Image and Vision Computing 150 (2024) 105201

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Visible thermal person re-identification via multi-branch modality residual complementary learning

Long Chen [a,b], Rui Sun [a,b,c,*], Yiheng Yu [a,b], Yun Du [a,b], Xudong Zhang [a,c]

[a] School of Computer Science and Information Engineering, Hefei University of Technology, No. 485 Danxia Road, Hefei 230009, China
[b] Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230009, China
[c] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education of the Peoples Republic of China, Hefei 230009, China

ARTICLE INFO

Keywords:
Visible thermal person re-identification
Modality residual complementary learning
Multi-branch feature learning
Multi-branch constraint loss

ABSTRACT

Visible-thermal Person Re-identification (VT Re-ID) is a challenging task in all-weather surveillance system. Existing methods concentrate on extracting the modality-shared features, ignoring the discriminative inter-modality complementary features. To tackle this issue, we propose a multi-branch modality residual complementary learning method which consists of the modality residual complementary learning (MRCL) module and the multi-branch feature learning (MBFL) module. The MRCL module can be easily integrated into existing CNN baselines and drive the network to focus on both intra-modality and inter-modality information. On one hand, we adopt the basic two-stream network to obtain the intra-modality features, on the other hand, we capture the inter-modality complementary features within the residual image obtained by cross-modality correlation saliency erasing operation. To handle the intra-modality variations, we employ the MBFL module to capture local spatial features and local channel features, then integrate them with global features to achieve part-to-part and high-level semantic information matching. Finally, the discriminability and robustness of the ultimate representations are enhanced by multi-branch constraint loss learning. Extensive experiments on RegDB and SYSU-MM01 datasets demonstrate the superiority of our proposed method compared with state-of-the-art methods.

1. Introduction

Person re-identification (ReID) [1,2,32–35] is an image retrieval task that that involves finding a person across multiple discontinuous camera views. It can be applied to the tracking and anomaly detection of person in the industrial and public security fields. Nowadays a large number of works are concentrated on single-modality person Re-ID, which has shown promising performance in both video and image domains. However, the applicability of single visible modality Re-ID is limited under low lighting conditions (e.g. during the night). In this instance, dual-mode cameras which work in visible mode during the day and thermal mode in the night have recently become widespread in practical video surveillance systems. Photos in a 24-h intelligent monitoring system are from many modalities. For example, the probe images might be captured by visible cameras during the day, while gallery images could be acquired by thermal cameras at night. Therefore, in this paper, we focus on the more challenging but practical cross-modality visible thermal person re-identification (VT Re-ID).

As shown in Fig. 1, the cross-modality discrepancy caused by differing imaging principles of visible and thermal cameras, as well as the intra-modality variations suffered from camera viewpoint changes, occlusion, and various pedestrian poses, are two main issues for VT Re-ID. To overcome these problems, some VT Re-ID methods [3–6,36–38] have been proposed to learn modality-shared features, and the advantages of two-stream networks have been explored.

Given the lack of color information in the thermal image compared to the visible image, we need to explore more additional information when matching the thermal image with the visible image for VT Re-ID. However, existing methods do not take full advantage of the underlying modality information. To be specific, existing methods utilize two-stream network to extract the modality-specific features for each modality independently and then learn the shareable features between them. The modality-shared features focus on the same local salient part and ignore the latent information which can help enhance performance. Nonetheless, cross-modality correlation is the key to explore the latent complementary information. Besides, previous approaches [7–10,22,23,39,40] only focused on global features which is lack of discriminative information with various granularities. However, due to

* Corresponding author at: School of Computer Science and Information Engineering, Hefei University of Technology, No. 485 Danxia Road, Hefei 230009, China.
E-mail address: sunrui@hfut.edu.cn (R. Sun).

https://doi.org/10.1016/j.imavis.2024.105201
Received 5 March 2024; Received in revised form 4 July 2024; Accepted 31 July 2024
Available online 2 August 2024

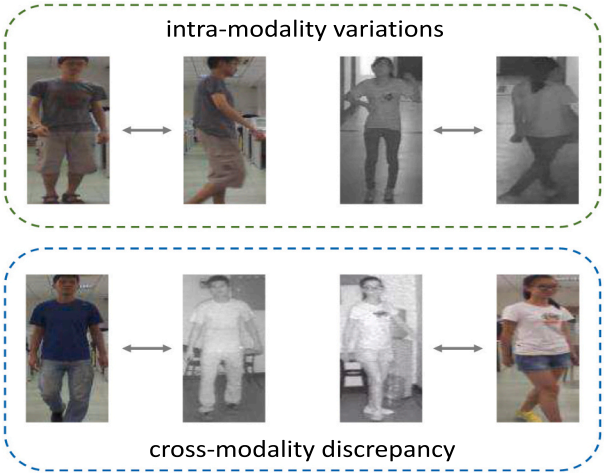

Fig. 1. Illustration of the key challenges in the VT Re-ID. The camera viewpoint changes, occlusion, and various pedestrian poses exacerbates the cross-modality discrepancy and intra-modality variations. Existing methods focus on modality-shared features but ignore the discriminative inter-modality complementary features.

the occlusion and misalignment, the global features may lose critical local information such as the pedestrian's body part, thereby affecting performance.

To address the above limitations, we propose modality residual complementary learning (MRCL) module with a cross-modality correlation saliency erasing operation (CM-CSEO) for VT Re-ID. The target of our MRCL module is to acquire the neglected inter-modality information for resolving the feature-level modality discrepancy. Specifically, the CM-CSEO aims at erasing the most dominant area of the person image based on the cross-modality correlation between pedestrian image and the feature of another modality for obtaining the residual image. Then we feed the remaining residual images to another two-stream network to capture complementary inter-modality features containing latent inter-modality information. We can get the final discriminative feature representation by fusing the intra-modality features with their complementary inter-modality features. Another advantage of the MRCL is that it can be easily inserted into any existing network to learn the neglected complementary information.

In addition, we design the multi-branch feature learning (MBFL) module to capture local spatial features and local channel features, then integrate them with global features to obtain more discriminative and robust ultimate representation for handling intra-modality variations. Local spatial features corresponding to individual body parts, as opposed to global features, can achieve part-to-part matching. Different from spatial partition, channel partition is conducted along the channel dimension. Besides, local channel features do not focus on individual body parts and instead corresponding to high-level semantic information such as hairstyle, body shape, etc. Fig. 2 illustrated the working block diagram of our approach. The main contributions can be summarized as follows:

1. We propose a modality residual complementary learning module to incorporate intra-modality feature and inter-modality feature which is often neglected by the two-stream network, thereby enhancing the discriminability and diversity of learned representation for VT Re-ID.
2. We introduce a multi-branch feature learning module to combine local spatial and channel features with the global features, which overcomes intra-modality variations by part-to-part matching and high-level semantic concept-to-concept matching.
3. To exploit the different attributes of both global features, local spatial features, local channel features and final fused features effectively, we design the multi-branch constraint loss including the global loss, local loss and hetero-center cross-modality constraint loss.
4. Extensive experiments on two publicly cross-modality datasets, SYSU-MM01 and RegDB demonstrate that our approach achieves promising performance. Especially when MRCL is inserted into other state-of-art models, it can effectively improve the performance.

2. Related work

2.1. Single-modality ReID

With the rapid development of deep learning, single-modality person Re-ID research has gotten remarkable advances. Nowadays, there are two types of deep learning-based person re-identification methods: feature learning and metric learning. Metric learning aims to optimize the learning of discriminative features by network models by designing different loss function. Feature learning extracts more robust features of pedestrians by introducing local feature learning [7] or using attention mechanisms to focus on key information about body parts [8]. In addition to this, several works enhanced the final feature representation by combining global and local features of pedestrians [9]. Generative adversarial network (GAN) is widely used for person re-identification tasks due to the good performance in generating images and learning features. To relieve the expensive costs of annotating new training images, Wei et al. [10] proposed a Person Transfer Generative Adversarial Network (PTGAN) to bridge the domain gap. More recently, approaches based on graph convolutional networks have also emerged to learn more discriminative and robust features by modeling graph relationships on

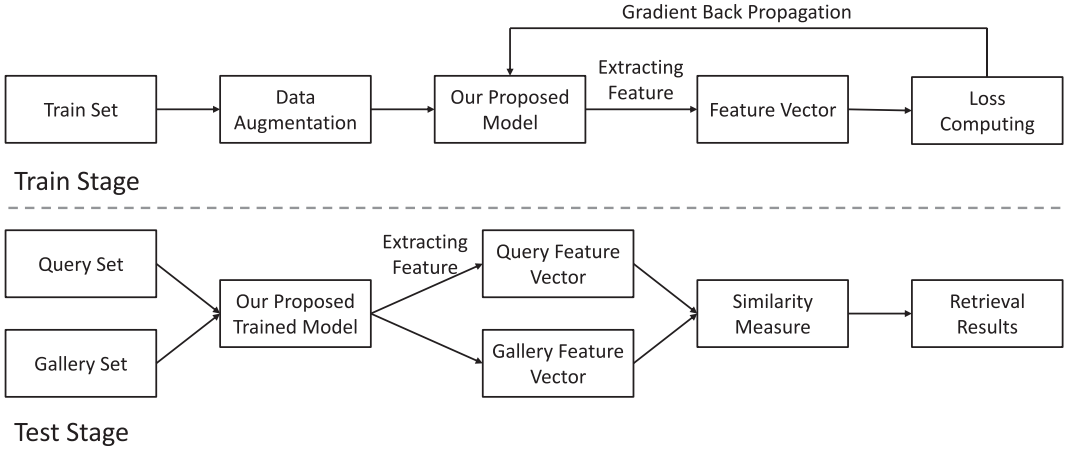

Fig. 2. The working block diagram of proposed multi-branch modality residual complementary learning method.

pedestrian images. Yang et al. [11] proposed a spatio-temporal graphical convolutional network (STGCN) to extract robust spatio-temporal information that complementary to appearance information.

2.2. Visible-thermal ReID

The VT Re-ID problem was first discussed by Wu et al. [12], they built the SYSU-MM01 dataset and presented the zero-padding network which was a one-stream network and learned modality sharable information adaptively. Ye et al. [13] proposed a two-stream CNN network (TONE) with hierarchical cross-modality metric learning to supervise the modality-specific features. Li et al. [15] introduced a supplemental X modality to transform visible-thermal dual-mode problem to X-visible-thermal three-mode problem for reducing cross-modality discrepancy. In addition to the research on both feature learning and metric learning, many scholars have also leveraged GAN to solve the VT Re-ID problem. Dai et al. [17] first proposed the cm-GAN based on a generative adversarial network involving a generator and a discriminator. The generator aimed to extract features from two different modalities, and then the discriminator was responsible for distinguishing whether the features were from thermal modality or visible modality. Wang et al. [19] also applied an alignment adversarial generative networks (AlignGAN) to employ generators to conduct joint alignment from pixels and features respectively.

2.3. Adversarial erasing learning

The core idea of adversarial erasing learning is that we can improve the performance by employing the erased image. Initially, the adversarial erasing operation was exploited extensively in data augmentation. Zhong et al. [20] randomly elected a rectangle region in the image and erased its pixels with random values. The robustness of the model to occlusions can be enhanced by using the erased images. Many researchers have extended adversarial erasing learning to other computer vision areas recently. Liu et al. [21] introduced a cross-modal attention-guided Erasing (CAGE) to accomplish cross-modal alignment of the textual and visual domain. Similar to [20], the CAGE constructed difficult samples by discarding the most dominating features from textual or visual domains, and then attempted to uncover complementary textual-visual correspondences by training the difficult samples. Different from them, we try to exploit erasure learning to mine discriminative information that is neglected by the network.

3. Approach

In this section, we will introduce the framework of our proposed multi-branch modality residual complementary learning method, as illustrated in Fig. 3. Our proposed method is mainly composed of three elements: (1) the modality residual complementary learning (MRCL) module, (2) the multi-branch feature learning (MBFL) module, (3) the multi-branch constraint loss which combines global loss, local loss and hetero-center cross-modality constraint loss.

3.1. The modality residual complementary learning module

In this subsection, we will present the structure of the MRCL module and how it addresses the cross-modality discrepancy in VT Re-ID. The MRCL module is made up of two components: a basic two-stream network for extracting prominent intra-modality features, and a complementary network with the cross-modality correlation saliency erasing operation for obtaining the inter-modality features.

1) *The basic two-stream*

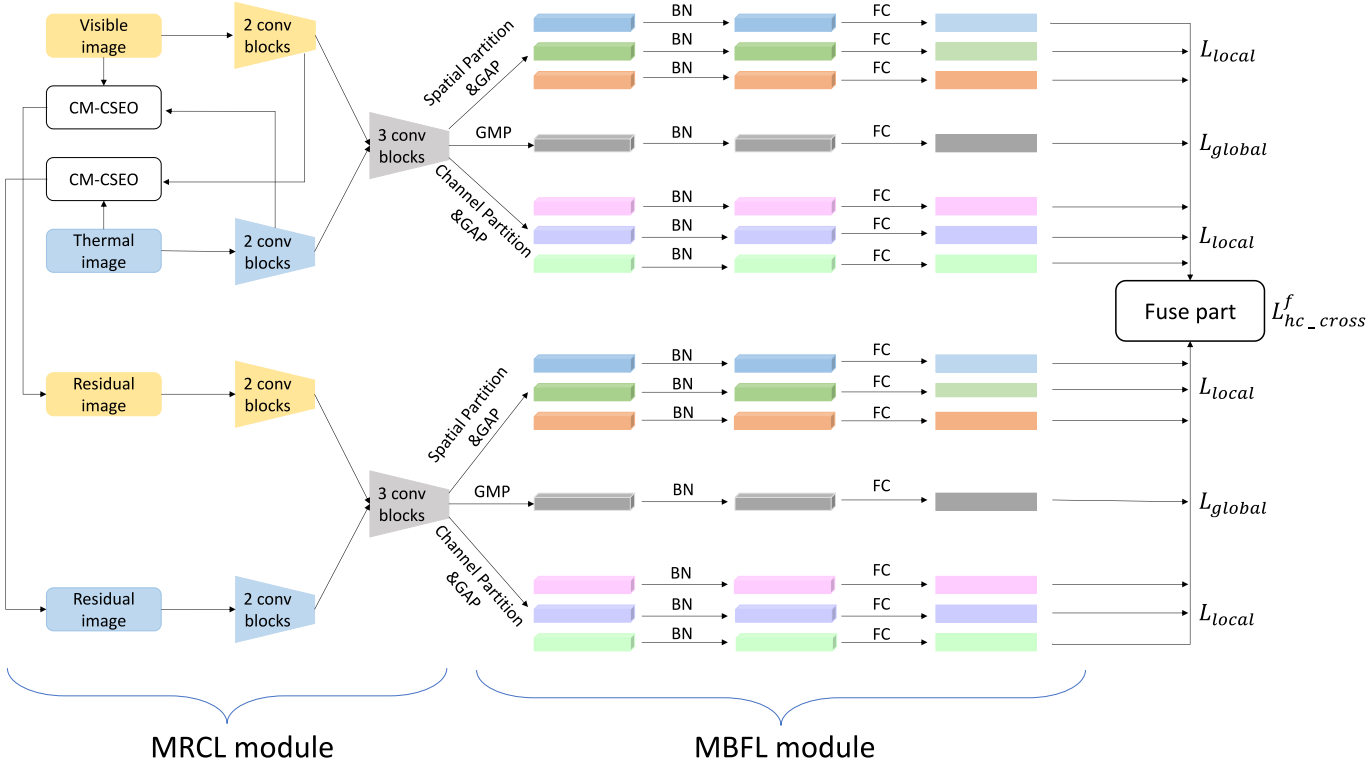

Fig. 3. The framework of our proposed multi-branch modality residual complementary learning method. GAP: Global Average Pooling, GMP: Global Max Pooling, BN: Batch Normalization, FC: Full Connection. Our framework mainly contains two components: MRCL module and MBFL module. The MRCL module includes a basic two-stream network for extracting prominent intra-modality features and a complementary network with the same structure as the basic two-stream network, but with the cross-modality correlation saliency erasing operation to obtain the inter-modality features. The MRFL contains the batch normalization neck and feature extraction part to obtain global feature, spatial local feature and channel local feature.

In VT Re-ID, the two-stream network comprised of feature extractor and feature embedding is a common approach. Because of the outstanding performance and the simple architecture, we adopt the ResNet50 as the backbone. We split the ResNet50 into two parts. The first two conv blocks in ResNet50 form a two-stream feature extractor with independent parameters for learning low-level modality-specific features from two different modalities. The following three conv blocks make up the sharable feature embedding, and their input is a 3D feature map generated by the feature extractor. These three conv blocks learn high-level features shared by different modalities that include enough spatial structure information and maps them to the common feature space. For simplicity in presentation, we utilize $F_v()$ to describe the visible-stream network and define $F_t()$ as the representation for the thermal-stream network. Given two different types of input, visible images I_v and thermal images I_t, the features captured by the two-stream network can be respectively denoted as.

$$V = F_V(I_V)\ T = F_t(I_t) \tag{1}$$

2) *The cross-modality correlation saliency erasing operation*

We apply the other two-stream network, the complementary network as illustrated in Fig. 3, to get complementary inter- modality features. The input to the complementary network determines the ability to obtain high-quality complementary inter-modality features. As a consequence, we design a cross-modality correlation saliency erasing operation (CM-CSEO) to help us erase the most prominent areas of highest correlation between the images and features from different modalities. The CM-CSEO consists of three steps: cross-modality correlation calculation, erasing region binarization, and significant area erasing operation. As previously mentioned, the inputs to the two-stream network are the visible image I_v and the thermal image I_t, and the modality-specific features obtained by the basic two-stream network are defined as S_v and respectively S_t.

a) *Cross-modality correlation calculation*

In general, we need to consider the image and the feature at every spatial location when calculating the correlation, so we specify the spatial positions (i,j) of the image and the modality-specific feature as $I(i,j)$ and $S(i,j)$ separately. The purpose of the cross-modality correlation calculation is to attain the semantic correlation between the image I and modality-specific feature S by calculating the similarity between them. We utilize dot product similarity which is easier to implement than cosine and Euclidean distance similarity in modern deep learning platforms. The corresponding correlation between the image I and modality-specific feature S is shown in the following Eq. (2).

$$C(i,j) = S(i,j)^T I(i,j)\ (1 \le i \le H, 1 \le j \le W) \tag{2}$$

The similarity value at each position in C reflect the degree of correlation between the image I and modality-specific features S. Darker colors in the correlation map represent higher correlations, as illustrated in Fig. 4.

b) *Erasing region binarization*

In most other fields, erasing operation is performed by applying a threshold to the corresponding correlation and obtaining a binarization mask. However, it frequently results in noncontinuous regions. As we all know, erasing the feature units discontinuously can be effective for fully connected layers, but it is less effective for convolutional layers, where features are correlated spatially. Erasing the correlation part discontinuously does not validly remove semantic information, as nearby features contain closely related information. Instead, erasing continuous regions can remove semantic information and consequently enforcing remaining units to learning potential feature. Our proposed MRCL module focus on convolution layer, so employing erasing region binarization is a better choice than the soft erasing mask.

The key to getting the region binarization mask is to fix and compare the value of the correlation of each region, as indicated in the Fig. 4. We leverage search kernel for performing region binarization mask to solve this problem. The search kernel is analogous to the convolutional kernel in a convolutional neural network. Formally, the size of the search kernel is established as $h_k \times w_k$, the horizontal stride is set as s_h, and the vertical stride is set as s_v. By sliding the search kernel according to the Eq. (3) below, we can figure out how many relevant regions there are.

$$N_{re} = \left(\lfloor\frac{H - h_k}{s_v}\rfloor + 1\right) \times \left(\lfloor\frac{W - w_k}{s_h}\rfloor + 1\right) \tag{3}$$

The value of each relevant region is considered to be the sum of all relevant correlation values of the item in the search kernel. Finally, we choose the area of the relevant region that has the highest value to be erased. We may get the erasing region binarization mask by setting the selected erased part to 0 and the other locations to 1. In this case, we can

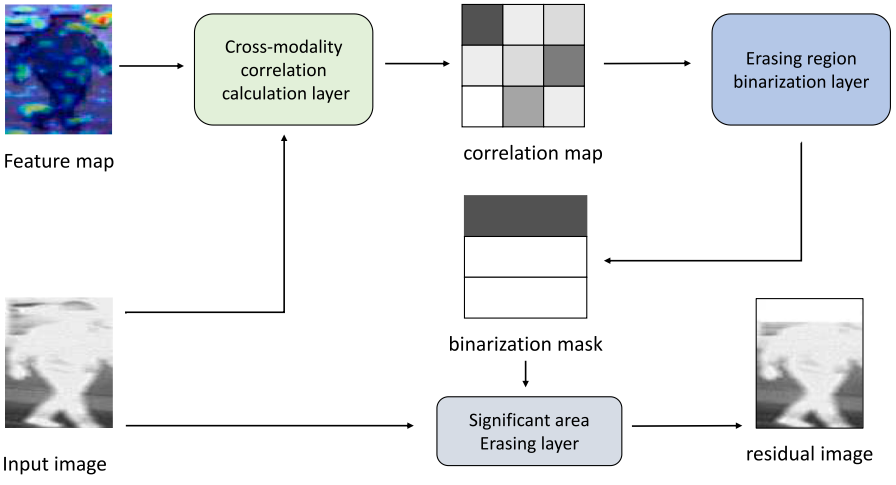

Fig. 4. Illustration of the cross-modality correlation saliency erasing operation. The CM-CSEO consists of three steps: cross-modality correlation calculation, erasing region binarization, and significant area erasing operation.

obtain the erasing region binarization mask $B \in B^{H\times W}$.

c) *Significant area erasing operation*

By dot-multiplying the binary region mask with the original image to get the residual image with no significant features, as demonstrated in Eq. (4).

$$X_r = X \odot B \tag{4}$$

3.2. The multi-branch feature learning module

In this subsection, we will introduce the MBFL module, a feature extraction part including global feature, spatial local feature and channel local feature extraction.

Most existing Re-ID works combined ID loss with triplet loss for training the models. We implemented a common module, batch normalization neck (BNNeck) which has been first proposed in [25], in both the global and local feature networks. After a global pooling layer, the triplet loss is employed to optimize feature vectors roughly distributed in the Euclidean space, while the ID loss is assigned to optimize feature vectors approximately distributed in the hypersphere space after a batch normalization layer and a fully connected layer.

We exploit the feature maps of pedestrian captured from the basic two-stream network and the complementary two-stream network as input to the multi-branch feature learning module, as shown in Fig. 3. A global feature extraction part, a spatial local feature extraction part, and a channel local feature extraction part are the three individual components that make up the feature extraction part. A single-branch network with a global max pooling layer (GMP) and a BNNeck makes up the global feature extraction component.

The spatial local feature extraction part is a three-branch network with a global average pooling layer (GAP) and a BNNeck on each branch. The feature maps are divided horizontally into three equal parts and served as inputs for the three-branch networks. The spatial local feature corresponding to individual body part can achieve part-to-part matching. The channel local feature extraction part is also a three-branch network, just like the spatial local feature extraction part. Relying on the number of channels in the feature map, we partition it into three equal parts at the channel level which are used as inputs for the three branches. The channel local features do not focus on individual body parts and instead corresponding to high-level semantic information such as hairstyle, body shape, etc. The global feature extraction network differs from the two local feature extraction networks in the pooling approach and the feature dimensionality.

3.3. The multi-branch constraint loss function

In this subsection, we introduce the multi-branch constraint loss consisting of the global loss, intra-modality constraint loss and the hetero-center cross-modality constraint loss. It not only takes advantage of both local and global features, but also focuses on both cross-modality discrepancy and intra-modality variations to improve the performance.

1) *Intra-modality constraint loss*

We design the intra-modality constraint loss to enable the feature representation robust to intra-modality variations in both visible and thermal modalities. For visible to thermal scenarios, we set visible image as anchor v_a and thermal image as positive t_p and negative objects t_n. Anchor t_a is thermal image and positive v_p and negative v_n objects are visible image for visible to thermal scenarios. By keeping the distance between the anchor sample and its negative sample larger than set margin ρ_1, the intra-modality constraint loss is defined by:

$$L_{intra} = \sum_{y_a\neq y_n}\left[\rho_1 - min\|v_a - v_n\|_2\right]_+ + \sum_{y_a\neq y_n}\left[\rho_1 - min\|t_a - t_n\|_2\right]_+ \tag{5}$$

2) *Hetero-center cross-modality constraint loss*

The cross-modality constraint loss hardly works if there are harmful triplets formed by outliers. Therefore, we utilize a hetero-center cross-modality constraint loss that takes the center of each class of individual data as the object for computation. In this manner, we can transfer the computation of the anchor point with all other objectives to the computation of center the anchor point center with all other centers. We have to first compute the data center for the features of every identity in both modalities.

$$c_v^i = \frac{1}{k}\sum_{j=1}^{k} v_j^i \quad c_t^i = \frac{1}{k}\sum_{j=1}^{k} t_j^i \tag{6}$$

where v_j^i and t_j^i correspond to the j^{th} visible image feature and thermal image feature of the i^{th} person from each batch respectively.

Based on the same PK sampling method as traditional triplet loss, we select P person identities at random, with each identity containing K images for each mini-batch. Accordingly, there are P visible image centers $\{c_v^i | i = 1,\ldots,P\}$ and P thermal image centers $\{c_t^i | i = 1,\ldots,P\}$. The computation is shown as the following Eq. (7):

$$\begin{aligned} L_{hc\_cross} = &\sum_{i=1}^{p}\left[\rho_2 + \left\|c_v^i - c_t^i\right\|_2 - \min_{j\neq i}\left\|c_v^i - c_n^j\right\|_2\right]_+ \\ &+ \sum_{i=1}^{p}\left[\rho_2 + \left\|c_t^i - c_v^i\right\|_2 - \min_{j\neq i}\left\|c_t^i - c_n^i\right\|_2\right]_+ \end{aligned} \tag{7}$$

where ρ_2 denote the predefined margin.

3) *Global loss*

As with local features, we also need to apply ID loss to enhance the robustness of global features. The global loss is comprised of three components, the hetero-center cross-modality constraint loss, intra-modality constraint loss and id loss, which is represented by

$$L_{global} = L_{hc\_cross}^g + L_{intra} + \sum_{i=1}^{M} L_{id}^i \tag{8}$$

where M represents the number of the global features and id loss is represented by

$$L_{id} = -\sum_{i=1}^{N} q_i log(p_i) \tag{9}$$

4) *The multi-branch constraint loss*

In order to take full advantage of local feature, global feature and final fused feature, we design the multi-branch constraint loss consisted of local loss, global loss and hetero-center cross-modality constraint loss. We employ id loss for each local feature and global loss for the global feature, but only hetero-center cross-modality constraint loss for the final fused feature. The multi-branch constraint loss is given as

$$\begin{aligned} L_{mb\_c} &= L_{hc\_cross}^f + L_{global} + L_{local} \\ &= L_{hc\_cross}^f + L_{hc\_cross}^g + L_{intra} + \sum_{i=1}^{M} L_{id}^i + \sum_{i=1}^{N} L_{id}^i \end{aligned} \tag{10}$$

where M represents the number of global features and N describes the number of local features.

4. Experiments

In this section, we employed two cross-modality re-identification datasets, RegDB [27] and SYSU-MM01 [12], to evaluate the

effectiveness of our proposed method. In addition, extensive ablation experiments were carried out to verify the role of each component of the network as well as the influence of various parameters.

4.1. Datasets

SYSU-MM01: It is the only large-scale cross-modality re-identification dataset with 287,628 visible images and 15,792 thermal images collected by two thermal cameras (cam3, 6) and four visible cameras (cam1, 2, 4, 5). The entire dataset consists of 491 identities, and each identity contains at least one visible image and thermal image. The dataset was split into two parts: a training set of 395 individuals with 22,258 visible and 11,909 thermal images, and a test set of 3803 visible and 301 thermal images of the remaining 95 individuals. SYSU-MM01 can fulfill different modes of experimental situations, namely the indoor-search mode and the all-search mode. We follow existing methods to conduct 10 trials of gallery set selection in the single-shot setting [3], and then evaluate the average retrieval performance.

RegDB: The image was captured using a dual-camera system with a visible camera and a thermal camera. The entire dataset consists of 412 persons, each corresponding to ten visible images and ten thermal images, for a total of 8240 images. In order to comply with the conditions of the evaluation protocol in [3,13], we split it into two halves, one for training and the other for testing. During the test process, we set all of the images in the gallery to the same modality, while the images in the probe were set to a different modality.

4.2. Evaluation metrics

We adopt three different evaluation metrics to measure the performance, namely the cumulative matching characteristic (CMC), the mean average precision (mAP) and the mean inverse negative penalty (mINP) [16].

$$mINP = \frac{1}{n}\sum_{i}(1 - NP_i) = \frac{1}{n}\sum_{i}\frac{|G_i|}{R_i^{hard}} \tag{11}$$

where R_i^{hard} denotes the ranked position of the most difficult-to-match samples, $|G_i|$ denotes the ranked position of table denotes the total number of correct matches for query *i*. Note that all the person features are L2 normalized for testing and metric learning conducted by our proposed multi-branch constraint loss.

4.3. Implementation details

The experiment was implemented in the Pytorch framework and deployed on NVIDIA GeForce 3090 GPU. We adopt the ResNet50 as the backbone and initialize the network parameters on ImageNet. Random flip and random crop are two widely used data augmentation techniques in content-based image retrieval methods [41] and image registration [42]. The person images are resized to 288*144 before being randomly flipped and cropped to their original size. We adopt a stochastic gradient descent (SGD) optimizer for optimization with a momentum parameter of 0.9. The initial value of the learning rate was set to 0.1. To improve the network's performance during the training phase, we utilized a warm-up learning rate strategy. The learning rate of epoch 0 is 0.01, and every epoch after that is increased by 0.01. From epoch 9 to epoch 29, the learning rate is 0.1, then decreases to 0.01 at epoch 30, then drops again at epoch 50, which is set at 0.001. For different datasets, we leverage different PK sampling strategies. We set P to 8 and K to 4 for RegDB while P to 6 and K to 8 for SYSU-MM01. Among the hetero-center cross-modality constraint loss and the intra-modality constraint loss, we set the default value of margin ρ_1 to 0.3 and ρ_2 to 0.5. For each local feature, the final dimension d of each local feature is set to 256.

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

We obtain experimental results on SYSU-MM01 and RegDB using the method proposed in this paper and compare them with the state-of-the-art VT Re-ID methods.

The experiment results on the SYSU-MM01 and RegDB datasets are shown in Table 1 and Table 2, respectively. We have applied the five CMC metrics of rank-1, rank-10, rank20, mAP and mINP to measure performance. The '-'in the Table 1 and Table 2 means that the corresponding results are not displayed in the original paper.

The experiments on SYSU-MM01 have two modes, all–search mode and indoor-search mode. The results illustrated in Table 1 demonstrate that our proposed method can achieve the best performance on SYSU-MM01 under two different modes. We can observe that compared to the baseline method, AGW [16], each metric has been improved by a substantial margin. Especially in the two crucial criteria Rank1 and mAP under the more challenging all-search mode, we achieve 61.75% and 59.51%. This indicates that our method can effectively extract ignored multi-granularity complementary features and enhance the richness of pedestrian information. When we inserted our MRCL in CAJ [4] without the channel-augmented joint learning strategy (only +J), our proposed method gets 68.75% Rank1, 65.65% mAP and 52.78% mINP on all-search mode, outperforming the current state-of-the-art method PMT [31]. The reason be attributed that the proposed MRCL can capture more identity-related complementary knowledge across modalities.

The experiments based on the RegDB dataset have two query settings, visible to thermal and thermal to visible. From Table 2, we observe that our proposed method can achieve 91.26% Rank1 / 88.79% mAP / 81.80% mINP and 89.35% Rank1 / 86.85% mAP / 79.96% mINP, which performs better than any of the existing methods under the two different query settings. This shows previous methods neglect discriminative features in the harsh thermal environment. Compared to AGW [16], our method has achieved a huge improvement in the metric mINP, from 50.19% to 81.80%. This can be attributed to the fact that, the thermal image contains less information, and our method can mine more effective detailed features by MRCL, MBFL and multi-branch constrains loss function. Performance also improves after inserting the MRCL in CAJ. In general, the experimental results on RegDB illustrate the effectiveness of our proposed method in a variety of query settings.

4.5. Ablation study

In this subsection, we have done comprehensive ablation experiments on SYSU-MM01 and RegDB to evaluate the effectiveness of each component of the network structure, which mainly includes the MRCL module, the MBFL module and the multi-branch constraint loss. All experiments in this subsection are carried out in all-search modes of SYSU-MM01 and visible to thermal query setting of RegDB. The baseline model is a two-stream backbone network based on ResNet-50. The first two stages are set as the modality-specific modules with independent parameters to learn the modality-specific feature. The remaining three stages are set as the modality-shared module with shared parameters to learn the modality-sharable feature. Apart from this, the baseline model only has global feature.

1) *The effectiveness of the CM-CSEO in the MRCL module*

In this subsection, we first discuss where to perform cross-modality correlation saliency erasing operation (CM-CSEO) on the network in order to obtain the best performing network structure. Moreover, we compare the network with the CM-CSEO to the basic two-stream network which only extracts the intra-modality feature. We also explore which fusion mechanism is better, summation(sum) or concatenation(cat).

The residual complementary learning module needs to satisfy two requirements: Firstly, we need to calculate the cross-modality

Table 1
Comparison to the state-of-the-art methods on the SYSU-MM01 dataset.

Method	Reference	SYSU-MM01									
		All-Search					Indoor-Search				
		R1	R10	R20	mAP	mINP	R1	R10	R20	mAP	mINP
Zero-Pad [12]	ICCV17	14.80	54.12	71.33	15.95	–	20.58	68.38	85.79	26.92	–
HCML [13]	AAAI18	14.32	53.16	77.95	23.12	–	24.52	73.25	86.73	30.08	–
cmGAN [17]	IJCAI18	26.97	67.51	80.56	27.80	–	31.63	77.23	89.18	42.19	–
D^2RL [18]	CVPR19	28.90	70.60	82.40	29.20	–	–	–	–	–	–
AlignGAN [19]	ICCV19	42.40	85.00	93.70	40.70	–	45.90	87.60	94.40	54.30	–
XIV [15]	AAAI20	49.92	89.79	95.96	50.73	–	–	–	–	–	–
AGW [16]	TPAMI21	47.50	–	–	47.65	35.30	54.17	–	–	62.97	59.23
HAT [30]	TIFS21	55.29	92.14	97.36	53.89	–	62.10	95.75	99.20	69.37	–
DML [24]	TCSVT22	58.40	91.20	96.90	56.10	–	62.40	95.20	98.70	69.50	–
MAUM [14]	CVPR22	61.60	–	–	60.00	–	67.10	–	–	73.60	–
SPOT [28]	TIP22	65.34	92.73	97.04	65.25	48.86	69.42	96.22	99.12	74.63	70.48
GALNet [29]	IVC23	58.76	93.17	97.63	57.67	–	61.45	95.07	98.41	68.53	–
CMTR [26]	TMM23	65.45	94.47	98.16	62.90	–	71.46	97.16	99.22	76.67	–
PMT [31]	AAAI23	67.53	95.36	98.64	64.98	51.86	71.66	96.73	99.25	76.52	72.74
Ours	–	61.75	93.10	97.35	59.51	42.35	63.50	94.85	97.76	69.52	65.12
Ours+J [4]	–	**68.57**	**95.42**	**98.23.**	**65.65**	**52.78**	**74.42**	**96.24**	**99.41**	**78.55**	**75.06**

Table 2
Comparison to the state-of-the art methods on the RegDB dataset.

Method	Venue	RegDB									
		Visible to Thermal					Thermal to Visible				
		R1	R10	R20	mAP	mINP	R1	R10	R20	mAP	mINP
Zero-Pad [12]	ICCV17	17.75	34.21	44.35	18.90	–	16.63	34.68	44.25	17.82	–
HCML [13]	AAAI18	24.44	47.53	56.78	20.80	–	21.70	45.02	55.58	22.24	–
cmGAN [17]	IJCAI18										
D^2RL [18]	CVPR19	43.40	66.10	76.30	44.10	–	–	–	–	–	–
AlignGAN [19]	ICCV19	57.90	–	–	53.60	–	56.30	–	–	53.40	–
XIV [15]	AAAI20	62.21	83.13	91.72	60.18	–	–	–	–	–	–
AGW [16]	TPAMI21	70.05	–	–	66.37	50.19	–	–	–	–	–
HAT [30]	TIFS21	71.83	87.16	92.16	67.56	–	–	–	–	–	–
DML [24]	TCSVT23	77.60	–	–	84.30	–	77.00	–	–	83.60	–
MAUM [14]	CVPR22	83.39	–	–	78.75	–	81.07	–	–	78.89	–
SPOT [28]	TIP22	80.35	93.48	96.44	72.46	56.19	79.37	92.79	96.01	72.26	56.06
GALNet [29]	IVC23	87.48	96.94	98.11	75.59	–	84.81	96.55	98.50	74.98	–
PMT [31]	AAAI23	84.83	–	–	76.55	–	84.16	–	–	75.13	–
CMTR [26]	TMM23	88.11	–	–	81.66	–	84.92	–	–	80.79	–
Ours	–	91.26	96.11	97.91	88.79	81.80	89.35	94.35	96.31	86.85	79.96
Ours+J [4]	–	**91.59**	**97.57**	**98.94**	**89.68**	**82.33**	**90.14**	**96.91**	**98.70**	**87.68**	**80.84**

correlation between the features and the input image from two different modalities, and then employ a significant area erasing operation on the image to acquire the residual image. Secondly, we need to transmit the residual image into another two-stream network to discover the potential inter-modality information. On the basis of these two conditions, we can put the CM-CSEO in four different positions of the network to form four different network structures. We name these four network structures E0, E1, E2, and E3 respectively. As shown in Fig. 3, the features used in the erasing operation in E1 are obtained by first two conv blocks, and then the residual image is fed into the complementary network, where first 2 conv blocks are used for extracting modality-specific features, and following 3 conv blocks are used to extract modality shared features. In comparison to E1, the difference between the other three structures E0, E2 and E3 is the location of the CM-CSEO, corresponding to after first conv block, after third block, and after fourth block, respectively.

From Table 3, we can observe that the best experimental results can be achieved with the E1 network structure for both SYSU-MM01 and RegDB. Although E0 and E2 are not as good as E1, they can also achieve acceptable results. However, E3 did not improve the performance a lot. We think this because the feature map extracted by each layer of the CNN corresponds to a different semantic level. The features extracted at the bottom layer of the CNN focus more on texture and structural information, while the higher-level features include more semantic information. We determine the location of significant feature erasing operation based on the correlation between the image and the features, and this is where texture and structure information can play a bigger role than more abstract semantic information. This is the reason why E3 is not able to achieve better performance.

Table 3
The results of different MRCL module structure on SYSU-MM01 and RegDB datasets.

Network	SYSU-MM01			RegDB		
	R = 1	mAP	mINP	R = 1	mAP	mINP
E0	57.10	52.86	39.53	87.00	84/90	75.88
E1	58.88	55.12	41.46	88.89	85.93	79.45
E2	57.89	53.26	40.12	88.36	84.52	75.61
E3	53.16	40.02	34.24	83.53	79.74	72.88

In Table 4, four-streams network denotes the addition of a same two stream network based on the baseline. Compared to baseline, the four streams network did not have improved a lot. “Random erasing” indicates that the input of a two-stream network is transformed into a random erased image on top of a four-streams network. “Random erasing” can improve model performance on both SYSU-MM01 and

Table 4
The effectiveness of CM-CSEO on SYSU-MM01 and RegDB datasets.

Network	Fusion	SYSU-MM01			RegDB		
		R = 1	mAP	mINP	R = 1	mAP	mINP
Baseline	–	51.10	50.12	36.13	81.35	78.28	71.23
Four-streams	Cat	51.63	50.75	36.86	82.53	79.73	71.01
Random erasing	Cat	52.54	51.13	37.24	83.94	80.95	72.12
CM-CSEO	Sum	52.96	51.32	37.89	83.56	80.36	72.85
CM-CSEO	Cat	58.88	55.12	41.46	88.89	85.93	79.45

Table 5
The effectiveness of different kind of local feature extraction part on SYSU-MM01 and RegDB datasets.

Network	SYSU-MM01			RegDB		
	R = 1	mAP	mINP	R = 1	mAP	mINP
Baseline	51.10	50.12	36.13	81.35	78.28	71.23
Channel part	55.78	52.68	38.41	85.75	82.12	75.62
Spatial part	56.02	53.56	39.15	86.35	83.65	76.81
All local part	59.94	54.97	41.26	89.96	86.02	79.96

RegDB datasets. However, the performance achieved by random erasing is far inferior to our proposed CM-CSEO. As can be seen from Table 4, CM-CSEO can effectively improve the performance, but the effect achieved by different fusion methods also varies greatly. "Sum" means element-wise sum and "Cat" means concatenation. The major difference between these two fusion strategies is that "Cat" changes the number of channels while "Sum" only increases characteristic patterns without changing the number of channels. Some performance improvement can be achieved when taking the "Sum" approach, but it is not obvious. In contrast to the "Sum" approach, CM-CSEO can greatly improve the model by using the "Cat" method.

2) *The effectiveness of the multi-branch feature learning module*

In this subsection, we prove the effectiveness of the multi-branch feature learning module. MBFL module can be categorized into three types: channel part which extracts only local channel features, spatial part which extracts only horizontal spatial features and all local part.

From Table 5, we can observe that different types of local feature extraction networks can effectively enhance the performance. Both the channel part and the spatial part can improve performance considerably. In contrast, the spatial part gets a slightly bigger boost. The most significant performance improvement is achieved when we combine the channel part and the spatial part to form the all local part. This illustrates that fused all local features can make a fuller contribution to the whole network.

3) *The effectiveness of the multi-branch constrain loss*

In this subsection, we performed ablation experiments of multi-branch constrain loss. From Table 6, we observe that both L_{intra} and $L^{g}_{hc\_cross}$ can effective improve the performance. Besides, $L^{g}_{hc\_cross}$ works a little better than L_{intra}. However, combining the two components can

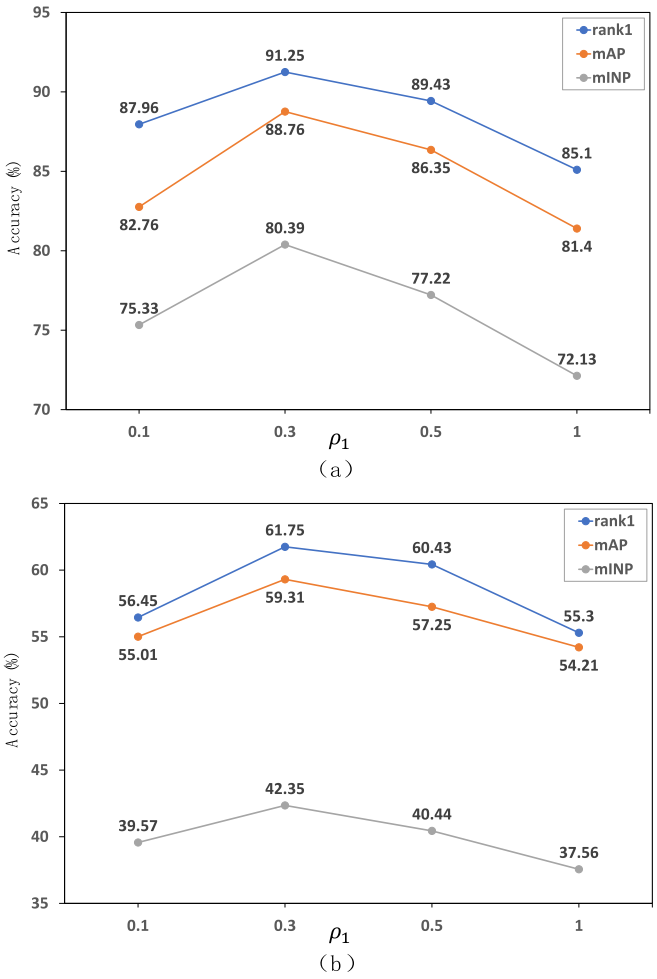

Fig. 5. The experimental results of different intra-modality constraint loss margins on (a) RegDB and (b) SYSU-MM01 datasets.

significantly improve performance on both datasets. It demonstrates that we need to concentrate on intra-class compactness and inter-class separability for both the intra-modality and cross-modality. Apart from this, we also confirm the effectiveness of the hetero-center cross-modality constraint loss for the final fused features ($L^{f}_{hc\_cross}$). When we adopt the multi-branch constraint loss completely, we can achieve the best results on all indicators.

4.6. Parameter discussion

In this subsection, we discuss some parameters in the network, the margin ρ_1 in the intra-modality constraint loss, the margin ρ_2 in the hetero-center cross-modality constraint loss, the dimension d of the local channel features and local spatial features in the multi-branch feature learning module and the number of local features in the multi-branch feature learning module.

Table 6
The effectiveness of the multi-branch constrain loss on RegDB and SYSU-MM01 datasets.

Loss function				SYSU-MM01			RegDB		
L_{id}	L_{intra}	$L^{g}_{hc\_cross}$	$L^{f}_{hc\_cross}$	R = 1	mAP	mINP	R = 1	mAP	mINP
√				50.17	49.17	34.87	81.61	78.93	72.65
√	√			55.48	53.08	36.00	86.46	83.16	75.23
√		√		56.51	54.26	36.88	87.53	84.23	77.85
√	√	√		58.48	56.08	38.05	89.82	86.61	80.39
√	√	√	√	61.75	59.51	42.35	91.26	88.79	81.80

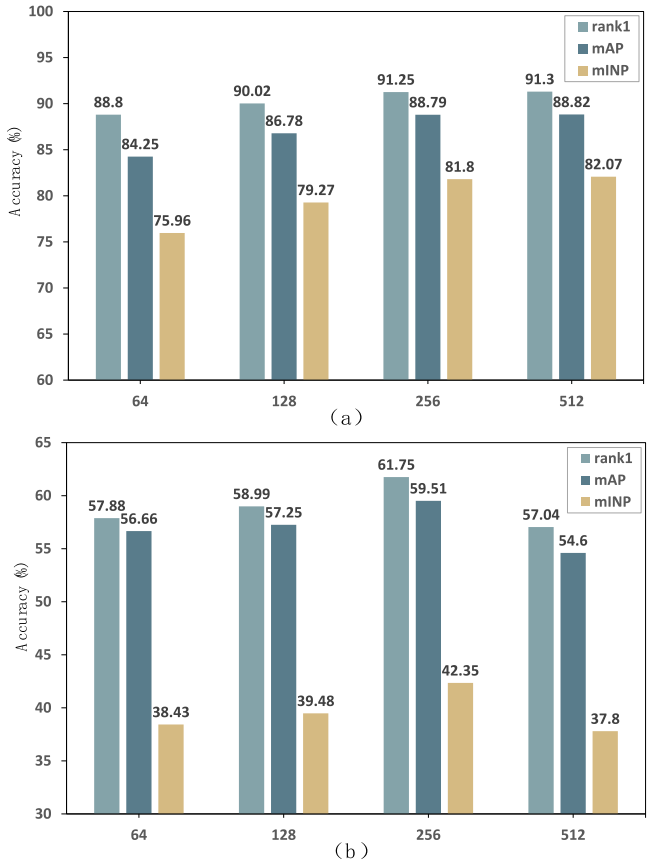

Fig. 6. The experimental results of different dimensions of local features on (a) RegDB and (b) SYSU-MM01 datasets.

1) *The effect of the margin ρ_1 in the intra-modality constraint loss*

In the intra-modality constraint loss, we predefine the ρ_1 to determine the minimal distance of the anchor image and its negative image. In this subsection, experiments are carried out using various values. of the margin ρ_1 in the intra-modality constraint loss on the RegDB and SYSU-MM01 datasets.

As demonstrated in Fig. 5, we could inspect that intra-modality constraint loss achieves the best performance when margin ρ_1=0.3 for both of RegDB and SYSU-MM01 datasets. It also proved that intra-modality constraint loss could help improve the performance to some extent with proper margin ρ_1.

2) *The effect of the margin ρ_2 in the hetero-center cross-modality constraint loss*

Different training data have different requirements for margin ρ_2 in hetero-center cross-modality constraint loss, so we discuss the values of the margin on two different datasets, RegDB and SYSU-MM01, in this subsection.

From (a) in Fig. 6, we find out that the experimental effect on RegDB gradually gets better as the margin ρ_2 gradually increases from 0 to 1, while the model performance progressively decreases as ρ_2 increases from 1 to 1.5. Therefore, the optimal value for the hetero-center cross-modality loss margin on the RegDB dataset is 1. As can be observed from Fig. 6 (b), the effect of changing the value on SYSU-MM01 is similar to the situation on RegDB, except that the best value is changed from 1 to 0.5. Therefore, the model has the best performance on SYSU-MM01 when the margin is 0.5.

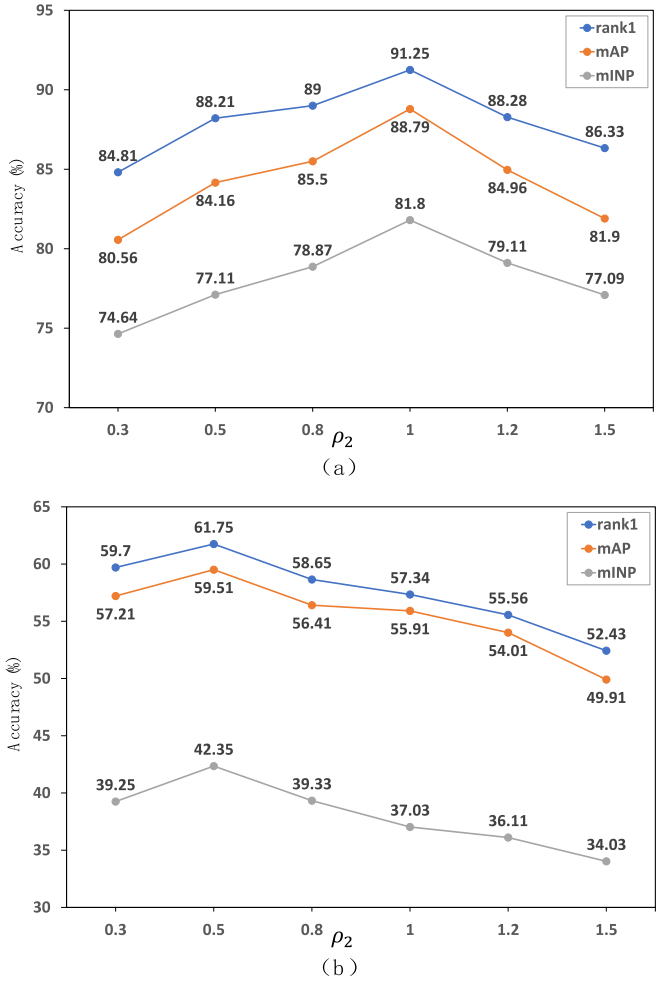

Fig. 7. The experimental results of different hetero-center cross-modality constraint loss margins on (a) RegDB and (b) SYSU-MM01 datasets.

3) *The effect of different local feature dimensions in the multi-branch feature learning module*

In the multi-branch feature learning module, we finally change the dimensions of local channel features and spatial features through the batch normalization layer. Therefore, in this subsection, we discuss the effect of the different dimensions of local features on two different datasets, RegDB and SYSU-MM01. As shown in Fig. 7(a), the performance of the model improves progressively as the local feature dimension increases on the RegDB dataset. When dimensions d is increased from 256 to 512, the model performance has improved a little, but this improvement is based on the premise that the feature dimension is doubled. Therefore, considering the performance of the model and the impact of feature dimensionality on network complexity, we set the final dimension d to 256 on the RegDB dataset. In Fig. 7(b), we can clearly observe that d = 256 performs the best effect on the rank1, mAP, and mINP criterion for the SYSU-MM01. Therefore, we set the dimension on SYSU-MM01 to 256.

4.7. Visualization analysis

In this subsection, we compare the visualization results of modality-specific feature maps extracted by baseline and our proposed MRCL module. Visible and infrared images with the same identity were selected as input images from datasets SYSU-MM01 and RegDB in Fig. 8 (a). As shown in Fig. 8, feature of baseline and feature of upper stream of MRCL almost pay attention to the same part. With the help of the CM-

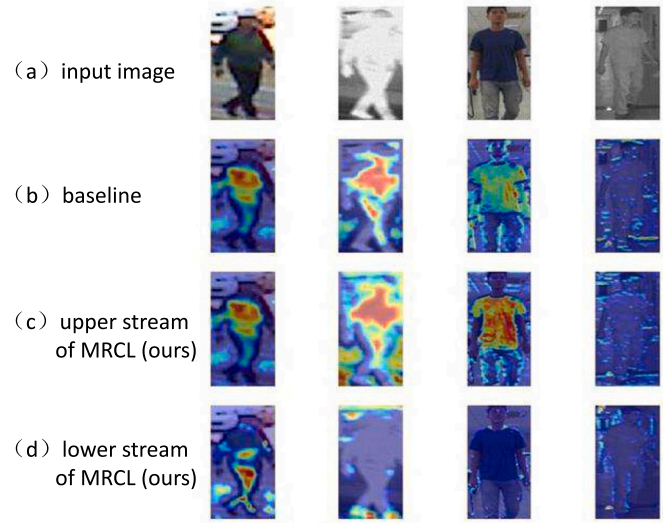

Fig. 8. Feature map visualization of baseline and MRCL.

CSEO, the lower stream of MRCL is able to mine potential complementary information in Fig. 8(d). With the complementary parts mined by the lower stream of MRCL, the final performance of our model can be improved significantly.

5. Conclusion

In this paper, we proposed an innovative visible thermal person re-identification (VT Re-ID) method named multi-branch modality residual complementary feature learning. Our method incorporates two modules, the modality residual complementary learning (MRCL) module and the multi-branch feature learning (MBFL) module. MRCL drives the model simultaneously considers the intra-modality features and inter-modality features containing important complementary information that was ignored by the conventional two-stream network. We can obtain diverse and discriminative features to reduce cross-modality discrepancy with the help of MRCL. Meanwhile, we also adopt the MBFL module to capture local spatial features and local channel features and integrate them with global features to acquire ultimate representations. The MBFL handles the intra-modality variations by part-to-part matching and high-level semantic concept-to-concept matching. We further introduce the multi-branch constraint loss function to utilize the different attributes of both global and local features thoroughly. Experimental results on two VT Re-ID datasets RegDB and SYSU-MM01 validate the superior performance of the proposed method.

CRediT authorship contribution statement

Long Chen: Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Rui Sun:** Writing – review & editing, Methodology, Investigation, Funding acquisition, Conceptualization. **Yiheng Yu:** Writing – review & editing, Investigation, Conceptualization. **Yun Du:** Writing – review & editing, Validation. **Xudong Zhang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302142 and Grant 61876057; in part by the China Postdoctoral Science Foundation under Grant 2022M720981; in part by the Anhui Province Natural Science Foundation under Grant No. 2208085MF158; and in part by the Key Research Plan of Anhui Province - Strengthening Police with Science and Technology under Grant 202004d07020012.

References

[1] J. Li, S. Zhang, Q. Tian, M. Wang, W. Gao, Pose-guided representation learning for person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2) (2022) 622–635.
[2] Y. Fu, et al., Horizontal pyramid matching for person re-identification, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 8295–8302.
[3] M. Ye, Z. Wang, X. Lan, P.C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking, Proc. Int. Joint Conf. Artif. Intell. 1 (2018) 1092–1099.
[4] M. Ye, W. Ruan, B. Du, M.Z. Shou, Channel Augmented Joint Learning for Visible-Infrared Recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 13547–13556.
[5] M. Ye, X. Lan, Q. Leng, J. Shen, Cross-modality person re-identification via modality-aware collaborative ensemble learning, IEEE Trans. Image Process. 29 (2020) 9387–9399.
[6] H. Liu, X. Tan, X. Zhou, Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification, IEEE Trans. Multimed. 23 (2021) 4414–4425.
[7] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision, 2018, pp. 480–496.
[8] X. Liu, H. Zhao, M. Tian, et al., Hydraplus-net: Attentive deep features for pedestrian analysis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 350–359.
[9] G. Wang, Y. Yuan, X. Chen, et al., Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 274–282.
[10] L. Wei, S. Zhang, W. Gao, et al., Person transfer Gan to bridge domain gap for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, p. 79.
[11] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, Q. Tian, Spatial-temporal graph convolutional network for video-based person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3289–3299.
[12] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, RGB-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.
[13] M. Ye, X. Lan, J. Li, P. Yuen, Hierarchical discriminative learning for visible thermal person re-identification, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018) 7501–7508.
[14] J. Liu, Y. Sun, F. Zhu, H. Pei, Y. Yang, W. Li, Learning memory-augmented unidirectional metrics for cross-modality person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 19344–19353.
[15] D. Li, X. Wei, X. Hong, Y. Gong, Infrared-visible cross-modal person re-identification with an x modality, Proc. AAAI Conf. Artif. Intell. 34 (04) (2020) 4610–4617.
[16] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C.H. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2022) 2872–2893.
[17] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training, Proc. Int. Joint Conf. Artif. Intell. 1 (3) (2018) 677–683.
[18] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S.I. Satoh, Learning to reduce dual-level discrepancy for infrared-visible person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 618–626.
[19] G.A. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3623–3632.
[20] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, Proceedings of the AAAI conference on artificial intelligence 34 (07) (2020) 13001–13008.
[21] X. Liu, Z. Wang, J. Shao, X. Wang, H. Li, Improving referring expression grounding with cross-modal attention-guided erasing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1950–1959.

[22] Y. Dai, X. Li, J. Liu, Z. Tong, L.-Y. Duan, Generalizable person re-identification with relevance-aware mixture of experts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 16145–16154.
[23] Y. Dai, J. Liu, Y. Sun, Z. Tong, C. Zhang, L.-Y. Duan, Idm: An intermediate domain module for domain adaptive person re-id, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 11864–11874.
[24] D. Zhang, Z. Zhang, Y. Ju, C. Wang, Y. Xie, Y. Qu, Dual mutual learning for cross-modality person re-identification, IEEE Trans. Circuits Syst. Video Technol. 32 (8) (2022) 5361–5373.
[25] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1487–1495.
[26] T. Liang, Y. Jin, W. Liu, Y. Li, Cross-modality transformer with modality mining for visible-infrared person re-Identification, IEEE Trans. Multimed. 25 (2023) 8432–8444.
[27] D.T. Nguyen, H.G. Hong, K.W. Kim, K.R.J.S. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17 (3) (2017) 605.
[28] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Trans. Image Process. 31 (2022) 2352–2364.
[29] Jiaqi Zhao, Hanzheng Wang, Yong Zhou, Rui Yao, Lixu Zhang, Abdulmotaleb El Saddik, Context-aware and part alignment for visible-infrared person re-identification, Image Vis. Comput. 138 (2023) 104791.
[30] M. Ye, J. Shen, L. Shao, Visible-infrared person re-identification via homogeneous augmented tri-modal learning, IEEE Trans. Inf. Forensics Secur. 16 (2020) 728–739.
[31] H. Lu, X. Zou, P. Zhang, Learning progressive modality-shared transformers for effective visible-infrared person re-identification, Proc. AAAI Conf. Artif. Intell. 37 (2) (2023) 1835–1843.
[32] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, TransReID: Transformer-based object re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 14993–15002.
[33] Q. Zhang, L. Wang, V.M. Patel, X. Xie, J. Lai, View-decoupled transformer for person re-identification under aerial-ground camera network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 22000–22009.
[34] W. Tan, C. Ding, P. Wang, M. Gong, K. Jia, Style interleaved learning for generalizable person re-identification, IEEE Trans. Multimed. 26 (2024) 1600–1612.
[35] P. Zhang, H. Yan, W. Wu, et al., Improving federated person re-identification through feature-aware proximity and aggregation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 2498–2506.
[36] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, C.-W. Lin, Structure-aware positional transformer for visible-infrared person re-identification, IEEE Trans. Image Process. 31 (2022) 2352–2364.
[37] P. Zhang, Y. Wang, Y. Liu, Z. Tu, H. Lu, Magic tokens: Select diverse tokens for multi-modal object re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 17117–17126.
[38] J. Shi, et al., Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2023, pp. 11184–11194.
[39] Y. Zhang, H. Wang, Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification, in: Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 2153–2162.
[40] K. Ren, L. Zhang, Implicit discriminative knowledge learning for visible-infrared person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2024, pp. 393–402.
[41] K. Arora, A. Kumar, A comparative study on content based image retrieval methods, Int. J. Latest Technol. Eng. Manag. Appl. Sci. 6 (4) (2017) 77–80.
[42] T. Kumari, P. Syal, A.K. Aggarwal, V. Guleria, Hybrid image registration methods: a review, Int. J. Adv. Trends Comput. Sci. Eng. 9 (2) (2020) 1134–1142.