Similarity Metric Learning For RGB-Infrared Group Re-Identification

Jianghao Xiong[1], Jianhuang Lai[1,2,3,4*]
[1]School of Computer Science and Engineering, Sun Yat-Sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Key Laboratory of Video and Image Intelligent Analysis and Applicaiton Technology,
Ministry of Public Security, China
xiongjh7@mail2.sysu.edu.cn, stsljh@mail.sysu.edu.cn

Abstract

Group re-identification (G-ReID) aims to re-identify a group of people that is observed from non-overlapping camera systems. The existing literature has mainly addressed RGB-based problems, but RGB-infrared (RGB-IR) cross-modality matching problem has not been studied yet. In this paper, we propose a metric learning method Closest Permutation Matching (CPM) for RGB-IR G-ReID. We model each group as a set of single-person features which are extracted by MPANet, then we propose the metric Closest Permutation Distance (CPD) to measure the similarity between two sets of features. CPD is invariant with order changes of group members so that it solves the layout change problem in G-ReID. Furthermore, we introduce the problem of G-ReID without person labels. In the weak-supervised case, we design the Relation-aware Module (RAM) that exploits visual context and relations among group members to produce a modality-invariant order of features in each group, with which group member features within a set can be sorted to form a robust group representation against modality change. To support the study on RGB-IR G-ReID, we construct a new large-scale RGB-IR G-ReID dataset CM-Group. The dataset contains 15,440 RGB images and 15,506 infrared images of 427 groups and 1,013 identities. Extensive experiments on the new dataset demonstrate the effectiveness of the proposed models and the complexity of CM-Group. The code and dataset are available at: https://github.com/WhollyOat/CM-Group.

1. Introduction

Group re-identification (G-ReID) is the problem of associating a group of people that appears in disjoint camera views. The significant importance in video surveillance has yield increasing attention and research efforts by the community [28, 30, 36]. Compared to single-person re-identification (ReID) which deals with a single person, generally G-ReID regards 2 to 6 people as a group and treats two groups with at least 60% the same individuals as the same group. Hence, the main challenge of G-ReID is to construct robust representations of groups with appearance changes and group topology changes.

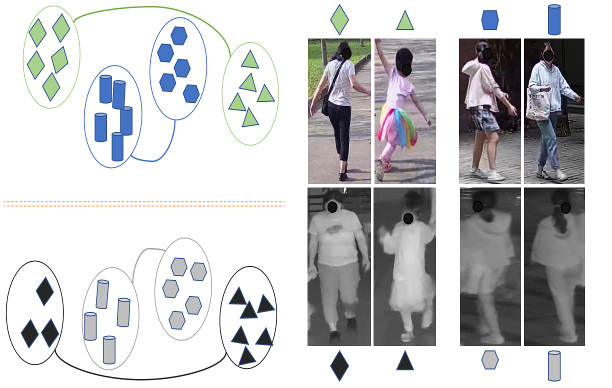


Figure 1. The members within a group are independent in appearance and the visual similarity of two people does not indicate whether they are in a group. In Figure 1b , images in the same column are from the same person. Images with the same color are from the same group.

The existing works based on current G-ReID datasets have made impressive strides, but there still remain several important issues need to be resolved. First, available datasets for G-ReID are limited in different aspects. For example, the extant largest dataset for G-ReID City1M [36] is synthesized and has huge domain gap between real images. The commonly used real datasets such as Road Group [28], DukeMTMC Group [28] and CSG [30] are limited in amount of groups and images. Therefore, for real

[*]Corresponding Author

applications, it is in need of a simulation of real scenarios which includes large amount of people and variable scenes.

Another challenge we notice is that although G-ReID tackles group matching problem, most deep-learning-based works rely on labels of individuals to train deep nets for feature extraction [8,40]. The fully supervised training scheme requires very large amounts of labour resources to make annotations, which is cost and time-consuming. This disadvantage makes it very hard to construct large-scale dataset and impedes development of G-ReID.

To facilitate the study of G-ReID towards real-world applications, we first introduce the RGB-infrared cross-modality group re-identification (RGB-IR G-ReID) problem. Since infrared mode is widely used by surveillance cameras at night, matching infrared images captured in dark scenes with RGB images captured in bright scenes has been an significant problem for ReID and G-ReID research. As shown in Figure 1, RGB images have a huge domain gap between infrared images. Meanwhile the appearances of group members are independent of each other. Therefore RGB-IR G-ReID not only handles modality discrepancy but also faces challenges from group retrieval. When person labels are available, we propose the Closest Permutation Matching (CPM) framework. We adopt a state-of-the-art RGB-IR ReID method MPANet to train a person feature extractor and model each group as a set of group member features. To measure the similarity of two groups, we calculate the Closest Permutation Distance (CPD) between two sets of extracted features. CPD is a new metric that represents the least distance of two sets of features under all permutations. In the weak-supervised case without person labels, we do not know the identities of group members, which makes it hard to train a person feature extractor. So we propose a Relation-aware Module (RAM) to extract order of group members which is invariant to modality changes. RAM calculates visual relations between individuals within a group to generate pseudo order and guide the network to learn intrinsic orderings within groups.

Furthermore, we have collected a new dataset called Cross-Modality Group ReID (CM-Group) dataset. Compared to existing G-ReID datasets, CM-Group has several new features. 1) CM-Group contains 15,440 RGB images, 15,506 infrared images, 427 groups and 1,013 persons, which is, to our best knowledge, the first RGB-IR cross-modality G-ReID dataset and the largest real-world G-ReID dataset. 2) The raw videos are captured by 6 cameras at 6 different scenes over a time span of 6 months, including large variations of illumination and viewpoint, clothes changes and scale changes. 3) All images are original frames of raw videos, *i.e.* all background information is reserved, which enables researchers to mine useful information in background. More details of CM-Group will be discussed in Section 4.

The main contributions of this work include:

- We propose the Closest Permutation Matching (CPM) to find the best match of group images with the permutation-invariant metric Closest Permutation Distance (CPD). The CPM is resistant to group layout changes and achieves excellent performances on CM-Group.
- We introduce the problem of G-ReID without person labels and propose the Relation-aware Module (RAM) to leverage mutual relations of group members. Our experiments show that RAM can extract a modality-invariant order of members in a group regardless of appearance and layout changes.
- We contribute a large-scale RGB-IR cross-modality G-ReID dataset CM-Group, which supports more comprehensive study on G-ReID.

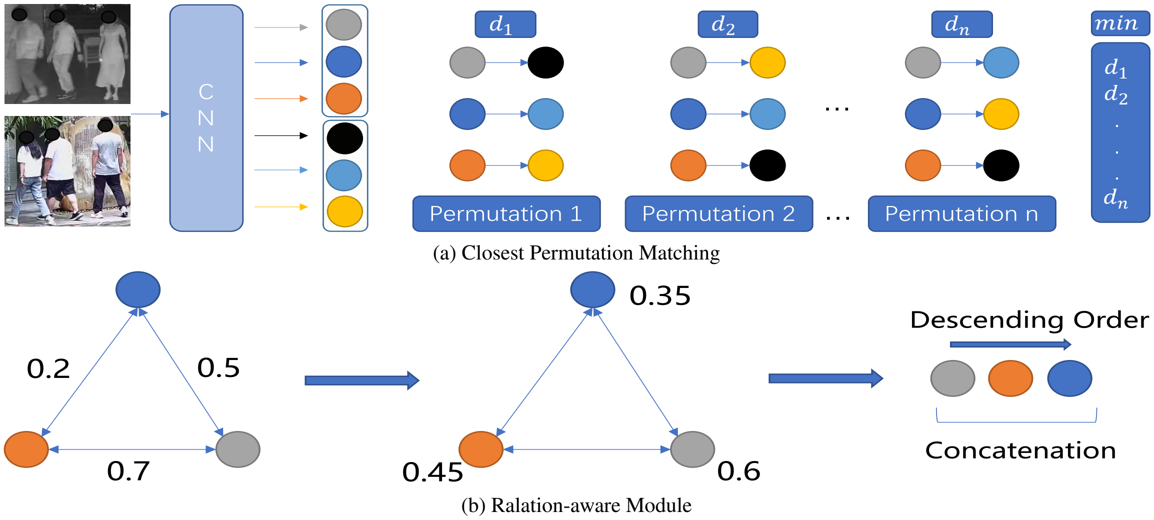
2. Related Work

2.1. RGB-Infrared Person Re-Identification

The modality discrepancy between RGB and infrared images is the main challenge to RGB-IR ReID, thus many methods have been proposed to reduce inter- and intra-modality variations. Wu *et al.* [25] release the first large-scale RGB-IR cross-modality ReID dataset SYSU-MM01 and propose a one-stream deep zero-padding network to learn domain-specific features. Ye *et al.* [33] propose a two-stream network BDTR that simultaneously learns inter- and intra-modality information to ensure the discriminability of features in cross-modality matching. Another two-stream network DDAG [32] is proposed to aggregate modality information by attention mechanism. Lu *et al.* [13] propose to transfer both modality-shared and modality-specific features within and cross modalities to extract discriminative shared and specific features of each modality. Wu *et al.* [26] introduce two attention modules to alleviate modality variance and extract modality-invariant patterns. Generative models have also been used in RGB-IR ReID. GAN-based methods, *e.g.* cmGAN [2], D^{2}RL [22], AlignGAN [20] and JSIA-ReID [21] adopt generative adversarial networks [6] to reduce modality discrepancy. Li *et al.* [11] and Wei *et al.* [24] use generator modules to generate the third modality to bridge RGB and infrared modality. RGB-IR ReID has made great progress. However, RGB-IR G-ReID has not been studied. Our CM-Group is the first RGB-IR dataset for G-ReID.

2.2. Group Re-Identification

Associating groups of people in non-overlapping camera views is first introduced in [38]. The authors combine two descriptors based on a code book of feature patterns to



(a) Closest Permutation Matching

(b) Ralation-aware Module

Figure 2. Figure 2a shows the framework of Closest Permutation Matching (CPM). CPM first trains a CNN to get a feature extractor, then calculates Closest Permutation Distance for all query-gallery image pairs. Figure 2b presents the Relation-aware Module (RAM) which constructs an ordering for each image using intra-group visual similarity. The pseudo order helps the network to learn latent relations among group members.

form the group representation. Since then, the literature of G-ReID is limited to a few works, *e.g*. Covariance Descriptor [1], Salience Channels [39] and PREF [12]. All above methods are based on hand-craft features.

Since Xiao *et al*. [28] publish two datasets DukeMTMC Group and Road Group, deep learning based methods have been the mainstream approach in G-ReID. To overcome layout and membership changes, most works extract feature of each person at the beginning of their networks. MGR [28] uses single-person features to construct multi-grain representations for a group of people and conducts multi-grain matching. DotSCN [9] fuses all single image pairs within groups for group representations. Graph Neural Network [16] (GNN) are applied in G-ReID by DotGNN [8], GCGNN [40] and MACG [30]. Specifically, DotGNN transfers the style of images in source domain to target domain and aggregates individual features into group features. GCGNN exploits neighborhood information as group context to build graph representations. MACG learns graph context using inter- and intra-group attention. The authors also present a large-scale G-ReID dataset CSG. With the development of transformer [18], transformer-based 3DT [36] utilize 3D layout information and the authors release the first synthesized G-ReID dataset City1M, which contains 3D annotations and is currently the largest G-ReID dataset.

Different with above methods, we model an image as a set of single-person features and focus on order of group members in the sets. We propose a set similarity metric for group matching and a weak-supervised relation-aware method to learn intrinsic orderings for group representations.

2.3. Permutation Invariant Models

The permutation invariant models are invariant to order changes of elements in the input, which is a useful property for modelling groups against layout changes. To handle cases when the inputs or outputs are permutation invariant sets, DeepSets [34] defines objective functions on sets that are invariant to permutations. Set Transformer [10] applies self-attention on sets and aggregates features by multi-head attention. JanossyPooling [15] and DuMLP-Pin [4] focus on feature aggregation method to obtain permutation-invariant features. GNN-based methods such as DGCNN [35], GraphTrans [27] and WGDL [37] aggregate node features into graph representations for graph classification regardless of permutations of nodes. The most relevant work with our approach is TAPNet [5], which explicitly considers graph topology by generating locality with neighboring information. However, how to effectively explore topological structure underneath unordered groups has not been addressed. To this end, we propose to use visual relations to define the order of individuals within a group.

3. Methodology

3.1. Problem Formulation and Overview

Let $\mathcal{V} = \{I^v\}$ and $\mathcal{R} = \{I^r\}$ denote the image set of RGB and infrared modality in a RGB-IR G-ReID dataset respectively. An image I of group G contains a set of people $S_I = \{p_1, p_2, ..., p_N\}$, where N is the number of people in image I and p_i $(i \in \{1, 2, ..., N\})$ denotes the bounding box of an individual. Figure 2 illustrates the proposed methods.

Under full-supervised setting, the ground-truth label set for each image I is denoted by $L_I = \{y_1, y_2, ..., y_N\}$, y_i corresponds to exact identity of p_i. We propose the Closest Permutation Matching (CPM) method to compute the similarity of two images by feature permutation and distance imputation. CPM is invariant to group layout changes and resistant to group member changes.

When the corresponding label for each person L_I is not available, *i.e.* only group labels and bounding boxes are provided, it turns to be a weak-supervised problem. We propose the Relation-aware Module (RAM) to learn the intrinsic order of group members from visual relations. The order is supposed to be consistent cross modality and lead to robust group representation.

3.2. Closest Permutation Matching

As shown in Figure 2a, we model a group as a set of people, thus the feature of an image I can be represented as a set $F_I = \{f_1, f_2, ..., f_N\}$, each feature f_i is extracted by a network $\mathcal{F}$.

First, we focus on feature extraction. The ground-truth labels of each person help to train a RGB-IR ReID network which extracts features of single person.

Definition 1. *A feature extractor $\mathcal{F}$ is regarded* ***Good*** *if for an feature set F of any person processed by $\mathcal{F}$, the intra-person distance is smaller than inter-person distance with any other person.*

Assuming we already have a Good feature extractor $\mathcal{F}$, we can obtain all RGB and infrared feature sets of each person. According to Definition 1, the following inequality holds:

$$\max\left\{D\left(f_a^{y_i}, f_b^{y_i}\right)\right\} < \min\left\{D\left(f_a^{y_i}, f_b^{y_j}\right)\right\}, \qquad (1)$$

where $D\left(\cdot\right)$ is a distance function, y_i and y_j are different identities, f_a and f_b are features of different images.

Given two images I_1 and I_2 with feature sets $F_{I_1} = \{f_1, f_2, ..., f_N\}$ and $F_{I_2} = \{f_1^*, f_2^*, ..., f_M^*\}$ respectively, we have to find the smallest distance between two sets. Here, we denote $D\left(\cdot\right)$ as the sum of Euclidean distance between elements at correspondence position of two sets, i.e. $D\left(\cdot\right) = \sum d\left(f_a, f_a^*\right)$. When N equals to M, the Closest Permutation Distance (CPD) between F_{I_1} and F_{I_2} is defined as

$$D_{cp}\left(F_{I_1}, F_{I_2}\right) = \min\left\{D\left(F_{I_1}, \pi F_{I_2}\right)\right\}, \qquad (2)$$

where $\pi \in \prod_{I_2}^{F}$, $\prod_{I_2}^{F}$ represents the set of all permutations of the feature set F_{I_2}.

When N is not equals to M, without loss of generality, let $N < M$. For each permutation π of F_{I_2}, we select the first N elements in F_{I_2} to form a subset E_{I_2}. The CPD between F_{I_1} and E_{I_2} is represented by

$$D_{cp}\left(F_{I_1}, E_{I_2}\right) = \min\left\{D\left(F_{I_1}, \pi^* E_{I_2}\right)\right\}, \qquad (3)$$

where $\pi^* \in \prod_{I_2}^{E}$. For the rest $M - N$ elements, we add a distance imputation term D_{im} which penalises the inequity of cardinality of F_{I_1} and F_{I_2}. D_{im} is defined as

$$D_{im} = \max\left\{d\left(f_a, f_a^*\right)\right\}, \qquad (4)$$

where $f_a \in F_{I_1}$ and $f_a^* \in E_{I_2}$. D_{im} stands for the largest distance of element pair between F_{I_1} and E_{I_2}.Then the CPD between F_{I_1} and F_{I_2} is defined as

$$D_{cp}\left(F_{I_1}, F_{I_2}\right) = \min\left\{D_{cp}\left(F_{I_1}, E_{I_2}\right) + (M - N)D_{im}\right\}. \qquad (5)$$

With CPM, a query image is compared with each gallery image separately and a rank list is generated based on CPD. According to Equation (1), if two images are from the same group, there is always a permutation that exactly matches each person correctly. Even if the group member changes slightly (with at least 60% members unchanged), true matches of at least 60% group members still makes their CPD smaller than the CPD between negative sample groups.

CPD is a permutation invariant metric to measure similarity of sets, and permutation invariance is essential when modeling groups as sets or graphs. Next we will show the advantage of CPM over aggregation-based permutation invariant models.

Theorem 1. *A function $f\left(X\right)$ operating on a set X having elements from a countable universe, is a valid set function, i.e., invariant to the permutation of instances in X, iff it can be decomposed in the form $\rho\left(\sum_{x \in X} \phi\left(x\right)\right)$, for suitable transformations ϕ and ρ.*

According to Theorem 1 [34], the permutation invariant models have the form $\rho\left(\sum_{x \in X} \phi\left(x\right)\right)$, which can be used to build permutation invariant group representations. This form takes overall features into account but ignores the corresponding relationships between individuals. As illustrated in Figure 3a, $\phi(A) + \phi(B) = \phi(C) + \phi(D)$ for suitable transformation ϕ , so that feature set $\{A, B\}$ and $\{C, D\}$ are supposed to belong to the same group even though they are from different groups. However, they can be well distinguished using CPD.

Figure 3b indicates that the aggregation procedure for group representations may disturb feature space and make features of different groups less discriminative. If features of all identities are messed up, the extra transformation of single features by aggregation may amplify the ambiguity, while matching-based method CPM are less influenced. When the features of single person extracted by a Good feature extractor are well distinguishable, the CPM is more powerful than aggregation-based methods.

3.3. Relation-Aware Module

Absence of person labels hinders the training of deep nets, which makes the performance of feature extraction

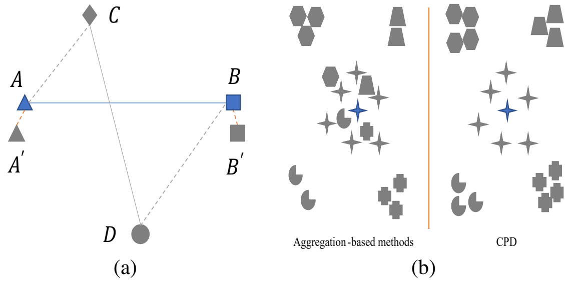


Figure 3. CPD can reduce false positive samples comparing to using aggregation-based methods.

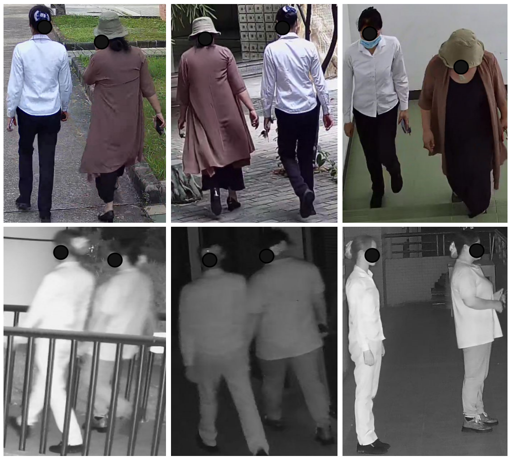


Figure 4. A demonstration of canonical ordering of a group. The woman in hat is stronger than the other, and the relation keeps the same cross viewpoints and modalities.

networks decrease a lot and the modality gap between RGB and infrared images not well-alleviated. It is straightforward to adopt weak-supervised methods in ReID such as [14, 19]. Here we focus on relations of group members rather than generating pseudo labels.

Following previous notations, we model an image I with a set of single-person features $F_I = \{f_1, f_2, ..., f_N\}$ without an ordering. The feature f_I of image I is the concatenation of elements in F_I with the ordering of input data. The problem of comparing two images is simplified to measure the similarity of two feature sets. We observe that in a group there are modality-invariant visual relations among group members to form its canonical ordering. For example, a group of three boys may vary in height, they can be ranked by height in descending order. The visual relations such as tall-short and fat-thin are robust cross modalities and camera-views. An illustration sample is displayed in Figure 4. The learned relations provide rules to assign an ordering to unordered feature set. Concatenation of single-person features in the modality-invariant order, *e.g.* from tall people to short people, forms a robust representation for groups. As shown in Figure 2b, the relation matrix R and relation score S are defined as

$$\begin{aligned} R &= (s_{ij}) \in \mathbb{R}^{N\times N}, \\ s_{ij} &= similarity\left(f_i, f_j\right), \\ S_i &= mean\left(R_{i,}\right), \end{aligned} \tag{6}$$

where $i, j \in N$, $R_{i,}$ is the ith row of R.

The similarity function measures the distance between features of any two members in a group. The relation score S_i of identity i is calculated by the average similarity with all other members, which indicates the overall relationship that can be regarded as the relative importance in the group. Then we get a rank list of all relation scores in the group and generate an ordering $\mathcal{O}_\mathcal{I}$ of the feature set F_I, which is supposed to be invariant to camera changes and modality changes. In the case of $N = 2$, we simply reverse the input order of f_I.

In training stage, we concatenate all single-person features in F_I with the learned ordering $\mathcal{O}_\mathcal{I}$ as

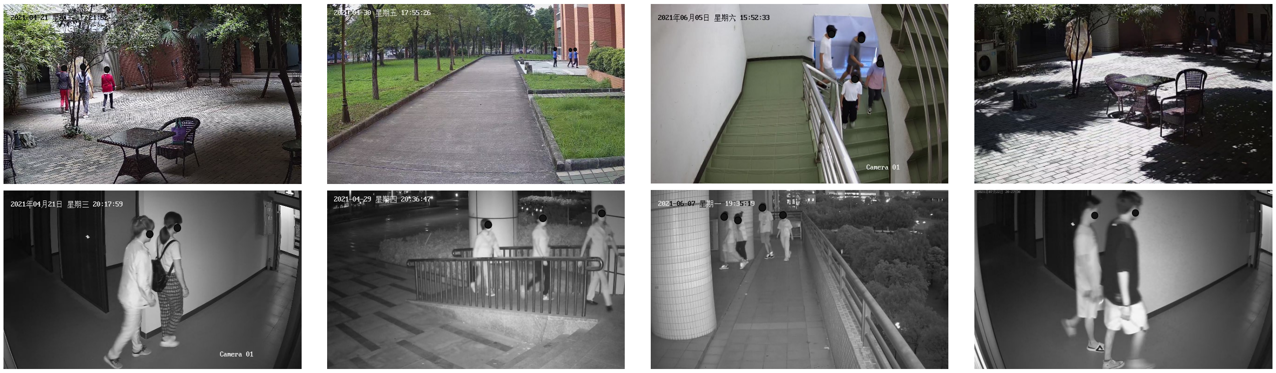
$$f_{\mathcal{O}_I} = concatenate\left\{f_1, f_2, ..., f_N\right\}_{\mathcal{O}_I}, \tag{7}$$

f_I with the input ordering and $f_{\mathcal{O}_I}$ are both reserved for training. In testing stage, the output feature is $f_{\mathcal{O}_I}$ which is irrelevant with the ordering of input data.

4. CM-Group Dataset

4.1. Dataset Collection

There are three main challenges to collect a large-scale RGB-IR G-ReID dataset. 1) To ensure large amount of groups, the amount of people have to be at least 2-3 times larger than groups. 2) The groups of people have to appear in camera views both in the daytime and in the evening. 3) It is very hard to annotate every person with automatic tools because the layout may change in different images. In this work, we present CM-Group, a large-scale real-world RGB-IR G-ReID dataset. We recruit 1,013 volunteers with every volunteer signs a consent letter for video recording and data collection for academic use. The volunteers are asked to walk in groups at 6 different scenes, three in the daytime and three at night. Each volunteer only belongs to one group. We use surveillance cameras to record videos and every group appears more than 30 seconds in each camera. Camera 1-3 are RGB cameras, which are placed on a road, a cortile and a stairwell, respectively. Camera 4-6 are infrared cameras, which are placed on a sidewalk, a corridor and a terrace, respectively. The raw videos are recorded from April to September, which covers different illumination conditions and weather. To obtain full annotations of all individuals and groups, we adopt YOLOv5 [17] to generate raw bounding boxes for every person, then we carefully annotate person labels and rectify wrong bounding boxes in each image by hand.



(a) Member changes

(b) Scale changes

(c) Clothes changes

(d) Extreme situation

Figure 5. Examples of RGB images and IR images in CM-Group. Figure 5a-Figure 5d present diverse challenging situations for RGB-IR G-ReID. Images are resized for better presentation.

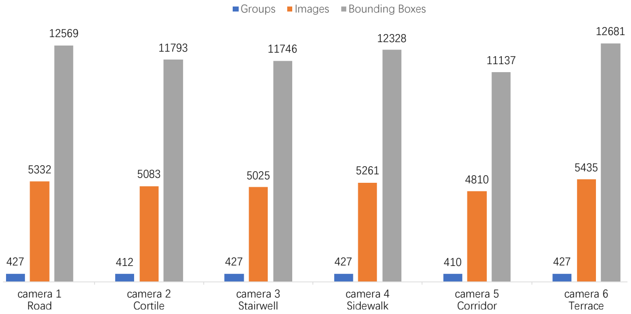


Figure 6. Statistics of groups, images and bounding boxes captured by each camera.

4.2. Dataset Description

The CM-Group dataset contains 30,946 images of 427 groups, with 1,013 annotated persons and 72,254 bounding boxes. The number of people in each group range from 2 to 5 and the average number of group members is 2.4. In each camera, we select 10-17 images for each group if available. The sample images and dataset statistics are shown in Figure 5 and Figure 6 respectively. Figure 5 displays several situations that are very common in real scenarios such as member changes, scale changes, clothes changes and extreme illumination conditions. It is clear that there are vast differences between RGB images and infrared images. To provide background information for future study, all images in CM-Group are original frames of raw videos.

The comparisons with three existing G-ReID datasets, *i.e*. Road Group, CSG, City1M, and one RGB-IR ReID dataset SYSU-MM01 are listed in Table 1. To our best knowledge, CM-Group is the first RGB-IR dataset for G-ReID and the largest real-world G-ReID dataset. In real-world G-ReID datasets, CM-Group includes more challenges, such as clothes changes and scale changes, to be addressed.

4.3. Evaluation Protocol

There are 427 groups in CM-Group. Following the setting in MSMT17 [23], we divide the dataset into training and testing set with a ratio of 1:3 to encourage more efficient training strategies. Accordingly, we have a fixed split of CM-Group with 107 groups for training and 320 groups for testing. The training set contains 17,282 bounding boxes of 233 identities, and the testing set contains 54,972 bounding boxes of 780 identities. In the testing set, RGB images from camera 1-3 form the gallery set, and infrared images from camera 4-6 form the query set.

To evaluate the performance of models, we use Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP). We randomly select one image for each group in each RGB camera to form the gallery and compute the average scores of 10 trials as final performance.

5. Experimental Results

5.1. Implementation Details

We conduct extensive evaluations on the CM-Group dataset. In our methods, we adopt ResNet-50 [7] pre-trained on ImageNet [3] as backbone for all experiments, where the stride size of the last convolution layer is set to 1. To train with full annotations including person labels, we assume a Good feature extraction network is provided. Thus we use the state-of-the-art RGB-IR ReID method MPANet [26] which is trained on CM-Group as feature extractor. The dataset for training is cropped to single-person images according to person labels in advance. In training stage, we use the hyperparameters and configurations as in MPANet. All images are resized to 384×128. We sample 128 images from the RGB modality and infrared modality to form a mini-batch. In each mini-batch, we randomly sample 16 identities and 8 images for each identity. The model is optimized by Adam method with a weight decay of 5×10^{-4}.

Dataset	R/S	Task	Modality	#Groups	#Identities	#Images	#Cameras
SYSU-MM01 [25]	Real	ReID	RGB-IR	-	491	45863	6
Road Group [28]	Real	G-ReID	RGB	162	1,099	324	2
CSG [30]	Real	G-ReID	RGB	1,558	3,500	3,989	Vary
City1M [36]	Synthetic	G-ReID	RGB	11,500	45,000	1,840,000	8
CM-Group (Ours)	Real	G-ReID	RGB-IR	427	1,013	30,946	6

Table 1. Comparisons with existing G-ReID and RGB-IR ReID datasets. “R/S” stands for real-world and synthetic dataset respectively.

Method	Label	Rank1	Rank10	Rank20	mAP
GIN [29]	S	0.50	4.67	9.92	1.37
DeepSets [34]	S	1.48	12.56	21.82	3.59
DuMLP [4]	S	1.55	12.46	20.62	3.47
Janossy [15]	S	1.60	12.52	20.70	3.50
ExpC-1 [31]	S	4.21	23.04	35.68	6.28
ST [10]	S	4.47	21.84	33.03	6.44
CPM	S	**57.68**	**86.07**	**92.51**	**55.23**
ExpC-1 [31]	W	5.28	26.77	39.85	7.50
MPANet [26]	W	15.32	55.60	68.78	19.31
Relabel	W	18.34	53.77	67.50	19.33
RAM	W	**27.92**	**66.35**	**78.66**	**27.14**

Table 2. Comparisons of CMC (%) and mAP (%) performances on CM-Group. “S/W” stands for supervised and weak-supervised respectively.

Method	Dataset	Rank1	mAP	Time
MACG [30]	CSG [30]	63.20	-	70h
CPM*	CSG [30]	88.57	51.45	14h
MACG [30]	CM-Group*	84.00	59.70	41h
CPM	CM-Group*	90.10	89.62	8h

Table 3. Comparisons with MACG in terms of CMC (%), mAP (%) performances and total running time on CSG and CM-Group*.

The initial learning rate is set to 3.5×10^{-4} and decays at the 80th and 120th epoch with a decay factor of 0.1. The total number of training epochs is set to 140. More details can be found in MPANet. All experiments are conducted on a single NVIDIA A100 GPU.

For the weak-supervised case without person labels, we embed RAM module into MPANet and adopt dot product to calculate person similarity. In training stage, we set the initial learning rate as 10^{-4} which decays at the 40th and 80th epoch with a decay factor of 0.1. The total number of training epochs is set to 100. The other settings are the same as full-supervised case.

5.2. Model Comparisons and Analysis

Since RGB-IR G-ReID has not been studied before, we turn to compare our CPM and RAM with permutation invariant methods on sets and graphs. The comparison results on CM-Group are listed in Table 2. For more details, please refer to the supplementary material.

Under Full-supervised Setting. We compare our CPM with permutation invariant approaches on CM-Group dataset. The compared methods include four set-based methods (DeepSets [34], Set Transformer [10], Janossy Pooling [15] and DuMLP-PIN [4]), and two GNN-based methods (GIN [29] and ExpC [31]). Note that we embed the compared methods into MPANet to aggregate single-person features for group representations. It is shown that the aggregation-based methods fail to acquire discriminative group representations and achieve poor Rank-1 accuracy and mAP. In contrast, our matching-based method CPM achieves the best Rank-1 accuracy of 57.68% and mAP of 55.23%.

Under Weak-supervised Setting. We compare RAM with the GNN-based method ExpC [31] and a base method MPANet [26] which regards a group of people as a whole. For more comparisons, we also design a pseudo-labelling method Relabel which assigns a pseudo label for each person based on intra-group relations in training stage. RAM achieves the Rank-1 accuracy of 27.92% and mAP of 27.14% which outperforms other methods in the weak-supervised setting.

Comparisons with MACG. We further compare CPM with the only open source G-ReID method MACG. The results are shown in Table 3. On CSG dataset, we remove the mutual learning module from MPANet to handle RGB images. The modified CPM outperforms MACG by a large margin. Specifically, the modified CPM achieves the Rank-1 accuracy of 88.57% and mAP of 51.45% on CSG dataset, significantly improving the Rank-1 accuracy by 25.37% over the MACG. The total running time in training and testing stage of the modified CPM is about 14 hours, which is five times faster than MACG. We emphasize that the modified MPANet only achieves the Rank-1 accuracy of 59.83% and mAP of 35.36% for ReID task on CSG dataset. If we use a better ReID network for person feature extraction, our CPM could gain better performance for G-ReID.

In terms of CM-Group dataset, it is estimated to take more than a month for MACG to finish the testing stage. Therefore we use an alternative to split the CM-Group with

Method	Rank1	Rank10	Rank20	mAP
Baseline	36.00	79.46	86.44	46.30
B+MPANet	35.95	77.58	87.37	46.02
B+CPD	43.78	77.41	86.24	41.46
B+MPANet+CPD	57.68	86.07	92.51	55.23

Table 4. Ablation study of CPM in terms of CMC (%) and mAP (%) performances on CM-Group.

Method	Rank1	Rank10	Rank20	mAP
Baseline	4.74	24.62	37.62	7.30
B+RAM	5.24	24.70	35.77	6.61
B+MPANet	11.97	47.18	62.62	15.45
B+MPANet+RAM	27.92	66.35	78.66	27.14

Table 5. Ablation study of RAM in terms of CMC (%) and mAP (%) performances on CM-Group.

357 groups for training and 50 groups for testing. Our CPM takes about 8 hours to achieve the Rank-1 accuracy of 90.10% and mAP of 89.62%, while MACG takes about 41 hours to achieve the Rank-1 accuracy of 84.00% and mAP of 59.70%.

5.3. Ablation Study

Modules in CPM. As shown in Table 4, we analyze the contribution of each module in CPM. All methods are trained on single-person images. The baseline method uses ResNet-50 as the backbone network followed by the BN neck and an FC layer as the classifier with Cross-Entropy loss. The group feature of an image is represented by the sum of all group member features. Compared with baseline, the MPANet does not solely improve the Rank-1 accuracy or mAP. The CPD improves the Rank-1 accuracy by 7.78% but reduces the mAP by 4.84%. When MPANet and CPD work together, the Rank-1 accuracy and mAP are significantly improved by 21.68% and 8.87% respectively. The results demonstrate that MPANet module effectively alleviates modality discrepancy and CPD module is useful for measuring similarities of person feature sets.

Effect of RAM. In Table 5, we evaluate the effectiveness of module RAM. The baseline method is all the same as described above but is trained on group images. Compared with the baseline, the RAM module improves the Rank-1 accuracy by 0.50% but reduces the mAP by 0.69%. The MPANet respectively improves the Rank-1 accuracy and mAP by 7.23% and 8.15%. MPANet and RAM together significantly improve the Rank-1 accuracy and mAP by 23.18% and 19.76% respectively. The results demonstrate that RAM module is able to extract modality-invariant ordering of person feature sets which helps to measure set similarities.

Method	Task	Rank1	Rank10	Rank20	mAP
ResNet [7]	R	17.73	46.04	56.56	16.80
ResNet+CPD	G	43.78	77.41	86.24	41.46
MPANet [26]	R	27.02	58.89	68.83	25.45
MPANet+CPD	G	57.68	86.07	92.51	55.23

Table 6. Improvements by CPD in terms of CMC (%) and mAP (%) performances on CM-Group. “R/G” stands for RGB-IR ReID and RGB-IR G-ReID respectively.

5.4. Discussions

Power of CPD. We evaluate the improvement made by CPD from two feature extraction networks on CM-Group. As shown in Table 6, both feature extraction networks can be integrated with CPD to achieve high performance in G-ReID task. The experiment results indicate that if we have trained a good feature extraction network, together with CPD our CPM framework will achieve strong performance for group retrieval.

Open Problems. Our research on CM-Group dataset leaves several open problems to the community. 1) Scale changes and clothes changes are common in CM-Group, which seriously disturb feature extraction. We seek for context information and group member relationships to build more robust group representations. However, the problem has not been addressed. 2) As we provide original images, how to use background information in G-ReID needs more attention. 3) CM-Group is flexible to be used in RGB-IR ReID task, effective methods on the new benchmark dataset requires further investigations.

6. Conclusion

In this paper, we introduce the RGB-IR cross-modality G-ReID problem. We model each group as a set of single person and use set similarity to represent group similarity. We analyze the limitation of permutation-invariant aggregation-based methods and propose the CPM framework. CPM uses CPD to measure group similarity and avoids aggregation functions adding noise to single-person features. Then we specially introduce the problem of G-ReID without person labels and propose the RAM module which extracts intrinsic relationships and orderings within groups. Finally, we contribute the first RGB-IR G-ReID dataset named CM-Group. The dataset is carefully designed and covers many challenging situations in real scenarios. The experimental results demonstrate the superiority of our proposed models CPM, RAM and CM-Group dataset.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (U22A2095, 62076258) and the Project of Ministry of Public Security of China (2019GABJC39).

References

[1] Yinghao Cai, Valtteri Takala, and Matti Pietikäinen. Matching groups of people by covariance descriptor. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, pages 2744–2747. IEEE Computer Society, 2010. 3

[2] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, page 6, 2018. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[4] Jiajun Fei, Ziyu Zhu, Wenlei Liu, Zhidong Deng, Mingyang Li, Huanjun Deng, and Shuo Zhang. Dumlp-pin: A dual-mlp-dot-product permutation-invariant network for set feature extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 598–606, 2022. 3, 7

[5] Hongyang Gao, Yi Liu, and Shuiwang Ji. Topology-aware graph pooling networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4512–4518, 2021. 3

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6, 8

[8] Ziling Huang, Zheng Wang, Wei Hu, Chia-Wen Lin, and Shin'ichi Satoh. Dot-gnn: Domain-transferred graph neural network for group re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 1888–1896, 2019. 2, 3

[9] Ziling Huang, Zheng Wang, Chung-Chi Tsai, Shin'ichi Satoh, and Chia-Wen Lin. Dotscn: Group re-identification via domain-transferred single and couple representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2739–2750, 2020. 3

[10] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753, 2019. 3, 7

[11] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4610–4617, 2020. 2

[12] Giuseppe Lisanti, Niki Martinel, Alberto Del Bimbo, and Gian Luca Foresti. Group re-identification via unsupervised transfer of sparse features encoding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2449–2458, 2017. 3

[13] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020. 2

[14] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. Weakly supervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2019. 5

[15] Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019. 3, 7

[16] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 3

[17] Ultralytics. Yolov5. https://github.com/ultralytics/yolov5, 2020. Accessed: 2021-09-06. 5

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[19] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2142–2156, 2020. 5

[20] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019. 2

[21] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12144–12151, 2020. 2

[22] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019. 2

[23] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 6

[24] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021. 2

[25] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 2, 7

[26] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 2, 6, 7, 8

[27] Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. *Advances in Neural Information Processing Systems*, 34:13266–13279, 2021. 3

[28] Hao Xiao, Weiyao Lin, Bin Sheng, Ke Lu, Junchi Yan, Jingdong Wang, Errui Ding, Yihao Zhang, and Hongkai Xiong. Group re-identification: Leveraging and integrating multi-grain information. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 192–200, 2018. 1, 3, 7

[29] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. 7

[30] Yichao Yan, Jie Qin, Bingbing Ni, Jiaxin Chen, Li Liu, Fan Zhu, Wei-Shi Zheng, Xiaokang Yang, and Ling Shao. Learning multi-attention context graph for group-based re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3, 7

[31] Mingqi Yang, Renjian Wang, Yanming Shen, Heng Qi, and Baocai Yin. Breaking the expression bottleneck of graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022. 7

[32] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020. 2

[33] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018. 2

[34] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 3, 4, 7

[35] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[36] Quan Zhang, Kaiheng Dang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Modeling 3d layout for group re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7512–7520, June 2022. 1, 3, 7

[37] Tong Zhang, Yun Wang, Zhen Cui, Chuanwei Zhou, Baoliang Cui, Haikuan Huang, and Jian Yang. Deep wasserstein graph discriminant learning for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10914–10922, 2021. 3

[38] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*, pages 1–11, 2009. 2

[39] Feng Zhu, Qi Chu, and Nenghai Yu. Consistent matching based on boosted salience channels for group re-identification. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 4279–4283. IEEE, 2016. 3

[40] Ji Zhu, Hua Yang, Weiyao Lin, Nian Liu, Jia Wang, and Wenjun Zhang. Group re-identification with group context graph neural networks. *IEEE Transactions on Multimedia*, 23:2614–2626, 2020. 2, 3