Received: 25 September 2022 | Revised: 5 December 2022 | Accepted: 4 January 2023

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12182

ORIGINAL RESEARCH

Multi-granularity re-ranking for visible-infrared person re-identification

Yadi Wang[1,2] | Hongyun Zhang[1,2] | Duoqian Miao[1,2] | Witold Pedrycz[3,4]

[1]Department of Computer Science and Technology, Tongji University, Shanghai, China

[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai, China

[3]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

[4]System Research Institute, Polish Academy of Sciences, Warsaw, Poland

Correspondence

Hongyun Zhang, Department of Computer Science and Technology, Tongji University, Shanghai, 201804 China.
Email: zhanghongyun@tongji.edu.cn

Funding information

Jiangxi "Double Thousand Plan"; National Natural Science Foundation of China, Grant/Award Numbers: 61976158, 62076182

Abstract

Visible-infrared person re-identification (VI-ReID) is a supplementary task of single-modality re-identification, which makes up for the defect of conventional re-identification under insufficient illumination. It is more challenging than single-modality ReID because, in addition to difficulties in pedestrian posture, camera shooting angle and background change, there are also difficulties in the cross-modality gap. Existing works only involve coarse-grained global features in the re-ranking calculation, which cannot effectively use fine-grained features. However, fine-grained features are particularly important due to the lack of information in cross-modality re-ID. To this end, the Q-center Multi-granularity K-reciprocal Re-ranking Algorithm (termed QCMR) is proposed, including a Q-nearest neighbour centre encoder (termed QNC) and a Multi-granularity K-reciprocal Encoder (termed MGK) for a more comprehensive feature representation. QNC converts the probe-corresponding modality features into gallery corresponding modality features through modality transfer to narrow the modality gap. MGK takes a coarse-grained mutual nearest neighbour as the dominant and combines a fine-grained nearest neighbour as a supplement for similarity measurement. Extensive experiments on two widely used VI-ReID benchmarks, SYSU-MM01 and RegDB have shown that our method achieves state-of-the-art results. Especially, the mAP of SYSU-MM01 is increased by 5.9% in all-search mode.

KEYWORDS

computer vision, recognition

1 | INTRODUCTION

Person re-identification aims at matching individual pedestrian images in a probe set from a large gallery set. Because of different body poses, shooting angles and environmental conditions, ReID becomes a challenging job. Many works are mainly suitable for RGB images captured by visible cameras and have achieved great success. However, visible cameras can only capture high-quality images under good illumination conditions, such as daytime. With the development of hardware devices, the dual-mode camera has been widely used for public security. The dual-mode camera can switch the day to night mode according to the light intensity or the predefined time point in the system. The visible camera is used to capture RGB images of three channels in the day mode, while the infrared camera is used to capture infrared images of the single channel in the night mode. Hence, visible-infrared person re-identification (VI-ReID) makes up for the single modality ReID under night and indoor conditions which lack light.

At present, the existing researches mainly focus on how to map the two modalities' features to the same feature space in the training stage. Due to the heterogeneity of data, the feature difference of the same pedestrian in different modalities is often greater than that of different pedestrians in the same modality, causing difficulty in matching. However, there are few studies on the re-ranking phase of cross-modality tasks, which greatly improves the matching effect and is also a very important part in the ReID process.

At the same time, due to the modality specificity in the VI-ReID task, the modality-shared features will lack some information. For example, the color information of the visible modality cannot be used in the matching process with the infrared modality, so it is particularly important to pay attention to the local details. Depending on the particularity of the human body structure, the distribution of various parts of the human body in the monitoring picture is relatively fixed. In the person re-ID, PCB architecture is often used to divide the extracted features horizontally to obtain the fine-grained features of the corresponding location. However, most networks use global features and fine-grained features in the training stage and only use global features for similarity measurement in the testing stage. To make effective use of the fine-grained characteristics of samples in the testing stage, this paper proposes a re-ranking algorithm based on multi-granularity for cross-modality person re-identification.

In this study, we propose an innovative re-ranking algorithm for the person VI-ReID, which considers features at different granularities both globally and locally. As shown in Figure 1a, in some cases, the ranking calculated by local features is more accurate than that calculated by global features. At the same time, because some fine-grained features under one modality may be unavailable under another modality in the cross-modality task, it is necessary to pay attention to other available fine-grained features. Fine-grained features should only be complementary, and the ranking list still needs to be dominated by global features. Therefore an improved multi-granularity k-reciprocal re-ranking algorithm (MGK) is proposed to make the measurement pay more attention to special features by modifying the calculation process of reciprocal neighbour. Specifically, the coarse and fine-grained features used in the calculation can be simultaneously obtained from one two-stream feature extraction network containing the PCB structure.

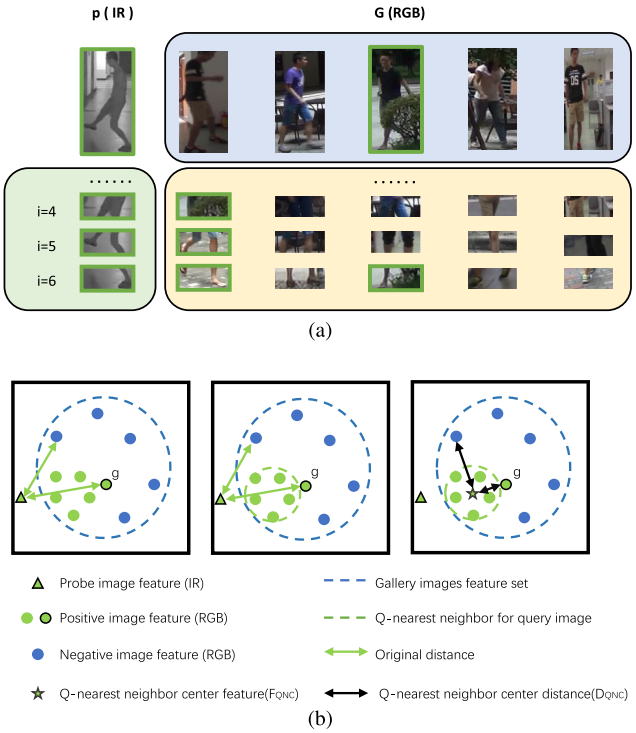

FIGURE 1 (a) The similarity ranking obtained by global features and different local feature blocks for re-ranking. The images in the green box are the local feature block of the probe images; The images in the blue box is the similarity ranking calculated by the global feature; The images in the yellow box are the similarity ranking calculated by local feature blocks (The more images in the front column, the higher the similarity); The images with the green edge has the same identity as the probe image. (b) Measure the cross-modality similarity by Q-nearest neighbour center encoding (QNC).

Although the feature extraction network aims to map the features of two different modalities into the same feature space, due to the difference between the two modalities, the features extracted from images with the same modality are more clustered even through the feature extraction network. As the left part of Figure 1b, the distance between the directly extracted probe image feature and the gallery set image g with the same identifier may be far. Thus, we introduce Q-nearest neighbuor centre encoding (QNC), which mitigates the difference of two modalities feature spaces in cross-modality samples. As the middle and right parts of Figure 1b, finding the Q-nearest neighbours with the same identity in another modality, and representing the probe central feature by the mean of neighbours' features. Therefore, the image feature encoder with the same identity as the probe is more similar to the probe's feature encoder, while the image feature encoder with different identity as the probe is more distinguishable from the probe's feature encoder.

We conduct extensive experiments to demonstrate the effectiveness of our proposed method. The contributions of this paper are three-fold:

- We propose a novel multi-granularity k-reciprocal re-ranking (MGK) algorithm for VI-ReID to enhance the comprehensiveness of measurement by aggregating local and global characteristics.
- We introduce a cross-modality feature approximation scheme (QNC) to extract the person similarity matrix, which smoothens the re-ranking and reduces the modality gap.
- The proposed method effectively improves the re-ranking performance in person VI-ReID on the two mainstream benchmark datasets SYSU-MM01 and RegDB in both rank one and mAP.

2 | RELATED WORK

2.1 | Person Re-identification

Person Re-identification aims at finding the matching with probe images in the gallery set. The matched pedestrians with the same identifier may be taken by different cameras or taken by the same camera at different times. Common ReID methods can be divided into two categories: methods based on representation learning and methods based on metric learning. After extracting pedestrian features, the former regard ReID as

a multi-classification problem and takes each pedestrian ID as a category; the latter calculates the similarity between two pedestrians and judge whether they have the same pedestrian ID.

The research scope of person ReID is very wide, it is divided into single modality ReID and cross-modality Re-ID. At present, the supervised single modality ReID research has achieved encouraging results. Some methods, such as ABD [1], SONA [2], VAL [3], whose experimental performance on public datasets is close to or even better than the manual matching performance. Therefore, lots of researches have shifted to cross-modality Re-ID, such as text-image Re-ID [4, 5], RGB-Depth Re-ID [6, 7] and VI Re-ID.

2.2 | Visible-infrared person Re-identification

Visible-infrared person Re-identification aims at matching individual pedestrian images in a probe set from a large gallery set, which consists of pictures taken by the other modality camera. Wu et al. [8] create the first large-scale RGB-IR person ReID dataset named SYSU-MM01 and propose a zero-padding network to finish the cross-modality person ReID task. To reduce the cross-modality difference, Dai et al. [9] propose cmGAN that extracts the feature vector by the generator and map it to the corresponding pedestrian identity. Many scholars deal with the difference between the two modes by improving the cross-modality loss function in the network. Ye et al. [10] propose Dual-Constrained Top-Ranking that considers both cross-modality and intra-modality variability. Hao et al. [11] propose DFE that constrains two modal feature distributions by JS divergence. HC loss [12] is proposed to achieve modality alignment by constraining the heterogeneous modality feature centre.

Other methods are based on modality transition, focusing on converting one modality image into the other modality image as real as possible. Wang et al. [13] first propose AlignGAN to transform the infrared images into RGB images by CycleGAN. D^{2}RL [14] is proposed to exploit variational autoencoders (VAEs) for style disentanglement, which is followed by GANs to generate another modality images. Choi et al. [15] propose Hi-CMD that automatically disentangles ID-discriminative factors for cross-modality matching. JSIA-ReID [16] propose set-level alignment, which distinguishes modality-specific features and mode invariant features, which is more conducive to judgment. However, modality transition only realises the one-to-one generation from the IR image to RGB image, but in fact, the same IR image may correspond to RGB images of multiple colours. Simultaneously it destroys part of the original information and leads to a certain degree of noise.

2.3 | Re-ranking for person Re-ID

Re-ranking aims at making the correctly matched images higher ranks after obtaining an initial ranking list. Le et al. [17] proposed Common Near-Neighbor Analysis, which extracts new features in the matching stage, and both consider the relative information and direct information of sample pairs. DCIA [18] was proposed to suppress the interference of context information in the sample and emphasise the function of content information. The correct search results in the gallery set should be similar, so the method based on the nearest neighbour is also widely used. Ye et al. [19] proposed a rank aggregation method that combines the result of the global-based rank method and local-based rank method by crossed k-nearest neighbours. Then they proposed the united similarity ranking aggregation method and dissimilarity ranking aggregation method, which also calculate quasi-dissimilar sets independently. Zhong et al. [20] proposed k-reciprocal encoding which considers not only the nearest neighbour relationship between the probe and gallery set but also the nearest neighbor relationship between the gallery set and probe set and add Jaccard distance to revise the initial rank list. However, the above methods are pointed at the re-ranking method under the single modality, and there is less research on cross-modality at present. Compared to the single modality, cross-modality analysis requires considering the modality gap caused by heterogeneous data characteristics. Jia et al. [21] proposed a similarity inference metric for VI-ReID that makes up intra-modality k-reciprocal neighbours by cross-modality k-nearest neighbours and adding the similarity graph reasoning as an auxiliary.

3 | PROPOSED METHOD

Given a RGB probe image p, and a gallery set $G = \{g_i | i = 1, 2, \ldots, N_G\}$ with N_g infrared person images, VI-ReID re-ranking aims to obtain a new matching sequence for p according to the similarity.

3.1 | Two-stream feature extractor

VI-ReID method based on feature fusion often uses a one-stream feature extraction network or a two-stream feature extraction network as the backbone. The difference is that the first one uses the network layer with the same structure and same parameters for feature extraction, while the next one uses the network layer with the same structure with different parameters which trained separately according to the modalities. As shown in Figure 2a, the first two convolution blocks of ResNet are independent for extracting the specific features of the two modalities, and the last three convolution blocks parameters are shared to map the different modality features to the same feature space. In the preprocessing stage, the single infrared channel repeats three times to ensure the consistency of the input image in two modalities. Different feature extraction branch networks have the same structure and different parameters. The two-stream feature extraction network extracts global features F_g and then PCB segments local features. Firstly,

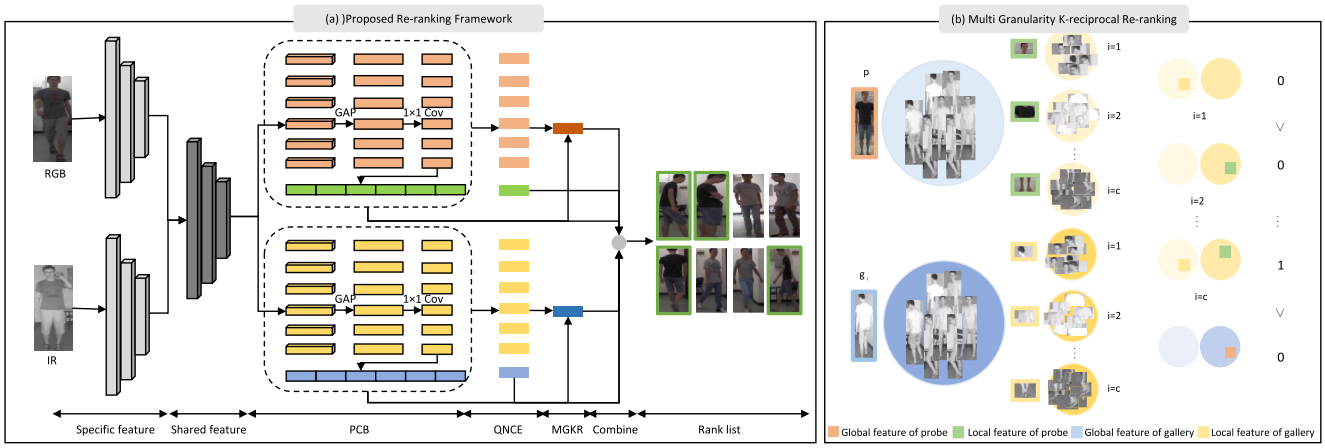

FIGURE 2 (a) The proposed Re-ranking Framework. RGB and IR images obtain initial features through different feature extraction networks, adopt QNC encoding and multi-granularity k-reciprocal encoding and calculate the ranking list after combination. (b) The partial visualisation process of multi-granularity k-reciprocal re-ranking. To judge whether g_i belongs to the k-reciprocal set of p, first find the local and global nearest neighbour sets and then judge whether the corresponding part is in the other's nearest neighbour set, such as $i = c$. If any part satisfies the reciprocal neighbor condition, g_i is in the reciprocal neighbor set of p.

the feature is horizontally divided into C feature blocks F_L, reducing dimension for each feature block by global average pooling and 1×1 convolution layer.

3.2 | Multi-granularity K-reciprocal Re-ranking

A person needs to pay attention to both coarse-grained global pedestrian features (such as body shape) and fine-grained local pedestrian features (such as shoes, backpacks, etc.). Especially among similar samples, local features are particularly important. Therefore, we improve the k-reciprocal neighbour loss in the single modality and propose a multi-granularity k-reciprocal re-ranking method that depends on both local and global information, as shown in Figure 2b. Firstly, the initial distance $d_c(p, g_i)$ between the probe image p and the gallery image g_i is defined as Mahalanobis distance:

$$d_c(p, g_i) = \sqrt{(F(q) - F(g_i))^T \Sigma^{-1} (F(q) - F(g_i))} \quad (1)$$

where Σ is the covariance matrix. Through the standardisation of the feature vector, Σ can be regarded as an identity matrix. Sorted by $d_c(p, g_i)$, we define n_l as the k nearest neighbours of the sample q computed from the corresponding local feature block at the same location:

$$n_l(p, k) = \left\{ g_l^0, g_l^1, \ldots, g_l^k \right\}, |n_l(p, k)| = k \quad (2)$$

The k-nearest neighbour of the sample calculated from all the local feature and global feature can be defined as $N_l(p, k)$ and $N_g(p, k)$ respectively:

$$\begin{aligned} N_l(p, k) &= \left\{ n_l^1(p, k) \cup n_l^2(p, k) \cup \cdots \cup n_l^c(p, k) \right\} \\ N_g(p, k) &= \left\{ g_g^0, g_g^1, \ldots, g_g^k \right\}, |N_g(p, k)| = k \end{aligned} \quad (3)$$

where c is the number of local feature blocks divided in PCB. g_l is the local block feature through the sample g^k, and g_g is the global feature. According to the previous description, we can define local k-nearest neighbours for a block as

$$r_l(p, k) = \{ (g_i \in n_l(p, k) \cap (p \in n_l(g_i, k))) \} \quad (4)$$

Accordingly, we can define local and global k-reciprocal neighbours respectively as

$$\begin{aligned} R_l(p, k) &= \left\{ r_l^i(p, k) | i = 1, 2 \cdots c \right\} \\ R_g(p, k) &= \left\{ (g_i \in N_g(p, k) \cap (p \in N_g(g_i, k)) \right\} \end{aligned} \quad (5)$$

Considering both global and local features can enhance the robustness of matching. As shown in Figure 1a, due to the influence of the shooting angle and light, the global characteristics of pedestrians with the same identity may be different, but their local characteristics may be similar. Therefore, we define R' as the union of the global k-nearest neighbour and multiple local k-nearest neighbours based on multi-granularity.

$$R'(p, k_1, k_2) = \left\{ R_g(p, k_1) \cup R_l(p, k_2) \right\} \quad (6)$$

where k_1 is the number of global nearest neighbours, and k_2 is the number of local nearest neighbours. According to the single-modality K-reciprocal encoding definition [20], extending R' to R'' can strengthen the robustness of the algorithm and help to supplement the matching samples missed in the initial ranking. In order to reduce the computation, we only traverse the global nearest neighbour.

$$R''(p,k_1,k_2) \leftarrow R'(p,k_1,k_2) \cup R_g\left(q,\frac{1}{2}k_1\right)$$
$$s.t.\left|R'(p,k_1,k_2) \cap R_g\left(q,\frac{1}{2}k_1\right)\right| \geq \frac{2}{3}\left|R_g\left(q,\frac{1}{2}k_1\right)\right| \quad (7)$$
$$\forall q \in R'(p,k_1,k_2)$$

where $|.|$ denotes the number of candidates in the set. Following the assumption that samples with the same identity have the same nearest neighbors, we define Jaccard distance as,

$$d_J(p,g_i) = 1 - \frac{|R''(p,k_1,k_2) \cap R''(g_i,k_1,k_2)|}{|R''(p,k_1,k_2) \cup R''(g_i,k_1,k_2)|} \quad (8)$$

where $|\cdot|$ denotes the number of candidates in the set. In the subsequent implementation, the R'' sets are represented as vectors according to the original k-reciprocal neighbor implementation, and the original distance between the two samples dominated by global features is used as the initial vector weight. Therefore, the overall reordering result is still dominated by global features, and local features are only used as supplements.

3.3 | Q-nearest neighbor center encoding

Due to the large difference between the two modalities, the feature extraction network may not be able to map them to the same feature space. For example, in Figure 3b, the distance between the feature vectors of different pedestrians in the same modality may be smaller than that of the same pedestrian in different modality. Therefore, we use the Q-nearest neighbour feature instead of the original feature representation F. The basic idea is that the characteristics of a modality sample p can be represented by the characteristics of the other modality samples, which are close to p. We define the Q-nearest neighbour centre encoding in the other modality for p as F_{QNC}. According to the transformation, the influence of cross-modality on feature measurement can be determined.

$$F_{QNC}(p,q) = \sum_{i=1}^{q}(F(g_i))|\ \forall\, g_i \in N(p,q) \quad (9)$$

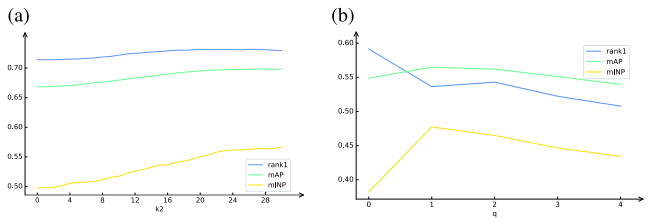

FIGURE 3 Parametric Analysis on SYSU-MM01. (a) Different k_2 in MGK. (b) Different q in QNC.

where $F_{QNC}(p, q)$ is the Q-nearest neighbour centre encoding for image p. $N(p, q)$ is the q-nearest neighbor set of p contains q images from gallery set. According to the previous definition, $N(p, q)$ is determined by Mahalanobis distance. The Q-nearest neighbour centre distance between p and g_i can be defined as

$$d_{QNC}(p,g_i,q) = \sqrt{\left(F_{QNC}(p,q) - F(g_i)\right)^T \Sigma^{-1}\left(F_{QNC}(p,q) - F(g_i)\right)} \quad (10)$$

According to Equation (10), the Q-nearest neighbour centre distance instead of the original distance as a measure can reduce the modality gap.

Algorithm 1 QCMR Algorithm

Input: Probe set global feature $F_{P,g}$, probe set local feature $F_{P,l}$, gallery set global feature $F_{G,g}$, gallery set local feature $F_{G,l}$
Calculate $d_c(p, g_i)$ as Equation (1)
Initializs q-nearest neighbour centre encoding $F'_{P,g}$ and $F'_{P,l}$, $F'_{G,g}$, $F'_{G,l}$ as Equation (10)
Calculate $d_{qnc}(p, g_i)$ as Equation (1)
for each p,g in P,G **do**
  Calculate k_1 nearest neighbours N_g F'_g and k_2 nearest neighbors N_g with F'_l as Equations (2) and (3).
  Calculate global k_1 reciprocal neighbours R_g and local k_2 reciprocal neighbours R_l as Equations (4) and (5)
  Calculate multi-granularity k-reciprocal R' as Equation (6)
end for
for each p,g in P,G **do**
  Calculate extended neighbours multi-granularity k-reciprocal R'' as Equation (7)
end for
for each p in P **do**
  Calculate Jaccard distance d_J as Equation (8)
end for
Calculate final distance d' as Equation (11)
Output: Distance between each probe and gallery set d'

3.4 | Q-center multi-granularity K-reciprocal Re-ranking

The proposed Q-center Multi-granularity k-reciprocal Re-ranking method (QCMR) can thus be derived by combining

QNC and MGK. To more comprehensively measure the distance between different features, we jointly aggregate Jaccard distance, Q-nearest neighbour centre distance and the original distance as the final distance d',

$$d'(p, g_i) = d_J(p, g_i) + \lambda_1 d_{QNC}(p, g_i) + \lambda_2 d_C(p, g_i) \quad (11)$$

where λ_1 and λ_2 donate the trade-off parameter for d_{QNC} and d_C, making the model more flexible. d_{QNC} and d_C are measured only by global features. Algorithm 1 provides the detailed description of our proposed re-ranking method and inference metric.

4 | EXPERIMENTS

4.1 | Experimental setting

Datasets. SYSU-MM01 is the first large-scale dataset in VI-ReID, which includes IR images from two infrared cameras and RGB images from four visible cameras. There are 491 pedestrians with different identities in total, including 296 for training, 99 for verification and 96 for testing. There are 30071 RGB images and 15792 IR images. The shooting scenes include indoor and outdoor. RegDB contains images captured by dual camera systems (one visible camera and one infrared camera). It contains a total of 412 pedestrians with different identities, and 10 RGB images and 10 IR images were taken for each person. The dataset was randomly divided into two halves, one for training and the other for testing. The procedure was repeated in 10 trials and stable results were obtained according to the average value.

Evaluation protocols. We use standard Cumulative Matching Characteristics curve, mean Average Precision (mAP) and mean inverse negative penalty (mINP) to evaluate the cross-modality person re-identification models. The difference from single modality and the images of one modality are regarded as probe samples, and the images of the other modality are regarded as gallery samples.

Implementation details. We implement the proposed method in PyTorch. The Resnet50 is adopted as the backbone network for feature extraction. We adopt the two-stream parameter shared feature extraction network [33] as the baseline, in which the feature extraction part mainly includes two-stream backbone network with partial parameter sharing and part-level feature extraction block, as shown in Section 3.1. The training loss includes hetero-center triplet loss and identity softmax loss for local features, and hetero-center triplet loss for global features. As the setting for Ref. [33], the first two convolution block parameters are independent, and the last three convolution block parameters are shared to extract modal specificity and modal sharing features, respectively. The training algorithm is optimised with SGD for 60 epochs with a learning rate of 0.1, adopting the warmup strategy for the first 10 epochs and decaying 10 times at 20 and 50 epochs.

4.2 | Comparison with state-of-the art methods

For a more comprehensive evaluation, our method is compared with a large number of existing VI-ReID methods, including methods based on feature mapping (such as TONE, BDTR and AWG) and methods based on modality transformation (such as AlignGAN, JSIA-ReID and Hi-CMD). Table 1 shows the performance on the SYSU-MM01 and RegDB. For SYSU-MM01, we test the performance on the all-search model and indoor-search model under a single shot. For RegDB, we test the performance on the visible-to-thermal model and thermal-to-visible model. Our method has a competitive performance in all metrics, including Map, Rank1 and mInp. Because the PCB structure can be flexibly transplanted to various training networks, we can easily obtain local features and global features at the same time, which is the basis of our method. Essentially, the QCMR is a post-processing procedure. Therefore, it can achieve better results when applied to a better feature extraction network.

4.3 | Ablation study

Our method consists of two components. All experiments are based on the same feature extraction network and the same weight. In all ablation experiments, we only trained the network once, and other experiments related to re-ranking are loaded with the same weight. Table 2 and Table 3 report the resultant rank 1, mAP and mINP values on the all-search SYSU-MM01 and thermal-to-visible RegDB dataset by adding one component at a time. The baseline follows the setting in Ref. [33] and is ranked only by cosine similarity. The baseline* obtains the ranking list by the raw k-reciprocal encoding (KR). Without the k-reciprocal neighbour encoding, the direct cosine similarity measurement of QNC will reduce the r1 value by about 6%, but the other two evaluation indicators will be improved. It is ineffective to directly calculate the cosine distance by the QNC, because we cannot guarantee the complete correctness of the initial sorting. Instead, the amount of information will be lost due to the modal transition, thus leading to a drop in rank 1. Therefore, the QNC can only be supplemented in the re-ranking stage as a supplement, and we prefer to regard MGK and QNC as a whole. Our multi-granularity k-reciprocal re-ranking method is better than the k-reciprocal re-ranking method in every index, especially mINP increased by 7%. The MGK extends some locally similar samples that are not originally in the neighbour nodes into the calculation of the Jaccard distance, which facilitates a more comprehensive measurement. Furthermore, since the Gaussian kernel of the pairwise distance is used to form a vector in the actual operation, the number of neighbour intersections between samples is implicitly represented in the calculation process. In other words, the similarity of global features is regarded as a measure of the number of local neighbours. So we do not count the number of neighbours

TABLE 1 Performance of the proposed method compared with state-of-the-arts

		SYSU-MM01						RegDB					
Setting		All search			Indoor search			Visible to thermal			Thermal to visible		
Method	Venue	Rank 1	mAP	mINP	Rank 1	mAP	mINP	Rank 1	mAP	mINP	Rank 1	mAP	mINP
Zero-Padding [8]	ICCV'17	14.8	16.0	-	20.6	27.0	-	17.8	18.9	-	16.6	17.8	-
TONE [22]	AAAI'18	12.5	14.4	-	20.8	26.4	-	-	-	-	-	-	-
HCML [22]	AAAI'18	14.3	16.2	-	24.5	30.1	-	24.2	20.1	-	21.7	22.2	-
D-HSME [23]	AAAI'18	20.7	23.1	-	-	-	-	50.8	47.0	-	52.0	46.2	-
BDTR [10]	IJCAI'18	17.0	19.7	-	-	-	-	33.6	32.8	-	33.5	31.8	-
cmGAN [9]	IJCAI'18	27.0	31.5	-	31.6	42.2	-	-	-	-	-	-	-
D^2LR [14]	CVPR'19	28.9	29.2	-	-	-	-	43.4	44.1	-	-	-	-
MAC [24]	MM'19	33.3	36.2	-	36.4	37.0	-	36.4	37.0	-	36.2	36.6	-
AlignGAN [13]	ICCV'19	42.4	40.7	-	45.9	54.3	-	57.9	53.6	-	56.3	53.4	-
X-modality [25]	AAAI'20	49.9	50.7	-	-	-	-	62.2	60.2	-	-	-	-
JSIA-ReID [16]	AAAI'20	38.1	36.9	-	43.8	52.9	-	48.5	49.3	-	48.1	48.9	-
cm-SSFT [26]	CVPR'20	61.6	63.4	-	70.5	72.6	-	62.2	63.0	-	-	-	-
Hi-CMD [15]	CVPR'20	34.9	35.9	-	-	-	-	70.9	66.0	-	89.3	81.5	-
DDAG [27]	ECCV'20	54.8	53.0	-	61.0	68.0	-	69.3	63.5	-	68.1	61.8	-
CIMA [28]	AAAI'21	57.2	59.3	-	66.6	74.7	-	78.8	69.4	-	77.9	69.4	-
NFS [29]	CVPR'21	56.9	55.5	-	62.8	69.8	-	-	-	-	-	-	-
VSD [30]	CVPR'21	60.0	58.8	-	66.1	73.0	-	73.2	71.6	-	71.8	70.1	-
MPANet [31]	CVPR'21	70.6	68.2	-	76.7	81.0	-	83.7	80.9	-	82.8	80.7	-
AGW [32]	TIPAMI'21	47.5	47.7	35.3	54.2	63.0	59.2	70.1	66.4	50.2	-	-	-
QCMR	-	**72.7**	**72.7**	**64.3**	**77.5**	**79.8**	**77.9**	**95.6**	**95.9**	**95.5**	**95.3**	**95.8**	**95.4**

Note: The bold value means the best value in the current indicator.

TABLE 2 Ablation study on the SYSU-MM01

Method	KR	MGK	QNC	r1	mAP	mINP
Baseline				59.14	54.86	38.22
baseline + QNC			√	53.64	56.49	47.74
baseline*	√			71.04	66.83	49.73
baseline* + QNC	√		√	71.30	66.74	49.64
baseline + MGK		√		**73.17**	69.85	56.38
baseline + QCMR		√	√	72.65	**72.70**	**64.33**

Note: The baseline* obtains the ranking by the raw k-reciprocal encoding (KR). The bold value means the best value in the current indicator.

TABLE 3 Ablation study on the RegDB

Method	KR	MGK	QNC	r1	mAP	mINP
baseline				92.61	87.95	79.57
baseline + QNC			√	92.50	92.54	90.42
baseline*	√			94.69	95.24	94.53
baseline* + QNC	√		√	94.70	95.27	94.60
baseline + MGK		√		94.70	95.36	94.65
baseline + QCMR		√	√	**95.28**	**95.83**	**95.43**

Note: The baseline* obtains the ranking by the raw k-reciprocal encoding (KR). The bold value means the best value in the current indicator.

between samples, but simply judge whether the two are neighbours, whether locally or globally.

When the two components are all used, mAP can increase by 5%, and mINP is even able to increase by 12%. Although rank 1 is down by 0.5% compared to MGK only, it is still an improvement compared to the baseline. Therefore, the two components reinforce each other.

Parametric Analysis in MGK. Figure 3a shows the results for different k_2 in the local reciprocal neighbour computation. As the value of k_2 continues to increase, the model performance will continue to improve and reach the optimal value when $k_2 = 27$.

Parametric Analysis in QNC. Figure 3b shows the results for different q in the cross-modality feature centre computation which tested on the baseline. We speculate that mAP rises first and then falls because the high q value causes some sample features which differ from the probe image identity as neighbours, resulting in a centre shift. In particular,

$q = 0$ means that the original feature is directly adopted instead of the QNC feature.

MGK on single-modality ReID. In the single-modality ReID task, there is also the problem that local feature ranks better. Therefore, we try to apply the MGK method to two widely used single-modality datasets to verify the generalisation of our method. As shown in the Table 4, MGK can also improve some indicators of the single-modality dataset, but the effect is not as significant as that of the VI-ReID. We speculate that due to the different causes of the local feature ranking better problem, and most of the local features are available in the single-modality task.

Parametric Analysis in QCMK. When applying both components at the same time, the utilisation position of the QNC is very important. We tested the nearest-neighbor modality transformation for global and local features separately and both simultaneously as Table 5. Only performing a neighbour transformation on local feature blocks achieves better results. Since the transformed features have more advantages in the neighbors' calculation, however, completely replacing the original features leads to information loss. Global features dominate the nearest neighbor statistics, while local features are complementary. Therefore, modality transformation only on local features effectively narrows the modality gap with less information reduction, leading to better results.

Besides, the results show that QNC improves mINP in all situations. Because the transformed features in the same modality pull into the distance between samples with the same identity, this makes difficult samples more likely to have higher rankings.

We also experimented with the values of λ_1 and λ_2 as Figure 4. By the grid search method with a gradient of 0.025, when λ_1 and λ_2 are both 0.05, the best mAP and rank 1 can be obtained. Too high λ_1 or λ_2 will reduce the overall performance of the model. Therefore, we speculate that MGKE plays a leading role, and the other two only play a complementary role.

Validity of Different Divisions in PCB. Figure 5 is the testing results of QCMK effectiveness for different local feature block divisions (p) in PCB, where the gray dotted line is the baseline with k-reciprocal re-ranking. At any value of p, both $q = 1$ and $q = 2$ achieve better results than the baseline, especially on mAP and mINP, which demonstrates the robustness of our method.

Smaller q-values lead to better results, which may be due to the excessive q leading to additional negative samples. What's more, with smaller p in PCB, the more obvious performance is improved. We deduce that this is because fewer feature block division easily makes it difficult for the two different modalities to map to the same space, so it is more necessary to perform modality feature transformation with QNC to narrow the modality gap.

Our method is easier to achieve better results on mAP and mINP metrics. Adding local features enriches the nearest-neighbour representation, which ensures that the original positive samples that are not in the Jaccard distance metric are included in the calculation. However, we still use the original features for distance measurement, which results in that the newly added positive samples cannot be ranked higher than the original positive samples in the nearest-neighbour set, thus making it difficult to improve rank 1. However, adding these positive samples that were not originally involved in the calculation has a positive contribution to the significant improvement of mAP. Although the samples added by our method cannot surpass the original best-matched samples, modality transformation can optimise the similarity of the original lower-ranked positive samples, resulting in an improvement in mINP.

TABLE 4 MGK on single-modality ReID

	Market-1501			CUHK03		
Method	**r1**	**mAP**	**mINP**	**r1**	**mAP**	**mINP**
PCB	92.72	77.64	60.77	60.20	53.93	49.70
PCB*	92.87	**78.27**	63.15	64.11	57.49	55.23
PCB + MGK	**92.89**	78.13	**64.32**	**66.27**	**58.13**	**56.98**

Note: The PCB* obtains the ranking by the raw k-reciprocal encoding (KR). The bold value means the best value in the current indicator.

TABLE 5 Comparison with different QNC strategies

Methods		**r1**	**r10**	**r20**	**mAP**	**mINP**
QNC_g	F_l	64.54	91.12	95.34	67.07	58.85
	QNC_l	64.44	90.9	94.99	66.99	58.86
F_g	F_l	**73.17**	92.58	96.27	69.85	56.38
	QNC_l	72.65	**94.45**	**98.09**	**72.70**	**64.33**

Note: The bold value means the best value in the current indicator.

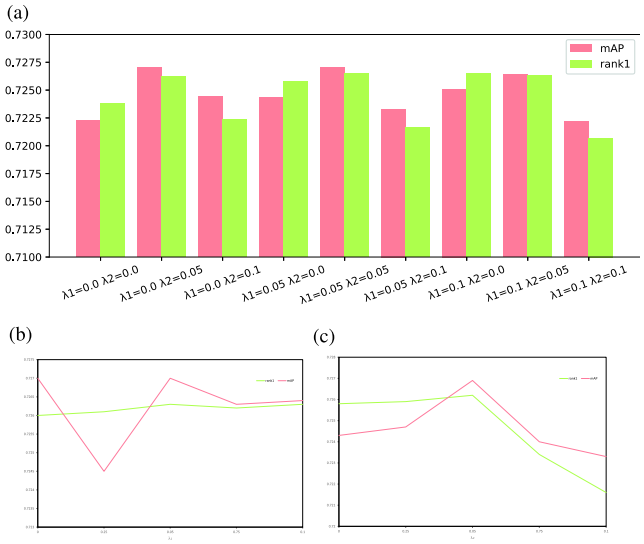

FIGURE 4 Different λ_1 and λ_2 in QCMK. (a) Grid search results of λ_1 and λ_2. (b) Test of λ_1 ($\lambda_2 = 0.05$). (c) Test of λ_2 ($\lambda_1 = 0.05$).

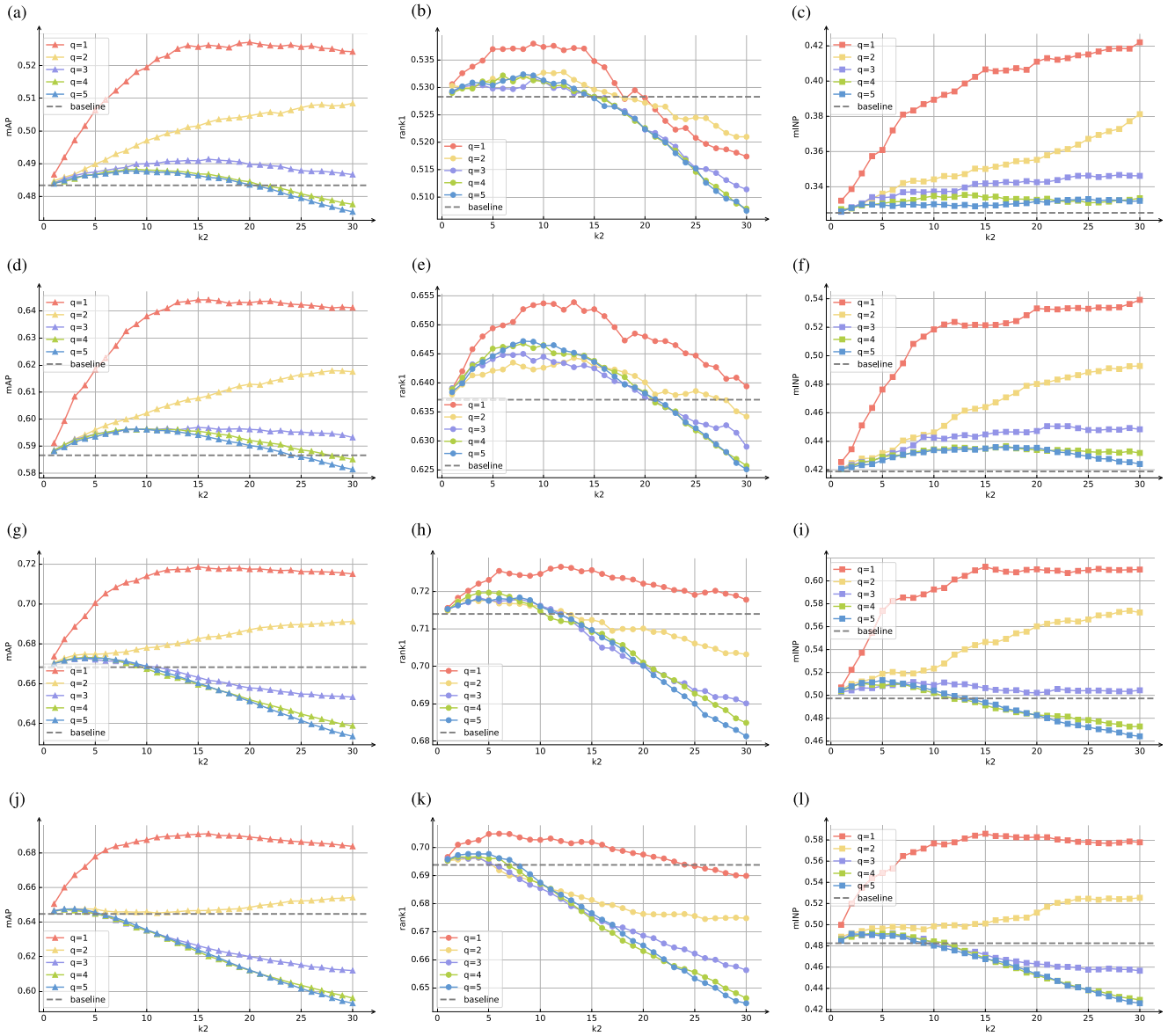

FIGURE 5 Different Number of Local Feature Divisions for SYSU-MM01. (a) mAP($p = 2$). (b) r1($p = 2$). (c) mINP($p = 2$). (d) mAP($p = 3$). (e) r1($p = 3$). (f) mINP($p = 3$). (g) mAP($p = 6$). (h) r1($p = 6$). (i) mINP($p = 6$). (j) mAP($p = 9$). (k) r1($p = 9$). (l) mINP($p = 9$).

5 | CONCLUSION

In this study, we propose an innovative re-ranking algorithm for cross-modality person re-identification, which utilises the local features and reduces the modality gap. Adding local reciprocal neighbours to the reciprocal neighbour set effectively supplements the differences between pedestrians with the same identity caused by the changes of the viewing angle, illumination and modality. Modal transformation improves the ranking of positive samples and enhances the robustness of the algorithm. Extensive experiments demonstrate that our proposed method brings significant improvements over the baseline. At present, our method mainly acts on mAP and mINP, but the improvement of rank one is weak. In the future, our research can focus more on how to improve the rank one in VI-ReID.

ACKNOWLEDGEMENT

This paper is partially supported by the National Natural Science Foundation of China (Serial No. 61976158 and No. 62076182), and the Jiangxi "Double Thousand Plan".

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

[SYSU-MM01]The data that support the findings of this study are openly available at http://isee.sysu.edu.cn/project/RGBIRReID.htm. [RegDB]The authors confirm that the data supporting the findings of this study are available at http://dm.dongguk.edu/link.html.

ORCID

Hongyun Zhang https://orcid.org/0000-0001-9781-5078

REFERENCES

1. Chen, T., et al.: Abd-net: attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8351–8361 (2019)
2. Bryan, X., et al.: Second-order Non-local Attention Networks for Person Re-identification. IEEE (2019)
3. Zhu, Z., et al.: Aware loss with angular regularization for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13114–13121 (2020)
4. Chen, D., et al.: Improving deep visual representation for person re-identification by global and local image-language association. In: Proceedings of the European Conference on Computer Vision, pp. 54–70. ECCV (2018)
5. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision, pp. 686–701. ECCV (2018)
6. Haque, A., Alahi, A., Fei-Fei, L.: Recurrent attention models for depth-based person identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1229–1238 (2016)
7. Karianakis, N., et al.: Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In: Proceedings of the European Conference on Computer Vision, pp. 715–733. ECCV (2018)
8. Wu, A., et al.: RGB-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5380–5389 (2017)
9. Dai, P., et al.: Cross-modality person re-identification with generative adversarial training. In: IJCAI, vol. 1, p. 6 (2018)
10. Ye, M., et al.: Visible thermal person re-identification via dual-constrained top-ranking. In: IJCAI, vol. 1, p. 2 (2018)
11. Hao, Y., et al.: Dual-alignment feature embedding for cross-modality person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 57–65 (2019)
12. 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
13. Mao, X., Li, Q., Xie, H.: Aligngan: Learning to Align Cross-Domain Images with Conditional Generative Adversarial Networks (2017). arXiv preprint arXiv:1707.01400
14. 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)
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. Wang, G.-A., et al.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12144–12151 (2020)
17. Li, W., et al.: Common-near-neighbor analysis for person re-identification. In: 2012 19th IEEE International Conference on Image Processing, pp. 1621–1624. IEEE (2012)
18. Garcia, J., et al.: Person re-identification ranking optimisation by discriminant context information analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1305–1313 (2015)
19. Ye, M., et al.: Coupled-view based ranking optimization for person re-identification. In: International Conference on Multimedia Modeling, pp. 105–117. Springer (2015)
20. Zhong, Z., et al.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)
21. Jia, M., et al.: A Similarity Inference Metric for Rgb-Infrared Cross-Modality Person Re-identification (2020). arXiv preprint arXiv: 2007.01504
22. Ye, M., et al.: Hierarchical discriminative learning for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
23. Hao, Y., et al.: HSME: hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8385–8392 (2019)
24. Ye, M., Lan, X., Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 347–355 (2019)
25. Li, D., et al.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 4610–4617 (2020)
26. Lu, Y., et al.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379–13389 (2020)
27. Ye, M., et al.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: European Conference on Computer Vision, pp. 229–247. Springer (2020)
28. Zhao, Z., et al.: Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3520–3528 (2021)
29. 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)
30. Tian, X., et al.: Farewell to mutual information: variational distillation for cross-modal person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1522–1531 (2021)
31. 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)
32. Ye, M., et al.: Deep learning for person re-identification: a survey and outlook. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
33. Liu, H., Tan, X., Zhou, X.: Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. IEEE Trans. Multimed. 23, 4414–4425 (2020). https://doi.org/10.1109/tmm.2020.3042080

How to cite this article: Wang, Y., et al.: Multi-granularity re-ranking for visible-infrared person re-identification. CAAI Trans. Intell. Technol. 1–10 (2023). https://doi.org/10.1049/cit2.12182