Received: 9 March 2022 | Revised: 13 July 2022 | Accepted: 2 August 2022

DOI: 10.1049/cvi2.12132

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Integration graph attention network and multi-centre constrained loss for cross-modality person re-identification

Di He[1] | Jingrui Zhang[1] | Zhong Zhang[1] | Shuang Liu[1] | Tariq S. Durrani[2]

[1]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China

[2]Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK

Correspondence

Shuang Liu, Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China.
Email: shuangliu.tjnu@gmail.com

Funding information

National Natural Science Foundation of China, Grant/Award Number: 2022KYCX032Z; Natural Science Foundation of Tianjin City, Grant/Award Numbers: 19JCZDJC31500, 20JCZDJC00180; Key Project of Graduate Scientific Research Innovation Project of Tianjin Normal University, Grant/Award Number: 2022KYCX032Z

Abstract

Cross-modality person re-identification is a challenging task due to the large visual appearance difference between RGB and infrared images. Existing studies mainly focus on learning local features and ignore the correlation between local features. In this paper, the Integration Graph Attention Network is proposed to learn the completed correlation between local features via the graph structure. To this end, the authors learn the coarse-fine attention weights to aggregate the local features by considering local detail and global information. Furthermore, the Multi-Centre Constrained Loss is proposed to optimise the feature similarity by constraining the centres of modality and identity. It simultaneously utilises three kinds of centre constraints, that is intra-identity centre constraint, modality centre constraint, and inter-identity centre constraint, in order to reduce the influence of modality information explicitly. The proposed method is evaluated on two standard benchmark datasets, that is SYSU-MM01 and RegDB, and the results demonstrate that the authors' method achieves better performance than the state-of-the-art methods, for example, surpassing NFS by 4.8% and 6.0% mAP on the single-shot setting in All-search and Indoor-search modes, respectively.

1 | INTRODUCTION

Person re-identification (Re-ID) is an important task, which aims to match the same pedestrian across different cameras [1–3]. This task can be applied to many practical application fields such as video surveillance, intelligent traffic supervision etc. With the development of deep learning, person Re-ID methods have made a great progress in recent years, and most of them are designed for processing RGB images captured by visible cameras [4–7]. However, visible cameras are difficult to capture discriminative appearance information under poor illumination condition. Hence, single modality person Re-ID methods cannot work well in the night scenario.

Compared with RGB images, infrared (IR) images could provide more appearance information under poor illumination condition, and therefore cross-modality person Re-ID is naturally proposed to apply RGB and IR images, simultaneously. The visual appearance difference between RGB and IR images is the main challenge for cross-modality person Re-ID, because IR images with one channel only contain the information of invisible electromagnetic radiation while three-channel RGB images include rich colour information of visible light. Furthermore, cross-modality person Re-ID inherits the challenges of single modality person Re-ID, such as the variations in poses and viewpoints. In a word, cross-modality person Re-ID is more challenging than single modality person Re-ID.

To address the above-mentioned issues, the existing cross-modality person Re-ID methods mainly focus on feature learning and metric learning. As for the feature learning, some methods design one-stream or two-stream networks to extract global features from RGB and IR images [8, 9, 11, 42]. Furthermore, modality-consistent features or images are usually learnt to reduce the modality gap, which is generated by various modality transformations, such as GAN, convolution operation, grayscale transformation and so on [13-17, 44]. Meanwhile, local features of pedestrian are utilised to explore the invariant body shape information for cross-modality person Re-ID [10, 18, 45]. However, these methods only extract local features from a single region and ignore the correlation between other features, which is difficult to learn complementary information between local features.

As for the metric learning, it is applied to reduce the appearance difference between RGB and IR images from the aspect of feature similarity optimisation. The cross-modality triplet loss, the hetero-centre loss and the contrastive loss are proposed to control the distance between cross-modality features [9, 10, 42]. Some methods map heterogeneous features into a common space so as to learn modality-shared metrics [19, 20]. However, these methods mix the modality information and the identity information of features in the process of metric learning, and they do not explicitly consider the influence of the modality information. Hence, the learnt metric functions are suboptimal for cross-modality person Re-ID.

To overcome the above-mentioned limitations, we propose a novel method named Integration Graph Attention Network (IGAT) for cross-modality person Re-ID, where IGAT is designed to learn the correlation between local features via the graph structure. To this end, we first extract the local features of pedestrian images and treat them as the nodes of graph. In order to model the completed correlation, we not only learn the correlation between local features, but also integrate the global correlation into the feature representation via learning the coarse-fine attention weights. Then, we apply the coarse-fine attention weights to aggregate the local features from the corresponding parts with the same modality. As a result, local detail and global information are injected into the final representation so as to obtain the complementary information.

Furthermore, to relieve the influence of modality information, we propose the Multi-Centre Constrained Loss (MCCL) to optimise the similarity between pedestrian images by constraining the centres of modality and identity. Specifically, as shown in Figure 1, MCCL consists of three components: 1) Intra-identity centre constraint: to increase the feature similarity between pedestrian images with the same identity, we directly pull the centres with the same identity from different modalities together. 2) Modality centre constraint: we also pull the centres of different modalities together to reduce the feature discrepancy caused by cross modality. 3) Inter-identity centre constraint: the centres of different identities are encouraged to be away from each other. It could improve the feature dissimilarity between pedestrian images from different identities in order to obtain discriminative features. In a word, the proposed MCCL explicitly reduces the influence of modality information by constraining different kinds of centres.

The main contributions of this work are summarised as follows:

1) We propose IGAT to obtain the completed correlation between local features by learning the coarse-fine attention weights.

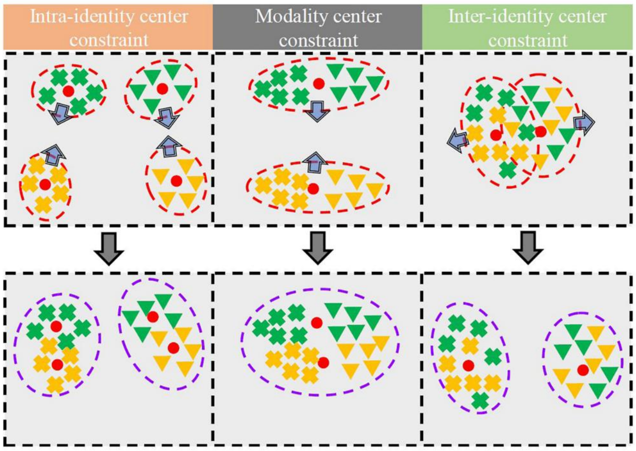
Intra-identity center constraint
Modality center constraint
Inter-identity center constraint

FIGURE 1 The illustration of Multi-Centre Constrained Loss. The top row and the bottom row are the schematic diagram before and after using the constraints, respectively. The points with the same shape denote the features belonging to the same identity, and the points with the same colour indicate the features belonging to the same modality

2) We propose MCCL to optimise the similarity between pedestrian images from different aspects by constraining different kinds of centres.
3) Extensive experimental results on the SYSU-MM01 and RegDB datasets show our method surpasses the state-of-the-art methods, which demonstrate the effectiveness of our method.

2 | RELATED WORK

2.1 | Cross-modality person Re-ID

In order to overcome the visual appearance difference between RGB and IR modalities, many approaches have been proposed to learn discriminative features for cross-modality person Re-ID [8, 21, 22]. Some of them design the specific network structures to obtain global features [8, 11]. For example, Ye et al. [11] present a two-stream network with non-local attention to extract global features. Some methods employ the generator module to generate modality alignment information [12, 16, 26]. Wang et al. [26] apply AlignGAN to transform real RGB images to fake IR images in order to obtain alignment features.

Furthermore, local features are introduced into cross-modality person Re-ID to extract the invariant body shape information from the pedestrian images of different modalities [18, 45]. Sun et al. [23] propose a whole-individual training (WIT) model to learn local features for VI-ReID, which is based on the idea of pulling the whole images and distinguishing the individuals. Ye et al. [18] exploit the intra-modality part relationship to enhance the feature representation. However, these local feature-based methods for cross-modality person Re-ID only learn the local information or their correlation, which results in learning incomplete correlation in the aggregation process. Different from the above methods, the proposed IGAT learns the completed correlation between local features, where the local and global features are both considered in the aggregation process.

In order to learn the accurate similarity measurement between cross-modality features, some methods reduce the modality gap by means of metric learning. Chen et al. [42] employ the contrastive loss, and Ye et al. [9] adopt the cross-modality triplet loss to optimise the deep networks, which is beneficial to extract modality-invariant features. Hao et al. [20] map the pedestrian images from two domains into a hypersphere and constrain the cross-modality variations by the hypersphere. Zhu et al. [10] propose the hetero-centre loss to reduce the intra-identity cross-modality variations by constraining the centres. Although these centre-based losses achieve promising results, they only consider one kind of centre constraint, which is difficult to handle the complex distributions of heterogeneous features. Different from them, the proposed MCCL considers three kinds of centre constraints of modality and identity so as to simultaneously reduce the intra-identity cross-modality variations and inter-modality variations, and increase the inter-identity variations.

2.2 | Graph Attention Network

Graph Attention Network (GAT) [29] remits the prior knowledge of graph structure from Graph Neural Network (GNN) [27, 28] and integrates the attention architecture into GNN. It assigns different weights to neighbour nodes for propagating information to centre nodes via masked self-attentional layers.

Recently, GAT has been applied into various tasks to exploit the dependency between nodes [30, 31]. Huang et al. [32] propose the target-dependent GAT to utilise dependency relationship among words for aspect level sentiment classification. Wang et al. [33] propose the relational GAT to encode syntax information for sentiment prediction. Yang et al. [25] propose HGAT based on a dual-level attention mechanism for short text classification. Chen et al. [24] exploit heterogeneous graph and node features to learn user profiles from limited labelled data.

As for the field of person Re-ID, Zhang et al. [37] present Heterogeneous Local Graph Attention Networks (HLGAT) to model the inter-local relation and the intra-local relation for person Re-ID. However, HLGAT ignores the aggregation of global information in the learning process of the local features. Different from HLGAT, the proposed IGAT models the dependency between local features from the local and global aspects.

3 | APPROACH

The framework of the proposed method is shown in Figure 2. It mainly consists of the Feature Extractor Module, the IGAT Module and MCCL. We detail each component in this section.

3.1 | Feature Extractor Module

The Feature Extractor Module is designed based on a two-stream network, which adopts ResNet-50 [34] as the backbone. Specifically, we adopt two individual ResNet-50, which are removed the last down-sampling operation for two modality streams. In each modality stream, the feature maps outputted from the last convolution layer are conducted by the average pooling. Meanwhile, the feature maps are divided into P uniform parts horizontally and then implemented by the average pooling for each part. Finally, a weight-shared fully connected (FC) layer is employed by the two modality streams to obtain the local features $f_L^p|_{p=1}^P$ and the global feature f_G.

3.2 | IGAT module

The local features have been demonstrated the effectiveness to viewpoint and posture changes [35–37]. Furthermore, the local features corresponding to the same part describe the pedestrian from different aspects, and therefore learning the correlation between local features could propagate useful information between them. As a result, the discrimination of local features is improved. Motivated by this, we design IGAT to model the completed correlation for local features.

The IGAT Module is connected after the Feature Extractor Module, and the local features $f_L^p|_{p=1}^P$ and the global feature f_G are the input of the IGAT Module. We utilise the local features to construct a graph where each local feature is treated as a node. Each node is updated by its neighbour nodes with the aggregation operation. The node after updating is formulated as:

$$\tilde{f}_L^{p,i} = \sigma\left(f_L^{p,i} + \sum_{j\in N_i} \alpha_p^{ij} V_p f_L^{p,j}\right) \tag{1}$$

where $f_L^{p,i}$ and $f_L^{p,j}$ are the p-th local features of the i-th and j-th pedestrian images respectively, V_p indicates the learnable transformation matrix for the p-th local feature, $\sigma(\cdot)$ denotes the non-linear transformation implemented by the LeakyReLU operation, and N_i is the neighbour node set of i, which contains the nodes with the same modality of the i-th pedestrian image. Here, α_p^{ij} is the attention weight, which reflects the correlation between the p-th local features of the i-th and j-th pedestrian images. It is usually formulated as [18]:

$$\alpha_p^{ij} = \frac{\exp\left(\sigma\left(\phi\left(W_p f_L^{p,i}, W_p f_L^{p,j}\right)\right)\right)}{\sum_{k\in N_i} \exp\left(\sigma\left(\phi\left(W_p f_L^{p,i}, W_p f_L^{p,k}\right)\right)\right)} \tag{2}$$

where ϕ denotes the cosine similarity function, and W_p represents the learnable transformation matrix for the p-th local feature.

From Equations (1) and (2), we can see that it only considers the correlation from the local aspect, and it may

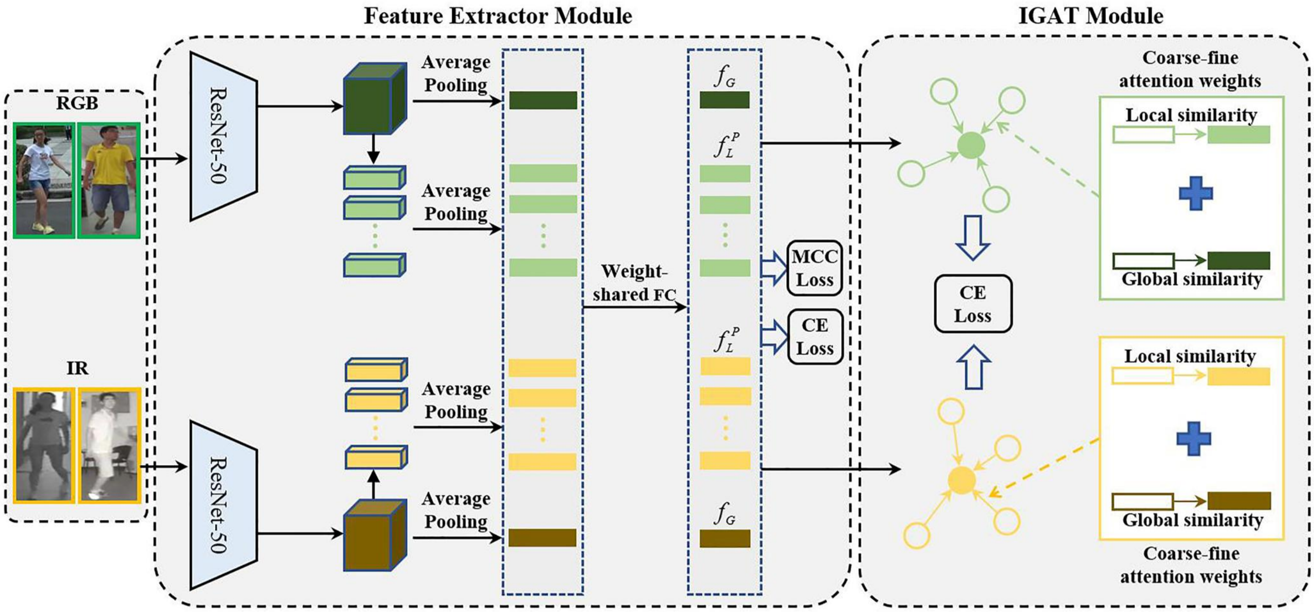

FIGURE 2 The framework of the proposed method. It contains two modules, that is, Feature Extractor Module and IGAT Module. The Feature Extractor Module treats ResNet-50 as the backbone and extracts the local features f_L^p and the global feature f_G from RGB and infrared images. The IGAT Module learns the completed correlation based on the local similarity and the global similarity using the coarse-fine attention weights. Furthermore, we propose Multi-Centre Constrained Loss to optimise the network via constraining three kinds of centres of modality and identity. IGAT, Integration Graph Attention Network

obtain some unexpected attention weights without considering global information. Figure 3 shows some local regions of RGB and IR images. From Figure 3a we can see that I_1 and I_3 are with the same identity, and I_1 and I_2 possess different identities. But the similarity between the local regions of I_1 and I_2 is higher than that of I_1 and I_3. Meanwhile, from the aspect of whole images, we can distinguish that I_1 and I_3 have the same identity, which is opposite to the judgement from the local aspect. We can draw the similar conclusion from IR images in Figure 3b.

Based on the observation of Figure 3, we inject the global information when learning the correlation between local features. We expect to utilise the global feature similarity to correct the mismatching caused by only considering the local similarity. Hence, we propose the coarse-fine attention weights to consider the similarity between local features and the similarity between global features. The coarse-fine attention weight between the p-th local features of the i-th and j-th pedestrian images is defined as:

$$\alpha_p^{ij} = \frac{\tilde{\alpha}_p^{ij}}{\sum_{k\in N_i}\tilde{\alpha}_p^{ik}} \tag{3}$$

$$\tilde{\alpha}_p^{ij} = \exp\left(\sigma\left(\phi\left(W_p f_L^{p,i}, W_p f_L^{p,j}\right) + \lambda\phi\left(W_G f_G^i, W_G f_G^j\right)\right)\right) \tag{4}$$

where W_G is the learnable transformation matrix for the global feature, λ is the balance parameter, and f_G^i and f_G^j indicate the global features of the i-th and j-th pedestrian images, respectively. Afterwards, we substitute Equation (3) into Equation (1) to obtain the aggregated local features.

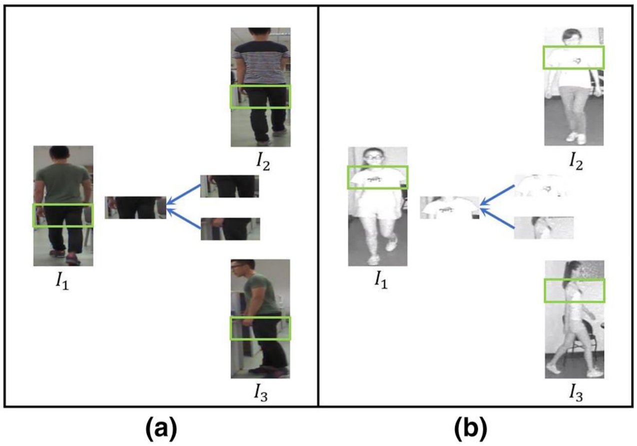

FIGURE 3 The similarity between some local regions

We employ the cross-entropy (CE) loss to supervise the learning process of IGAT:

$$L_{id}^{IGAT} = \sum_{p=1}^{P} L_{id}^p \tag{5}$$

where L_{id}^p is the CE loss for the p-th aggregated local feature.

3.3 | Multi-centre constrained loss

In the field of cross-modality person Re-ID, it is common that the similarity between RGB and IR images with the same identity is not high enough to distinguish because of the influence of heterogeneous modalities. The metric learning is effective to reduce the modality gap; however, the existing metric learning methods for cross-modality Re-ID do not explicitly handle the influence of modality information.

Hence, we propose MCCL to simultaneously consider the influence of modality information and identity information by constraining multiple centres of modality and identity. Specifically, MCCL includes three kinds of centre constraints in order to achieve comprehensive similarity optimisation. First, we apply the intra-identity centre constraint to pull the centres with the same identity from different modalities together in order to increase the similarity of cross modality features with the same identity [10]. It is defined as:

$$L^{p}_{tra} = \sum_{i=1}^{N} \| c_{R}^{p,i} - c_{I}^{p,i} \|_2^2 \tag{6}$$

where N denotes the number of identities, $\|\cdot\|_2$ denotes the Euclidean distance, and $c_R^{p,i}$ and $c_I^{p,i}$ are the centres (mean vectors) of the p-th local features for the i-th identity of RGB images and IR images, respectively.

In order to further reduce the modality gap, we propose the modality centre constraint from the macro perspective. The modality centre constraint is expected to pull the centres of two modalities together, which is convenient to transform the heterogeneous features into the homogeneous features. It is defined as:

$$L^{p}_{m} = \| c_{R}^{p} - c_{I}^{p} \|_2^2 \tag{7}$$

where c_R^p and c_I^p are the centres of the p-th local features of all RGB and IR images, respectively. Different from computing multiple centres for the p-th local features in the intra-identity centre constraint, the modality centre constraint only requires to compute one centre for each modality.

Finally, we propose the inter-identity centre constraint to push the centres of different identities away so as to increase the differentiation of features. The intra-identity centre constraint and the modality centre constraint mainly focus on improving the similarity between the pedestrian images of cross modality. As a complement, the inter-identity centre constraint is designed to increase the dissimilarity between the pedestrian images with different identities. We define two kinds of forms for the inter-identity centre constraint, and as shown in Figure 4a the first one is:

$$\tilde{L}^{p}_{ter} = \sum_{i,j} \max\left(r^{p,i} + r^{p,j} - d_p^{i,j}, 0 \right) \tag{8}$$

where $r^{p,i}$ is the maximum of all distances between the centre and the features for the p-th local features of i-th identity, so is $r^{p,j}$ for the j-th identity. Here, the margin $d_p^{i,j}$ is the Euclidean distance between the centres of the i-th identity and the j-th identity for the p-th local features.

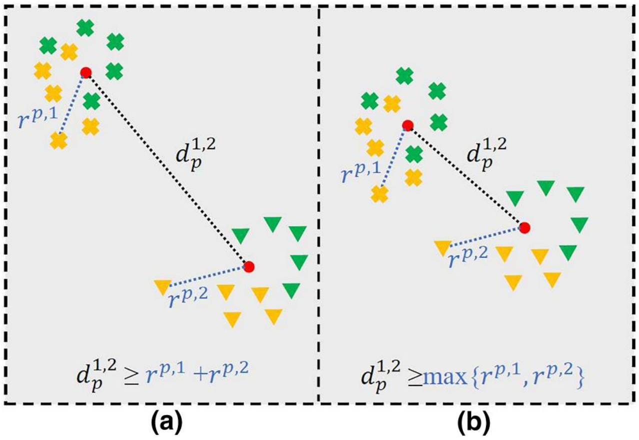
$r^{p,1}$
$d_p^{1,2}$
$r^{p,2}$
$d_p^{1,2} \geq r^{p,1} + r^{p,2}$
(a)
$r^{p,1}$
$d_p^{1,2}$
$r^{p,2}$
$d_p^{1,2} \geq \max\{r^{p,1}, r^{p,2}\}$
(b)

FIGURE 4 The schematic diagram of the inter-identity centre constraint. The points with the same shape denote the features belonging to the same identity, and the yellow points and the green points indicate the features belonging to RGB and infrared images, respectively. The red circles represent the centres

Furthermore, we narrow the margin in Equation (8) as shown in Figure 4b and then obtain another form of inter-identity centre constraint:

$$L^{p}_{ter} = \sum_{i,j} \max\left(\max\left(r^{p,i}, r^{p,j}\right) - d_p^{i,j}, 0 \right) \tag{9}$$

Equation (9) relaxes the margin restriction and it does not introduce any extra parameters.

Figure 5 shows the loss trend of $\tilde{L}_{ter}$ and L_{ter} in the training process, where we can see that L_{ter} has faster convergence speed than $\tilde{L}_{ter}$. Meanwhile, in the ablation study, we conduct experiments to validate that L_{ter} is more effective than $\tilde{L}_{ter}$. In a word, the proposed MCCL for local features is defined as:

$$L^{p}_{MCC\_L} = \beta_1 L^{p}_{tra} + \beta_2 L^{p}_{m} + \beta_3 L^{p}_{ter} \tag{10}$$

where β_1, β_2 and β_3 are the weight parameters. We not only adopt MCCL on the local features using Equation (10) but also on the global features denoted as L_{MCC_G}. Hence, MCCL on the local and global features is formulated as:

$$L_{MCC} = \sum_{p=1}^{P} L^{p}_{MCC\_L} + L_{MCC\_G} \tag{11}$$

3.4 | Optimisation

To optimise the proposed framework in an end-to-end way, the overall loss is defined as:

$$Loss = \mu_1 L_{id}^{IGAT} + \mu_2 L_{id}^{T} + \mu_3 L_{MCC} \tag{12}$$

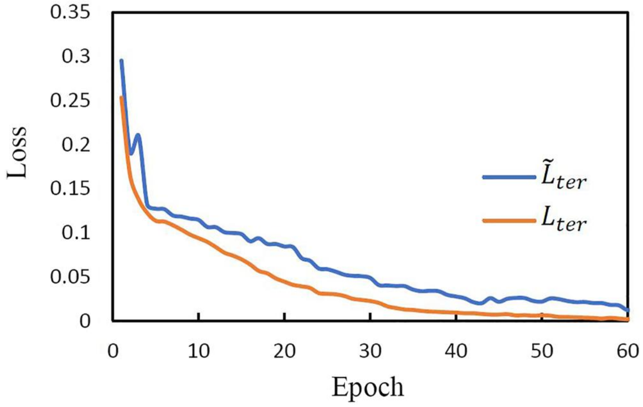

FIGURE 5 The loss trend of L_{ter} and $\tilde{L}_{ter}$ in the training process

where μ_1, μ_2 and μ_3 are the parameters to control the weights of different components. Here, L_{id}^{T} denotes the sum of CE losses for the local features and the global features in the Feature Extractor Module. Finally, the result obtained by calculating Equation (12) is back-propagated to the model so as to optimise the model.

4 | EXPERIMENTS

4.1 | Datasets

SYSU-MM01 [8] is a large-scale cross-modality person Re-ID dataset, which contains 301/3010 (*single-shot*/*multi-shot*) RGB images and 3803 IR images of 96 identities in the test set, and 22,258 RGB images and 11,909 IR images of 391 identities in the training set.

RegDB cross-modality person Re-ID dataset [38] includes 8240 images of 412 identities. Each identity has 10 RGB images and 10 IR images. Two hundred and six identities are randomly selected from all 412 identities to construct the training set, and the remaining identities constitute the test set. RegDB provides two types of evaluation modes according to different modality match settings. One is *Visible to Thermal* (*V-T*), which searches RGB images of the same identity from IR images, and the other one is *Thermal to Visible* (*T-V*), which queries IR images of the same identity from RGB images.

4.2 | Implementation details

All the pedestrian images are resized to 288 × 144 and augmented by the random horizontal flipping and the random cropping. The batch size is set to 64, which contains four identities and each identity carries eight RGB images and eight IR images. The weight-shared FC layer in the Feature Extractor Module reduces the dimension of both the local features and the global features from 2048 to 512. The number of the local features P is set to 6. Besides, we set the balance parameter λ in Equation (4) to 0.2. The weights of MCCL β_1, β_2 and β_3 in Equation (10) are set to 1, 0.5, and 0.5, respectively. The weights of different losses μ_1, μ_2 and μ_3 in Equation (12) are set to 0.1, 1, and 0.5, respectively. To enhance the stability of graph learning, we adopt the multi-head attention strategy [29] in IGAT, and the number of multi-head is set to 4.

The proposed network is optimised by the stochastic gradient descent (SGD) scheme [46]. The number of epochs is set to 60 in the training process. The initial learning rate is set to 0.01 and lasted for 30 epochs. Afterwards, the learning rate is changed to 0.001 for the remaining epochs.

4.3 | Ablation study

In this subsection, we design the ablation study to validate the effectiveness of each component of our method. We choose the most challenging *single-shot* setting on SYSU-MM01 and the *V-T* mode on RegDB to evaluate the performance. The results of ablation study are shown in Table 1. *BS* refers to the baseline, which adopts the Feature Extractor Module supervised by the CE losses. *BS* + GAT indicates that modelling the local correlation without considering the similarity between the global features, and its attention weights are computed by Equation (2).

For SYSU-MM01, it is obvious that the performance of our method (Ours) achieves the best results and the following conclusions can be drawn.

Effectiveness of IGAT. The performance of *BS* + GAT surpasses *BS* by 2.2% rank-1 accuracy and 2.4% mAP, which illustrates the importance of learning the correlation between local features. The performance of *BS* + IGAT further brings 2.3% and 1.7% increments on rank-1 accuracy and mAP compared with *BS* + GAT. It is because the proposed IGAT models the dependency between local features from the local and global aspects. Specifically, the IGAT module not only considers the correlation between local features but also injects global information when learning the correlation so as to obtain more precise attention weights, namely, the coarse-fine attention weights. Meanwhile, it further proves the effectiveness of adding global information to attention weights of the local features.

Effectiveness of MCCL. The performance of $BS + L_{tra}$, $BS + L_m$, $BS + L_{ter}$ and $BS{+}\tilde{L}_{ter}$ all achieves better than that of *BS* due to adding different centre constraints. Afterwards, the performance further gains by using two or three different kinds of centre constraints. Hence, each component in MCCL prompts the network to obtain higher performance, which demonstrates the effectiveness of MCCL.

Effectiveness of the margin for the inter-identity centre constraint. As shown in Figure 5, we can see that the loss curve of Equation (9) is smoother and faster than Equation (8) because Equation (9) relaxes the margin restriction, which makes convergence in the training process more stable. Furthermore, in Table 1, $BS + L_{ter}$ improves Rank-1 and mAP compared with $BS + \tilde{L}_{ter}$ and so does *BS* + MCCL compared with $BS + L_{tra} + L_m{+}\tilde{L}_{ter}$. It can be concluded that narrowing the margin of inter-identity, that is Equation (9) is more effective.

TABLE 1 Ablation study on SYSU-MM01 and RegDB

	SYSU-MM01		RegDB	
Methods	R1	mAP	R1	mAP
BS	46.8	46.1	60.5	58.1
BS + GAT	49.0	48.5	63.5	59.7
BS + IGAT	51.3	50.2	66.7	61.8
BS + L_{tra}	56.3	54.8	80.1	72.2
BS + L_m	52.3	51.2	65.7	64.6
BS + L_{ter}	51.2	51.8	63.0	62.9
BS + $\tilde{L}_{ter}$	49.8	48.5	61.8	59.2
BS + L_{tra} + L_m	58.2	57.9	81.5	73.5
BS + L_{tra} + L_m + $\tilde{L}_{ter}$	58.4	57.7	81.8	73.7
BS + MCCL	59.5	59.4	83.0	75.1
Ours	60.6	60.3	84.1	75.4

Note: R1 denotes Rank-1 accuracy (%).

Through the above analysis, we can further prove the effectiveness of our proposed method, namely, the proposed IGAT learns completed correlation between local features by considering both local detail and global information and the proposed MCCL constrains the centres of modality and identity to optimise the similarity of features so as to explicitly overcome the influence of modality information. Note that as for the RegDB dataset, we can obtain the similar conclusions mentioned above.

4.4 | Comparison with the state-of-the-art methods

In this subsection, we compare our method with the state-of-the-art methods on two standard benchmark datasets, that is, SYSU-MM01 and RegDB, to verify the effectiveness of our method. The compared results are listed in Table 2 and Table 3, and these compared methods are mainly classified into three categories: 1) Zero-Padding [8] uses the one-stream structure for feature extraction, 2) some methods (TONE [19], BDTR [9], D-HSME [20], MAC [39], MSR [40], JSIA-ReID [14], DAPR [41], AGW [11], CMAlign [49] and NFS [42]) adopt the two-stream structure for feature extraction, and 3) some methods (cmGAN [12], D^{2}RL [13], AlignGAN [26], Xmodal [16] and Hi-CMD [15]) employ the specific way for image generation. Furthermore, some methods adopt local features for cross-modality person ReID, that is, DDAG [18], TSLFN + HC [10] and FBP-AL [50].

Results on SYSU-MM01. As shown in Table 2, in SYSU-MM01, our method achieves 60.6% rank-1 accuracy and 60.3% mAP under the *single-shot* setting, and 65.5% rank-1 accuracy and 53.8% mAP under the *multi-shot* setting, which outperforms the compared state-of-the-art methods. For the three methods (DDAG [18], TSLFN + HC [10] and FBP-AL [50]) using local features, DDAG mines intra-modality part-level context cues using local features and FBP-AL learns more fine-grained information by part representations, while our method learns the correlation between local features from the local and global aspects. TSLFN + HC adopts the intra-identity centre constraint to reduce the modality gap, while our method simultaneously utilises three different kinds of centre constraints. Hence, the performance of our method is better than other local feature learning methods.

Results on RegDB. From Table 3, it can be seen that our method achieves 84.1% rank-1 accuracy and 75.4% mAP in the *V-T* mode and 83.1% rank-1 accuracy and 76.0% mAP in the *T-V* mode, which exceeds the second best method by a large margin. It demonstrates that our method possesses high generalisation ability to different cross-modality person Re-ID datasets.

Recently, MPANet [48] modifies the network backbone to build a new baseline where they propose to embed the attention mechanisms into the ResNet50 network and utilises mutual learning to enable different modalities to interact with each other. It achieves the state-of-the-art performance for cross-modality person Re-ID. While our method does not change the backbone and only utilises the original ResNet50 network. Furthermore, it does not focus on the interaction of two modality streams. GLMC [47] applies the cross-entropy loss and the triplet loss to supervise the whole global branch, where they treat cross-modality person Re-ID as a classification task and a rank task, and focusses on learning the global and local features of pedestrians. Compared to Ref. [47], we only treat cross-modality person Re-ID as a classification task.

Since building the correlations among pedestrian features is beneficial for learning completed information, we focus on the learning of the correlations of pedestrian features. To this end, we propose IGAT to consider the correlation between local features via the graph structure. The IGAT module injects global information when learning the correlation between local features so as to obtain more precise attention weights, namely, the coarse-fine attention weights. Moreover, we propose MCCL to optimise the similarity between pedestrian images from different aspects by constraining different kinds of centres, so as to reduce the discrepancies among different modalities and make the features with the same identity compact and the features with the different identity far away.

4.5 | Parameter analysis

In this subsection, we conduct a series of experiments to study the influence of several key parameters for the proposed method including the balance parameter of local detail and global information λ in Equation (4), the weights of three components in MCCL β_1, β_2 and β_3 in Equation (10), and the weights of different losses μ_1, μ_2 and μ_3 in Equation (12). The experiments are conducted under the *single-shot* setting on SYSU-MM01. The experimental results can be generalised to *multi-shot* settings of SYSU-MM01 and RegDB.

TABLE 2 Comparison on SYSU-MM01

		All-search								*Indoor-search*							
		Single-shot				*Multi-shot*				*Single-shot*				*Multi-shot*			
Methods	**Venue**	**R1**	**R10**	**R20**	**mAP**	**R1**	**R10**	**R20**	**mAP**	**R1**	**R10**	**R20**	**mAP**	**R1**	**R10**	**R20**	**mAP**
Zero-padding [8]	ICCV'17	14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
TONE [19]	AAAI'18	14.3	53.2	69.2	16.2	–	–	–	–	24.52	73.25	86.73	30.08	–	–	–	–
BDTR [9]	IJCAI'18	17.0	55.4	72.0	19.7	–	–	–	–	–	–	–	–	–	–	–	–
cmGAN [12]	IJCAI'18	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
D-HSME [20]	AAAI'19	20.7	62.8	78.0	23.2	–	–	–	–	–	–	–	–	–	–	–	–
D^{2}RL [13]	ICCV'19	28.9	70.6	82.4	29.2	–	–	–	–	–	–	–	–	–	–	–	–
MAC [39]	MM'19	33.3	79.0	90.9	36.2	–	–	–	–	36.4	62.3	71.6	37.0	–	–	–	–
AlignGAN [26]	ICCV'19	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
Hi-CMD [15]	CVPR'20	34.9	77.6	–	35.9	–	–	–	–	–	–	–	–	–	–	–	–
MSR [40]	TIP'20	37.4	83.4	93.3	38.1	43.9	86.9	95.7	30.5	39.6	89.3	97.7	50.9	46.6	93.6	98.8	40.1
JSIA-ReID [14]	AAAI'20	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
Xmodal [16]	AAAI'20	49.9	89.8	96.0	50.7	–	–	–	–	–	–	–	–	–	–	–	–
DAPR [41]	IVC'21	46.0	87.9	96.0	43.9	47.7	89.9	96.6	34.5	46.2	89.2	96.7	55.8	51.4	92.9	98.5	44.1
AGW [11]	TPAMI'21	47.5	84.4	92.1	47.7	54.6	90.2	96.2	41.8	54.2	91.1	96.0	63.0	61.2	92.9	97.6	53.8
CMAlign [49]	ICCV'21	55.4	–	–	54.1	–	–	–	–	58.5	–	–	66.3	–	–	–	–
NFS [42]	CVPR'21	56.9	91.3	96.5	55.5	63.5	94.2	97.8	48.6	62.8	96.5	99.1	69.8	70.3	97.7	99.5	61.5
GLMC* [47]	TNNLS'21	64.4	93.9	97.5	63.4	66.7	95.9	98.6	54.4	67.4	98.1	99.8	74.0	77.5	97.7	99.6	67.2
MPANet‡ [48]	CVPR'21	70.6	96.2	98.8	68.2	75.6	97.9	99.4	62.9	76.7	98.2	99.6	81.0	84.2	99.7	99.9	75.1
DDAG [18]	ECCV'20	54.8	90.4	95.8	53.0	–	–	–	–	61.0	94.1	98.4	68.0	–	–	–	–
TSLFN + HC [10]	Neurocomputing'20	57.0	91.5	96.8	55.0	62.1	93.7	97.9	48.0	59.7	92.1	96.2	65.0	69.8	95.9	98.0	57.9
FBP-AL [50]	TNNLS'21	54.1	86.0	93.0	50.2	–	–	–	–	–	–	–	–	–	–	–	–
Ours	–	60.6	94.4	98.5	60.3	65.5	95.8	98.9	53.8	66.7	99.0	99.8	75.8	74.6	99.7	99.8	69.3

Note: R1, R10 and R20 denote Rank-1, Rank-10 and Rank-20 accuracies (%), respectively. Here, * means the multi-task learning is used, and ‡ indicates that the attention mechanism module is added to the backbone network and the mutual learning method is used.

The weight parameter λ controls global information in the aggregation process of local features. As shown in Figure 6, we show that the accuracy varies with the balance parameter λ. On the one hand, when λ gradually increases, we can see that the performance improves, which indicates that global information aggregation is beneficial for the local feature attention weights. On the other hand, we can see that the performance decreases when λ is larger than 0.2, which indicates that too much global information for the aggregation of local features appears the error interference phenomenon. In a word, introducing too much global information and little global information cannot offer the accurate completed correlation, which leads to a suboptimal performance. Thus, we obtain the best results when λ is set to 0.2.

As shown in Figure 7 and Figure 8, for β_1, β_2 and β_3 in Equation (10) we fix two parameters to the optimal values and investigate the impact of the remaining one for the convenience of display. We also apply the same method to investigate the impact of μ_1, μ_2 and μ_3 in Equation (12) as shown in Table 4. When β_1, β_2 and β_3 are set to 1, 0.5 and 0.5 respectively, the performance of the network achieves the best. The optimal values of μ_1, μ_2 and μ_3 are 0.1, 1 and 0.5, respectively.

4.6 | Visualisation

In this subsection, we first visualise the similarity score of RGB-IR positive and negative pairs as shown in Figure 9. The difference between the distributions of RGB-IR positive and negative pairs for BS + IGAT is larger than that of BS, and therefore the correct matching is more probably to occur. It demonstrates that IGAT is beneficial to learn discriminative features.

We also report t-SNE [43] visualisation of 10 randomly selected identities on RegDB. The feature distributions of BS, $BS + L_{tra}$ and BS + MCCL are shown in Figure 10. Comparing $BS + L_{tra}$ with BS, we can see that the modality gap is alleviated largely. After adding the modality centre constraint and the

TABLE 3 Comparison on RegDB

		V-T		*T-V*	
Methods	**Venue**	**R1**	**mAP**	**R1**	**mAP**
Zero-padding [8]	ICCV'17	17.8	18.9	16.7	17.9
TONE [19]	AAAI'18	24.4	20.8	21.7	22.2
BDTR [9]	IJCAI'18	33.5	31.8	32.7	31.1
MAC [39]	MM'19	36.4	37.3	36.2	36.6
D^{2}RL [13]	CVPR'19	43.4	44.1	–	–
D-HSME [20]	AAAI'19	50.9	47.0	50.2	46.2
AlignGAN [26]	ICCV'19	57.9	53.6	56.3	53.4
MSR [40]	TIP'20	48.4	48.7	–	–
JSIA-ReID [14]	AAAI'20	48.5	49.3	48.1	48.9
Xmodal [16]	AAAI'20	62.2	60.2	–	–
DDAG [18]	ECCV'20	69.3	63.4	68.0	61.8
Hi-CMD [15]	CVPR'20	70.9	66.0	–	–
DAPR [41]	IVC'21	61.5	59.4	–	–
AGW [11]	TPAMI'21	70.0	66.4	71.6	65.2
FBP-AL [50]	TNNLS'21	74.0	68.2	70.1	66.6
CMAlign [49]	ICCV'21	74.2	67.6	72.4	65.5
NFS [42]	CVPR'21	80.5	72.1	78.0	69.8
GLMC* [47]	TNNLS'21	91.8	81.4	91.1	81.1
MPANet‡ [48]	CVPR'21	83.7	80.9	82.8	80.7
Ours	–	84.1	75.4	83.1	76.0

Note: Here, * means the multi-task learning is used, and ‡ indicates that the attention mechanism module is added to the backbone network and the mutual learning method is used.

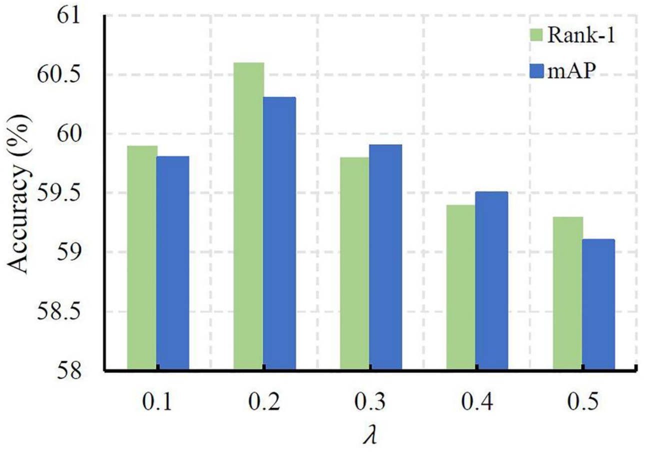

FIGURE 6 The experimental results with different λ

inter-identity centre constraint, the distance between cross modality features is pulled closer and the features with the same identity become more compact, which validates that the influence of the modality information is relieved by MCCL.

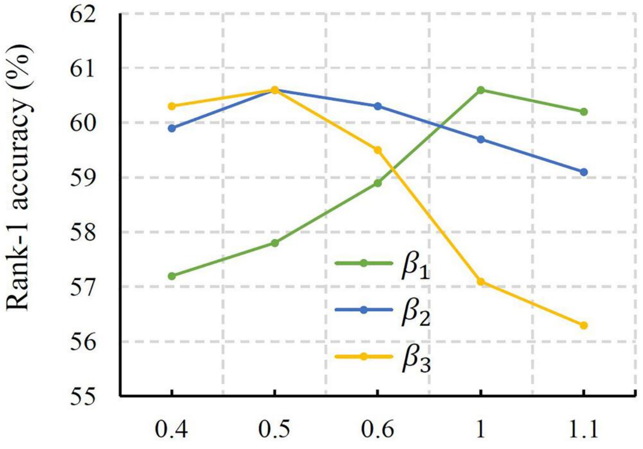

FIGURE 7 The rank-1 accuracy with different β_1, β_2 and β_3

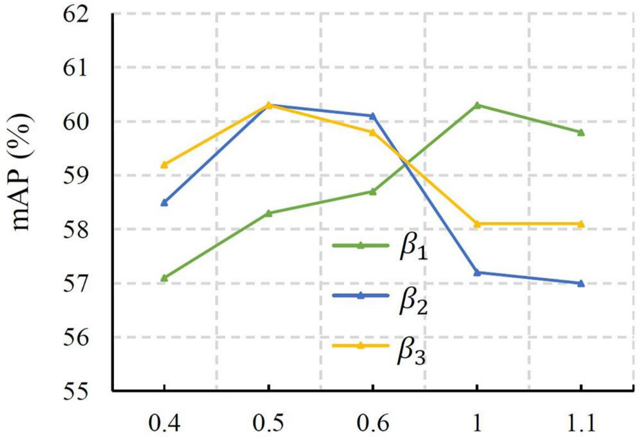

FIGURE 8 The mAP with different β_1, β_2 and β_3

TABLE 4 The experimental results with different μ_1, μ_2 and μ_3

μ_1	0.01	0.05	0.1	0.2
Rank-1 (%)	59.7	59.8	**60.6**	60.2
mAP (%)	59.7	59.8	**60.3**	60.0
μ_2	0.8	0.9	1	1.1
Rank-1 (%)	57.3	58.8	**60.6**	60.6
mAP (%)	56.2	57.5	**60.3**	60.1
μ_3	0.3	0.4	0.5	0.6
Rank-1 (%)	55.2	57.9	**60.6**	59.5
mAP (%)	54.2	57.4	**60.3**	60.0

5 | CONCLUSION

In this paper, we have proposed IGAT and MCCL for cross-modality person Re-ID. The proposed IGAT considers both local detail and global information to construct completed correlation between local features. To explicitly overcome the influence of modality information, we propose MCCL which

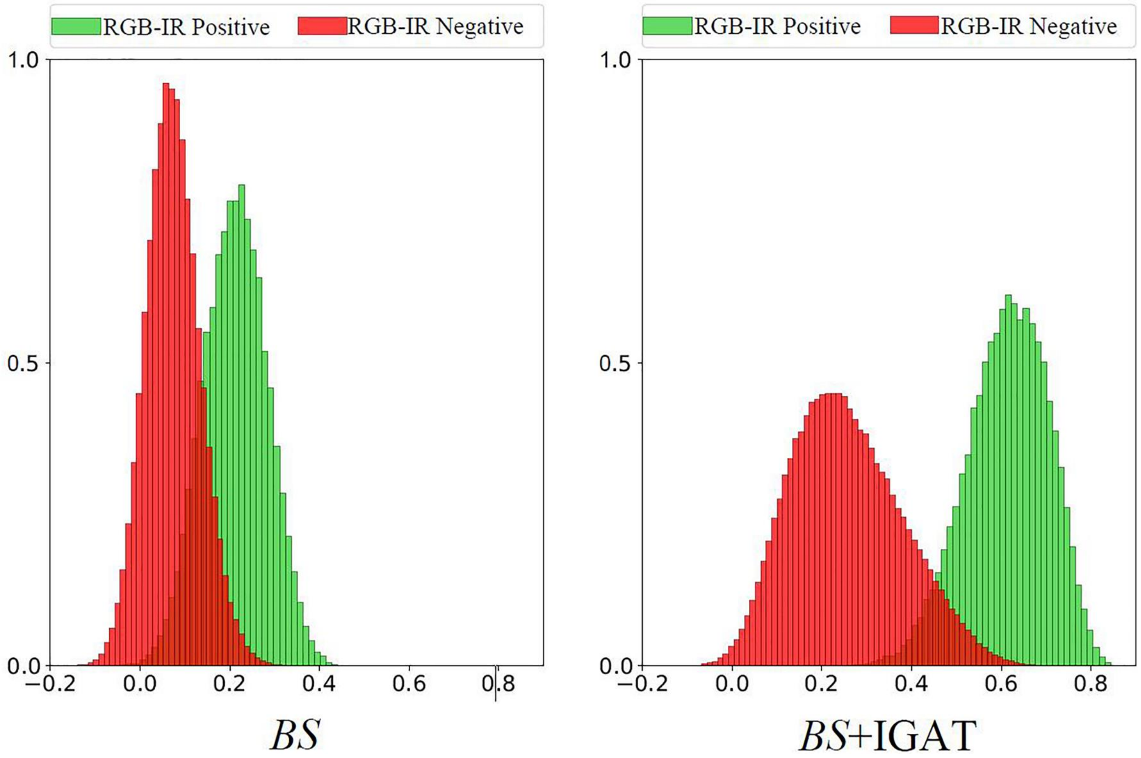

FIGURE 9 The visualisation of similarity score of RGB-IR positive and negative pairs on SYSU-MM01. The *x* axis is the similarity score of cross-modality image pair (the image pair with the same identity is called RGB-IR positive and the image pair with different identities is called RGB-IR negative). The *y* axis is the statistical normalisation value for each similarity score. IR, infrared

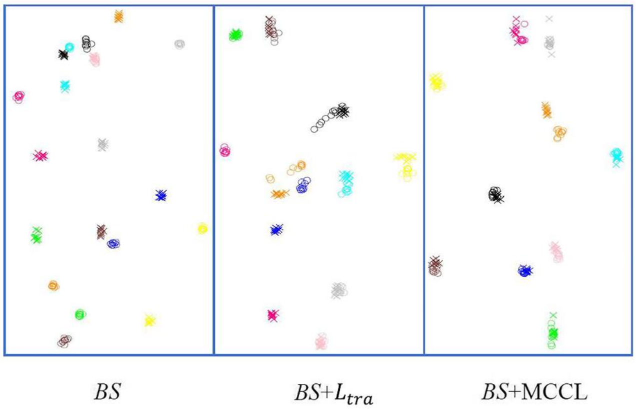

FIGURE 10 Visualisations of the feature distribution generated from different models. The colour and shape of points indicate the identity and modality, respectively. The figure is best viewed in colour with PDF magnification

constrains the centres of modality and identity to optimise the similarity of features. Extensive experimental results on two standard datasets have demonstrated the proposed method surpasses the state-of-the-art methods. Moreover, the proposed method is good at handling heterogeneous data, and therefore we believe that our method has great potential to generalise to other related research fields, such as cross-modality image retrieval, domain adaptation image classification, and so on.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under Grant No. 62171321, Natural Science Foundation of Tianjin under Grant No. 20JCZDJC00180 and No. 19JCZDJC31500, and the Key Project of Graduate Scientific Research Innovation Project of Tianjin Normal University under Grant No. 2022KYCX032Z.

CONFLICT OF INTEREST

No conflict of interest statement.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in SYSU-MM01 at https://www.isee-ai.cn/project/RGBIRReID.htm, reference number [8] and RegDB at http://dm.dongguk.edu/link.html, reference number [38].

ORCID

Shuang Liu https://orcid.org/0000-0002-9027-0690

REFERENCES

1. Saghafi, M.A., et al.: Review of person re-identification techniques. IET Comput. Vis. 8(6), 455–474 (2014). https://doi.org/10.1049/iet-cvi.2013.0180
2. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future (2016). arXiv preprint arXiv:1610.02984
3. Fumera, B.L.G., Roli, F.: Multi-stage ranking approach for fast person re-identification. IET Comput. Vis. 12(4), 513–519 (2018). https://doi.org/10.1049/iet-cvi.2017.0240

4. Chen, Y., Zheng, W., Lai, J.: Mirror representation for modeling view-specific transform in person re-identification. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 3402–3408 (2015)
5. Suh, Y., et al.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 402–419 (2018)
6. Yu, H., Wu, A., Zheng, W.: Unsupervised person re-identification by deep asymmetric metric embedding. IEEE Trans. Pattern Anal. Mach. Intell. 42(4), 956–973 (2018). https://doi.org/10.1109/tpami.2018.2886878
7. Zhang, J.A., Wang, Q., Yuan, Y.: Metric learning by simultaneously learning linear transformation matrix and weight matrix for person re-identification. IET Comput. Vis. 13(4), 428–434 (2019). https://doi.org/10.1049/iet-cvi.2018.5402
8. Wu, A., et al.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5390–5399 (2017)
9. Ye, M., et al.: Visible thermal person re-identification via dual-constrained top-ranking. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1092–1099 (2018)
10. Zhu, Y., et al.: Hetero-center loss for cross-modality person re-identification. Neurocomputing 386, 97–109 (2020). https://doi.org/10.1016/j.neucom.2019.12.100
11. Ye, M., et al.: Deep learning for person re-identification: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 44(6), 2872–2893 (2021). https://doi.org/10.1109/TPAMI.2021.3054775
12. Dai, P., et al.: Cross-modality person re-identification with generative adversarial training. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 677–683 (2018)
13. Wang, Z., et al.: 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, pp. 618–626 (2019)
14. Wang, G., et al.: Cross-modality paired-images generation for RGB-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12144–12151 (2020)
15. Choi, S., et al.: HI-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10257–10266 (2020)
16. Li, D., et al.: Infrared-visible cross-modal person re-identification with an X modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4610–4617 (2020)
17. Ye, M., Shen, J., Shao, L.: Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Trans. Inf. Forensics Secur. 16, 728–739 (2020). https://doi.org/10.1109/tifs.2020.3001665
18. Ye, M., et al.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 229–247 (2020)
19. Ye, M., et al.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7501–7508 (2018)
20. Hao, Y., et al.: HSME: hypersphere manifold embedding for visible thermal personre-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8385–8392 (2019)
21. Ye, M., et al.: Bi-directional center-constrained top-ranking for visible thermal person re-identification. IEEE Trans. Inf. Forensics Secur. 15, 407–419 (2020). https://doi.org/10.1109/tifs.2019.2921454
22. Jia, M., et al.: A similarity inference metric for RGB-Infrared cross-modality person re-identification. In: Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence, pp. 1026–1032 (2020)
23. Sun, J., et al.: Visible-infrared cross-modality person re-identification based on whole-individual training. Neurocomputing 440, 1–11 (2021). https://doi.org/10.1016/j.neucom.2021.01.073
24. Chen, W.J., et al.: Semi-supervised user profiling with heterogeneous graph attention networks. In: Proceedings of the International Joint Conference on Artificial Intelligence, vol. 19, pp. 2116–2122 (2019)
25. Yang, T.C., et al.: HGAT: heterogeneous graph attention networks for semi-supervised short text classification. ACM Trans. Inf. Syst. 39(3), 1–29 (2021). https://doi.org/10.1145/3450352
26. Wang, G., et al.: RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3622–3631 (2019)
27. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 729–734 (2005)
28. Scarselli, F., et al.: The graph neural network model. IEEE Trans. Neural Network. 20(1), 61–80 (2008). https://doi.org/10.1109/tnn.2008.2005605
29. Veličković, P., et al.: Graph attention networks. In: Proceedings of the International Conference on Learning Representations (2018)
30. Mu, N., et al.: Graph attention networks for neural social recommendation. In: Proceedings of the IEEE International Conference on Tools with Artificial Intelligence, pp. 1320–1327 (2019)
31. Zhu, K., Cao, M.: A semantic subgraphs based link prediction method for heterogeneous social networks with graph attention networks. In: Proceedings of the International Joint Conference on Neural Networks, vol. 8, pp. 1–8 (2020)
32. Huang, B., Carley, K.M.: Syntax-aware aspect level sentiment classification with graph attention networks (2019). arXiv preprint arXiv:1909.02606
33. Wang, K., et al.: Relational graph attention network for aspect-based sentiment analysis (2020). arXiv preprint arXiv:2004.12362
34. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
35. Sun, Y., et al.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision, pp. 480–496 (2018)
36. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8295–8302 (2019)
37. Zhang, Z., Zhang, H., Liu, S.: Person re-identification using heterogeneous local graph attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12136–12145 (2021)
38. Nguyen, D.T., et al.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors 17(3), 605 (2017). https://doi.org/10.3390/s17030605
39. Ye, M., Lan, X., Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp. 347–355 (2019)
40. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Trans. Image Process. 29, 579–590 (2020). https://doi.org/10.1109/tip.2019.2928126
41. Zhang, P., et al.: Beyond modality alignment: learning part-level representation for visible-infrared person re-identification. Image Vis. Comput. 108, 104118 (2021). https://doi.org/10.1016/j.imavis.2021.104118
42. Chen, Y., et al.: Neural feature search for RGB-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 587–597 (2021)
43. Laurens, V.D.M., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. 9(11), 2579–2605 (2008)
44. Xia, D.X., et al.: Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network. Neurocomputing 443, 35–36 (2021). https://doi.org/10.1016/j.neucom.2021.02.088
45. Cai, X., et al.: Dual-modality hard mining triplet-center loss for visible infrared person re-identification. Knowl. Base Syst. 215, 106772 (2021). https://doi.org/10.1016/j.knosys.2021.106772
46. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT, pp. 177–186 (2010)

47. Zhang, L.Y., et al.: Global–local multiple granularity learning for cross-modality visible–infrared person reidentification. IEEE Transact. Neural Network Learn. Syst., 1–11 (2021). https://doi.org/10.1109/tnnls.2021.3085978
48. Wu, Q., et al.: Discover cross-modality nuances for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4330–4339 (2021)
49. Park, H., et al.: Learning by aligning: visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12046–12055 (2021)
50. Wei, Z.Y., et al.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Transact. Neural Network Learn. Syst., 1–2 (2021). https://doi.org/10.1109/tnnls.2021.3059713

How to cite this article: He, D., et al.: Integration graph attention network and multi-centre constrained loss for cross-modality person re-identification. IET Comput. Vis. 1–12 (2022). https://doi.org/10.1049/cvi2.12132