimedia Communications: 18th International Forum, IFTC 2021, Shanghai, China, December 3–4, 2021, Revised Selected Papers*. Springer, 297–311.

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

[21] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. 2018. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7948–7956.

[22] Munaga VNK Prasad, Ramadoss Balakrishnan, et al. 2022. Spatio-temporal association rule based deep annotation-free clustering (STAR-DAC) for unsupervised person re-identification. *Pattern Recognition* 122 (2022), 108287.

[23] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. 2019. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8080–8089.

[24] Min Ren, Lingxiao He, Xingyu Liao, Wu Liu, Yunlong Wang, and Tieniu Tan. 2021. Learning Instance-level Spatial-Temporal Patterns for Person Re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14930–14939.

[25] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. 2021. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4390–4403.

[26] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry Steven Davis, and Wen Gao. 2017. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE transactions on pattern analysis and machine intelligence* 40, 5 (2017), 1167–1181.

[27] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10981–10990.

[28] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. 2019. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8933–8940.

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

[30] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. 2021. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2764–2772.

[31] Wenhao Wang, Fang Zhao, Shengcai Liao, and Ling Shao. 2022. Attentive WaveBlock: Complementarity-enhanced mutual networks for unsupervised domain adaptation in person re-identification and beyond. *IEEE Transactions on Image Processing* 31 (2022), 1532–1544.

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

[33] Chih-Wei Wu, Chih-Ting Liu, Wei-Chih Tu, Yu Tsao, Yu-Chiang Frank Wang, and Shao-Yi Chien. 2020. Space-time guided association learning for unsupervised person re-identification. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2261–2265.

[34] Yiming Wu, Xintian Wu, Xi Li, and Jian Tian. 2021. MGH: metadata guided hypergraph modeling for unsupervised person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1571–1580.

[35] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. 2019. Unsupervised person re-identification by soft multilabel learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2148–2157.

[36] Jongmin Yu and Hyeontaek Oh. 2022. Graph-structure based multi-label prediction and classification for unsupervised person re-identification. *Applied Intelligence* 52, 12 (2022), 14281–14293.

[37] Shengming Yu, Zhaopeng Dou, and Shengjin Wang. 2023. Prompting and Tuning: A Two-Stage Unsupervised Domain Adaptive Person Re-identification Method on Vision Transformer Backbone. *Tsinghua Science and Technology* 28, 4 (2023), 799–810.

[38] Minying Zhang, Kai Liu, Yidong Li, Shihui Guo, Hongtao Duan, Yimin Long, and Yi Jin. 2021. Unsupervised domain adaptation for person re-identification via heterogeneous graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 3360–3368.

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

[40] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. 2020. Robust partial matching for person search in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6827–6835.

[41] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. 2018. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European conference on computer vision (ECCV)*. 172–188.

[42] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 598–607.

[43] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera style adaptation for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5157–5166.