IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

HPILN: a feature learning framework for cross-modality person re-identification

ISSN 1751-9659
Received on 6th June 2019
Revised 19th August 2019
Accepted on 16th September 2019
doi: 10.1049/iet-ipr.2019.0699
www.ietdl.org

Yun-Bo Zhao[1] ✉, Jian-Wu Lin[1], Qi Xuan[1], Xugang Xi[2]

[1]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, People's Republic of China
[2]School of Automation, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China
✉ *E-mail: ybzhao@ieee.org*

Abstract: Most video surveillance systems use both RGB and infrared cameras, making it a vital technique to re-identify a person cross the RGB and infrared modalities. This task can be challenging due to both the cross-modality variations caused by heterogeneous images in RGB and infrared, and the intra-modality variations caused by the heterogeneous human poses, camera position, light brightness etc. To meet these challenges, a novel feature learning framework, hard pentaplet and identity loss network (HPILN), is proposed. In the framework existing single-modality re-identification models are modified to fit for the cross-modality scenario, following which specifically designed hard pentaplet loss and identity loss are used to increase the accuracy of the modified cross-modality re-identification models. Based on the benchmark of the SYSU-MM01 dataset, extensive experiments have been conducted, showing that the authors' method outperforms all existing ones in terms of cumulative match characteristic curve and mean average precision.

1 Introduction

Person re-identification (Re-ID) is the technique of identifying an individual from a surveillance camera previously shown up from other non-overlapping cameras [1], which has recently become a research focus due to its practical importance. Typical Re-ID uses only RGB cameras, i.e. identifying an individual from RGB cameras based on previously recorded RGB camera videos/images, and hence the name RGB–RGB Re-ID [2–6]. However, in many cases, both RGB and infrared (IR) cameras are used, and consequently it becomes necessary to develop Re-ID methods capable of cross RGB and IR modalities, that is, either identifying an individual from RGB cameras based on previously recorded IR camera videos/images or identifying an individual from IR cameras based on previously recorded RGB cameral videos/images, both being referred to as RGB–IR Re-ID [7–11].

RGB–IR Re-ID has not been well studied to date, with few literature being reported. To name just a few, in [7], a deep zero-padding network is proposed to automatically extract the common features between RGB and IR modalities. In [8], a dual-path convolutional neural network (CNN) with top-ranking loss is proposed, which simultaneously handles both the cross- and intra-modality variations. In [9], a cross-modality generative adversarial network (cmGAN) approach with cross-modality triplet loss is proposed. In [10], a single image input method is proposed to simplify the CNN structure. In [11], a Dual-level Discrepancy Reduction Learning (D^2RL) scheme is proposed to decompose the mixed modality and appearance discrepancies. A dedicated dataset for RGB–IR Re-ID called SYSU-MM01 has been collected [7], as shown in Fig. 1.

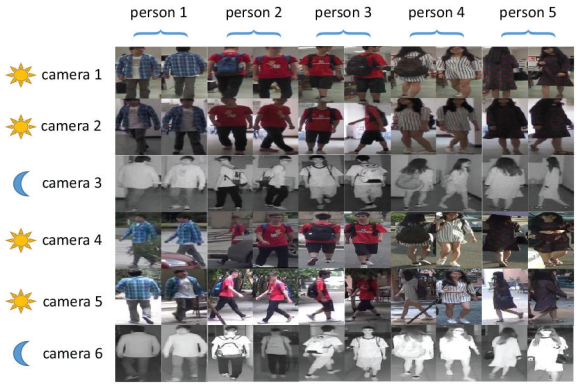

Fig. 1 *RGB and IR images in SYSU-MM01 dataset. The first, second, fourth, fifth and third, sixth rows are captured by RGB and IR cameras, respectively*

RGB–IR Re-ID is challenging mainly owing to the great cross- and intra-modality variations as illustrated in Fig. 2. By 'cross-modality variations' we mean that RGB and IR images are essentially heterogeneous as the former consists of three channels of colour information while the latter only one. By 'intra-modality variations' we mean that the image quality including the camera view, resolution, light brightness, human body pose etc. can still be significantly different even within the same RGB or IR modality, as long as multiple heterogeneous cameras and different monitoring scenarios are involved.

To meet the above challenges, a novel feature learning framework based on hard pentaplet and identity loss network (HPILN) is proposed. Specifically, we select existing RGB–RGB Re-ID models as the feature extraction module in our framework [2–6], and then design the hard pentaplet (HP) loss to compensate for the deficiencies of the RGB–RGB Re-ID network in the cross-modality Re-ID task. The HP loss considers the following two aspects: (i) a pentaplet loss, consisting of the global and cross-modality triplet loss where the former can simultaneously handle cross- and intra-modality variations and the latter can increase the

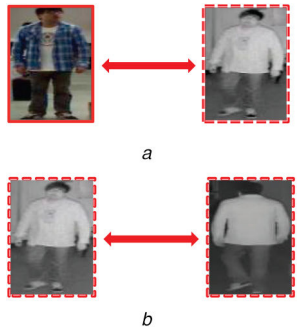

Fig. 2 *Cross- and intra-modality variations in RGB–IR Re-ID. Solid and dotted lines are for the RGB and IR domains, respectively*
(a) Cross-modality variations, ***(b)*** Intra-modality variations

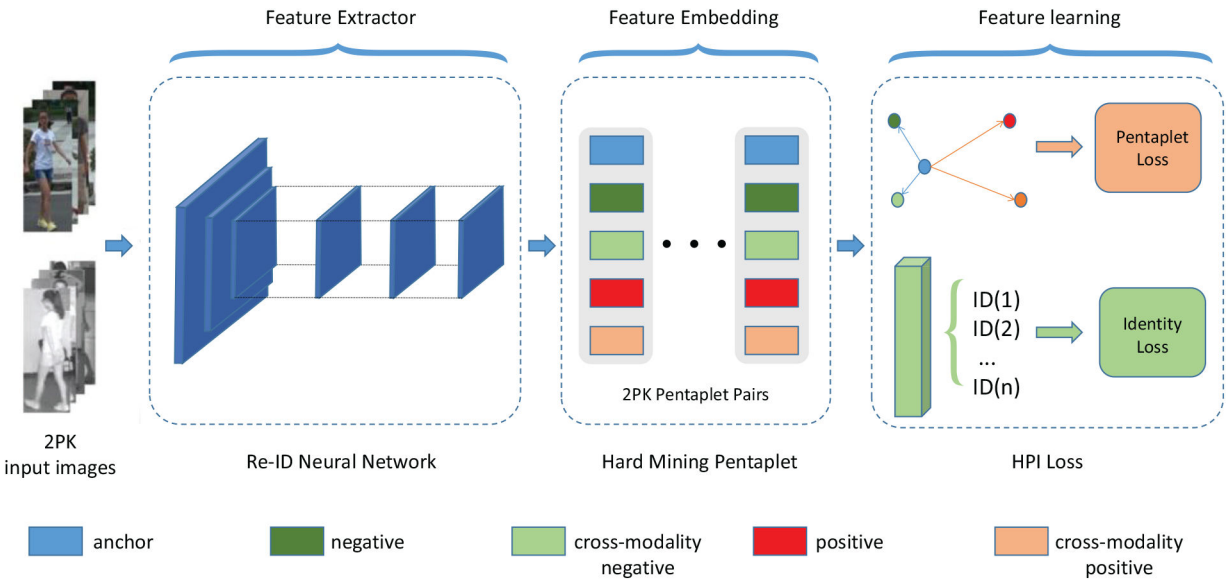

Fig. 3 *Proposed feature learning framework based HPI loss for RGB–IR Re-ID. The framework consists of three major components: (i) the Re-ID neural network, which extracts the common features of RGB and IR images; (ii) the hard mining sampling method, which obtains the hardest pentaplet pair sets; (iii) the HPI loss for feature learning, which consists of pentaplet loss and identity loss. 2PK is the training batch size. In each training batch, P individuals are randomly selected, and each person randomly selects K RGB images and K IR images. Rectangles of different colours below the image represent the different elements in the pentaplet pair*

ability to handle cross-modality variations. (ii) An improved hard mining sampling method by selecting the hardest global triplet and the hardest cross-modality triplet to form the hardest pentaplet pair and to contribute to the convergence of the CNN.

The main contributions of this study can be summarised as follows.

- An end-to-end feature learning framework is proposed yielding the state-of-the-art performance on the RGB–IR Re-ID dataset SYSU-MM01.
- The proposed RGB–RGB Re-ID model migration to RGB–IR Re-ID task provides a superior feature extraction method for future improvements.
- A novel loss function called HP loss is proposed, which is capable of simultaneously handling the cross- and intra-modality variations in RGB–IR Re-ID.

The remainder of the paper is organised as follows. Section 2 discusses related works on Re-ID. Our method is detailed in Section 3, which is then verified experimentally in Section 4. Section 5 concludes the paper.

2 Related works

In this section, we discuss related works on single-modality and multi-modality Re-ID.

2.1 Single-modality person Re-ID

In the single-modality person Re-ID study, most attentions have been paid to RGB–RGB Re-ID.

For RGB–RGB Re-ID, hand-designed descriptors are often used to extract pedestrian features such as colour and texture information. In [12], pedestrians are segmented from the background, and then the weighted colour histogram and the maximally stable colour regions are calculated for the pedestrian body part. Recently, the mainstream of Re-ID is to design the loss function and CNNs based on deep learning methods. The design of the loss function may depend on either metric learning or representation learning. The purpose of metric learning is to learn the similarity between two pedestrian images through a deep CNN, where the similarity is usually represented by the Euclidean distance. Frequently used metric learning methods include contrastive loss [13], triplet loss [14], hard triplet (HT) loss [15], and quadruplet loss [16]. Representation learning uses identity tags to automatically extract pedestrian representation features, including identity loss [17] and verification loss [18]. In addition, three types of special networks have been designed for Re-ID, i.e. either global-based or part-based or attention-based. Global-based networks aggregate global-level features into a global vector [2, 4]. Part-based networks divide the pedestrian image into different parts, and the local feature vectors of different parts are merged into a vector [3, 5, 6]. Attention-based networks focus on automatically finding local salient regions for computing deep features [19, 20]. These existing single-modality Re-ID models have rarely been applied to RGB–IR Re-ID to date and efforts need to be taken for such a migration.

2.2 Multi-modality person Re-ID

Existing multi-modality fusion Re-ID focuses on RGB–D modules [21–23], visible-thermal (VT) modules [8, 24] and RGB–IR modules [7]. RGB–D Re-ID combines human RGB image and depth information, and depth information is used to provide more stable body information to reduce the impact of changed clothes or extreme illumination on Re-ID. RGB–IR and VT Re-ID are based on the principle of IR imaging, enabling Re-ID to take place at night. The difference is that the RGB–IR Re-ID transmits and collects IR light through the IR camera to obtain IR images, while the VT Re-ID capturing the heat emitted by the human body to obtain IR images. However, depth cameras and thermal cameras are rare in surveillance systems. In contrast, IR cameras have been widely deployed. Most surveillance cameras in the real world are visible light cameras during the day and become IR cameras at night. Therefore, from the perspective of practical applications, RGB–IR Re-ID can be of more value.

3 Proposed method

This study addresses RGB–IR Re-ID by a feature learning framework based on HP loss and identity loss as shown in Fig. 3. The framework consists of three parts: (i) Re-ID neural network for feature extraction; (ii) the hard mining sampling method to find hardest pentaplet pair sets after getting feature embedding; (iii) hard pentaplet and identity (HPI) loss for feature learning. Specifically, the Re-ID neural network is taken from the existing RGB–RGB Re-ID CNN, which can also extract the representation feature of IR person images. By calculating the Euclidean distance of the feature embedding, the hard mining sampling method maximises training and ensures model convergence. The HP loss

enables the network to handle cross- and intra-modality variations simultaneously, and the HP loss and the identity loss are integrated into multiple losses to facilitate the process of feature learning.

3.1 Re-ID neural network

A typical RGB–RGB Re-ID model-based CNN is shown in Fig. 4. The CNN part is used as a feature extractor to obtain feature embedding, allowing different designs in different RGB–RGB Re-ID models. Most Re-ID models have at least two fully connected layers (FC-1 and FC-2 for short), where FC-2 is used for identity loss and the output of FC-1 is used as feature embedding supervised by ranking loss based on metric learning. Cross entropy loss and its variants are often used as identity loss, and ranking loss typically uses a loss function based on metric learning, such as HT loss. Joint training of identity and ranking loss can learn more discriminative feature embedding.

In our framework, we slightly adjust the structure of the RGB–RGB Re-ID model. First, identity loss is usually expressed using softmax loss, meaning that the dimension of FC-2 has to be changed to fit for the number of identities in the SYSU-MM01 training set. Second, the ranking loss used in the RGB–RGB Re-ID model does not consider cross-modality variations. Therefore, we design a new ranking loss called HP loss, which can better extract the discriminative features of heterogeneous images, as detailed in Section 3.2.2.

The modified network has two major advantages: (i) an existing RGB–RGB Re-ID neural network is used as the feature extraction module in the RGB–IR Re-ID task. Compared to the classification model, the customised RGB–RGB Re-ID model can learn more discriminative features, since the latter is designed by taking consideration of the person characteristics, such as the spatial distribution of pedestrian body parts used by the part-based Re-ID models. (ii) An HP loss is designed as the ranking loss for cross-modality Re-ID, which considers both the cross- and intra-modality variations and can, therefore, better learn the common features of heterogeneous images.

3.2 HP loss

We first introduce the HT loss and then discuss our proposed HP loss.

3.2.1 HT loss: The HT loss [15] is an improved loss function for triplet loss [14]. The triplet loss is widely used in person Re-ID, vehicle retrieval, and face recognition. In the person Re-ID task, for the anchor image x^a in the candidate triplet set $\{x_i^a, x_i^p, x_i^n\}, i \in [1, N]$, x^p is a positive sample image of the same identity, and x^n is a negative sample image of a different identity. Using CNN as the feature extractor, the image x is mapped into the d-dimensional Euclidean space. The feature embedding vector has the form of $f(x) \in \mathbb{R}^d$. The following Euclidean distance between feature embedding measures the similarity of two images:

$$d(x_i, x_j) = \| f(x_i) - f(x_j) \|_2 \tag{1}$$

The triplet loss is obtained as follows:

$$L_{\text{trp}} = \sum_i^N \left[d(x_i^a, x_i^p)^2 - d(x_i^a, x_i^n)^2 + \alpha \right]_+ \tag{2}$$

where $[z]_+ = \max(z, 0)$. For $\{x_i^a, x_i^p, x_i^n\}$, the ith pair of triplets, $d(x_i^a, x_i^p)$ represents the Euclidean distance between positive samples (x_i^a, x_i^p), and $d(x_i^a, x_i^n)$ represents the Euclidean distance between negative samples (x_i^a, x_i^n). Note that the square of the Euclidean distance is used for the triplet loss. α is a hyperparameter that forces the positive and negative sample pairs to separate in the Euclidean space.

Alexander Hermans *et al.* proposed a HT loss by improving the sampling method [15], which improves the training speed and accuracy in many retrieval tasks. With this loss, each batch randomly samples P-identity person, and each person randomly samples K images, thus PK images for each batch. For each image in the batch, select the hardest positive and negative samples to form the hardest triplet. The hardest positive sample represents the positive sample with the maximum Euclidean distance from the anchor, while the hardest negative sample represents the negative sample with the minimum Euclidean distance from the anchor. The HT loss can be expressed as follows:

$$L_{\text{htrp}} = \overbrace{\sum_{i=1}^{P} \sum_{a=1}^{K}}^{\text{all anchor}} \left[\alpha + \overbrace{\max_{p=1,\ldots,K} d(x_i^a, x_i^p)}^{\text{hardest positive}} - \underbrace{\min_{\substack{j=1,\ldots,P \\ n=1,\ldots,K \\ j \neq i}} d(x_i^a, x_j^n)}_{\text{hardest negative}} \right]_+ \tag{3}$$

Unlike (2), the HT loss does not calculate the square of the Euclidean distance, thus making the training more stable.

3.2.2 HP loss: As shown in Fig. 5*a*, the HT loss focuses on reducing the intra-class distance and increasing the inter-class distance, which is effective in the conventional retrieval task. However, the HT loss does not perform very well in RGB–IR Re-ID task. As shown in Fig. 2*a*, the same person in different modalities can be dissimilar. The HT loss does not consider cross-modality factors, and hence the training model does not deal well with cross- and intra-modality variations at the same time.

To address the huge cross- and intra-modality variations in cross-class or intra-class, we propose a hard global triplet (HGT)

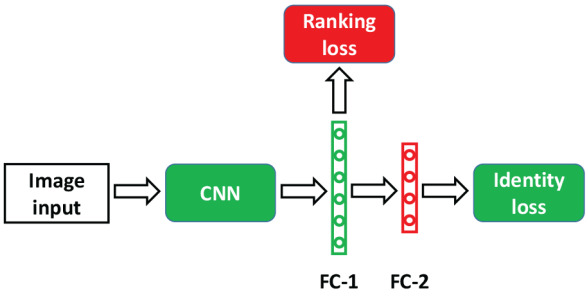

Fig. 4 *Typical RGB–RGB Re-ID CNN model. Green and red represent the unchanged and changed parts, respectively*

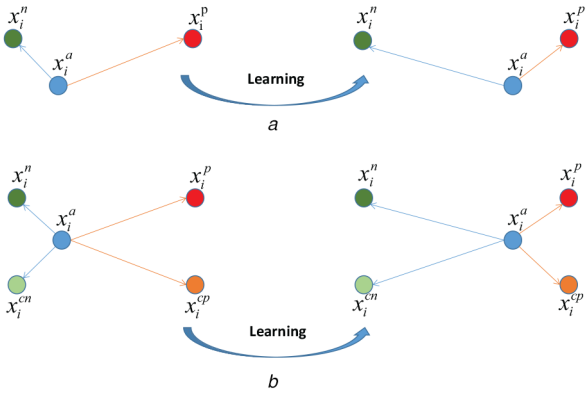

Fig. 5 *Geometry representation of HT loss and HP loss in Euclidean space*

(a) The HT loss minimises the distance between the anchor x_i^a and a positive x_i^p, and maximises the distance between the anchor x_i^a and a negative x_i^n, ***(b)*** In addition to the function of HT loss, HP loss can minimise the distance between an anchor x_i^a and a cross-modality positive x_i^{cp}, and maximises the distance between the anchor x_i^a and a cross-modality negative x_i^{cn}

loss based on a cross-modality batch (cm-batch) structure. Specifically, in each cm-batch, P individuals are randomly selected, each person randomly selects K RGB images and K IR images. For an anchor image x_i^a, the sum of cross-modality negative set x^{cn} and intra-modality negative set x^{in} constitutes the global negative set x^n, and the sum of the cross-modality positive set x^{cp} and intra-modality positive set x^{ip} constitutes the global positive set x^p. The HGT loss is computed as follows:

$$L_{\text{hgt}} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\overbrace{\left[\alpha + \overbrace{\max_{\substack{p=1,\ldots,2K\\ p\neq a}} d(x_i^a, x_i^p)}^{\text{hardest global positive}} - \underbrace{\min_{\substack{n=1,\ldots,2K\\ j=1,\ldots,P\\ j\neq i}} d(x_i^a, x_j^n)}_{\text{hardest global negative}}\right]_+}^{\text{all anchor}} \tag{4}$$

where α is a hyperparameter, $x_i^a \in x^a, x_i^p \in x^p, x_i^n \in x^n$, x_i^j represents the ith image of the jth person in the corresponding set of anchors. For any x_i^a in the cm-batch, the hardest global positive or negative may be the same or different modality.

Although HGT loss can handle cross- and intra-modality variations at the same time, usually cross-modality variations are much larger than intra-modality variations. We thus design a hard cross-modality triplet (HCT) loss to handle cross-modality variations. The HCT loss is computed as follows:

$$L_{\text{hct}} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\overbrace{\left[\alpha + \overbrace{\max_{\text{cp}\in A} d(x_i^a, x_i^{\text{cp}})}^{\text{hardest cross}-\text{modality positive}} - \underbrace{\min_{\substack{\text{cn}\in A\\ k=1,\ldots,K\\ k\neq i}} d(x_i^a, x_k^{\text{cn}})}_{\text{hardest cross}-\text{modality negative}}\right]_+}^{\text{all anchor}} \tag{5}$$

where $A = \{1, 2, \ldots, K\}$ for $a \geq K$, and $A = \{K+1, K+2, \ldots, 2K\}$ otherwise. The meaning of x_i^j is consistent with (4).

Our proposed HP loss consists of hard global and cross-modality triplet loss. For an anchor image x_i^a in cm-batch, the hardest global triplet pair $\{x_i^a, x_j^p, x_k^n\}$ and the hardest cross-modality triplet pair $\{x_i^a, x_h^{\text{cp}}, x_t^{\text{cn}}\}$ can be obtained by hard sampling methods, i.e. combining the hardest triplet pairs above to obtain a hardest pentaplet pair $\{x_i^a, x_j^p, x_k^n, x_h^{\text{cp}}, x_t^{\text{cn}}\}$. Note that x_j^p and x_h^{cp}, x_k^n and x_t^{cn} may be the same image. The HP loss can be expressed as follows:

$$L_{\text{hp}} = \frac{1}{2\times P\times K}(L_{\text{hgt}} + L_{\text{hct}}) \tag{6}$$

As shown in Fig. 5*b*, using the HP loss the distribution of human images in Euclidean space is more discriminative. The HP loss has two major advantages: (i) The HP loss can handle intra-modality and deeper cross-modality variations simultaneously. (ii) The HP sampling method uses a limited number of images to generate sufficient hardest pentaplet pairs, which enriches the training samples and speeds up model convergence.

3.3 HP with identity loss

We use identity loss to handle intra-class variations. As shown in Figs. 2*a* and *b*, there may be large variations in person images of the same identity. Given the success of identity loss in cross-modality Re-ID task, identity loss enables the CNN framework to extract the identity-specific information to reduce intra-class variations. We regard the same person in the heterogeneous modality as the same class, and the identity loss is then expressed by softmax loss, as follows:

$$L_{\text{id}} = \frac{1}{2\times P\times K}\sum_{i=1}^{2PK} -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \tag{7}$$

where $2PK$ is the number of images in cm-batch, $\boldsymbol{f}$ is designed as the output vector of the last FC layer, $\boldsymbol{f}_j, j \in [1, H]$ is the jth part of class score vector, H is the number of classes, y_i is the class label of the input image x_i, and f_{y_i} is the class score of x_i.

We add identity loss to our framework to learn a more robust feature representation. HPI loss is combined with HP loss and identity loss, which can be expressed as follows:

$$L_{\text{HPI}} = L_{\text{hp}} + L_{\text{id}} \tag{8}$$

4 Experimental results

4.1 Datasets and settings

The publicly available SYSU-MM01 dataset is adopted for evaluation. This dataset contains 491 identities with 287,628 RGB images and 15,792 IR images in total, captured by four RGB cameras (cameras 1, 2, 4, and 5) and two IR cameras (cameras 3 and 6). RGB and IR cameras work in bright and dimly lit environment, respectively. In addition, cameras 1, 2, 3 and 4, 5, 6 are for indoor and outdoor usages, respectively.

4.2 Evaluation protocol

The SYSU-MM01 dataset contains the training set and the test set, consisting of 395 and 96 persons, respectively. Note that a person does not appear in the two sets simultaneously.

In the training stage, all images in the training set can be used for training. In the test stage, the RGB images are for the gallery set and the IR images are for the query set. There are two verification methods: all-search and indoor-search. For the former, the test set contains all images from indoors and outdoors, while for the latter, the test set contains only indoor images. For each method, there are multi-shot and single-shot settings. For every identity in the gallery set, we randomly select 1/10 images from the RGB camera as single- and multi-shot setting, respectively. For the query set, all IR images are selected.

For a given query image, we match it by calculating the similarity between the images from the query set and the gallery set. The matching of the Re-ID is performed between cameras at different positions, and hence the query images will skip the gallery images from the cameras at the same position. After sorting the similarities, we use cumulative match characteristic (CMC) [25] and mean average precision (mAP) to calculate the accuracy.

4.3 Implementation details

We use NVIDIA GeForce 1080Ti graphics cards with Pytorch computing framework to implement our algorithm. Five RGB–RGB Re-ID neural networks were used to verify the superiority of our algorithms: Res-Mid, Multiple Granularity Network (MGN), Part-based Convolutional Baseline (PCB), Batch Feature Erasing (BFE), Multi-Level Factorisation Net (MLFN). As shown in Table 1, the input image size and the output embedding feature dimension are different due to the difference of the model. The IR image is padding to three channels, which copies the information of one channel. Adam optimiser [26] is used to train 10 k iterations.

Since our HP loss requires slightly different cm-batches, we sample a $2PK$ batch by randomly sampling P identities, and each person randomly samples K RGB images and K IR images. In our experiment, P is set to 8, K is set to 4, and the size of cm-batch is calculated to be 64. For input images, the methods of random horizontal flip and random cropping is used to expand the amount

of data. We set margin α in HP loss in the range [0.3, 0.6, 0.9, 1.2, 1.5, 1.8] and evaluate our method by experimenting with other hyper-parameters.

4.4 Comparison with the state-of-the-arts

We evaluated our HPILN method against 17 previous methods on the SYSU-MM01 dataset in Table 2. For performance measure, the rank-1, -10, -20 accuracies of CMC and mAP are used to show the clear performance superiority of our HPILN method. The comparison takes advantage of seven state-of-the-art methods: zero-padding [7], cmGAN [9], bi-directional dual-constrained top-ranking (BDTR) [8], inter-channel pair between the visible-light and thermal images + multi-scale Retinex (IPVT-1 + MSR) [10], D^2RL [11], bi-directional center-constrained top-ranking (eBDTR) [27] and D-hypersphere manifold embedding (HSME) [28].

Table 1 Settings for different model training: input image width and height ($W \times H$), feature dimension (Dim), training batch size (batch size) and learning rate (Lr)

Model	$W \times H$	Dim	Batch size	Lr
Res-Mid	224 × 224	3072	64	3×10^{-4}
MGN	128 × 384	2048	64	3×10^{-4}
PCB	224 × 224	12,288	64	3×10^{-4}
BFE	128 × 256	1024	64	3×10^{-4}
MLFN	224 × 224	1024	64	3×10^{-4}

In addition, other existing methods are used for comparison, including handcrafted features such as Histograms of Oriented Gradient (HOG) [29] and Local Maximal Occurrence (LOMO) [30], cross-domain models such as Common Discriminant Feature Extraction (CDFE) [31] and Camera coRrelation Aware Feature augmenTation (CRAFT) [32], canonical correlation analysis (CCA) [33], one-stream and two-stream networks [7], and metric learning method Local Fisher Discriminant Analysis (LFDA) [34]. Most of the results were obtained from the references [7–11, 27, 28].

We use Res-Mid [2], MGN [6], PCB [3], BFE [5], MLFN [4] as feature extractors in our HPILN method. To our best knowledge, these models are the best available methods of RGB–RGB Re-ID in the past two years. Table 3 records their precision on the Market1501 [35], CUHK03 [36] and DukeMTMC-reID [37, 38] datasets.

In Table 2, the results of five rows on the bottom show the performance of the HPILN method, which applies HPI loss to five models. It is seen that our method is significantly better than all existing methods in the SYSU-MM01 benchmark, where the five models based on HPI loss have higher rank-1, -10, -20 and mAP in all verification modes and setting than existing methods. Specifically, the BFE model-based HPI loss performs the best in most of the indicators, which outperforms the second best method (D^2RL) on all-search single-shot setting in terms of the rank 1 and mAP metric 12.46% (41.36–28.9) and 13.75% (42.95–29.2), respectively.

Table 2 Comparison results on the SYSU-MM01 dataset. Our method exceeds the existing methods on the rank 1, 10, 20 and mAP metrics

Method	All-search single-shot				All-search multi-shot				Indoor-search single-shot				Indoor-search multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
HOG + Euclidean	2.76	18.25	31.91	4.24	3.82	22.77	37.63	2.16	3.22	24.68	44.52	7.25	4.75	29.06	49.38	3.51
HOG + CRAFT	2.59	17.93	31.50	4.24	3.58	22.90	38.59	2.06	3.03	24.07	42.89	7.07	4.16	27.75	47.16	3.17
HOG + CCA	2.74	18.91	32.51	4.28	3.25	21.82	36.51	2.04	4.38	29.96	50.43	8.70	4.62	34.22	56.28	3.87
HOG + LFDA	2.33	18.58	33.38	4.35	3.82	20.48	35.84	2.20	2.44	24.13	45.50	6.87	3.42	25.27	45.11	3.19
LOMO + CCA	2.42	18.22	32.45	4.19	2.63	19.68	34.82	2.15	4.11	30.60	52.54	8.83	4.86	34.40	57.30	4.47
LOMO + CRAFT	2.34	18.70	32.93	4.22	3.03	21.70	37.05	2.13	3.89	27.55	48.16	8.37	2.45	20.20	38.15	2.69
LOMO + CDFE	3.64	23.18	37.28	4.53	4.70	28.23	43.05	2.28	5.75	34.35	54.90	10.19	7.36	40.38	60.33	5.64
LOMO + LFDA	2.98	21.11	35.36	4.81	3.86	24.01	40.54	2.61	4.81	32.16	52.50	9.56	6.27	36.29	58.11	5.15
one-stream	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
two-stream	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.02	21.49	22.49	72.22	88.61	13.92
zero-padding	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.64
cmGAN	26.97	67.51	80.56	27.80	31.49	72.74	85.01	22.27	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
BDTR	17.01	55.43	71.96	19.66	—	—	—	—	—	—	—	—	—	—	—	—
eBDTR	27.82	67.34	81.34	28.42	—	—	—	—	—	—	—	—	—	—	—	—
D-HSME	20.68	62.74	77.95	23.12	—	—	—	—	—	—	—	—	—	—	—	—
IPVT-1 + MSR	23.18	51.21	61.73	22.49	—	—	—	—	—	—	—	—	—	—	—	—
D^2RL	28.9	70.6	82.4	29.2	—	—	—	—	—	—	—	—	—	—	—	—
Res-Mid + HPI	40.49	83.61	93.13	41.64	**47.70**	87.99	95.34	35.15	45.65	90.76	97.77	56.19	50.79	93.03	97.86	46.21
MGN + HPI	39.77	79.78	90.14	41.12	44.86	82.54	91.61	34.88	44.06	87.77	95.59	54.52	50.55	89.99	96.06	44.90
PCB + HPI	33.29	80.66	91.42	35.15	38.55	82.86	92.82	28.16	39.70	88.26	96.68	50.49	46.86	90.31	96.85	40.93
MLFN + HPI	33.34	78.54	89.66	36.13	39.45	83.21	92.45	29.52	36.25	85.07	94.51	47.99	41.99	86.34	95.20	38.43
BFE + HPI	**41.36**	**84.78**	**94.51**	**42.95**	47.56	**88.13**	**95.98**	**36.08**	**45.77**	**91.82**	**98.46**	**56.52**	**53.05**	**93.71**	**98.93**	**47.48**

Table 3 Performance of RGB–RGB Re-ID models on Market1501, CUHK03, and DukeMTMC-reID datasets

Method	Market1501		CUHK03		DukeMTMC-reID	
	r1	mAP	r1	mAP	r1	mAP
Res-Mid	89.87	75.55	43.51	47.14	63.88	80.43
MGN	95.7	86.9	66.8	66	88.7	78.4
PCB	92.4	77.3	61.3	54.2	81.9	65.3
BFE	94.4	85	72.1	67.9	88.7	75.8
MLFN	90	74.3	52.8	47.8	81.0	62.8

Table 4 Effectiveness of fusion loss on the SYSU-MM01 dataset. Rank-1 accuracies (%) in all/indoor-search mode and single/multi-shot setting

Model	Loss	All-search		Indoor-search	
		Single-shot	Multi-shot	Single-shot	Multi-shot
Res-Mid	identity loss	32.68	38.58	37.41	45.23
	HP loss	36.06	42.32	40.46	44.08
	HPI loss	40.49	47.70	45.65	50.79
MGN	identity loss	27.29	31.05	33.47	38.19
	HP loss	36.68	41.62	41.95	48.37
	HPI loss	39.77	44.86	44.06	50.55
PCB	identity loss	11.22	15.67	8.7	12.76
	HP loss	26.51	32.40	33.61	40.42
	HPI loss	33.29	38.55	39.70	46.86
MLFN	identity loss	28.44	33.23	30.19	34.82
	HP loss	30.62	35.43	31.28	36.69
	HPI loss	33.34	39.45	36.25	41.99
BFE	identity loss	25.69	32.02	29.65	36.68
	HP loss	38.89	45.69	44.51	52.51
	HPI loss	41.36	47.56	45.77	53.05

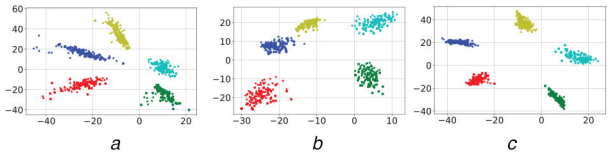

Fig. 6 *Comparison among identity loss, HP loss and HPI loss. In this toy experiment, we modified Res-Mid to learn a 2D feature on a subset of the SYSU-MM01 dataset. In detail, we set the output of dimension of the last FC layer as two and visualise the learned features. The five colour points represent five identity classes, the circular and star shapes represent RGB modality and IR modality, respectively*
(a) Identity loss, ***(b)*** HP loss, ***(c)*** HPI loss

4.5 Effectiveness of fusion loss

To verify the effectiveness of fusion identity loss and HP loss, we compared the rank-1 precision of identity loss, HP loss and HPI loss on the SYSU-MM01 dataset. We report the results with five models in Table 4, and it shows that the combination of identity loss is effective. It is clear that the RGB–RGB Re-ID models based on identity loss can also achieve excellent precision, even the Res-Mid-based identity loss performance is better than the second best method (D^2RL) in rank-1 accuracies. In addition, although HP loss has shown excellent performance, HPI loss which integrates identity loss and HP loss further improves the accuracy. We speculate that the fusion of identity loss further enhances the feature discrimination of HP loss.

In order to verify the above speculation, we conducted a toy experiment to illustrate the differences of features in 2D Euclidean space learned by identity loss, HP loss, and HPI loss, respectively, shown in Fig. 6. Using only identity loss to train the model, the learned features are slightly separable, which are not discriminative enough, since Fig. 6*a* still shows large cross-modality variations and small inter-class discrimination. Fig. 6*b* shows that there is a large margin between the dot clusters, which means HP loss learned discriminative large-margin features. For HPI loss combined with HP loss and identity loss, Fig. 6*c* shows that the same classes are clustered together and there is a significant separation between the different classes. The reason why the HPI loss performance superior is that HP loss handles the cross- and intra-modality variations to learn the distinguishing large-margin features and identity loss assists HP loss to further reduce intra-class distance.

4.6 Comparison with other advanced loss

To demonstrate the effect of our loss function, we compare our HP loss and HPI loss with other advanced loss functions in Re-ID, including HT loss [15], HT with identity (HTI) loss, centre loss [39], and identity loss [17]. The performance of the contrast methods was reported in Figs. 7 and 8.

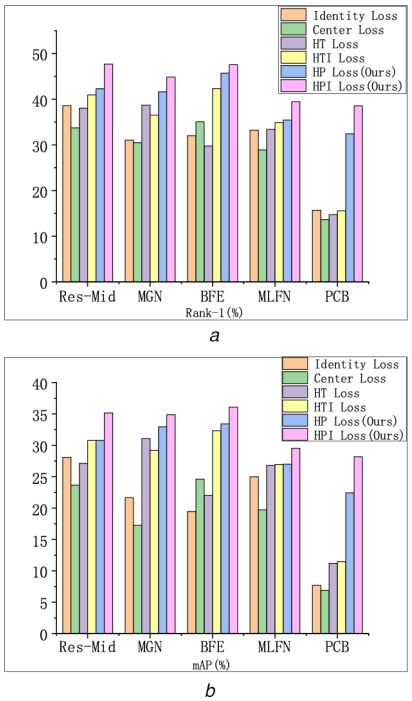

Fig. 7 *Performance of different loss functions. We tested rank-1 and mAP with the all-search multi-shot setting on five models*
(a) Rank-1, ***(b)*** mAP

Results shown in Fig. 7 illustrate that HP loss and HPI loss have better performance on rank-1 and mAP than other loss functions. We tested five models, and the rank-1 and mAP of the PCB-based HPI loss were 22.88% and 16.69% higher than the second loss (except HP loss), respectively.

Fig. 8 shows the CMC curves of different models under different losses in the SYSU-MM01 dataset. The CMC curve can more fully reflect the performance of the model. We tested four models under the all-search single-shot setting. In all tested models, HP loss and HPI loss performed better than other losses, and our method is not only higher than the existing method in rank-1 but also maintains a lead in ranks 1–50.

The reason why we are better than other methods is that we consider cross-modality variations to better extract common features in heterogeneous modality. In addition, we use a more reasonable image sampling method to balance the number of input images so that the CNN does not focus on certain modality images.

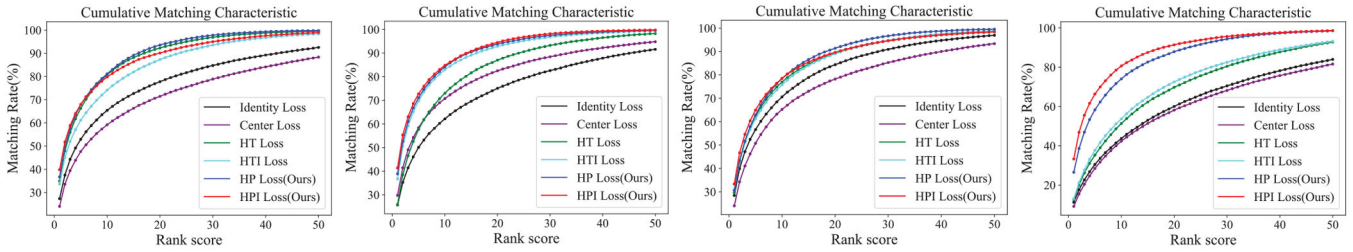

Fig. 8 *CMC curve of different models under different loss functions*
(a) MGN, ***(b)*** BFE, ***(c)*** MLFN, ***(d)*** PCB

Table 5 Rank-1 of classification models and RGB–RGB Re-ID models in HPILN framework

Type	Model	All-search	
		Single-shot	Multi-shot
global-based	Res-Mid	40.49	47.70
	MLFN	33.34	39.45
part-based	PCB	33.29	38.55
	MGN	39.77	44.86
	BFE	41.36	47.56
attention-based	hacnn	1.07	1.53
	mudeep	9.35	11.78
classification-based	Resnet50	5.36	5.69
	Densenet121	13.59	16.10

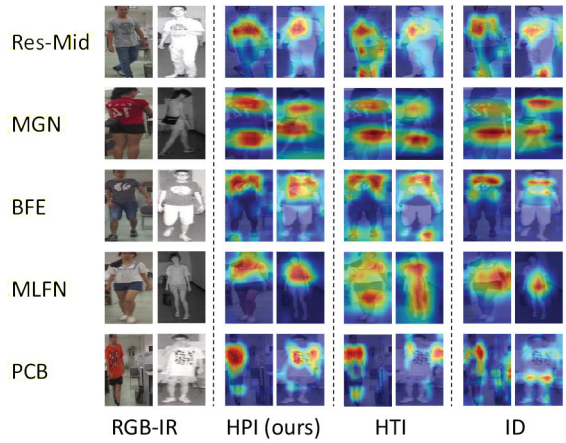

Fig. 9 *Visualisation of features for cross-modality Re-ID. Each pair of images contains visible and IR images of the same person. Each row represents the different models, and columns 2, 3, and 4 represent the activation map of the first column of images by the HPI, HTI, and ID loss training models, respectively*

4.7 Analysis of model selection

In HPILN framework, the RGB–RGB Re-ID models were first adopted as the feature extractor for RGB–IR Re-ID task. We tested the performance of the classification models and some RGB–RGB Re-ID models in HPILN framework. The RGB–RGB Re-ID models include different types: global-based networks, part-based networks, and attention-based networks. We record the results in Table 5.

In the HPILN framework, the RGB–RGB Re-ID model is more effective for RGB–IR Re-ID tasks than classification models. We chose Resnet50 [40] and Densenet121 [41] as classification models, which perform well on ImageNet dataset. From Table 5, we observed that classification models do not achieve good accuracy in the HPILN framework compared to the RGB–RGB Re-ID model. The reason is that the RGB–RGB Re-ID model is designed for person images. Although IR images and RGB images are very different, heterogeneous images also have certain common features, such as body shape and clothing shape. Therefore, the RGB–RGB Re-ID model performs well in RGB–IR Re-ID tasks.

However, not all RGB–RGB Re-ID models perform well in RGB–IR Re-ID tasks. We tested two attention-based RGB–RGB Re-ID models: mudeep [19] and hacnn [20]. From Table 5, mudeep and hacnn have lower precision on SYSU-MM01. Both mudeep and hacnn use the attention mechanism which automatically focuses on local salient areas for computing deep features. We find that attentional mechanisms are more difficult to train in cross-modality Re-ID because there are few similar local regions in heterogeneous images.

4.8 Visualising activation map

To illustrate the effectiveness of our model, we use activation maps to visualise the features learned by different models. The activation map is obtained using the Grad-cam method [42]. Specifically, the feature map of the last convolutional layer is first obtained, and the weighted sum in the channel dimension is then calculated. The resulting activation map is superimposed on the original image.

Five models (Res-Mid, MGN, BFE, MLFN, and PCB) are tested and trained by the HPI, HTI, and identity loss, with the results being shown in Fig. 9. We find that the model has a strong response to the human body, especially in significant areas such as clothes. This shows that the model can automatically exclude background interference and concentrate on extracting the features of the human body. Compared with other loss functions, the models trained by the HPI loss have more similar response positions on RGB and IR images, meaning that the HPI loss can better learn the common features of heterogeneous images. In addition, our method can better extract features since more discriminative regions are recognised by our method.

5 Conclusion

A novel feature learning framework based on HPILN is proposed for RGB–IR person Re-ID. In the framework, an existing RGB–RGB Re-ID model is used as the feature extractor; HP loss is used to learn the discriminative large-margin features in order to handle cross- and intra-modality variations and the identity loss are combined to extract identity-specific information to learn the separation features. The experimental results show that our method achieves state-of-the-art performance on SYSU-MM01 dataset.

6 References

[1] Zheng, L., Yang, Y., Hauptmann, A.G.: 'Person re-identification: past, present and future'. arXiv preprint arXiv:1610.02984, 2016
[2] Yu, Q., Chang, X., Song, Y.Z., *et al.*: 'The devil is in the middle: exploiting mid-level representations for cross-domain instance matching'. arXiv preprint arXiv:1711.08106, 2017
[3] Sun, Y., Zheng, L., Yang, Y., *et al.*: 'Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)'. European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 480–496
[4] Chang, X., Hospedales, T.M., Xiang, T.: 'Multi-level factorisation net for person re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2109–2118
[5] Dai, Z., Chen, M., Zhu, S., *et al.*: 'Batch feature erasing for person re-identification and beyond'. arXiv preprint arXiv:1811.07130, 2018
[6] Wang, G., Yuan, Y., Chen, X., *et al.*: 'Learning discriminative features with multiple granularities for person re-identification'. ACM Int. Conf. on Multimedia, Seoul, South Korea, 2018, pp. 274–282
[7] Wu, A., Zheng, W.S., Yu, H.X., *et al.*: 'RGB-infrared cross-modality person re-identification'. Proc. IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 5380–5389

[8] Ye, M., Wang, Z., Lan, X., *et al.*: 'Visible thermal person re-identification via dual-constrained top-ranking'. Proc. Int. Joint Conf. on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 1092–1099

[9] Dai, P., Ji, R., Wang, H., *et al.*: 'Cross-modality person re-identification with generative adversarial training'. Proc. Int. Joint Conf. on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 677–683

[10] Kang, J.K., Hoang, T.M., Park, K.R., *et al.*: 'Person re-identification between visible and thermal camera images based on deep residual CNN using single input', *IEEE Access*, 2019, **7**, pp. 57972–57984

[11] Wang, Z., Wang, Z., Zheng, Y., *et al.*: 'Learning to reduce dual-level discrepancy for infrared-visible person re-identification'. Proc. IEEE Int. Conf. on Artificial Intelligence Organisation/IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019, pp. 618–626

[12] Farenzena, M., Bazzani, L., Perina, A., *et al.*: 'Person re-identification by symmetry-driven accumulation of local features'. Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, San Fransisco, USA, 2010, pp. 2360–2367

[13] Varior, R.R., Haloi, M., Wang, G.: 'Gated Siamese convolutional neural network architecture for human re-identification'. Proc. European Conf. on Computer Vision, Amsterdam, Netherlands, 2016, pp. 791–808

[14] Schroff, F., Kalenichenko, D., Philbin, J.: 'FaceNet: a unified embedding for face recognition and clustering'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Boston, USA, 2015, pp. 815–823

[15] Hermans, A., Beyer, L., Leibe, B.: 'In defense of the triplet loss for person re-identification'. arXiv preprint arXiv:1703.07737, 2017

[16] Chen, W., Chen, X., Zhang, J., *et al.*: 'Beyond triplet loss: a deep quadruplet network for person re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 403–412

[17] Xiao, T., Li, H., Ouyang, W., *et al.*: 'Learning deep feature representations with domain guided dropout for person re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 1249–1258

[18] Chen, H., Wang, Y., Shi, Y., *et al.*: 'Deep transfer learning for person re-identification'. Proc. 2018 IEEE Fourth Int. Conf. on Multimedia Big Data (BigMM), Xi'an, China, 2018, pp. 1–5

[19] Qian, X., Fu, Y., Jiang, Y.G., *et al.*: 'Multi-scale deep learning architectures for person re-identification'. Proc. 2017 IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, 2017, pp. 5409–5418

[20] Li, W., Zhu, X., Gong, S.: 'Harmonious attention network for person re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2285–2294

[21] Wu, A., Zheng, W.S., Lai, J.H.: 'Robust depth-based person re-identification', *IEEE Trans. Image Process.*, 2017, **26**, (6), pp. 2588–2603

[22] Barbosa, I.B., Cristani, M., Del Bue, A., *et al.*: 'Re-identification with RGB-D sensors'. Proc. European Conf. on Computer Vision, Florence, Italy, 2012, pp. 433–442

[23] Chattopadhyay, P., Sural, S., Mukherjee, J.: 'Information fusion from multiple cameras for gait-based re-identification and recognition', *IET Image Process.*, 2015, **9**, (11), pp. 969–976

[24] Kniaz, V.V., Knyaz, V.A., Hladuvka, J., *et al.*: 'Thermalgan: multimodal color-to-thermal image translation for person re-identification in multispectral dataset'. Proc. European Conf. on Computer Vision (ECCV), Munich, Germany, 2018

[25] Moon, H., Phillips, P.J.: 'Computational and performance aspects of PCA-based face-recognition algorithms', *Perception*, 2001, **30**, (3), pp. 303–321

[26] Kingma, D.P., Ba, J.: 'Adam: a method for stochastic optimization'. arXiv preprint arXiv:1412.6980, 2014

[27] Ye, M., Lan, X., Wang, Z., *et al.*: 'Bi-directional center-constrained top-ranking for visible thermal person re-identification', *IEEE Trans. Inf. Forensics Sec.*, 2020, **15**, pp. 407–419

[28] Hao, Y., Wang, N., Li, J., *et al.*: 'HSME: hypersphere manifold embedding for visible thermal person re-identification'. Proc. AAAI Conf. on Artificial Intelligence, Honolulu, Hawaii, USA, 2019, vol. 33, pp. 8385–8392

[29] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. Proc. IEEE Int. Conf. on Computer Vision & Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, vol. 1, pp. 886–893

[30] Liao, S., Hu, Y., Zhu, X., *et al.*: 'Person re-identification by local maximal occurrence representation and metric learning'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 2197–2206

[31] Lin, D., Tang, X.: 'Inter-modality face recognition'. Proc. European Conf. on Computer Vision, Graz, Austria, 2006, pp. 13–26

[32] Chen, Y.C., Zhu, X., Zheng, W.S., *et al.*: 'Person re-identification by camera correlation aware feature augmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, **40**, (2), pp. 392–408

[33] Rasiwasia, N., Costa Pereira, J., Coviello, E., *et al.*: 'A new approach to cross-modal multimedia retrieval'. Proc. 18th ACM Int. Conf. on Multimedia, Firenze, Italy, 2010, pp. 251–260

[34] Pedagadi, S., Orwell, J., Velastin, S., *et al.*: 'Local fisher discriminant analysis for pedestrian re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 3318–3325

[35] Zheng, L., Shen, L., Tian, L., *et al.*: 'Scalable person re-identification: a benchmark'. Proc. IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 1116–1124

[36] Li, W., Zhao, R., Xiao, T., *et al.*: 'DeepReID: deep filter pairing neural network for person re-identification'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 152–159

[37] Ristani, E., Solera, F., Zou, R., *et al.*: 'Performance measures and a data set for multi-target, multi-camera tracking'. Proc. European Conf. on Computer Vision, Amsterdam, Netherlands, 2016, pp. 17–35

[38] Yoon, K., Song, Y.M., Jeon, M.: 'Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views', *IET Image Process.*, 2018, **12**, (7), pp. 1175–1184

[39] Wen, Y., Zhang, K., Li, Z., *et al.*: 'A discriminative feature learning approach for deep face recognition'. Proc. European Conf. on Computer Vision, Amsterdam, Netherlands, 2016, pp. 499–515

[40] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778

[41] Huang, G., Liu, Z., Van Der Maaten, L., *et al.*: 'Densely connected convolutional networks'. Proc. IEEE Int. Computer Society Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4700–4708

[42] Selvaraju, R.R., Cogswell, M., Das, A., *et al.*: 'Grad-CAM: visual explanations from deep networks via gradient-based localization'. Proc. IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 618–626