Top-Push Constrained Modality-Adaptive Dictionary Learning for Cross-Modality Person Re-Identification

Peng Zhang, Jingsong Xu, Qiang Wu, *Senior Member, IEEE*,
Yan Huang and Jian Zhang, *Senior Member, IEEE*

***Abstract*—Person re-identification aims to match person captured by multiple non-overlapping cameras that mainly mean standard RGB cameras. In contemporary surveillance, cameras of different modalities such as infrared cameras and depth cameras are introduced because of their unique advantages in poor illumination scenarios. However, re-identifying the persons across such cameras of different modalities is extremely difficult and, unfortunately, seldom discussed. It is mainly caused by extremely different appearances of the person shown under such different camera modalities. In this paper, we tackle this challenging cross-modality people re-identification through a top-push constrained modality-adaptive dictionary learning. The proposed model asymmetrically projects the heterogeneous features from dissimilar modalities onto a common space. In this way, the modality-specific bias is mitigated. Thus, the heterogeneous data can be simultaneously enforced by a shared dictionary in a canonical space. Moreover, a top-push ranking graph regularization is embedded in the proposed model to improve the discriminability, which efficiently further boosts the matching accuracy. In order to implement the proposed model, an iterative process is developed in this paper to optimize these two processes jointly. Extensive experiments on the benchmark SYSU-MM01 and BIWI RGBD-ID person re-identification datasets show promising results which outperform state-of-the-art methods.**

***Index Terms*—Cross-modality person re-identification, data bias, asymmetric mapping, top-push constrained dictionary learning, domain adaptation.**

I. Introduction

PERSON re-identification (Re-ID) is a challenging retrieval problem in the video surveillance area, which aims to associate the person of interest across multiple disjoint cameras. It has drawn extensive research interests because of its promising potentials in security surveillance, *e.g.*, searching and tracking suspicious persons for criminal investigation. Different approaches from the perspective of either feature representation or metric learning [1]–[4] have been developed to tackle the issue. These approaches mainly depend on the fundamental assumption that person images are collected in the daytime by RGB cameras deployed in disjoint regions since RGB cameras are cheap and informative. Though these images are collected from different cameras, they approximately obey the same distribution because these cameras have the same attribute. Benefiting from the robust representative power, the data differences caused by cameras are merely considered as heterogeneous in previous works.

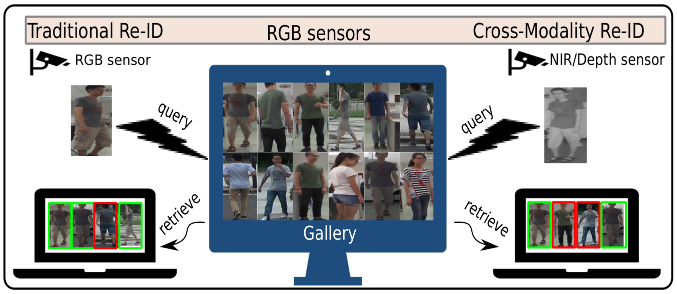


Fig. 1. An example of traditional person Re-ID versus cross-modality person Re-ID.

However, people are hard to be captured by RGB cameras in poor lighting scenarios, *e.g.*, night or cloudy. In these cases, from the perspective of applications, alternative sensors whose image-forming principle is invariant to visible light are necessary such as near-infrared (NIR) camera [5] and RGB-D camera [6], [7] as shown in Fig. 1. These different sensors take their advantages to sense images in both the day time (good lighting) and the night time (poor lighting). However, this case raises another problem of how we can identify the persons with such images taken by different types of sensors. We term the case that re-identifying person across different types of sensors as cross-modality person Re-ID. Due to the differences of image-forming principle between different types of sensors, people's appearance between RGB and NIR/RGB-D cameras are heterogeneous. This violates the prior assumption that images from different cameras obey the same distribution. Compared to the traditional Re-ID, heterogeneous/cross-modality person Re-ID suffers larger data biases, which cannot be generalized by treating images across cameras equally as previous methods. Along with the existing problem of cross camera-view matching in traditional person Re-ID, cross-camera-modality creates another layer of difficulty to Re-ID.

In this paper, we alleviate the data biases between modalities and improve the representative power of feature via asymmetric feature learning and discriminative dictionary learning.

Manuscript submitted November 18, 2018

Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang and Jian Zhang are with the Global Big Data and Technologies Centre, School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: Peng.Zhang-2@student.edu.au, Jingsong.Xu@uts.edu.au, Qiang.Wu@uts.edu.au, Jian.Zhang@uts.edu.au, Yan.Huang-3@student.uts.edu.au)

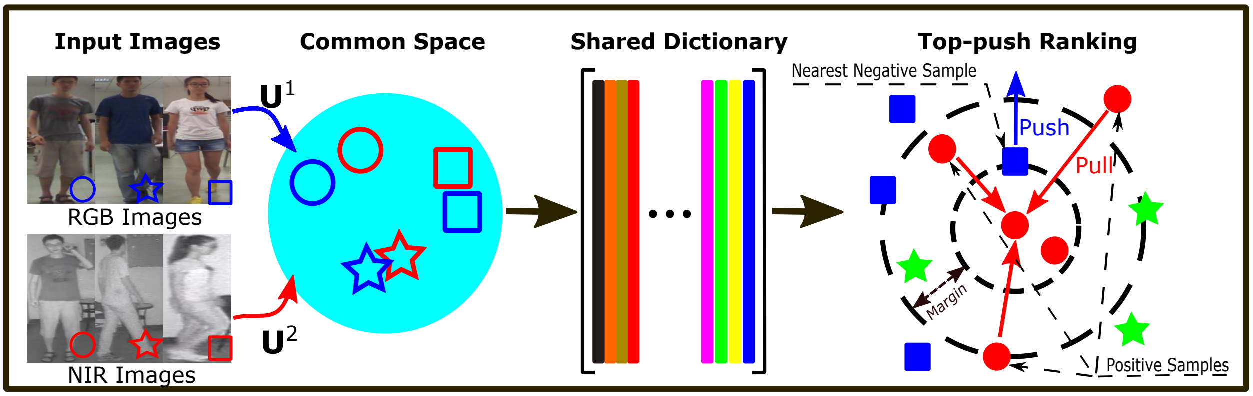


Fig. 2. Illustration of the proposed TCMDL model. Objects with the same shape (hallow/solid) represent the same person. Input images from cross modalities are mapped into a subspace in which a shared dictionary is learned. Meanwhile, the encoding coefficients are regularized by a top-push ranking constraint embedded Laplacian-like graph. In the figure, we take cross-modality person Re-ID using images from RGB and NIR sensors as an example. However, it can be extended to other cross-modality cases.

Our motivations are three-fold. Firstly, the coupled metric learning [8]–[11] is able to bridge gaps between heterogeneous data. It motivates us to learn a pair of asymmetric mapping matrices to project original feature representation from different modalities into a shared subspace. Since samples from each modality are transformed to the shared subspace by an independent matrix, *i.e.*, asymmetrically, data heterogeneity across modalities is thus mitigated in the subspace as shown in Fig. 3. This is essential for distance measurement across modalities. Secondly, the success of discriminative dictionary learning [4] inspires us to impose Laplacian-like graph regularization to perform retrieval task with a ranking formulation. Finally, triplet constraint [12] is usually utilised in the classical ranking scheme. However, inter-person feature differences are more ambiguous in the cross-modality setting. Thus more stringent regularization is necessary. Inspired by top-push ranking [3] which forces intra-class difference to be smaller than the minimum inter-class difference, we reformulate the top-push ranking to a Laplacian-like graph and integrate it to a unified objective. Benefiting of the above three aspects, the unified objective can simultaneously mitigate the data biases across modalities and keep the powerful feature representation ability of dictionary as well as the discriminative ability of top-push constraint.

Based on the above motivations, we propose a Top-push Constrained Modality-adaptive Dictionary Learning (TCMDL) model for cross-modality person Re-ID as illustrated in Fig. 2. This model simultaneously learns the latent subspace and discriminative dictionary for cross-modality retrieval problem. In detail, our model consists of four parts. One is asymmetric feature and dictionary learning, which jointly map the heterogeneous data into a common subspace and learn a shared dictionary for the heterogeneous data. Since data biases are alleviated by asymmetric feature learning, the projected features can be represented by a shared dictionary. Different from cross-view dictionary learning that learns two distinct dictionaries, we represent the same person across modalities with a shared dictionary in the common learned subspace. The rest three parts are regularization terms, aiming to avoid information loss while performing feature mapping, keep consistent information of the same person across modalities and preserve discriminative ability, respectively. Especially, we reformulate top-push distance learning model into a Laplacian-like graph and impose it to the coding coefficients through dictionary learning. It is critical for cross-modality person Re-ID to differentiate minor variations. As far as we know, this is the first work to integrate asymmetric feature mapping and discriminative dictionary learning into a uniform framework and achieve these two purposes simultaneously to solve the cross-modality person Re-ID problem.

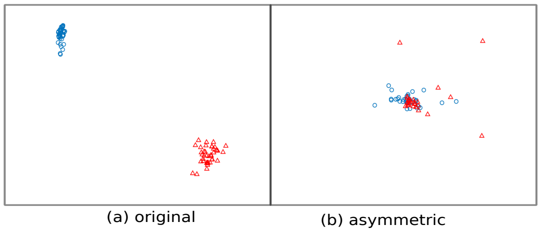


Fig. 3. Illustration of how our asymmetric mapping bridges the data gaps. We performed PCA on samples from BIWI RGBD-ID dataset [13] for visualization. Each shape (circle or triangle) represents samples from one modality. (a) original data distribution, (b) distribution in the shared space learned by asymmetric mapping.

In summary, the contributions of this paper are

- We propose to join asymmetric feature mapping and discriminative dictionary learning in a unified scheme for heterogeneous person Re-ID. It alleviates data biases across modalities in the projected subspace, and thus heterogeneous data can be represented by a shared discriminative dictionary.
- Moreover, a top-push ranking constraint is reformulated and integrated into the unified model, which makes the dictionary learning more effective to person Re-ID.

II. Related Works

In this section, we review literature in two fields which are most related to our work: person Re-ID including single-modality and cross-modality scenarios, and dictionary learning for cross-modality matching.

A. Person Re-ID

1) Single-Modality Person Re-ID: Previous approaches for person Re-ID mainly target on solving the single-modality matching problems where images are collected by visual light cameras. These methods can be categorized into three classes. One is to craft or learn effective feature representations empirically by characterizing colour, texture [1], gradient [14], attribute [15] and spatial-temporal information [3], etc. Another one is to learn discriminative metrics to measure the similarity between image pairs collected from distinct cameras [1], [3]. In addition to the above methods, deep neural network based approaches learn effective feature representations which achieve promising results on many benchmarks [16]–[20]. These studies promote the development of person Re-ID using visual RGB images. Rather than using visual light cameras, RGB-D sensors [6], [7], [21] are also utilized that target on tackling Re-ID under variants such as clothing and illumination. These approaches usually extract hand-crafted features from depth images and compute the matching score between them. Since the benefits of depth sensors, these methods gain great success on long-term person Re-ID [22] in scenarios where RGB features are failed.

Though RGB-D images captured by Kinect are adopted in early person Re-ID studies, they are treated as an alternative of RGB in the scenarios of cloth changing and poor light condition, and never used to query person using depth images while the gallery is RGB images. This means that most of the previous works using no matter visual-light cameras or depth sensors only focus on person Re-ID in a single modality. And, these models cannot sort out person Re-ID across different camera modalities such as matching person shot by RGB camera v.s. NIR/Depth camera. It is because models for single-modality person Re-ID heavily rely on the information in one modality only, and there is no mechanism in the existing single-model Re-ID methods to tackle the data gaps caused by the modality change.

2) Cross-Modality Person Re-ID: Recently, person Re-ID across different modalities is proposed because of the need to match different types of person images collected in the day by visual-light cameras and in the night by NIR or Depth sensors. The first work on the problem was published in 2017 by Wu *et al.* [5]. They first discussed the cross-modality person Re-ID using RGB and NIR images, which learned domain-specific and domain-shared feature via a one-stream network by padding zeros to images. Though the work achieves promising results, data gaps still exist due to the hard threshold used to determine whether a node is domain-specific. In 2018, Ye *et al.* [23] proposed to map modality-specific features from two modalities into a consistent space and learn modality-shared features using a two-stream network (TONE), which improves performances than only using TONE. Furthermore, Ye *et al.* [24] proposed a novel bi-directional dual-constrained top-ranking loss to optimize the two-stream network, which further improves the performance. Though both works achieve promising performances, they rely on the two-stream CNN network. As we know, training CNN is quite time-consuming. For example, it takes approximately 10.9 hours to fine tune TONE [23] with AlexNet as the backbone on one 16G Quadro GPU. And, it will take over 3.7 times longer time if GPU is unavailable. The longer training time really impedes the system development and update progress [25], [26], especially for the scenarios where the model requires frequent redeployment for end-users or adjustment for vendors. Though the efficiency of CNN training has been improved a lot recently, heavy GPU expense is still a barrier. For example, Mikami *et al.* [25] significantly reduce the time on training ResNet-50 from 29 hours to 224s. However, they use 2176 Tesla V100 GPUs. This will be a concern to many cases where heavy computing resources are not available. In addition, deep learning-based methods typically rely on large-scale training samples. However, it is practically difficult for the case of cross-modality person Re-ID, where it is hard to guarantee sufficient training data on both modalities. In fact, the scale of available datasets for cross-modality person re-identification is relatively small at present. This is mainly because of practical challenges to capture the same person appearing in different types of cameras/sensors. Moreover, they are easy to be over-fitting, especially on small-scale datasets such as BIWI RGBD-ID dataset due to the limitation of labelled samples.

Different from above, the proposed method targets on the scenarios where light-weight computing power is essential and heavy GPU based server is not available. We propose a light-weight model which mitigates date bias by jointly optimizing asymmetric mapping and discriminative shared-dictionary learning in an explicit way. The proposed model projects person images from different modalities into a common latent subspace by asymmetric mapping and reconstructs the mapped features in the subspace using a shared dictionary. The design explicitly mitigates data biases across different modalities. Moreover, we adapt top-push ranking as regularization that makes the learned dictionary discriminative and further improves the performance of cross-modality person Re-ID.

B. Dictionary Learning for Cross-Modality Matching

Benefiting of great expressive ability, dictionary learning and its variations have been applied to many fields during the last decades. Among them, several works attempt to achieve domain adaptation for the cross-modality matching task. Shekhar *et al.* [27] proposed to jointly map heterogeneous data into a common space and represented data using a shared dictionary in a common space for object detection. To preserve discriminative ability, they regularized the dictionary rather than encoding coefficients, which are significantly different from us. More recently, Liu *et al.* [28] proposed semi-supervised coupled dictionary learning for Re-ID that learns two separated dictionaries to encode images from different domains to address alignment problem. Inspired by [28], [29] proposed a cross-view projective dictionary learning method for Re-ID, which also learned two distinct dictionaries for each camera view. However, their model should be supervised by using paired samples across views, and paired dictionaries should be learned which highly rely on the expressive of the paired dictionaries. Peng *et al.* [30] proposed to conduct transfer learning to achieve cross-dataset person re-identification. In 2017, Zhou *et al.* [31] proposed a joint model for person Re-ID that performs dictionary and metric learning simultaneously. However, they targeted the traditional person Re-ID and treated images from different cameras equally.

Motivated by the current works, we also attempt to reconstruct data across modalities by building a dictionary. To make models discriminative and better correlate data across modalities, the existing methods either regularize the shared dictionary [27] or adopt two dictionaries corresponding to two modalities, and then link them with coefficients [28], [29]. Different from existing methods, our method links the data across modalities through asymmetric mapping. To make it supervised, the proposed method simultaneously regularizes the learned coefficients through top-push constrained Laplacian graph using a single shared dictionary. This joint design improves discriminability for cross-modality person Re-ID.

III. The Proposed Approach

In this section, we formulate the proposed TCMDL model for cross-modality person Re-ID. The proposed TCMDL involves four parts: joint asymmetric mapping and dictionary learning, energy-preserving regularization, cross-view consistency regularization and top-push constrained Laplacian graph regularization. It learns a shared discriminative dictionary in a common subspace by joint asymmetric feature mapping and top-push constraint regularized dictionary learning. Optimization and complexity analysis of the model are then presented.

A. Overview

The aim of cross-modality person Re-ID is to retrieve the person of interest from volumes of gallery images captured by a series of disjoint cameras in the surveillance network. Currently, most algorithms are developed for the two-camera single-modality setting, in which both cameras are based on visible light. In contrast, we consider the cross-modality person Re-ID and develop the TCMDL model in a general multi-modality way not only for the cross-modality but also for multi-modality scenarios.

Fig. 3 gives the pipeline of the proposed TCMDL for cross-modality person Re-ID. In the figure, two modalities, images captured by RGB cameras and NIR sensors are taken as an example. Here, different modalities refer to different styles of person images taken by different sensors such as RGB cameras, NIR sensors and depth sensors. In Fig. 3, features extracted from RGB images and NIR images are simultaneously input to the model and mapped into a common subspace by a pair of asymmetric mapping matrices in which a shared dictionary is learned to reconstruct features from both modalities. Due to data gaps are mitigated when performing asymmetric mapping as shown in Fig. 2, the features from two domains (modalities) can be represented using the shared dictionary in the common subspace. To keep discriminability, we impose a top-ranking regularization on the encoding coefficients with respect to features from each domain. This regularization term ensures that the distance between positive samples is smaller than any pair of their corresponding negative sample.

B. Objectives

Without loss of generality, we denote training sets collected from P modalities across disjoint cameras as $\mathbf{X}^p = [\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_{N_p}^p] \in \mathbf{R}^{n_p \times N_p} (p = 1, \ldots, P)$ respectively, where each $\{\mathbf{x}_i^p; l_i^p\}$ $(i = 1, \ldots, N_p)$ corresponds to a n_p-dimensional feature of the i-th image from the p-th camera and l_i^p is its class label.

1) Joint Asymmetric Mapping and Dictionary Learning (AsyDic): Due to large data gaps between P cameras, we wish to learn a set of mapping matrices $\mathbf{U}^p \in \mathbf{R}^{n \times n_p}, n \leq min(n_1, n_2, \ldots, n_p)$ to project the heterogeneous features in terms of each sensor modality into a common low-dimensional subspace in which dictionary learning can be performed by learning a shared K-atom dictionary $\mathbf{D} \in \mathbf{R}^{n \times K}$. Thus, the dictionary learning in the learned subspace is

$$\mathcal{J}_1(\mathbf{U}^p, \mathbf{D}, \mathbf{A}^p) = \sum_{p=1}^{P} (\|\mathbf{U}^p \phi(\mathbf{X}^p) - \mathbf{D}\mathbf{A}^p\|_F^2 + \alpha \|\mathbf{A}^p\|_F^2) \quad \text{s.t.} \quad \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K \tag{1}$$

where $\phi(\cdot)$ denotes the feature representation function, which can be hand-crafted feature or data-driven feature learned by deep neural networks, $\mathbf{A}^p \in \mathbf{R}^{K \times N_p}$ is the encoding coefficients spanned over the shared dictionary $\mathbf{D}$, $\mathbf{d}_i$ is the i-th column of $\mathbf{D}$, $\|\cdot\|_F$ denotes Frobenious norm of a matrix and α is the trade-off parameter. It is worth noting that we regularize the coefficient $\mathbf{a}_i^p \in \mathbf{R}^K$ corresponding to each sample $\mathbf{x}_i^p$ with l_2-norm rather than l_1-norm. This is because less sparsity benefits identification and improves computation efficiency as described in [32].

2) Energy-preserving Regularization: To avoid information loss of original signals, it is fashionable to impose an energy-preserving regularization [27], [33], defined as

$$\mathcal{J}_2(\mathbf{U}^p) = \sum_{p=1}^{P} \|\mathbf{U}^{p\top} \mathbf{U}^p \phi(\mathbf{X}^p) - \phi(\mathbf{X}^p)\|_F^2 \tag{2}$$

where superscript $\top$ denotes matrix transpose operation.

3) Cross-view Consistency Regularization: Intuitively, the learned mapping matrices $\mathbf{U}^p (p = 1, \ldots P)$ are arbitrarily inconsistent due to distinct data distribution, which focus more on alleviating data gap between cameras while sacrificing discriminativeness. This is inconsistent to our expectation since images of the same person from different cameras are inherently correlated. As in [11], we add another regularization term to keep the cross-view consistency, given as

$$\mathcal{J}_3(\mathbf{U}^p) = \sum_{i \neq j} \|\mathbf{U}^i - \mathbf{U}^j\|_F^2 \tag{3}$$

It is worth to point out that the term is specifically designed to the case when data from two modalities can be represented using the same feature descriptor, *i.e.*, the person images across two domains do not vary too much. This term requires dimensional consistency between mapping matrices. However, the term can be omitted when applied to other domain independent cross-modality matching cases or tasks such as RGB-depth person Re-ID, text-image retrieval and cross-biometric recognition.

4) Top-push Constrained Laplacian Graph Regularization (Top-push): To make the learned dictionary discriminative, Laplacian graph regularization is usually imposed to minimize the difference between coefficient vectors of samples from the same class and maximize the difference between coefficient vectors of visually similar samples from different classes, *e.g.*, [33] and [34]. In person Re-ID, we also hope the learned dictionary be discriminative that the distance between samples from different people should be larger than that of the same person by a margin of ρ. Different from previous works, we consider a top-push ranking metric embedded graph regularization inspired by the success of ranking matching in Re-ID. Since the coefficient vectors of samples from different modalities are learned in a common subspace spanned on a shared dictionary, $\mathbf{A}^p (p = 1 \dots P)$ can be treated equally. Thus, we define $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^P]$ and its class labels l_i corresponding to the i-th sample $\mathbf{a}_i$ in $\mathbf{A}$. Following the principle, ranking methods minimizing the hinge loss of triplets achieve significant success [12]. Compared to triplet loss, the top-push constraint enhances top-rank matching, which only considers the relationship between the distance of the positive pair and minimum distance of its related negative pairs, given as

$$\min \sum_{l_i = l_j} \max\{\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - \min_{l_i \neq l_k} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) + \rho, 0\} \quad (4)$$

where $\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) = (\mathbf{a}_i - \mathbf{a}_j)^\top \mathbf{W}^\top \mathbf{W}(\mathbf{a}_i - \mathbf{a}_j)$ is the squared Mahalanobis distance, which indicates distance of a positive pair is closer than its corresponding negative pairs.

It is reasonable to reformulate the top-push constraint into a Laplacian graph representation by ignoring some constant terms, scaling coefficients and imposing a regularization term on $\mathbf{W}$ inspired by [4], denoting as

$$\mathcal{J}_4(\mathbf{A}, \mathbf{W}) = trace(\mathbf{WALA}^\top \mathbf{W}^\top) + \gamma \|\mathbf{W}\|_F^2 \quad (5)$$

where $\gamma (\gamma \geq 0)$ is the trade-off parameter, $trace(\cdot)$ represents trace of a matrix, $\mathbf{L}$ is the Laplacian matrix, defined as $\mathbf{L} = \mathbf{G} - (\mathbf{S} + \mathbf{S}^\top)/2$, $\mathbf{G}$ is a diagonal matrix whose i-th diagonal element is $g_{ii} = \sum_{j=1, j \neq i} \frac{s_{ij} + s_{ji}}{2}$, and s_{ij} is the entry of the weight matrix $\mathbf{S}$ of graph edges which denotes the similarity between the adjacent pairwise vertices $(\mathbf{a}_i, \mathbf{a}_j)$, defining as

$$s_{ij} = \begin{cases} \varepsilon \left[\mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) - \underset{\substack{k \in [1, N], \\ l_i \neq l_k}}{min} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_j, \mathbf{a}_k) + \rho \right]_{l_i = l_j} & , \ i \neq j, \\ -\varepsilon \left[\underset{\substack{k \in [1, N], \\ l_i = l_k}}{max} \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_k) - \mathcal{D}_{\mathbf{W}}(\mathbf{a}_i, \mathbf{a}_j) + \rho \right]_{l_i \neq l_j} & , \ i \neq j, \\ 0, \ i = j, \end{cases} \quad (6)$$

where $\varepsilon[\cdot]$ is an indicator function whose value is zero for negative argument, and one otherwise, $N = \sum_{p=1}^{P} N_p$.

To take the benefits of Eq. 1, 2, 3 and 5 that simultaneously complete asymmetric mapping and discriminative dictionary learning, our overall optimization objective is

$$\begin{aligned} \min_{\mathbf{U}^p, \mathbf{D}, \mathbf{A}^p, \mathbf{W}} \quad & \mathcal{J}_1 + \lambda_1 \mathcal{J}_2 + \lambda_2 \mathcal{J}_3 + \lambda_3 \mathcal{J}_4 \\ \text{s.t.} \quad & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \dots, K; \lambda_1, \lambda_2, \lambda_3, \gamma \geq 0 \end{aligned} \quad (7)$$

where λ_1, λ_2, and λ_3 are three trade-off parameters which balance the contributions of different terms. The objective jointly learns mapping matrices with respect to each modality to transform samples from heterogeneous domains into a shared subspace and a discriminative encoding dictionary to reconstruct features in the shared space. Similarity scores between samples then can be computed by Mahalanobis distance of encoding coefficients.

C. Optimization

To optimize the objective function Eq. 7, we first make some simplification and rewrite it to a compact form. For convenience, we define some auxiliary matrices

$$\begin{aligned} \mathbf{U} = [\mathbf{U}^1, \mathbf{U}^2, \cdots, \mathbf{U}^P], \mathbf{A} = [\mathbf{A}^1, \cdots, \mathbf{A}^P] \\ and \ \phi(\mathbf{X}) = \begin{bmatrix} \phi(\mathbf{X}^1) & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \phi(\mathbf{X}^2) & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \phi(\mathbf{X}^P) \end{bmatrix}, \end{aligned} \quad (8)$$

where $\mathbf{0}$ is a zero matrix whose entries are all 0. Therefore, the first term $\mathcal{J}_1$ can be simplified as

$$\mathcal{J}_1(\mathbf{U}, \mathbf{D}, \mathbf{A}) = \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{DA}\|_F^2 + \alpha \|\mathbf{A}\|_F^2 \quad (9)$$

By ignoring some constant terms, according to [27], the second term $\mathcal{J}_2$ can be rewritten as

$$\mathcal{J}_2(\mathbf{U}) = -trace(\mathbf{U}\phi(\mathbf{X})\phi(\mathbf{X})^\top \mathbf{U}^\top) \quad (10)$$

And the third term can be rewritten as

$$\mathcal{J}_3(\mathbf{U}) = trace(\mathbf{UZU}^\top) \quad (11)$$

where

$$\mathbf{Z} = \begin{bmatrix} (P-1)\mathbf{I} & -\mathbf{I} & \cdots & -\mathbf{I} \\ -\mathbf{I} & (P-1)\mathbf{I} & \cdots & -\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I} & -\mathbf{I} & \cdots & (P-1)\mathbf{I} \end{bmatrix}$$

and $\mathbf{I}$ denotes the identity matrix.

By substituting Eq. [9, 10, 11] and Eq. 5 into Eq. 7, the optimization problem can be finally simplified as

$$\begin{aligned} \min_{\mathbf{U}, \mathbf{D}, \mathbf{A}, \mathbf{W}} & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{DA}\|_F^2 - trace\{\mathbf{U}(\lambda_2 \mathbf{Z} - \lambda_1 \phi(\mathbf{X})\phi(\mathbf{X})^\top)\mathbf{U}^\top\} \\ & + \lambda_3 trace(\mathbf{WALA}^\top \mathbf{W}^\top) + \alpha \|\mathbf{A}\|_F^2 + \gamma \|\mathbf{W}\|_F^2 \\ \text{s.t.} \quad & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \dots, K; \lambda_1, \lambda_2, \lambda_3, \gamma \geq 0 \end{aligned} \quad (12)$$

It is clear that the objective function in Eq.12 is not jointly convex to variables $\mathbf{U}$, $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{W}$. The formulation cannot be directly solved by convex optimization. Following by [4], [29], [32], [33], [35], we adopt an iteration optimization procedure which alternatively optimizes one variable by fixing others, as follows:

(1) *Initialization.* Considering efficiency of the optimization, some initializations on variables are made based on empirical experience: (a) Mapping matrices, *i.e.*, $\mathbf{U}^p (p = 1, \dots, P)$ and $\mathbf{W}$, are all initialized as identity matrix; (b) The shared dictionary $\mathbf{D}$ and the corresponding coefficients $\mathbf{A}$ are initialized

by solving the standard dictionary learning problem in which input feature matrix $\mathbf{X}$ is defined as in Eq. 9.

(2) *Given* $\mathbf{D}, \mathbf{A}$ *and* $\mathbf{W}$*, update* $\mathbf{U}$. By ignoring the irrelevant terms regarding variable $\mathbf{U}$, we can rewrite the optimization problem Eq. 12 as

$$\begin{aligned} \min_{\mathbf{U}} \mathcal{L}(\mathbf{U}) = \min_{\mathbf{U}} \ & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{DA}\|_F^2 \\ & + trace\{\mathbf{U}(\lambda_2 \mathbf{Z} - \lambda_1 \phi(\mathbf{X})\phi(\mathbf{X})^\top)\mathbf{U}^\top\} \end{aligned} \tag{13}$$

By setting $\frac{\partial \mathcal{L}(\mathbf{U})}{\partial \mathbf{U}} = 0$, we get the analytical solution of $\mathbf{U}$: $\mathbf{U} = \mathbf{DA}\phi(\mathbf{X})^\top \mathbf{\Omega}^{-1}$, where $\mathbf{\Omega} = [(1 - \lambda_1)\phi(\mathbf{X})\phi(\mathbf{X})^\top + \lambda_2 \mathbf{Z}]$. When $\mathbf{U}$ is obtained, $\mathbf{U}^p$ can be computed by splitting $\mathbf{U}$ into slices according to Eq. 8.

(3) *Given* $\mathbf{U}, \mathbf{A}$ *and* $\mathbf{W}$*, update* $\mathbf{D}$. The optimization problem is reduced to

$$\begin{aligned} \min_{\mathbf{D}} \ & \|\mathbf{U}\phi(\mathbf{X}) - \mathbf{DA}\|_F^2, \\ \text{s.t.} \ & \|\mathbf{d}_i\|_2^2 \leq 1, \forall i, i = 1, \ldots, K \end{aligned} \tag{14}$$

Define $\tilde{\mathbf{X}} \triangleq \mathbf{U}\phi(\mathbf{X})$, the quadratic constrained least square problem can be solved using the Lagrange dual approach. As in [36], the optimal solution of Eq. 14 is $\mathbf{D}^* = \tilde{\mathbf{X}}\mathbf{A}^\top(\mathbf{AA}^\top + \Lambda)^+$, where superscript $+$ denotes pseudo inverse operation and Λ is a diagonal matrix whose diagonal entries are dual variables.

Algorithm 1: Top-push Constrained Modality-Adaptive Dictionary Learning

```
Input: Features of training images from $P$ modalities:
    $\phi(\mathbf{X}^p), p = 1, \ldots, P$, parameters $\alpha, \gamma, \rho, \eta$, and
    $\lambda_i, i \in \{1, 2, 3\}$, iteration number $T$.
Output: The feature learning matrices $\mathbf{U}^p, p = 1, \ldots, P$,
    the learned shared dictionary $\mathbf{D}$ and the learned
    projection matrix $\mathbf{W}$.
Initialize $\mathbf{U} = [\mathbf{U}^1, \mathbf{U}^2, \ldots, \mathbf{U}^P], \mathbf{D}, \mathbf{W}$ and
    $\mathbf{A} = [\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^P]$ as described in subsection III-C;
for $\kappa = 1 \longrightarrow T$ do
    Update $\mathbf{U}^p$ by Eq. 13;
    Update $\mathbf{D}$ according to Eq. 14;
    Calculate Laplacian matrix $\mathbf{L}$ using current $\mathbf{W}$ and
        $\mathbf{A}$ by Eq. 5 and 6;
    while Non-convergence do
        Update $\mathbf{A}$ by Eq.15;
    end
    Calculate Laplacian matrix $\mathbf{L}$ using current $\mathbf{W}$ and
        $\mathbf{A}$ by Eq. 5 and 6;
    while Non-convergence do
        Update $\mathbf{W}$ by Eq. 18;
    end
end
```

(4) *Given* $\mathbf{U}, \mathbf{D}$ *and* $\mathbf{W}$*, update* $\mathbf{A}$. With $\mathbf{U}, \mathbf{D}$ and $\mathbf{W}$ fixed, we obtain the following loss function

$$\mathcal{F}(\mathbf{A}) = \|\tilde{\mathbf{X}} - \mathbf{DA}\|_F^2 + \lambda_3 trace(\mathbf{WALA}^\top \mathbf{W}^\top) + \alpha\|\mathbf{A}\|_F^2 \tag{15}$$

It is noticeable that the Laplacian matrix $\mathbf{L}$ in term $trace(\mathbf{WALA}^\top \mathbf{W}^\top)$ explicitly depends on variable $\mathbf{A}$ during the iterations, which causes the objective Eq. 15 intractable. Inspired by [4], we pre-calculate $\mathbf{L}$ using the prior $\mathbf{A}$ which ensures the objective to be convergent, and then update $\mathbf{L}$ with the new $\mathbf{A}$. Since $\mathbf{L}$ is non-positive semi-definite, we alternatively optimize objective Eq. 15 with gradient descent method, given as

$$\mathbf{A}^{(t)} := \mathbf{A}^{(t-1)} - \eta \bigtriangledown \mathcal{F}(\mathbf{A}^{(t-1)}), \quad t \geq 1 \tag{16}$$

where $\mathbf{A}^{(t)}$ denotes the t-th step to update variable $\mathbf{A}$, $\eta(\eta \geq 0)$ is the learning rate, and the gradient of Eq. 15 with respect to $\mathbf{A}$ is calculated by

$$\bigtriangledown \mathcal{F}(\mathbf{A}) = 2\mathbf{\Theta A} + \lambda_3 \mathbf{W}^\top \mathbf{WA}(\mathbf{L}^\top + \mathbf{L}) - 2\mathbf{D}^\top \tilde{\mathbf{X}} \tag{17}$$

where $\mathbf{\Theta} \triangleq \mathbf{D}^\top \mathbf{D} + \alpha \mathbf{I}$. When updating $\mathbf{A}$ in the t-th iteration, $\mathbf{L}$ is firstly pre-computed with fixed $\mathbf{A}^{(t-1)}$. After obtaining $\mathbf{A}^{(t)}$ by Eq. 17, $\mathbf{L}$ is subsequently updated.

(5) *Given* $\mathbf{U}, \mathbf{D}$ *and* $\mathbf{A}$*, Update* $\mathbf{W}$. When $\mathbf{U}, \mathbf{D}$ and $\mathbf{A}$ are fixed, the objective function in Eq.12 can be rewritten as

$$\mathcal{H}(\mathbf{W}) = \lambda_3 trace(\mathbf{WALA}^\top \mathbf{W}^\top) + \gamma\|\mathbf{W}\|_F^2 \tag{18}$$

As in step (3), the Laplacian matrix $\mathbf{L}$ is also relevant to $\mathbf{W}$, so $\mathbf{L}$ should be kept fixed when updating $\mathbf{W}$. Gradient descent method is also utilized to optimize $\mathbf{W}$, and the corresponding gradient is deduced as

$$\bigtriangledown \mathcal{H}(\mathbf{W}) = \mathbf{W}[\lambda_3 \mathbf{A}(\mathbf{L}^\top + \mathbf{L})\mathbf{A}^\top + 2\gamma \mathbf{I}] \tag{19}$$

After obtaining $\mathbf{W}$, we update $\mathbf{L}$ subsequently according to Eq. 4 and Eq. 5.

D. Complexity Analysis

The complete algorithm is summarized in **Algorithm 1**. In practice, the objective Eq. 12 can converge to the local optimum after $T = 30$ iterations. According to the procedure, computational costs are mainly caused by inverse operations in Eq. 13 and learning dictionaries in Eq. 14, which is $\mathcal{O}((\sum_{p=1}^P n_p)^3)$ and $\mathcal{O}(K^3)$ respectively in each iteration. Thus, the computational complexity in T iterations is $\mathcal{O}(T[(\sum_{p=1}^P n_p)^3 + K^3])$.

E. Matching for Heterogeneous Person Re-ID

Given a query person feature vector $\phi(x^p)$ from the p-th modality and gallery person feature vectors from the g-th modality $\phi(\mathbf{x}_i^g), i = 1, \ldots, N_g$, the encoding coefficients $\mathbf{a}^p$ and $\mathbf{a}_i^g$ can be computed by

$$\begin{aligned} \mathbf{a}^p &= \arg\min_{\mathbf{a}} \|\mathbf{U}^p \phi(\mathbf{x}^p) - \mathbf{Da}\|_2^2 + \alpha\|\mathbf{a}\|_2^2 \\ \mathbf{a}_i^g &= \arg\min_{\mathbf{a}} \|\mathbf{U}^g \phi(\mathbf{x}_i^g) - \mathbf{Da}\|_2^2 + \alpha\|\mathbf{a}\|_2^2 \end{aligned} \tag{20}$$

with respect to the shared dictionary $\mathbf{D}$. Then, the similarity scores between the query person and gallery persons can be calculated by

$$Score(i) = -\|\mathbf{W}(\mathbf{a}^p - \mathbf{a}_i^g)\|_2, \forall i, i = 1, \ldots, N_g \tag{21}$$

Thus, we assign the query sample $\mathbf{x}^p$ to the category corresponding to the largest score in the gallery set.

IV. Experiments

In this section, we evaluate our method TCMDL (AsyDic +Top-push) on two benchmark datasets: NIR versus VIS Re-ID dataset SYSU-MM01 [5] and the classical RGB-D person Re-ID dataset BIWI RGBD-ID [13], [21]. Moreover, two variants of the proposed method, *i.e.*, Dic+Top-push by removing asymmetric mapping and AsyDic+Triplet by changing top-push constraint to classical triplet ranking constraint, are compared to demonstrate the benefits of combining AsyDic and top-push constraint. In the paper, two popular evaluation metrics are adopted: Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) [18]. We reported the rank-k accuracy on both datasets which is the cumulative identification rate of the true matches in the top k ranks. Moreover, mAP is reported on both datasets which are mean of average precision scores for each query. We randomly repeat our evaluation for 10 times and report the average performances. All of the experiments are performed using Matlab on a desktop with a configuration of 64-bit OS, Intel(R) Core(TM) i5-6300U CPU @ 2.4 GHz and 8GB RAM.

Baselines. In the paper, four types of approaches are compared, which are state-of-the-art metric learning approaches for person Re-ID, cross-modality retrieval models, dictionary learning approaches and popular deep learning models. The representative metric learning approaches include KISSME [37] and XQDA [2]. The cross-modality retrieval methods include CCA [38], GMA [39], SCM [40] and CRAFT [41]. We also compare with supervised dictionary learning method DicRW [4] and unsupervised cross-dataset transfer learning UMDL [30]. For deep learning methods, we compare four state-of-the-art models on thermal-visible[1] person Re-ID, DeepZero [5], TONE [23], BCTR [24] and BDTR [24]. For KISSME, XQDA and the dictionary-based methods, we compute matching scores using the Mahalanobis distance. For all other methods, matching scores are directly calculated by Euclidean distance. Since the feature dimensions of depth images and RGB images in BIWI RGBD-ID dataset are different, PCA is firstly applied to get a fixed dimensional (*i.e.,* 80) feature.

A. Experiments on SYSU-MM01

SYSU-MM01 is the only public dataset for cross-modality person Re-ID. The dataset includes 287,628 RGB images from 4 VIS cameras in bright environments and 15,792 NIR images from 2 NIR cameras in dark environments of 491 valid identities. Some examples are shown in Fig. 4. It is clear that images captured by NIR sensors are different from those captured by VIS sensors from perceptual experience. In NIR images, color and texture information which are critical for traditional person Re-ID using VIS images are seriously degraded. The large data biases cannot be generalized by previous approaches for person Re-ID and thus incurs the cross-modality Re-ID problem. Although only two modalities are in our experiment, *i.e.*, NIR images and VIS images, the proposed method is also suitable for the scenarios where multiple modalities are available.

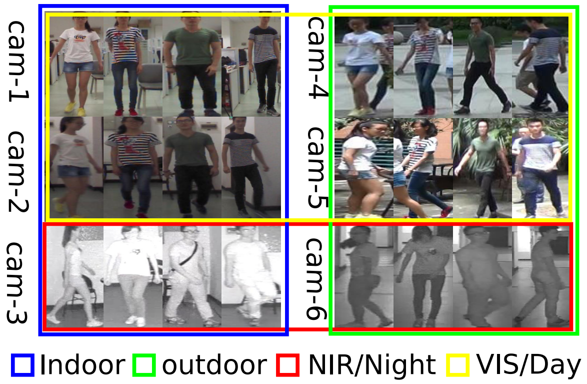


Fig. 4. Examples of samples in SYSU-MM01 dataset. Images from cameras 1-3 in the blue box are captured on indoor scenes while images from camera 4-6 in the green box are captured on outdoor scenes. Cameras 1, 2, 4, 5 are visual light sensors and cameras 3, 6 are near-infared sensors. Every column represents images from the same person.

1) Setting: **Feature Representation.** We use two kinds of feature representations $\phi(\cdot)$ to evaluate our approach, *i.e.*, LOMO [2] and Deep Zero-Padding (DZP) [5]. LOMO is a state-of-the-art hand-crafted feature representation for classical single-modality person Re-ID which characterizes person using color and texture information. DZP is learned by a one-stream network which extracts features of heterogeneous data by learning domain-specific nodes.

Evaluation Protocol. We follow the evaluation protocol of [5] which 296 fixed identities are for training and another 96 identities for testing. Differently, we leverage one image per identity for training in the training set, which is identical for single-shot person Re-ID. During testing, we follow the two validation modes of [5], *all-search* mode and *indoor-search* mode. In both modes, all images of NIR images from two NIR cameras form the probe set. Particularly, images from all VIS cameras form the gallery set for *all-search* mode while images from VIS camera #1 and #2 deployed indoor form the gallery set for *indoor-search* mode. In both modes, we follow the single-shot setting in [5] that only chooses one image for each identity in the gallery set and all images in the probe set (3803 query images).

Parameter Setting. In our experiments, we empirically set balance parameters λ_1 and λ_2 as 0.002 and 0.001, parameters for regularization terms α and γ as 0.05. λ_3 is set to $\beta/N(\rho)$ where $N(\rho)$ is the number of triplet sets, β is set to 800 and 50 empirically for LOMO and DZP respectively. As in metric learning approaches [3], [12], the margin ρ is simply set to 1. More detailed parameter analysis is in subsection IV-C.

2) Evaluation: **LOMO Feature Representation.** We extract LOMO using the code[2] provided by [2] with default parameters. All images from SYSU-MM01 dataset are resized into 160 × 60 due to varying bounding box size and thus generate 35722-dimensional features. To overcome dimension curse, we perform PCA on LOMO vectors from each modality respectively and reduce dimension to 300.

[1] In this paper, we use thermal-visible and infrared-RGB exchange.

[2] Code is available on http://www.cbsr.ia.ac.cn/users/scliao/projects/lomo_xqda/index.html

TABLE I
RESULTS USING LOMO (%). '-' MEANS RESULT IS NOT REPORTED.

Method		*all-search*					*indoor-search*				
		mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
Metric Learning	Euclidean	3.63	1.60	7.30	14.07	26.62	8.52	2.56	12.22	22.90	41.43
	KISSME [37]	4.43	1.76	9.12	17.55	32.46	9.78	3.20	14.18	26.27	48.56
	XQDA [2]	4.22	1.88	9.06	17.17	31.15	9.27	3.16	13.65	25.18	45.82
Cross-Modality Retrieval	CCA [38]	3.57	1.37	6.84	13.06	25.54	8.09	2.06	11.40	22.43	41.36
	GMA [39]	4.22	1.97	8.64	16.18	29.68	9.36	2.90	13.51	25.48	46.25
	SCM [40]	4.08	2.00	8.87	16.42	30.00	9.31	3.10	13.61	25.36	46.27
	CRAFT [41]	3.55	1.53	7.17	13.69	25.93	8.09	2.31	11.40	21.42	40.26
Dictionary Learning	DicRW(Dic+Triplet) [4]	4.06	1.84	8.86	15.65	29.03	8.93	3.10	13.30	24.43	43.29
	UMDL [30]	4.61	2.46	10.22	18.24	32.19	9.35	3.09	13.25	24.86	45.18
Ours	**Dic+Top-push**	4.30	1.99	8.97	16.49	30.04	9.94	3.45	14.71	27.04	48.21
	AsyDic+Triplet	4.97	2.48	10.86	19.60	35.08	11.72	4.52	**18.34**	**31.87**	**54.56**
	TCMDL	**5.07**	**2.61**	**11.21**	**20.17**	**35.65**	**11.79**	**4.65**	18.33	31.65	54.24

TABLE II
RESULTS USING DZP (%). '-' MEANS RESULT IS NOT REPORTED.

Method		*all-search*					*indoor-search*				
		mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
Metric Learning	Euclidean [5]	15.95	14.80	-	54.12	71.33	26.92	20.58	-	68.38	85.79
	KISSME [37]	14.81	12.24	34.69	50.47	69.30	27.36	16.44	45.99	64.08	82.46
	XQDA [2]	18.42	15.87	40.95	57.57	75.72	31.16	20.04	51.08	68.70	86.15
Cross-Modality Retrieval	CCA [38]	19.10	16.71	42.43	58.46	76.12	32.11	21.46	51.68	68.91	85.91
	GMA [39]	12.02	10.34	29.69	43.25	59.66	21.93	13.40	35.91	51.78	71.18
	SCM [40]	4.70	2.60	10.78	19.44	34.17	10.07	3.56	15.19	27.60	48.86
	CRFAT [41]	4.79	3.08	11.14	19.00	32.07	10.10	3.96	15.87	27.24	46.97
Dictionary Learning	DicRW(Dic+Triplet) [4]	17.60	14.41	38.92	54.48	72.55	30.60	20.47	50.23	68.25	85.46
	UMDL [30]	17.45	15.35	39.68	55.04	72.31	28.67	18.82	46.33	63.05	80.60
Ours	**Dic+Top-push**	17.66	15.85	40.75	56.42	73.36	30.98	21.51	49.05	64.72	81.52
	AsyDic+Triplet	**19.32**	16.57	42.62	58.62	**77.23**	32.19	21.49	52.02	69.37	86.20
	TCMDL	19.30	**16.91**	**42.74**	**58.83**	76.64	**32.27**	**21.60**	**54.26**	**71.38**	**87.91**

Table I lists the CMC and mAP results using LOMO for *all-search* mode and *indoor-search* mode, respectively. We can observe cross-modality retrieval methods such as GMA, SCM and ours achieve better performances than classical metric learning for person Re-ID. Compared to using baseline Euclidean metric, these methods achieve 0.37 (from 1.6% to 1.97%), 0.40% (from 1.60% to 2.00%) and 1.01% (from 1.6% to 2.61%) improvements of rank-1 accuracy in *all-search* mode and 0.34% (from 2.56% to 2.90%), 0.54% (from 2.56% to 3.10%), 2.09% (from 2.56% to 4.65%) improvements of rank-1 accuracy in *indoor-search* mode respectively. This is because cross-modality methods can mitigate data biases between heterogeneous data while classical metric learning methods cannot generalize large differences. However, CCA and CRAFT achieve poor performance due to severe noises, *i.e.* color information. Among the cross-modality methods, our model achieves better performances in all modes, especially in larger ranks, *e.g.* more than rank-10. Compared to the single-modality dictionary learning DicRW [4], our model improves the performance with a large margin in both modes (from 1.84% to 2.61% for *all-search* and from 3.10% to 4.65% for *indoor-search*). This validates the effectiveness of our model to mitigate data biases for cross-modality person Re-ID task. In particular, our method TCMDL improves rank-1 accuracy from 1.99% to 2.61% and mAP from 4.30% to 5.07% for *all-search*, improves rank-1 accuracy from 3.45% to 4.65% and mAP from 9.94% to 11.79% for *indoor-search* compared to Dic+Top-push. The results show the effectiveness of asymmetric mapping. In another aspect, our method TCMDL improves rank-1 accuracy from 2.48% to 2.61% and mAP from 4.97% to 5.07% for *all-search*, improves rank-1 accuracy from 4.52% to 4.65% and mAP from 11.72% to 11.79% for *indoor-search* compared to AsyDic+Triplet. The results demonstrate the effectiveness of top-push constrained regularization.

However, performances of the proposed methods are still limited, which the rank-1 and mAP are 2.61% and 5.07% for *all-search* mode and 4.65% and 11.79% for *indoor-search* respectively. This is because LOMO is characterized by color and texture information which are degraded seriously in NIR images. The imbalanced feature information from different cameras restricts the identification accuracy. Moreover, performances in *indoor-search* mode are much better than that in *all-search* mode because illumination and background interferences can be better controlled under indoor scenarios.

Deep Zero-Padding Feature Representation. In this paper, 256-dimension DZP[3] extracted by [5] are utilized. Both *all-search* mode and *indoor-search* mode are adopted with respect to the protocol above. The experiments are also conducted 10 times with random sample selection. Table II gives the results when using the DZP. Compared to LOMO, DZP is

[3] Available on http://isee.sysu.edu.cn/project/RGBIRReID.htm

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS FOR THERMAL-VISIBLE RE-ID ON SYSU-MM01 DATASET.

Methods	mAP	rank-1	rank-10	rank-20
DeepZero [5]	15.95	14.80	54.12	71.33
TONE [23]	14.42	12.52	50.72	68.60
BCTR [24]	19.15	16.12	54.90	71.47
BDTR [24]	**19.66**	**17.01**	55.43	71.96
Ours	19.30	16.91	**58.83**	**76.64**

specially designed for cross-modality person Re-ID, which results in better performances than that in Table I when using the same method. As in Table II, our model achieves the best performances no matter CMC ranks or mAP. Though larger data biases have been already migrated when learning DZP, our model obtains about 3.4% (from 15.95% to 19.36%) and 6.4 % (from 26.92% to 33.35%) mAP improvements compared to the baseline using Euclidean metric. It shows strong feature representation augmentation power of our model. In particular, our model outperforms other cross-modality models with a large margin, *e.g.*, more than 13% and 19% rank-1 accuracy improvement than CRFAT for *all-search* mode and *indoor-search* mode, respectively. In addition, our model achieves more than 2% mAP improvement than DicRW, which shows strong feature representation ability when data biases exist. Similar to the conclusion using LOMO, the proposed model outperforms Dic+Top-push and AsyDic+Triplet in most cases. This validates the benefits of combining asymmetric mapping and top-pushed constrained dictionary learning together.

Table III compares the performances of *all-search* mode with the state-of-the-art thermal-visible person Re-ID methods on SYSU-MM01 dataset. These methods either use one-stream or two-stream neural networks to mitigate data biases across RGB visual images and infrared images. From the table, it is easy to observe that our method achieves equivalent performances with these methods. Especially, it outperforms DeepZero and TONE with a large margin since we use deep zero-padding features as input and further mitigate modality biases by jointly performing asymmetric mapping and discriminative dictionary learning.

Computation Analysis. On average, it takes 314.1s and 1196.4s to train the proposed TCMDL using DZP for *indoor-search* and *all-search* mode, respectively. And, it averagely takes 470.0s and 1303.1s to train TCMDL using LOMO for the two modes on SYSU-MM01 dataset. It costs more time to train using LOMO than DZP since LOMO is in higher dimensions after dimension reduction using PCA. However, the proposed method is much more efficient than CNN based approaches. For instance, TONE costs 10.9h for *all-search* mode, which is about 30 times longer than the proposed method.

B. Experiments on BIWI RGBD-ID dataset

The BIWI RGBD-ID dataset [13], [21] is originally built for long-term person Re-ID that uses depth information instead of RGB images since depth image is robust to cloth changing and illumination. Different from [7], [13], [21], we evaluate cross-modality person Re-ID on the dataset, which considers depth images as query set and corresponding RGB images as gallery set. The dataset includes video sequences captured from 50 different people by a Kinect at about 10 frames per second. For each sequence of a subject, there are about 300 frames of RGB images and their corresponding depth images as well as skeletons. In the dataset of BIWI RGBD-ID, these sequences are originally provided in two groups corresponding to two folders "*Training*" and "*Testing*" which are denoted as *TR* and *TE* in the following. *TE* is sub-divided into two groups named as "*Still*" and "*Walking*". Only 28 out of 50 persons appeared in both *TR* and *TE* (*i.e., Still and Walking*), and they are collected on a different day which thus most subject dressed differently. In *TR*, the subject performs actions such as walking and rotation. In *Still*, people stand in front of the sensor and move slightly. In *Walking*, people walk frontally and diagonally against the Kinect. Some examples are shown in Fig. 5.

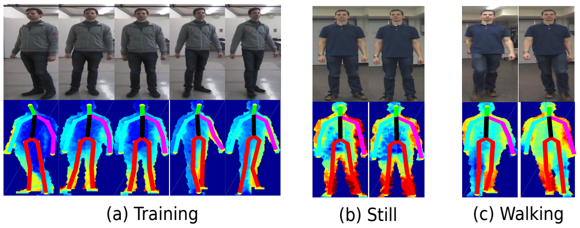


Fig. 5. Examples of images in the BIWI RGBD-ID dataset. Images in the top row are RGB images and in the bottom row are depth images (shown by pseudo-color) as well as skeletons.

1) Setting: **Feature Representation.** Due to the large gaps between RGB images and depth images, we extract features with different methods for them. For depth images, we first convert them into point clouds as in [13] and describe body shape and skeleton of people using 510-dimension Eigen-depth feature [7] and 13-dimension skeleton-based feature [21]. Some examples of RGB images of persons in the dataset and their corresponding point clouds are shown in Fig. 6. For RGB images, we also hope the visual features could describe body shape rather than colour information because the color is varying between *TR* and *TE* groups. In this paper, we utilized LBP [42] and HOG [43] to describe texture and body silhouette, respectively. When extracting LBP and HOG features, we firstly resize all RGB images to 128 × 48 and convert them to grayscale images. 8 × 8 cells are used for LBP and HOG feature extraction, which results in 5664 and 2700 dimensional features.

Evaluation Protocol. To evaluate the proposed method on BIWI RGBD-ID dataset, we follow the data protocol in [7] which images of 28 people who appeared in both *TR* and *TE (Walking and Still)* are used for testing, and the remaining 22 subjects who only appeared in *TR* are used for training the model. Different from Wu *et al.* [7], we take RGB images to construct the gallery set and take depth images as the probe. Thus, three groups of testing sets are built, *i.e.*, images of the 28 persons who appeared in all of *TR*, *Walking* and *Still*. We termed the three groups of testing sets as subset #1, #2 and subset #3. Since depth information is not complete for all frames in one sequence, we selected samples as advised in

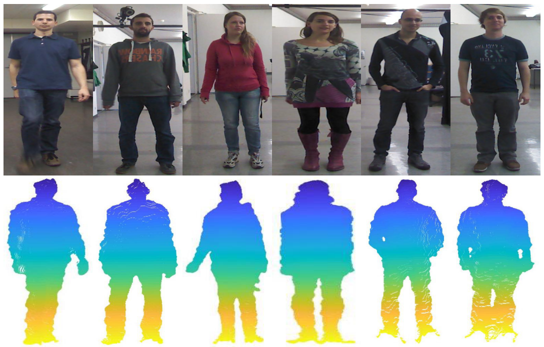


Fig. 6. Examples of point clouds. Images in top row is the RGB images in BIWI RGBD-ID dataset and images in bottom row are their corresponding visualization of point clouds.

[13] and 5 frames are used for each sequence.

Parameter Setting. For BIWI RGBD-ID dataset, we also empirically set trade-off parameters λ_1 as 0.002, parameters for regularization terms α and γ as 0.005. λ_3 is set to $\beta/N(\rho)$ where $N(\rho)$ is the number of triplet sets, β is set to 0.8. As in metric learning approaches [3], [12], the margin ρ is simply set to 1. Since the large gaps between different feature representation from RGB images and depth images, we omit the cross-modality consistency term in Eq. 10 and set the dimension of asymmetric mapping matrices to be 80. Considering the number of people in the dataset, we set the dictionary size to 50.

2) Evaluation: Table IV-VI list results on the three testing subsets of BIWI RGBD-ID dataset, *i.e.*, *TR*, *Walking* and *Still*. It is easy to observe that our proposed method achieves significant results on all the three subsets. In detail, our proposed method achieves 10.71%, 7.14% and 7.14% accuracy on the three subsets at rank-1, which outperforms baseline Euclidean with a large margin. For other ranks, our method also achieves top and stable performances in most cases. This shows the effectiveness of our proposed method.

In another aspect, results on subset #1 and subset #3 are better than that on subset #2. It is reasonable because data distribution of subset #1 is more similar than other subsets and subset #3 suffers less motion variation than other subsets. As in the tables, cross-modality methods achieve relevantly good results than single-modality methods. It is because asymmetric mapping bridges the data gaps to some content. It is interesting that XQDA achieves remarkable performances in some cases in benefits of powerful discriminative ability of metric learning. However, it still cannot reach the performance of the proposed method due to the large gaps between modalities.

Computation Analysis. As describe before, we train TCMDL on BIWI RGBD-ID dataset with the same hardware and software configuration. It averagely takes 16.7s by using 220 samples, which is much faster than the training on SYSU MM01. We suppose the main reason is that BIWI RGBD-ID includes much fewer samples.

TABLE IV
RESULTS ON BIWI RGBD-ID SUBSET #1 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	9.82	2.14	10.71	24.29	54.29
KISSME [37]	15.97	3.57	22.86	35.71	74.29
XQDA [2]	19.70	6.43	26.43	47.86	**90.00**
CCA [38]	18.62	7.14	26.43	37.86	70.71
GMA [39]	14.42	3.57	17.86	39.29	73.57
CRAFT [41]	12.03	0.71	18.57	40.00	72.86
Ours	**19.96**	**10.71**	**29.29**	**48.57**	84.29

TABLE V
RESULTS ON BIWI RGBD-ID SUBSET #2 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	10.76	0.00	12.14	40.00	75.00
KISSME [37]	13.82	3.57	18.57	42.14	75.71
XQDA [2]	17.31	6.43	17.86	**47.86**	74.29
CCA [38]	15.51	4.29	20.00	35.00	**77.86**
GMA [39]	14.13	3.57	17.86	36.43	72.14
CRAFT [41]	16.50	4.29	**24.29**	42.14	70.71
Ours	**18.88**	**7.14**	20.00	41.43	77.14

C. Parameter Analysis

1) Analysis on SYSU-MM01: As in the objective function Eq. 12, our model includes four parts, and corresponding three parameters λ_1, λ_2 and $\lambda_3 = \beta/N(\rho)$ to balance contributions of each part. Fig. 7(a-c) shows the performance changes under two evaluation metrics (*i.e.*, rank-1 and mAP) in terms of the three different trade-off parameters on the SYSU-MM01 dataset when using DZP, respectively. From the figure, it can be observed that our model is less sensitive to λ_1 and λ_2 than λ_3. However, we can still find that both rank-1 and mAP performance rise with increasing of λ_1 and achieve peak values at 0.002. After that, they show a downward trend with small fluctuation. And, λ_2 varies in a similar way with λ_1 except achieves peak performance at 0.001. Though the performance is relatively stable when the choices of λ_1 and λ_2 in suitable ranges, it is easy to get better performance when setting λ_1 and λ_2 to 0.002 and 0.001 from the above analysis. From Fig. 7(c), we can see that the performances fluctuate drastically when β is less than 20. However, it can achieve stable performance when β is larger than 20. To balance each part in the objective, we empirically set parameters $\lambda_1 = 0.002, \lambda_2 = 0.001$ and $\beta = 50$ when using DZP. As for LOMO, λ_1 and λ_2 comply the same variation rule and thus can be set to same value. But for β, we empirically find the optimum value is 800.

TABLE VI
RESULTS ON BIWI RGBD-ID SUBSET #3 (%).

Method	mAP	rank-1	rank-5	rank-10	rank-20
Euclidean	10.01	0.71	13.57	25.71	64.29
KISSME [37]	13.26	5.00	16.43	30.71	67.86
XQDA [2]	16.32	5.00	21.43	**43.57**	72.86
CCA [38]	13.40	5.00	14.29	27.86	58.57
GMA [39]	14.23	3.57	21.43	39.29	68.57
CRAFT [41]	**17.73**	6.43	**22.86**	37.86	71.43
Ours	17.53	**7.14**	20.00	42.14	**76.43**

Dictionary size K of $\mathbf{D}$ is another important parameter. Fig. 7(d) shows rank-1 and mAP accuracies for K in $[100, 500]$. We can observe that our model is not sensitive to dictionary

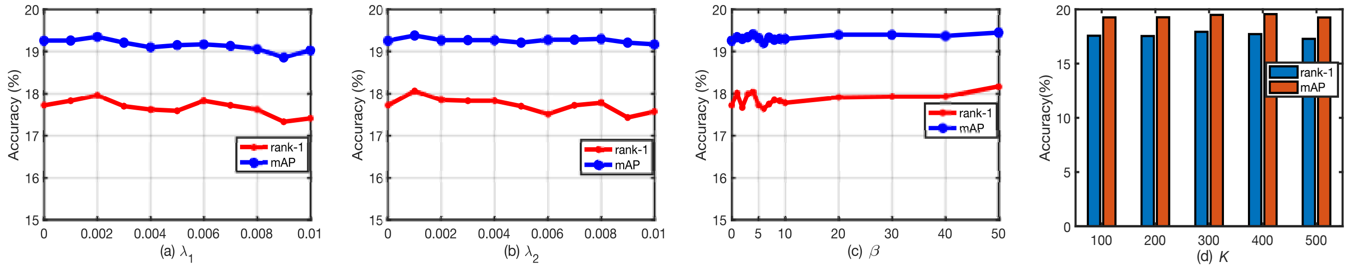


Fig. 7. Parameter analysis using DZP for all-search mode. Rank-1 and mAP accuracy with different parameters (a) λ_1, (b) λ_2, (c) β and dictionary size (d) K are reported.

size in suitable ranges. Considering the running time, we set $K = 300$ in all our experiments.

2) Analysis on BIWI RGBD-ID: Since different feature representations are used to describe RGB images and depth images on BIWI RGBD-ID dataset, we do not use cross-view consistency regularization as described in Sec. III-B3. Thus, only two parameters λ_1 and $\lambda_3 = \beta/N(\rho)$ are left to control the contribution of two regularization terms. Fig. 8(a) illustrates two evaluation metrics rank-1 and mAP against λ_1. It can be observed that both rank-1 and mAP performance rise with increasing of λ_1 and achieve the highest performance at 0.002. After that, the graph shows a downward trend with a small fluctuation. Fig. 8 (b) displays the relationship between accuracies measured by rank-1 and mAP and β. From the figure, both rank-1 and mAP performance show an upward trend and achieve peak performance at 0.8. After that, rank-1 performance drops drastically while mAP performance shows a downward trend with small fluctuations. From the above analysis, it is supposed to set $\lambda_1 = 0.002$ and $\beta = 0.8$ on BIWI RGBD-ID dataset. Fig. 8(c) illustrates the rank-1 and mAP accuracies for K in range $[20, 70]$. We can observe that both rank-1 and mAP performance achieve the highest performance when $K = 50$. This is much smaller than K for SYSU-MM01 dataset. We suppose the reason is that the number of people in BIWI RGBD-ID dataset is relatively small, which only has 22 for training and 28 for testing.

D. Effects of Energy Terms

As in the proposed objective Eq. 7, there are four components which control dictionary learning in shared subspace, prevent information loss, keep mapping matrices consistent and force dictionary discriminative, respectively. Basically, we want to reconstruct data from different modalities using a shared dictionary in a common latent subspace. Considering the propose, reconstruction error is minimized to ensure the learned dictionary to be representative. Three penalty terms are adopted to regularize the dictionary learning. We evaluate the effects of these regularization terms on both SYSU MM01 and BIWI RGBD-ID dataset and report results in Fig. 7 and Fig. 8. The energy-preserving regularization controlled by λ_1 tries to preserve as much information while performing the asymmetric mapping. As in Fig. 7(a) and Fig. 8(a), the performance will drop if λ_1 is too small due to too much information loss. However, there will be redundant if this term is too large. The cross-view consistency term aims to present the learned mapping matrices for different modalities varying too much. However, it incurs the mapping matrices to be the same if the term dominates the whole objective and thus cause poor performance. This is verified by results in Fig. 7(b). The last penalty term attempts to make the learned dictionary discriminative. As in Fig. 7(c) and Fig. 8(b), both small and large β will cause poor performance. This is because small β causes the term contributing less to the whole objective and thus makes the learned dictionary less discriminative while large β incurs over-fitting. Thus, it is important to choose appropriate trade-off parameters to balance the contribution of these different penalty terms.

V. Conclusions and Future Work

In this paper, we propose a top-push constrained modality-adaptive dictionary learning model for cross-modality person Re-ID. It aims to address the challenging long-term person Re-ID when people are captured by different types of sensors, *e.g.*, the query is captured by NIR/depth sensor in the night while the gallery images are collected by visual RGB sensors. Our model simultaneously maps the features from different modalities into a common subspace and learns a shared dictionary for the projected features from all modalities in the subspace. In the subspace, data biases across modalities are alleviated, and features from each modality can be encoded by the shared dictionary. In particular, we impose a top-push constraint to the coding coefficients which improves the discriminative ability of the learned dictionary. We evaluate the proposed top-push constrained modality-adaptive dictionary learning on two cross-modality person Re-ID datasets, SYSU-MM01 dataset and BIWI RGBD-ID dataset. Experiments on both datasets show promising performances and thus demonstrate the effectiveness of the proposed method.

However, the model is not an end-to-end approach, and the robustness of input features is still challenged by various facts such as camera view differences and pose indeterminacy. These factors cause the body-part misalignment problem that substantially influence the performance of the model. To solve the problem, two approaches can be investigated in the future: 1) hard body segmentation or automatic body parsing which can mitigate the problem caused by misalignment, and 2) Combining CNN models and our work together to build an end-to-end framework, and training them simultaneously. For

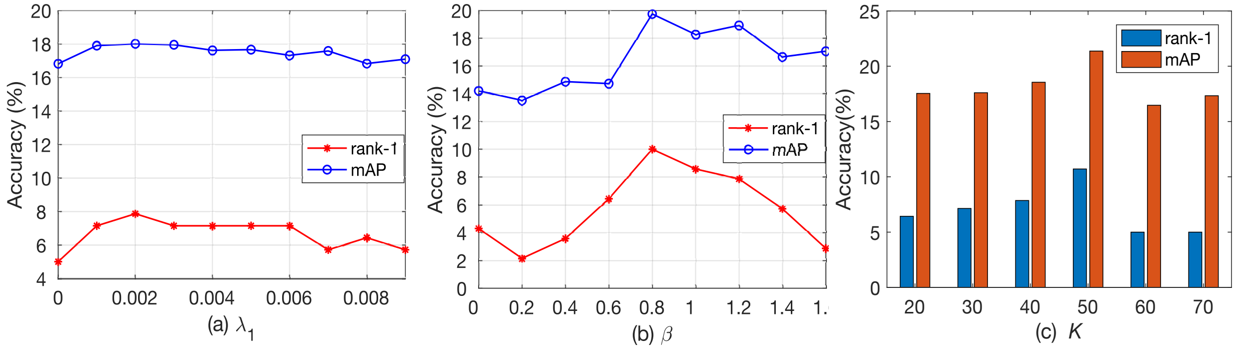


Fig. 8. Parameter analysis on subset #1 of BIWI RGBD-ID dataset. Rank-1 and mAP accuracy in terms of different parameters (a) λ_1, (b) β, and dictionary size (c) K are reported.

instance, a two-branch CNN backbone is adapted to extract features from both modalities and join them together by performing asymmetric mapping at the end of the backbone. This is achievable benefiting from the increasing scale of the Re-ID dataset and the development of GPU techniques in the future.

References

[1] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person re-identification using kernel-based metric learning methods,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 1–16.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[3] J. You, A. Wu, X. Li, and W.-S. Zheng, “Top-push video-based person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1345–1353.

[4] D. Cheng, X. Chang, L. Liu, A. G. Hauptmann, Y. Gong, and N. Zheng, “Discriminative dictionary learning with ranking metric embedded for person re-identification,” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 964–970.

[5] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, “Rgb-infrared cross-modality person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5380–5389.

[6] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, “Re-identification with rgb-d sensors,” in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2012, pp. 433–442.

[7] A. Wu, W.-S. Zheng, and J.-H. Lai, “Robust depth-based person re-identification,” *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, 2017.

[8] B. Li, H. Chang, S. Shan, and X. Chen, “Coupled metric learning for face recognition with degraded images,” *Adv. Mach. Learn.*, pp. 220–233, 2009.

[9] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, “Coupled bilinear discriminant projection for cross-view gait recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, 2019.

[10] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, “Coupled patch alignment for matching cross-view gaits,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, 2019.

[11] H. Yu, A. Wu, and W. Zheng, “Cross-view asymmetric metric learning for unsupervised person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 994–1002.

[12] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2501–2514, 2016.

[13] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool, “One-shot person re-identification with a consumer depth camera,” in *Person Re-Identification*. Springer, 2014, pp. 161–181.

[14] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 262–275, 2008.

[15] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, “Improving person re-identification by attribute and identity learning,” *Pattern Recognit.*, vol. 95, pp. 151–161, 2019.

[16] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with domain guided dropout for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1249–1258.

[17] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based cnn with improved triplet loss function,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1335–1344.

[18] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, “Person re-identification in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1367–1376.

[19] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, “Multi-pseudo regularized label for generated data in person re-identification,” *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391–1403, 2018.

[20] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, “Sbsgan: Suppression of inter-domain background shift for person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[21] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, “3d reconstruction of freely moving persons for re-identification with a depth sensor,” in *Proc. IEEE Conf. Appl. Rob. Autom. (ICRA)*, 2014, pp. 4512–4519.

[22] P. Zhang, Q. Wu, J. Xu, and J. Zhang, “Long-term person re-identification using true motion from videos,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018, pp. 494–502.

[23] M. Ye, X. Lan, J. Li, and P. C. Yuen, “Hierarchical discriminative learning for visible thermal person re-identification,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7501–7508.

[24] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 1092–1099.

[25] H. Mikami, H. Suganuma, P. U-chupala, Y. Tanaka, and Y. Kageyama. (2018) Imagenet/resnet-50 training in 224 seconds. [Online]. Available: https://arxiv.org/abs/1811.05233v1

[26] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, “Imagenet training in minutes,” in *Proc. Int. Conf. Parallel Comput. (ICPP)*, 2018, pp. 1–10.

[27] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, “Generalized domain-adaptive dictionaries,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 361–368.

[28] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, “Semi-supervised coupled dictionary learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3550–3557.

[29] S. Li, M. Shao, and Y. Fu, “Cross-view projective dictionary learning for person re-identification.” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 2155–2161.

[30] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, “Unsupervised cross-dataset transfer learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1306–1315.

[31] Q. Zhou, S. Zheng, H. Ling, H. Su, and S. Wu, “Joint dictionary and metric learning for person re-identification,” *Pattern Recognit.*, vol. 72, pp. 196–206, 2017.

[32] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang, "Joint semantic and latent attribute modelling for cross-class transfer learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1625–1638, 2017.

[33] J. Lu, G. Wang, and J. Zhou, "Simultaneous feature and dictionary learning for image set based face recognition," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4042–4054, 2017.

[34] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, 2011.

[35] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4516–4524.

[36] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 801–808.

[37] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2288–2295.

[38] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia. (ACM MM)*, 2010, pp. 251–260.

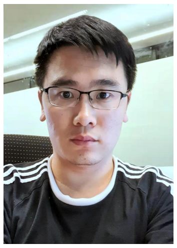
[39] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2160–2167.

[40] D. Zhang and W. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2014, pp. 2177–2183.

[41] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, 2018.

[42] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, 2005, pp. 886–893.

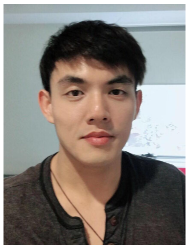


Peng Zhang received B.S. and M.S. degrees from the School of Information Science and Engineering in Shandong University (SDU), China. He is currently a Ph.D. student with the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), NSW, Australia. His research interests include gait recognition, person reidentification and deep learning.

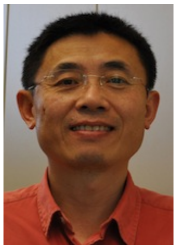
Jingsong Xu received the B.S. degree in computer science and the Ph.D. degree in pattern recogni- tion from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2014, respectively. He is currently a Research Fellow with the Global Big Data Technologies Center, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include computer vision, pattern recognition, and machine learning.

Qiang Wu received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Australia, in 2004.

He is currently an Associate Professor and a Core Member of the Global Big Data Technologies Centre, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. His research outcomes are applied span over fields such as video security surveillance, biometrics, video data analysis, and humancomputer interaction. His research outcomes have been published in many premier international conferences, including ECCV, CVPR, ICIP, and ICPR, and the major international journals, such as IEEE TIP, IEEE TSMC-B, IEEE TCSVT, IEEE TIFS, PR, PRL, and Signal Processing.



Yan Huang received B.S. and M.S. degrees from the School of Computer Science and Engineering in Sichuan University (SCU), China, and Beihang University (BUAA), China, respectively. He is currently a Ph.D. student with the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), NSW, Australia. His research interests include deep learning and person re-identification.



Jian Zhang (SM'04) received the B.S. degree in electronics from East China Normal University, China, the M.S. degree in computer science from Flinders University, Australia, and the Ph.D. degree in electrical engi- neering from the University of New South Wales (UNSW), Australia. From 2004 to 2011, he was a Principal Researcher and a Project Leader with Data61, Australia, and a Conjoint Associate Professor with the School of Computer Science and Engineering, UNSW. He is currently an Associate Professor with the Global Big Data Technologies Centre, University of Technology Sydney, Australia. He has authored or co-authored over 140 paper publications, book chapters, and six issued U.S. and Chinese patents. His current interests include social multimedia signal processing, large-scale image and video content analytics, retrieval and mining, 3D-based computer vision, and intelligent video surveillance systems.

Dr. Zhang was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2006 to 2015. He has been an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the EURASIP Journal on Image and Video Processing since 2016.