RGB-Infrared Cross-Modality Person Re-Identification

Ancong Wu[1] , Wei-Shi Zheng[2,3*] , Hong-Xing Yu[2] , Shaogang Gong[5] , and Jianhuang Lai[2,4]
[1]School of Electronics and Information Technology, Sun Yat-sen University, China
[2]School of Data and Computer Science, Sun Yat-sen University, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Guangdong Province Key Laboratory of Information Security, China
[5]Queen Mary University of London, United Kingdom
wuancong@mail2.sysu.edu.cn, wszheng@ieee.org, xKoven@gmail.com,
s.gong@qmul.ac.uk, stsljh@mail.sysu.edu.cn

Abstract

Person re-identification (Re-ID) is an important problem in video surveillance, aiming to match pedestrian images across camera views. Currently, most works focus on RGB-based Re-ID. However, in some applications, RGB images are not suitable, e.g. in a dark environment or at night. Infrared (IR) imaging becomes necessary in many visual systems. To that end, matching RGB images with infrared images is required, which are heterogeneous with very different visual characteristics. For person Re-ID, this is a very challenging cross-modality problem not studied so far. In this work, we address the RGB-IR cross-modality Re-ID problem and contribute a new multi-modality Re-ID dataset named SYSU-MM01, including RGB and IR images of 491 identities from 6 cameras, giving in total 287,628 RGB images and 15,792 IR images. To explore the RGB-IR Re-ID problem, we evaluate existing popular cross-domain models, including three commonly used neural network structures (one-stream, two-stream and asymmetric FC layer) and analyse the relation between them. We further propose deep zero-padding for training one-stream network towards automatically evolving domain-specific nodes in the network for cross-modality matching. Our experiments show that RGB-IR cross-modality matching is very challenging but still feasible using the proposed model with deep zero-padding giving the best performance. Our dataset is available at http://isee.sysu.edu.cn/project/RGBIRReID.htm.

1. Introduction

Person re-identification (Re-ID) is an important field in visual surveillance. A large number of models for RGB-based Re-ID problem have been proposed in literature, including feature learning [27, 44, 21], distance metric learning [50, 14, 20, 26, 21, 22], and end-to-end learning [19, 1, 43]. Most Re-ID methods are developed based on RGB-RGB matching, which is the most common single-modality Re-ID problem.

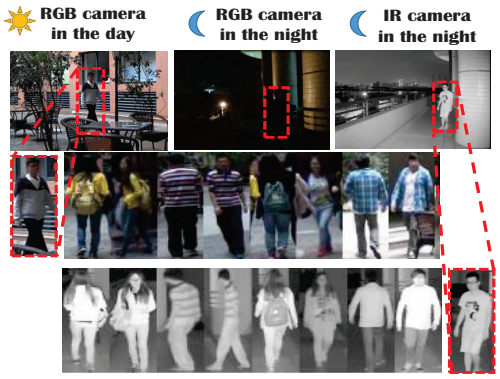

Figure 1. Examples of RGB images and infrared (IR) images captured in two outdoor scenes in the day time and in the night , respectively. The images in every two columns are of the same person. Captured by devices receiving light of different wavelength, RGB images and IR images of the same person look very different.

However, RGB-RGB Re-ID can be limited in surveillance when lighting is either poor or unavailable, e.g. at night, RGB images become uninformative (Figure 1). In such case, imaging devices regardless of visible light should be applied. Infrared (IR) cameras are commonly used in video surveillance systems. Most cameras are able to automatically switch from RGB to IR mode in the dark. Only a few Re-ID methods consider using IR images. Jungling et al. [12] proposed an IR-IR video matching method for Re-ID, which can only be applied in the dark. Depth images captured by RGB-D cameras such as Kinect are also regardless of visible light, but they are rarely deployed because they are more expensive, used indoor only and with

*Corresponding author

distance limitations. Thus, it is necessary to study RGB-IR cross-modality matching in 24-hour surveillance systems.

RGB and IR images are intrinsically distinct. See Figure 1, RGB images in the first row have three channels containing colour information of visible light, while IR images in the second row have one channel containing information of invisible light. Thus, they can be regarded as heterogeneous data. We call this problem the *RGB-IR cross-modality Re-ID* problem. Despite RGB-IR Re-ID is common and significant in real-world applications, to our best knowledge it is rarely explored and remains an open issue.

RGB-IR Re-ID is a very challenging problem due to the great differences between two modalities. From imaging principle aspect, the wavelength range of RGB and IR images is different. In existing Re-ID works, colour information is the most important appearance cue for identifying persons. However, in the RGB-IR Re-ID problem, this cue can hardly be used. As shown in Figure 1, even human can hardly recognise the persons by colour information. This leads to severe data misalignment within the same class. Moreover, viewpoint change, pose and exposure problems which cause large intra-class discrepancy in RGB-based Re-ID also bring difficulties to RGB-IR cross-modality Re-ID, resulting in a much more challenging problem.

In this work, we first identify the challenge of RGB-IR by conducting extensive evaluations on existing popularly used cross-modality methods. For this purpose, we have collected a new dataset called RGB-IR Re-ID dataset. The comparison with existing most commonly used Re-ID datasets is shown in Table 1. It contains 287,628 RGB images and 15,792 IR images of 491 persons captured in 6 cameras. To our best knowledge, this new RGB-IR Re-ID dataset provides for the first time a meaningful benchmark for the study of cross-modality RGB-IR Re-ID problem.

For cross-modality matching tasks, domain-specific modelling is important for extracting shared features for matching because of the domain shift. Considering using neural networks for cross-modality matching, we investigate and analyse the relation between different neural network structures, including two-stream structure and asymmetric FC layer structure, in which the domain-specific modelling exists but is designed artificially. Alternatively, we propose a deep zero-padding method for training one-stream network tending to evolve domain-specific structures automatically. Extensive experiments show the effectiveness of deep zero-padding, which outperforms the compared hand-crafted feature and deep models.

The contributions of this paper are: (1) We contribute for the first time a standard benchmark SYSU-MM01 for supporting the study of RGB-IR cross-modality Re-ID. We conducted extensive experiments to evaluate popular baseline deep learning architectures for cross-modality RGB-IR Re-ID. (2) We analyse three different network structures (one-stream network, two-stream network and asymmetric FC layer structure) and give insights on their effectiveness for RGB-IR Re-ID. (3) We propose *deep zero-padding* for evolving domain-specific structure automatically in one-stream network optimized for RGB-IR Re-ID tasks. Our experiments show that this approach for RGB-IR cross-modality Re-ID outperforms not only a standard one-stream network but also a two-stream network with explicit cross-domain learning and extra computational costs.

Table 1. Comparison between SYSU-MM01 with existing Re-ID datasets. (-/- denotes the RGB#/IR#.)

Datasets	ID#	images#	cameras#	RGB	IR
VIPER [7]	632	1,264	2	yes	no
iLIDS [49]	119	476	2	yes	no
CAVIAR [5]	72	610	2	yes	no
PRID2011 [10]	200	971	2	yes	no
CUHK01 [18]	972	1,942	2	yes	no
SYSU [8]	502	24,448	2	yes	no
CUHK03 [19]	1467	13,164	6	yes	no
Market [48]	1501	32,668	6	yes	no
MARS [47]	1261	1,191,003	6	yes	no
SYSU-MM01	491	287,628/15,792	6	yes	yes

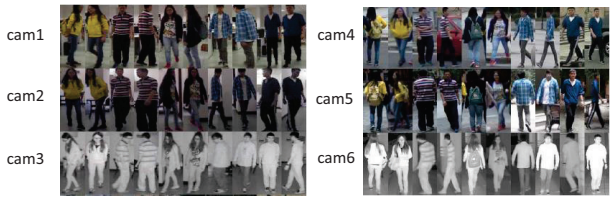

Figure 2. Examples of RGB images and infrared (IR) images in our SYSU-MM01 dataset. Cameras 1-3 on the left are indoor scenes and cameras 4-6 on the right are outdoor scenes. Every two columns are of the same person.

2. SYSU-MM01 Dataset

2.1. Dataset Description

SYSU-MM01 contains images captured by 6 cameras, including two IR cameras and four RGB ones. Different from RGB cameras, IR cameras work in dark scenarios. We show the details in Table 2, and some samples from each camera view in Figure 2. RGB images of camera 1 and camera 2 were captured in two bright indoor rooms (room 1 and room 2) by Kinect V1. For each person, there are at least 400 continuous RGB frames with different poses and viewpoints. IR images of camera 3 and camera 6 are captured by IR cameras in the dark. The IR images have only one channel and is different from 3-channel RGB images. Camera 3 is placed in room 2 in dark environment, while camera 6 is placed in an outdoor passage with background clutters. Camera 4 and 5 are RGB surveillance cameras placed in two outdoor scenes named gate and garden.

Observing the samples of the dataset, we can see clearly that the images of IR cameras (camera 3 and 6) are distinct

Table 2. SYSU-MM01 dataset overview.

Cam	location	(in/out)door	lighting	ID#	RGB#/ID	IR#/ID
1	room1	indoor	bright	259	400+	-
2	room2	indoor	bright	259	400+	-
3	room2	indoor	dark	486	-	20
4	gate	outdoor	bright	493	20	-
5	garden	outdoor	bright	502	20	-
6	passage	outdoor	dark	299	-	20

from those of RGB ones, in terms of both colour and contrast. Specifically, although camera 2 and 3 are in the same scenario, the images of them suffer from dramatic colour shift and contrast degrading. For example, the first person's yellow clothes is distinct from her black trousers under the RGB camera, but this colour distinction is nearly eliminated under IR camera (Column 1,2, Row 2,3 in Figure 2). Moreover, IR image has only one channel and might lose some texture details. The exposure of IR image is also an issue when it is captured at different distances. These all introduce difficulty for RGB-IR cross-modality Re-ID problem.

2.2. Evaluation Protocol

There are 491 valid IDs in SYSU-MM01 dataset. We have a fixed split using 296 identities for training, 99 for validation and 96 for testing. During training, all images of the 296 persons in training set in all cameras can be applied.

In the testing stage, samples from RGB cameras are for gallery set, and those from IR cameras are for probe set. We design two modes, *all-search* mode and *indoor-search* mode. For all-search mode, RGB cameras 1, 2, 4 and 5 are for gallery set and IR cameras 3 and 6 are for probe set. For indoor-search mode, RGB cameras 1 and 2 (excluding outdoor cameras 4 and 5) are for gallery set and IR cameras 3 and 6 are for probe set, which is less challenging.

For both modes, we adopt single-shot and multi-shot settings. That is, for every identity under an RGB camera, we randomly choose one/ten image(s) of the identity to the gallery set for single-shot/multi-shot setting. As for probe set, all images are used. Given a probe image, matching is conducted by computing similarities between the probe image and gallery images. Notice that matching is conducted between cameras in different locations (locations are shown in Table 2). Camera 2 and camera 3 are in the same location, so probe images of camera 3 skip the gallery images of camera 2. After computing similarities, we can get a ranking list according to descending order of similarities.

For indicating the performance, we use Cumulative Matching Characteristic (CMC) [30] and mean average precision (mAP). Notice that, for CMC under multi-shot setting, only the maximum similarity in all gallery images of the same person is taken to compute the rank list. We repeat the above evaluation 10 times with random split of gallery and probe set and compute the average performance finally.

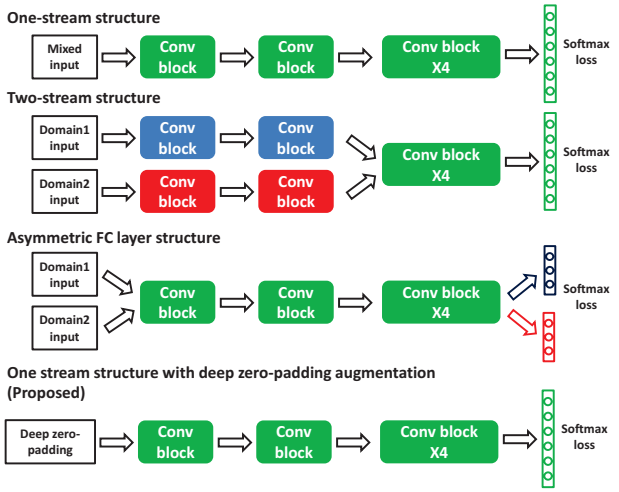

Figure 3. Four types of network structures in our evaluation. The structure of conv blocks depends on the selected base network (ResNet [9] in our evaluation). The colour of conv blocks and FC layers indicates whether the parameters are shared or not. Red and blue means specific and green means shared.

3. Network Structure Comparison on Cross-Modality Modelling

We investigate deep learning network architectures for the task of RGB-IR cross-modality Re-ID. In particular, we examine three commonly adopted network structures for visual recognition and cross-modality learning. We further exploit the idea of *deep zero-padding* for model training and give insights on its impact on cross-modality matching task.

3.1. Common Deep Model Network Structures

In the past few years, a large number of deep models have been proposed for visual matching and cross-modality modelling, and have achieved satisfactory performance in many tasks. The most commonly used structures can mainly be categorized into 3 types. All structures that we are going to discuss are shown in Figure 3.

One-stream Structure. One-stream structure is the most commonly used in vision tasks. As shown in the first network in Figure 3, there is single input and all parameters are shared in the whole network. Representative networks include AlexNet [15], VGG [36], GoogleNet [38], ResNet [9] and so on, which perform well in classification, detection, tracking and many other tasks. In the field of Re-ID, JSTL-DGD [43], one of the state-of-the-art network, uses one-stream structure as well. Generally, in these tasks, the inputs to the network are RGB images, which are of the same modality. So sharing all parameters in the network is appropriate for these tasks.

Two-stream Structure. Two-stream structure is commonly used in cross-modality matching tasks. As shown in the second network in Figure 3, there are two inputs, corre-

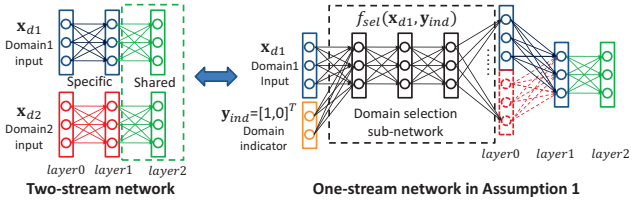

Figure 4. Explanation of how one-stream network can represent two-stream network in Assumption 1 with domain indicator and domain selection sub-network in forward propagation (best viewed in colour).

sponding to data in two different domains. In the shallower layers, the parameters of network are specific for each domain. In the deeper layers, shared parameters are used. The generalized similarity net [24] proposed by Lin *et al.* for cross-domain visual matching including the Re-ID task is one of the representative structure of this type. Two-stream structure is also favorable in Re-ID tasks, for example, Ahmed's net [1], SIR-CIR net [40], gated siamese net [39], etc. Note that except for Lin's structure [24], most of them prefer sharing parameters in domain-specific layers. This is not exactly identical to our definition of two-stream structure. The reason may be, although the images are from different cameras, they are all of the same modality of RGB images. Different from one-stream structure for single-modality tasks, two-stream structure does two things, domain adaptation and discriminative feature learning. It is assumed that the domain-specific network can extract shared features for different domains, and then the shared network can extract discriminative features for matching.

Asymmetric FC Layer Structure. Asymmetric FC layer model is also used in multi-domain tasks, for example, MDNet [31] for multi-domain tracking, CVDCA [2] for Re-ID, etc. As shown in the third network in Figure 3, the structure shares nearly all parameters except for the last FC layer. This design assumes that the feature extraction for different domains can be the same and domain adaptation is achieved in feature level. This order of feature extraction and domain adaptation is different from two-stream structure.

3.2. Analysis of Network Structures

– Connection of One-stream and Two-stream Structures in special case. The three structures discussed above seem to be different, we find interestingly that all structures can be represented by one-stream structure in the *forward propagation process* when the following assumption is hold:

Assumption 1. *A domain selection sub-network would exist somewhere in a network, which can automatically select samples of the corresponding domain as input. And domain selection sub-network is fixed.*

Under Assumption 1, we firstly give a simple example how one-stream network can perform as two-stream network in forward propagation. As shown in Figure 4, on the left is a simplified two-stream network: two fully connected networks, each with a specific layer (blue and red) and a shared layer (green). On the right is a one-stream network which can be conditionally equivalent to the two-stream one in forward propagation, in which there is a domain selection sub-network for selecting the following domain-specific structure. We first define some symbols for illustration. Let $\mathbf{x}_{d1} \in \mathbb{R}^d$ and $\mathbf{x}_{d2} \in \mathbb{R}^d$ denote the input of domain1 and domain2, respectively. We define a domain indicator $\mathbf{y}_{ind}$ as a vector with two elements, of which the value is $[1,0]^T$ or $[0,1]^T$ indicating domain1 or domain2, respectively. Let $f_{sel}(\mathbf{x}, \mathbf{y}_{ind})$ denote the domain selection sub-network, implementing the following function:

$$f_{sel}(\mathbf{x}, \mathbf{y}_{ind}) = \begin{cases} [\mathbf{I}_d, \mathbf{O}_d]^T \mathbf{x}, & \mathbf{y}_{ind} = [1,0]^T \\ [\mathbf{O}_d, \mathbf{I}_d]^T \mathbf{x}, & \mathbf{y}_{ind} = [0,1]^T. \end{cases} \quad (1)$$

The equation above suggests that if the domain selection sub-network is fixed, the two-stream network can be represented by one-stream network in forward propagation.

– Analysis of One-stream Structure in General Case. The assumption we hope above is less feasible. Now, we drop this assumption and analyse the domain-specific property of one-stream network. For cross-modality matching tasks, domain-specific modelling is important for extracting shared components for matching because of domain shift. Generally, in neural networks, *e.g.*, two-stream and asymmetric FC layer structure, this is modelled by domain-specific structures. Thus we intend to analyse the domain-specific modelling in one-stream network. Our analysis is based on the following relaxed assumption:

Assumption 2. *As shown in Figure 5, for a one-stream network dealing with inputs of two domains, we categorize the output nodes of each layer into three types, domain1-specific nodes, domain2-specific nodes and shared nodes. The categorization depends on whether the response of the node is domain-specific. Let $\mathbf{x}_{d1}^{(l)}$ and $\mathbf{x}_{d2}^{(l)}$ denote the input to layer $l+1$ of domain1 and domain2, respectively. For example, $\mathbf{x}_{d1}^{(0)}$ and $\mathbf{x}_{d2}^{(0)}$ are inputs of the whole network. Let $\eta_i^{(l)}$ denote the i-th node in layer l and $f_{out}(\mathbf{x}^{(0)}, i, l)$ denote the output of $\eta_i^{(l)}$ with the network input $\mathbf{x}^{(0)}$, we have:*

$$f_{out}(\mathbf{x}^{(0)}, i, l) = \sigma(\sum_j w_{j,i}^{(l-1)} f_{out}(\mathbf{x}^{(0)}, j, l-1) + b_i^{(l-1)}), \quad (2)$$

where $\sigma(\cdot)$ is the activation function, $w_{j,i}^{(l-1)}$ and $b_i^{(l-1)}$ are weight and bias parameters of layer $l-1$. The type of node $\eta_i^{(l)}$ is defined by

$$type(\eta_i^{(l)}) = \begin{cases} domain1-specific, & f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) \equiv 0 \\ domain2-specific, & f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) \equiv 0 \\ shared, & otherwise. \end{cases} \quad (3)$$

For domain1-specific nodes, we use identity sign in

$f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) \equiv 0$, *which means that for any input of domain2, the output of node* $\eta_i^{(l)}$ *is always zero.*

Under Assumption 2, we define some symbols for analysis. Let L denote the loss function. Let $o_i^{(l+1)}$ denote the output of the i-th node before activation function in layer $l+1$, $\mathbf{x}^{(l)}$ denote the input to layer $l+1$ and $\mathbf{w}_i^{(l)}$ and $b_i^{(l)}$ denote the weight and bias parameters, *i.e.*, $o_i^{(l+1)} = (\mathbf{w}_i^{(l)})^T\mathbf{x}^{(l)} + b_i^{(l)}$. Using the above defined categorization, without loss of generality, $\mathbf{x}^{(l)}$ can be factorized into three parts[1] $\mathbf{x}^{(l)} = [\mathbf{x}^{(l),1spe}; \mathbf{x}^{(l),2spe}; \mathbf{x}^{(l),s}]$ in which the three components denote the domain1-specific, domain2-specific and shared nodes, respectively. We can also denote $\mathbf{w}_i^{(l)}$ as $\mathbf{w}_i^{(l)} = [\mathbf{w}^{(l),1spe}; \mathbf{w}^{(l),2spe}; \mathbf{w}^{(l),s}]$.

For an input of the network $\mathbf{x}_{d1}^{(0)}$ in domain1, according to the categorization definition, $\mathbf{x}_{d1}^{(l),2spe} = \mathbf{0}$ because for the output of each domain2-specific node, $f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) \equiv 0$. In the forward propagation process, the output of layer $l+1$ is

$$o_i^{(l+1)} = (\mathbf{w}_i^{(l),1spe})^T\mathbf{x}_{d1}^{(l),1spe} + (\mathbf{w}_i^{(l),s})^T\mathbf{x}_{d1}^{(l),s} + b_i^{(l)}. \quad (4)$$

For an input of the network $\mathbf{x}_{d2}^{(0)}$ in domain2, similarly, we have

$$o_i^{(l+1)} = (\mathbf{w}_i^{(l),2spe})^T\mathbf{x}_{d2}^{(l),2spe} + (\mathbf{w}_i^{(l),s})^T\mathbf{x}_{d2}^{(l),s} + b_i^{(l)}. \quad (5)$$

In the back propagation process, for input of the network $\mathbf{x}_{d1}^{(0)}$ in domain1,

$$\frac{\partial L}{\partial \mathbf{w}_i^{(l),1spe}} = \frac{\partial L}{\partial o_i^{(l+1)}}\frac{\partial o_i^{(l+1)}}{\partial \mathbf{w}_i^{(l),1spe}} = \frac{\partial L}{\partial o_i^{(l+1)}}\mathbf{x}_{d1}^{(l),1spe}, \quad (6)$$

$$\frac{\partial L}{\partial \mathbf{w}_i^{(l),s}} = \frac{\partial L}{\partial o_i^{(l+1)}}\frac{\partial o_i^{(l+1)}}{\partial \mathbf{w}_i^{(l),s}} = \frac{\partial L}{\partial o_i^{(l+1)}}\mathbf{x}_{d1}^{(l),s}, \quad (7)$$

$$\frac{\partial L}{\partial \mathbf{w}_i^{(l),2spe}} = \frac{\partial L}{\partial o_i^{(l+1)}}\frac{\partial o_i^{(l+1)}}{\partial \mathbf{w}_i^{(l),2spe}} = \frac{\partial L}{\partial o_i^{(l+1)}}\mathbf{x}_{d1}^{(l),2spe} = \mathbf{0}. \quad (8)$$

From the analysis above, we have two conclusions: (1) In forward propagation, as shown in Figure 5, the weight parameters $\mathbf{w}_i^{(l),1spe}$ (blue connections) and $\mathbf{w}_i^{(l),2spe}$ (red connections) only have impact on input of corresponding domain, which is similar to the domain-specific parameters in two-stream networks. While for $\mathbf{w}_i^{(l),s}$ (green connections), it has impact on both two domains, which is similar to the shared parameters in two-stream networks. Thus, the network can implicitly control the domain-specific structure by domain-specific nodes and control the shared structure by shared nodes. (2) In backward propagation, if a node is domain2-specific, with input in domain1, its corresponding weight parameters will not be updated because the gradient is zero. That means the training samples of the other domain would not influence the implicit domain-specific structure. Note that for an input $\mathbf{x}_{d2}^{(0)}$, the same conclusion can be drawn in a similar way.

[1]";" means concatenation of vectors.

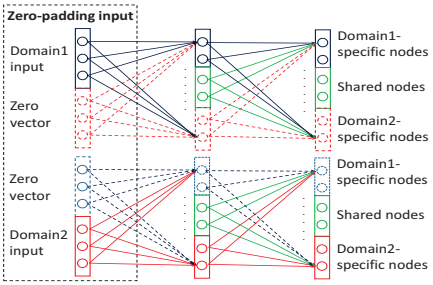

Figure 5. Explanation of deep zero-padding method. In each layer, the blue nodes denote the domain1-specific nodes, the red nodes denote the domain2-specific nodes, the green nodes denote the shared nodes and the dotted line nodes denote zero values.

Remark 1. A one-stream network may implicitly learn and evolve the domain-specific and shared structures in the network if the three types of nodes defined by Equation (3) are assumed to be existed in the network.

Remark 2. Considering two-stream structure and asymmetric FC layer structure, they are designed manually and fixed during training. Moreover, the domain-specific structure of two domains are decoupled, while the shared structure is completely identical. In contrast, if one-stream structure can implicitly learn the structure, the implicit structures corresponding to different domains are partially coupled by shared nodes and shared bias parameters (Equations (4) and (5)), which can provide more flexibility in training for cross-modality matching tasks.

4. Deep Zero-Padding

4.1. Analysis of Zero-Padding as Network Input

Since the node type we define in the last section (Equation (3)) is very optimal based on the assumption that $f_{out}(\mathbf{x}_{d1}^{(0)}, i, l) \equiv 0$ or $f_{out}(\mathbf{x}_{d2}^{(0)}, i, l) \equiv 0$, and how to make the network learn such nodes with the domain-specific property in training stage remains an important problem. In most cases, one-stream network is applied in single-domain tasks, which treats all samples equally so that generally domain-specific nodes may not be learned.

As analyzed in the previous sections, the structures of two-stream network and asymmetric FC layer network are designed manually and fixed during training, while one-stream network can evolve the network structure implicitly by learning domain-specific nodes, which may generate more optimal structure. For this purpose, we propose to use zero-padding input to stimulate the domain-specific response. As shown in Figure 5, for inputs from two domains $\mathbf{x}_{d1} \in \mathbb{R}^d$ and $\mathbf{x}_{d2} \in \mathbb{R}^d$, we apply zero-padding as follows:

$$\mathbf{x}_{d1}^{pad} = [\mathbf{x}_{d1}^T, \mathbf{O}_{1\times d}]^T, \ \mathbf{x}_{d2}^{pad} = [\mathbf{O}_{1\times d}, \mathbf{x}_{d2}^T]^T. \quad (9)$$

If we regard the network input as a prior-layer (or called the 0-th layer), then all the nodes in such a prior-layer will be definitely categorized as domain-specific nodes according to our definition in Equation (3). Now, what is the case for the nodes on the next layer? Indeed, it is hard to mathematically tell what it is, but we find that with the zero-padding as network input, the nodes in the networks are more possibly becoming domain-specific nodes. Here we continue the analysis in Section 3.2, after applying activation function $\sigma(\cdot)$ to Equations (4) and (5) we have:

$$\begin{aligned} x_{d1,i}^{(l+1)} &= \sigma((\mathbf{w}_i^{(l),1spe})^T \mathbf{x}_{d1}^{(l),1spe} + (\mathbf{w}_i^{(l),s})^T \mathbf{x}_{d1}^{(l),s} + b_i^{(l)}), \\ x_{d2,i}^{(l+1)} &= \sigma((\mathbf{w}_i^{(l),2spe})^T \mathbf{x}_{d2}^{(l),2spe} + (\mathbf{w}_i^{(l),s})^T \mathbf{x}_{d2}^{(l),s} + b_i^{(l)}). \end{aligned} \tag{10}$$

where $x_{d1,i}^{(l+1)}$ and $x_{d2,i}^{(l+1)}$ denote the output of the i-th node in layer $l+1$ with input from domain1 and domain2.

We find that for $x_{d1,i}^{(l+1)}$, there is an independent projection $(\mathbf{w}_i^{(l),1spe})^T \mathbf{x}_{d1}^{(l),1spe}$ which has no overlap with the part inside the activation function for $x_{d2,i}^{(l+1)}$. This means $\mathbf{w}_i^{(l),1spe}$ becomes a free variable independent of the terms in the activation function for $x_{d2,i}^{(l+1)}$. Thus it can provide more flexibility for neural networks to make $f_{out}(\mathbf{x}_{d1}^{(0)}, i, l+1) > 0$ and $f_{out}(\mathbf{x}_{d2}^{(0)}, i, l+1) \equiv 0$ to meet the requirement for domain1-specific nodes as compared to Equation(3). It would be more easier for neural network to spread the domain specific-nodes in deeper layers. Since the zero-padding not only appears in the input, but also can spread in the network. We call this method *deep zero-padding*.

Of course, our deep zero-padding only makes neural networks more flexible in spreading domain-specific nodes in the network but not force to. Actually, our neural networks learning empirically support this. As shown in Figure 7 and Figure 8, deep zero-padding helps the network learn domain-specific nodes more easily than that without zero-padding. The details will be illustrated later in Section 4.2.

4.2. Deep Zero-Padding for RGB-IR Re-ID

Back to our task of RGB-IR cross-modality Re-ID, for convolutional neural network, channel is corresponding to node in FC layer. For images, zero-padding is conducted in channel level. As illustrated in Figure 6, RGB image is converted to gray-scale image and placed in the first channel, and then zero-padding image is placed in the second channel. For infrared image, it is placed in the second channel and zero-padding image is placed in the first channel.

To show the effectiveness of deep zero-padding, we visualize the feature map of ResNet-6 in our experiments and compare the differences between deep zero-padding and original single-channel input. In Figure 7, we compute the average feature maps of 50 different persons on our dataset, and show all the 16 feature maps of the first and second convolution layers. As defined in Equation (3), we can categorize the domain-specific channels indicated by the bounding boxes. It is evident that, deep zero-padding helps to learn more domain-specific channels than single-channel input.

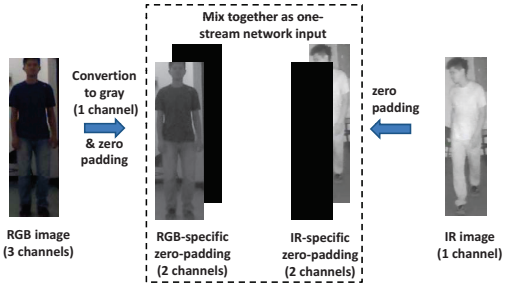

Figure 6. Deep zero-padding for RGB and infrared (IR) images.

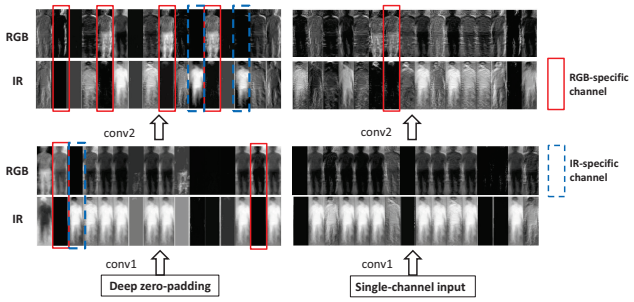

Figure 7. Feature maps of the first and second convolution layers of ResNet-6 with deep zero-padding and single channel input. In each layer, the first row shows feature maps of RGB input and the second row shows those of IR input. It is evident that domain-specific channels on the left learned by deep zero-padding are much more than those learned by single-channel input.

To quantify the domain-specific nodes in the whole network, we calculate the proportion of domain-specific nodes in each layer. Both a small (strict) threshold and a large (loose) threshold were set to determine whether the node is domain-specific according to Equation (3). The relation between the proportion of domain-specific nodes and layer depth is shown in Figure 8. It can be observed that, the domain-specific nodes mainly appear in shallower layers. It is reasonable that the network prefers shared structure after layer 6. Using deep zero-padding generally helps to generate more domain-specific nodes, while the proportions without zero-padding are low in most layers.

The analysis above and experiment results in Table 3 show that, the network can learn domain-specific nodes easier with deep zero-padding and achieve better performance.

4.3. Comparison of Cross-Modality Learning

While cross-modality matching task has not drawn much attention in Re-ID problem, it has been studied a lot in other fields like information retrieval and face verification. Cross-modality retrieval (e.g. text-image, tag-image) plays an important role in information retrieval. The models for cross-modality retrieval can be classified into real-value represen-

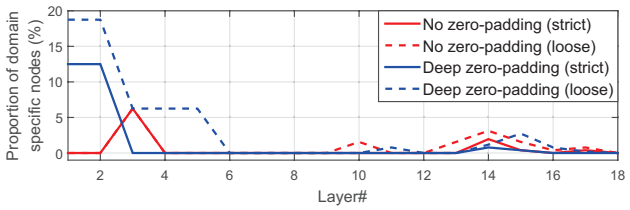

Figure 8. Relation between proportion of domain-specific nodes and layer depth. The x-axis denotes layer depth from bottom to top of the network, and the y-axis denotes the proportion of domain-specific nodes. The strict threshold is $T = 0.01\ std(x_i^{(l)})$ and the loose threshold is $T = 0.05\ std(x_i^{(l)})$ ($std(x_i^{(l)})$ is the standard deviation of the output of the i-th node in layer l). Generally, the proportion of domain-specific nodes using deep zero-padding is higher than that without zero-padding.

tation and binary representation learning [41]. The former one includes subspace learning methods [23, 28, 53] and deep learning frameworks [42, 6, 13, 11], while the latter one includes linear models [37, 34, 52, 46] and non-linear models [25, 45, 29]. Generalized similarity measure proposed by Lin *et al*. [24] is for cross-domain visual matching tasks, including RGB-RGB Re-ID task. Matching visual face versus near infrared ones (VIS-NIR) [16, 51] is rather related to RGB-IR cross-modality Re-ID. Compared with VIS-NIR face recognition, RGB-IR Re-ID is much more challenging, due to lack of important colour information. The remaining useful cues may be body shape, which differs greatly with different viewpoints and poses.

In some multi-domain learning methods, *e.g*., HFA [17], CRAFT [3], zero-padding in feature level is applied and proved to be effective. These works are closely related to our deep zero-padding. In comparison, our zero-padding is done in raw image level and the domain-specific and shared learning are done by deep neural network.

5. Experiments

We conducted extensive evaluations of existing Re-ID and cross-domain matching models as baselines on our SYSU-MM01 dataset. Then, we evaluated and analysed the effectiveness of deep models, including the proposed deep zero-padding and three commonly used network structures discussed in Section 3. See Section 2.2 for detailed evaluation protocol.

5.1. Compared Models

Baseline Models. We evaluated three favorable handcrafted features and cross-domain metric learning models as baselines. Handcrafted features included HOG [4], LOMO [21] and HIPHOP [3]. Metric learning methods included KISSME [14], LFDA [32] and cross-domain models CCA [33], CDFE [23], GMA [35], SCM [46] and CRAFT [3].

Deep Models. We evaluated four deep models shown in Figure 3, including one-steam network, two-stream network, asymmetric FC layer network and the proposed deep zero-padding method (network structure is the same as one-stream network). We applied residual block in ResNet [9] as the base convolution block for all the four structures. The number of filters for each block is 16, 16, 64, 128, 256 and 512, respectively. The next layer is an FC layer of 256 dimensions used as feature. For these four networks, the loss function was softmax loss as in ResNet [9], which is commonly used and relatively stable. All of the hyper parameters were kept the same. As for the input of the first three networks, the images were converted to single-channel gray-scale image and adjusted the size to 224×224. For our proposed method, the input was further processed by deep zero-padding as introduced in Section 3.

5.2. Model Comparisons and Analysis

Deep Models v.s. Baseline Models. We show comparative results in Table 3, including the rank-1, 10, 20 accuracies of CMC [30] and mean average precision (mAP).

From Table 3, we can see clearly that deep models outperformed baseline models with large margins. Specifically, the proposed deep zero-padding outperformed all the baselines by nearly 10% in terms of rank-1 accuracies.

All baseline models, *i.e*., handcrafted features with cross-domain metric learning methods, performed poorly: even the rank-1 accuracy of the best case failed to reach 10%. Note that LOMO feature contains rich colour information, and it performs very well in RGB-RGB Re-ID problem. Hence the results indicate that in RGB-IR matching, due to different imaging principles, the colour's discrimination degrades largely. Although body shape and clothes textures can be used for identifying persons, low-level features are not discriminative enough for the RGB-IR cross-modality person Re-ID problem. As for the deep models, however, the best rank-1 accuracy can achieve 20.58% for indoor-search. Although the problem is challenging, deep models are feasible to deal with it.

Deep Zero-Padding v.s. Other Deep Strategies. Although deep models outperformed baseline models on the whole, there were gaps among their performances to some extent.

In Table 3 we can see that the deep zero-padding outperformed two-stream network and asymmetric FC layer structure. Taking rank-1 accuracy in all-search mode under single-shot setting for instance, the gaps between deep zero-padding and two-stream/asymmetric FC layer are 3.15%/5.50%. It is possibly because two-stream and asymmetric FC layer structure require careful architecture design for this task, so their structures may not be optimal.

While with the same one-stream ResNet-6 network structure, deep zero-padding also showed superiority over the original one. The differences mainly took place in the training stage. Deep zero-padding generated two domain-specific channels in the input layer, helping the network

Table 3. Performance under all-search and indoor-search mode. r1, r10, r20 denote rank-1, 10, 20 accuracies (%).

Feature	Metric	All-search								Indoor-search							
		Single-shot				Multi-shot				Single-shot				Multi-shot			
		r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
One-stream network (deep zero-padding)	Euclidean	**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**
One-stream network	Euclidean	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 network (2 specific + 4 shared blocks)	Euclidean	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
Asymmetric FC layer network	Euclidean	9.30	43.26	60.38	10.82	13.06	52.11	69.52	6.68	14.59	57.94	78.68	20.33	20.09	69.37	85.80	13.04
Lin's	GSM	5.29	33.71	52.95	8.00	6.19	37.15	55.66	4.38	9.46	48.98	72.06	15.57	11.36	51.34	73.41	9.03
HIPHOP	CRAFT	1.80	14.56	26.29	3.40	1.92	16.00	28.31	1.77	2.86	23.40	41.94	7.16	3.01	25.53	44.97	3.43
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
	KISSME	2.12	16.21	29.13	3.53	2.79	18.23	31.25	1.96	3.11	25.47	46.47	7.43	4.10	29.32	50.59	3.61
	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
	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
	CDFE	2.09	16.68	30.51	3.75	2.47	19.11	34.11	1.86	2.80	23.39	44.46	6.91	3.28	27.31	48.61	3.24
	GMA	1.07	10.42	20.91	2.52	1.03	10.29	20.73	1.39	1.84	17.97	36.14	5.64	1.80	18.10	35.79	2.63
	SCM	1.86	15.16	28.27	3.57	2.40	17.45	31.22	1.66	3.30	25.82	46.23	7.52	3.90	28.84	51.64	3.22
	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
LOMO	Euclidean	1.75	14.14	26.63	3.48	1.96	15.06	27.30	1.85	2.24	22.53	41.53	6.64	2.24	22.79	41.80	3.31
	KISSME	2.23	18.95	32.67	4.05	2.65	20.36	34.78	2.45	3.83	31.09	52.86	8.94	4.46	34.35	58.43	4.93
	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
	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
	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
	GMA	1.04	10.45	20.81	2.54	0.99	10.50	21.06	1.47	1.79	17.90	36.01	5.63	1.71	18.11	36.17	2.88
	SCM	1.54	14.12	26.27	3.34	1.66	15.17	28.41	1.57	2.86	24.34	44.53	7.06	2.89	25.81	48.33	3.02
	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

Table 4. Comparison of deep zero-padding and similar networks. r1 and r10 denote rank-1 and 10 accuracies (%).

Feature	Metric	Single-shot			Multi-shot		
		r1	r10	mAP	r1	r10	mAP
One-stream network	Euclidean	12.04	49.68	13.67	16.26	58.14	8.59
One-stream network (domain indicator)	Euclidean	11.79	50.86	13.83	15.46	59.19	8.72
One-stream network (deep zero-padding)	Euclidean	**14.80**	**54.12**	**15.95**	**19.13**	**61.40**	**10.89**

evolve the domain-specific structure implicitly during the training stage. Since the domain-specific and shared structures/components are learned by the network automatically, the implicit structure of the one-stream network may be more suitable for RGB-IR Cross-modality Re-ID problem than the two-stream structure and asymmetric FC layer structure in our experiments, which are set manually.

We also evaluated generalized similarity measure (GSM), a closely related cross-domain two-stream deep model proposed by Lin *et al*. [24]. We used the codes released by the authors in the experiments. Lin's network is not as effective as our two-stream model. Although it can achieve good performance for RGB-RGB Re-ID as reported in [24], it is inferior for dealing with the much more challenging RGB-IR cross-modality Re-ID problem.

Deep Zero-Padding v.s. Domain Indicator. As illustrated in Section 3, one-stream network may be able to work as two-stream network with existence of domain indicator. So we pad two additional channels to the input images as domain indicators. For an RGB image, the first channel is padded with all pixels equal to 255, and second channel with 0, while for an IR image the first channel is padded with 0 and the second is padded with 255. This padding procedure explicitly provides domain indicators to the network. The performance comparison is shown in Table 4.

Table 4 shows that our proposed deep zero-padding method achieved the best performance. Input with domain indicator only achieved comparable performance as original one-stream network. This result indicates that deep zero-padding can exploit domain information more effectively.

6. Summary

This paper is the first to identify the RGB-IR cross-modality person re-identification problem and introduce a new multi-modality Re-ID dataset named SYSU-MM01. The great difference between RGB and IR images makes RGB-IR cross-modality Re-ID a very challenging problem. We discuss and evaluate three common network structures for cross-domain tasks including one-stream structure, two-stream structure and asymmetric FC layer structure. We analyse the connection between one-stream and two-stream structure and find that one-stream network can learn and evolve the structure implicitly if there exist domain-specific and shared nodes. We propose deep zero-padding to help one-stream network be more likely to automatically evolve domain-specific nodes that make the implicit network structure more suitable for the task, and this is an alternative and flexible way for cross-modality modelling as compared to the manually designed fixed structure of two-stream network. Experiments show that one-stream network trained by deep zero-padding achieved the best performance.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC (61522115, 61472456, 61573387, 61661130157, U1611461), the Royal Society Newton Advanced Fellowship (NA150459), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 4

[2] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. Yuen. An asymmetric distance model for cross-view feature mapping in person re-identification. *TCSVT*, 2015. 4

[3] Y. C. Chen, X. Zhu, W. S. Zheng, and J. H. Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017. 7

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 7

[5] S. C. Dong, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 7

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 2

[8] C.-C. Guo, S.-Z. Chen, J.-H. Lai, X.-J. Hu, and S.-C. Shi. Multi-shot person re-identification with automatic ambiguity inference and removal. In *ICPR*, 2014. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 7

[10] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, 2011. 2

[11] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang. Deep compositional cross-modal learning to rank via local-global alignment. In *ACMMM*, 2015. 7

[12] K. Jungling and M. Arens. Local feature based person reidentification in infrared image sequences. In *AVSS*, 2010. 1

[13] A. Karpathy, A. Joulin, and F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014. 7

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 7

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 3

[16] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*, 2009. 7

[17] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *TPAMI*, 2014. 7

[18] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 2

[19] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2

[20] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 1

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1, 7

[22] S. Liao and S. Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, 2015. 1

[23] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, 2006. 7

[24] L. Lin, G. Wang, W. Zuo, F. Xiangchu, and L. Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *TPAMI*, 2016. 4, 7, 8

[25] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, 2015. 7

[26] G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *ICDSC*, 2014. 1

[27] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 2014. 1

[28] X. Mao, B. Lin, D. Cai, X. He, and J. Pei. Parallel field alignment for cross media retrieval. In *ACMMM*, 2013. 7

[29] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *TPAMI*, 2014. 7

[30] H. Moon and P. J. Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception-London*, 2001. 3, 7

[31] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 4

[32] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 7

[33] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACMMM*, 2010. 7

[34] M. Rastegari, J. Choi, S. Fakhraei, H. Daumé III, and L. S. Davis. Predictable dual-view hashing. In *ICML*, 2013. 7

[35] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. 7

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[37] L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. In *ICML*, 2008. 7

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 3

[39] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 4

[40] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 4

[41] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016. 7

[42] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 2016. 7

[43] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1, 3

[44] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 1

[45] D. Zhai, H. Chang, Y. Zhen, X. Liu, X. Chen, and W. Gao. Parametric local multimodal hashing for cross-view similarity search. In *IJCAI*, 2013. 7

[46] D. Zhang and W.-J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, 2014. 7

[47] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2

[48] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2

[49] W. S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009. 2

[50] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1

[51] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li. Matching nir face to vis face using transduction. *TIFS*, 2014. 7

[52] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao. Linear cross-modal hashing for efficient multimedia search. In *ACMMM*, 2013. 7

[53] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013. 7