Signal Processing: Image Communication 87 (2020) 115933

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

An efficient framework for visible–infrared cross modality person re-identification

Emrah Basaran [a,*], Muhittin Gökmen [b], Mustafa E. Kamasak [a]

[a] Istanbul Technical University, Computer Engineering Department, Maslak, 34467, Istanbul, Turkey
[b] MEF University, Computer Engineering Department, Maslak, 34396, Istanbul, Turkey

ARTICLE INFO

Keywords:
Person re-identification
Cross modality person re-identification
Local Zernike moments

ABSTRACT

Visible–infrared cross-modality person re-identification (VI-ReId) is an essential task for video surveillance in poorly illuminated or dark environments. Despite many recent studies on person re-identification in the visible domain (ReId), there are few studies dealing specifically with VI-ReId. Besides challenges that are common for both ReId and VI-ReId such as pose/illumination variations, background clutter and occlusion, VI-ReId has additional challenges as color information is not available in infrared images. As a result, the performance of VI-ReId systems is typically lower than that of ReId systems. In this work, we propose a four-stream framework to improve VI-ReId performance. We train a separate deep convolutional neural network in each stream using different representations of input images. We expect that different and complementary features can be learned from each stream. In our framework, grayscale and infrared input images are used to train the ResNet in the first stream. In the second stream, RGB and three-channel infrared images (created by repeating the infrared channel) are used. In the remaining two streams, we use local pattern maps as input images. These maps are generated utilizing local Zernike moments transformation. Local pattern maps are obtained from grayscale and infrared images in the third stream and from RGB and three-channel infrared images in the last stream. We improve the performance of the proposed framework by employing a re-ranking algorithm for post-processing. Our results indicate that the proposed framework outperforms current state-of-the-art with a large margin by improving Rank-1/mAP by 29.79%/30.91% on SYSU-MM01 dataset, and by 9.73%/16.36% on RegDB dataset.

1. Introduction

Person re-identification (ReId) can be defined as the retrieval of the images of a person from a gallery set, where the gallery and query sets consist of the images captured by different cameras with different field-of-views. It is an essential problem for various real-world scenarios, especially for security. Therefore, this problem has recently attracted considerable attention of researchers working in the fields of computer vision and machine learning.

The illumination conditions, visible body parts, occlusion, and background complexity can vary extremely in images captured by different cameras. Due to the dynamic nature of the scene, these conditions can also change in images recorded by the same camera. These challenges and typically low-resolution images make ReId a challenging computer vision problem. In the literature on this subject, many different methods have been proposed in the last decade that address ReId from various aspects. The vast majority of the methods have been developed considering only the images captured by visible cameras. However, in dark or poorly illuminated environments, visible cameras cannot capture features that can distinguish people.

Most of the surveillance cameras used at night or in the dark usually operate in infrared mode in order to cope with poor illumination. Therefore, matching the person images captured by visible and infrared cameras is an important issue for video surveillance or miscellaneous applications. This issue is studied in literature as visible–infrared cross-modality person re-identification (VI-ReId) [1–5]. VI-ReId is the problem of retrieving the images of a person from a gallery set consisting of RGB (or infrared) images, given an infrared (or RGB) query image. For ReId, one of the most important cues in person images is obtained from the color. Therefore, the lack of color information in infrared images makes VI-ReId a very challenging problem.

In this paper, we show that ResNet [6] architectures trained with the use of RGB and infrared images together can outperform the current state-of-the-art. These architectures are widely used for image classification and other computer vision problems. They can learn the common feature representations for RGB and infrared images of the same individual as well as the distinctive properties between the individuals better than the existing methods proposed for VI-ReId. In this study, we introduce a four-stream framework built with ResNet architectures. There is no weight sharing between the ResNets in the framework, and

* Corresponding author.
E-mail address: basaranemrah@itu.edu.tr (E. Basaran).

https://doi.org/10.1016/j.image.2020.115933
Received 4 June 2019; Received in revised form 18 February 2020; Accepted 21 June 2020
Available online 29 June 2020
0923-5965/© 2020 Elsevier B.V. All rights reserved.

in order to obtain different and complementary features as much as possible, we train each of them using a different representation of input images. In the first stream, RGB images converted to grayscale and infrared images are used together. In this way, features are extracted using only the shape and pattern information. In the second stream, to take advantage of the color information, RGB images and the three-channel infrared images (obtained by repeating the infrared image) are used. As mentioned above, one of the most important cues for ReId is obtained from the color information. However, due to the lack of color in infrared images, local shape and pattern information is critical for VI-ReId. Therefore, in addition to the features obtained from the raw (RGB, grayscale and infrared) images in the first two streams of the framework, in the last two streams, features are derived from the local pattern maps. In the third stream, we first apply the local Zernike moments (LZM) transformation [7] on grayscale and infrared images to expose the local patterns in the images. With the LZM transformation, different numbers of local pattern maps are generated by computing the Zernike moments around each pixel. We train the ResNet in the third stream by using the LZM pattern maps of grayscale and infrared images. In the last stream of the framework, we generate the pattern maps by exposing the local patterns separately in R, G, and B channels.

The contributions of this paper are summarized as follows:

- We propose a novel framework that consists of four streams. In order to obtain complementary features from each stream, we train a ResNet architecture in each stream by using a different representation of input images.
- This work is the first study employing the LZM transformation for ReId and training a deep convolutional neural network with the LZM pattern maps.
- Our framework[1] outperforms current state-of-the-art with a large margin by improving Rank-1/mAP by 29.79%/30.91% on SYSU-MM01 dataset, and by 9.73%/16.36% on RegDB dataset.

2. Related work

In recent years, deep convolutional neural networks (DCNN) have led to significant progress in many different computer vision areas. Person re-identification (ReId) is one of the challenging computer vision problems. In order to cope with issues such as illumination conditions, variations in pose, closure, and background clutter, researchers have developed many DCNN frameworks tackling ReId in different ways [8–12].

In the majority of the studies on ReId, only the images in the visible spectrum (RGB-based) are taken into account when addressing the challenges, and the developed methods are intended for RGB-based images only. There are only a few studies for visible–infrared or visible–thermal cross-modality person re-identification. However, they are important issues for video surveillance or miscellaneous applications that have to be performed in poorly illuminated or dark environments. In [1], the authors have analyzed different network structures for VI-ReId, including one-stream and two-stream architectures and asymmetric fully-connected layer. They have found the performance of the one-stream architecture to be better. In addition, they have proposed deep zero padding to contribute to the performance of the one-stream network. Since the authors train the architectures using grayscale input images, they do not take advantage of the color of RGB images. Kang et al. [5] propose a one-stream model for cross-modality re-identification. They create single input images by placing visible and infrared images in different channels or by concatenating them. Using these input images consisting of positive and negative pairs, they train a DCNN employing two-class classification loss. In the study, some pre-processing methods are also analyzed. Ye et al. [2] use a two-stream DCNN to obtain modality-specific features from the images. One stream of the network is fed with RGB while the other is fed with infrared (or thermal) images, and on top of the streams, there is a shared fully connected layer in order to learn a common embedding space. The authors train the two-stream model using multi-class classification loss along with a ranking loss proposed by them. A very similar two-stream model is employed in [3], but this model is trained utilizing contrastive loss instead of ranking loss used in [2]. To improve the discrimination ability of the features extracted by the two-stream model, they propose a hierarchical cross-modality metric learning. In this method, modality-specific and modality-shared metrics are jointly learned in order to minimize cross-modality inconsistency. In [13], a generative adversarial network (GAN) is proposed for visible–infrared re-identification. The authors use a DCNN, which is trained employing both multi-class classification and triplet losses, as a generator to extract features from RGB and infrared images. The discriminator is constructed as a modality classifier to distinguish between RGB and infrared representations. Another work utilizing a variant of GAN has introduced by Kniaz et al. [14] for visible–thermal re-identification. In their method, a set of synthetic thermal images are generated for an RGB probe image by employing a GAN framework. Then, the similarities between the synthetic probe and the thermal gallery images are calculated. Wang et al. [4] propose a network dealing with the modality discrepancy and appearance discrepancy separately. They first create a unified multi-spectral representation by projecting the visible and infrared images to a unified space. Then, they train a DCNN using the multi-spectral representations by employing triplet and multi-class classification losses. There are also some other works [8,15] trying to improve RGB-based ReId by fusing the information from different modalities.

In the literature, most of the existing studies on cross-modality matching or retrieval have been performed for heterogeneous face recognition (HFR) and text-to-image (or image-to-text) matching in the past decade. Studies on HFR, which are more relevant to VI-ReId, can be reviewed in three perspectives [16]: image synthesis, latent subspace, and domain-invariant features. In the first group, before the recognition, face images are transformed into the same domain [17–19]. The approach followed by the studies in the second group is the projection of the data of different domains into a common latent space [20–22]. In the recent works exploring domain-invariant feature representations, deep learning based methods are proposed. In [23], the features are obtained from a DCNN pre-trained on visible spectrum images, and metric learning methods are used to overcome inconsistencies between the modalities. He et al. [24] build a model in which the low-level layers are shared, and the high-level layers are divided into infrared, visible, and shared infrared–visible branches. Peng et al. [25] propose to generate features from facial patches and utilize a novel cross-modality enumeration loss while training the network. In [26], Mutual Component Analysis (MCA) [27] is integrated as a fully-connected layer into DCNN, and an MCA loss is proposed.

In this study, we propose a framework that utilizes the pattern maps generated by local Zernike moments (LZM) transformation [7]. As robust and holistic image descriptors, Zernike moments (ZM) [28] are commonly used in many different computer vision problems, i.e., character [29], fingerprint [30] and iris [31] recognition, and object detection [32]. LZM transformation is proposed by Sariyanidi et al. [7] to utilize ZM at the local scale for shape/texture analysis. In this transformation, images are encoded by calculating the ZM around each pixel. Thus, the local patterns in the images are exposed, and a rich representation is obtained. In literature, LZM transformation is used in various studies such as face recognition [7,33–35], facial expression [36,37] and facial affect recognition [38], traffic sign classification [39], loop closure detection [40,41], and interest point detector [42]. In these works, the features are obtained directly from the LZM pattern maps. Unlike them, in this study, we use the LZM pattern maps as input images to train DCNNs. This is the first study training DCNNs with the LZM pattern maps. However, there have been some attempts employing

[1] Project webpage: https://github.com/emrahbasaran/cross-re-id.

Table 1
ResNet architectures used in this study. The building blocks are given in brackets. Each row in the brackets indicates the kernel sizes and the number of output channels of convolution layers used in the building blocks.

50-layer	101-layer	152-layer
7×7, 64, stride 2		
3×3 max pool, stride 2		
$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 8$
$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
Global average pool		

ZM in convolutional networks. Mahesh et al. [43] initialize the trainable kernel coefficients by utilizing ZM with different moment orders. In [44] and [45], learning-free architectures are built constructing the convolutional layers by using ZM. Sun et al. [46] introduce a novel concept of Zernike convolution to extend convolutional neural networks for 2D-manifold domains.

3. Proposed method

In this section, we begin by giving the details of ResNet [6] architectures. Then, we describe the LZM transformation [7] and introduce our four-stream framework proposed for VI-ReId. Finally, we explain the ECN re-ranking algorithm [10] used as post-processing.

3.1. ResNet architecture

We show the performance of our proposed framework by using three different ResNet architectures which have different depths. These architectures have 50, 101, and 152 layers (we call these architectures as ResNet-50, ResNet-101, and ResNet-152, respectively, in the rest of the paper), and their details are provided in Table 1. Each ResNet model shown in Table 1 starts with a 7×7 convolution layer followed by a 3×3 max pooling. The next layers of the models consist of a different number of stacked building blocks with a global average pooling layer at their top. In Table 1, we show the building blocks in brackets where each row in the brackets indicates the kernel sizes and the number of output channels of convolution layers used in the building blocks.

3.2. LZM transformation

Zernike moments (ZM) [28] are commonly used to generate holistic image descriptors for various computer vision tasks [29–31]. Independent holistic characteristics of the images are exposed with the calculation of ZM of different orders [47]. In order to reveal the local shape and texture information from the images by utilizing ZM, local Zernike moments (LZM) transformation is proposed in [7].

Zernike moments of an image are calculated using orthogonal Zernike polynomials [28]. These polynomials are defined in polar coordinates within a unit circle and represented as follows:

$$V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{-\hat{j}m\theta}, \quad \hat{j} = \sqrt{-1} \tag{1}$$

where $R_{nm}(\rho)$ are real-valued radial polynomials, and ρ and θ are the radial coordinates calculated as

$$\rho_{ij} = \sqrt{x_i^2 + y_j^2}, \tag{2}$$

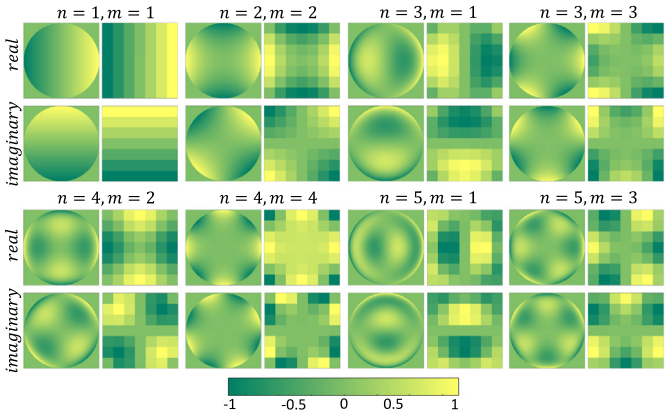

Fig. 1. Zernike polynomials (odd columns), and the corresponding 7×7 LZM filters (even columns) generated using the values up to $n = 5$, $m = 3$.

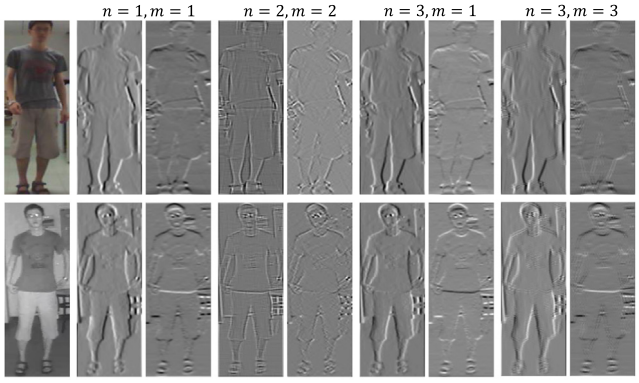

Fig. 2. Pattern maps generated using the LZM filters given in Fig. 1 up to $n = 3$. First column shows the RGB (first row) and infrared (second row) input images. Even and odd columns show the real and imaginary components, respectively. Pattern maps are normalized for better visualization.

$$\theta_{ij} = tan^{-1}(y_j/x_i). \tag{3}$$

Here, x_i and y_j are the pixel coordinates scaled to the range of $[-1, 1]$. $R_{nm}(\rho)$ polynomials are defined as

$$R_{nm}(\rho) = \sum_{k=0}^{\frac{n-|m|}{2}} (-1)^k \frac{(n-k)!}{k!(\frac{n+|m|-2k}{2})!(\frac{(n-|m|-2k)}{2})!}\rho^{n-2k}. \tag{4}$$

In (1) and (4), n and m are the moment order and the repetition, respectively. They take values such that $n - m$ is even, $0 \leq n$ and $0 \leq m \leq n$. Real and imaginary components of some $V_{nm}(\rho,\theta)$ Zernike polynomials are shown in Fig. 1. Finally, ZM of an image $f(i,j)$ are calculated as:

$$Z_{nm} = \frac{2(n+1)}{\pi(N-1)^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} V_{nm}(\rho_{ij},\theta_{ij}) f(i,j), \tag{5}$$

In the LZM transformation, images are encoded in $k \times k$ neighborhood of each pixel, as follows:

$$Z_{nm}^k(i,j) = \frac{2(n+1)}{\pi(k-1)^2} \sum_{p,q=-\frac{k-1}{2}}^{\frac{k-1}{2}} V_{nm}^k(p,q) f(i-p, j-q) \tag{6}$$

where $V_{nm}^k(p,q)$ represents the Zernike polynomials used as $k \times k$ filtering kernels and defined as

$$V_{nm}^k(i,j) = V_{nm}(\rho_{ij},\theta_{ij}). \tag{7}$$

We give real and imaginary components of some $V_{nm}^{k}(i,j)$ filters in Fig. 1. In Fig. 2, we show a few pattern maps generated using the LZM filters. The filters constructed using $m = 0$ are not used in the LZM transformation since the imaginary components of Zernike polynomials are not generated. According to the moment order n, the number of complex filters is calculated using

$$K(n) = \begin{cases} \frac{n(n+2)}{4} & \text{if } n \text{ is even,} \\ \frac{(n+1)^2}{4} & \text{if } n \text{ is odd.} \end{cases} \tag{8}$$

Since we create a pattern map with each of the real and imaginary components of complex filters, the total number of local pattern maps becomes $2 \times K(n)$.

3.3. Person re-identification framework

The proposed framework for VI-ReId is shown in Fig. 3. This framework includes four streams, and in each stream, a separate ResNet architecture is trained using both infrared and RGB images. Since each of the ResNets accepts a different representation of input images, different and complementary features are obtained from each stream.

There is a large visual difference between infrared and RGB images since infrared images have a single channel with invisible light information. This visual difference is reduced by converting RGB images to grayscale in the upper stream of the proposed framework, and the training is performed using infrared and grayscale images. In this way, $ResNet_{gray\_ir}$ in the upper stream of the framework learns to extract robust features for the person images by utilizing only the shape and texture information.

For ReId, one of the most important cues for person images is obtained from the color. Therefore, in the second stream of the proposed framework, $ResNet_{rgb\_ir}$ is trained using RGB and three-channel infrared (created by repeating the infrared channel) images. Unlike the one that is trained in the upper stream, the model trained in this stream uses color information of the RGB images and extracts discriminative features for the images in different modalities with visually large differences.

In VI-ReId, due to the lack of color in infrared images, the local shape and pattern information has great importance while matching the infrared and RGB images. Therefore, in addition to the two streams mentioned above, there are two other streams in the proposed framework in order to obtain stronger features related to local shape and pattern information. In the third stream, first, the local patterns are exposed by applying the LZM transformation on the grayscale and infrared images. Then, $ResNet_{LZM(gray\_ir)}$ is trained using the generated LZM pattern maps. As indicated in Section 3.2, a number of complex pattern maps are obtained as a result of the LZM transformation. The input tensor $I_{LZM(gray\_ir)}$ for $ResNet_{LZM(gray\_ir)}$ is prepared by concatenating the real and imaginary components of the complex maps. $I_{LZM(gray\_ir)}$ is denoted as

$$I_{LZM(gray\_ir)} = T_{LZM}(I_{gray/ir}) \tag{9}$$

where $I_{gray/ir}$ represents the grayscale or infrared image and $T_{LZM}(I)$ is defined as

$$T_{LZM}(I) = [re_{11}^{k}(I), im_{11}^{k}(I), re_{22}^{k}(I), im_{22}^{k}(I), \dots, re_{nm}^{k}(I), im_{nm}^{k}(I)]. \tag{10}$$

In Eq. (10), $re_{nm}^{k}(I)$ and $im_{nm}^{k}(I)$ are the real and imaginary components of a complex LZM pattern map, which is generated using the moment order n, the repetition m, and the filter size k. In the last stream of the proposed framework, LZM transformation is applied separately for the R, G and B channels, and the pattern maps obtained from these channels are concatenated to form the input tensor $I_{LZM(rgb\_ir)}$ for $ResNet_{LZM(rgb\_ir)}$, such as

$$I_{LZM(rgb\_ir)} = [T_{LZM}(I_{rgb}^{r}), T_{LZM}(I_{rgb}^{g}), T_{LZM}(I_{rgb}^{b})]. \tag{11}$$

Here, I_{rgb}^{r}, I_{rgb}^{g}, and I_{rgb}^{b} are the R, G, and B channels of the input image I. For infrared images, as mentioned earlier, the infrared channel is repeated for R, G, and B, such as

$$I_{LZM(rgb\_ir)} = [T_{LZM}(I_{ir}), T_{LZM}(I_{ir}), T_{LZM}(I_{ir})]. \tag{12}$$

In order to train the ResNet architectures in the proposed framework, multi-class classification with softmax cross-entropy loss is employed. Thus, in addition to learning to extract common features for the images in different modalities, the ResNet models also learn to extract features that express differences between individuals. As shown in Fig. 3, the loss for each stream is calculated separately during the training. In the evaluation phase, the final representations for the person images are constructed by concatenating the feature vectors obtained from each stream and normalized using ℓ_2-norm.

3.4. Re-ranking

In recent studies [10,48,49], it has been shown that re-ranking techniques have a significant contribution to the performance of person re-identification. Therefore, in this study, we utilize Expanded Cross Neighborhood (ECN) re-ranking algorithm [10] as a post-processing element to further improve the performance. In this algorithm, the distance between a probe image p and a gallery image g_i from a gallery set G with B images $G = \{g_i | i = 1, 2, \dots, B\}$ is defined as

$$ECN(p, g_i) = \frac{1}{2M} \sum_{j=1}^{M} d(pN_j, g_i) + d(g_i N_j, p). \tag{13}$$

Here, pN_j and $g_i N_j$ are the jth neighbors in the expanded neighbor sets $N(p, M)$ and $N(g_i, M)$ of the probe and ith gallery images, respectively. $d(\cdot)$ represents the distance between the images, and M is the total number of neighbors in a set. In order to construct the expanded neighborhood sets, first, an initial rank list $\mathcal{L}(p, G) = \{g_1^o, \dots, g_B^o\}$ in increasing order is created for each image by calculating the pairwise Euclidean distances between all images in the probe and gallery sets. Then, the expanded neighbor set $N(p, M)$ for a probe image p is given as

$$N(p, M) \leftarrow \{N(p, t), N(t, q)\} \tag{14}$$

where the set $N(p, t)$ consists of the t nearest neighbors of probe p, and the set $N(t, q)$ contains the q nearest neighbors of each of the images in $N(p, t)$ such that:

$$\begin{aligned} N(p, t) &= \{g_i^o | i = 1, 2, \dots, t\} \\ N(t, q) &= \{N(g_i^o, q), \dots, N(g_t^o, q)\} \end{aligned} \tag{15}$$

The expanded neighbor set $N(g_i, M)$ for gallery image g_i is obtained in the same way. In Eq. (13), as suggested by the authors of [10], we compute the distance between the pairs using a list comparison similarity measure proposed in [50] and defined as

$$R(\mathcal{L}_i, \mathcal{L}_j) = \sum_{b=1}^{B} [K + 1 - pos_i(b)]_+ \times [K + 1 - pos_j(b)]_+. \tag{16}$$

In this equation, $[\cdot]_+ = max(\cdot, 0)$, K is the number of nearest neighbors to be considered, and $pos_i(b)$ and $pos_j(b)$ show the position of image b in the rank lists $\mathcal{L}_i$ and $\mathcal{L}_j$, respectively. The distance d between the pairs is calculated as $d = 1 - R$ after scaling the range of the values of R between 0 and 1. For the parameters t, q, and K of ECN re-ranking algorithm, we use the same setting given in [10] such that $t = 3$, $q = 8$, and $K = 25$.

4. Experimental results

4.1. Datasets

In this work, we evaluate the proposed framework on two different cross-modality re-identification datasets, SYSU-MM01 [1] and RegDB [15]. Additionally, we perform experiments on Market-1501 [51] to expose the performance of the framework for re-identification in the visible domain.

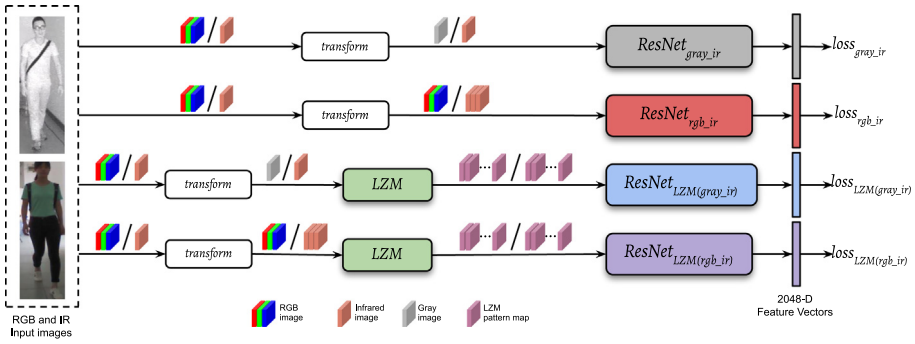

Fig. 3. Our proposed VI-ReId framework. There is no weight sharing between ResNets in the framework, and each ResNet is trained with a different representation of input images. In the first stream, grayscale and infrared input images are used, while RGB and three-channel infrared images (created by repeating infrared channel) are used in the second stream. In the other streams, LZM pattern maps are used as input images. These maps are obtained from grayscale and infrared images in the third stream and separately from R, G, and B channels in the fourth stream. We employ multi-class classification with softmax cross-entropy loss for each stream separately during the training. In the evaluation phase, the final representations are constructed by concatenating the feature vectors obtained from each stream. Best viewed in color.

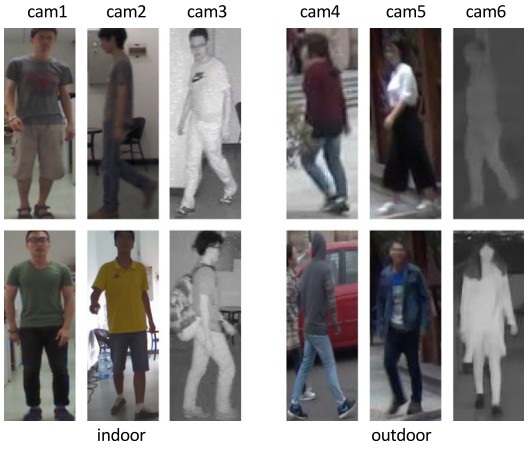

Fig. 4. Sample images from SYSU-MM01 dataset.

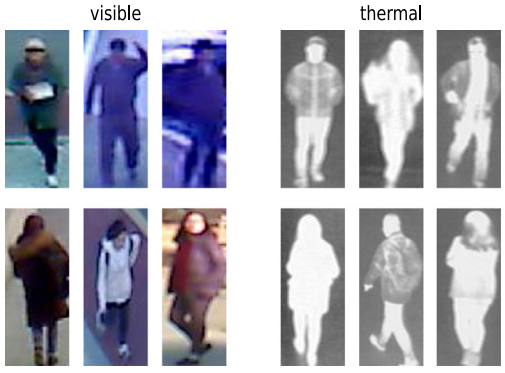

Fig. 5. Sample images from RegDB dataset.

4.1.1. SYSU-MM01

The SYSU-MM01 [1] is a dataset collected for VI-ReId problem. The images in the dataset were obtained from 491 different persons by recording them using four RGB and two infrared cameras. Within the dataset, the persons are divided into three fixed splits to create training, validation, and test sets. In the training set, there are 20284 RGB and 9929 infrared images of 296 persons. The validation set contains 1974 RGB and 1980 infrared images of 99 persons. The testing set consists of the images of 96 persons where 3803 infrared images are used as query and 301 randomly selected RGB images are used as gallery. For the evaluation, there are two search modes, all-search and indoor-search. In the all-search mode, the gallery set consists of the images taken from RGB cameras (cam1, cam2, cam4, cam5), and the probe set consists of the images taken from infrared cameras (cam3 and cam6). In the indoor-search mode, the gallery set contains the RGB images only taken from the cameras (cam1 and cam2) located in the indoor. In the standard evaluation protocol of the dataset, the performance is reported using cumulative matching characteristic (CMC) curves and mean average precision (mAP). CMC and mAP are calculated with 10 random splits of the gallery and probe sets, and the average performance is reported. We follow the single-shot setting for all the experiments. In Fig. 4, some sample images from SYSU-MM01 are shown.

4.1.2. RegDB

RegDB [15] consists of images captured using RGB and thermal dual-camera system. There are 412 persons in the dataset and each person has 20 images, 10 RGB and 10 thermal. We perform the experiments on this dataset by following the evaluation protocol proposed by Ye et al. [3]. In this protocol, the training and the test sets are created by randomly splitting the dataset into two parts. The gallery set is then created with thermal images, and the query set is created with RGB images. The results are reported with mAP and CMC, and the splitting process is repeated for 10 trials to ensure that the results are statistically stable. Sample images from RegDB are given in Fig. 5.

4.1.3. Market-1501

Market-1501 dataset [51] has 32668 images captured from 1501 persons using one low-resolution and five high-resolution cameras. In the standard evaluation protocol of Market-1501, 12936 images of 751 persons are reserved for the training operations. The images of the remaining 750 persons constitute the query and gallery sets with 3368 and 19732 images, respectively. CMC and mAP metrics are used to report the performance on this dataset, as in SYSU-MM01 and RegDB.

4.2. Training the networks

While training the ResNet networks in the proposed framework, we set the resolution of all RGB, infrared, and thermal input images to 492 × 164. In the experiments where the LZM transformation is used, we normalize the images to have zero mean. For SYSU-MM01 and Market-1501, we perform the training operations using the own training set of each dataset and set the maximum number of iterations to 100K. However, for RegDB, we follow a two-step training strategy similar to [11] to avoid overfitting. Since the training set of RegDB has a limited number of images, models easily over-fit the training set. In

Table 2
Results on SYSU-MM01 dataset obtained with different values of n and k parameters of the LZM transformation.

Method	n	k	Rank1	Rank10	Rank20	mAP
$R50_{LZM(gray\_ir)}$	2	5	30.77	78.74	89.67	32.28
	3	3	29.18	76.14	88.25	30.79
	3	5	31.45	**79.27**	**90.48**	32.57
	3	7	**31.48**	78.05	89.29	**32.64**
	4	5	30.35	78.79	89.93	31.92
	5	5	29.23	75.97	88.39	31.06

the first step, the networks are trained for 100K iterations using a new set created by combining the training sets of SYSU-MM01 and RegDB. In the second step, the networks are fine-tuned for 10K iterations using only the training set of RegDB. For all the training operations (including the fine-tuning on RegDB) performed in this work, we use Nesterov Accelerated Gradient [52] and set the mini-batch size to 8, momentum to 0.9 and weight decay to 0.0005. We choose 0.01 as the initial value of the learning rate and decay it 10 times using the exponential shift with the rate of 0.9, such that

$$lr\_new = lr\_init \times (1 - \frac{iter}{max\_iter})^{0.9}. \quad (17)$$

Here, lr_init and lr_new are the initial and the updated learning rates, respectively. $iter$ is the current number of iterations and max_iter is the maximum number of iterations.

4.3. Re-identification performance on SYSU-MM01

In this section, we call the ResNet models in the streams of the proposed framework as RX_{gray_ir}, RX_{rgb_ir}, $RX_{LZM(gray\_ir)}$, and $RX_{LZM(rgb\_ir)}$, respectively, to avoid confusion while reporting the experimental results. The X in the labels indicates which ResNet model is used and takes one of the values 50, 101, and 152. In this section, we show the experimental results under all-search mode.

As mentioned in Section 3.2, the number of filters used in the LZM transformation depends on the moment degree n (Eq. (8)). Therefore, we have first carried out experiments with model $R50_{LZM(gray\_ir)}$ in order to determine the optimal values for both the moment degree n and the filter size k. The results obtained with different parameter sets are given in Table 2. According to this table, the results are close to each other, but slightly better results are obtained when $3/5$ and $3/7$ are used for the n/k pair. For this reason, we prefer to use $n = 3$ and $k = 5$ to generate the LZM pattern maps for the rest of the study.

Table 3 shows the results obtained by training a ResNet-50 in each stream of the proposed framework. According to the results, better performance is achieved with $R50_{gray\_ir}$ and $R50_{LZM(gray\_ir)}$, where grayscale & infrared input images are used, compared to those of $R50_{rgb\_ir}$ and $R50_{LZM(rgb\_ir)}$ using RGB & infrared input images. Reducing the visual differences between infrared and RGB images by converting RGB images to grayscale makes a positive contribution to the performance. When the features obtained from the first two streams and the last two streams are combined, a significant performance increase is observed, as can be seen in the third and sixth rows of Table 3. With the concatenation of the features extracted by $R50_{gray\_ir}$ and $R50_{rgb\_ir}$, 4.5% improvement is achieved for Rank-1 as compared to $R50_{gray\_ir}$. Likewise, the concatenation of the features extracted by $R50_{LZM(gray\_ir)}$ and $R50_{LZM(rgb\_ir)}$ improves Rank1 by 3.3% as compared to $R50_{LZM(gray\_ir)}$. This shows that we generate complementary features with the models that are trained with grayscale & infrared and RGB & infrared input images. In the last row of Table 3, the results achieved by combining the features obtained in all streams of the proposed framework are given. When they are compared to those of $R50_{gray\_ir} + R50_{rgb\_ir}$ and $R50_{LZM(gray\_ir)} + R50_{LZM(rgb\_ir)}$, it is seen that the improvement obtained for Rank-1 is 5.3% and 3.7%, respectively. Therefore, taking this performance improvement into account, we can conclude that the models trained with the LZM pattern maps generate features that are different from and complementary to the ones produced by $R50_{gray\_ir}$ and $R50_{rgb\_ir}$.

Table 3
Results on SYSU-MM01 dataset obtained when ResNet-50 initialized with scaled Gaussian distribution [53] is used in all the streams of the framework.

Method	Rank1	Rank10	Rank20	mAP
$R50_{gray\_ir}$	28.59	75.95	87.57	30.42
$R50_{rgb\_ir}$	26.20	70.53	83.24	28.48
$R50_{gray\_ir} + R50_{rgb\_ir}$	33.06	79.67	89.87	35.11
$R50_{LZM(gray\_ir)}$	31.45	79.27	90.48	32.57
$R50_{LZM(rgb\_ir)}$	28.32	73.09	84.98	30.44
$R50_{LZM(gray\_ir)} + R50_{LZM(rgb\_ir)}$	34.73	81.13	91.26	36.30
All	**38.39**	**83.75**	**92.47**	**39.67**

Table 4
Results on SYSU-MM01 dataset obtained when pre-trained ResNet-50 is used in all the streams of the framework.

Method	Rank1	Rank10	Rank20	mAP
$R50_{gray\_ir}$	38.94	85.47	93.99	39.63
$R50_{rgb\_ir}$	33.87	76.99	88.66	35.21
$R50_{gray\_ir} + R50_{rgb\_ir}$	42.58	86.61	94.54	43.38
$R50_{LZM(gray\_ir)}$	37.30	84.56	93.60	37.93
$R50_{LZM(rgb\_ir)}$	35.09	80.11	91.18	36.67
$R50_{LZM(gray\_ir)} + R50_{LZM(rgb\_ir)}$	41.50	86.44	94.73	42.33
All	**45.00**	**89.06**	**95.77**	**45.94**

While performing the experiments whose results are given in Tables 2 and 3, we have initialized the weights of ResNet-50 with scaled Gaussian distribution proposed in [53]. However, many ReId studies [9–11] in the literature use pre-trained models instead of the models with randomly initialized weights. The vast majority of these pre-trained models are trained using ImageNet [54], which is a large-scale image classification dataset. In this way, the information of the model capable of image classification is utilized for ReId. By following the same approach, we have conducted experiments using ResNet-50 models, which are pre-trained on ImageNet, for each stream of our framework. The results are given in Table 4. In the third and fourth streams of the framework, the number of channels of the input images is greater than three because the LZM pattern maps are used. Therefore, in the experiments, we remove the first convolutional layer (conv1) of the pre-trained ResNet-50. Then, we insert a randomly initialized convolutional layer whose number of input channels matches the number of LZM pattern maps. For the first stream of the framework, we create three-channel input images by repeating the grayscale/infrared images for R, G, and B channels. If the results given in Tables 3 and 4 are compared, it is seen that we achieve a significant performance improvement for each stream by using pre-trained ResNet-50 models. With the concatenation of the features from four streams, we improve Rank-1 and mAP by 6.6% and 6.3% as compared to the results in the last row of Table 3 and reach to 45% and 45.9%, respectively.

According to recent research on computer vision tasks such as image classification, it is observed that deeper networks are more accurate [55]. In order to show the performance of our framework with deeper networks, we have conducted experiments using ResNet-101 and ResNet-152, which have more layers than ResNet-50. In the experiments, we have used the models pre-trained on ImageNet, and we show the results in Tables 5 and 6. As can be seen in the third and sixth rows of Tables 5 and 6, significant performance improvement is obtained with the concatenation of the features generated in the first two streams and in the last two streams, as in Tables 3 and 4. The highest results are achieved when the features from all the streams are combined. Using ResNet-101 and ResNet-152, we improve the best Rank1 result given in Table 4 (last row) by 1.8% and 2.4% and reach to 46.80% and 47.35%, respectively.

Table 5
Results on SYSU-MM01 dataset obtained when pre-trained ResNet-101 is used in all the streams of the framework.

Method	Rank1	Rank10	Rank20	mAP
$R101_{gray\_ir}$	40.41	85.02	93.80	41.21
$R101_{rgb\_ir}$	35.71	79.34	89.69	38.01
$R101_{gray\_ir} + R101_{rgb\_ir}$	43.70	86.89	94.71	45.35
$R101_{LZM(gray\_ir)}$	39.98	86.39	94.88	40.94
$R101_{LZM(rgb\_ir)}$	37.41	81.92	92.17	39.25
$R101_{LZM(gray\_ir)} + R101_{LZM(rgb\_ir)}$	43.58	88.35	96.02	44.94
All	**46.80**	**89.99**	**96.62**	**48.21**

Table 6
Results on SYSU-MM01 dataset obtained when pre-trained ResNet-152 is used in all the streams of the framework.

Method	Rank1	Rank10	Rank20	mAP
$R152_{gray\_ir}$	38.88	85.01	93.80	40.23
$R152_{rgb\_ir}$	36.27	78.63	88.74	38.33
$R152_{gray\_ir} + R152_{rgb\_ir}$	43.54	86.49	94.40	45.30
$R152_{LZM(gray\_ir)}$	40.68	84.89	93.32	41.02
$R152_{LZM(rgb\_ir)}$	39.65	81.69	91.36	40.49
$R152_{LZM(gray\_ir)} + R152_{LZM(rgb\_ir)}$	44.69	87.00	94.71	45.21
All	**47.35**	**89.10**	**95.67**	**48.32**

With the results given so far, we have demonstrated that the performance is significantly boosted by using the features learned from the LZM pattern maps. This shows that the models trained with the LZM pattern maps generate features that are different from and complementary to the ones generated by the other models. To further show the contribution of the features extracted from the LZM pattern maps, we have performed additional experiments and give the results in Table 7. In this table, $RX_{\{g+r\}}$ and $RX_{\{LZM(g+r)\}}$ represent the concatenation of the features of the first two and the last two streams, such that $RX_{\{g+r\}} = [RX_{gray\_ir}, RX_{rgb\_ir}]$ and $RX_{\{LZM(g+r)\}} = [RX_{LZM(gray\_ir)}, RX_{LZM(rgb\_ir)}]$. The first row of Table 7 shows the results obtained using ResNet-50 for each stream of the proposed framework. We have computed the results given in the second row of the table by using $R101_{gray\_ir}$ and $R101_{rgb\_ir}$ for the third and fourth streams of the framework instead of ResNet-50 models that utilize the LZM pattern maps. In this way, there is a 0.8% improvement for Rank-1. However, the improvement becomes 1.7% when $R101_{gray\_ir}$ and $R101_{rgb\_ir}$ are replaced with $R101_{LZM(gray\_ir)}$ and $R101_{LZM(rgb\_ir)}$, as shown in the third row. Similarly, better results are achieved when $R152_{LZM(gray\_ir)}$ and $R152_{LZM(rgb\_ir)}$ are used for the third and fourth streams compared to using $R152_{gray\_ir}$ and $R152_{rgb\_ir}$, as shown in the fourth and fifth rows. The results of the experiments, where the ResNet-50 models are used for the first and second streams of the framework, are given in the first five rows of Table 7. The next five rows and the last five rows show the results obtained using ResNet-101 and ResNet-152 for the first two streams, respectively. Like the results given in the first five rows, these results also demonstrate that the LZM transformation plays an important role. Using the models trained with the LZM pattern maps, better performance improvements are achieved compared to the other models with the same depth and trained with grayscale & infrared and RGB & infrared images. This verifies that, by exposing the texture information from the images, the LZM transformation enables the ResNet architectures to learn different as well as complementary features.

4.3.1. Comparison with the state-of-the-art

Tables 8 and 9 show our results in comparison with the state-of-the-art. In [1], zero-padding is utilized to enable a one-stream network to learn domain-specific structures automatically. TONE+HCML [3] and TONE+XQDA [3] use a two-stage framework that includes feature learning and metric learning. BCTR [2] and BDTR [2] have a two-stream framework that employs separate networks for RGB and infrared images to extract domain-specific features. Kang et al. [5] train a one-stream network using single input images generated by placing visible and infrared images in different channels or by concatenating them. In [13], a generative adversarial network is used in order to extract common features for the images in different domains. D^{2}RL [4] train a one-stream network with multi-spectral images generated by projecting RGB and infrared images to a unified space. Different from these works, in this study, we train multiple ResNet architectures by using the different representations of the input images and generate complementary features from each one of them for RGB and infrared images. As can be seen from Tables 8 and 9, we outperform the current state-of-the-art with a large margin. When we use ResNet-152 architectures in all the streams, our framework improves VI-ReId performance on SYSU-MM01 under all-search mode by 14.09%/12.10% and under indoor-search mode by 14.48%/16.01% in Rank-1/mAP. By concatenating all the feature vectors generated with ResNet-50, ResNet-101, and ResNet-152, we obtain at least 1.5% additional improvements for Rank-1 and mAP under both search modes. When we perform the re-ranking [10], the margin of the improvement further increases. We achieve 63.05%/67.13% and 69.06%/76.95% under all-search and indoor-search modes, respectively.

Table 7
Results on SYSU-MM01 dataset obtained using ResNet-50, ResNet-101 and ResNet-152 models in different combinations. $RX_{\{g+r\}}$ and $RX_{\{LZM(g+r)\}}$ represent the concatenation of the features from the first and second, and third and fourth streams, respectively.

Method	Rank1	Rank10	Rank20	mAP
$R50_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	45.00	89.06	95.77	45.94
$R50_{\{g+r\}} + R101_{\{g+r\}}$	45.80	88.51	95.53	46.94
$R50_{\{g+r\}} + R101_{\{LZM(g+r)\}}$	46.69	**90.32**	96.59	47.79
$R50_{\{g+r\}} + R152_{\{g+r\}}$	46.02	88.25	95.44	47.17
$R50_{\{g+r\}} + R152_{\{LZM(g+r)\}}$	**47.51**	89.59	96.01	48.15
$R101_{\{g+r\}} + R50_{\{g+r\}}$	45.80	88.51	95.53	46.94
$R101_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	46.23	89.22	96.11	47.42
$R101_{\{g+r\}} + R101_{\{LZM(g+r)\}}$	46.80	89.99	**96.62**	48.21
$R101_{\{g+r\}} + R152_{\{g+r\}}$	45.84	88.13	95.34	47.50
$R101_{\{g+r\}} + R152_{\{LZM(g+r)\}}$	47.24	89.12	96.09	48.27
$R152_{\{g+r\}} + R50_{\{g+r\}}$	46.02	88.25	95.44	47.17
$R152_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	46.57	89.16	95.97	47.75
$R152_{\{g+r\}} + R101_{\{g+r\}}$	45.84	88.13	95.34	47.50
$R152_{\{g+r\}} + R101_{\{LZM(g+r)\}}$	47.01	89.63	96.33	**48.37**
$R152_{\{g+r\}} + R152_{\{LZM(g+r)\}}$	47.35	89.10	95.67	48.32

Table 8
State-of-the-art comparison on SYSU-MM01 dataset under all-search mode.

Method	Rank1	Rank10	Rank20	mAP
Lin et al. [a] [56]	5.29	33.71	52.95	8.00
One-stream [1]	12.04	49.68	66.74	13.67
Two-stream [1]	11.65	47.99	65.50	12.85
Zero-padding [1]	14.80	54.12	71.33	15.95
TONE+XQDA[a] [3]	14.01	52.78	69.06	15.97
TONE+HCML[a] [3]	14.32	53.16	69.17	16.16
BCTR [2]	16.12	54.90	71.47	19.15
BDTR [2]	17.01	55.43	71.96	19.66
Kang et al. [5]	23.18	51.21	61.73	22.49
cmGAN [13]	26.97	67.51	80.56	27.80
eBDTR(ResNet50) [57]	27.82	67.34	81.34	28.42
D^{2}RL [4]	28.90	70.60	82.40	29.20
Ye et al. [58]	31.41	73.75	86.29	33.18
MAC [59]	33.26	79.04	90.09	36.22
$R50_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	45.00	89.06	95.77	45.94
$R101_{\{g+r\}} + R101_{\{LZM(g+r)\}}$	46.80	89.99	96.62	48.21
$R152_{\{g+r\}} + R152_{\{LZM(g+r)\}}$	47.35	89.10	95.67	48.32
All	48.87	90.73	**96.72**	49.85
All + re-ranking	**63.05**	**93.62**	96.30	**67.13**

[a]Indicates the results copied from [2].

Table 9
State-of-the-art comparison on SYSU-MM01 dataset under indoor-search mode.

Method	Rank1	Rank10	Rank20	mAP
Lin et al. [a] [56]	9.46	48.98	72.06	15.57
One-stream [1]	16.94	63.55	82.10	22.95
Two-stream [1]	15.60	61.18	81.02	21.49
Zero-padding [1]	20.58	68.38	85.79	26.92
cmGAN [13]	31.63	77.23	89.18	42.19
eBDTR(ResNet50) [57]	32.46	77.42	89.62	42.46
MAC [59]	33.37	82.49	93.69	44.95
Ye et al. [58]	37.62	83.27	93.56	46.32
$R50_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	49.66	92.47	97.15	59.81
$R101_{\{g+r\}} + R101_{\{LZM(g+r)\}}$	53.01	94.05	**98.44**	62.86
$R152_{\{g+r\}} + R152_{\{LZM(g+r)\}}$	52.10	93.69	98.06	62.33
All	54.28	94.22	98.22	63.92
All + re-ranking	**69.06**	**96.30**	97.16	**76.95**

[a]Indicates the results copied from [1].

4.4. Re-identification performance on RegDB

In the experiments on the RegDB dataset, we employ only the ResNet-50 models due to the relatively low number of training images. As noted in Section 4.2, we train these models in two steps. In the first step, the training is carried out using a set created by combining SYSU-MM01 and RegDB training sets. Then, in the second step, the models are fine-tuned using only the images from the RegDB dataset. With such a training strategy, we aim to avoid overfitting.

According to the results given in Table 10, the models trained with LZM pattern maps exhibit lower performance than the others. As shown in Figs. 1 and 2, LZM filters used perform high-pass filtering and reduce the low-frequency components of the images. This reduction results in LZM pattern maps to have less information compared to the original images. For this reason, in order to learn features with the same level of discriminating information (compared to using RGB or grayscale images), more training images will be needed in the training process performed with LZM pattern maps. In the first phase of the two-step training strategy, we expand the training set using the images from the SYSU-MM01 database. However, the images in the SYSU-MM01 and RegDB databases were recorded at different domains and so have different features from each other. Therefore, the models trained with the LZM pattern maps are not fed as many different images as they need in order to learn the specific features of the images in RegDB. As a result, models trained with the LZM pattern maps have lower performance than others. On the other hand, when the features obtained in all the streams of the proposed framework are combined, it is observed that Rank-1 and mAP are improved by 1.91% and 1.84%, respectively. This demonstrates that different and complementary features can be learned for also the RegDB dataset by using LZM pattern maps. Similar to [3], we have conducted additional experiments to evaluate the performance of our framework by following a different gallery/query setting where the RGB images are used as the gallery and the thermal images used as the query set. The results are given in Table 11. It is observed that the features of the models $R50_{LZM(rgb\_ir)}$ and $R50_{LZM(gray\_ir)}$ contribute to the performance by improving the Rank-1 and mAP by 2.32% and 2.22%, respectively.

4.4.1. Comparison with the state-of-the-art

In Table 12, our results are compared with the state-of-the-art. It is seen that we outperform the current state-of-the-art with a large margin by improving Rank-1 and mAP by 6.18% and 11.06%, respectively. The improvements become 9.73% and 16.36% when we perform the re-ranking.

Table 10
Results on RegDB dataset under visible to thermal setting.

Method	Rank1	Rank10	Rank20	mAP
$R50_{gray\_ir}$	43.94	67.28	77.50	45.28
$R50_{rgb\_ir}$	49.48	69.60	79.30	50.42
$R50_{gray\_ir} + R50_{rgb\_ir}$	55.12	75.37	83.85	56.22
$R50_{LZM(gray\_ir)}$	38.78	61.15	71.54	40.14
$R50_{LZM(rgb\_ir)}$	36.85	59.25	70.39	38.62
$R50_{LZM(gray\_ir)} + R50_{LZM(rgb\_ir)}$	44.11	65.94	75.84	45.70
All	**57.03**	**76.10**	**84.34**	**58.06**

Table 11
Results on RegDB dataset under thermal to visible setting.

Method	Rank1	Rank10	Rank20	mAP
$R50_{gray\_ir}$	45.62	69.13	78.80	45.51
$R50_{rgb\_ir}$	48.36	68.57	78.53	49.19
$R50_{gray\_ir} + R50_{rgb\_ir}$	54.85	74.05	82.91	55.34
$R50_{LZM(gray\_ir)}$	39.54	63.99	74.08	40.10
$R50_{LZM(rgb\_ir)}$	37.51	61.61	71.75	38.13
$R50_{LZM(gray\_ir)} + R50_{LZM(rgb\_ir)}$	45.18	67.50	77.57	45.53
All	**57.17**	**76.62**	**84.88**	**57.56**

Table 12
State-of-the-art comparison on RegDB dataset.

Method	Rank1	Rank10	Rank20	mAP
Lin et al. [a] [56]	17.28	34.47	45.26	15.06
One-stream[a] [1]	13.11	32.98	42.51	14.02
Two-stream[a] [1]	12.43	30.36	40.96	13.42
Zero-padding[a] [1]	17.75	34.21	44.35	18.90
TONE+XQDA[a] [3]	21.94	45.05	55.73	21.80
TONE+HCML [3]	24.44	47.53	56.78	20.80
BCTR [2]	32.67	57.64	66.58	30.99
BDTR [2]	33.47	58.42	67.52	31.83
eBDTR (AlexNet) [57]	34.62	58.96	68.72	33.46
Ye et al. [58]	35.42	53.75	64.08	36.42
MAC [59]	36.43	62.36	71.63	37.03
D^2RL [4]	43.40	66.10	76.30	44.10
D-HSME [60]	50.85	73.36	81.66	47.00
$R50_{\{g+r\}} + R50_{\{LZM(g+r)\}}$	57.03	**76.10**	**84.34**	58.06
+ re-ranking	**60.58**	67.71	77.13	**63.36**

[a]Indicates the results copied from [2].

4.5. Re-identification performance on Market-1501

In this section, we have performed experiments on the Market-1501 dataset to expose the performance of the LZM transformation for the ReId problem in the visible domain. In these experiments, we use the ResNet-50 model and give the results in Table 13. The performance of ReId in the visible domain depends on the effective use of both low and high-frequency information of the images. However, as mentioned in the previous section, the LZM filters used perform high-pass filtering by reducing the low-frequency components. Therefore, lower performance is obtained with the features extracted from the LZM pattern maps. Results given in Table 13 indicate that the features of $R50_{LZM(gray\_ir)}$ and $R50_{LZM(rgb\_ir)}$ do not contribute to the performance obtained with the features of $R50_{gray\_ir}$ and $R50_{rgb\_ir}$.

4.6. Discussion on execution time and memory utilization

We have used the Chainer framework [61] for the implementation of the ResNet models used in this study and carried out the experiments on a PC with an Intel Core i7-4790 CPU (3.60 GHz x 8), 32 GB RAM and an Nvidia GeForce GTX 1080Ti GPU. We have performed all the LZM calculations on the GPU. In this section, we show the execution time and the GPU memory utilization of the proposed framework by using ResNet-50 models.

Table 13
Results on Market-1501 dataset.

Method	Rank1	Rank10	Rank20	mAP
$R50_{gray}$	69.09	90.32	94.06	44.71
$R50_{rgb}$	85.18	96.05	97.33	68.81
$R50_{gray} + R50_{rgb}$	88.57	**96.85**	**98.28**	**72.91**
$R50_{LZM(gray)}$	60.96	87.38	91.30	35.86
$R50_{LZM(rgb)}$	76.31	92.22	95.31	53.36
$R50_{LZM(gray)} + R50_{LZM(rgb)}$	80.52	94.36	96.08	58.03
All	**88.78**	**96.85**	97.83	71.83

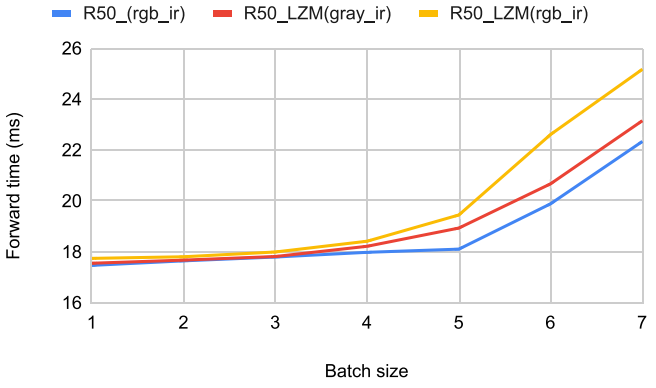

Fig. 6. Forward pass times of $R50_{rgb\_ir}$, $R50_{LZM(gray\_ir)}$, and $R50_{LZM(rgb\_ir)}$.

RX_{gray_ir} and RX_{rgb_ir} models are trained using the input images with 3 channels. For the other models, $RX_{LZM(gray\_ir)}$ and $RX_{LZM(rgb\_ir)}$, the input images have 8 and 24 channels, respectively. Therefore, in this section, we initially compare the forward pass times of $R50_{rgb\_ir}$, $R50_{LZM(gray\_ir)}$, and $R50_{LZM(rgb\_ir)}$. The results are given graphically in Fig. 6, where the graphs of $R50_{LZM(gray\_ir)}$ and $R50_{LZM(rgb\_ir)}$ show the total time spent on LZM transformation and ResNet-50. In Fig. 6, it is seen that there is no significant difference in the forward pass times of the models until the batch size is four. When the batch size is five or more, the difference begins to occur. This is because the GPU utilization is less than 100% for all the three models when the batch size is four or less. After GPU utilization reaches 100%, delays occur in $R50_{LZM(gray\_ir)}$ and $R50_{LZM(rgb\_ir)}$ compared to the $R50_{rgb\_ir}$.

Figs. 7a and 7b show forward pass time and GPU memory utilization when running 1, 2, 3, and 4 ResNet-50 models together on the same GPU. The last two of these models are $R50_{LZM(gray\_ir)}$ and $R50_{LZM(rgb\_ir)}$, respectively. According to Fig. 7a, when the batch size is one, the running of the four models increases the execution time by 1.7 times compared to the running of a single model. This rate increases if the number of images in the batch increases. However, since there is no weight sharing between the models, each model can perform feature extraction independently. Therefore, the execution time can be easily pulled down using more GPUs. When the batch size is one, a single ResNet-50 needs 827 MB of memory, and this memory requirement increases to 2541 MB when the batch size is seven. As shown in Fig. 7b, the memory requirement is directly proportional to the number of models running together. Due to the lack of enough GPU memory, we were unable to run four models together for larger batch sizes.

5. Conclusion

In this study, we have introduced a four-stream framework for VI-ReId using ResNet architectures. In each stream of the framework, we train a ResNet by using a different representation of input images in order to obtain complementary features as much as possible from each

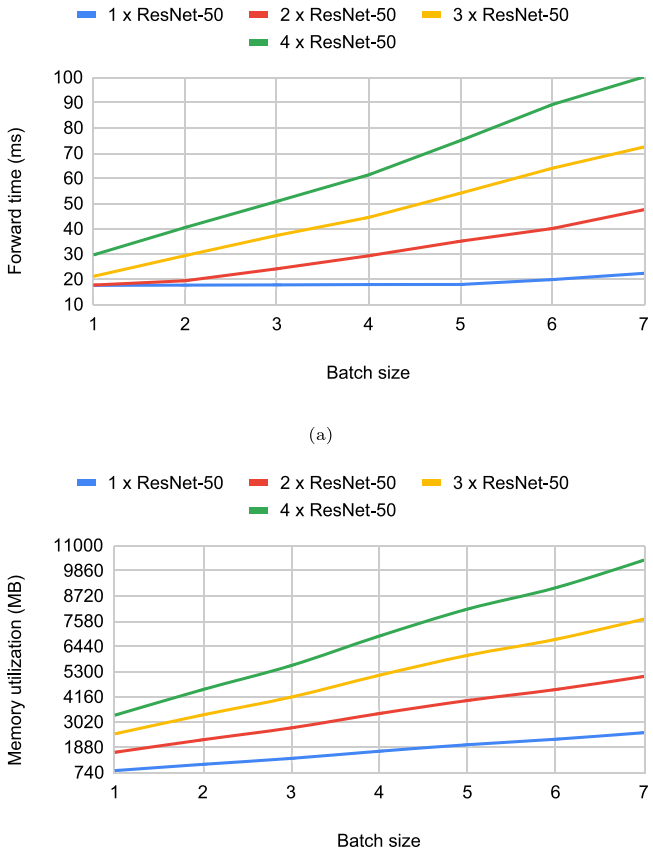

Fig. 7. Execution time (a) and memory utilization (b) when running 1, 2, 3, and 4 ResNet-50 models together on the same GPU. Due to the lack of enough GPU memory, we can only show the results when the batch size is up to 7.

stream. While grayscale and infrared input images are used to train the ResNet in the first stream, RGB and three-channel infrared images are used in the second stream. The first stream learns the features by using only the shape and texture information. The second stream uses the color information of the RGB images and learns to extract common features for the images with visually large differences. Unlike the first two streams, the input images in the other two streams are local pattern maps generated by employing the LZM transformation. Due to the lack of color, which provides the most important cues, in infrared images, the local shape and texture information is critical for VI-ReId. In the third and fourth streams, we expose this information from the images by generating the LZM pattern maps and train the ResNets using these maps as input images. With the exhaustive experiments performed employing three different ResNet architectures with different depths, we have demonstrated that each stream extracts different and complementary features and provides a significant contribution to the performance. Our framework outperforms, with a large margin, the current state-of-the-art on SYSU-MM01 and RegDB datasets. We further increase the improvement margin by utilizing re-ranking.

CRediT authorship contribution statement

Emrah Basaran: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing - original draft. **Muhittin Gökmen:** Methodology, Resources, Supervision, Validation, Writing - review & editing. **Mustafa E. Kamasak:** Resources, Supervision, Validation, Writing - review & editing.

Declaration of competing interest

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

References

[1] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.

[2] M. Ye, Z. Wang, X. Lan, P.C. Yuen, Visible thermal person re-identification via dual-constrained top-ranking, in: IJCAI, 2018, pp. 1092–1099.

[3] M. Ye, X. Lan, J. Li, P.C. Yuen, Hierarchical discriminative learning for visible thermal person re-identification, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[4] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S. Satoh, Learning to reduce dual-level discrepancy for infrared-visible person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 618–626.

[5] J.K. Kang, T.M. Hoang, K.R. Park, Person re-identification between visible and thermal Camera images based on deep residual cnn using single input, IEEE Access 7 (2019) 57972–57984.

[6] 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, 2016, pp. 770–778.

[7] E. Sarıyanidi, V. Dağlı, S.C. Tek, B. Tunc, M. Gökmen, Local zernike moments: A new representation for face recognition, in: 2012 19th IEEE International Conference on Image Processing, IEEE, 2012, pp. 585–588.

[8] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapes, S. Escalera, Tri-modal person re-identification with rgb, depth and thermal features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 301–307.

[9] Y. Chen, X. Zhu, S. Gong, Person re-identification by deep learning multi-scale representations, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2590–2600.

[10] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 420–429.

[11] M.M. Kalayeh, E. Basaran, M. Gökmen, M.E. Kamasak, M. Shah, Human semantic parsing for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1062–1071.

[12] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 994–1003.

[13] P. Dai, R. Ji, H. Wang, Q. Wu, Y. Huang, Cross-modality person re-identification with generative adversarial training, in: IJCAI, 2018, pp. 677–683.

[14] V.V. Kniaz, V.A. Knyaz, J. Hladuvka, W.G. Kropatsch, V. Mizginov, Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.

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

[16] R. He, J. Cao, L. Song, Z. Sun, T. Tan, Cross-spectral face completion for nir-vis heterogeneous face recognition, 2019, arXiv preprint arXiv:1902.03565.

[17] L. Song, M. Zhang, X. Wu, R. He, Adversarial discriminative heterogeneous face recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[18] J. Lezama, Q. Qiu, G. Sapiro, Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6628–6637.

[19] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.

[20] Z. Lei, S. Liao, A.K. Jain, S.Z. Li, Coupled discriminant analysis for heterogeneous face recognition, IEEE Trans. Inf. Forensics Secur. 7 (6) (2012) 1707–1716.

[21] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194.

[22] Y. Jin, J. Li, C. Lang, Q. Ruan, Multi-task clustering elm for vis-nir cross-modal feature learning, Multidimens. Syst. Signal Process. 28 (3) (2017) 905–920.

[23] S. Saxena, J. Verbeek, Heterogeneous face recognition with cnns, in: European Conference on Computer Vision, Springer, 2016, pp. 483–491.

[24] R. He, X. Wu, Z. Sun, T. Tan, Wasserstein cnn: Learning invariant features for nir-vis face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1761–1773.

[25] C. Peng, N. Wang, J. Li, X. Gao, Dlface: Deep local descriptor for cross-modality face recognition, Pattern Recognit. 90 (2019) 161–171.

[26] Z. Deng, X. Peng, Z. Li, Y. Qiao, Mutual component convolutional neural networks for heterogeneous face recognition, IEEE Trans. Image Process. 28 (6) (2019) 3102–3114.

[27] Z. Li, D. Gong, Q. Li, D. Tao, X. Li, Mutual component analysis for heterogeneous face recognition, ACM Trans. Intell. Syst. Technol. (TIST) 7 (3) (2016) 28.

[28] M.R. Teague, Image analysis via the general theory of moments, JOSA 70 (8) (1980) 920–930.

[29] C. Kan, M.D. Srinath, Invariant character recognition with zernike and orthogonal fourier-mellin moments, Pattern Recognit. 35 (1) (2002) 143–154, http://dx.doi.org/10.1016/S0031-3203(00)00179-5.

[30] H.L. Zhai, F. Di Hu, X.Y. Huang, J.H. Chen, The application of digital image recognition to the analysis of two-dimensional fingerprints, Anal. Chim. Acta 657 (2) (2010) 131–135.

[31] C.-W. Tan, A. Kumar, Accurate iris recognition at a distance using stabilized iris encoding and zernike moments phase features, IEEE Trans. Image Process. 23 (9) (2014) 3962–3974.

[32] A. Bera, P. Klesk, D. Sychel, Constant-time calculation of zernike moments for detection with rotational invariance, IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2018) 537–551.

[33] T. Alasag, M. Gokmen, Face recognition in low resolution images by using local Zernike moments, in: Proceedings of the International Conference on Machine Vision and Machine Learning, Beijing, China, 2014, pp. 21–26.

[34] E. Basaran, M. Gökmen, M.E. Kamasak, An efficient multiscale scheme using local zernike moments for face recognition, Appl. Sci. 8 (5) (2018) 827.

[35] E. Basaran, M. Gokmen, An efficient face recognition scheme using local zernike moments (LZM) patterns, in: Asian Conference on Computer Vision, Springer, 2014, pp. 710–724.

[36] X. Fan, T. Tjahjadi, A dynamic framework based on local zernike moment and motion history image for facial expression recognition, Pattern Recognit. 64 (2017) 399–406.

[37] B.S.A. Gazioğlu, M. Gökmen, Facial expression recognition from still images, in: International Conference on Augmented Cognition, Springer, 2017, pp. 413–428.

[38] E. Sariyanidi, H. Gunes, M. Gökmen, A. Cavallaro, Local zernike moment representation for facial affect recognition, in: BMVC, Vol. 2, 2013, p. 3.

[39] E. Başaran, M. Gökmen, Traffic sign classification with quantized local zernike moments, in: 2013 21st Signal Processing and Communications Applications Conference (SIU), IEEE, 2013, pp. 1–4.

[40] E. Sariyanidi, O. Sencan, H. Temeltas, Loop closure detection using local zernike moment patterns, in: Intelligent Robots and Computer Vision XXX: Algorithms and Techniques, Vol. 8662, International Society for Optics and Photonics, 2013, p. 866207.

[41] C. Erhan, E. Sariyanidi, O. Sencan, H. Temeltas, Patterns of approximated localised moments for visual loop closure detection, IET Comput. Vis. 11 (3) (2016) 237–245.

[42] G. Özbulak, M. Gökmen, A rotation invariant local zernike moment based interest point detector, in: Seventh International Conference on Machine Vision (ICMV 2014), Vol. 9445, International Society for Optics and Photonics, 2015, p. 94450E.

[43] V.G. Mahesh, A.N.J. Raj, Z. Fan, Invariant moments based convolutional neural networks for image analysis, Int. J. Comput. Intell. Syst. 10 (1) (2017) 936–950.

[44] Y. Yoon, L.-K. Lee, S.-Y. Oh, Semi-rotation invariant feature descriptors using zernike moments for mlp classifier, in: Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE, 2016, pp. 3990–3994.

[45] J. Wu, S. Qiu, Y. Kong, Y. Chen, L. Senhadji, H. Shu, Momentsnet: A simple learning-free method for binary image recognition, in: Image Processing (ICIP), 2017 IEEE International Conference on, IEEE, 2017, pp. 2667–2671.

[46] Z. Sun, J. Lu, S. Baek, Zernet: Convolutional neural networks on arbitrary surfaces via zernike local tangent space estimation, 2018, arXiv preprint arXiv:1812.01082.

[47] C.-W. Chong, P. Raveendran, R. Mukundan, A comparative analysis of algorithms for fast computation of zernike moments, Pattern Recognit. 36 (3) (2003) 731–742.

[48] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, R. Hu, Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing, IEEE Trans. Multimed. 18 (12) (2016) 2553–2566.

[49] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[50] R.A. Jarvis, E.A. Patrick, Clustering using a similarity measure based on shared near neighbors, IEEE Trans. Comput. 100 (11) (1973) 1025–1034.

[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[52] Y. Bengio, N. Boulanger-Lewandowski, R. Pascanu, Advances in optimizing recurrent networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8624–8628.

[53] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[55] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, J.E. Gonzalez, Skipnet: Learning dynamic routing in convolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 409–424.

[56] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1089–1102.

[57] M. Ye, X. Lan, Z. Wang, P.C. Yuen, Bi-directional center-constrained top-ranking for visible thermal person re-identification, IEEE Trans. Inf. Forensics Secur. 15 (2019) 407–419.

[58] M. Ye, Y. Cheng, X. Lan, H. Zhu, Improving night-time pedestrian retrieval with distribution alignment and contextual distance, IEEE Trans. Ind. Inf. 16 (1) (2020) 615–624.

[59] 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, 2019, pp. 347–355.

[60] Y. Hao, N. Wang, J. Li, X. Gao, HSME: hypersphere manifold embedding for visible thermal person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8385–8392.

[61] S. Tokui, K. Oono, S. Hido, J. Clayton, Chainer: a next-generation open source framework for deep learning, in: Proceedings of Workshop on Machine Learning Systems (LearningSys) in the Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS), Vol. 5, 2015, pp. 1–6.