IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDP7087

Publicized:2024/08/26

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Partial Enhancement and Channel Aggregation for Visible-Infrared Person Re-Identification

Weiwei JING[†] *and* **Zhonghua LI**[††*], *Nonmembers*

SUMMARY Visible-infrared person re-identification (VI-ReID) aims to achieve cross-modality matching between the visible and infrared modalities, thus enabling usage in all-day monitoring scenarios. Existing VI-ReID methods have indeed achieved promising performance by considering the global information for identity-related discriminative learning. However, they often overlook the importance of local information, which can contribute significantly to learning identity-specific discriminative cues. Moreover, the substantial modality gap typically poses challenges during the model training process. In response to the aforementioned issues, we propose a VI-ReID method called partial enhancement and channel aggregation (PECA) and make efforts in the following three aspects. Firstly, to capture local information, we introduce the global-local similarity learning (GSL) module, which compels the encoder to focus on fine-grained details by increasing the similarity between global and local features within various feature spaces. Secondly, to address the modality gap, we propose an inter-modality channel aggregation learning (ICAL) approach, which progressively guides the learning of modality-invariant features. ICAL not only progressively alleviates modality gap but also augments the training data. Additionally, we introduce a novel instance-modality contrastive loss, which facilitates the learning of modality-invariant and identity-related features at both the instance and modality levels. Extensive experiments on the SYSU-MM01 and RegDB datasets have shown that PECA outperforms state-of-the-art methods.

key words: *computer vision, person re-identification, cross-modality, contrastive learning*

1. Introduction

Traditional person re-identification (ReID) methods [1]–[5] aim to connect images of the same person captured by different cameras, essentially constituting a single-modality image retrieval task. However, in practical applications, a significant number of dark scenarios are encountered, where surveillance cameras capture numerous infrared images [6], [7]. In light of this, researchers have shifted their focus towards visible-infrared person re-identification (VI-ReID) [8]–[10]. In contrast to single-modality ReID approaches, VI-ReID endeavors to perform matching between visible and infrared modalities. Specifically, given a visible (infrared) query image, VI-ReID seeks to find images in the gallery composed of infrared (visible) images that share the same identity as the query image. As illustrated in Figure 1, the substantial gap between the visible and infrared modalities pose significant intra-class variations to VI-ReID [11], [12].

Currently, VI-ReID methods primarily focus on mitigating the impact of modality gap between visible and infrared images while learning modality-invariant and identity-related feature representations [13], [14]. For instance, CAJ [13] approaches the problem from a data processing perspective and enhances the model's robustness to modality gap by employing techniques such as random channel exchangeable augmentation, random channel erasing, random grayscale augmentation, and horizontal flip operation augmentation. FMCNet [14] addresses the modality gap from a feature-based perspective. Specifically, it first employs a single-modality feature decomposition module to decompose the single-modality features into modality-specific and modality-shared features. Subsequently, a feature-level modality compensation module is used to generate modality-specific features from the modality-shared features. Finally, a shared-specific feature fusion module combines the generated features with existing features. Although existing methods have made significant progress, there is still ample room for improvement in their performance.

In this paper, we propose a method called partial enhancement and channel aggregation (PECA) to address the VI-ReID task, which makes efforts in the following three aspects. Firstly, existing VI-ReID methods typically focus solely on global features [13]–[16], while disregarding local features. However, as illustrated in Figure 1, local features often contain important discriminative information that can help model learn identity-related features and bridge the modality gap. To address this, we introduce a global-local similarity learning (GSL) module, which aims to learn global features while capturing abundant local information. Specifically, GSL first maps global features into various feature spaces and then compels the encoder to attend to fine-grained details by increasing the similarity between multiple global features and multiple local features.

Secondly, in response to modality gap, existing methods [15], [17], [18] usually design a two-stream network to separately extract features from the visible and infrared modalities. Subsequently, a shared encoder is employed to extract modality-shared features. However, the substantial modality gap impedes the learning of inter-modality similarities, making the model optimization challenging. To address this issue, we propose an inter-modality channel aggregation learning (ICAL) approach to facilitate modality-invariant feature learning. ICAL first generates an auxiliary modality and then progressively guides the model to bridge the modality gap by reducing the disparities between the real modality and the auxiliary modality separately. This approach offers dual

[†]The author is with the School of Semiconductor and Physics, North University of China, Taiyuan, 030051, Shanxi, China

[††]The author is with the School of Sport And Physical Education, North University of China, Taiyuan, 030051, Shanxi, China

[*]Corresponding author (E-mail: lzh228829@163.com)

Fig. 1: Each column of images has the same identity. (a) There is a significant gap between visible modality and infrared modality. (b) Several images with different identities only show differences in local areas (circled in red).

advantages: on one hand, the progressive learning process from easy to difficult is more manageable and intuitive; on the other hand, the auxiliary modality serves as data augmentation, thereby further enhancing the model's performance. Moreover, ICAL exhibits a high level of versatility, as it can be readily integrated into any existing VI-ReID method.

Additionally, we introduce a novel instance-modality contrastive loss. The widely used triplet loss [15], [18], [19] typically focuses solely on the similarity between instances while neglecting modality-level similarity. In contrast to triplet loss, the instance-modality contrastive loss not only enhances the positive similarity and reduces the negative similarity at the instance-level but also guides the model to extract modality-invariant and identity-related features at the modality-level. Furthermore, this approach offers a broader optimization scope compared to triplet loss.

Our main contributions are summarized below:

- We propose a global-local similarity learning (GSL) module, which compels the encoder to focus on fine-grained details by increasing the similarity between global and local features within multiple feature spaces.
- To enhance the model's robustness to modality gap, we propose an inter-modality channel aggregation learning (ICAL) algorithm to generate an auxiliary modality, progressively guiding the learning of modality-invariant features.
- We introduce an instance-modality contrastive loss, which aims to learn modality-invariant and identity-related features at both the instance and modality levels.
- We conduct extensive experiments on two cross-modality datasets, and the experimental results demonstrate the effectiveness of each component and the superiority of the proposed method.

2. Related work

2.1 Visible-infrared person re-identification

As VI-ReID demonstrates potential in practical applications, it is increasingly garnering attention. Wu et al. [6] proposed the first VI-ReID method, which handles modality gap from the channel perspective and uses a single-stream network to extract modality-shared features. Subsequent research efforts, such as TONE+HCML [20] and BDTR [18], adopt dual-stream networks to simultaneously consider modality-specific and modality-shared features, incorporating contrastive losses to guide the learning process. Inspired by GAN [21], [22], cmGAN [23] employs a generative adversarial approach to encourage the encoder to extract modality-invariant feature representations. MAC [24] introduces a collaborative learning scheme to regularize both modality-shared and modality-specific classifiers. This collaborative learning idea has also been proven effective in subsequent studies [25]. DDAG [26] simultaneously explores intra-modality part-level and cross-modality graph-level contextual cues to learn discriminative features. To align the feature distributions of the two modalities and learn modality-invariant features, JSIA [27] generates cross-modality paired-images and performs both global set-level and fine-grained instance-level alignments. Subsequent research focuses on dealing with modality-shared and modality-specific features separately. For example, cm-SSFT [19] models the affinities of different modality samples based on shared features and propagates both shared and specific features between modalities. Hi-CMD [16] decouples identity-related and identity-unrelated features from the images and employs only identity-related features for identification.

In recent works, researchers have introduced various strategies in the VI-ReID field. For instance, Chen et al. propose the neural feature search [28] to achieve automated feature selection in VI-ReID. Tian et al. present the variational self-distillation [29] strategy to fit mutual information, allowing the information bottleneck to capture the intrinsic correlation between features and labels. To mitigate the impact of modality gap, Ye et al. design the channel exchangeable augmentation strategy [13], which generates color-independent images by randomly swapping color channels. This method can be integrated into existing augmentation operations to continually improve the model's robustness to color variations. Huang et al. propose a method called modality adaptive mixup and invariant decomposition (MID) [30], which generates mixed images between visible and infrared modalities to alleviate modality gap. To learn certain modality-specific information related to humans, FMCNet [14] prompts the model to generate modality-specific features from modality-shared features and utilizes a shared-specific feature fusion module to combine the existing features with the generated features. Existing methods have achieved promising performance but often focus solely on global features, neglecting to capture fine-grained details. In contrast, our approach not only emphasizes local information but also integrates the two modalities from the channel perspective to gradually bridge the modality gap.

2.2 Memory-based learning

Memory-based learning has been widely applied in unsuper-

vised learning [31]–[34]. For instance, Moco [32] utilizes memory to increase the quantity of stored keys for improved contrastive learning. XBM [35] leverages historical features in memory for enhanced hard mining. In the context of single-modality unsupervised ReID, Wang et al. [33] employ memory to store camera centroids and assign positive and negative camera centroids to each sample, thus alleviating the impact of camera variations during the optimization process. Pang et al. [34] use memory to store cluster centers and utilize all centers in memory as reference points to estimate the reliability of pseudo-labels. In contrast to the aforementioned methods, we employ memory to address modality gap. Specifically, we store all modality centroids in a memory, which serves as guidance for subsequent optimization. Additionally, we update the modality centroids during the optimization process.

3. Partial enhancement and channel aggregation

For VI-ReID, during the training phase, we are given a visible image set $\{x_i^v, y_i^v\}_{i=1}^N$ and an infrared image set $\{x_i^{ir}, y_i^{ir}\}_{i=1}^M$, where x_i^v and x_i^{ir} represent samples from the visible and infrared image sets respectively, and y_i^v and y_i^{ir} denote their corresponding identity labels. N and M are the numbers of images contained in the two sets, respectively. For ease of description, we will no longer distinguish between samples, images, and instances. They all refer to cross-modality data.

The overall framework of our proposed partial enhancement and channel aggregation (PECA) is shown in Figure 2. PECA consists of three novel components: global-local similarity learning (GSL), inter-modality channel aggregation learning (ICAL), and instance-modality contrastive loss L_{imcl}. GSL guides the model to focus on fine-grained details, ICAL aims to progressively enhance the model's robustness to modality gap, and L_{imcl} encourages the model to extract modality-invariant and identity-related features. In the subsequent sections, we introduce each of these components separately. We use CNN models to extract features f_i^v and f_i^{ir} from x_i^v and x_i^{ir} respectively. In the subsequent sections, we refer to f_i^v and f_i^{ir} collectively as f_i.

During the inference phase, we are provided with a query set and a gallery. For each query image in the query set, we rank all images in the gallery in descending order based on the similarity of their features to the features of the query image. In the re-identification process, we consider that images ranked higher in the sequence are more likely to share the same identity with the query image.

3.1 Global-local similarity learning

Existing methods often focus solely on global features [13]–[16] while neglecting local discriminative information. To fully leverage local information to aid identification, we propose global-local similarity learning (GSL) to encourage the model to extract global features that are rich in local information. As shown in Figure 2, for each visible or infrared image, we first extract the features of the overall image and the local image respectively to obtain a global feature f_i and three local features (f_i^u, f_i^m and f_i^l). Where f_i^u, f_i^m and f_i^l represent the upper, middle and lower features of the image respectively. Subsequently, we input the global feature f_i into three separate MLPs with non-shared parameters. We introduce three different MLPs with the purpose of mapping the global feature f_i into three distinct local features (upper, middle, and lower). Considering that these three local features serve different semantic purposes, we use three separate MLPs to accomplish these tasks tailored to their respective objectives. The outputs of the three MLPs are then concatenated to obtain the feature f_i^g. Meanwhile, we concatenate f_i^u, f_i^m, and f_i^l to obtain feature f_i^s. Finally, we applied a global-local similarity loss to f_i^g and f_i^s:

$$L_{sim} = -\frac{f_i^g}{||f_i^g||_2} \cdot stopgrad(\frac{f_i^s}{||f_i^s||_2}) \tag{1}$$

where $stopgrad$ is the stop gradient operation [36]. That is, the local features branch does not receive the gradient from L_{sim}. This approach has been validated to effectively prevent model collapse [36]. L_{sim} will make f_i^g approach f_i^s. When the global feature contains sufficient local information, the value of L_{sim} is small. Due to the fact that L_{sim} incorporates local information into the global features, we focus only on the global features in the subsequent optimization. In the following sections, all mentions of "features" refer to "global features".

3.2 Inter-modality channel aggregation learning

Existing methods typically directly align the visible and infrared modalities [15], [17], [18]. However, the model's initial performance is poor, and there are significant gaps between the two modalities, making the optimization process particularly challenging. To address this issue, we have designed inter-modality channel aggregation learning (ICAL) to gradually guide the model's optimization process.

Inspired by CAJ [13], we process the data from the channel perspective. However, unlike CAJ, we do not randomly select one channel from the visible image to replace the other channels. Instead, we randomly select one infrared image and replace one or two channels of a visible image with the randomly selected infrared image, provided that both images belong to the same identity. For example, for a visible image x_i^v and an infrared image x_i^{ir} from the same person, we can get a variety of enhanced images:

$$\begin{aligned}
x_i^{v,RG} &= (x_i^{v,r}, x_i^{v,g}, x_i^{ir}) \\
x_i^{v,RB} &= (x_i^{v,r}, x_i^{v,ir}, x_i^{v,b}) \\
x_i^{v,GB} &= (x_i^{v,ir}, x_i^{v,g}, x_i^{v,b}) \\
x_i^{v,R} &= (x_i^{v,r}, x_i^{ir}, x_i^{ir}) \\
x_i^{v,G} &= (x_i^{ir}, x_i^{v,g}, x_i^{ir}) \\
x_i^{v,B} &= (x_i^{ir}, x_i^{ir}, x_i^{v,b})
\end{aligned} \tag{2}$$

where $x_i^{v,r}$, $x_i^{v,g}$ and $x_i^{v,b}$ are the red, green and blue channels

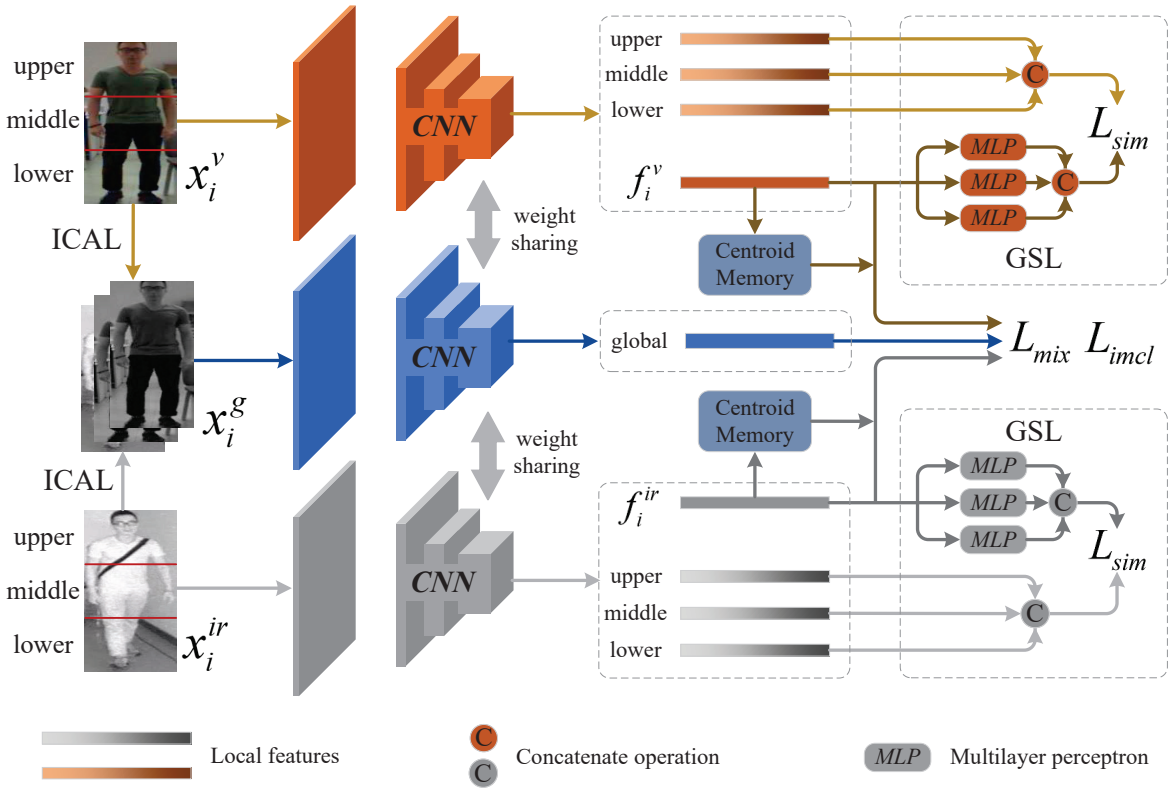


Fig. 2: The input data, from top to bottom, consists of visible images, ICAL-generated images, and infrared images. We have designed a three-stream encoder, where the first layer is responsible for learning modality-specific features, and the subsequent layers share their weights. The visible images and infrared images are further divided into upper, middle, and lower parts, which are also fed into the encoder. Additionally, we introduce a centroid memory to store modality centroids based on global features. In GSL, we use a multilayer perceptron (MLP) to map global features into different local feature spaces.

of the visible image respectively. From Eq. 2, it is evident that ICAL can easily be combined with CAJ [13] and other traditional data augmentation methods and integrated into any existing VI-ReID method. Based on the generated images mentioned above, we can progressively optimize the model:

$$L_{mix} = \max[||D(x_i^v, x_i^g) - m||_2, 0] + \max[||D(x_i^{ir}, x_i^g) - m||_2, 0] \tag{3}$$

where x_i^g represents any image generated by a visible image x_i^v and an infrared image x_i^{ir} from the same person in Eq. 2. m is the distance margin. $D(\cdot,\cdot)$ is the Euclidean distance function used to measure the distance between two sample features. To stabilize the convergence, all features are normalized.

ICAL has a dual advantage: on one hand, it can generate an auxiliary modality with minimal time cost. On the other hand, based on an auxiliary modality composed of generated images $\{x_i^g\}$ that lie between the visible and infrared modalities, ICAL encourages the model to produce similar features for real images and generated images from the same person, as specified in Eq. 3. In other words, ICAL progressively enhances the model's robustness to the modality gap by increasing the similarity between features from different modalities.

3.3 Instance-modality contrastive loss

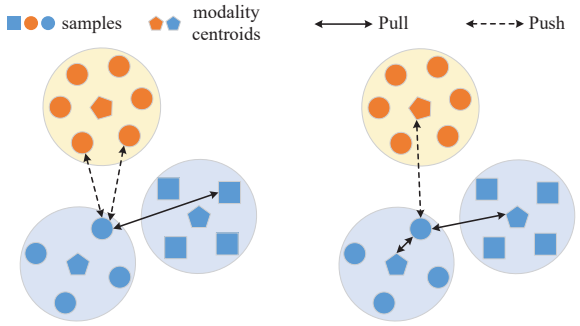
To learn identity related feature representations, existing VI-ReID methods [15], [37] typically use cross-entropy loss to optimize the model:

$$L_{ce} = -\frac{1}{n}\sum_{i=1}^{n}\log P(y_i^v|C(E(x_i^v))) - \frac{1}{m}\sum_{j=1}^{m}\log P(y_j^{ir}|C(E(x_j^{ir}))) \tag{4}$$

where E represents encoding operation, and C represents classification layer. y_i^v and y_j^{ir} represent the identity labels corresponding to samples x_i^v and x_j^{ir}, respectively. In addition, triplet loss [15], [18], [19] is also widely used to guide the encoder to extract discriminable features. Unlike triplet loss, we design an instance-modality contrastive loss in this paper:

$$L_{imcl} = L_{ins} + \lambda_m L_{mod} \tag{5}$$

where L_{imcl} includes two parts: instance-similarity contrastive loss L_{ins} and modality-similarity contrastive loss L_{mod}, and λ_m represents the weight of L_{mod}. L_{imcl} has a



(a) L_{ins} (b) L_{mod}

Fig. 3: Illustration of L_{ins} and L_{mod}. Different colors indicate different identities, and different shapes indicate different modalities.

wider optimization range compared to triple loss.

As shown in Figure 3, the purpose of the instance-similarity contrastive loss L_{ins} is not only to handle hard samples but also to increase intra-class similarity and inter-class separability. For any sample x_i from any modality, we refer to the sample with the same identity as x_i and having the farthest distance as the hard positive sample of x_i. Additionally, we define a set Q of $|Q|$ samples that have different identities from x_i and are closest to x_i as the hard negative set of x_i. For sample x_i, L_{ins} is defined as:

$$L_{ins} = E[-\log \frac{\exp(f_i^T \cdot f_j/\tau_{ins})}{\sum_{f_k \in f_j \cup Q} \exp(f_i^T \cdot f_k/\tau_{ins})}] \tag{6}$$

where f_i is the feature of sample x_i, f_j is the feature of hard positive sample x_j of x_i, and τ_{ins} is the temperature hyper-parameter [38] of L_{ins}. L_{ins} aims to increase the similarity between the features of x_i and x_j, and reduce the similarity between the features of x_i and the features of all samples in the set Q.

As shown in Figure 3, in addition to the instance-similarity contrastive loss L_{ins}, we also introduce the modality-similarity contrastive loss L_{mod} to impose constraints on the model at the modality-level. For each identity, we calculate all modality centroids within that identity. For example, the centroid of the visible modality of the i-th identity is defined as:

$$m_i^v = \frac{1}{n_i^v} \sum_{i=1}^{n_i^v} f_i^v \tag{7}$$

where f_i^v is the feature of any sample in the visible modality of the i-th identity, and n_i^v is the number of samples in the visible modality of the i-th identity. Similarly, the centroid of the infrared modality of the i-th identity is defined as:

$$m_i^{ir} = \frac{1}{n_i^{ir}} \sum_{i=1}^{n_i^{ir}} f_i^{ir} \tag{8}$$

where f_i^{ir} is the feature of any sample in the infrared modality of the i-th identity, and n_i^{ir} is the number of samples in the infrared modality of the i-th identity. As shown in Figure 2, we use centroid memory to store all modality centroids and update them during the optimization process. For example, for any input image x_i from any modality, we update its modality centroid through moving average:

$$m_i \leftarrow (1-\alpha)m_i + \alpha f_i \tag{9}$$

where f_i is the feature of sample x_i, and m_i is the modality centroid to which x_i belongs, and α is an updating rate.

For any sample x_i from any modality, we refer to all modality centroids with the same identity label as x_i as the positive modality centroid set P, and all modality centroids with the same modality label but different identity labels as x_i as the negative modality centroid set U. For sample x_i, the modality-similarity contrastive loss L_{mod} is defined as:

$$L_{mod} = E[-\frac{1}{|P|} \sum_{m_i \in P} \log \frac{\exp(f_i^T \cdot m_i/\tau_{mod})}{\sum_{m_k \in P \cup U} \exp(f_i^T \cdot m_k/\tau_{mod})}] \tag{10}$$

where f_i is the feature of sample x_i, and $|P|$ is the number of modality centroids contained in set P. τ_{mod} is the temperature hyper-parameter [38] of L_{mod}. L_{mod} aims to reduce the impact of modality gap and further enhance the intra-class compactness of the learned feature representations.

Finally, the overall loss of PECA is defined as:

$$L_{PECA} = L_{ce} + L_{sim} + L_{mix} + L_{imcl} \tag{11}$$

4. Experiments

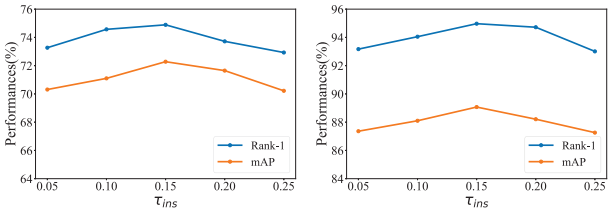
4.1 Experimental setting

4.1.1 Datasets and evaluation protocols

We evaluated the proposed method on the SYSU-MM01 [6] and RegDB [7] datasets.

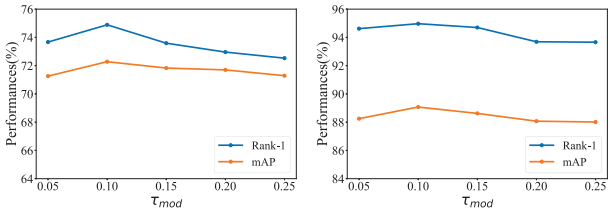
SYSU-MM01 is a challenging cross-modality dataset that includes images captured by 4 visible and 2 infrared cameras, containing data from 491 identities. The training set comprises 22,258 visible images and 11,909 infrared images from 395 identities. During the testing phase, this dataset offers two testing modes: all search and indoor search, with the former being more challenging compared to the latter.

RegDB is a small-scale cross-modality dataset comprising images captured by one visible and one thermal infrared camera, containing data from 412 identities. Each identity consists of 10 visible images and 10 thermal infrared images. Following the approach of previous methods [13], [20], we select images from 206 identities as the training set and use the remaining images as the test set. This partitioning is



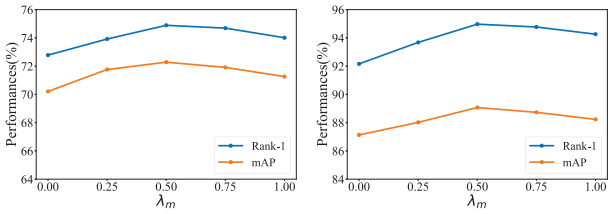
(a) SYSU-MM01(All Search) (b) RegDB (Visible to Infrared)

Fig. 4: Parameter analysis of τ_{ins}.



(a) SYSU-MM01(All Search) (b) RegDB (Visible to Infrared)

Fig. 5: Parameter analysis of τ_{mod}.



(a) SYSU-MM01(All Search) (b) RegDB (Visible to Infrared)

Fig. 6: Parameter analysis of λ_m.

repeated ten times, and the average performance across the ten repetitions is reported as the final performance.

We evaluate the performance using the cumulated matching characteristics (CMC) and the mean average precision (mAP).

4.1.2 Implementation details

In the data preprocessing stage, all images are resized to 288x144, and random flipping and cropping are applied to augment all images. For visible images, we utilize random channel exchangeable augmentation and random channel erasing techniques [13]. For infrared images, we input three replicated channels into the network. The encoder follows AGW [39] and employs a ResNet-50 [40] pretrained on ImageNet [41] as the backbone. MLP contains two fully connected layers, excluding BN and ReLU. We use stochastic gradient descent (SGD) optimizer to train the model and set the initial learning rate to 0.1. For L_{ins}, we set $|Q|$=50. We use the warm-up strategy [42] in the first 10 epochs. We set the distance margin m to 0.7, and set the updating rate α to 0.8. For each minibatch, we randomly select 8 identities and then sample 4 visible images and 4 infrared images from each identity. During the training phase, a total of 100 epochs are conducted. In the test phase, only the encoder is used for the inference.

4.2 Parameter analysis

In this subsection, we analyze the impact of hyper-parameters τ_{ins}, τ_{mod} and λ_m on performance on SYSU-MM01 [6] and RegDB [7] dataset. For SYSU-MM01, we analyze the performance under the all search mode, which is more challenging compared to the indoor search mode. For RegDB, we analyze the performance under the visible to infrared mode.

4.2.1 τ_{ins} and τ_{mod}

As shown in Figure 4 and Figure 5, we analyze the performance variation of the model for different values of τ_{ins} and τ_{mod} within the range of [0.05, 0.25]. We find that when τ_{ins} = 0.15 and τ_{mod} = 0.10, the model achieves optimal performance on both datasets. This verifies the generalization of τ_{ins} and τ_{mod}. The temperature hyperparameter primarily adjusts the flexibility of the probability distribution. When τ_{ins} and τ_{mod} take on either very high or very low values, the model performance is usually suboptimal. This is because a larger temperature hyperparameter makes the probability distribution too flat and increases uncertainty, while a smaller temperature hyperparameter makes the probability distribution too sharp, making the model overly confident. Additionally, since the modality-similarity contrastive loss aims to handle large modality gap, τ_{mod} should be slightly smaller than τ_{ins} to provide sufficiently high prediction values for the positive modality centroids with different modality labels from the features, thereby mitigating the impact of modality gap. Therefore, theoretically, the model can achieve optimal performance when τ_{ins} and τ_{mod} are approximately 0.15 and 0.10, respectively.

4.2.2 λ_m of L_{mod}

In Figure 6, we illustrate how the performance varies as λ_m is varied from 0 to 1 via plots of Rank-1 and mAP. We find that when λ_m = 0.50, the model achieves optimal performance. This is because setting λ_m close to 1 reduces the relative weight of L_{ins}, thereby affecting the model's ability to learn identity-related features. Conversely, when λ_m is set too low, L_{mod} cannot sufficiently contribute to the model's optimization, leaving the model significantly impacted by modality gap. When λ_m = 0 and L_{mod} are not effective, the model achieves the worst performance. This preliminarily verifies the effectiveness of L_{mod}.

4.3 Comparison with state-of-the-art methods

In this subsection, we compare our PECA with state-of-the-art VI-ReID methods on the SYSU-MM01 and RegDB

Table 1: Comparison to the state-of-the-art VI-ReID methods on the SYSU-MM01 dataset. The optimal and suboptimal performance are represented in bold and italic, respectively. "*" denotes results reproduced under our conditions. "-" indicates that the experimental result was not reported in the original paper.

Methods	Reference	All Search				Indoor Search			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad	ICCV17	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML	AAAI18	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
MSR	TIP19	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
Hi-CMD*	CVPR20	34.60	76.58	88.36	35.21	36.87	85.26	92.07	44.76
cm-SSFT	CVPR20	61.60	89.20	93.90	63.20	70.50	94.90	97.70	72.60
AGW	TPAMI21	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
MPANet	CVPR21	70.58	96.21	98.80	68.24	76.74	98.21	*99.57*	80.95
NFS	CVPR21	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
CAJ	ICCV21	69.88	95.71	98.46	66.89	76.26	97.88	99.49	80.37
CM-NAS	ICCV21	61.99	92.87	97.25	60.02	67.01	97.02	99.32	72.95
DTRM	TIFS22	63.03	93.82	97.56	58.63	66.35	95.58	98.80	71.76
MID	AAAI22	60.27	92.90	-	59.40	64.86	96.12	-	70.12
DART	CVPR22	68.72	96.39	*98.96*	66.29	72.52	97.84	99.46	78.17
FMCNet	CVPR22	66.34	-	-	62.51	68.15	-	-	74.09
MAUM	CVPR22	71.68	-	-	68.79	76.97	-	-	81.94
PMT	AAAI23	67.53	95.36	98.64	64.98	71.66	96.73	99.25	76.52
CAL	ICCV23	*74.66*	*96.47*	-	*71.73*	*79.69*	**98.93**	-	*83.68*
PECA	Ours	**74.89**	**96.57**	**99.03**	**72.28**	**79.96**	*98.90*	**99.78**	**83.96**

Table 2: Comparison to the state-of-the-art VI-ReID methods on the RegDB dataset. The optimal and suboptimal performance are represented in bold and italic, respectively. "*" denotes results reproduced under our conditions. "-" indicates that the experimental result was not reported in the original paper.

Methods	Reference	Visible to Infrared				Infrared to Visible			
		Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad	ICCV17	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML	AAAI18	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
MSR	TIP19	48.43	70.32	79.95	48.67	-	-	-	-
Hi-CMD*	CVPR20	68.96	86.03	90.87	65.89	70.12	87.23	90.64	64.31
cm-SSFT	CVPR20	72.30	-	-	72.90	71.00	-	-	71.70
AGW	TPAMI21	70.05	86.21	91.55	66.37	70.49	87.21	91.84	65.90
MPANet	CVPR21	83.70	-	-	80.90	82.80	-	-	80.70
NFS	CVPR21	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
CAJ	ICCV21	85.03	95.49	*97.54*	79.14	84.75	95.33	*97.51*	77.82
CM-NAS	ICCV21	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
DTRM	TIFS22	79.09	92.25	95.66	70.09	78.02	91.75	95.19	69.56
MID	AAAI22	87.45	95.73	-	84.85	84.29	93.44	-	81.41
DART*	CVPR22	82.53	92.16	96.05	74.93	80.25	92.09	95.23	72.19
FMCNet	CVPR22	89.12	-	-	84.43	88.38	-	-	83.86
MAUM	CVPR22	87.87	-	-	85.09	86.95	-	-	84.34
PMT*	AAAI23	83.69	93.66	96.89	75.82	83.27	93.25	96.28	75.02
CAL	ICCV23	*94.51*	**99.70**	-	*88.67*	**93.64**	**99.46**	-	*87.61*
PECA	Ours	**94.97**	*99.23*	**99.56**	**89.07**	*93.06*	*98.97*	**99.51**	**90.22**

datasets, including Zero-Pad [6], HCML [20], MSR [15], Hi-CMD [16], cm-SSFT [19], AGW [39], MPANet [37], NFS [28], CAJ [13], CM-NAS [43], DTRM [44], MID [30], DART [45], FMCNet [14], MAUM [46], PMT [47] and CAL [48].

As shown in Table 1, on the SYSU-MM01 dataset, MAUM [46] and CAL [48] have achieved superior performance. The comprehensive performance of our method surpasses the above methods on SYSU-MM01 dataset. Specifically, for Rank-1 and mAP, in the all search testing mode, PECA surpasses CAL by 0.23% and 0.55%, respectively, and in the indoor search testing mode, PECA outperforms CAL by 0.27% and 0.28%, respectively. This improvement is attributed to PECA's consideration of local discriminative information and the design of progressive data augmentation methods to address modality gap.

As shown in Table 2, on the RegDB dataset, FMCNet [14], MAUM [46] and CAL [48] achieve relatively superior performance. The performance of our proposed method, PECA, on the RegDB dataset significantly outperforms FMCNet [14] and MAUM [46]. Specifically, for Rank-1 and mAP, PECA in the Visible to Infrared testing mode performs 5.85% and 4.64% better than FMCNet, respectively. In the Infrared to Visible testing mode, PECA outperforms FMCNet by 4.68% in Rank-1 and 6.36% in mAP. Moreover, our method achieves performance comparable to CAL [48]

Table 3: The ablation study results for each component in PECA.

Methods	L_{ins}	L_{mod}	GSL	ICAL	SYSU-MM01(All Search)				RegDB (Visible to Infrared)			
					Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
M1	×	×	×	×	65.12	92.37	96.03	61.25	80.23	92.10	95.87	73.06
M2	✓	×	×	×	65.98	92.79	96.30	62.17	81.25	92.57	95.88	73.86
M3	✓	✓	×	×	67.23	93.25	96.39	62.83	83.17	93.69	96.85	75.32
M4	✓	✓	✓	×	72.86	96.12	98.16	68.25	88.96	96.03	97.81	83.19
M5	✓	✓	×	✓	70.23	95.83	98.62	66.83	85.36	95.27	97.22	80.07
PECA	✓	✓	✓	✓	74.89	96.57	99.03	72.28	94.97	99.23	99.56	89.07

on the RegDB dataset.

Based on the comprehensive experimental results, it can be observed that our PECA outperforms existing VI-ReID methods in terms of overall performance on the SYSU-MM01 and RegDB datasets.

4.4 Ablation study

In this subsection, we conduct a series of experiments to validate the effectiveness of each component. The performance of all methods is shown in Table 3. Following CAJ [13], M1 optimizes the model using cross-entropy loss and triplet loss without utilizing random grayscale augmentation. M2 replaces the triplet loss with L_{ins}. M3 builds upon M2 by introducing L_{mod}. M4 and M5 further incorporate GSL and ICAL, respectively, on top of M3.

4.4.1 Effectiveness of L_{ins}

L_{ins} is designed to handle hard samples and increase intra-class similarity and inter-class separability. As shown in Table 3, when M2 replaces the triplet loss in M1 with L_{ins}, the performance improves. Specifically, in the SYSU-MM01 (All Search) scenario, M2 outperforms M1 by 0.86%, 0.42%, 0.27%, and 0.92% in terms of Rank-1, Rank-10, Rank-20, and mAP, respectively. In the RegDB (Visible to Infrared) scenario, M2 achieves 1.02%, 0.47%, 0.01%, and 0.80% higher performance than M1 for the four metrics. This validates the effectiveness and superiority of L_{ins}.

4.4.2 Effectiveness of L_{mod}

L_{mod} aims to reduce the impact of modality gap and further improve intra-class compactness. As shown in Table 3, we find that with the help of L_{mod}, M3 outperforms M2. Specifically, in the SYSU-MM01 (All Search) scenario, M3 achieves 1.25%, 0.46%, 0.09%, and 0.66% higher performance than M2 for Rank-1, Rank-10, Rank-20, and mAP, respectively. In the RegDB (Visible to Infrared) scenario, M3 demonstrates 1.92%, 1.12%, 0.97%, and 1.46% improvement over M2 for the four metrics. This validates the effectiveness of L_{mod}.

4.4.3 Effectiveness of GSL

GSL aims to guide the model to extract global features that are rich in fine-grained information. As shown in Table 3,

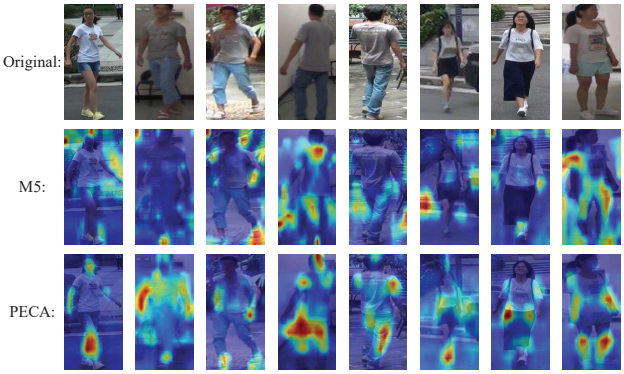
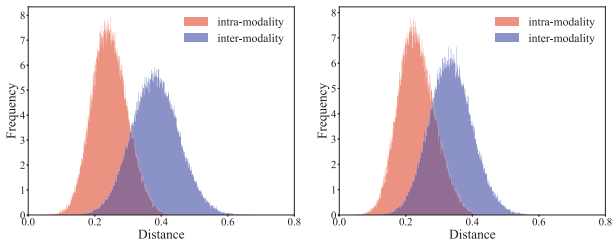


Fig. 7: The visual explanations of the models.



(a) M4 (b) PECA

Fig. 8: The distance distributions of intra-modality and inter-modality.

we observe that in both the SYSU-MM01 (All Search) and RegDB (Visible to Infrared) scenarios, M4 achieves Rank-1 and mAP metrics that are more than 5% higher than M3. Additionally, we find that PECA's performance in both of these scenarios is significantly better than M5. We visualize the visual explanations [49] of the models as shown in Figure 7. The attention of M5 is typically more scattered and may even focus on identity-irrelevant information such as the background. In contrast, PECA with the introduction of GSL tends to focus more on discriminative identity-related local information in the image. This validates that GSL can enhance the model's recognition performance by embedding fine-grained information in global features.

4.4.4 Effectiveness of ICAL

The significant improvement of M5 over M3, and the superiority of PECA over M4, after the introduction of ICAL is

attributed to ICAL's gradual guidance in bridging the modality gap. To further understand the effectiveness of ICAL in alleviating modality gap, we visualize the inter-modality feature distances and intra-modality feature distances in M4 and PECA, respectively. As shown in Figure 8, compared to M4, PECA can bring the distribution of inter-modality feature distances closer to the distribution of intra-modality feature distances, thus further reducing the impact of modality gap.

Additionally, we find that PECA outperforms all methods listed in Table 3, validating the effectiveness of the combination of all components. From a qualitative perspective, as shown in Figure 7, PECA focuses more on identity-related information compared to M5 by introducing GSL. This validates that GSL can effectively integrate with other components. On the other hand, as shown in Figure 8, PECA significantly reduces the difference between inter-modality feature distances and intra-modality feature distances by introducing ICAL. This validates that ICAL can organically combine with other components to mitigate the impact of modality gap. This further confirms that combining these advantageous components can enhance the overall performance of the model.

5. Conclusion

In this paper, we propose a partial enhancement and channel aggregation (PECA) method to address the VI-ReID problem. PECA consists of three components: global-local similarity learning (GSL), inter-modality channel aggregation learning (ICAL), and instance-modality contrastive loss. We conduct an analysis of the key hyper-parameters in PECA and perform ablation studies on each component. The results verify that GSL can enhance recognition performance by embedding local information in global features, ICAL can effectively reduce the impact of modality gap, and instance-modality contrastive loss can further improve the model's performance by learning modality-invariant and identity-related features at both instance and modality levels. Extensive experimental results on the SYSU-MM01 and RegDB datasets validate the effectiveness and superiority of PECA.

Acknowledgment

Thanks to all who have supported this work during the research process. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] R. Sun, Q. Liang, Z. Yang, Z. Zhao, and X. Zhang, "Triplet Attention Network for Video-Based Person Re-Identification," IEICE TRANSACTIONS on Information and Systems, 104(10), 1775-1779, 2021.

[2] R. Sun, Z. Yang, L. Zhang, Y. Yu, "Orthogonal Deep Feature Decomposition Network for Cross-Resolution Person Re-Identification," IEICE TRANSACTIONS on Information and Systems, 105(11), 1994-1997, 2022.

[3] Z. Pang, J. Guo, W. Sun, Y. Xiao, M. Yu, "Cross-domain person re-identification by hybrid supervised and unsupervised learning," Applied Intelligence, 52(3), 2987-3001, 2022.

[4] B. Gaikwad, A. Karmakar, "End-to-end person re-identification: Real-time video surveillance over edge-cloud environment," Computers and Electrical Engineering, 99, 107824, 2022.

[5] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, J. Kautz, "Joint discriminative and generative learning for person re-identification," In proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2138-2147, 2019.

[6] 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, pp. 5380-5389, 2017.

[7] D. T. Nguyen, H. G. Hong, K. W. Kim, K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," Sensors, 17(3), 605, 2017.

[8] X. Cheng, R. Li, Y. Sun, Y. Zhou, K. Dong, "Gray Augmentation Exploration with All-Modality Center-Triplet Loss for Visible-Infrared Person Re-Identification," IEICE TRANSACTIONS on Information and Systems, 105(7), 1356-1360, 2022.

[9] Z. Pang, C. Wang, L. Zhao, Y. Liu, G. Sharma, "Cross-modality Hierarchical Clustering and Refinement for Unsupervised Visible-Infrared Person Re-Identification," IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[10] Y. Gavini, A. Agarwal, B. M. Mehtre, "Thermal to Visual Person Re-Identification Using Collaborative Metric Learning Based on Maximum Margin Matrix Factorization," Pattern Recognition, 134, 109069, 2023.

[11] Z. Pang, C. Wang, H. Pan, L. Zhao, J. Wang, M. Guo, "MIMR: Modality-Invariance Modeling and Refinement for unsupervised visible-infrared person re-identification," Knowledge-Based Systems, 285, 111350, 2024.

[12] Z. Wei, X. Yang, N. Wang, X. Gao, "Syncretic modality collaborative learning for visible infrared person re-identification," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 225-234, 2021.

[13] M. Ye, W. Ruan, B. Du, M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13567-13576, 2021.

[14] Q. Zhang, C. Lai, J. Liu, N. Huang, J. 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 2022, pp. 7349-7358.

[15] Z. Feng, J. Lai, X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," IEEE Transactions on Image Processing, 29, 579-590, 2019.

[16] S. Choi, S. Lee, Y. Kim, T. Kim, C. 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, pp. 10257-10266, 2020.

[17] X. Hao, S. Zhao, M. Ye, J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16403-16412, 2021.

[18] M. Ye, Z. Wang, X. Lan, P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," In IJCAI, Volume. 1, p. 2, 2018.

[19] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13379-13389, 2020.

[20] M. Ye, X. Lan, J. Li, P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," In Proceedings of the AAAI conference on Artificial Intelligence, Volume. 32, No. 1, 2018.

[21] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, "Generative adversarial networks: An overview," IEEE signal processing magazine, 35(1), 53-65, 2018.

[22] Z. Pang, J. Guo, Z. Ma, W. Sun, Y. Xiao, "Median stable clustering and global distance classification for cross-domain person re-identification," IEEE Transactions on Circuits and Systems for Video Technology, 32(5), 3164-3177, 2021.
[23] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, "Cross-modality person re-identification with generative adversarial training," In IJCAI, Volume. 1, No. 3, p. 6, 2018.
[24] M. Ye, X. Lan, Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," In Proceedings of the 27th ACM International Conference on Multimedia, pp. 347-355, 2019.
[25] Z. Pang, J. Guo, W. Sun, S. Li, "Biclustering collaborative learning for cross-domain person re-identification," IEEE Signal Processing Letters, 28, 2142-2146, 2021.
[26] M. Ye, J. Shen, D. J Crandall, L. Shao, J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person re-identification," In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, Proceedings, Part XVII 16, pp. 229-247, 2020.
[27] G. A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z. G. Hou, "Cross-modality paired-images generation for RGB-infrared person re-identification," In Proceedings of the AAAI Conference on Artificial Intelligence, Volume. 34, No. 07, pp. 12144-12151, 2020.
[28] Y. Chen, L. Wan, Z. Li, Q. Jing, Z. Sun, "Neural feature search for rgb-infrared person re-identification," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 587-597, 2021.
[29] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1522-1531, 2021.
[30] Z. Huang, J. Liu, L. Li, K. Zheng, Z. J. Zha, "Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification," In Proceedings of the AAAI Conference on Artificial Intelligence, Volume. 36, No. 1, pp. 1034-1042, 2022.
[31] Z. Pang, L. Zhao, Q. Liu, C. Wang, "Camera Invariant Feature Learning for Unsupervised Person Re-Identification," IEEE Transactions on Multimedia, 2022.
[32] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, "Momentum contrast for unsupervised visual representation learning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729-9738, 2020.
[33] M. Wang, B. Lai, J. Huang, X. Gong, X S. Hua, "Camera-aware proxies for unsupervised person re-identification," In Proceedings of the AAAI Conference on Artificial Intelligence, Volume. 35, No. 4, pp. 2764-2772, 2021.
[34] Z. Pang, C. Wang, J. Wang, L. Zhao, "Reliability modeling and contrastive learning for unsupervised person re-identification," Knowledge-Based Systems, 263, 110263, 2023.
[35] X. Wang, H. Zhang, W. Huang, M R. Scott, "Cross-batch memory for embedding learning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6388-6397, 2020.
[36] X. Chen, K. He, "Exploring simple siamese representation learning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750-15758, 2021.
[37] Q. Wu, P. Dai, J. Chen, C. W. Lin, Y. Wu, F. Huang, B. Zhong, R. Ji, "Discover cross-modality nuances for visible-infrared person re-identification," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4330-4339, 2021.
[38] G. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network," arXiv 2015, arXiv:1503.02531.
[39] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S C. Hoi, "Deep learning for person re-identification: A survey and outlook," IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 2872-2893, 2021.
[40] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
[41] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255, 2009.
[42] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," IEEE Transactions on Multimedia, 22(10), 2597-2609, 2019.
[43] C. Fu, Y. Hu, X. Wu, H. Shi, T. Mei, R. He, "CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11823-11832, 2021.
[44] M. Ye, C. Chen, J. Shen, L. Shao, "Dynamic tri-level relation mining with attentive graph for visible infrared re-identification," IEEE Transactions on Information Forensics and Security, 17, 386-398, 2021.
[45] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, X. 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, pp. 14308-14317, 2022.
[46] 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/CVF Conference on Computer Vision and Pattern Recognition, pp. 19366-19375, 2022.
[47] H. Lu, X. Zou, P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," In Proceedings of the AAAI conference on Artificial Intelligence, Volume. 37, No. 2, pp. 1835-1843, 2023.
[48] J. Wu, H. Liu, Y. Su, W. Shi, H. Tang, "Learning concordant attention via target-aware alignment for visible-infrared person re-identification," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11122-11131, 2023.
[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," In Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, 2017.

Weiwei Jing received the M.S. degree from North University of China, P.R. China. Now, she works in School of Semiconductors and Physics. Her research interests encompass deep learning and person re-identification. She has authored several scholarly papers in these domains.

Zhonghua Li received the M.S. degree from North University of China, P.R. China. Now, he works in School of Sport And Physical Education. He has long been dedicated to research in image retrieval, particularly in the area of person re-identification. He has published several academic papers in this field.