IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 10, 2021, accepted March 2, 2021, date of publication March 15, 2021, date of current version March 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065984

Part Relational Mean Model for Group Re-Identification

PING HU[1,2,3,4], HONGWEI ZHENG[1,2,3], AND WEISHI ZHENG[4], (Member, IEEE)

[1]Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China
[2]Key Laboratory of GIS and RS Application Xinjiang Uygur Autonomous Region, Urumqi 830011, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]School of Computer Science, Sun Yat-sen University, Guangzhou 510006, China

Corresponding author: Hongwei Zheng (hzheng@ms.xjb.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant NSFC-U1803120.

ABSTRACT Most current research on pedestrian re-identification (ReID) is focusing on single-person ReID. However, people are rarely alone and often walk together in groups. Therefore, there is an urgent need to study the problem of group ReID (G-ReID). G-ReID is challenging because of the difficulties related to the differences in group appearance caused by changes in the group layout and membership. In this paper, we have proposed a part-based minus-average relational and arithmetic mean descriptor (PRM) algorithm to obtain a robust representation of groups. Based on local features, we have designed the arithmetic mean descriptor and the minus-average relational descriptor to solve the G-ReID problem caused by changes in the number of group members and their relative positions within the group. Moreover, the minus-average relational descriptor can also be used to describe the differences in the appearance of group members. Considering the rarity of G-ReID datasets and the need to improve the applicability of the G-ReID algorithm in real scenarios, we have collected a new dataset called the Bus Rapid Transit (BRT) G-ReID dataset. Extensive experimental results demonstrate the effectiveness of the PRM algorithm and indicate that it outperforms state-of-the-art algorithms by 7.5% for the cumulative matching feature (CMC-1) on the i-LIDS MCTS group dataset and by 19.4% for the CMC-1 on the Road Group dataset and it outperforms the baseline by 2.4% for the CMC-1 on the BRT dataset.

INDEX TERMS Group re-identification, relational model, part-based CNN.

I. INTRODUCTION

Person re-identification (ReID) has attracted considerable attention due to its wide range of applications, such as in security and surveillance. Existing research has focused on re-identifying individuals; however, searching for a certain group of persons has rarely been studied. It is usual for a group of people to walk along a street together. As illustrated in Fig. 1, the same group was captured by cameras at different Bus Rapid Transit (BRT) stations in the city center. In this paper, the research objects of the pedestrian ReID task are defined as a group, such as a couple travelling together, students walking together after school, colleagues who have the same work schedule and parents walking with children after school.

The associate editor coordinating the review of this manuscript and approving it for publication was Hongwei Du.

Unlike individual ReID, the aim of group re-identification (G-ReID) is to associate a certain group with different camera views. In addition to the traditional challenges of ReID, such as low resolutions, pose changes, illumination variations and blurred vision, G-ReID poses some unique challenges [1]. Changes in the number of group members and in the relative positions of members within the group can cause differences in the appearance of the group, which is not conducive to group image matching. Therefore, G-ReID is a more challenging task because of the deformable characteristics of groups.

To solve G-ReID, early researchers used traditional manual design methods [2]–[7] to extract group features. Most existing methods view the input group image as an entire unit and extract global or semiglobal features. However, it is not suitable to treat the group as a whole and extract its global or semiglobal features because the changes in the relative positions of group members can alter the visual content of

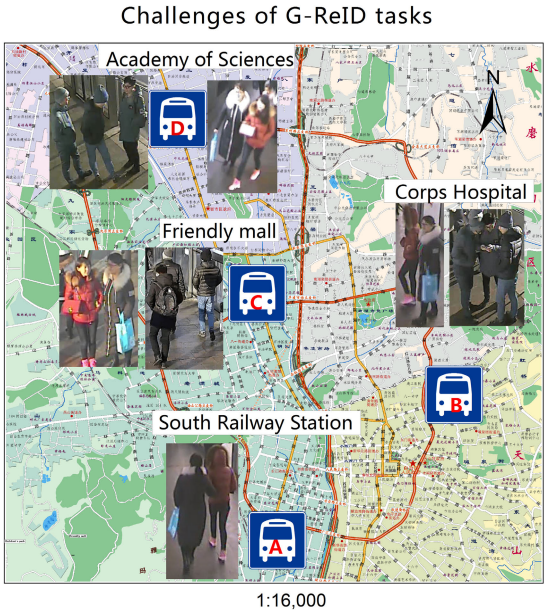


FIGURE 1. Challenges of G-ReID. In addition to traditional challenges, such as low resolutions, pose changes, illumination variations and blurred vision, G-ReID has some unique challenges. Changes in the number of group members and in the relative positions of members within the group can cause deformable characteristics of the group appearance, as shown in Fig. 1. In the figure, there are two groups. One group consists of two women and the other group consists of three men. The group image taken at a Friendly Mall site shows that a woman in red is to the left of another woman in black. The image of the same group captured by another camera at the site of the Academy of Sciences shows that the woman in red is on the right side of the woman in black.

the group. Because of the powerful ability of convolutional neural networks to describe local features, recent studies have shown excellent performance of some deep learning-based methods in visual recognition tasks [8], [9]. Inspired by these works, a large number of deep learning techniques have been applied to single-pedestrian ReID. tasks [10], [11]. Nevertheless, few works have utilized deep learning methods for G-ReID. Deep learning can obtain a good feature summary by gradually summarizing shallow features into deep features. Shallow features represent the local details of objects and deep features express high-level semantic information. In addition, people usually distinguish pedestrians by local features. Therefore, considering the strong recognition of local features and the summary ability of deep features, we use local deep features to describe the outward appearance of a group.

Since the appearance of a group is affected by changes in the number of group members and in the relative positions of the members within the group, the goal is to construct descriptors that are not affected by these deformable characteristics. The part-based minus-average relational and arithmetic mean descriptor (PRM) algorithm are designed for the challenges of the G-ReID. Based on local features, we have designed the arithmetic mean descriptor and the minus-average relational descriptor to solve the G-ReID problem caused by changes in the number of group members and their relative positions within the group. Moreover, the minus-average relational descriptor can also be used to describe the differences in the appearance of group members. Additionally, we input the features obtained by the minus-average relational descriptor and the features obtained by the arithmetic mean descriptor into the cross-entropy loss function. We then apply the gradient descent algorithm to optimize the objective function and obtain 12 classifiers to describe group features. Consequently, the PRM algorithm task is formed for group feature extraction.

The main contributions of this paper include the following: 1) We have proposed a PRM algorithm for G-ReID. 2) Considering the rarity of G-ReID datasets and the need to improve the applicability of the G-ReID algorithm in real-life scenarios, we have collected a new dataset denoted as the BRT G-ReID dataset. Our extensive experimental results demonstrate the effectiveness of the PRM algorithm and indicate that it outperforms state-of-the-art algorithms by 7.5% for the cumulative matching feature (CMC-1) on the i-LIDS MCTS group dataset and by 19.4% for the CMC-1 on the Road Group dataset, and it outperforms the baseline by 2.4% for the CMC-1 on the BRT dataset.

II. RELATED WORKS

A. GROUP SEMANTICS

1) MULTIDISCIPLINARY USE OF GROUP SEMANTICS

The research on group semantics involves multidisciplinary fields. Different disciplines that study group semantics have different research perspectives. The studies on group semantics in social humanities [5] [12]–[17] aim to use existing group semantic algorithms to analyse social phenomena and provide technical support for social services. The research on group semantics in the field of computer vision aims to innovate the group semantic algorithm model. To improve the accuracy and efficiency of the subtask, these group semantic models are applied to computer vision subdivision tasks, such as target detection [18], [19], ReID [20], target tracking [21]–[26] and G-ReID [2]–[7] .

2) G-ReID USING GROUP SEMANTICS

We summarize the existing G-ReID tasks as follows. Zheng *et al.* [4] proposed the center rectangular ring ratio-occurrence descriptor (CRRRO) and block based ratio-occurrence descriptor (BRO). Cai *et al.* [5] proposed a covariance descriptor for the appearance matching of group images. The covariance descriptor is a discriminative descriptor that captures both the appearance and statistical properties of image regions. Zhu *et al.* [6] formulated G-ReID as a patch matching task and proposed to learn an ensemble of ''salience channels'' that are robust to illumination variations and that can filter out unreliable and noninformative patch matches. Lisanti *et al.* [3] proposed a novel encoding scheme based on dictionary learning to perform G-ReID. To circumvent the poor detection performance caused by occlusions, Koperski *et al.* [7] used fixed regions of interest and

employed codebook-based visual representations. In terms of the extraction of group features, early researchers used traditional manual design methods. In addition, most existing methods view the input group image as an entire unit and extract global or semiglobal features.

B. LOCAL FEATURE REPRESENTATION

Local features can consider the geometric properties of data [27], [28]. In addition, local features constitute global features, thus, local features can effectively represent the intrinsic structural relationship. Specifically, there is a spatial relationship between local features. The attribute of consistency of spatial information contributes to feature expression [29]. Therefore, we want to apply local feature information to group appearance modeling. Local features include traditional local features and deep local features.

1) ReID USING LOCAL FEATURES

Pedestrian ReID has been performed with traditional local features. Gray and Hai [30] presented an algorithm for performing viewpoint-invariant pedestrian ReID by using the ensemble of localized features (ELF) representation. Bak *et al.* [20] proposed a new appearance model based on spatial covariance regions extracted from human body parts. The new spatial pyramid scheme was applied to capture the relationships between human body parts to obtain a discriminative human signature. In [20], Farenzenal *et al.* proposed features that model three complementary aspects of the human appearance. Farenzena *et al.* [31] computed a simple vector of attributes that consists of the pixel coordinates for each pixel of an image. These local descriptors are then turned into Fisher vectors that represent the group image. Ma *et al.* [32] used a local Fisher discriminant analysis algorithm to achieve pedestrian ReID. Pedagadi *et al.* [33] proposed an effective feature representation called local maximal occurrence (LOMO) and a subspace and metric learning algorithm called cross-view quadratic discriminant analysis (XQDA). Liao *et al.* [34] proposed a decision function for verification that can be viewed as a joint model of a distance metric and a locally adaptive thresholding rule. The hand-designed descriptors are used in these works to express pedestrian. Actually, hand-designed descriptors would be disturbed by human factors and less intelligence.

The research on ReID that uses deep local features includes the following. Chen *et al.* [35] proposed a polynomial feature map to describe the matching within each subregion and injected all the feature maps into a unified framework. Yao *et al.* [36] proposed a deep representation learning procedure named the part loss network (PL-Net) to minimize both the empirical classification risk and the representation learning risk. Sun *et al.* [37] proposed a uniform partition strategy, namely, a part-based convolutional baseline (PCB), that achieves competitive results with state-of-the-art algorithms, which validate it as a strong convolutional baseline for person retrieval. Suh *et al.* [38] proposed a two-stream network and a bilinear-pooling layer. Each local feature of the part-aligned map is obtained by a bilinear mapping of the corresponding local appearance and body part descriptors. Sun *et al.* [39] proposed a visibility-aware part model (VPM) that learns to perceive the visibility of regions through self-supervision. The visibility awareness allows the VPM to extract region-level features and compare two images with a focus on their shared regions. Researchers use deep learning technology adaptively learn the local feature weight matrix, and perform well on the task of ReID.

2) G-ReID USING LOCAL FEATURES

G-ReID works have also involved local region-based descriptors. Zheng *et al.* [2] proposed a center rectangular ring ratio-occurrence descriptor and a block-based ratio-occurrence descriptor. Lisanti *et al.* [3] divided the entire group image into uniform small blocks, extracted features based on the small blocks and represented the group image in the form of block sets. However, such methods extract back information at different scales, which increases interference information when expressing group appearance. Moreover, treating the group as a whole and extracting its global or semiglobal features may not yield good performance because changes in the relative positions of group members can alter the visual content of the group.

Current evaluations show that the performance of traditional local feature operators is far inferior to the performance of deep feature operators. The main reason is that deep learning has good feature summary ability. Deep learning can gradually summarize shallow features into deep features. Shallow features represent the local details of objects and deep features express high-level semantic information. Considering the above factors, we intend to detect and clip the members of a group to avoid interference from background information. Finally, considering the stability of human body structure information and the advantages of deep learning, we use the PCB network [40] to implement local feature localization and local feature extraction.

C. RESEARCH ON TARGET RELATIONALITY

In real-life scenarios, we often observe changes in the relative positions of group members within a group. We define the changes in the structure of group members as the relationality. Deng *et al.* [41] developed a new model that allows the encoding of flexible relations between labels. This model introduced hierarchy and exclusion (HEX) graphs and a new formalism. Ding *et al.* [42] proposed the HEX model to allow for soft or probabilistic relations between labels. Inspired by recent advances in the relational representation learning of knowledge bases and convolutional object detection networks, Zhang *et al.* [43] proposed a visual translation embedding network (VTransE) for visual relation detection. Chen *et al.* [44] learned a novel similarity function that consists of multiple subsimilarity measurements, each of which is in charge of a subregion. Li *et al.* [45] designed a multi-scale context-aware network (MSCAN) to learn the powerful features of the full body and body parts and to capture local

context knowledge well by stacking multiscale convolutions in each layer. Chen *et al.* [46] proposed an algorithm that not only improves the learning of global visual features via a supervision of the overall description but also enforces semantic consistencies between the local visual and linguistic features, which is achieved by building global and local image-language associations. Fei *et al.* [47] proposed a new saliency learning algorithm based on a three-stream convolutional neural network (CNN) that is first presented to learn the distinctive features of the upper body, lower body and global body. Hu *et al.* [48] proposed an object relation module that processes a set of objects simultaneously through an interaction between their appearance feature and geometry, which allows their relations to be modeled. Huang *et al.* [49], [50] represented the coupling relations between every two group members according to the differences in their personal features to efficiently signify the co-occurrence of the two members with only their discrepancies.

The above studies have achieved target classification and ReID by enhancing the relationality among the targets. Inspired by these previous works and the efficient expression ability of deep convolution, we first detect the members of a group, clip a single pedestrian according to the detection frame and extract the local features of the single pedestrian. According to the local features of the group members, the minus-average relational and arithmetic mean descriptors are constructed to strengthen the relationality among group members to solve the problem of G-ReID.

III. APPROACH

The proposed framework of this work is shown in Fig. 3 and Fig. 4. The framework consists of a group feature classifiers training stage and a group feature matching stage [51]. In the training stage, based on local feature location and extraction, we obtain six arithmetic mean features h_j^M by using the arithmetic mean descriptor and six minus-average relational features h_j^R by using the minus-average relational descriptor. After dimension reduction, we input these feature vectors into the softmax multiobjective classification function in Eq. (3) and Eq. (4) and use the gradient descent algorithm to optimize the cross-entropy loss function in Eq. (5) and Eq. (6) to obtain the weight matrix W. In the training stage, we obtain 12 feature classifiers of the PRM algorithm. In the matching stage, we extract features from the probe image h^{Gp} and the gallery images h^{Gg} via the PRM model and calculate the distances between the probe feature and the gallery feature to re-identify the group ID of the probe image according to the distances. To facilitate reading, we show the notations used in the PRM algorithm in TABLE 1. Please refer to TABLE 1 for the notations used throughout the paper.

TABLE 1. Descriptions of the key notations used in this paper.

P	Notation for the part-based algorithm strategy
R	Notation for the minus-average relational descriptor
M	Notation for the arithmetic mean descriptor
U	Notation for the max relational descriptor
G	Notation for the global descriptor $R \oplus M$
Φ	Represent a fully connected operation
S	Notation for the cosine similarity function
p	Number of person parts, $p = 6$
b	Bias vector
c	The category of the training set, $c = 200$
m	Number of group members
h_j^M	j-th part feature vector expressed by M
h_j^R	j-th part feature vector expressed by R
h_{ij}	Features of the j-th part of the i-th pedestrian
h_{kj}	Features of the j-th part of the k-th person
$\hat{h}_j^R$	Feature h_j^R after fully connected operation Φ
$\hat{h}_j^M$	Feature h_j^M after fully connected operation Φ
h^G	Feature of a group image described by $\hat{h}_j^R \oplus \hat{h}_j^M$
h^{Gp}	Feature from the probe set described by G
h^{Gg}	Feature from the gallery set described by G
W^R	Weight matrix that corresponds to the feature vector $\hat{h}^R$
W^M	Weight matrix that corresponds to the feature vector $\hat{h}^M$

A. DETECTION OF G-ReID

The G-ReID image of the existing database contains multiple pedestrians, and a single pedestrian has no detection frame-labeling information. Therefore, before the G-ReID can work, we must detect the pedestrians and the quality of the pedestrian detection directly affects the accuracy of the G-ReID. The workflow of G-ReID is to first detect the members of the group, crop a single pedestrian picture based on the detection bounding box and design the group features based on the features of the single pedestrian. Considering the efficiency and accuracy of the detection algorithm, we use SSD [52] to detect a single pedestrian in a group. To improve the accuracy of pedestrian detection, we pretrain the SSD model on the INRIA [53] pedestrian datasets.

B. GROUP FEATURE EXPRESSION

Group feature expression is an important stage of the G-ReID task. Good group feature expression can effectively improve the accuracy of G-ReID. G-ReID must solve the problem of the group appearance changes caused by the layout of group members and changes in the number of group members. To solve these problems, we have proposed the PRM algorithm. The PRM algorithm uses the PCB network structure [40] to implement local feature localization and the extraction of group members. Subsequently, to solve the unique problem of G-ReID, we use an arithmetic mean descriptor and minus-average relational descriptor to work with the local features.

1) LOCAL FEATURE LOCATION AND EXTRACTION

In daily life, people usually distinguish pedestrians by local features, such as differences in hairstyles, facial contours and body weight. Therefore, considering the highly discriminative nature of local features, we use local feature descriptors to describe group images. PCB is a baseline for the person retrieval task [40]. p is a parameter in the baseline algorithm, and its specific meaning is the number of horizontally divided blocks of a detected pedestrian. When $p = 6$, the pedestrian retrieval task of the baseline algorithm has the

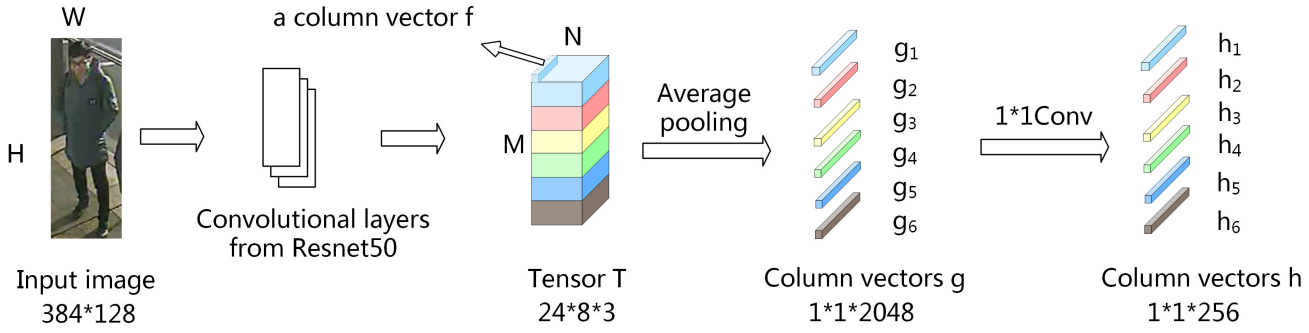


FIGURE 2. PCB feature extraction network structure. The size of a single pedestrian picture is unified into an image with an aspect ratio of 3 : 1 and size of 384 × 128. The tensor T is formed after the ResNet50 convolutional network and the T level is divided into 6 bands. The size of the tensor T is 24 × 8 × 3. Then, by using traditional average pooling, T is changed to a column vector g of 1 × 1×2, 048 dimensions and g is transformed into a 1 × 1 × 256 column vector h by the 1 × 1 convolution kernel. The detected single pedestrian is finally expressed as 6 feature column vectors h.

highest accuracy. Therefore, parameter p is also set to 6 in our research method. Considering the robust performance of the PCB framework, we use the PCB network structure to realize the local feature location and extraction of a single pedestrian in a group. The network structure of PCB is shown in Fig. 2.

2) ARITHMETIC MEAN DESCRIPTOR

The changes in the relative positions of group members lead to variations in the appearance of the same group. Our algorithm aims to design a consistent appearance representation of the same group image regardless of the changes in group member positions. In addition, a change in the number of group members leads to a change in the group image appearance features, which is not conducive to later group image feature matching. Therefore, we must find the descriptor of the group image to solve these problems. The arithmetic mean descriptor proposed in this paper can solve the G-ReID problem caused by changes in the number and relative positions of group members, as shown in Eq. (1).

Based on the location and extraction of local features, we must find an effective method of expressing group appearance features combined with local features. The number of members in a group is variable. If the local feature vectors of the members in a group are simply concatenated, then the dimensions of feature expression in different groups will be different, which is not conducive to feature matching between different groups. If we use the arithmetic mean operator to calculate group features, then the feature dimensions of each group image are all the same. The feature dimensions of different group images are not affected by the number of members in the group, which is conducive to later-stage cosine similarity calculations between group images. Additionally, the group image represented by the arithmetic mean descriptor can robustly work with differences in the appearance of the group image caused by changes in the relative positions of group members. Moreover, the dimension of the group appearance feature h_j^M obtained by the arithmetic mean descriptor is fixed and is not related to the change in the number of pedestrians, thereby avoiding the dimension disaster caused by the increase in the number of members in the group.

$$h_j^M = M(h_j) = \frac{\sum_{i=1}^{m} h_{ij}}{m} \tag{1}$$

3) MINUS-AVERAGE RELATIONAL DESCRIPTOR

We use the minus-average relational descriptor h_j^R to describe the differences in the appearance of group members, as shown in Eq. (2). The algorithm flow of the minus-average relational descriptor is to subtract the corresponding local visual features of the group members and calculate the mean. Therefore, the minus-average relational descriptor can better describe the appearance differences of the members in a group, and reflect the relationship between group members. In addition, the minus-average relational descriptor also can solve the variation in the group appearance caused by changes in the number and relative positions of the group members. The physical mechanism is the same as the arithmetic mean descriptor's mechanism because it uses the same mean operation.

$$h_j^R = R(h_j) = \frac{\sum_{i=1}^{m} \sum_{k=i+1}^{m} |h_{ij} - h_{kj}|}{C_m^2} \tag{2}$$

Accordingly, the minus-average relational descriptor not only has the advantages of the arithmetic mean descriptor but also can describe the appearance differences of group members, which contribute to the correct judgement of the G-ReID system.

4) PRM G-ReID ALGORITHM

The PRM algorithm is an integration of the technical strategies described above. Based on local feature location and extraction, we obtain six arithmetic mean features h_j^M through the arithmetic mean descriptor and six minus-average relational features h_j^R through the minus-average relational descriptor. We reduce the dimension by using the fully connected layer and obtain the 1 × 200 dimensional vectors $\hat{h}_j^R$

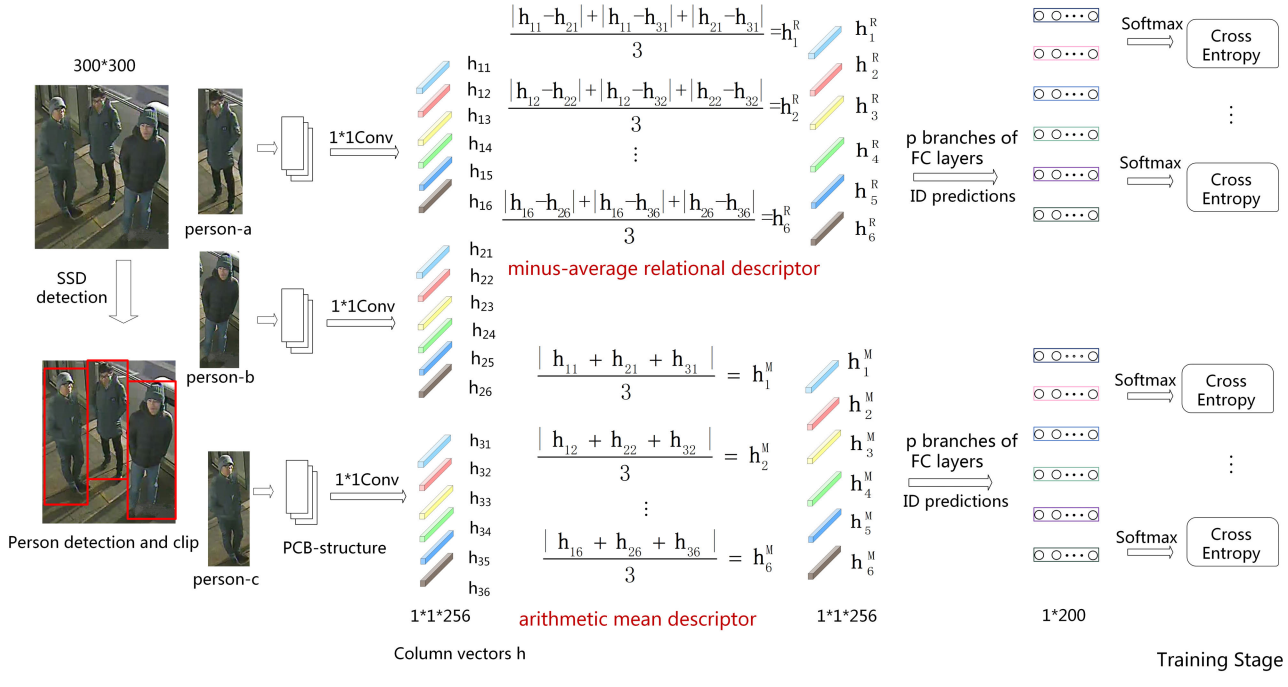


FIGURE 3. Architecture of the PRM algorithm. After entering a group of images, we use the SSD detection algorithm pretrained on the INRIA dataset to detect the pedestrians. Subsequently, we crop the detected pedestrians, set the image size to 384 × 128 and set the image ratio to 3: 1. We place the detected single pedestrian into the PCB network structure to achieve local feature localization and feature extraction. Specifically, the detected single pedestrian is divided into an even six blocks and the local feature information for these six blocks is extracted to form six column vectors h with a size of 1 × 1 × 256. We perform the arithmetic mean and minus-average relational operations on h to obtain the arithmetic mean descriptor h^M and the minus-average relational descriptor h^R. We reduce the dimension using the fully connected layer and obtain the 1 × 200 dimensional vectors $\hat{h}^M$ and $\hat{h}^R$. Then we input these features into the softmax multiobjective classification functions to obtain 200 probability values. Next, we input these probability values into the cross-entropy loss function and use the gradient descent algorithm to optimize and obtain the weight matrix W^M of the arithmetic mean descriptor and the weight matrix W^R of the minus-average relational descriptor. Finally, the entire process of the PRM algorithm has been completed.

and $\hat{h}_j^M$. These feature vectors are subsequently input into the softmax multiobjective classification function shown in Eq. (3) and Eq. (4). Next, we input the probability value into the cross-entropy loss function in Eq. (5) and Eq. (6) and use the gradient descent algorithm to optimize the cross-entropy loss function and calculate the weight matrix W^R and W^M. Finally, the entire process of the PRM algorithm has been completed. In the group feature expression step, we train 6 arithmetic mean classifiers and 6 minus-average relational classifiers to classify group images. In this way, the PRM algorithm not only solves the problem of the group appearance changes caused by the changes in the number and relative positions of group members but also describes the relations among the group members. Fig. 3 illustrates the architecture of the PRM algorithm.

$$P(\hat{y} = k|\hat{h}_j^R) = \frac{exp(w_k \hat{h}_j^R + b_k)}{\sum_{i=1}^{c} exp(w_i \hat{h}_j^R + b_i)} \tag{3}$$

$$P(\hat{y} = k|\hat{h}_j^M) = \frac{exp(w_k \hat{h}_j^M + b_k)}{\sum_{i=1}^{c} exp(w_i \hat{h}_j^M + b_i)} \tag{4}$$

where, $\hat{h}_j^R = \Phi(h_j^R)$, and $\hat{h}_j^M = \Phi(h_j^M)$. $P(\hat{y} = k|\hat{h}_j^R)$ is the probability that the minus-average relational feature $\hat{h}_j^R$ of the j-th part of a group image belongs to the k-th category. $P(\hat{y} = k|\hat{h}_j^M)$ is the probability that the arithmetic mean features $\hat{h}_j^M$ of the j-th part of a group image belongs to the k-th category.

$$\hat{\mathcal{L}}_j^R(x) = -\sum_{k=1}^{c} \mathcal{I}(y,k) log(P(\hat{y} = k|\hat{h}_j^R)) \tag{5}$$

$$\hat{\mathcal{L}}_j^M(x) = -\sum_{k=1}^{c} \mathcal{I}(y,k) log(P(\hat{y} = k|\hat{h}_j^M)) \tag{6}$$

$$\mathcal{L}_j(x) = \sum_{x \in \chi} \hat{\mathcal{L}}_j^R(x) + \sum_{x \in \chi} \hat{\mathcal{L}}_j^M(x) \tag{7}$$

where, $\mathcal{I}(y,k)$. If $y = k$, then return 1, otherwise, return 0. x is a single group image. χ is the set of all training sample images. $\hat{\mathcal{L}}_j^R(x)$ is the loss of the j-th part of one sample that corresponds to the minus-average relational descriptor. $\hat{\mathcal{L}}_j^M(x)$ is the loss of the j-th part of one sample that corresponds to the arithmetic mean descriptor. $\mathcal{L}_j(x)$ is the sum of the total loss of the j-th part of all samples.

C. GROUP FEATURE MATCHING

The research on G-ReID can be divided into two stages. The first stage is feature extraction and classifier training and the second stage is feature matching. In the first stage

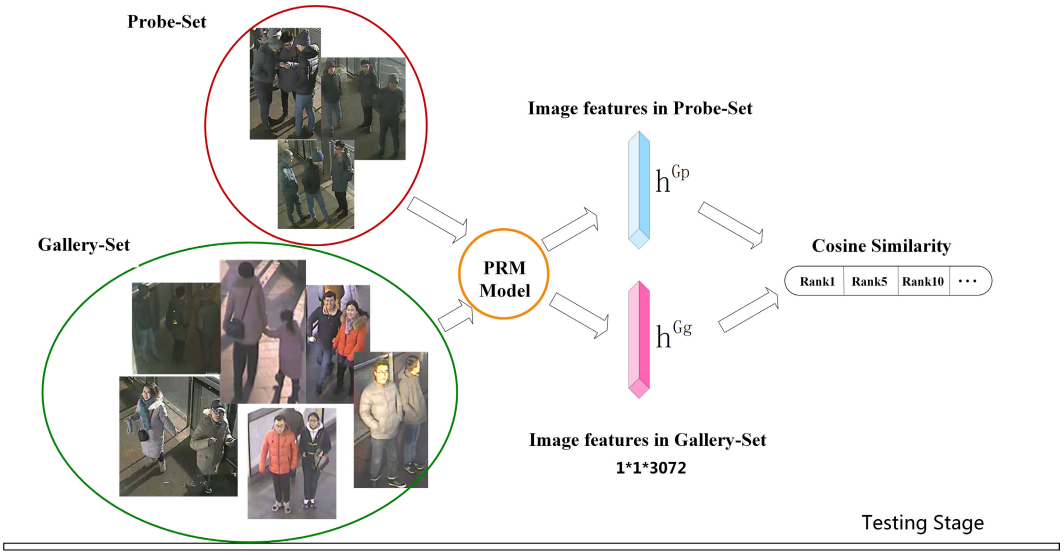


FIGURE 4. The architecture of group feature matching. The figure shows the group image matching process of the G-ReID. For each group of pedestrian data in the BRT test dataset, one image is randomly selected in the gallery set, and the remaining images are selected as the probe set. We use the PRM model to extract group image features from the probe set and gallery set. We use the cosine similarity function to calculate the feature distance and sort the calculation results. Finally, we use CMC as the evaluation index of the PRM model.

of G-ReID, 12 feature classifiers are trained by the PRM algorithm. In the feature-matching stage of the group image, we use Eq.(8) to concatenate 12 local feature vectors from 12 classifiers as the overall appearance features of a group image.

$$h^G = G(\hat{h}_j^R, \hat{h}_j^M) = \hat{h}_j^R \oplus \hat{h}_j^M \tag{8}$$

where, $G(\cdot)$ represents the global descriptor for the concatenation between R and M. $\oplus$ is the concatenation operator.

The test process is shown in Fig. 4. First, we extract a picture from each sample in the test set to form the gallery set. Then, the remaining sample images in the test set constitute the probe set. In this way, there is only one image of the same group in the gallery set and several images of the same group in the probe set. Third, we extract a picture from the probe set and input it into the PRM model to obtain the feature vector of the probe set group image. Next, we input all pictures in the gallery set to the PRM model and obtain all the group feature vectors of the gallery set. The cosine similarity in Eq. (9) is then used to calculate the distance between the feature vectors h^{Gp} of the probe set image and all the image feature vectors h^{Gg} in the gallery set. Because the cosine similarity considers the direction of a vector, it is mostly used in the feature matching of pedestrian ReID. When the cosine similarity is greater, the similarity between the two images is higher. We use the cumulative matching feature (CMC) as the evaluation index of the PRM model. The feature matching process is shown in Fig. 4.

$$\mathcal{S} = \frac{h^{Gg} \odot h^{Gp}}{\|h^{Gg}\| \cdot \|h^{Gp}\|} \tag{9}$$

IV. BRT DATASET

The BRT dataset is used to assess the G-ReID. Because the sampling location of the G-ReID dataset is the transfer station of the city's BRT system, the G-ReID dataset is denoted as the BRT dataset. The BRT dataset expands the research scale of G-ReID research from small-scale extensions of blocks and schools to large-scale urban centers, which is beneficial for the spatial visualization of G-ReID results and understanding group behavior. We describe aspects such as the necessity of BRT dataset collection, the BRT dataset overview and sampling scheme, the ground truth and the challenges of the BRT dataset.

A. NECESSITY OF BRT DATASET COLLECTION

First, the BRT dataset is a G-ReID dataset, and a large number of current pedestrian ReID datasets are single-pedestrian ReID datasets. In real life, people often travel together and their travel activities have social attributes. The task of G-ReID is different from that of single-pedestrian ReID. The research on G-ReID has unique difficulties caused by variations in the number and positions of group members. Second, in the BRT G-ReID dataset, the sampling points cover 5 urban administrative areas in the city center. In previous research, the sampling point referred only to a local space area, such as an airport transfer hall or a corner of a campus. The BRT dataset brings many benefits for future G-ReID research. On the one hand, the BRT dataset is convenient for spatiotemporal feature modeling and matching, which is helpful for improving the efficiency and accuracy of G-ReID. On the other hand, the spatial visualization of G-ReID results is conducive to mining and understanding the potential semantic

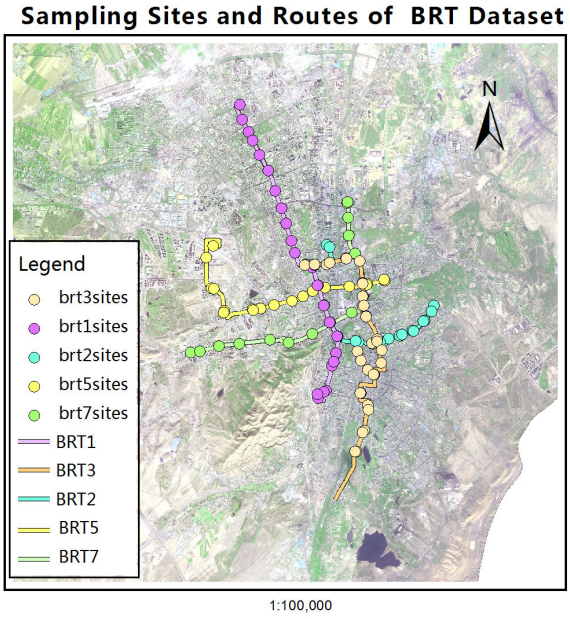


FIGURE 5. Urban spatial distribution of the sampling lines and sites of the BRT dataset. The circles in different colors represent the different BRT stops and the lines in different colors represent the different operating lines of the BRT system. The group images come from the camera group at these sites, which all have real spatial and geographical coordinates.

information of human activities. In addition, the previous ReID datasets involve only internal campus scenes. Students travel in a single way among the school, canteen and dormitory. However, the BRT dataset is different and reflects the style of public travel among people of different ages, genders, occupations and nationalities. Such datasets are the true reflection of society and can combine ReID tasks with tracking to form many trajectories, thereby contributing to the use of sociological, psychological and human geographical knowledge to mine semantic information.

B. BRT DATASET OVERVIEW AND SAMPLING SCHEME

1) BRT DATASET OVERVIEW

The BRT system is a new type of public transportation system between rapid rail transit and conventional public transportation. The collection line of the BRT dataset involves five BRT operating lines in the city. The sampling stations cover the central urban area of the city and include 53 stations, such as Haojia Town, Cultural Palace, Mingyuan, Mobile Company, Youyi, Hongshan, Academy of Sciences and Railway Bureau. The details of the sampling stations and sampling lines of the BRT dataset are shown in Fig. 5.

2) BRT DATASET SAMPLING SCHEME

The project team arranged 13 groups that consisted of 6-8 people in each group. Each group had a sampling route. The team members combined freely to ensure the sufficient diversity and quantity of the group images. During the sample collection process, we recorded the time spent in the station and the station name. In addition, group members transferred to different bus routes at least once. Fig. 6 shows a schematic diagram of the group sampling. The groups appear at different blue and red sites. The blue and red lines indicate the sampling routes.

C. GROUND TRUTH OF THE BRT DATASET

To simplify the research, we do not consider the spatial layout of the cameras in the BRT station when we divide the dataset. We label pedestrians collected by different types of cameras at the same site with the same group label. The dataset is divided into the training set and test set. There are 200 types of group pictures in the training set, each of which is a group, with a total of 1, 870 images. There are also 200 types of group pictures in the test set, each of which is a group, with a total of 1, 340 images. The total number of images in the test set and training set is 3, 210. First, we extract a picture from each sample of the test set to form the gallery set. The remaining sample images in the test set constitute a probe set. In this way, the gallery set has only one image of the same group, while the probe set has several images of the same group. We statistically compare BRT with two existing G-ReID datasets, namely, i-LIDS MCTS[7] and Road Group [58] in TABLE 2.

TABLE 2. Statistical comparisons between the BRT and existing G-ReID datasets.

Datasets	i-LIDS MCTS[54]	Road Group[55]	BRT
Image	274	324	3, 210
Group	64	162	400
Camera	8	2	53
Coordinate	\	\	√

D. CHALLENGES OF THE BRT DATASET

The BRT dataset is collected at a real scene and has practical application value. The real scene is complex and changeable. The collected BRT dataset is subject to interference from natural factors, such as changes in the light, illumination and imaging color. It is also influenced by human factors, such as group members blocking each other, changes in affiliations of the group members, etc. The challenges of the BRT dataset are summarized in 9 aspects, as shown in Fig. 7.

V. EXPERIMENTAL RESULTS

In this section, we introduce the experiment from the following five aspects: the evaluation protocol; pretraining the SSD detector; experiments on the BRT dataset; experiments on the i-LIDS MCTS and Road Group datasets; implementation details and experimental setup.

A. EVALUATION PROTOCOL

Pedestrian ReID approximates matching retrieval or sorting tasks. The basic goal is to use the algorithm model to calculate the distance between the probe image and all images in the gallery set. Then, according to the distance, we obtain a sorted list. The Rank 1 accuracy and cumulative matching characteristic curve are commonly used evaluation

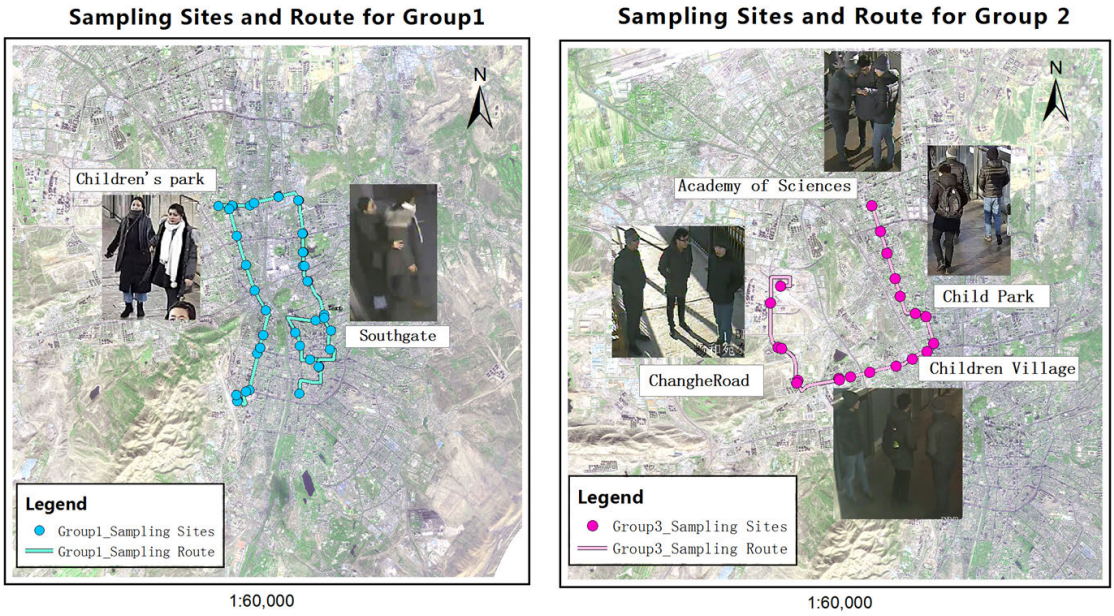


FIGURE 6. Schematic diagram of the group sampling. Group 1 on the left consists of two people. The picture on the left shows that group 1 was photographed by cameras at different BRT stations. Two girls in group 1 passed through Children's Park and South Gate separately. The blue line indicates the operating route of BRT 1. It is roughly inferred that group 1 got on at Southgate and off at Children's Park. The resolution of group 1 captured by the two sites is different. On the right is group 2, which is composed of three people. The figure on the right shows that group 2 passed the four stations: Changjiang Road, Children's Village, Children's Park and the Academy of Sciences. The picture on the right is the distribution map of the re-identification results of group 2 on the urban real geographic map. The pink line indicates the sampling route of group 2. The appearance of group 2 captured by different cameras varies.

FIGURE 7. The challenges of the BRT dataset are summarized in 9 aspects. (a) Self occlusion (b) Relative position changes of the group members (c) Significant color differences (d) Group local appearance information (e) Group members of different sizes (f) Strong light changes (g) Changes in the affiliations of the group members (h) Interference with pedestrian information (i) Group members with different body postures.

indicators to evaluate, quantify and verify pedestrian ReID algorithms. Rank 1 can be regarded as the traditional classification accuracy. However, in actual application scenarios, by solely depending on the ReID algorithm, we cannot achieve a very high Rank 1 value. Therefore, we cannot truly complete the pedestrian identity consistency matching task across cameras. In this case, having the algorithm return a sorted list is a more practical application of ReID. The user

selects the correct matching result from the first N objects on the list. If N is much smaller than the size of the probe set, the ReID algorithm can greatly reduce the labor cost and improve the efficiency of the matching task. The mathematical formula is defined as follows.

$$cmc(N) = \sum_{n=1}^{N} r(n) \tag{10}$$

where $r(n)$ represents the probability that the n-th element on the sorted list is consistent with the identity of the target to be queried. With N as the abscissa and CMC(N) as the ordinate, the CMC curve can be drawn and it is easy to find that CMC(1) is the same as Rank 1.

B. IMPLEMENTATION DETAILS AND EXPERIMENTAL SETUP

The backbone network of PCB uses ResNet50 [8]. The batch size is set to 32 and the person images are resized to 384×128 as inputs. The total training of the PRM algorithm is 60 epochs and the basic learning rate is 0.05. After 30 epochs of training, the learning rate decays to 0.005. Because the i-LIDS MCTS dataset does not have a training set, the article [4] treats the entire i-LIDS MCTS dataset as the test set. Therefore, to ensure that the evaluation ground truths are the same, we use the G-ReID model trained on the BRT dataset and test it directly on the i-LIDS MCTS dataset. We use the labeled detection frame of the Road Group dataset to obtain the information of a single person. Next, we execute the PRM algorithm on the Road Group dataset. The final result is obtained by averaging the results of 10 random splits. We use the cumulative matching characteristics (CMC) as the evaluation metric.

C. PRETRAINING THE SSD DETECTOR

The BRT dataset does not have bounding-box detection. Therefore, before the pedestrian ReID work is carried out, the members of the group must be detected. To improve the pedestrian detection accuracy of the SSD, we pretrain the SSD detector on pedestrian datasets, such as PRW [56] and INRIA [53]. INRIA is currently the most used static pedestrian detection database and provides original pictures and corresponding annotation files. These high-resolution images come from GRAZ-01 and Google. The PRW pedestrian ReID dataset is an extension of the Maretk1501 dataset. The test results are shown in Fig. 8. Pretraining the SSD model on the INRIA pedestrian dataset is more helpful for the G-ReID task, because the INRIA dataset is similar to the BRT dataset, and includes pedestrians with complete bodies, while the PRW pedestrian dataset is different, and contains many cropped images at different scales.

D. EXPERIMENTS ON THE BRT DATASET

1) EVALUATION COMPONENT OF THE PRM ALGORITHM ON THE BRT DATASET

Because the final PRM algorithm is determined by a combination of strategies such as detector pretraining, PCB local feature localization and extraction strategies and group feature descriptors, we must identify the optimal combination of these strategies to ensure the best performance of the PRM algorithm. We test and compare the components of the PRM algorithm on the BRT dataset. TABLE 3 shows the test results for the component of the algorithms on the BRT dataset, where R-k (k = 1, 5, 10) denotes the Rank-k accuracy(%).

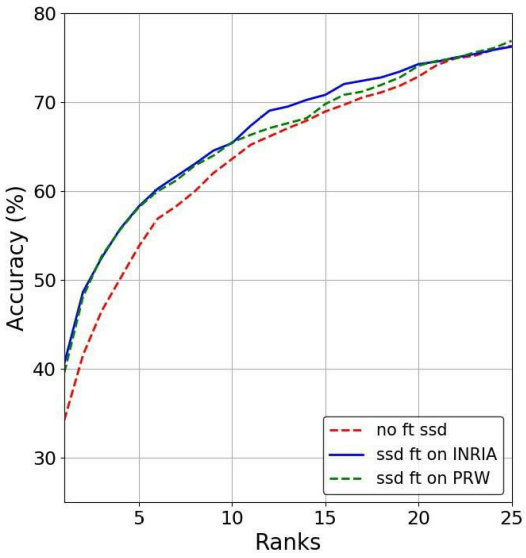


FIGURE 8. Contribution of pretrained pedestrian detection to improve the accuracy of G-ReID. The abscissa is the rank and the ordinate is the accuracy. The red polyline represents the accuracy of G-ReID with the SSD model without additional fine-tuned training. The blue and green polylines represent the accuracy of G-ReID with the SSD model pretrained for pedestrian detection on the INRIA and PRW datasets, respectively.

The PRM algorithm is based on the baseline algorithm, by adding some innovative elements. For the baseline algorithm, we choose the ResNet50 [8] and PCB [37] networks because the ResNet50 has low complexity and good performance. It is the baseline of the ILSVRC [57] and COCO2015 [54] competitions and is ranked first in ImageNet detection, local positioning and segmentation tasks. The PCB algorithm is a recognized baseline in the field of ReID. We evaluate and compare the PRM algorithm with the existing baseline algorithms on the BRT dataset. The R50B algorithm does not perform detection and inputs the entire picture into the ResNet50 network to extract group features. The PCB algorithm does not perform detection and sends the entire picture directly to the PCB network to extract group features. As shown in TABLE 3, the performance of the G-ReID algorithm that uses the PCB network is better than the performance of the G-ReID algorithm that uses R50B, because the ResNet50 feature extraction network does not specify the classification target. The targets in these images are people, animals, vehicles, objects etc. By contrast, the PCB baseline algorithm is designed for single-person ReID. Therefore, we choose the PCB network as the basis for the design of the PRM algorithm.

FIGURE 9. Example of the correct matching of G-ReID on the BRT dataset. The blue box indicates the pedestrians who should be probed during the test phase and the red box indicates the pedestrians that the probe image matches correctly in the gallery set. The ranking images are sorted from left (Rank 1) to right (Rank 5).

TABLE 3. Ablation study of the proposed PRM method. The matching accuracy values (%) at Rank(r) = 1, 5, 10 are shown on the BRT datasets. The best results are shown in black boldface font.

	BRT-dataset		
Method	Rank 1	Rank 5	Rank 10
R50B	47.4	68.2	78.0
PCB	48.2	66.6	74.7
SSD-PCB	34.2	53.8	63.6
SSD-INR-PCB	41.5	59.2	65.2
SSD-INR-PCB+M	49.5	68.3	75.2
SSD-INR-PCB+M+U	45.7	64.5	72.7
PRM(SSD-INR-PCB+M+R)	**50.6**	**70.9**	**78.1**

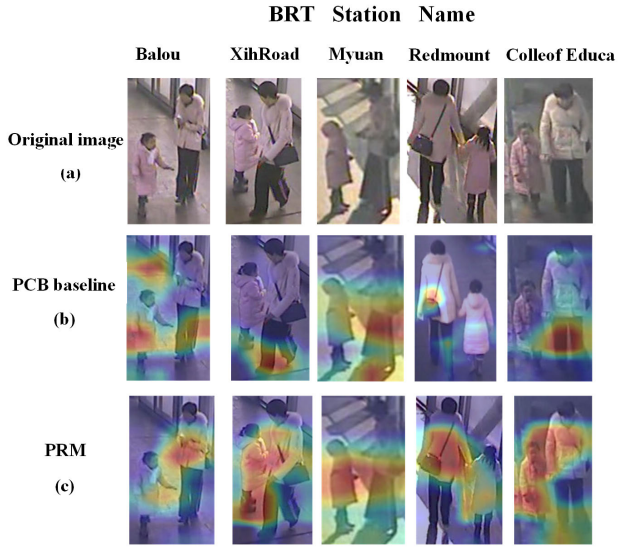


FIGURE 10. Visualization of the class activation maps (CAMs). The CAM of (a) indicates the original images from the BRT dataset. The CAMs of (b) and (c) are generated by the PCB baseline and PRM model, respectively.

The PRM algorithm adds relational descriptors based on the baseline PCB algorithm. The SSD-PCB algorithm detects pedestrians by using an SSD detector without pretraining and inputs the detected pedestrians into the PCB network to extract group features. A lower pedestrian detection rate leads to a lower ReID accuracy rate. The SSD-INR-PCB+M denotes the strategy of using the INRIA dataset to pretrain the SSD model of pedestrian detection and using the arithmetic mean descriptor to extract group features. The SSD-INR-PCB+M+U algorithm uses the SSD pedestrian detection model pretrained on the INRIA dataset, the arithmetic mean descriptor and the max relational descriptors to extract group features. In the architecture of the max relational descriptor, we first calculate the difference between the local features and then select the max feature value. The PRM (SSD-INR-PCB+M+R) denotes the strategy of using the INRIA dataset to pretrain the SSD model of pedestrian detection and using the minus-average relational feature and the arithmetic mean descriptor to extract features. Theoretically, the max relational descriptor mainly gives attention to the salient line features of the group, for example, the overall outline information of a pedestrian. However, the minus-average relational descriptor and the arithmetic mean descriptor mainly concentrate on the overall information of the group. The content of the overall information is more abundant than the contour information. Therefore, the expression ability of the minus-average relational descriptor and arithmetic mean descriptor are stronger than the expression ability of the max relational descriptor. The performance of PRM is better than the performance of SSD-INR-PCB+M, which indicates

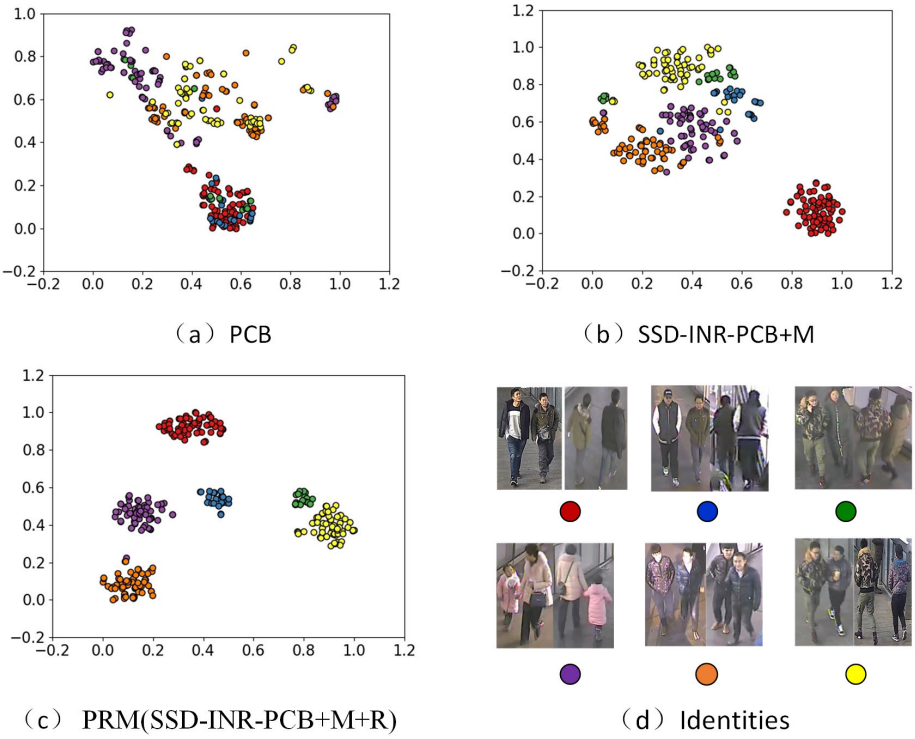


FIGURE 11. The t-SNE visualization of high-dimensional group features obtained by the PCB, SSD-INR-PCB+M and PRM from the BRT dataset. Each color represents an identity randomly chosen from the unseen training BRT dataset.

that the minus-average relational descriptor is an effective accumulation factor. The final evaluation results show that the PRM algorithm performs the best.

2) VISUALIZATION

The visualization of the PRM model is presented for the evaluation on the BRT dataset. To express the algorithm performance intuitively, we show a typical test-phase case in Fig. 9. The blue box image from the probe set is the group image that must be queried during the test phase. The red frame picture from the gallery set is the group image that was correctly matched during the test phase. The matching process is performed to calculate the cosine distance between the blue box image and all the group images in the gallery set and then to sort the calculation results. When the cosine similarity is higher, the similarity of the two images is greater. We sort the calculated similarity values, the value with the highest similarity is named Rank 1 and the value with the second-highest similarity is named Rank 2. The sorted results are displayed from left (Rank 1) to right (Rank 5). The evaluation of the visualization also demonstrates the performance of the algorithm. Fig. 9 is the visualization effect of the PRM model that continuously hits an index target on the BRT dataset during the retrieval of the gallery set.

a: VISUALIZATION OF THE CLASS ACTIVATION MAP

We visualize the class activation maps (CAM) in Fig. 10 by using Grad-CAM [58]. The visual group images were taken at different sites of the BRT and the members of the groups in the images were accompanied by mutual occlusion and relative position changes. We use the PCB baseline algorithm to process the entire group image to obtain the CAM activation map (b). We use the PRM algorithm to detect the group members and construct relations for the detected group members to obtain the CAM activation map (c). Compared with the PCB baseline algorithm, the proposed method has a higher activation in the same discriminative area. The PRM algorithm can more effectively capture the relationship information between group members. The PRM algorithm can also show independent hotspots in group members and thermal transitions between group members. This indicates that the proposed method can focus on more discriminative cues.

b: THE T-SNE VISUALIZATION OF GROUP FEATURES

To determine whether the PRM descriptors are effective in group classification from the perspective of feature dimensionality reduction, we use the PCB, SSD-INR-PCB+M and PRM algorithms to extract group features and then use the t-SNE algorithm to visualize these extracted features. The t-SNE algorithm [59] creates a single map that reveals structures at many different scales. This is particularly important for high-dimensional data that lie on several different but also related low-dimensional manifolds, such as images of objects from multiple classes seen from multiple viewpoints. We randomly selected 6 groups of data and used the t-SNE

TABLE 4. Comparison with state-of-the-art G-ReID methods. The best results (%) are in black boldface font.

	i-LIDS MCTS				Road Group			
Method	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
CRRRO-BRO [2]	23.0	44.6	57.5	73.2	17.8	34.6	48.1	62.2
Covariance [5]	26.5	52.5	66.0	90.9	38.0	61.0	73.1	82.5
PREF [3]	31.1	49.5	60.3	70.2	43.0	68.7	77.9	85.2
PRM	**38.6**	**58.7**	**69.4**	**82.8**	**62.4**	**75.9**	**82.1**	**88.3**

algorithm to perform a visualization on the training set. The PCB algorithm does not perform detection but rather sends the entire picture directly to the PCB network to extract group features. The test results show that the arithmetic mean descriptor and the minus-average relational descriptor are effective at group classification tasks. Moreover, the performance of the PRM algorithm based on the two descriptors is the best. The PRM algorithm can pull the images from the same identity closer while pushing different identities away from each other.

E. EXPERIMENTS ON THE I-LIDS MCTS AND ROAD GROUP DATASET

In this section, we introduce a comparison between the PRM model and the state-of-the-art algorithms for G-ReID on the i-LIDS MCTS dataset [54] and Road Group dataset [55].

1) DATASET DESCRIPTION

As shown in Fig. 12, the i-LIDS MCTS dataset [54] was captured by the multicamera CCTV network in the airport arrival hall. The dataset was captured from two non-overlapping camera views. A total of 64 groups were extracted and 274 images were cropped. For most groups, four images are available that are from different camera views or from the same camera but are captured at different locations and different times. The capture of these images was subject to large variances in light and occlusion.

The Road Group dataset [55] consists of 162 group pairs taken from a 2-camera view of a crowded road scene. The bounding box coordinates of a total of 1, 099 pedestrians are also provided. The Road Group dataset includes severe object occlusions and large variations in group layout.

2) COMPARISON WITH STATE-OF-THE-ART METHODS

To evaluate the G-ReID performance, we compare the PRM model with the following state-of-the-art methods: CRRRO-BRO [2]; Covariance [5]; and PREF [3]. In the previous methods, the group features were designed by hand. The CRRRO-BRO descriptor attempts to obtain a stable representation against a relative position change between the couple and BRO descriptor is robust to the changes in non-center-rotation. In this case, CRRRO-BRO achieves decent performance on the i-LIDS MCTS dataset, while the most groups in this dataset contain two pedestrians. The principle of the covariance descriptor [5] is to measure the similarity of two groups by calculating the difference in the covariance matrix. Because the calculation of the covariance

FIGURE 12. Snapshots of the utilized datasets. The left is the i-LIDS MCTS dataset. The right is the Road Group dataset. Each row of each dataset shows a few snapshots with the same group ID. Each column represents different groups.

matrix is based on local pixel values, it is highly susceptible to interference from background information. Consequently, the performance of the covariance descriptor is limited. PREF (pooling residuals of encoded feature) [3] uses a feature dictionary to express single-person features and then transfers them for group appearance coding. The effect of PREF is limited because changes in group appearance are more complicated than changes in individual pedestrian appearance. As shown in TABLE , the accuracy of the G-ReID is higher with the PRM model than with the state-of-the-art algorithms because the PRM algorithm solves the problems of changes in the number and relative positions of the members within the group. Therefore, PRM is valid for most group datasets, regardless of whether the group in the dataset contains 2 pedestrians or multiple pedestrians.

VI. CONCLUSION

In this paper, we have proposed the PRM algorithm. First, it is based on the local features that are conducive to expressing the internal structure of the human body. Second, the arithmetic mean descriptor and the minus-average relational descriptor solve the G-ReID problem caused by changes in the number and relative positions of group members. Third, the minus-average relational descriptor can describe the differences in the appearance of the group members. The PRM algorithm has a simple structure and has effective performance on the G-ReID task evaluated by test experiments. In addition, considering the rarity of G-ReID datasets and the requirement to improve the applicability of the G-ReID

algorithm in real-life scenarios, we have contributed the BRT G-ReID dataset.

In the open world, G-ReID must work with more complex scenarios and more diverse problems. The PRM algorithm is a supervised learning method based on a group dataset with label information. However, there are very few datasets for G-ReID with label information. Therefore, future work can consider using semi-supervised learning and transfer learning methods to address variations in group appearance. Moreover, in real-life scenarios, considerable multi-source information is available for G-ReID. For example, the geographic spatio-temporal information is also a beneficial constraint condition. Therefore, in the next stage, we will integrate spatio-temporal constraint information with image appearance information to address the challenges of G-ReID.

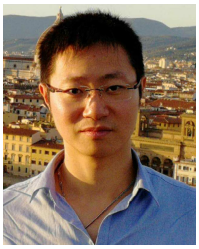
REFERENCES

[1] Z. Wang, W. Liu, Y. Matsui, and S. Satoh, "Effective and efficient: Toward open-world instance re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4789–4790.
[2] W. S. Zheng, S. Gong, and T. Xiang, "Group association: Assisting re-identification by visual context," *J. Vis.*, vol. 14, p. 1132, 2014.
[3] G. Lisanti, N. Martinel, A. Del Bimbo, and G. L. Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2449–2458.
[4] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.
[5] Y. Cai, V. Takala, and M. Pietikainen, "Matching groups of people by covariance descriptor," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2744–2747.
[6] F. Zhu, Q. Chu, and N. Yu, "Consistent matching based on boosted salience channels for group re-identification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4279–4283.
[7] M. Koperski, S. Bak, and P. Carr, "Groups re-identification with temporal context," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 209–217.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[10] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737
[11] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.
[12] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino, "Analyzing groups: A social signaling perspective," in *Video Analytics for Business Intelligence*. 2012.
[13] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PLoS ONE*, vol. 5, no. 4, Apr. 2010, Art. no. e10047.
[14] M. L. Federici, A. Gorrini, L. Manenti, and G. Vizzari, "Data collection for modeling and simulation: Case study at the University of Milan-Bicocca," in *Proc. Int. Conf. Cellular Automata*, 2012, pp. 699–708.
[15] D. Brscic and T. Kanda, "Changes in usage of an indoor public space: Analysis of one year of person tracking," *IEEE Trans. Hum.-Mach. Syst.*, vol. 45, no. 2, pp. 228–237, Apr. 2015.
[16] J. Sochman and D. C. Hogg, "Who knows who–inverting the social force model for finding groups," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 830–837.
[17] Z. Yucel, F. Zanlungo, C. Feliciani, A. Gregorj, and T. Kanda, "Identification of social relation within pedestrian dyads," *PLoS ONE*, vol. 14, no. 10, Oct. 2019, Art. no. e0223656.
[18] S. D. Khan, G. Vizzari, S. Bandini, and S. Basalamah, "Detection of social groups in pedestrian crowds using computer vision," Complex Syst. Artif. Intell. Res. Centre, Univ. degli Studi di Milano–Bicocca, Milan, Italy, Tech. Rep., 2015.
[19] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 995–1008, May 2016.
[20] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2010, pp. 435–440.
[21] S. J. Mckenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Comput. Vis. Image Understand.*, vol. 80, no. 1, pp. 42–56, 2000.
[22] J. S. Marques, P. M. Jorge, A. J. Abrantes, and J. M. Lemos, "Tracking groups of pedestrians in video sequences," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2003, p. 101.
[23] K. O. Arras, B. Lau, S. Grzonka, M. Luber, and O. M. Mozos, "Range-based people detection and tracking for socially enabled service robots," in *Towards Service Robots for Everyday Environments*. Berlin, Germany: Springer, 2012.
[24] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino, "Joint individual-group modeling for tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 746–759, Apr. 2015.
[25] K. Gilholm, S. Godsill, S. Maskell, and D. Salmond, "Poisson models for extended target and group tracking," *Proc. SPIE*, vol. 5913, Sep. 2005, Art. no. 59130R.
[26] B. Lau, K. O. Arras, and W. Burgard, "Tracking groups of people with a multi-model hypothesis tracker," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 3180–3185.
[27] F. Luo, H. Huang, Y. Duan, J. Liu, and Y. Liao, "Local geometric structure feature for dimensionality reduction of hyperspectral imagery," *Remote Sens.*, vol. 9, no. 8, p. 790, Aug. 2017.
[28] G. Shi, H. Huang, and L. Wang, "Unsupervised dimensionality reduction for hyperspectral imagery via local geometric structure feature learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1425–1429, Aug. 2020.
[29] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial-spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.
[30] D. Gray and T. Hai, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.
[31] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.
[32] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by Fisher vectors for person re-identification," in *Proc. 12th Int. Conf. Comput. Vis.*, 2012, pp. 413–422.
[33] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local Fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.
[34] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[35] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.
[36] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[37] Y. Sun, L. Zheng, Y. Yi, and T. Qi, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.
[38] Y. M. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 402–419.
[39] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.
[40] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.

[41] J. Deng, N. Ding, Y. Jia, and A. Frome, "Large-scale object classification using label relation graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 48–64.
[42] N. Ding, J. Deng, K. P. Murphy, and H. Neven, "Probabilistic label relation graphs with ising models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1161–1169.
[43] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5532–5540.
[44] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1268–1277.
[45] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.
[46] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, "Improving deep visual representation for person re-identification by global and local image-language association," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 54–70.
[47] W. Fei, Z. Zhao, and F. Su, "Deep global and local saliency learning with new re-ranking for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2018, pp. 1–6.
[48] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.
[49] Z. Huang, Z. Wang, W. Hu, C.-W. Lin, and S. Satoh, "DoT-GNN: Domain-transferred graph neural network for group re-identification," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1888–1896.
[50] Z. Huang, Z. Wang, C.-C. Tsai, S. Satoh, and C.-W. Lin, "DotSCN: Group re-identification via domain-transferred single and couple representation learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 15, 2020, doi: 10.1109/TCSVT.2020.3031303.
[51] X. Wang and R. Zhao, "Person re-identification: System design and evaluation overview," in *Person Re-Identification*. 2014.
[52] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[53] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, 2005, pp. 886–893.
[54] T.-Y. Lin, M. Maire, S. Belongie, and J. Hays, "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*. [Online]. Available: https://arxiv.org/abs/1405.0312
[55] W. Lin, Y. Li, H. Xiao, J. See, J. Zou, H. Xiong, J. Wang, and T. Mei, "Group re-identification with multi-grained matching and integration," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogint.*, 2019, pp. 1–15.
[56] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," 2017, *arXiv:1604.02531*. [Online]. Available: https://arxiv.org/abs/1604.02531
[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[58] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, and D. Parikh, "Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1611.07450*. [Online]. Available: https://arxiv.org/abs/1611.07450
[59] L. Van Der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

PING HU received the degree in cartography and geographic information systems from Xinjiang University, in 2009. She is currently pursuing the Ph.D. degree in cartography and geographic information with the Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences. She worked with the Urumqi Land Resources Survey and Planning Institution where she was mainly responsible for county-level land use spatial plan projects. Her research interests include spatio-temporal data mining and group re-identification.

HONGWEI ZHENG is currently a Researcher with the Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences. He has published more than 100 articles in the fields of computer vision, remote sensing and geography, including interdisciplinary research articles in high-impact journals and at the top conferences. His research interests include the development of machine learning and large-scale computational algorithms, including automated learning and decision making in high-dimensional and spatio-temporal dynamic worlds in interdisciplinary fields.



WEISHI ZHENG (Member, IEEE) received the Ph.D. degree in applied mathematics from Sun Yat-sen University, in 2008. He is currently a Professor with Sun Yat-sen University. He has published more than 100 articles in major journals, such as TIP, TSMC-By, and PR, and at top conferences, such as ICCV, IJCAI, and AAAI. His research interests include person/object association and activity understanding in visual surveillance.