International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01290-1

RGB-IR Person Re-identification by Cross-Modality Similarity Preservation

Ancong Wu[1] · **Wei-Shi Zheng**[1,2,3] · **Shaogang Gong**[5] · **Jianhuang Lai**[1,4]

Received: 3 June 2019 / Accepted: 30 December 2019

Abstract

Person re-identification (Re-ID) is an important problem in video surveillance for matching pedestrian images across non-overlapping camera views. Currently, most works focus on RGB-based Re-ID. However, RGB images are not well suited to a dark environment; consequently, infrared (IR) imaging becomes necessary for indoor scenes with low lighting and 24-h outdoor scene surveillance systems. In such scenarios, matching needs to be performed between RGB images and IR images, which exhibit different visual characteristics; this cross-modality matching problem is more challenging than RGB-based Re-ID due to the lack of visible colour information in IR images. To address this challenge, we study the RGB-IR cross-modality Re-ID (RGB-IR Re-ID) problem. Rather than applying existing cross-modality matching models that operate under the assumption of identical data distributions between training and testing sets to handle the discrepancy between RGB and IR modalities for Re-ID, we cast learning shared knowledge for cross-modality matching as the problem of cross-modality similarity preservation. We exploit same-modality similarity as the constraint to guide the learning of cross-modality similarity along with the alleviation of modality-specific information, and finally propose a Focal Modality-Aware Similarity-Preserving Loss. To further assist the feature extractor in extracting shared knowledge, we design a modality-gated node as a universal representation of both modality-specific and shared structures for constructing a structure-learnable feature extractor called Modality-Gated Extractor. For validation, we construct a new multi-modality Re-ID dataset, called SYSU-MM01, to enable wider study of this problem. Extensive experiments on this SYSU-MM01 dataset show the effectiveness of our method. Download link of dataset: https://github.com/wuancong/SYSU-MM01.

Keywords Person re-identification · Cross-modality model · RGB-infrared matching

1 Introduction

Person re-identification (Re-ID) is an important problem in video surveillance for which the available solutions have undergone fast-growing development in recent years, from feature design (Gray and Tao 2008; Farenzena et al. 2010; Liu et al. 2012; Liao et al. 2015; Matsukawa et al. 2016; Zheng et al. 2015; Xiong et al. 2014) to distance metric learning (Xiong et al. 2014; Paisitkriangkrai et al. 2015; Gray and Tao 2008; Chen et al. 2015, 2018; Liao and Li 2015; Zheng et al. 2013; Köstinger et al. 2012; Pedagadi et al. 2013; Li et al. 2013; Liao et al. 2015) and end-to-end deep learning (Li et al. 2014; Ahmed et al. 2015; Wu et al. 2016; Xiao et al. 2016; Zhao et al. 2017; Zheng et al. 2017; Song et al. 2018; Chen et al. 2018; Sun et al. 2018). Currently, in most cases of Re-ID, it is assumed that the appearance (e.g., clothing) of a person remains unchanged in the short term; consequently,

Communicated by Bernt Schiele.

✉ Wei-Shi Zheng
wszheng@ieee.org

Ancong Wu
wuancong@gmail.com

Shaogang Gong
s.gong@qmul.ac.uk

Jianhuang Lai
stsljh@mail.sysu.edu.cn

[1] Sun Yat-sen University, Guangzhou, China

[2] Pengcheng Laboratory, Shenzhen, China

[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China

[4] Guangdong Province Key Laboratory of Information Security, Guangzhou, China

[5] Queen Mary University of London, London, UK

Published online: 03 February 2020

the majority of studies focus on cross-view matching between RGB images, i.e., RGB-RGB cross-view matching of visual imagery.

However, the capabilities of such single-modality RGB-RGB Re-ID are limited when pedestrians appear in poor lighting or dark conditions, e.g., at night. RGB images become uninformative (not merely noisy) at night (see Fig. 1). In such a case, relying on visible light results in unreliable and less meaningful cross-view matching. Many modern surveillance cameras can automatically switch between RGB and infrared (IR) modes at any time when the lighting conditions change significantly. Therefore, it is necessary to solve the interesting problem of enabling an IR image of a person captured in a dark camera view to be matched with an RGB image from a disjoint bright camera view. We call this problem *RGB-IR Re-ID*.

This work addresses the *RGB-IR Re-ID* problem. RGB-IR Re-ID has rarely been studied and remains a challenging problem due to the significant visual differences between the two modalities. There are two factors contributing to the difficulty of the problem. First, there are intrinsic differences between RGB and IR images caused by the different wavelength ranges used in the imaging process. As shown in Fig. 1, RGB images (the first row) have three channels, containing colour information obtained from visible light, while IR images (the second row) have only one channel, containing information obtained from invisible light. Consequently, it is highly improbable to find image patches with the same colour in RGB and IR images, meaning that colour information, which is the most important appearance cue (Liu et al. 2012; Liao et al. 2015) used to identify people in existing Re-ID methods, has become uninformative. Second, variations of viewpoint and person poses, which already make single-modality RGB-based Re-ID a challenging problem, can cause even greater difficulties in RGB-IR Re-ID because of the severe imagery misalignment across images in the two modalities captured from the same subject.

RGB-IR Re-ID is a cross-modality matching problem. Handling the intra-class imaging discrepancies caused by cross-modality transformation is the key challenge. While RGB images and IR images are visually different, they actually share some information (such as shape) in images of the same object. Therefore, it is possible and critical to extract the *shared knowledge* in two modalities for cross-modality matching. A common technique for cross-modality matching is to minimise the gap in some feature space between different modalities by identity classification and feature distribution alignment (e.g., He et al. 2017, 2019; Ye et al. 2018; Dai et al. 2018). However, these techniques assume that the data distributions are identical for training and testing, whereas this assumption is invalid for Re-ID since there is a discrepancy between these two data distributions due to non-overlapping person identities in the training and testing sets. To visualise the effect of the distribution discrepancy between training and testing, we evaluate a ResNet-50 model (Sun et al. 2018) trained on an RGB-IR person Re-ID dataset called SYSU-MM01 that is introduced later. We show the distributions of the training and testing sets in the feature space after dimensionality reduction by t-SNE (Maaten and Hinton 2008) in Fig. 2. The distribution discrepancy between training and testing sets is significant in the feature space.

In this work, we do not operate under the assumption of identical data distributions between training and testing data, and mine the shared knowledge for cross-modality matching in the similarity space, because similarity

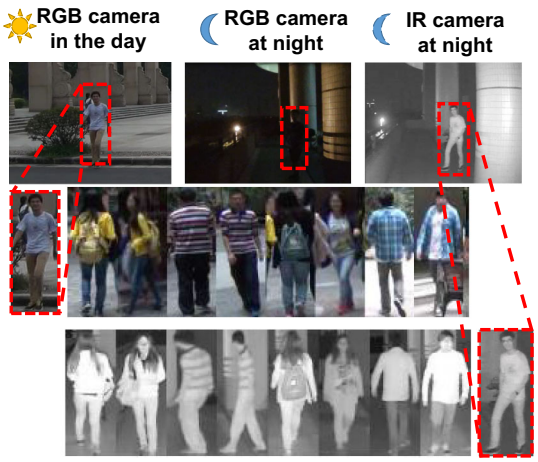


Fig. 1 Examples of RGB images and infrared (IR) images captured in two outdoor scenes during the day and at night, respectively. The images in every two columns are of the same person. Captured by devices that detect light of different wavelengths, RGB images and IR images of the same person look very different (best viewed in colour)

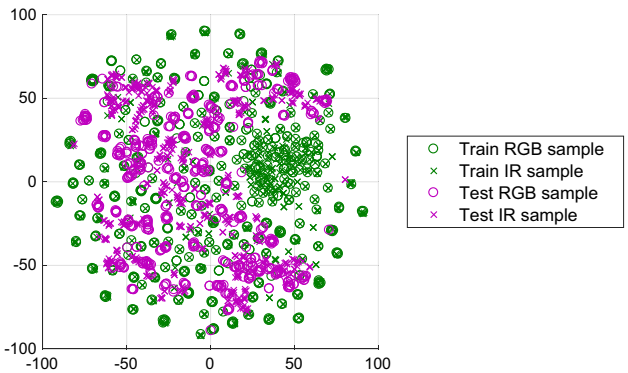


Fig. 2 Visualisation of features in the training set and testing set by dimensionality reduction using t-SNE (Maaten and Hinton 2008). The features are extracted by a ResNet-50 model (Sun et al. 2018) trained on an RGB-IR Re-ID dataset SYSU-MM01 introduced later. The figure shows that the feature distributions are different for the training set and testing set because of non-overlapping identities during training and testing

value is relative information between samples and does not require the assumption of identical training and testing data distributions. In particular, we expect that the shared knowledge for cross-modality matching should be valid for same-modality matching, so that the modality-specific information in the shared feature space of the two modalities can be alleviated. To this end, we cast learning effective shared knowledge across RGB and infrared modalities as a cross-modality similarity preservation problem. We guide learning cross-modality matching by same-modality matching as a constraint for regularisation in terms of similarity preservation and propose a *Focal Modality-Aware Similarity-Preserving Loss*.

To further assist the extraction of shared knowledge for matching, we propose a structure-learnable network called *Modality-Gated Extractor* by using *Modality-Gated Node* as a universal representation of modality-specific and shared structures. When constructing feature extractor for cross-modality matching, it is important to learn appropriate shared and modality-specific structures. In contrast, in existing neural networks for cross-modality matching (e.g., Ye et al. 2018; He et al. 2017), fixed shared and modality-specific model structures are used. In these structures, the parameters indicating whether and to what extent a node should be shared are manually designed. Since these parameters cannot be learned based on training data, the designed structures are suboptimal and cannot be dynamically adjusted to suit the data to better extract shared knowledge across modalities. By introducing modality-gated nodes in our model, we enable our feature extraction network to implicitly learn both modality-specific and shared network structures from training data without manual intervention.

To study the RGB-IR Re-ID problem, due to the lack of public RGB-IR benchmark datasets, we construct a new SYSU Multiple-Modality Re-ID dataset (SYSU-MM01). Compared to existing commonly used single-modality Re-ID datasets, as summarised in Table 1, this new RGB-IR cross-modality Re-ID dataset provides, for the first time, a meaningful public benchmark for the study of RGB-IR Re-ID. It contains 30,071 RGB images and 15,792 IR images corresponding to 491 person IDs captured from 6 camera views. Compared with another multi-modality pedestrian dataset RegDB (Nguyen et al. 2017) in Table 1, our SYSU-MM01 dataset contains more samples from more identities captured in more cameras. Moreover, for capturing IR images, RegDB (Nguyen et al. 2017) used thermal cameras, whereas our SYSU-MM01 used near infrared (NIR) cameras, and thus SYSU-MM01 is more practical for surveillance systems. Extensive experiments on this SYSU-MM01 dataset show the effectiveness of our proposed framework for RGB-IR Re-ID when compared to contemporary methods for Re-ID, cross-modality matching and domain adaptation.

In summary, the contributions of this work are as follows: (1) This work presents an early comprehensive attempt to address the challenging RGB-IR cross-modality Re-ID problem for matching images of persons captured under normal lighting with those captured in dark environments (e.g., for 24-h surveillance). (2) We cast mining shared knowledge for cross-modality matching as the problem of cross-modality similarity preservation, and we propose a Focal Modality-Aware Similarity-Preserving Loss, which does not operate under the assumption of identical distributions between the training and testing data. (3) We propose a modality-gated node for constructing a structure-learnable deep neural network to assist in learning more effective shared and modality-specific structures in a data-driven manner for RGB-IR Re-ID. (4) We construct, for the first time, a public benchmark dataset called SYSU-MM01 for studying RGB-IR Re-ID. Extensive experiments were conducted to

Table 1 Comparison between SYSU-MM01 and existing Re-ID datasets

Dataset	#IDs	#Images	#Cams	RGB	IR
ViPER (Gray et al. 2007)	632	1264	2	Yes	No
iLIDS (Zheng et al. 2009)	119	476	2	Yes	No
PRID2011 (Hirzer et al. 2011)	200	971	2	Yes	No
CUHK01 (Li et al. 2012)	972	1942	2	Yes	No
SYSU (Guo et al. 2014)	502	24,448	2	Yes	No
CUHK03 (Li et al. 2014)	1467	13,164	6	Yes	No
Market (Zheng et al. 2015)	1501	32,668	6	Yes	No
MARS (Zheng et al. 2016)	1261	1,191,003	6	Yes	No
DukeMTMC (Ristani et al. 2016)	1404	36,411	8	Yes	No
RegDB (Nguyen et al. 2017)	412	4120/4120	2	Yes	Yes (thermal)
SYSU-MM01	491	30,071/15,792	6	Yes	Yes (NIR)

(#/# denotes the numbers of RGB/IR images)

evaluate the proposed model against a wide range of computer vision models for cross-modality matching.

2 Related Work

2.1 Single-Modality Re-ID

Most existing works rely solely on RGB visual appearance features. Among them, colour is most frequently used and is often encoded in histograms (Gray and Tao 2008; Farenzena et al. 2010; Liu et al. 2012; Liao et al. 2015), as in SDALF (Farenzena et al. 2010) and LOMO (Liao et al. 2015). Texture-based features are also employed, such as HOG features (Zheng et al. 2015) and LBP features (Xiong et al. 2014). Some other types of hand-crafted features have also been developed, such as the covariance-based GOG descriptor (Matsukawa et al. 2016) and custom pictorial structures (Dong et al. 2011). Recently, more advanced feature learning methods, such as saliency learning (Zhao et al. 2017), mirror representation (Chen et al. 2015, 2018), pose prior feature learning (Wu et al. 2015), invariant colour feature learning (Yang et al. 2014; Kviatkovsky et al. 2013), dictionary learning (Jing et al. 2015; Karanam et al. 2015), attribute learning (Shi et al. 2015; Su et al. 2016) and binary representation learning (Chen et al. 2017; Zhu et al. 2017), have been studied.

In addition to feature representations, a large number of metric/subspace learning models (Gray and Tao 2008; Prosser et al. 2010; Köstinger et al. 2012; Zheng et al. 2013, 2015, 2016; Pedagadi et al. 2013; Li et al. 2013; Xiong et al. 2014; Paisitkriangkrai et al. 2015; Liao et al. 2015; Chen et al. 2017; Liao and Li 2015; You et al. 2016; Zhang et al. 2016; Li et al. 2015; Wang et al. 2016; Bak and Carr 2017) have been developed to achieve more reliable matching; such models include RDC (Zheng et al. 2013), KISSME (Köstinger et al. 2012), LADF (Li et al. 2013), LFDA (Pedagadi et al. 2013), MLAPG (Liao and Li 2015) and DNS (Zhang et al. 2016). In particular, deep learning methods for Re-ID (Li et al. 2014; Ahmed et al. 2015; Xiao et al. 2016; Zhao et al. 2017; Zheng et al. 2017; Song et al. 2018; Chen et al. 2018; Sun et al. 2018; Yang et al. 2019; Yin et al. 2020) have received substantial attention in recent years due to their more powerful deep features, which enable superior performance compared to hand-crafted features, especially when large training datasets are available. Various other problems have also been studied, such as unsupervised learning (Kodirov et al. 2016; Yu et al. 2017, 2018; Wei et al. 2018; Li et al. 2018), re-ranking (Lisanti et al. 2015; Zhong et al. 2017) and person search (Xiao et al. 2017).

Given the fast development of recent imaging devices, Re-ID studies have also been extended beyond methods relying on RGB visual images. For instance, depth-based Re-ID methods (Wu et al. 2017; Haque et al. 2016) have been exploited for the case of clothing changes. However, depth-sensing devices have not yet been widely deployed in practical applications. Similarly, very few Re-ID methods use IR images, with the exception of the method of Jungling and Arens (2010), who studied IR-IR video matching for Re-ID at night but did not consider RGB-IR matching.

In summary, the overwhelming majority of techniques for single-modality person Re-ID in the literature are not suitable for cross-modality matching.

2.2 Cross-Modality Matching Models

Very few works have studied cross-modality Re-ID. Recently, TONE + HCML (Ye et al. 2018) and BDTR (Ye et al. 2018), which are based on two-stream networks, were developed for RGB-thermal Re-ID. Although RGB-thermal matching can also be used for person Re-ID across day and night conditions, thermal imaging devices are far more expensive than IR cameras, especially for large-scale surveillance systems; therefore, such methods are impractical for wide use. cmGAN (Dai et al. 2018) was developed for cross-modality Re-ID based on a one-stream network with adversarial training. The task of matching visual (VIS) face images with near-infrared (NIR) ones (Lei and Li 2009; Zhu et al. 2014; He et al. 2017, 2019) is related to the task of cross-modality RGB-IR Re-ID. However, compared with VIS-NIR face recognition, RGB-IR Re-ID is a much more challenging problem because the visual appearance variations between RGB and IR images of pedestrians are much more significant than those between VIS and NIR images of faces due to pose variations and the lack of colour information.

Unlike our method, these cross-modality methods assume that the distributions of the training and testing data are identical, which is not valid for Re-ID because of non-overlapping person identities between training and testing; in comparison, we do not operate under this assumption and learn our cross-modality model using data similarity information. Moreover, these related methods do not have the capabilities of structure learning in a deep CNN framework.

In addition, for solving the RGB-IR person Re-ID problem, the general cross-modality models for information retrieval and face verification are related. Cross-modality retrieval models can be categorised into real-valued representation learning and binary representation learning. Methods for real-valued representation learning include CCA (Rasiwasia et al. 2010), CDFE (Lin and Tang 2006), GMA (Sharma et al. 2012), MMD (Zhu et al. 2014), DeepCCA (Andrew et al. 2013), Corr-AE (Fang et al. 2014), deep-SM (Wei et al. 2017), and MDNN (Wang et al. 2016), whilst methods for binary representation learning include SCM (Zhang and Li 2014), QCH (Wu et al. 2015), and SePH (Lin et al. 2015). More recently, deep domain adaptation methods

have been developed, e.g., an MMD-based domain adaptation net (Long et al. 2015), DeepCORAL (Sun and Saenko 2016), and ADDA (Tzeng et al. 2017). These methods aim to minimise the distance between the feature distributions in two modalities. In contrast to these models, which align two modalities in the feature space, a characteristic of our model is to use same-modality matching as a constraint to guide the learning of cross-modality matching in the similarity space, which does not require the assumption of identical data distributions in training and testing, as held by the above discussed methods.

In general, for cross-modality image matching, several deep frameworks have been developed, e.g., a generalised similarity net (Lin et al. 2017), Castrejon's net (Castrejon et al. 2016) and a multi-view deep network (MvDN) (Kan et al. 2016). In these methods, the modality-specific and shared structures are manually designed and remain fixed during training; hence, these structures are likely to be suboptimal and do not provide flexibility for identifying information shared across modalities for the challenging RGB-IR Re-ID task. In comparison, we propose a structure-learnable framework, in which the modality-specific and shared structures are more flexibly determined in a data-driven way.

2.3 Cross-Modality Convolutional Network Structures

To lay the foundation for introducing our proposed modality-gated node in Sect. 4, we revisit convolutional neural network (CNN) structures for cross-modality matching. Generally, these structures can be categorised into three types, as shown in Fig. 3.

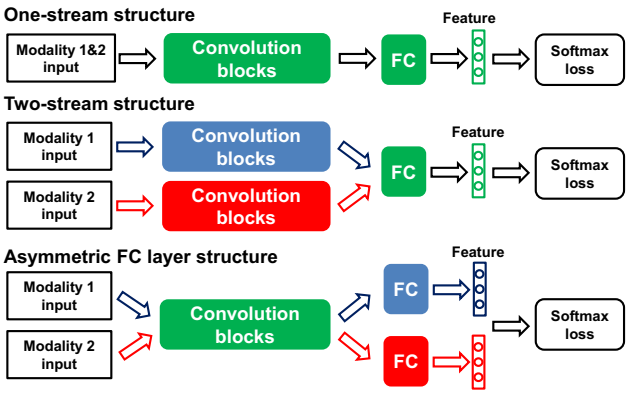


Fig. 3 Three commonly used CNN structures for cross-modality image matching. The colours of the convolution blocks and fully connected (FC) layers indicate whether the parameters are shared. Red and blue indicate modality-specific parameters, and green indicates shared parameters (best viewed in colour) (Color figure online)

One-Stream Structure As a preliminary approach, the commonly used one-stream network structure can be applied for cross-modality image matching. As shown in the first network in Fig. 3, there is a single input stream, and all parameters are shared. This structure is usually applied for single-modality data. Representative networks include AlexNet (Krizhevsky et al. 2012), VGG (Simonyan and Zisserman 2015), GoogleNet (Szegedy et al. 2015), and ResNet (He et al. 2016). For Re-ID, most networks for matching RGB pedestrian images have a one-stream structure, such as JSTL-DGD (Xiao et al. 2016), PCB (Sun et al. 2018), the part-aligned representation (Zhao et al. 2017) and Ahmed's Siamese network (Ahmed et al. 2015). cmGAN (Dai et al. 2018) for cross-modality Re-ID also uses a one-stream structure.

Two-Stream Structure As shown in the second network in Fig. 3, in the two-stream structure, there are two input streams, corresponding to data from two different modalities. In the shallower layers, the parameters of the network are specific to a particular modality, while in the deeper layers, shared parameters are used. A two-stream network uses modality-specific structures in shallow layers to alleviate the modality gap at a low level. Representative networks include Lin's generalised similarity net (Lin et al. 2017) for cross-modality image matching, Castrejon's net (Castrejon et al. 2016) for cross-modality retrieval and MvDN (Kan et al. 2016) for cross-view classification. As for RGB-thermal Re-ID, both TONE + HCML (Ye et al. 2018) and BDTR (Ye et al. 2018) use the two-stream network structure.

Asymmetric FC Layer Structure. As shown in the third network in Fig. 3, nearly all parameters are shared except in the last fully connected (FC) layer, with the purpose of alleviating modality gap at feature level. Representative methods include CVDCA (Chen et al. 2017), CAMEL (Yu et al. 2017) for Re-ID and IDR (He et al. 2017), WCNN (He et al. 2019) for VIS-NIR face recognition.

Ultimately, networks for cross-modality image matching are constructed by modality-specific and shared structures.

A preliminary result from our research was reported in Wu et al. (2017). In this work, we significantly extend our research. We do not operate under the assumption of identical distributions between training and testing data, and we cast learning shared knowledge for cross-modality matching as a cross-modality similarity preservation problem. In addition, we extend our deep zero padding model by developing a structure-learnable deep CNN framework based on modality-gated nodes. Our early model based on the deep zero padding method (Wu et al. 2017) is a special case of our proposed framework in which fixed modality-gated nodes are used in the input layer of a one-stream network.

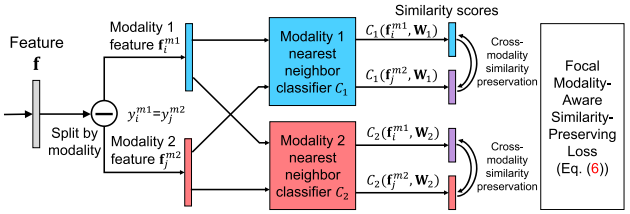


Fig. 4 Illustration of cross-modality similarity preservation. First, cross-modality positive sample pairs $(\mathbf{f}_i^{m1}, \mathbf{f}_j^{m2})$ (with identity labels $y_i^{m1} = y_j^{m2}$) are sampled, where $\ominus$ denotes the splitting of sample features by modality. Then, the similarity scores for same-modality and cross-modality matching are obtained using two modality-specific nearest neighbour classifiers, C_1 and C_2, for modality 1 and modality 2, respectively. Finally, the Focal Modality-Aware Similarity-Preserving Loss is applied to minimise the difference between same-modality similarities and cross-modality similarities

3 Learning Cross-Modality Similarity Preservation

For cross-modality matching, we aim to extract shared knowledge to bridge the two different modalities. For this purpose, feature distribution alignment (e.g., He et al. 2017, 2019; Ye et al. 2018; Dai et al. 2018) is commonly used with the assumption of identical data distributions between training and testing. However, for Re-ID, the non-overlapping identities in training and testing lead to data distribution discrepancies in training and testing, which violates the assumption of feature distribution alignment, as shown in Fig. 2. To address this challenge, we do not operate under this assumption, and cast mining such shared knowledge for cross-modality matching as the problem of cross-modality similarity preservation. The overview of cross-modality similarity preservation is shown in Fig. 4.

3.1 Modality-Aware Similarity-Preserving Loss

To eliminate the discrepancy between two modalities in the similarity space, we expect that when performing cross-modality matching and same-modality matching, the retrieval results can be consistent in some feature space. Therefore, we force the cross-modality similarity and same-modality similarity between two objects to be as equivalent as possible, i.e., *cross-modality similarity preservation*, so modality-specific information can be alleviated in the shared feature space due to the consistent constraint between the retrieval results of cross-modality matching and those of same-modality matching. For this purpose, we propose a Modality-Aware Similarity-Preserving Loss as follows.

To detail the proposed loss function, for an object J_k, we first assume that a pair of synchronised RGB image $\mathbf{I}^{m1}(J_k)$ and IR image $\mathbf{I}^{m2}(J_k)$ is given. For any two images $\mathbf{I}_1$ and $\mathbf{I}_2$, we aim to learn a function $f_{sim}(\mathbf{I}_1, \mathbf{I}_2)$ to compute the similarity between them. For two objects J_k and J_l, we expect that in some feature space the cross-modality similarity $f_{sim}(\mathbf{I}^{m1}(J_k), \mathbf{I}^{m2}(J_l))$ and $f_{sim}(\mathbf{I}^{m2}(J_k), \mathbf{I}^{m1}(J_l))$ can be preserved as the same-modality similarity $f_{sim}(\mathbf{I}^{m1}(J_k), \mathbf{I}^{m1}(J_l))$ and $f_{sim}(\mathbf{I}^{m2}(J_k), \mathbf{I}^{m2}(J_l))$; that is, cross-modality matching is constrained and guided by same-modality matching. Optimally speaking, the cross-modality similarity preservation is formulated as

$$\sum_{J_k, J_l \in \mathcal{J}} (f_{sim}(\mathbf{I}^{m1}(J_k), \mathbf{I}^{m1}(J_l)) - f_{sim}(\mathbf{I}^{m1}(J_k), \mathbf{I}^{m2}(J_l)))^2 + (f_{sim}(\mathbf{I}^{m2}(J_k), \mathbf{I}^{m2}(J_l)) - f_{sim}(\mathbf{I}^{m2}(J_k), \mathbf{I}^{m1}(J_l)))^2, \tag{1}$$

where $\mathcal{J}$ is a set of objects.

However, in practice, for RGB-IR Re-ID, it is difficult to have the synchronised RGB and IR image, as they are not simultaneously captured by the same camera. To overcome this problem, we introduce a relaxed version of the cross-modality similarity preservation by using RGB and IR image pairs of the same identity. More specifically, let $\{\mathbf{I}_i^{m1}, y_i^{m1}\}_{i=1}^{n_1}$ and $\{\mathbf{I}_j^{m2}, y_j^{m2}\}_{j=1}^{n_2}$ denote the training samples from modality 1 and modality 2, respectively, where $\mathbf{I}_i^{m1}$ and $\mathbf{I}_j^{m2}$ are images and y_i^{m1} and y_j^{m2} are identity labels. Let f_{ex} denote the model for feature extraction. The features are $\mathbf{f}_i^{m1} = f_{ex}(\mathbf{I}_i^{m1}; \mathbf{2})$ and $\mathbf{f}_j^{m2} = f_{ex}(\mathbf{I}_j^{m2}; \mathbf{2})$. In our case, the features $\mathbf{f}_i^{m1}$ and $\mathbf{f}_j^{m2}$ are normalised by the ℓ_2-norm; thus, the inner product of the two features is the cosine similarity.

To model same-modality and cross-modality matching, which are equivalent to similarity-based nearest neighbour classification, we then introduce two modality-specific nearest neighbour classifiers C_1 and C_2 for modality 1 and modality 2, respectively, as follows:

$$C_1(\mathbf{f}, \mathbf{W}_1) = \mathbf{W}_1^\top \mathbf{f}, \quad C_2(\mathbf{f}, \mathbf{W}_2) = \mathbf{W}_2^\top \mathbf{f}, \tag{2}$$

where $\mathbf{f}$ is the feature to be classified. For modality 1, $\mathbf{W}_1 = [\mathbf{f}_{id,1}^{m1}, \mathbf{f}_{id,2}^{m1}, \ldots, \mathbf{f}_{id,K}^{m1}]$, where each column is a feature vector $\mathbf{f}_{id,k}^{m1}$ of a sample of ID k in modality 1. K is the number of identities considered by classifier C_1. $C_1(\mathbf{f}, \mathbf{W}_1)$ is a similarity score vector computed by the inner product operation and consists of the similarity scores between the feature $\mathbf{f}$ and the feature corresponding to each identity in modality 1. For modality 2, $C_2(\mathbf{f}, \mathbf{W}_2)$ is defined similarly with $\mathbf{W}_2 = [\mathbf{f}_{id,1}^{m2}, \mathbf{f}_{id,2}^{m2}, \ldots, \mathbf{f}_{id,K}^{m2}]$. During training, in each iteration for a mini-batch, $\mathbf{W}_1$ and $\mathbf{W}_2$ are constructed by feature vectors extracted from the samples in the current mini-batch. The gradients of loss functions with respect to $\mathbf{W}_1$ and $\mathbf{W}_2$ are propagated to parameters $\mathbf{2}$ of the feature extractor through the feature vectors $\mathbf{f}_{id,k}^{m1}$ and $\mathbf{f}_{id,k}^{m2}$.

Given features $\mathbf{f}_i^{m1}$ of modality 1 and $\mathbf{f}_j^{m2}$ of modality 2 for the same identity (i.e., $y_i^{m1} = y_j^{m2}$) as probes, $\mathbf{W}_1$ is regarded as the gallery set of K identities in modality 1. For

same-modality matching and cross-modality matching, the similarity score vectors $C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1)$ and $C_1(\mathbf{f}_j^{m2}, \mathbf{W}_1)$ can be computed using the nearest neighbour classifier C_1. We guide the learning of cross-modality similarity given the constraint of preserving same-modality similarity by forcing the similarity score vector $C_1(\mathbf{f}_j^{m2}, \mathbf{W}_1)$ to be as close as possible to $C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1)$. The objective is similar for classifier C_2. For cross-modality similarity preservation, we minimise the following objective function:

$$L_{MSP} = \sum_{(i,j)\in\mathcal{P}} ||C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1) - C_1(\mathbf{f}_j^{m2}, \mathbf{W}_1)||^2 + ||C_2(\mathbf{f}_j^{m2}, \mathbf{W}_2) - C_2(\mathbf{f}_i^{m1}, \mathbf{W}_2)||^2, \tag{3}$$

where $\mathcal{P} = \{(i, j)|y_i^{m1} = y_j^{m2}, i \in \{1, \ldots, n_1\}, j \in \{1, \ldots, n_2\}\}$ is the index pair set of all cross-modality positive sample pairs. Negative sample pairs are not used because they are from different identities, so the similarity scores output by the classifiers should be different. We call L_{MSP} the *Modality-Aware Similarity-Preserving Loss*.

Focal Modality-Aware Similarity Preserving Loss In the above Modality-Aware Similarity-Preserving Loss (Eq. 3), the terms corresponding to all positive sample pairs are equally weighted for learning. However, during training, when the classification results from classifiers C_1 and C_2 are not correct, preserving the corresponding similarities may not always provide valuable information. Non-significant information may be learned if the similarity preservation is enforced. Thus, in the learning process, we should focus on the reliable positive sample pairs that are correctly classified and neglect unreliable sample pairs that are not correctly classified. To address this problem, we dynamically adjust the weight of each cross-modality positive sample pair $(\mathbf{f}_i^{m1}, \mathbf{f}_j^{m2})$ by introducing two confidence factors, $p_{i,j}^{C_1}$ and $p_{i,j}^{C_2}$, for classifiers C_1 and C_2, respectively, which are defined as follows:

$$\begin{aligned} p_{i,j}^{C_1} &= f_{sm}(C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1), y_i^{m1}) \cdot f_{sm}(C_1(\mathbf{f}_j^{m2}, \mathbf{W}_1), y_j^{m2}), \\ p_{i,j}^{C_2} &= f_{sm}(C_2(\mathbf{f}_j^{m2}, \mathbf{W}_2), y_j^{m2}) \cdot f_{sm}(C_2(\mathbf{f}_i^{m1}, \mathbf{W}_2), y_i^{m1}), \end{aligned} \tag{4}$$

where f_{sm} is a softmax function defined as

$$f_{sm}(\mathbf{s}, y) = \frac{\exp(s_y)}{\sum_{k=1}^{K} \exp(s_k)}, \tag{5}$$

in which $\mathbf{s}$ is the similarity score vector, s_k is the k-th element of the similarity score vector $\mathbf{s}$ and y denotes an identity label. For example, $f_{sm}(C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1), y_i^{m1})$ is the probability of correctly classifying the feature $\mathbf{f}_i^{m1}$ as the identity y_i^{m1} with classifier C_1. If $\mathbf{f}_i^{m1}$ is correctly classified with high confidence, then $f_{sm}(C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1), y_i^{m1})$ is close to 1. Similar interpretations hold for the other terms involving f_{sm}. If both samples $\mathbf{f}_i^{m1}$ and $\mathbf{f}_j^{m2}$ in a positive sample pair can be correctly classified by C_1 with high confidence, then the confidence factor $p_{i,j}^{C_1}$ is close to 1. Thus, the confidence factors $p_{i,j}^{C_1}$ and $p_{i,j}^{C_2}$ have values ranging between 0 and 1 and indicate the reliability of the sample pair $(\mathbf{f}_i^{m1}, \mathbf{f}_j^{m2})$ with respect to C_1 and C_2, respectively.

By introducing the confidence factors $p_{i,j}^{C_1}$ and $p_{i,j}^{C_2}$ into the expression for the Modality-Aware Similarity-Preserving Loss given in Eq. (3), we obtain

$$L_{FMSP} = \sum_{(i,j)\in\mathcal{P}} p_{i,j}^{C_1} ||C_1(\mathbf{f}_i^{m1}, \mathbf{W}_1) - C_1(\mathbf{f}_j^{m2}, \mathbf{W}_1)||^2 + p_{i,j}^{C_2} ||C_2(\mathbf{f}_j^{m2}, \mathbf{W}_2) - C_2(\mathbf{f}_i^{m1}, \mathbf{W}_2)||^2. \tag{6}$$

Hence, the effect of each sample pair $(\mathbf{f}_i^{m1}, \mathbf{f}_j^{m2})$ on cross-modality similarity preservation is dynamically adjusted by the confidence factors. We call L_{FMSP} the *Focal Modality-Aware Similarity-Preserving Loss*. In Fig. 4, we show the process of cross-modality similarity preservation.

3.2 A Cross-Modality Matching Framework

We apply the Focal Modality-Aware Similarity-Preserving Loss for training a feature extractor in a cross-modality matching framework shown in Fig. 5.

In the training stage, the training images $\{\mathbf{I}_i\}$ and the corresponding identity labels $\{y_i\}$ and modality labels $\{y_i^{mod}\}$ are required. We minimise the following loss function:

$$L = L_{cls} + \lambda L_{FMSP}, \tag{7}$$

where L_{cls} is the softmax cross-entropy loss that is commonly used for classification, L_{FMSP} is the Focal Modality-Aware Similarity-Preserving Loss, and λ is a trade-off parameter. L_{cls} is designed for learning discriminative features for classification without considering the discrepancy between two modalities. L_{FMSP} can alleviate the discrepancy between two modalities for cross-modality matching by cross-modality similarity preservation. They are complementary to each other.

Visualisation A result of cross-modality similarity preservation is that the ranking lists of identities are consistent for cross-modality matching and same-modality matching. To visualise the effect of the Focal Modality-Aware Similarity-Preserving Loss, several examples of same-modality matching and cross-modality matching using different losses are shown in Fig. 6, where (a) "Ours" is our Focal Modality-Aware Similarity-Preserving Loss, (b) "Softmax" is the Softmax loss and (c) "Softmax + MMD" is a combination

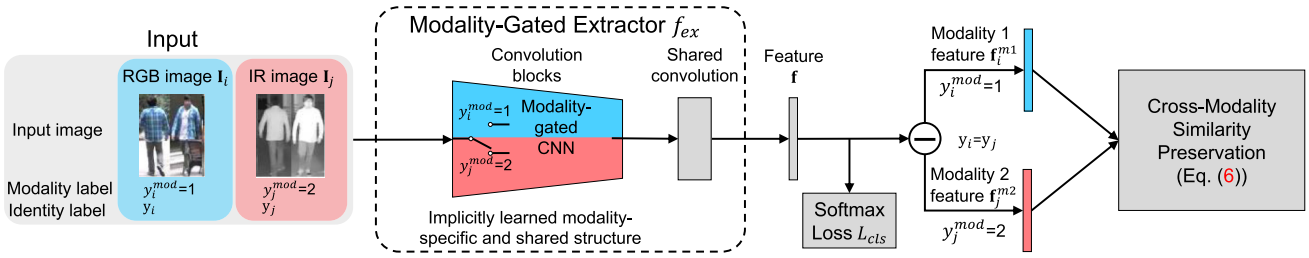


Fig. 5 Overview of the cross-modality matching framework. The framework consists of two parts: cross-modality similarity preservation and a Modality-Gated Extractor. Cross-modality similarity preservation is achieved by a Focal Modality-Aware Similarity-Preserving Loss (see Sect. 3.1) for guiding cross-modality feature learning. The Modality-Gated Extractor consists of a Modality-Gated CNN followed by a shared convolution layer for feature extraction. The Modality-Gated CNN (see Sect. 4.3) is constructed using modality-gated nodes (see Sect. 4.2) and can flexibly learn modality-specific and shared structures that are suited to the training data. The softmax loss L_{cls} is applied for learning discriminative features for classification. The implementation details of the framework are illustrated in Sect. 6.2. Note that, $\ominus$ denotes the splitting of sample features by modality (best viewed in colour)

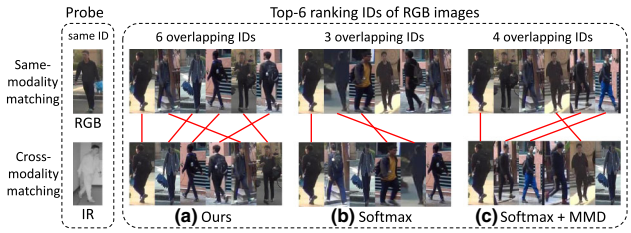


Fig. 6 Comparison of ranking results from same-modality matching and cross-modality matching based on features learned using three different losses: **a** ours, **b** Softmax, and **c** Softmax + MMD (Gretton et al. 2012). RGB images in the training set are retrieved based on either a probe RGB image or a probe IR image of the same identity. The top-6 ranking lists of identities are shown, where red lines indicate the same identity. This shows that our Focal Modality-Aware Similarity-Preserving Loss results in more overlapping identities in the ranking lists, demonstrating that the consistency between same-modality and cross-modality ranking is better than those in the cases of "Softmax" and "Softmax + MMD" (Color figure online)

of the Softmax and MMD losses (Gretton et al. 2012) that is representative for domain adaptation. RGB images in the training set are retrieved for either a probe RGB image or a probe IR image of the same identity. The top-6 ranking lists of identities are shown, where red lines indicate the same identity. We also calculate the proportions of overlapping identities in the top-6 ranking lists of identities on the testing set. The proportions are 38.0% for "Ours", 31.8% for "Softmax" and 31.6% for "Softmax + MMD". Our method can retrieve more overlapping identities in the top ranking lists, indicating that the cross-modality similarity is more consistent with same-modality similarity. In this way, the modality discrepancy in the similarity space is alleviated, i.e., the modality-specific information in the shared feature space of two modalities is alleviated. Experiments in Sect. 6.5 show that cross-modality similarity preservation can improve the performance of cross-modality matching significantly.

4 Modality-Gated Nodes

To determine how to assist the feature extractor in extracting shared knowledge for cross-modality matching, we first analyse the functionality of deep neural networks for cross-modality image matching by means of defining the concepts of modality-specific and shared nodes. Then, we propose the modality-gated node, a generalised structure that can represent both modality-specific and shared nodes. We use modality-gated nodes to construct a structure-learnable feature extractor called Modality-Gated Extractor, which is able to construct more complex modality-specific structures compared with the manually designed and fixed modality-specific structures used in existing methods (Lin et al. 2017; Kan et al. 2016; He et al. 2017, 2019).

4.1 Modality-Specific and Shared Nodes

As discussed in Sect. 2.3, the commonly used structures for cross-modality image matching, namely, the one-stream structure, the two-stream structure and the asymmetric FC layer structure, all consist of both modality-specific and shared structures. Although these three structures all seem to be different, the two-stream structure can actually be represented as a one-stream structure if a *modality selection block* is employed in the network, as shown in Fig. 7.

Modality Selection Block The modality selection block f_{sel} is defined as follows:

$$f_{sel}(\mathbf{x}, y^{mod}) = \begin{cases} [\mathbf{E}_d, \mathbf{O}_d]^\top \mathbf{x}, & y^{mod} = 1, \\ [\mathbf{O}_d, \mathbf{E}_d]^\top \mathbf{x}, & y^{mod} = 2, \end{cases} \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input, $y^{mod} \in \{1, 2\}$ is a modality label indicating modality 1 or modality 2, $\mathbf{E}_d \in \mathbb{R}^{d \times d}$ is the identity matrix, and $\mathbf{O}_d \in \mathbb{R}^{d \times d}$ is the zero matrix.

We attempt to analyse the two-stream structure in the same way as the one-stream structure. To enable the decomposition of a network into modality-specific and shared structures for analysis, we define modality-specific nodes and shared nodes as the basic network components.

Definition 1 For a neural network with input from two modalities, modality 1 and modality 2, the nodes in each layer can be categorised into three types: modality-1-specific nodes, modality-2-specific nodes and shared nodes. Let $\mathbf{x}_{(l)}^{m1}$ and $\mathbf{x}_{(l)}^{m2}$ denote the inputs to layer $l+1$ from modality 1 and modality 2, respectively. In particular, $\mathbf{x}_{(0)}^{m1}$ and $\mathbf{x}_{(0)}^{m2}$ are the inputs to the network. Let $\eta_{(l),i}$ denote the i-th node in layer l, and let $\eta_{(l),i}(\mathbf{x}_{(0)})$ denote the output of $\eta_{(l),i}$ with the network input $\mathbf{x}_{(0)}$:

$$\eta_{(l),i}(\mathbf{x}_{(0)}) = \sigma\left(\sum_j w_{(l),j,i}\,\eta_{(l-1),j}(\mathbf{x}_{(0)}) + b_{(l),i}\right), \tag{9}$$

where $\sigma(\cdot)$ is the activation function, and $w_{(l),j,i}$ and $b_{(l),i}$ are the weight and bias parameters of layer l, respectively. The type of node $type(\eta_{(l),i})$ is defined as follows:

$$type(\eta_{(l),i}) = \begin{cases} modality\text{-}1\text{-}specific, & \eta_{(l),i}(\mathbf{x}_{(0)}^{m1}) \not\equiv 0 \\ & and\ \eta_{(l),i}(\mathbf{x}_{(0)}^{m2}) \equiv 0, \\ modality\text{-}2\text{-}specific, & \eta_{(l),i}(\mathbf{x}_{(0)}^{m2}) \not\equiv 0 \\ & and\ \eta_{(l),i}(\mathbf{x}_{(0)}^{m1}) \equiv 0, \\ shared, & otherwise. \end{cases} \tag{10}$$

Note that for modality-1-specific nodes, we use the identity sign in $\eta_{(l),i}(\mathbf{x}_{(0)}^{m2}) \equiv 0$, which means that for any input from modality 2, the output of node $\eta_{(l),i}$ is always zero. A similar condition holds for modality-2-specific nodes.

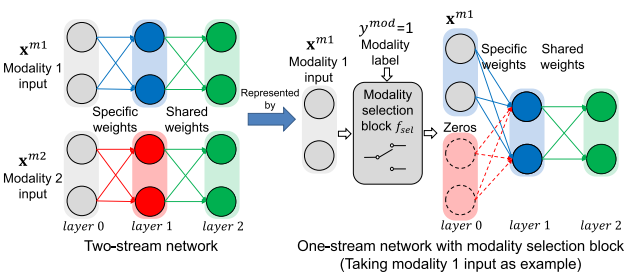


Fig. 7 Explanation of how a one-stream network with a modality selection block can be used to represent a two-stream network. The modality selection block modifies the input by padding it with zeros in different positions for different modalities so that only the weights for non-zero nodes corresponding to the modality of the input data will take effect (best viewed in colour)

We assume that nodes of all three types exist in a network, and analyse the properties of modality-specific and shared nodes, where the corresponding derivations are presented in the section of "Properties of Modality-Specific and Shared Nodes" in the "Appendix".

Properties of Modality-Specific Nodes In forward propagation, the modality-specific weight parameters $\mathbf{w}_{(l+1),i}^{1spe}$ and $\mathbf{w}_{(l+1),i}^{2spe}$ affect only the input from corresponding modality. In backward propagation, these parameters can only be updated based on input from corresponding modality.

Properties of Shared Nodes In forward propagation, the shared weight parameters $\mathbf{w}_{(l+1),i}^{sh}$ affect both modalities. In backward propagation, they are updated based on inputs from both modalities.

Analysis Based on the properties described above, we find that all of the existing networks for cross-modality matching constructed with modality-specific and shared structures as discussed in Sect. 2.3 can be represented by a one-stream network consisting of modality-specific and shared nodes. We take a two-stream network as an example. As shown in Fig. 7, according to Definition 1, in layer 0 of the one-stream network with a modality selection block, the first two nodes are modality-1-specific nodes, and the last two nodes are modality-2-specific nodes. The nodes in the subsequent layer 1 and layer 2 are shared nodes.

However, it is difficult to manually determine how many nodes in a network should be of the modality-specific or shared type because this number depends on the task-specific data distribution. We overcome this problem by designing a modality-gated node based on a one-stream network, as described in the following section.

4.2 Modality-Gated Nodes

While modality-specific and shared nodes are the keys to modality-specific modelling in neural networks, constructing modality-specific structures using modality-specific nodes (e.g., the unshared convolution layers in the two-stream structure) as defined in Sect. 4.1 is a "hard" strategy. This "hard" strategy only allows a node to be either unshared or shared between the two modalities; there is no way to express the extent to which a node can be shared, i.e., partial sharing. Thus, this approach is not sufficiently flexible for the challenging RGB-IR Re-ID task.

Therefore, we aim to develop a model structure that allows partial sharing between two modalities with different weights. To achieve such "soft" modelling, we propose the modality-gated node as a universal structure that can represent both modality-specific and shared nodes and has learnable parameters that control the degree of modality

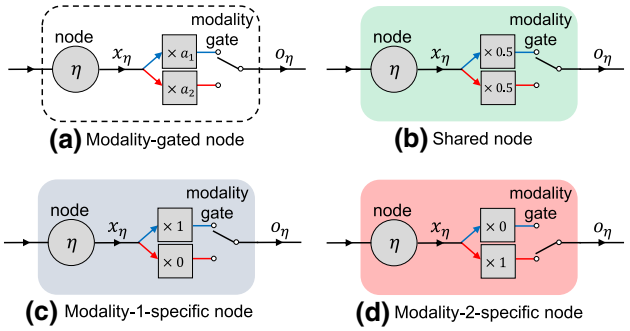


Fig. 8 Structure of a modality-gated node (sub-figure **a**) and how it can represent a shared node (sub-figure **b**), a modality-1-specific node (sub-figure **c**) or a modality-2-specific node (sub-figure **d**) with different values of the coefficients a_1 and a_2 (best viewed in colour)

specificity. We regard such a modality-gated node as a *soft modality-specific node*.

The structure of a modality-gated node is shown in (a) of Fig. 8. For a normal node η, the original output of the node is x_η. Then, there are two branches with modality selection weights of a_1 and a_2, with values ranging from 0 to 1, by which x_η will be multiplied. The modality gate is controlled by the modality label $y^{mod} \in \{1, 2\}$, which indicates whether the sample belongs to modality 1 or modality 2. The output of the modality-gated node, o_η, is determined as follows:

$$o_\eta = \begin{cases} a_1 x_\eta, & y^{mod} = 1, \\ a_2 x_\eta, & y^{mod} = 2, \end{cases} \quad where\ a_1 + a_2 = 1,\ a_1, a_2 \geq 0. \tag{11}$$

To avoid the case in which both a_1 and a_2 are 0, resulting in a dead node, we subject a_1 and a_2 to the constraints $a_1 + a_2 = 1$ and $a_1, a_2 \geq 0$. In the optimisation process, to avoid directly clipping the values of a_1 and a_2 to satisfy the constraints, we parameterise a_1 and a_2 using two parameters a_1' and a_2' as follows:

$$a_1 = \frac{|a_1'|}{|a_1'| + |a_2'|},\ a_2 = \frac{|a_2'|}{|a_1'| + |a_2'|}, \tag{12}$$

where a_1' and a_2' are unconstrained non-zero real numbers.

Forward Propagation Analysis In a modality-gated node, the key parameters are the modality selection weights a_1 and a_2, which take continuous values ranging from 0 to 1. With different values of the modality selection weights, modality-gated nodes can represent nodes with different degrees of modality specificity. Sub-figures (b)∼(d) in Fig. 8 show three examples of different types of nodes that modality-gated nodes can represent when the modality selection weights are varied. Given two modalities (modality 1 and modality 2), three special cases are listed below:

(1) When $a_1 = 0.5$ and $a_2 = 0.5$, the modality-gated node represents a shared node (see Fig. 8b).
(2) When $a_1 = 1$ and $a_2 = 0$, the modality-gated node represents a modality-1-specific node (see Fig. 8c).
(3) When $a_1 = 0$ and $a_2 = 1$, the modality-gated node represents a modality-2-specific node (see Fig. 8d).

In these three cases, the conditions in Definition 1 are strictly satisfied; thus, modality-gated nodes are able to represent both modality-specific and shared nodes.

The analysis shows that modality-specific and shared nodes are the extreme cases of modality-gated nodes. When the modality selection weights a_1 and a_2 have values between 0 and 1, the node is a soft modality-specific node, which is partially shared by both modalities; such a node tends to show higher modality 1 (or modality 2) specificity when the weight a_1 (or a_2) is higher. Therefore, modality-gated nodes provide more flexible means of constructing more complex modality-specific structures compared with the manually designed, fixed modality-specific structures used in existing methods (e.g., Lin et al. 2017; Kan et al. 2016; He et al. 2017, 2019, as discussed in Sect. 2.3).

Backward Propagation Analysis Let us further analyse the behaviour of modality-gated nodes in backward propagation. We compute the derivative of the node output o_η with respect to x_η:

$$\frac{\partial o_\eta}{\partial x_\eta} = \begin{cases} a_1, & y^{mod} = 1, \\ a_2, & y^{mod} = 2. \end{cases} \tag{13}$$

When a gradient flow passes through the modality gate, it is weighted by a_1 and a_2 for modality 1 and modality 2, respectively. Thus, the learning processes for the two modalities are different but partially shared.

Since the modality selection weights a_1 and a_2 can be learned through end-to-end training, a network constructed of modality-gated nodes can evolve both modality-specific and shared structures by learning from data without requiring any manual design. During training, the nodes can automatically evolve into soft modality-specific nodes under the guidance provided by the loss function.

Remarks: Connection with Deep Zero Padding Deep zero padding (Wu et al. 2017) was exploited in our previous work. In this approach, modality-specific nodes are generated in the input layer by padding the input images with zeros; in this way, the input layer acts as a modality selection block (Sect. 4.1). This approach is a special case in which fixed modality-gated nodes are applied only in the input layer; thus, this approach is not as flexible as the approach formulated in this work, in which the entire network is constructed from learnable modality-gated nodes.

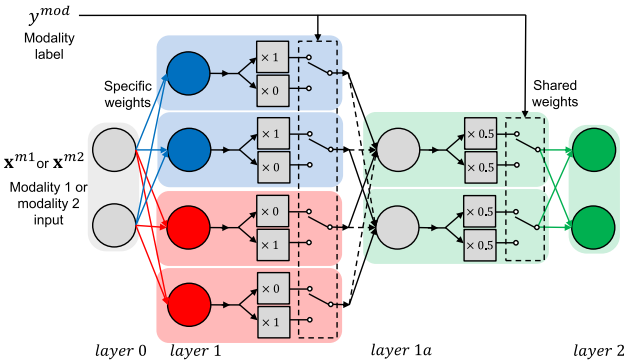


Fig. 9 A one-stream network consisting of modality-gated nodes that can serve as a representation of the two-stream network shown in Fig. 7 for the case of forward propagation. We assume that in this special case, the modality selection weights are set as shown in this figure. In layer 1a, the black solid lines denote weights of 1, and the black dotted lines denote weights of 0. This is a simple example showing the structural representation ability of modality-gated nodes (best viewed in colour)

Connection with Existing Network Structures As discussed in Sect. 2.3, most neural networks for cross-modality modelling, e.g., networks with the two-stream structure or the asymmetric FC layer structure, consist of both modality-specific nodes and shared nodes. Modality-gated nodes can be used to represent soft modality-specific nodes and provide sufficient flexibility to allow a network to evolve both modality-specific and shared structures; thus, the network can act as a structure-learnable feature extractor. An example of how a one-stream network consisting of modality-gated nodes can be used to represent a two-stream network is shown in Fig. 9, and the corresponding analysis is presented in the section of "Structure Representation Ability" in the "Appendix".

4.3 Modality-Gated Extractor

To extract features from input images, we use modality-gated nodes to construct a CNN as the basis for the Modality-Gated Extractor shown in the dashed box in Fig. 5. We build the modality-gated CNN based on a given backbone model (e.g., ResNet He et al. 2016) by replacing all nodes in the backbone model with modality-gated nodes. A subsequent shared convolution layer is included for further learning of shared feature representation. The modality-gated CNN and the shared convolution layer form the Modality-Gated Extractor f_{ex}.

The nodes in fully connected (FC) network correspond to channels in the CNN. For each feature map channel $\mathbf{X}_\eta$, the output $\mathbf{O}_\eta$ of a modality-gated node as in Eq. (11) is

$$\mathbf{O}_\eta = \begin{cases} a_1\mathbf{X}_\eta, & y^{mod} = 1, \\ a_2\mathbf{X}_\eta, & y^{mod} = 2, \end{cases} \quad where\ a_1 + a_2 = 1,\ a_1, a_2 \geq 0, \tag{14}$$

where a_1 and a_2 are scalars with values in [0, 1] corresponding to RGB and IR feature map channels, respectively.

In the training stage, the modality-gated nodes in Modality-Gated Extractor assist our Focal Modality-Aware Similarity-Preserving Loss to learn effective features for cross-modality matching, as shown in Fig. 5. In the testing stage, given a test image $\mathbf{I}_i$ with a corresponding modality label y_i^{mod}, the Modality-Gated Extractor is used for feature extraction. To measure the similarity between an RGB image and an IR image, the cosine distance is computed.

5 An RGB-IR Person Re-ID Dataset

5.1 Dataset Description

Since there is currently no available RGB-IR person Re-ID dataset collected by surveillance cameras, we collected a new multi-modality Re-ID dataset called SYSU-MM01 for evaluating the RGB-IR cross-modality person Re-ID (RGB-IR Re-ID) problem. SYSU-MM01 contains images captured by 6 cameras, including two IR cameras and four RGB ones. Unlike RGB cameras, IR cameras detect near infrared (NIR) light and function well under dark conditions. We present the details of the dataset in Table 2 and show some examples from each camera view in Fig. 10. The RGB images from camera 1 and camera 2 were captured in two bright indoor rooms (room 1 and room 2). For each person, at least 400 RGB images were captured, of different poses and from dif-

Table 2 An overview of the SYSU-MM01 dataset

Cam	Location	In-/outdoors	Lighting	#IDs	#RGB/ID	#IR/ID
1	Room 1	Indoors	Bright	259	20+	–
2	Room 2	Indoors	Bright	259	20+	–
3	Room 2	Indoors	Dark	486	–	20
4	Gate	Outdoors	Bright	493	20	–
5	Garden	Outdoors	Bright	502	20	–
6	Passage	Outdoors	Dark	299	–	20

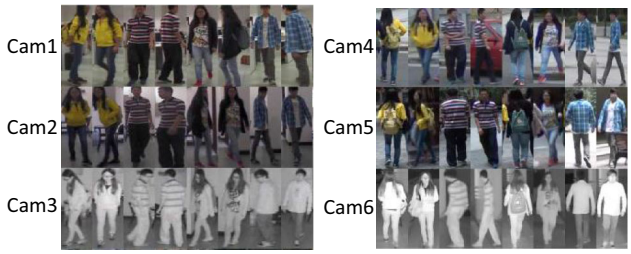


Fig. 10 Examples of RGB images and IR images in our SYSU-MM01 dataset. Cameras 1, 2, 4 and 5 were RGB cameras, and cameras 3 and 6 were IR cameras. Cameras 1 to 3, as shown on the left, captured indoor scenes, and cameras 4 to 6, as shown on the right, captured outdoor scenes. Every two columns contain images of the same person (best viewed in colour)

ferent viewpoints. The IR images from camera 3 and camera 6 were captured in the dark. IR images have only one channel, unlike RGB images, which consist of three channels. Camera 3 was in room 2 in a dark environment, while camera 6 was in an outdoor passage with background clutter. Camera 4 and camera 5 were RGB surveillance cameras in two outdoor scenes, named gate and garden.

The images captured by IR cameras (camera 3 and camera 6) are distinct from RGB images in terms of both colour and exposure. Specifically, although camera 2 and camera 3 were both placed indoors, their images show dramatic colour shifts and exposure differences. For example, for the first person shown in Fig. 10, her yellow clothes are distinct from her black trousers in RGB images, but this colour distinction is nearly eliminated in IR images (columns 1 and 2, rows 2 and 3 in Fig. 10). IR images have only one channel and may lose some textural details. The exposure of IR images captured at different distances is also an issue. These concerns all make the RGB-IR Re-ID problem challenging.

5.2 Evaluation Protocol

There are 491 valid identities in the SYSU-MM01 dataset. We have established a fixed split with 395 identities for training and 96 for testing. During training, all images of the 395 persons in the training set from all camera views can be used.

For cross-modality person Re-ID matching, we have designed two search modes, namely, the *all-search* mode and the *indoor-search* mode. For the all-search mode, the gallery images are from RGB cameras 1, 2, 4 and 5, and the probe images are from IR cameras 3 and 6. For the indoor-search mode, the gallery images are from RGB cameras 1 and 2 (excluding the outdoor views from cameras 4 and 5), and the probe images are from IR cameras 3 and 6.

Matching For both modes, we adopted single-shot setting. For every identity in each RGB camera view, we randomly chose one image of that person to form the gallery set for the single-shot setting. For the probe set, all images were used.

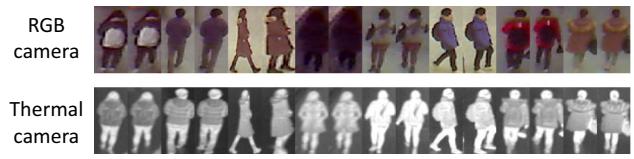


Fig. 11 Examples of RGB images and thermal images in the RegDB (Nguyen et al. 2017) dataset. The images were captured by two cameras, an RGB camera and a thermal camera. Every two columns contain images of the same person (best viewed in colour)

Given a probe image, matching was conducted by computing the similarities between the probe image and the gallery images. Note that matching was conducted only between camera views from different locations, as shown in Table 2. Camera 2 and camera 3 captured data of different modalities, but they were deployed in the same indoor scene; therefore, for the probe images from camera 3, the gallery images from camera 2 were skipped.

Measurement For a quantitative evaluation, we computed the cumulative matching characteristic (CMC) curve and the mean average precision (mAP) based on the ranking list obtained by measuring the similarities between the probe and gallery images. The above evaluation was repeated 10 times with random splits of the gallery and probe sets, and the average performance is reported.

6 Experiments

We conducted extensive comparative evaluations of our cross-modality matching framework against existing Re-ID, cross-modality matching and domain adaptation models on our SYSU-MM01 dataset for RGB-IR Re-ID. In addition, we evaluated on an RGB-thermal person dataset called RegDB (Nguyen et al. 2017) for cross-modality RGB-thermal matching.

6.1 Experimental Settings

Datasets Our evaluations were conducted on our SYSU-MM01 dataset and the RegDB dataset (Nguyen et al. 2017). A description of the SYSU-MM01 dataset is presented in Sect. 5.1. RegDB (Nguyen et al. 2017) contains 8,240 images associated with 412 identities captured by an RGB camera and a thermal camera. For each identity, there are 10 RGB images and 10 thermal images. Some examples of RegDB (Nguyen et al. 2017) are shown in Fig. 11.

Compared with RegDB (Nguyen et al. 2017), our SYSU-MM01 includes images from more cameras and samples with more variations. The thermal images in RegDB are not as suitable for surveillance applications as the NIR images in SYSU-MM01 because of the high price of thermal cameras.

Although RegDB was not collected specifically for the Re-ID task, we additionally tested our method on this dataset to achieve a more extensive evaluation of cross-modality matching.

Evaluation Protocols For our SYSU-MM01 dataset, we followed the evaluation protocol introduced in Sect. 5.2. For RegDB (Nguyen et al. 2017), we followed the evaluation protocol reported in Ye et al. (2018). Half of the identities were used for training, and the remaining identities were used for testing. The thermal images were used as the gallery set, and the RGB images were used as the probe set.

6.2 Implementation Details

Backbone One-stream Model In our implementation, we adopted PCB (Sun et al. 2018), which is based on ResNet-50 (He et al. 2016), as the backbone one-stream model. Compared with ResNet-50 (He et al. 2016), in PCB (Sun et al. 2018), the global average pooling layer is replaced by a 1×1 convolution layer for channel number reduction. The IR images were converted into three-channel images by replicating their single channel three times. The input images were resized to 384×128 as inputs to the model. Our Modality-Gated Extractor and the baseline models considered for comparison were all based on this backbone one-stream model.

Model Structure and Hyperparameters Our Modality-Gated Extractor was based on the backbone one-stream model introduced above. As shown in Fig. 5, a Modality-Gated Extractor is formed of a modality-gated CNN and a shared convolution layer. The structure of the modality-gated CNN was based on ResNet-50 with the global average pooling layer discarded and all nodes replaced with modality-gated nodes. Following the structural design of PCB (Sun et al. 2018), the shared convolution layer was a 1×1 convolution layer for reducing the number of channels from 2048 to 256. Then, the feature map was equally split into 6 horizontal stripes, and global average pooling was applied to each stripe. Finally, the output feature map was reshaped as a vector and normalised using the ℓ_2-norm.

The default value of λ, the weight of the Focal Modality-Aware Similarity-Preserving Loss L_{FMSP}, was 10.0. In the modality-gated nodes, the values of the modality selection weights a_1 and a_2 were initialised to 0.5 with $a_1' = a_2' = 1.0$, i.e., all nodes were initialised as shared nodes. For training, we followed the strategy described in Sun et al. (2018).

6.3 Compared Methods

We mainly compared our method with existing Re-ID, cross-modality matching, domain adaptation frameworks and CNN structures for cross-modality image matching discussed in Sect. 2.3.

Cross-Modality Methods We compared our method with several recently developed deep cross-modality methods, including cross-modality RGB-thermal Re-ID method TONE + HCML (Ye et al. 2018), BDTR (Ye et al. 2018) and cmGAN (Dai et al. 2018); the cross-modality face recognition method IDR (He et al. 2017); cross-modality matching methods DeepCCA (Andrew et al. 2013) and generalised similarity measure (GSM) (Lin et al. 2017); the domain adaptation methods MMD (Gretton et al. 2012) and DeepCORAL (Sun and Saenko 2016); and a method using a triplet loss (Schroff et al. 2015) for domain alignment (Softmax + triplet). Among these methods, TONE + HCML (Ye et al. 2018) and BDTR (Ye et al. 2018) are based on the two-stream network structure, and cmGAN (Dai et al. 2018) is based on the one-stream network structure. TONE + HCML (Ye et al. 2018), BDTR (Ye et al. 2018) and cmGAN (Dai et al. 2018) have previously been evaluated on our SYSU-MM01 dataset and RegDB (Nguyen et al. 2017); thus, we directly used their reported results. The generalised similarity measure (GSM) approach proposed by Lin et al. (2017). (referred to as Lin's method) is based on a deep framework with a two-stream structure. For MMD (Gretton et al. 2012) (denoted by DeepMMD when implemented with a deep neural network), DeepCORAL (Sun and Saenko 2016) and Softmax + triplet (Schroff et al. 2015), since these methods do not require specific network structures, we adopted our backbone one-stream model and used their loss functions to ensure fair comparisons. By contrast, DeepCCA (Andrew et al. 2013), IDR (He et al. 2017) and Lin's method (Lin et al. 2017) require specifically designed structures in the feature extractors; therefore, we used the codes released by the authors.

We also evaluated the three CNN structures for cross-modality matching shown in Fig. 3 and discussed in Sect. 2.3, including the one-stream network structure, the two-stream network structure and the asymmetric FC layer network structure, all of which were based on the same backbone one-stream model in our evaluation. The "FC" layer in Fig. 3 corresponded to the last 1×1 convolution layer in the backbone one-stream model. For the one-stream network structure, we simply used the backbone one-stream model, i.e., the PCB model (Sun et al. 2018). For the two-stream and asymmetric FC layer network structures, we used the same convolution blocks as in the one-stream network, and the modality-specific parts used unshared parameters (i.e., twice the number of parameters). The softmax loss was used, as in the backbone model PCB (Sun et al. 2018).

Metric/Subspace Models We combined favourable hand-crafted features with metric/subspace learning models for RGB-IR cross-modality Re-ID to show that features learned from data are superior to conventional hand-crafted features in solving the problem. The hand-crafted features considered for comparison included HOG (Dalal and Triggs 2005), LOMO (Liao et al. 2015) and HIPHOP (Chen et al.

Table 3 Performances in the all-search and indoor-search modes on SYSU-MM01

Method	All-search				Indoor-search			
	mAP	r1	r5	r10	mAP	r1	r5	r10
Ours	44.98	43.56	74.61	86.25	57.50	48.62	79.01	89.50
DeepMMD (Gretton et al. 2012)*	41.69	40.70	71.77	83.46	51.89	42.29	74.38	86.17
DeepCORAL (Sun and Saenko 2016)*	40.83	39.30	70.32	82.48	50.84	41.03	72.77	84.82
Softmax + triplet (Schroff et al. 2015)*	41.10	38.96	71.78	84.03	53.83	43.50	78.15	89.36
One-stream (PCB (Sun et al. 2018))	38.39	36.91	68.88	81.67	50.46	40.38	73.70	85.55
Two-stream	39.59	38.10	68.38	80.30	50.75	40.82	72.83	84.16
Asymmetric FC	40.03	38.71	69.70	81.80	50.62	40.55	72.62	85.07
Deep zero padding	41.77	40.95	71.61	82.59	52.13	42.31	73.87	85.60
cmGAN (Dai et al. 2018)	27.80	26.97	–	67.51	42.19	31.63	–	77.23
BDTR (Ye et al. 2018)	19.66	17.01	–	55.43	–	–	–	–
TONE + HCML (Ye et al. 2018)	16.16	14.32	–	53.16	–	–	–	–
Lin's (Lin et al. 2017)	9.96	7.28	24.48	38.39	19.49	10.30	32.21	50.42
IDR (He et al. 2017)	14.84	13.13	34.99	50.20	26.38	16.56	44.48	62.31
DeepCCA (Andrew et al. 2013)	14.92	12.10	33.67	47.94	25.57	16.20	41.00	55.97
HIPHOP + CRAFT (Chen et al. 2018)	3.67	1.88	8.01	14.78	8.95	2.95	12.81	23.52
HOG (Dalal and Triggs 2005) + XQDA (Liao et al. 2015)	5.69	3.68	13.11	23.18	11.25	4.57	17.28	30.71
HOG (Dalal and Triggs 2005) + LFDA (Pedagadi et al. 2013)	5.37	3.36	12.73	21.92	10.06	3.84	15.34	26.78
HOG (Dalal and Triggs 2005) + CCA (Rasiwasia et al. 2010)	4.59	2.91	11.58	20.01	11.45	4.70	18.56	31.30
HOG (Dalal and Triggs 2005) + CDFE (Lin and Tang 2006)	4.54	2.32	10.33	19.01	9.86	3.30	15.25	27.69
HOG (Dalal and Triggs 2005) + GMA (Sharma et al. 2012)	2.78	1.01	5.28	10.31	6.93	1.74	8.97	17.63
HOG (Dalal and Triggs 2005) + SCM (Zhang and Li 2014)	3.94	2.05	8.51	16.12	9.94	3.54	15.42	27.88
LOMO (Liao et al. 2015) + XQDA (Liao et al. 2015)	6.54	3.75	15.57	26.63	12.89	4.93	20.81	37.11
LOMO (Liao et al. 2015) + LFDA (Pedagadi et al. 2013)	6.52	4.40	16.05	26.54	13.50	5.82	22.54	36.85
LOMO (Liao et al. 2015) + CCA (Rasiwasia et al. 2010)	5.49	4.45	15.37	25.27	15.13	7.71	26.12	41.82
LOMO (Liao et al. 2015) + CDFE (Lin and Tang 2006)	6.74	3.96	16.66	29.05	13.21	5.00	22.56	38.50
LOMO (Liao et al. 2015) + GMA (Sharma et al. 2012)	2.81	1.07	5.23	10.29	6.96	1.94	8.89	17.69
LOMO (Liao et al. 2015) + SCM (Zhang and Li 2014)	4.26	2.40	10.13	18.06	10.95	4.30	17.39	30.46

Here, r1, r5, and r10 denote the rank-1, rank-5, and rank-10 accuracies (%). "*" denotes re-implementation using the same backbone model as our method. "–" denotes a result not reported in the published paper

2018) features. For the 1-channel IR images, when the hand-crafted feature extraction methods required 3-channel images, the existing channel was duplicated three times. The metric/subspace models considered for comparison included XQDA (Liao et al. 2015), LFDA (Pedagadi et al. 2013); the cross-modality methods CCA (Rasiwasia et al. 2010), CDFE (Lin and Tang 2006), GMA (Sharma et al. 2012), and CRAFT (Chen et al. 2018); and the cross-modality binary representation learning method SCM (Zhang and Li 2014).

6.4 Model Comparison and Analysis

The experimental results on our SYSU-MM01 dataset are reported in Table 3, including the rank-1, rank-5, and rank-10 accuracies and the mAP values of our method and all compared methods in the all-search and indoor-search modes. The results obtained on RegDB (Nguyen et al. 2017) are reported in Table 4.

Comparison with Deep Cross-Modality Frameworks Our method outperformed all related deep models considered for comparison. Compared to frameworks using related losses with the same backbone model as our method, DeepMMD (Gretton et al. 2012), DeepCORAL (Sun and Saenko 2016) and Softmax + triplet (Schroff et al. 2015) show minor improvements. However, they are not as effective as our method because they focus only on the alignment of the feature distributions between the two modalities; in comparison, we do not operate under the assumption of identical distributions between training and testing data, and we guide the learning of shared knowledge for cross-modality matching by imposing same-modality matching as a constraint for cross-modality similarity preservation.

Table 4 Performance on RegDB (Nguyen et al. 2017)

Method	mAP	r1	r5	r10
Ours	64.50	65.07	76.70	83.71
DeepMMD (Gretton et al. 2012)*	52.51	51.82	67.79	76.77
DeepCORAL (Sun and Saenko 2016)*	51.51	52.06	66.43	74.78
Softmax + triplet (Schroff et al. 2015)*	50.00	50.12	63.50	71.33
One-stream (PCB (Sun et al. 2018))	48.29	47.60	61.70	69.81
Two-stream	49.11	48.18	62.72	71.38
Asymmetric FC	48.56	47.18	62.57	71.41
Deep zero padding	50.32	50.05	63.88	72.65
BDTR (Ye et al. 2018)	31.83	33.47	–	58.42
TONE + HCML (Ye et al. 2018)	20.80	24.44	–	47.53
Lin's (Lin et al. 2017)	15.06	17.28	–	34.47

Here, r1, r5, and r10 denote the rank-1, rank-5, and rank-10 accuracies (%). "*" denotes re-implementation using the same backbone model as our method. "–" denotes a result not reported in the published paper

The TONE + HCML (Ye et al. 2018), BDTR (Ye et al. 2018) and cmGAN (Dai et al. 2018) methods for cross-modality RGB-thermal Re-ID and the IDR (He et al. 2017) method for VIS-NIR face recognition did not perform sufficiently well for RGB-IR Re-ID. The reason is because they use either a two-stream (TONE + HCML and BDTR), one-stream (cmGAN) or asymmetric FC layer (IDR) network structure, all of which are fixed structures and cannot learn modality-specific and shared network structures flexibly as using modality-gated nodes in our method. Moreover, ours is also different from these methods by specially considering the preservation of cross-modality similarity for learning effective shared knowledge.

Deep Models Versus Metric/Subspace Models From Table 3, it can be observed that deep models can outperform metric/subspace models by large margins. All cases of hand-crafted features combined with metric/subspace learning models performed poorly. Even the rank-1 accuracy in the best case did not reach 10%. LOMO features contain rich colour and texture information and perform well for the RGB-based Re-ID problem. However, these features nevertheless failed here because of the heterogeneity of RGB and IR image data due to the different imaging processes; consequently, the colour information becomes non-discriminative. Although HOG features also capture both texture and shape information, they failed as well. Our empirical results suggest that these hand-crafted low-level features are not suitable for the RGB-IR cross-modality Re-ID task even when combined with metric learning. In comparison, deep models are more feasible.

Comparison with Other Network Structures. Tables 3 and 4 show that compared to the backbone one-stream model, the improvement achieved with our method is clear; our method performed approximately 7% better in terms of the rank-1 accuracy and mAP on our SYSU-MM01 database, and the corresponding improvement on RegDB (Nguyen et al. 2017) was greater than 16%. Our method also outperformed the two-stream and asymmetric FC layer network structures in which the modality-specific and shared structures were manually designed. By means of our proposed modality-gated nodes, our method can learn modality-specific and shared network structures from data without prior knowledge. The results suggest that our Modality-Gated Extractor can evolve a more effective network structure for extracting shared knowledge compared to networks with fixed modality-specific and shared structures.

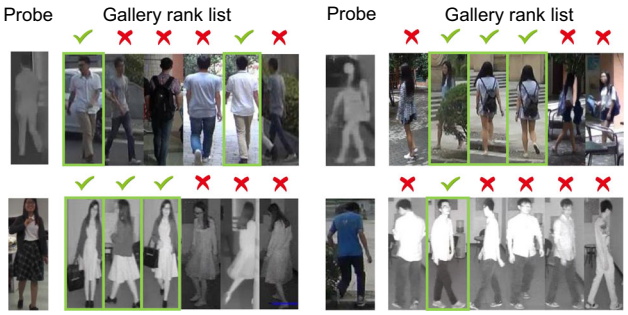


Fig. 12 Matching examples of the top-1 (leftmost) to top-6 (rightmost) gallery ranking lists of our method on SYSU-MM01. In the first row, the probe images are IR images, and the gallery images are RGB images. In the second row, the probe images are RGB images, and the gallery images are IR images. The green bounding boxes indicate correct matches. It can be observed that even human viewers have difficulty distinguishing different people without colour information (best viewed in colour) (Color figure online)

Matching Examples We show some matching examples from the top-1 to top-6 ranking lists obtained with our method on SYSU-MM01 in Fig. 12. First, from a brief look at the images, it is clear that even a human viewer has difficulty

in identifying which person is the correct match based on only colour information because there is no exact theoretical relation between the greyscale values in RGB images and IR images. However, there are still some useful cues for matching. From the top-ranked images, we can analyse what cues may be useful for distinguishing individuals. For example, the images in the top left group depict thin men with short hair in short-sleeved shirts, whereas the images in the top right group show women of medium stature with long hair carrying small bags.

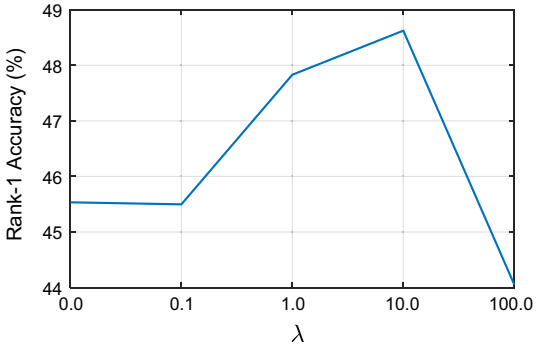


Fig. 13 Effect of λ on the rank-1 accuracy achieved in the indoor-search mode on SYSU-MM01. λ is the weight controlling the effect of the Focal Modality-Aware Similarity-Preserving Loss L_{FMSP} in our loss function L (Eq. 7)

6.5 Further Analysis

Component Evaluation To validate the effectiveness of each component in our proposed framework, we evaluated two key components: the modality-gated nodes and the Focal Modality-Aware Similarity-Preserving Loss L_{FMSP}. We applied both these components with the backbone one-stream model as the baseline. The comparison results are reported in Table 5, in which the following notations are used. "One-stream" refers to the backbone one-stream model with the Softmax cross-entropy loss L_{cls} in Eq. (7). "L_{FMSP}" denotes the Focal Modality-Aware Similarity-Preserving Loss in Eq. (6). "Modality-Gated Extractor" refers to the backbone one-stream model with modality-gated nodes. "L_{MSP}" denotes the Modality-Aware Similarity-Preserving Loss in Eq. (3) (without the confidence factors used in L_{FMSP}). "+" denotes the combination of two components in the framework. "Modality-Gated Extractor + L_{FMSP}" is the full model of our proposed method.

To evaluate the effectiveness of the Focal Modality-Aware Similarity-Preserving Loss L_{FMSP} (Eq. 6), we compared "one-stream" with "one-stream + L_{FMSP}" and compared "Modality-Gated Extractor" with "Modality-Gated Extractor + L_{FMSP}". Compared with "one-stream", the improvement in the mAP of "one-stream + L_{FMSP}" is approximately 6% on both datasets. The Focal Modality-Aware Similarity-Preserving Loss L_{FMSP} can guide cross-modality feature learning by cross-modality similarity preservation to mine shared knowledge. Moreover, to validate the effectiveness of the confidence factors in L_{FMSP} in Eq. (6), we compared "Modality-Gated Extractor + L_{MSP}" with "Modality-Gated Extractor + L_{FMSP}". The results show that without the confidence factors, the performance dropped because the confidence factors can help to emphasise the contribution of reliable sample pairs during cross-modality similarity preservation while avoiding learning incorrect information from unreliable sample pairs. To evaluate the effectiveness of modality-gated nodes, we compared "one-stream" with "Modality-Gated Extractor" and compared "one-stream + L_{FMSP}" with "Modality-Gated Extractor + L_{FMSP}". Compared with "one-stream", the improvement in the mAP of "Modality-Gated Extractor" is approximately 5% on both datasets. Using modality-gated nodes in a one-stream network can enable the learning of suitable modality-specific and shared structures in the network and thus enable the extraction of better features for cross-modality matching, leading to improved performance. The full model of our proposed method, "Modality-Gated Extractor + L_{FMSP}", achieved the best performance.

Effect of the Parameter λ in the Loss In our loss function L (Eq. 7), λ is the weight for controlling the effect of the Focal Modality-Aware Similarity-Preserving Loss L_{FMSP}. To further analyse the effect of λ, we varied its value from 0 to 100

Table 5 Evaluation of various components of our method on SYSU-MM01 ("All-search" and "Indoor-search") and RegDB (Nguyen et al. 2017)

Method	All-search		Indoor-search		RegDB	
	mAP	r1	mAP	r1	mAP	r1
One-stream	38.39	36.91	50.46	40.38	48.29	47.60
One-stream + L_{FMSP}	43.18	41.34	56.27	46.49	54.10	53.54
Modality-Gated Extractor	43.03	41.37	55.34	45.53	52.65	53.88
Modality-Gated Extractor + L_{MSP}	43.23	41.55	55.13	45.06	59.91	61.29
Modality-Gated Extractor + L_{FMSP} (our full model)	44.98	43.56	57.50	48.62	64.50	65.07

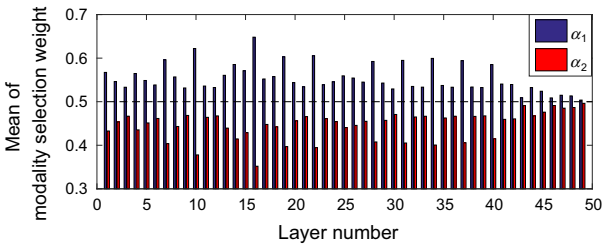
Here, r1, r5, and r10 denote the rank-1, rank-5, and rank-10 accuracies (%), respectively. See the text in the first paragraph of Sect. 6.5 for detailed definitions of the other notations

and evaluated the resulting performance of our method on SYSU-MM01. The rank-1 accuracies achieved in the indoor-search mode with different λ values are plotted in Fig. 13. Similar conclusions can be drawn from the results obtained in other settings. As λ was increased from 0 to 10.0, the performance improved; the best performance was achieved with $\lambda = 10.0$. Empirically, the best value of the parameter λ in our method is between 1.0 and 10.0.

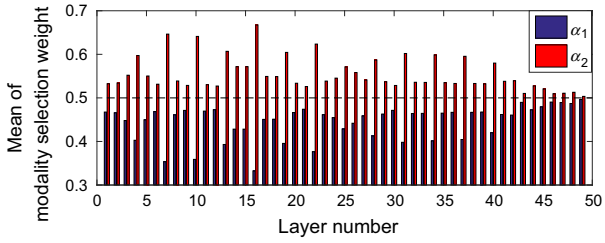
Comparison with Deep Zero Padding Our preliminary method based on deep zero padding was considered in the comparisons presented in Tables 3 and 4, with all hyperparameters being the same as in the method proposed in this study. Our proposed method clearly outperformed deep zero padding, in which fixed modality-gated nodes are used in the input layer to generate modality-specific nodes in the network and training is performed using the Softmax loss. Our Focal Modality-Aware Similarity-Preserving Loss can mine the shared knowledge for cross-modality matching, and our Modality-Gated Extractor is more flexible than deep zero padding for learning a network structure that is suited to the data. Note that the results of deep zero padding reported here are better than those reported in our preliminary work (Wu et al. 2017) because we used a larger neural network as the backbone and the amount of data used for training was increased. More specifically, we used data of all 395 identities in the training set in this work, while image data of only 296 people were used in Wu et al. (2017).

Visualisation of the Modality Selection Weights The modality selection weights a_1 and a_2 are important parameters in modality-gated nodes (Eq. 11). These parameters determine to what extent a node tends to be shared or specific to one modality, and in turn, they implicitly determine the modality-specific and shared structures in the Modality-Gated Extractor. When $a_1 = a_2 = 0.5$, the node is a shared node, as the sum of a_1 and a_2 is constrained by $a_1 + a_2 = 1$. By contrast, the farther a_1 or a_2 is from 0.5, the more the node tends to be specific to one modality. We regard nodes with modality selection weights such that $a_1 \geq a_2$ as soft modality-1-specific nodes and nodes with $a_1 < a_2$ as soft modality-2-specific nodes. We show the means of the modality selection weights for soft modality-1-specific nodes and soft modality-2-specific nodes and the proportions of these two types of nodes in all nodes of each layer of our Modality-Gated Extractor in Fig. 14. In Fig. 14, (a) shows the weights of soft modality-1-specific nodes, (b) shows the weights of soft modality-2-specific nodes and (c) shows the proportions of these two types of nodes in all nodes of each layer.

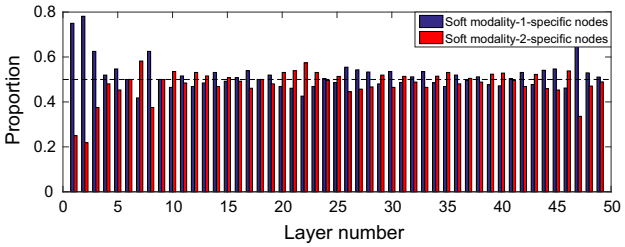
For cross-modality matching, intuitively, it can be assumed that more modality-specific structures will be needed in the shallower layers for reducing the modality gap, whereas more shared structures will be needed in the deeper layers for learning the shared representation for matching. The means of the modality selection weights in Fig. 14 are generally consistent with this assumption. The modality selection weights are closer to 0.5 in the deeper layers (after the 40th layer) and farther from 0.5 in the shallower layers, indicating that the nodes in deeper layers tend to be shared, while the nodes in shallower layers tend to be more specific to a particular modality. The proportion of soft modality-1-specific nodes or soft modality-2-specific nodes is close to 0.5 for most layers except for the first several layers. In the first several layers, there are more soft modality-1-specific (RGB-specific) nodes than soft modality-2-specific (IR-specific) nodes, and this is probably because the ImageNet-pretrained convolution filters for extracting low-level features such as edges and textures can be shared by both RGB images and IR images. As the filters are pretrained on RGB images, most of them tend to be more specific to RGB modality, although they are also partially shared for IR modality.



(a) Soft modality-1-specific nodes



(b) Soft modality-2-specific nodes



(c) Proportions of soft modality-1-specific/modality-2-specific nodes

Fig. 14 Means of the modality selection weights a_1 and a_2 for the modality-gated nodes (Eq. 11) and the proportions of two types of modality-gated nodes in all nodes of each layer in our Modality-Gated Extractor. **a** shows the weights of soft modality-1-specific nodes, **b** shows the weights of soft modality-2-specific nodes and **c** shows the proportions of these two types of nodes in all nodes of each layer. The x-axis shows the layer number. In **a** or **b**, the y-axis shows the mean of the corresponding modality selection weights a_1 or a_2. In **c**, the y-axis shows the proportions of soft modality-1-specific nodes and soft modality-2-specific nodes in all nodes of each layer. Please see the last paragraph of Sect. 6.5 for detailed analysis

7 Conclusion

In this work, we study the problem of matching images of persons captured under normal lighting with those acquired in very dark environments, where IR images are used instead of RGB images, e.g., person Re-ID across day and night conditions in 24-h surveillance systems. We make an early attempt to address the RGB-IR cross-modality Re-ID (RGB-IR Re-ID) problem and introduce a new multi-modality Re-ID dataset, SYSU-MM01. The significant disparities between RGB and IR image data make RGB-IR Re-ID a very challenging problem compared with conventional single-modality RGB-based Re-ID problem.

In contrast to previous cross-modality modelling methods, we do not operate under the assumption of identical distributions between training and testing data for Re-ID and cast mining shared knowledge for cross-modality matching as the problem of cross-modality similarity preservation. For this purpose, we propose a Focal Modality-Aware Similarity-Preserving Loss. Furthermore, to facilitate the extraction of the shared knowledge, we overcome the limitations of the manually designed modality-specific and shared model structures used in existing cross-modality matching methods by proposing the modality-gated node as a generalisation of both modality-specific and shared nodes; we use our proposed modality-gated nodes to construct a structure-learnable feature extraction model Modality-Gated Extractor. Extensive experiments on a new SYSU-MM01 benchmark dataset show the effectiveness of our method compared to a wide range of methods for person Re-ID, cross-modality matching and domain adaptation.

Acknowledgements This work was supported partially by the National Key Research and Development Program of China (2016YFB1001002), NSFC(U1911401, U1811461, U1611461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong Project (No. 2018B030312002), and Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03). The corresponding author and principal investigator for this paper is Wei-Shi Zheng.

Appendix

Properties of Modality-Specific and Shared Nodes In Sect. 4.1, we analyse the properties of modality-specific and shared nodes, which can be derived as follows.

Let $\mathbf{x}_{(l)}$ denote the input to layer $l+1$, let $o_{(l+1),i}$ denote the output of the i-th node before the activation function in layer $l+1$, and let $\mathbf{w}_{(l+1),i}$ and $b_{(l+1),i}$ denote the weight and bias parameters, respectively, i.e., $o_{(l+1),i} = (\mathbf{w}_{(l+1),i})^\top \mathbf{x}_{(l)} + b_{(l+1),i}$. Using the previously defined types of nodes, without loss of generality, $\mathbf{x}_{(l)}^{m1}$ and $\mathbf{x}_{(l)}^{m2}$ can be factorised into three parts as follows: $\mathbf{x}_{(l)}^{m1} = [\mathbf{x}_{(l)}^{m1,1spe}; \mathbf{x}_{(l)}^{m1,2spe}; \mathbf{x}_{(l)}^{m1,sh}]$ and $\mathbf{x}_{(l)}^{m2} = [\mathbf{x}_{(l)}^{m2,1spe}; \mathbf{x}_{(l)}^{m2,2spe}; \mathbf{x}_{(l)}^{m2,sh}]$, where ";" denotes vector concatenation and the three components correspond to modality-1-specific, modality-2-specific and shared nodes, respectively. We denote $\mathbf{w}_{(l+1),i}$ as $\mathbf{w}_{(l+1),i} = [\mathbf{w}_{(l+1),i}^{1spe}; \mathbf{w}_{(l+1),i}^{2spe}; \mathbf{w}_{(l+1),i}^{sh}]$.

Let L denote the loss function of the network. For modality-2-specific nodes, given a network input $\mathbf{x}_{(0)}^{m1}$ of modality 1, the output is $\mathbf{x}_{(l)}^{m1,2spe} = \mathbf{0}$; this can be derived in accordance with the definition of types of nodes in Eq. (10). Therefore, in the forward propagation process, the output of layer $l+1$ is

$$o_{(l+1),i} = (\mathbf{w}_{(l+1),i}^{1spe})^\top \mathbf{x}_{(l)}^{m1,1spe} + (\mathbf{w}_{(l+1),i}^{sh})^\top \mathbf{x}_{(l)}^{m1,sh} + b_{(l+1),i}. \tag{15}$$

In the backward propagation process, the derivatives of the loss function L with respect to the weights are

$$\frac{\partial L}{\partial \mathbf{w}_{(l+1),i}^{1spe}} = \frac{\partial L}{\partial o_{(l+1),i}} \frac{\partial o_{(l+1),i}}{\partial \mathbf{w}_{(l+1),i}^{1spe}} = \frac{\partial L}{\partial o_{(l+1),i}} \mathbf{x}_{(l)}^{m1,1spe}, \tag{16}$$

$$\frac{\partial L}{\partial \mathbf{w}_{(l+1),i}^{2spe}} = \frac{\partial L}{\partial o_{(l+1),i}} \frac{\partial o_{(l+1),i}}{\partial \mathbf{w}_{(l+1),i}^{2spe}} = \frac{\partial L}{\partial o_{(l+1),i}} \mathbf{x}_{(l)}^{m1,2spe} = \mathbf{0}, \tag{17}$$

$$\frac{\partial L}{\partial \mathbf{w}_{(l+1),i}^{sh}} = \frac{\partial L}{\partial o_{(l+1),i}} \frac{\partial o_{(l+1),i}}{\partial \mathbf{w}_{(l+1),i}^{sh}} = \frac{\partial L}{\partial o_{(l+1),i}} \mathbf{x}_{(l)}^{m1,sh}. \tag{18}$$

For a network input $\mathbf{x}_{(0)}^{m2}$ of modality 2, formulations similar to those above can be derived. Since the bias parameter $b_{(l+1),i}$ can be included in the weight parameter $\mathbf{w}_{(l+1),i}$ by simply padding 1 into the layer input $\mathbf{x}_{(l)}$, the bias parameter is not analysed individually.

Based on the derivations above, the properties of modality-specific and shared nodes are analysed in Sect. 4.1.

Structure Representation Ability In the last paragraph of Sect. 4.2, we analyse the connection between one-stream networks consisting of modality-gated nodes and the existing network structures. To show the structure representation ability of modality-gated nodes, we take a simple two-stream network as an example. In Fig. 9, a one-stream network equivalent to the two-stream network in Fig. 7 with respect to forward propagation is shown. Inputs $\mathbf{x}^{m1}$ and $\mathbf{x}^{m2}$ are fed into the same nodes in layer 0. In layer 1, there are four modality-gated nodes, of which two are modality-1-specific nodes and the others are modality-2-specific nodes. In layer 1a, there are two shared modality-gated nodes. The black solid lines denote weights of 1, and the black dotted lines denote weights

of 0. The modality label y^{mod} controls the modality gate. In this way, the weights represented in blue and red correspond to modality-specific nodes, and the weights represented in green correspond to shared nodes; thus, this one-stream structure consisting of modality-gated nodes is identical to the two-stream structure in Fig. 7. Therefore, it is possible to learn a two-stream structure. More generally, modality-gated nodes can represent soft modality-specific nodes and provide a network with sufficient flexibility to evolve into any modality-specific and shared structures.

Modality-Gated Nodes versus Channel Attention For our modality-gated nodes, each node (or channel in a convolutional network) is weighted by two modality selection weights, one for each modality. Some neural networks apply a channel attention mechanism (e.g., SENet Hu et al. 2018) and also use different weights for different channels. According to our analysis, the modality selection weights of modality-gated nodes can be adjusted to form either modality-specific or shared structures in a network for processing data from two modalities. By contrast, simply applying the channel attention mechanism does not allow modality-specific and shared structures to be learned because there is no specific modelling for different modalities.

References

Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3908–3916).

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning (ICML)* (pp. 1247–1255).

Bak, S., & Carr, P. (2017). One-shot metric learning for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1571–1580).

Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Learning aligned cross-modal representations from weakly aligned data. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2940–2949).

Chen, D., Xu, D., Li, H., Sebe, N., & Wang, X. (2018). Group consistent similarity learning via deep CRF for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 8649–8658).

Chen, D., Yuan, Z., Hua, G., Zheng, N., & Wang, J. (2015). Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1565–1573).

Chen, J., Wang, Y., Qin, J., Liu, L., & Shao, L. (2017). Fast person re-identification via cross-camera semantic binary transformation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5330–5339).

Chen, Y. C., Zheng, W. S., & Lai, J. (2015). Mirror representation for modeling view-specific transform in person re-identification. In *International joint conferences on artificial intelligence (IJCAI)*.

Chen, Y. C., Zheng, W. S., Lai, J. H., & Yuen, P. (2017). An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, *27*(8), 1661–1675.

Chen, Y. C., Zhu, X., Zheng, W. S., & Lai, J. H. (2018). Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *40*(2), 392–408.

Dai, P., Ji, R., Wang, H., Wu, Q., & Huang, Y. (2018). Cross-modality person re-identification with generative adversarial training. In *International joint conferences on artificial intelligence (IJCAI)*.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 886–893).

Dong, S. C., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. (2011). Custom pictorial structures for re-identification. In *British machine vision conference (BMVC)* (pp. 68.1–68.11).

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2360–2367).

Feng, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *ACM international conference on multimedia (ICM)* (pp. 7–16).

Gray, D., Brennan, S., & Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, *3*, 1–7.

Gray, D., & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision (ECCV)* (pp. 262–275).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, *13*, 723–773.

Guo, C. C., Chen, S. Z., Lai, J. H., Hu, X. J., & Shi, S. C. (2014). Multi-shot person re-identification with automatic ambiguity inference and removal. In *International conference on pattern recognition (ICPR)* (pp. 3540–3545).

Haque, A., Alahi, A., & Li, F. F. (2016). Recurrent attention models for depth-based person identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1229–1238).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

He, R., Wu, X., Sun, Z., & Tan, T. (2017). Learning invariant deep representation for NIR-VIS face recognition. In *Association for Advancement of Artificial Intelligence (AAAI)*.

He, R., Wu, X., Sun, Z., & Tan, T. N. (2019). Wasserstein CNN: Learning invariant features for NIR-VIS face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *41*(7), 1761–1773.

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on image analysis* (pp. 91–102).

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2018). Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7132–7141).

Jing, X. Y., Zhu, X., Wu, F., You, X., Liu, Q., Yue, D., Hu, R., & Xu, B. (2015). Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 695–704).

Jungling, K., & Arens, M. (2010). Local feature based person reidentification in infrared image sequences. In *IEEE international conference on advanced video and signal-based surveillance (AVSS)* (pp. 448–455).

Kan, M., Shan, S., Chen, X. (2016). Multi-view deep network for cross-view classification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4847–4855).

Karanam, S., Li, Y., & Radke, R. J. (2015). Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *IEEE international conference on computer vision (ICCV)* (pp. 4516–4524).

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Person re-identification by unsupervised $\ell 1$ graph learning. In *European conference on computer vision (ECCV)* (pp. 178–195).

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2288–2295).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Annual conference on neural information processing systems (NeurIPS)* (pp. 1097–1105).

Kviatkovsky, I., Adam, A., & Rivlin, E. (2013). Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(7), 1622–1634.

Lei, Z., & Li, S. Z. (2009). Coupled spectral regression for matching heterogeneous faces. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1123–1128).

Li, M., Zhu, X., & Gong, S. (2018). Unsupervised person re-identification by deep learning tracklet association. In *European conference on computer vision (ECCV)* (pp. 737–753).

Li, W., Zhao, R., & Wang, X. (2012). Human reidentification with transferred metric learning. In *Asian confererence on computer vision (ACCV)* (pp. 31–44).

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 152–159).

Li, X., Zheng, W. S., Wang, X., Xiang, T., & Gong, S. (2015). Multi-scale learning for low-resolution person re-identification. In *IEEE international conference on computer vision (ICCV)* (pp. 3765–3773).

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., & Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3610–3617).

Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2197–2206).

Liao, S., & Li, S. Z. (2015). Efficient PSD constrained asymmetric metric learning for person re-identification. In IEEE international conference on computer vision (ICCV) (pp. 3685–3693).

Lin, D., & Tang, X. (2006). Inter-modality face recognition. In *European conference on computer vision (ECCV)* (pp. 13–26).

Lin, L., Wang, G., Zuo, W., Xiangchu, F., & Zhang, L. (2017). Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *39*(6), 1089–1102.

Lin, Z., Ding, G., Hu, M., & Wang, J. (2015). Semantics-preserving hashing for cross-view retrieval. In IEEE conference on computer vision and pattern recognition (CVPR) (pp. 3864–3872).

Lisanti, G., Masi, I., Bagdanov, A. D., & Bimbo, A. D. (2015). Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *37*(8), 1629–1642.

Liu, C., Gong, S., Loy, C.C., & Lin, X. (2012). Person re-identification: What features are important? In European conference on computer vision workshop (ECCVW) (pp. 391–401).

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning (ICML)* (pp. 97–105).

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, *9*, 2579–2605.

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). Hierarchical Gaussian descriptor for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1363–1372).

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

Paisitkriangkrai, S., Shen, C., & Hengel, A.V.D. (2015). Learning to rank in person re-identification with metric ensembles. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1846–1855).

Pedagadi, S., Orwell, J., Velastin, S., & Boghossian, B. (2013). Local Fisher discriminant analysis for pedestrian re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3318–3325).

Prosser, B., Zheng, W. S., Gong, S., Xiang, T., & Mary, Q. (2010). Person re-identification by support vector ranking. In B*ritish machine vision conference (BMVC)* (pp. 21.1–21.11).

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *ACM multimedia (ACMMM)* (pp. 251–260).

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision workshop (ECCVW)* (pp. 17–35).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 815–823).

Sharma, A., Kumar, A., Daume, H., & Jacobs, D. W. (2012). Generalized multiview analysis: A discriminative latent space. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2160–2167).

Shi, Z., Hospedales, T. M., & Xiang, T. (2015). Transferring a semantic representation for person re-identification and search. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4184–4193).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR)*.

Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1179–1188).

Su, C., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *European conference on computer vision (ECCV)* (pp. 475–491).

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision workshop (ECCVW)* (pp. 443–450).

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling. In *European conference on computer vision (ECCV)* (pp. 480–496).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–9).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 7167–7176).

Wang, W., Yang, X., Ooi, B. C., Zhang, D., & Zhuang, Y. (2016). Effective deep learning-based multi-modal retrieval. *The International Journal on Very Large Data Bases (VLDB)*, *25*(1), 79–101.

Wang, X., Zheng, W. S., Li, X., & Zhang, J. (2016). Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, *26*(8), 1447–1460.

Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer gan to bridge domain gap for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 79–88).

Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., et al. (2017). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics*, *47*(2), 449–460.

Wu, A., Zheng, W. S., & Lai, J. H. (2017). Robust depth-based person re-identification. *IEEE Transactions on Image Processing (TIP)*, *26*(6), 2588–2603.

Wu, A., Zheng, W. S., Yu, H. X., Gong, S., & Lai, J. (2017).RGB-infrared cross-modality person re-identification. In *IEEE international conference on computer vision (ICCV)* (pp. 5390–5399).

Wu, B., Yang, Q., Zheng, W. S., Wang, Y., & Wang, J. (2015). Quantized correlation hashing for fast cross-modal search. In *International joint conferences on artificial intelligence (IJCAI)*.

Wu, S., Chen, Y. C., Li, X., Wu, A., You, J., & Zheng, W. S. (2016). An enhanced deep feature representation for person re-identification. In *IEEE winter conference on applications of computer vision (WACV)* (pp. 1–8).

Wu, Z., Li, Y., & Radke, R. J. (2015). Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *37*(5), 1095–1108.

Xiao, T., Li, H., Ouyang, W., & Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In I*EEE conference on computer vision and pattern recognition (CVPR)* (pp. 1249–1258).

Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2017). Joint detection and identification feature learning for person search. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3376–3385).

Xiong, F., Gou, M., Camps, O., & Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *European conference on computer vision (ECCV)* (pp. 1–16).

Yang, Q., Wu, A., & Zheng, W. S. (2019). Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*,. https://doi.org/10.1109/TPAMI.2019.2960509.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., & Li, S. Z. (2014). Salient color names for person re-identification. In *European conference on computer vision (ECCV)* (pp. 536–551).

Ye, M., Lan, X., Li, J., & Yuen, P. C. (2018). Hierarchical discriminative learning for visible thermal person re-identification. In *Association for Advancement of Artificial Intelligence (AAAI)*.

Ye, M., Wang, Z., Lan, X., & Yuen, P. C. (2018). Visible thermal person re-identification via dual-constrained top-ranking. In *International joint conferences on artificial intelligence (IJCAI)*.

Yin, J., Wu, A., & Zheng, W. S. (2020). Fine-grained person re-identification. *International Journal of Computer Vision (IJCV)*. https://doi.org/10.1007/s11263-019-01259-0.

You, J., Wu, A., Li, X., & Zheng, W. S. (2016). Top-push video-based person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1345–1353).

Yu, H. X., Wu, A., & Zheng, W. S. (2018). Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*,. https://doi.org/10.1109/TPAMI.2018.2886878.

Yu, H. X., Wu, A., & Zheng, W. S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE international conference on computer vision (ICCV)* (pp. 994–1002).

Zhang, D., & Li, W. J. (2014). Large-scale supervised multimodal hashing with semantic correlation maximization. In *Association for Advancement of Artificial Intelligence (AAAI)*.

Zhang, L., Xiang, T., & Gong, S. (2016). Learning a discriminative null space for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1239–1248).

Zhao, L., Li, X., Zhuang, Y., & Wang, J. (2017). Deeply-learned part-aligned representations for person re-identification. In *IEEE international conference on computer vision (ICCV)* (pp. 3239–3248).

Zhao, R., Oyang, W., & Wang, X. (2017). Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *39*(2), 356–370.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In European conference on computer vision (ECCV) (pp. 868–884).

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *IEEE international conference on computer vision (ICCV)* (pp. 1116–1124).

Zheng, W. S., Gong, S., & Xiang, T. (2009). Associating groups of people. In *British machine vision conference (BMVC)* (pp. 23.1–23.11).

Zheng, W. S., Gong, S., & Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(3), 653–668.

Zheng, W. S., Gong, S., & Xiang, T. (2016). Towards open-world person re-identification by one-shot group-based verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *38*(3), 591–606.

Zheng, W. S., Li, X., Xiang, T., Liao, S., Lai, J., & Gong, S. (2015). Partial person re-identification. In *IEEE international conference on computer vision (ICCV)* (pp. 4678–4686).

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE international conference on computer vision (ICCV)* (pp. 3774–3782).

Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. In *IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 3652–3661).

Zhu, F., Shao, L., & Yu, M. (2014). Cross-modality submodular dictionary learning for information retrieval. In *ACM international conference on information and knowledge management (CIKM)* (pp. 1479–1488).

Zhu, J. Y., Zheng, W. S., Lai, J. H., & Li, S. Z. (2014). Matching nir face to vis face using transduction. *IEEE Transactions on Information Forensics and Security (TIFS)*, *9*(3), 501–514.

Zhu, X., Wu, B., Huang, D., & Zheng, W. S. (2017). Fast open-world person re-identification. *IEEE Transactions on Image Processing (TIP)*, *27*(5), 2286–2300.