Visible-Infrared Person Re-Identification via Specific and Shared Representations Learning

Aihua Zheng · **Juncong Liu** · **Zi Wang** · **Lili Huang** · **Chenglong Li** · **Bing Yin**

Received: date / Accepted: date

Abstract The primary goal of visible-infrared person re-identification (VI-ReID) is to match pedestrian photos obtained during the day and night. The majority of existing methods simply generate auxiliary modality to reduce the modality discrepancy for cross-modality matching. They capture modality-invariant representations but ignore the extraction of modality-specific representations that can aid in distinguishing among various identities of the same modality. To alleviate these issues, this work provides a novel Specific and Shared Representations Learning (SSRL) model for VI-REID to learn modality-specific and modality-shared representations. We design a shared branch in SSRL to bridge the image-level gap and learn modality-shared representations, while a specific branch to retain the discriminative information of visible images to learn modality-specific representations. In addition, we propose intra-class aggregation and inter-class separation learning strategies to optimize the distribution of feature embeddings at a fine-grained level. Extensive experimental results on two challenging benchmark datasets SYSU-MM01 and the RegDB demonstrate the superior performance of SSRL over state-of-the-art methods.

A. Zheng and C. Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: ahzheng214@foxmail.com; lcl1314@foxmail.com).
J. Liu, Z. Wang, and L. Huang are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: liujuncong1115@163.com; ziwang1121@foxmail.com; hill_ahu@ahu.edu.cn)
B. Yin is with iFLYTEK Research, No.666 West Wangjiang Road, Hefei City, Anhui Province, China (e-mail:bingyin@iflytek.com).
Corresponding author: Lili Huang.

Keywords Person Re-identification · Cross-modality · Specific Representations · Shared Representations

1 Introduction

Single-modality person re-identification (Re-ID) is a pedestrian matching problem between query and gallery photos from separate cameras, which has received much attention in computer vision [1–6]. Visible cameras play a limited role in night monitoring and security work, therefore visible infrared person re-identification (VI-ReID) [7] is proposed to match the image of people captured by visible and infrared cameras. This task aims to solve not only intra-modality discrepancies caused by different camera perspectives and human poses in single-modality visible person Re-ID but also inter-modality discrepancies by various spectral cameras.

Reducing the modality discrepancy and learning modality-invariant representations is a significant challenge. Cross-modality matching using grayscale images is a popular method to eliminate the color discrepancy [7, 8]. However, modality-specific representations are discarded along with color information, and specific representations are a crucial decision-making accordance for retrieval, which can help to widen the inter-class discrepancy. Using Generative Adversarial Networks (GANs) to generate concrete visualizations [9, 10] to eliminate image-level gaps allows the task to be reduced as much as possible to a single-modal task, but with the inevitable generation of noise. Another method is to build complex networks [11, 12] for learning modality-invariant representations, but due to the huge modality discrepancy, heterogeneous modality features cannot be well projected into a unified space. In summary, extract-

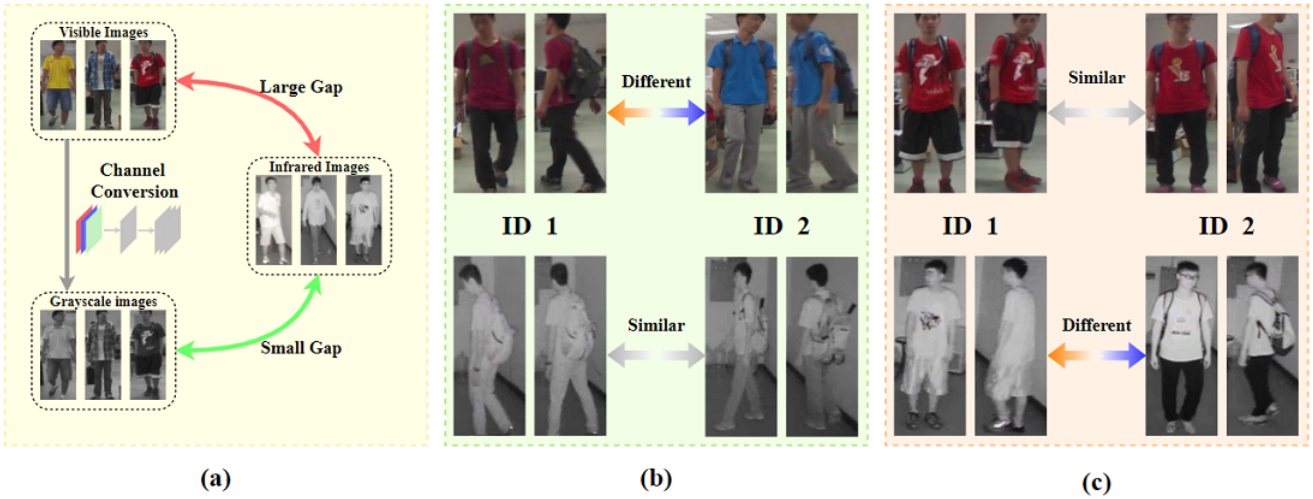

Fig. 1 (a) The conversion of visible light images to grayscale images. After that, the large gap between visible and infrared images shrinks to the small gap between infrared and grayscale images. (b) Different IDs are similar under visible images but different under infrared photos. The discriminative information of visible light images can help the network distinguish between different IDs. (c) Different IDs are different under visible photos but similar under infrared images. The shared representations of infrared images can help find the same identity.

ing and decoupling specific and shared representations is a challenge for current methods. Although [13] have made a preliminary attempt, it cannot completely decouple specific and shared representations.

This paper proposes a specific and shared representations learning (SSRL) model for VI-ReID to mitigate the discrepancy between heterogeneous pedestrian images while better capturing modality-invariant and modality-specific representations. Specifically, in order to learn modality-specific representations, we employ the traditional RGB-NIR branch as our specific branch. The color information in RGB images is an important decision-making accordance for retrieval which may lead to large inter-class differences and slight intra-class differences. Therefore, it is necessary to decompose modality-specific representations. Specific branch takes visible and infrared images as inputs and feature extractors with different parameters to ensure that the extracted features are modality-specifc representations. We design a shared branch, as seen in Fig. 1 (a). We can obtain grayscale images directly from visible images using channel conversion, which converts three-channel visible images into single-channel grayscale images and replicates them to the three channels. The converted grayscale images alleviated the modality discrepancy with the infrared images, and then fed them into the feature extractor. The shared branch feature extractor is parameter-shared and used to learn modality-invariant representations. Our dual-branch structure has two main benefits. First, the specific branch improves extraction to modalitiy-specific representations by preserving color information from visible images. As illustrated in Fig. 1 (b), when infrared images from different IDs are similar and visible images from different IDs are different, it is required to rely on specific representations in the visible light images to separate them. Second, grayscale operations for shared branch alleviates the modality discrepancy, allowing the parameter-shared feature extractor to capture the modality-shared representations more effectively. As illustrated in Fig. 1 (c), when visible images from different IDs are similar and infrared images from different IDs are different, it is vital to rely on shared representations to bring the features of different modalities of the same ID closer. In addition, we design effective intra-class aggregation and inter-class separation learning strategies to optimize the distribution of feature embeddings on a fine-grained level. It constrains the distance of different class centers from both the same modality and cross-modality and plays a role in expanding the intra-class distance and minimizing the inter-class distance. In the early stages of training, substantial disparities exist between modality-specific features, rendering them unsuitable for cross-modality tasks. However, we mitigated this issue by using intra-class agregation learning (IAL) loss, which constrains the differences between modality-specific and modality-shared features. As training progresses, these differences gradually diminish, while both types of features continue to contribute effectively to the network's accurate identification of individual identities. During the training process, we employed CE loss and inter-class separation learning (ISL) loss to enhance the representation capability of the final features by ex-

tracting modality-specific feature and combining it with modality-shared feature through concatenation. This approach shows advantages in increasing the distinctiveness between features from different IDs and, to some extent, improves the ability of the feature extraction network to extract ID-related information.

Our main contributions are summarized below:

- A novel specific and shared representations learning model termed SSRL is proposed for the VI-ReID task, which contains a specific branch and a shared branch to learn modality-specific and modality-shared representations.
- The intra-class aggregation and inter-class separation learning strategies are further developed to optimize the distribution of feature embeddings on a fine-grained level.
- Extensive experimental results on the SYSU-MM01 and the RegDB datasets show that our proposed method achieves a new state-of-the-art performance.

2 Related Work

2.1 Single-Modality Person Re-Identification

Single-modality person Re-ID aims at matching the person images captured by different cameras in the daytime, while all the images are from the same visible modality. Existing works have shown desirable performance on the widely-used datasets with deep learning technique [14–16]. [17] propose an attribute-person recognition (APR) network, a multi-task network which learns a re-ID embedding and at the same time predicts pedestrian attributes. [18] propose a network named Part-based Convolutional Baseline (PCB) which conducts uniform partition on the conv-layer for learning part-level features and an adaptive pooling method named Refined Part Pooling (RPP) to improve the uniform partition. [19] formulate a harmonious attention CNN model for joint learning of pixel and regional attention to optimize reID performance with misaligned images. Due to the tremendous discrepancy between visible and infrared images, single-modality solutions are not suitable for cross-modality person re-identification, which creates a demand for the development of VI-ReID solutions. [20] propose a general framework, namely JoT-GAN, to jointly train GAN and the re-id model.

2.2 Visible-Infrared Person Re-Identification

The main challenge in VI-ReID is appearance discrepancy, including large intra-class and slight inter-class variations. The existing VI-ReID methods are divided into three categories: representation learning, metric learning, and image generation. The approach of representation learning based on feature extraction mainly explores how to construct a reasonable network architecture, which can extract the robust and discriminating features shared by the two modality images. [7] firstly define the VI-REID problem and contribute a new multiple modality Re-ID dataset SYSU-MM01 for research, and they propose deep zero-padding for training one-stream network towards automatically evolving domain-specific nodes in the network for cross-modality matching. [21] offer a novel Modality Confusion Learning Network (MCLNet) to confuse two modalities, ensuring that the optimization is explicitly concentrated on the modality-irrelevant perspective. [22] propose a novel hierarchical cross-modality matching model which could simultaneously handle both cross-modality discrepancy and cross-view variations, as well as intra-modality intra-person variations. For exploring the potential of both the modality-shared information and the modality-specific characteristics to boost the re-identification performance, [13] propose a novel cross-modality shared-specific feature transfer algorithm to tackle the above limitation. To learn discriminative feature representations, a dual-path network with a novel bi-directional dual-constrained top-ranking loss is developed in [23]. Metric learning is to learn the similarity of two pictures through the network, and the key is to design a reasonable measurement method or loss function. [24] design a novel loss function, called Hetero-Center loss, to constrain the distance between two centers of heterogenous modality. [25] propose the hetero-center triplet loss to constrain the distance of different class centers from the same modality and cross-modality. [26] use Sphere Softmax to learn a hypersphere manifold embedding and constrain the intra-modality variations and cross-modality variations on this hypersphere. [27] propose the dual-modality triplet loss to constrain both the cross-modality and intra-modality and address the cross-modality discrepancy and intra-modality variations. [28] propose a novel loss function called HP loss, which can simultaneously handle the cross-modality and intra-modality variations. Image generation means using GANs or other methods to reduce discrepancies by transforming one modality into another. [10] propose generating cross-modality paired images and performing global set-level and fine-grained instance-level alignments. [29] propose a Hierarchical Cross-Modality Disentanglement method that extracts pose- and illumination-invariant features for cross-modality matching. [9] translate an infrared image into its visible counterpart and a visible image into its infrared version, which can unify the rep-

resentations for images with different modalities. [30] proposes a novel and end-to-end Alignment Generative Adversarial Network (AlignGAN) for the RGB-IR RE-ID task. [31] introduce an auxiliary X modality as an assistant and reformulate infrared-visible dual-mode cross-modality learning as an X-Infrared-Visible three-mode cross-modal learning problem.

3 Method

This section will detail the SSRL model proposed for visible-infrared person Re-ID. The proposed SSRL model is illustrated in Fig. 2. We first propose a dual-branch structure containing a shared branch to alleviate modality discrepancy to learn modality-invariant representations and a specific branch to learn more accurate modality-specific representations. Then, we use the intra-class aggregation and inter-class separation learning strategies to further optimize the distribution of features and aggregate instances with the same identity.

3.1 Baseline

We adopt ResNet-50 [32] as the backbone, in which each branch contains a pre-trained model. At the same time, we use max pooling to obtain fine-grained features. Inspired by the work of PCBs [18,33] in extracting discriminant features, the work divides the feature map horizontally into sections and feeds each part into a classifier to learn local clues and sets the part to 6. Following the state-of-the-art methods, we utilize identity loss [34] $\mathcal{L}_{ID}$ and hetero-center triplet loss [25] $\mathcal{L}_{HCT}$ to constrain the network, the baseline learning loss is denoted as $\mathcal{L}_{base}$.

$$\mathcal{L}_{base} = \mathcal{L}_{ID} + \mathcal{L}_{HCT}. \tag{1}$$

3.2 Dual-Branched Structure

3.2.1 Specific Branch

The color information in the visible light image is crucial discriminative information, which can help the network expand the difference among classes as much as possible. The specific branch takes the original RGB and NIR images as input and employs two parameter-independent feature extractors to ensure that the network learns specific characteristics more effectively. Through CNN, global max pooling (GMP), and batch normalization (BN) operations, feature vectors are input to the fully connected layer for identity classification. Since the input to the specific branch includes RGB images, the network pays more attention to learning discriminative color information. For this branch, identity loss is formulated as follows:

$$\mathcal{L}_{ID}^{specific} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{e^{\boldsymbol{W}_{y_n}^T \boldsymbol{f}_n^V}}{\sum_{u=1}^{U} e^{\boldsymbol{W}_u^T \boldsymbol{f}_n^V}}, \tag{2}$$

where N denotes the sample number in a batch, y_n and f_n^V are the identity and the feature vector of the n-th pedestrian image, U is the number of identities, and $\boldsymbol{W}_u$ denotes the u^{th} column of the weights. Under the supervision of identity loss, the specific branch can extract information about a specific identity for classification.

3.2.2 Shared Branch

The visible image has three channels, which contain the visible light color information of red, green and blue, while the infrared image has only one channel, which contains the intensity information of near-infrared light. From the perspective of imaging principles, the wavelength ranges of the two are also different. Different sharpness and lighting conditions can produce very different effects on the two types of images. The large modality discrepancy makes it very challenging to extract modality-shared features directly using feature extractors. Thus we introduced an additional shared branch. The shared branch first performs a grayscale transformation operation on the visible light image to reduce modality discrepancy. This is done for a given visible image x_v^i with three channels $\mathcal{R}$, $\mathcal{G}$, $\mathcal{B}$, we take the $\mathcal{R}(x)$, $\mathcal{G}(x)$ and $\mathcal{B}(x)$ values for each pixel of the visible image x_v^i. The corresponding grayscale pixel point $\mathcal{G}(x)$ can therefore be calculated as:

$$\mathcal{G}(x) = \alpha_1\mathcal{R}(x) + \alpha_2\mathcal{G}(x) + \alpha_3\mathcal{B}(x), \tag{3}$$

where α_1, α_1 and α_1 are each set to 0.299, 0.587 and 0.114. The grayscale image is then restored to a three-channel image and sent to the shared branch along with the infrared image. The shared branch's two feature extractor parameters are shared. The shared branch removes color information from the network, allowing it to focus on learning modality-invariant representations such as texture and structural information. The identity loss of the shared branch $\mathcal{L}_{ID}^{shared}$ is the same as that of the specific branch.

3.2.3 Fusion Features

We employ concatenation to fuse the obtained specific and shared features from dual-branch cooperative learning on visible and infrared images from specific and

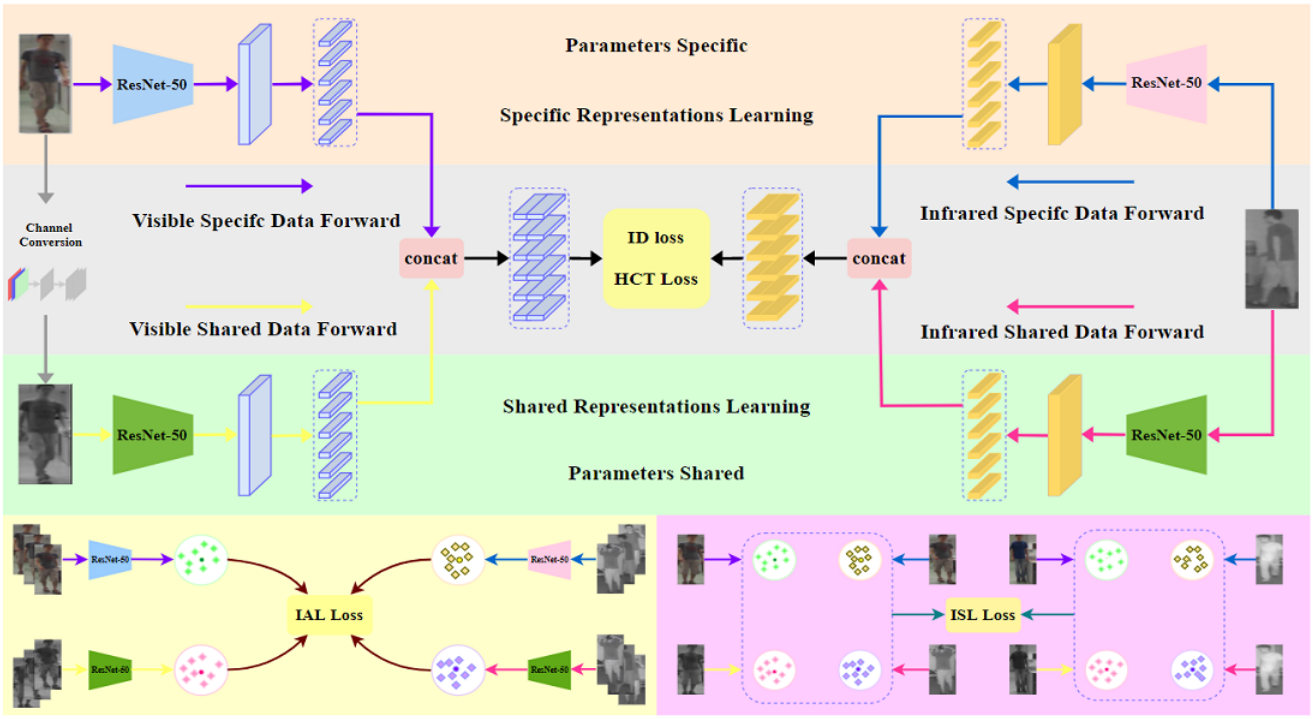

Fig. 2 Framework of the proposed Specific and Shared Representations Learning (SSRL) model. The cross-modality images are fed into a dual-branch structure, with one branch dedicated to learning specific representations and the other dedicated to learning shared representations and fusing the specific representations and shared representations. For the fusion feature, the identity (ID) loss is leveraged to enhance the discriminative power of the embedding features and the hetero-center triplet (HCT) loss is leveraged to constrain the distance of different class centers from both the same modality and cross-modality. For the specific feature and shared feature, intra-class aggregation learning (IAL) loss and inter-class separation learning (ISL) loss are further developed to optimize the distribution of feature embeddings on a fine-grained level.

shared branches. The discriminative information of specific features in the fusion feature can assist the model in to separating different identities. The invariant information of the shared features in the fusion feature can assist the model in identifying the same identity in different modalities. As a result, for synchronous classification learning, we implement an identity classification layer for the fusion features. The fusion feature $\mathcal{L}_{ID}^{fusion}$ has the same identity loss computation as the specific branch. Add the above identity loss to obtain the final identity loss function:

$$\mathcal{L}_{ID}^{final} = \mathcal{L}_{ID}^{specific} + \mathcal{L}_{ID}^{shared} + \mathcal{L}_{ID}^{fusion}, \tag{4}$$

and the hetero-center triplet loss $\mathcal{L}_{HCT}$ is also used for metric learning of the fusion feature.

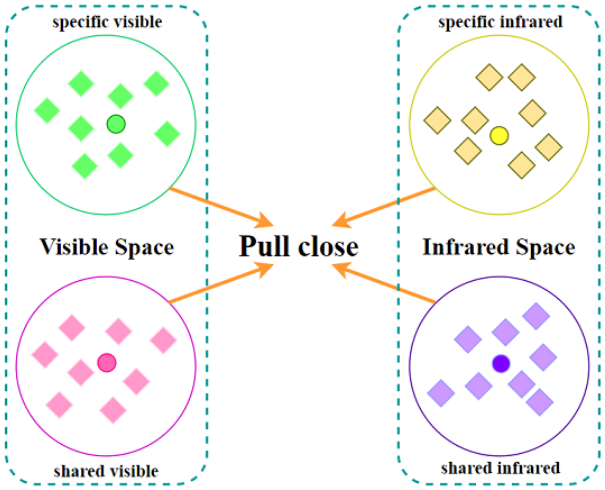

Fig. 3 Illustration of intra-class aggregation learning strategy. The distribution of the four features is shown by four different colors of diamonds. The circle represents the center of the feature distribution of a modality. The orange line represents intra-class aggregation.

3.3 Intra-Class Aggregation Learning

Due to the camera viewpoint, clothing, posture, and other factors, the distance between sample pairs among classes is frequently longer than the distance between sample pairs intra-classes. We apply intra-class aggregation learning strategy to limit the distance in each class to increase cross-modality intra-class similarity.

By intra-class aggregation learning strategies, we aggregate different modalities of the same identity. It is challenging to restrict the distance in each class's distribution of visible-specific features, infrared-specific features, visible-shared features, and infrared-shared features, therefore, take the centers of these feature distributions and penalize the center distance. As shown in Fig. 3, we suppose that there are P × K images of P identities in a mini-batch, where each identity contains K images. The feature distribution center of identity in visible-specific features is calculated as follows:

$$c_{vspe}^{p} = \frac{1}{K}\sum_{k=1}^{K}(vspe)_{k}^{p}, p \in [1, P], \tag{5}$$

$(vspe)_{k}^{p}$ denotes the feature vector of the k-th image output. Infrared-specific features, visible-shared features, and infrared-shared feature distribution centers are calculated as c_{ispe}^{p}, c_{vsha}^{p}, c_{isha}^{p} in the same way. We introduce an intra-class aggregation constraint loss to handle the distance among different modalities of the same identity, which can be interpreted as:

$$\begin{aligned}\mathcal{L}_{IAL} = \sum_{p=1}^{P}\Big[&\left\|\boldsymbol{c}_{vspe}^{p} - \boldsymbol{c}_{ispe}^{p}\right\|_2^2 + \left\|\boldsymbol{c}_{vspe}^{p} - \boldsymbol{c}_{vsha}^{p}\right\|_2^2 \\ &+ \left\|\boldsymbol{c}_{vspe}^{p} - \boldsymbol{c}_{isha}^{p}\right\|_2^2 + \left\|\boldsymbol{c}_{ispe}^{p} - \boldsymbol{c}_{vsha}^{p}\right\|_2^2 \\ &+ \left\|\boldsymbol{c}_{ispecifc}^{p} - \boldsymbol{c}_{isha}^{p}\right\|_2^2 + \left\|\boldsymbol{c}_{vsha}^{p} - \boldsymbol{c}_{isha}^{p}\right\|_2^2\Big],\end{aligned} \tag{6}$$

as shown in Fig. 3, the visible-specific features, infrared-specific features, visible-shared features, and infrared-shared features represented by the four colors are focused closer by intra-class aggregation learning strategy. Our goal is to reduce the distance between different modality centers of the same identity, thereby suppressing cross-modality variations.

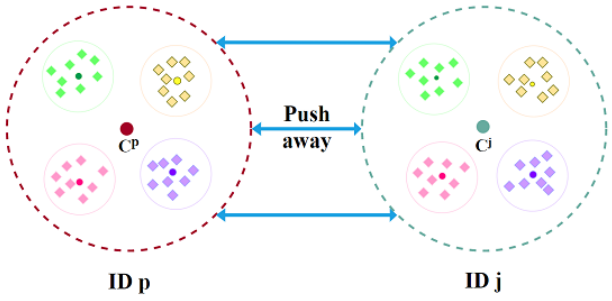

Fig. 4 Illustration of inter-class separation learning strategy. Different dotted circles represent the distribution of features for different identities. The blue line represents inter-class aggregation.

3.4 Inter-Class Separation Learning

Intra-class aggregation learning strategy can only ensure that samples of cross-modality of the same identity are aggregated together, but the model also needs to ensure the dissimilarity among different identities. The extracted modality-specific representations can help different IDs to achieve inter-class separation. By inter-class separation learning strategies, we separate different identities. We first calculate the center of all samples for each identity and then constrain the distance of the distribution of different identity features by cosine distance. The inter-class separation loss is calculated as follows:

$$\mathcal{L}_{ISL} = \sum_{p=1}^{P}\sum_{\mathrm{j}\neq p}^{P}\max\left[0, 1 - \cos(c^{p}, c^{j}) - m\right], \tag{7}$$

where c^{p}, c^{j} denote the center of the(c_{vspe}^{p}, c_{ispe}^{p}, c_{vsha}^{p}, c_{isha}^{p}), (c_{vspe}^{j}, c_{ispe}^{j}, c_{vsha}^{j}, c_{isha}^{j}), $\cos(\cdot,\cdot)$ represents the cosine distance in centers of different identities, m is a margin term. As shown in Fig. 4, with the inter-class separation learning strategy, the feature distributions of different identities are separated.

3.5 Objective Function

As mentioned above, the core goal of SSRL is to jointly learn the two branches, obtain modality-specific representations and modality-invariant representations, and make full use of them. Combined with the losses mentioned above, we finally define the total loss of the overall network as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ID}^{final} + \lambda_1\mathcal{L}_{HCT} + \lambda_2\mathcal{L}_{IAL} + \lambda_3\mathcal{L}_{ISL}, \tag{8}$$

where λ_1, λ_2 and λ_3 are the weights of $\mathcal{L}_{HCT}$, $\mathcal{L}_{IAL}$ and $\mathcal{L}_{ISL}$.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our proposed method on two publicly available VI-ReID datasets(SYSU-MM01 [7] and RegDB [51]).

SYSU-MM01 dataset contains 287,628 visible images and 15,729 infrared images, captured by 4 RGB cameras and 2 thermal imaging cameras, with a total of 491 valid IDs, of which 296 identities are used for training, 99 for verification, and 96 for testing. During the testing

Table 1 Comparison with the state-of-the-arts on SYSU-MM01 datasets and RegDB datasets. R=1, R=10 denotes the Rank-1, Rank-10 accuracy. Rank-k accuracy (%) and mAP (%) are reported. Herein, the best and second best results are indicated by red and green fonts. († This paper reports a higher accuracy by using the transferred graph features. We use backbone features for inference for a fair comparison).

Method	SYSU-MM01						RegDB					
	All Search			Indoor Search			VIS to IR			IR to VIS		
	R=1	R=10	mAP	R=1	R=10	mAP	R=1	R=10	mAP	R=1	R=10	mAP
Zero-Pad [7]	14.80	54.12	15.95	20.58	68.38	26.92	17.75	34.21	18.90	16.63	34.68	17.82
cmGAN [35]	26.97	67.51	27.80	31.63	77.23	42.19	-	-	-	-	-	-
HSME [26]	20.68	32.74	23.12	-	-	-	50.85	73.36	47.00	50.15	72.40	46.16
Hi-CMD [29]	34.94	77.58	35.94	-	-	-	70.93	86.39	66.04	-	-	-
DDAG [11]	54.75	90.39	53.02	61.02	94.06	67.98	69.34	86.19	63.46	68.06	85.15	61.80
AGW [1]	47.50	84.39	47.65	54.17	91.14	62.97	70.05	86.21	66.37	70.49	87.12	65.90
NFS [36]	56.91	91.34	55.45	62.79	96.53	69.79	80.54	91.96	72.10	77.95	90.45	69.79
MSO [37]	58.70	92.06	56.42	63.09	96.61	70.31	73.60	88.60	66.90	74.60	88.70	67.50
MCLNet [21]	65.40	93.33	61.98	72.56	96.98	76.58	80.31	92.70	73.07	75.93	90.93	69.49
SMCL [38]	67.39	92.87	61.78	68.84	96.55	75.56	83.93	-	79.83	83.05	-	78.57
ADP [39]	69.88	95.71	66.89	76.26	97.88	80.37	85.03	95.49	79.14	84.75	95.33	77.82
MPANet [40]	70.58	96.21	68.24	76.64	98.21	80.95	82.80	-	80.70	83.70	-	80.90
PIC [41]	57.50	-	55.10	60.40	-	67.70	83.60	-	79.60	79.50	-	77.40
MID [42]	60.27	92.90	59.40	64.86	96.12	70.12	87.45	95.73	84.85	84.29	93.44	81.41
SPOT [43]	65.34	92.73	62.25	69.42	96.22	74.63	80.35	93.48	72.46	79.37	92.79	72.26
FMCNet [44]	66.34	-	62.51	68.15	-	74.09	89.12	-	84.43	88.38	-	83.86
DART [45]	68.72	96.39	66.29	72.52	97.84	78.17	83.60	-	75.67	81.97	-	73.78
MMN [46]	70.60	96.20	66.90	76.20	97.20	79.60	91.60	97.70	84.10	87.50	96.00	80.50
DCLNet [47]	70.97	-	65.18	73.51	-	76.80	81.20	-	74.30	78.00	-	70.60
CIFT† [48]	71.77	-	67.64	78.65	-	82.11	92.17	-	86.96	90.12	-	84.41
MAUM [49]	71.68	-	68.79	76.97	-	81.94	87.87	-	85.09	86.95	-	84.34
DEEN [50]	74.70	97.60	71.80	80.30	99.00	83.30	91.10	97.80	85.10	89.50	96.80	83.40
SSRL(ours)	72.68	96.42	68.28	77.58	98.58	79.50	93.64	98.70	82.55	93.52	98.66	82.43

phase, infrared images were used to search for visible images. Samples from the visible camera are used in the gallery set, and samples from the infrared camera are used in the probe set. This dataset contains two modes: all-search mode and indoor-search mode. For all-search mode, visible cameras 1, 2, 4, and 5 are used for the gallery set, and infrared cameras 3 and 6 are used for the probe set. For indoor search mode, use visible cameras 1 and 2 for the gallery set and infrared cameras 3 and 6 for the probe. A detailed description of the experimental settings can be found in [7].

RegDB dataset contains 412 person identities, each with 10 visible light and 10 infrared images, a total of 4120 visible images and 4120 infrared images. The 10 images of each individual vary in body pose, capture distance, and lighting conditions. However, in the 10 images of the same person, the camera's weather conditions, viewing angles, and shooting angles are all the same, and the pose of the same identity varies little, so the task of the RegDB dataset is less complicated. Following the evaluation protocol proposed by [22], we randomly sampled 206 identities for training, and the remaining 206 identities were used for testing. The training/test segmentation process is repeated 10 times.

LLCM dataset utilizes a 9-camera network deployed in low-light environments and contains 46,767 bounding boxes containing 1,064 identities. The training set contains 30921 bounding boxes of 713 identities (16946 bounding boxes from the VIS modality, 13975 bounding boxes from the IR modality), and the test set contains 13909 bounding boxes of 351 identities (8680 bounding boxes from the VIS modality, 7166 bounding boxes from the IR modality.)

Evaluation metrics. For performance evaluation, we employed the widely known Cumulative Matching Characteristic (CMC) [52] curve and mean Average Precision (mAP).

4.2 Implementation details

The proposed method is implemented by the PyTorch framework and an NVIDIA Tesla V100 GPU. Building on existing person Re-ID work, a pre-trained ResNet-50 model is used as the backbone for a fair comparison. Specifically, the stride of the last convolutional block is changed from 2 to 1 to obtain fine-grained feature maps. In the training phase, batch-size is set to 32, containing 16 visible light and 16 infrared images from 8 identities. For each identity, 2 visible and 2 infrared images are selected randomly. For infrared images, three copied channels are fed into the network. The input images are resized to 288 × 144 and padded with 10, then randomly left-right flipped and cropped to 288 × 144

Table 2 Comparison with the state-of-the-arts on LLCM dataset. R1, R10 denotes the Rank-1, Rank-10 accuracy. Rank-k accuracy (%) and mAP (%) are reported. Herein, the best and second best results are indicated by red and green fonts. (The symbol of "*" represents the methods that we reproduced with the random erasing technique).

Method	LLCM					
	IR to VIS			VIS to IR		
	R1	R10	mAP	R1	R10	mAP
DDAG [11]	40.3	71.4	48.4	48.0	79.2	52.3
DDAG* [11]	41.0	73.4	49.6	48.5	81.0	53.0
AGW [1]	43.6	74.6	51.8	51.5	81.5	55.3
LbA [53]	43.8	78.2	53.1	50.8	84.3	55.6
LbA* [53]	44.6	78.2	53.8	50.8	84.6	55.9
AGW* [1]	46.4	77.8	54.8	56.0	84.9	59.1
CAJ [39]	48.8	79.5	56.6	56.5	85.3	59.8
DART [45]	52.2	80.7	59.8	60.4	87.1	63.2
MMN [46]	52.5	81.6	58.9	59.9	88.5	62.7
DEEN [50]	54.9	84.9	62.9	62.5	90.3	65.8
SRRL(ours)	52.3	81.2	58.8	60.1	87.8	62.6

and randomly erased [54] for data augmentation. We use the stochastic gradient descent (SGD) optimizer for optimization with the momentum parameter set to 0.9. We set the initial learning rate of the two datasets to 0.1 and guide the network using a warmup learning rate strategy [55]. For the λ_1, λ_2 and λ_3 in Eq. (8), we set them to 1.0, 1.0 and 2.0, respectively.

4.3 Comparison with State-of-the-art Methods

Results on SYSU-MM01: Tab. 1 illustrates the comparison results on the SYSU-MM01. It can be seen that the proposed method has reached the state-of-the-art result in two settings. Our model achieves 72.68% rank-1 and 68.28% mAP in all search setting and achieves 77.58% rank-1 and 79.50% mAP in indoor setting. Most of the indicators reach the second accuracy.

Results on RegDB: Tab. 1 illustrates the comparison results on the RegDB. Our SSRL achieves superior performance on both visible-to-infrared and infrared-to-visible settings. Specifically, we achieve Rank-1 accuracy of 93.64% and mAP of 82.55% in infrared to visible mode, and Rank-1 accuracy of 93.52% and mAP of 82.43% in visible to infrared mode. Hence, our SSRL model is robust against different datasets and query settings.

Results on LLCM: Tab. 2 illustrates the comparison results on the LLCM. The images in the LLCM dataset are captured in complex low-light environments, which contain severe illumination variations. Our SSRL does not have measures for low-light environments, but also achieves decent performance.

Table 3 The Ablation study of different components on SYSU-MM01 (*all search*). Base: baseline, DB: dual-branch structure, IAL: intra-class aggregationg learning strategy, ISL: intra-class separation learning strategy. Rank-1 accuracy(%), Rank-10 accuracy(%) and mAP(%) are reported.

Base	DB	IAL	ISL	SYSU-MM01		
				R=1	R=10	mAP
✓	-	-	-	58.69	91.64	53.97
✓	✓	-	-	68.81	94.64	64.86
✓	✓	✓	-	71.42	95.85	67.22
✓	✓	-	✓	69.60	96.37	65.68
✓	✓	✓	✓	72.68	96.42	68.28

Table 4 Analysis of the effectiveness of Dual-Branch Structure on SYSU-MM01 datasets under the *all search* mode. Specific Branch: Use only specific branch, Shared Branch: Use only shared branch, Dual-Branch: Use both specific branch and shared branch. Rank-1 accuracy(%), Rank-10 accuracy(%) and mAP(%) are reported.

Networks	SYSU-MM01		
	R=1	R=10	mAP
Specific Branch	58.69	91.64	53.97
Shared Branch	46.23	87.98	41.09
Dual-Branch	68.81	94.64	64.86

Table 5 Comparison of different augmented modalities on SYSU-MM01 dataset under the all-search setting. Rank-1 accuracy(%), Rank-10 accuracy(%) and mAP(%) are reported.

Method	SYSU-MM01		
	R=1	R=10	mAP
Base	58.69	91.64	53.97
Base + X modality	61.61	94.24	58.39
Base + Gray modality	62.98	93.45	59.90
Ours	68.81	94.64	64.86

Table 6 Comparison of our model with the baseline method in FLOPs and Params.

Model	FLOPs	Params
Base	10.36	54.53 M
Ours	20.72	84.33 M

4.4 Ablation Study

Effectiveness of each component. We evaluate the performance of each component on SYSU-MM01 datasets to verify the effectiveness of each component of SSRL. The ablation experiment is conducted on SYSU-MM01 datasets in the all-search single-shot mode. The results are demonstrated in Tab. 3. Compared with the baseline model, the dual-branch structure improves the Rank-1 accuracy and mAP by 10.12% and 10.89%. The reason for the improvement is mainly because the dual-branch structure fully extracts specific features and shared features. After intra-class aggregation and inter-class separation learning strategy, the performance is greatly improved by 3.87% and up to 72.68% Rank-1.

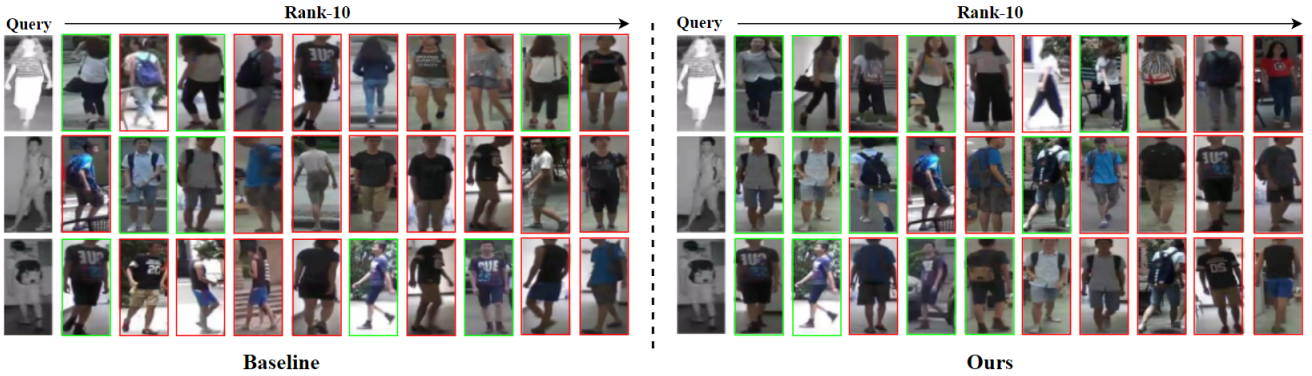

Fig. 5 The Rank-10 retrieval results were obtained by the baseline and the proposed SSRL model on the SYSU-MM01 datasets. For each retrieval case, the query images of the first column are the NIR images, and the gallery images are the VIS images. The retrieved VIS images with green bounding boxes have the same identities with the query images, and those with red bounding boxes have different identities with the query images.

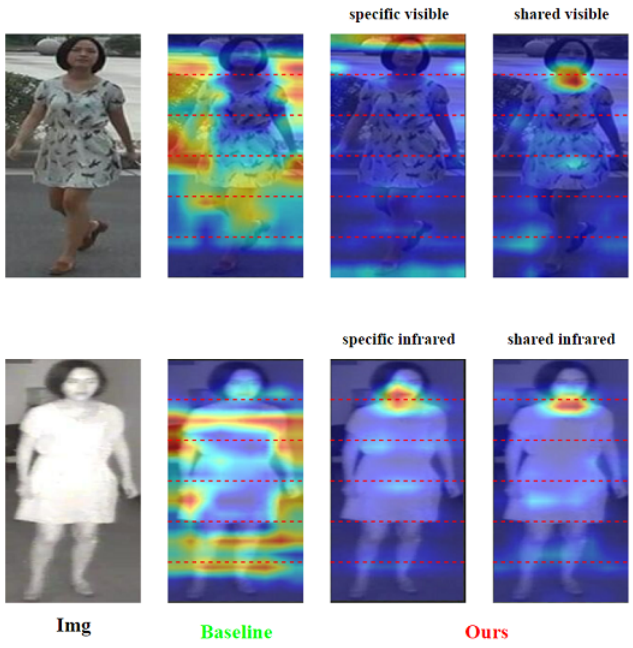

Fig. 6 Visualization of Feature Response Maps within baseline and the proposed SSRL model. For each example, the 1st images respectively show the RGB and NIR images, the 2nd images are the baseline corresponding feature response maps, while the 3rd and the 4th images are specific branch and shared branch corresponding feature response maps.

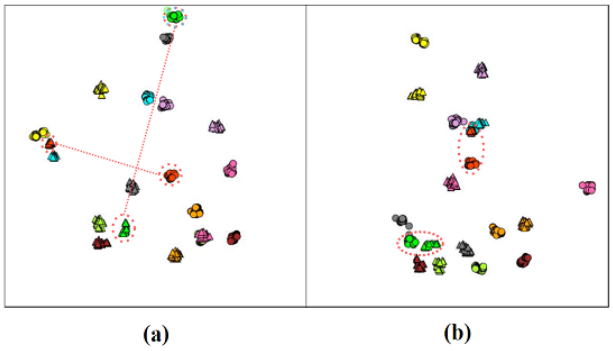

Fig. 7 Visualization of learned features, where each color represents an identity in the testing set. The circles and triangles indicate the features extracted from the visible and infrared modalities. A total of 10 persons are selected from the test set. The samples with the same color are from the same person. (a) Features extracted by our SSRL model excluding the IAL loss and ISL loss. (b) Features extracted by our SSRL model.

Effectiveness of dual-branch structure. We evaluate the performance of the dual-branch structure to verify that the dual-branch structure is better than the single-branch structure. The ablation experiment is conducted on SYSU-MM01 dataset in the all-search single-shot mode. For a fair comparison, we keep other structures the same and only use $\mathcal{L}_{id}$ and $\mathcal{L}_{Hc-Tri}$ in the training phase. The results are demonstrated in Tab. 4. It can be seen that accuracy is inferior when only using a shared branch or specific branch, but outstanding performance gains can be achieved when the specific and shared features of the dual-branch structure are fully utilized.

Comparison of Different Augmented Modalities. To verify the superiority of the dual-branch structure augmented, we compare it with other existing modality augmentation strategies, including X modality [31], and Grayscale modality [56]. The ablation experiment is conducted on SYSU-MM01 datasets in the all-search single-shot mode. The results are demonstrated in Tab. 5. Our dual-branch modality augmentation method outperforms other augmentation methods.

4.5 Visualization

Retrieval result. We compare the SSRL approach with the baseline on the SYSU-MM01 dataset using the single-shot setting and the all-search mode in order to

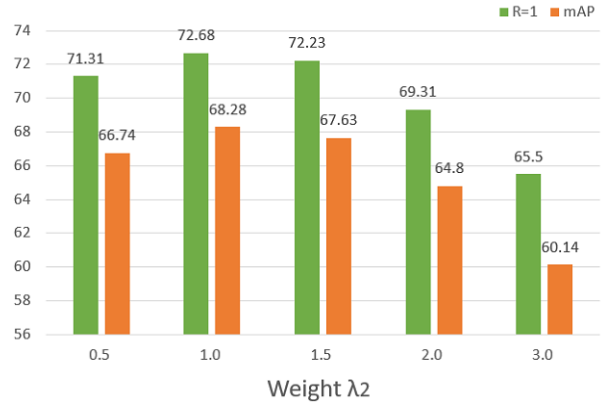

Fig. 8 The effect of parameter λ_2 on SYSU-MM01 datasets under the all-search mode. Rank-1 and mAP (%) are reported.

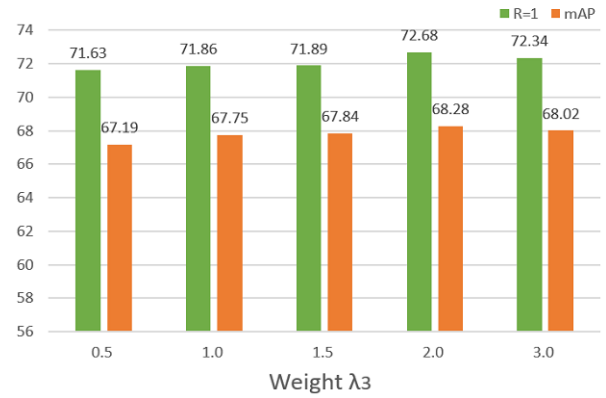

Fig. 9 The effect of parameter λ_3 on SYSU-MM01 datasets under the all-search mode. Rank-1 and mAP (%) are reported.

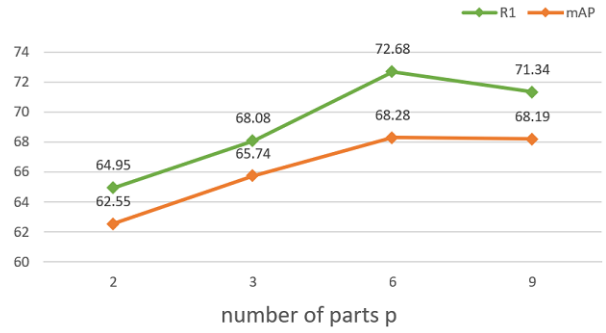

Fig. 10 The effect of partition strips p on SYSU-MM01 datasets under the all-search setting. Rank-1 and mAP (%) are reported.

further highlight the advantages of our suggested SSRL model. The results of the acquired Rank-10 ranking are displayed in Fig. 5. In general, the ranking list can be greatly improved by the proposed SSRL method, with more accurately recovered images placed in the top spots.

Attention to patterns. One of the key goals to the VI-REID task is to improve the discriminability of features. We visualize the pixel-level pattern mapping learnt by SSRL to illustrate further that it can learn modality-specific and modality-invariant features. We apply Grad-Cam [57] to visualize these areas by highlighting them on the image. Since our approach uses the PCB method, we separate the feature map into 6 sections, as shown in Fig. 6. In the RGB/IR modality, comparing each part of the segmentation, the baseline model focuses only on some unconsidered areas. By contrast, our SSRL model focuses on specific and shared information separately through a dual-branch structure.

Feature distribution. To further analyze the effectiveness of IAL loss and ISL loss, we use t-SNE [58] to transform high-dimensional features vectors into two-dimensional vectors. As shown in Fig. 7, compared to the visualization results of (a), the features extracted from SSRL including IAL loss and ISL loss are better clustered together. The distance between the centers and boundaries among different identities are more obvious, verifying that our work is more discriminating.

4.6 Parameters Analysis

The proposed SSRL involves two key parameters, including intra-class aggregation loss weight λ_2 and inter-class separation loss weight λ_3. As seen in Fig. 8 and Fig. 9, the two parameters are investigated by setting them to various values. Setting λ_2 to 1.0 and λ_3 to 2.0 achieve the best performance for $\mathcal{L}_{IAL}$ and $\mathcal{L}_{ISL}$ respectively. Then, since our method adopts the PCB method, we analyze the performance of our SSRL model with a different number of parts p. P represents the number of blocks being sliced. Performance is optimum when p is 6, as illustrated in Fig. 10. The proposed SSRL network includes specific-branch and shared-branche, so we compared it with the baseline in FLOPs and Params. As shown in Tab. 6, the amount of calculations and parameters has increased compared with the baseline, but it can be seen from Tab. 3 that there is a huge improvement in accuracy.

5 Conclusion

In this paper, we propose a novel specific and shared representations learning (SSRL) model. It consists of a shared branch to learn modality-invariant representations based on bridging the gap at the image level, and a specific branch to learn modality-specific representations under the premise of retaining the discriminative information of visible light images. Through intra-class aggregation and inter-class separation learning, which

reduces the intra-class distance and increases the inter-class distance, the distribution of feature embeddings at the fine-grained level is optimized. Comprehensive experiments on two VI-REID datasets demonstrate that the proposed method performs superior to the state-of-the-art methods.

Table 7 List of abbreviations.

Abbreviations	Definition
VI-ReID	visible-infrared person re-identification
SSRL	specific and shared representations learning
Re-ID	re-identification
ID	identity loss
HCT	hetero-center triplet
IAL	intra-class aggregation learning
ISL	inter-class separation learning
mAP	mean Average Precision

Declarations

Availability of data and material All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

Authors' Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Aihua Zheng, Juncong Liu and Zi Wang. The first draft of the manuscript was written by Juncong Liu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work is supported by the National Key R&D Program of China (2022ZD0160605), the National Natural Science Foundation of China (61976002), the University Synergy Innovation Program of Anhui Province (GXXT-2022-036), the Natural Science Foundation of Anhui Province (No. 2208085J18), and the Natural Science Foundation of Anhui Higher Education Institution (No. 2022AH040014).

Acknowledgements I would like to express my gratitude to several individuals who have supported me throughout the process of completing this paper. Firstly, I thank my academic supervisor, Aihua Zheng, for her invaluable guidance, feedback, and encouragement. I am also grateful to Anhui Provincial Key Laboratory of Multimodal Cognitive Computation for providing me with the necessary resources and facilities. Finally, I appreciate all the participants who generously gave their time and shared their experiences for the purpose of this study.

References

1. Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.
2. Song Bai, Peng Tang, Philip HS Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 740–749, 2019.
3. Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9021–9030, 2020.
4. Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6639–6648, 2019.
5. Meng Liu, Leigang Qu, Liqiang Nie, Maofu Liu, Lingyu Duan, and Baoquan Chen. Iterative local-global collaboration learning towards one-shot video person re-identification. *IEEE Transactions on Image Processing*, 29:9360–9372, 2020.
6. Asmat Zahra, Nazia Perwaiz, Muhammad Shahzad, and Muhammad Moazam Fraz. Person re-identification: A retrospective on domain specific open challenges and future trends. *Pattern Recognition*, page 109669, 2023.
7. Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.
8. Haojie Liu, Daoxun Xia, Wei Jiang, and Chao Xu. Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification. *arXiv preprint arXiv:2204.04842*, 2022.
9. Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019.
10. Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12144–12151, 2020.
11. Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020.
12. Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, 128(6):1765–1785, 2020.
13. Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389, 2020.

14. Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152, 2020.
15. Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020.
16. Jiahuan Zhou, Bing Su, and Ying Wu. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2909–2918, 2020.
17. Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019.
18. Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018.
19. Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018.
20. Zhongwei Zhao, Ran Song, Qian Zhang, Peng Duan, and Youmei Zhang. Jot-gan: A framework for jointly training gan and person re-identification model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–18, 2022.
21. Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16403–16412, 2021.
22. Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
23. Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, volume 1, page 2, 2018.
24. Yuanxin Zhu, Zhao Yang, Li Wang, Sai Zhao, Xiao Hu, and Dapeng Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 386:97–109, 2020.
25. Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification. *IEEE Transactions on Multimedia*, 23:4414–4425, 2020.
26. Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8385–8392, 2019.
27. Haijun Liu, Jian Cheng, Wen Wang, Yanzhou Su, and Haiwei Bai. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing*, 398:11–19, 2020.
28. Yun-Bo Zhao, Jian-Wu Lin, Qi Xuan, and Xugang Xi. Hpiln: a feature learning framework for cross-modality person re-identification. *IET Image Processing*, 13(14):2897–2904, 2019.
29. Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 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*, pages 10257–10266, 2020.
30. Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019.
31. Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020.
32. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
33. Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018.
34. Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.
35. Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 6, 2018.
36. Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 587–597, 2021.
37. Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu, Yidong Li, and Congyan Lang. Mso: Multi-feature space joint optimization network for rgb-infrared person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5257–5265, 2021.
38. Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 225–234, 2021.
39. Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021.
40. Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021.
41. Xiangtao Zheng, Xiumei Chen, and Xiaoqiang Lu. Visible-infrared person re-identification via partially interactive collaboration. *IEEE Transactions on Image Processing*, 31:6951–6963, 2022.
42. Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1034–1042, 2022.
43. Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022.
44. Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022.
45. Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14308–14317, 2022.
46. Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 788–796, 2021.
47. Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Not all pixels are matched: Dense contrastive learning for cross-modality person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5333–5341, 2022.
48. Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 381–398. Springer, 2022.
49. Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19366–19375, 2022.
50. Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023.
51. Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.
52. Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.
53. Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12046–12055, 2021.
54. Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.
55. Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.
56. Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020.
57. Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
58. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.