Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification

Mang Ye, Xiangyuan Lan, Zheng Wang, *Member, IEEE*, and Pong C. Yuen, *Senior Member, IEEE*

***Abstract*— Visible thermal person re-identification (VT-REID) is a task of matching person images captured by thermal and visible cameras, which is an extremely important issue in night-time surveillance applications. Existing cross-modality recognition works mainly focus on learning sharable feature representations to handle the cross-modality discrepancies. However, apart from the cross-modality discrepancy caused by different camera spectrums, VT-REID also suffers from large cross-modality and intra-modality variations caused by different camera environments and human poses, and so on. In this paper, we propose a dual-path network with a novel bi-directional dual-constrained top-ranking (BDTR) loss to learn discriminative feature representations. It is featured in two aspects: 1) end-to-end learning without extra metric learning step and 2) the dual-constraint simultaneously handles the cross-modality and intra-modality variations to ensure the feature discriminability. Meanwhile, a bi-directional center-constrained top-ranking (eBDTR) is proposed to incorporate the previous two constraints into a single formula, which preserves the properties to handle both cross-modality and intra-modality variations. The extensive experiments on two cross-modality re-ID datasets demonstrate the superiority of the proposed method compared to the state-of-the-arts.**

***Index Terms*— Person re-identification (REID), cross-modality, visible thermal (VT), top-ranking.**

I. Introduction

PERSON re-identification (re-ID) is a recognition task which aims at matching a specific query person from a set of person images captured by other disjoint surveillance cameras [1]. It has achieved increasing attention in research community due to its importance in intelligent video surveillance applications [2]–[4]. Most of current research progresses in re-ID field are mainly conducted on visible camera module, i.e., all the person images are captured by visible cameras in the daytime [5]–[7] under well-lighted environment. However, the general visible cameras usually cannot capture valid appearance information under low lighting conditions (e.g. in the night time), which limits the applicability of single visible modality re-ID for practical surveillance scenarios [8]–[10]. In this paper, we mainly focus on the cross-modality matching problem named visible thermal person re-identification (VT-REID), which aims at matching person images captured by visible and thermal cameras with different spectrums. Given a visible (thermal) image of a specific query person, VT-REID[1] needs to search out the corresponding thermal (visible) images from a gallery set, as illustrated in Fig 1. VT-REID plays an important role in night-time video surveillance applications.

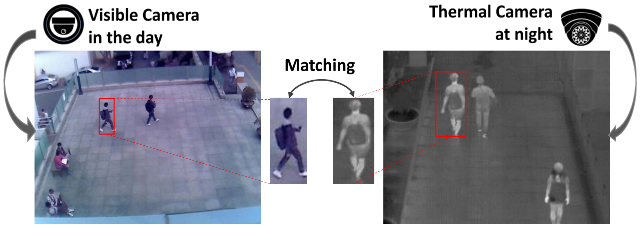


Fig. 1. Visible thermal person re-identification (VT-REID), matching person images across different spectrum cameras.

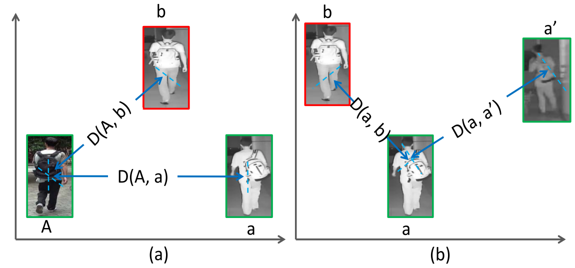


Fig. 2. Intra-class distance is larger than the inter-class distance caused by (a) Cross-modality variation and (b) Intra-modality variation. The color of the bounding box denotes the person identity.

Related cross-modality recognition has been widely studied in VIS-NIR face recognition [12], [13], face sketch recognition [14], [15] and so on. Typically, they mainly focus on addressing the *cross-modality discrepancy* [16], caused by different reflective visible spectrums and sensed emissivities of visible and thermal cameras. However, besides the cross-modality discrepancy, VT-REID also suffers from 1) large cross-modality variations caused by the different cross-camera views and similar visual appearance of different persons ($D(A,a) > D(A,b)$ in Fig. 2), and 2) large intra-modality variations caused by different human poses and viewpoints ($D(a,a') > D(a,b)$ in Fig. 2). Due to cross-modality and intra-modality variations, the visual appearance difference between the visible and thermal images are much more significant than those of face images [11]. It means that large

Manuscript received December 20, 2018; revised May 29, 2019; accepted May 29, 2019. Date of publication June 6, 2019; date of current version September 16, 2019. This work was supported in part by the Hong Kong Research Grant Council General Research Fund (RGC/HKBU 12200518). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Clinton Fookes. *(Corresponding author: Pong C. Yuen.)*

M. Ye, X. Lan, and P. C. Yuen are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: mangye@comp.hkbu.edu.hk; pcyuen@comp.hkbu.edu.hk; xiangyuanlan@life.hkbu.edu.hk).

Z. Wang is with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: wangz@nii.ac.jp).

Digital Object Identifier 10.1109/TIFS.2019.2921454

[1] It's also called RGB-infrared person re-ID in [11].

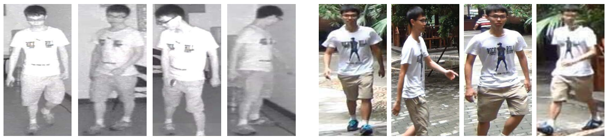


Fig. 3. Intra-class variations. Images represent the same person identity captured from two modalities.

amount of intra-class distances are much larger than that of inter-class distances. Therefore, their methods usually have limited performance for VT-REID [11].

To our best knowledge, three pioneer works have studied the VT-REID problem. In particular, Wu *et al.* [11] proposed a deep zero-padding method with one-stream network for multi-modality sharable feature learning. However, it only utilizes the identity information, which limits the discriminability of the learned features since the training and testing sets do not have identity overlap. A cross-modality generative adversarial network is presented in [17], but the adversarial training process is easy to get stuck for large-scale training. Another two-stage framework with feature learning and metric learning is introduced in in [18], but it needs additional human intervention which is unsuitable for large-scale practical applications [19]. In this paper, we propose an end-to-end learning method to learn discriminative features, which simultaneously considers both the cross-modality and intra-modality variations.

In particular, we introduce a dual-path network trained with bi-directional dual-constrained top-ranking (BDTR) loss for VT-REID. On one hand, the dual-path network aims at learning multi-modality sharable features, which contains a visible path and a thermal path for two different modalities. The network parameters of the shallow layers are independent to capture the modality-specific information, which tackles the cross-modality discrepancy problem. Then, a shared fully connected layer is introduced to learn the shared embedding space. In this manner, we simultaneously consider the modality commonality and discrepancy. On the other hand, the designed learning objective contains two constraints: 1) cross-modality top-ranking constraint, which focuses on handling the large cross-modality variations. Its main idea is that the distance of *anchor to all the cross-modality positive samples within the batch* should be smaller than the *anchor to all the cross-modality negative samples within the batch* by a predefined margin. 2) intra-modality top-ranking constraint, which aims at on addressing the intra-modality variations. Under the same sampled batch of the cross-modality top-ranking constraint, the intra-modality constraint ensures that distance between the *anchor's furthest-positive* and *its nearest-negative* within the same modality should also be large enough. Meanwhile, a bi-directional training strategy (*visible to thermal relationship and thermal to visible relationship*) is employed to enhance the robustness of the learned feature representation.

In addition, there are also large intra-class variations in VT-REID, as shown in Fig. 3. The ranking loss only exploits the relationships among persons across heterogenous modalities, which cannot guarantee the invariance of the learned features. Therefore, we further aggregate the identity loss into the dual-constrained top-ranking loss to model the identity-invariant information. The main idea is that images of the same person identity across heterogenous modalities are treated as the same class. The identity loss also helps to stabilize the overall training process since the network may have totally different parameters for two paths caused by heterogenous modalities.

A preliminary conference version [20] has been published in IJCAI-ECAI 2018. In this journal version, we have following three major improvements: Firstly, we give a detailed explanation and analysis about the rationale intra-modality constraint for the cross-modality person re-identification problem. Secondly, we propose a center-constrained variant which incorporates previous two constraints into a single formula. It preserves the properties to handle both cross-modality and intra-modality variations simultaneously. Finally, comprehensive analysis and experimental evaluation are presented to illustrate the superiority of the proposed method.

The main contributions can be summarized as follows:

- We introduce a novel bi-directional dual-constrained top-ranking loss to simultaneously consider the cross-modality and intra-modality variations, which provides new insights to enhance the discriminability of the learned representation for VT-REID.
- We propose a bi-directional center-constrained top-ranking loss which incorporates two constraints into a single formula, achieving better performance.
- We achieve state-of-the-art performances on two public datasets, which provides a superior baseline in this research field. We have also systematically evaluated different network structure designs and loss functions for cross-modality person re-identification.

II. Related Work

A. Single-Modality Re-ID

Most existing Re-ID works are designed for single visible modality, where person images are captured by visible cameras [21], [22]. Current Re-ID methods can be roughly categorized into two aspects: feature representation learning [23]–[25] and distance metric learning [26]–[29]. Recently, end-to-end deep learning [30]–[33] has achieved increasing attention due to its superior performance. A detailed overview about person Re-ID in single visible modality can be found in [34]. Most of these techniques developed for single visible domain are unsuitable for the cross-modality person re-identification problem [11].

B. Multi-Modality Re-ID

Previously, several multi-modal fusion methods have been introduced to improve the performance of single visible modality person re-identification [35]. In particular, they try to integrate the visible and thermal domains [36] or RGB and depth information [6], [37]. The additional information provided by other spectral cameras (depth camera, thermal camera) helps to enrich the visual representation of the standard RGB images [38]. In addition, semantic attributes information is also aggregated with visible images to improve the person re-identification performance [39], [40]. Rather than

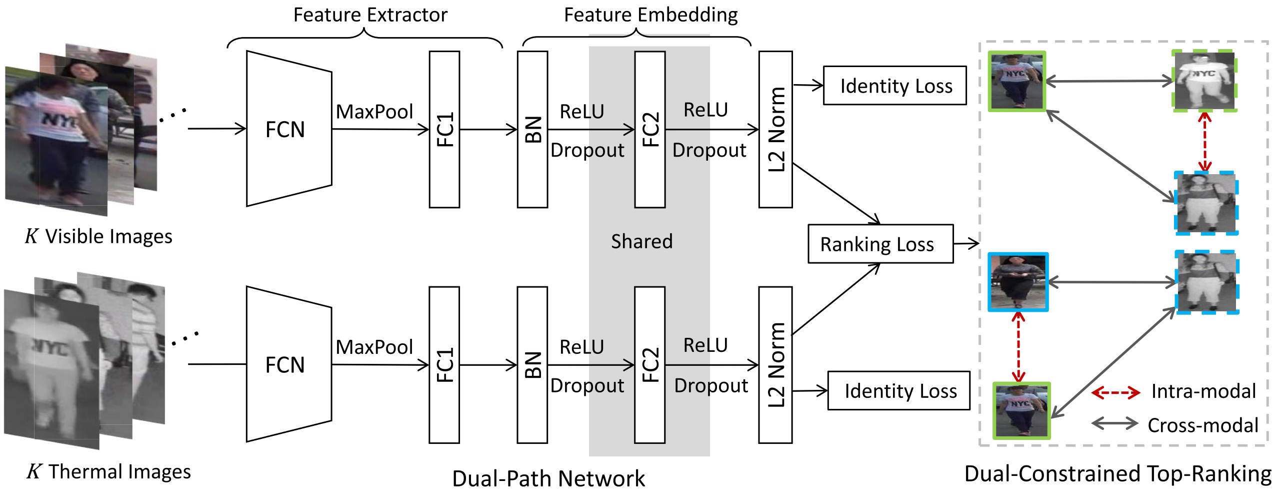


Fig. 4. The framework of the proposed method. In particular, K denotes the batch size, and totally $2*K$ images are fed into the network at each step. The framework contains two main components: dual-path network for feature extraction (one path for visible images and the other for thermal images) and learning objective with bi-directional dual-constrained top-ranking loss. The parameters in the shallow layers (*feature extractor*) are different to model the modality-specific information and the parameters of the deep embedding FC layer (*feature embedding*) are shared to learn multi-modality sharable features. A bi-directional dual-constrained top-ranking loss is introduced for network training after ℓ_2 normalization. Meanwhile, we combine the identity loss to stabilize the training process which also improves the performance.

multi-modality fusion, this paper addresses the cross-modality person re-identification problem.

For cross-modality person re-identification, several works have studied the text-to-image person retrieval problem [41]–[43]. However, their methods cannot be directly applied for VT-REID due to the data difference. For VT-REID, Ye *et al.* [18] introduced a two-stage framework to learn discriminate features and distance metrics in a sequential manner. In addition, Wu *et al.* [11] proposed a deep zero-padding network [44] for modality sharable feature representations learning. Recently, another cross-modality generative adversarial network is presented in [17], which jointly discriminates the modalities and person identities.

C. Deep Cross-Modality Matching

Cross-modality matching problem has been widely studied in general text-to-image retrieval [45] and heterogenous face recognition [12], [13], [46]. In this section, we mainly discuss the cross-modality matching models with deep learning techniques due to their advantages in various vision tasks [47]–[49].

For heterogenous (NIR-VIS) face recognition, Wu *et al.* [50], [51] studied the invariant feature representation learning with deep neural network for NIR-VIS face recognition. Sarfra *et al.* [16] proposed a deep matching model by learning a two-layer non-linear mapping function on top of hand-crafted features. In comparison, the VT-REID task is a much more challenging problem because the visual appearance variations between the visible and thermal person images are much more significant than those of face images [11]. It results in much larger intra-class variations compared with the face recognition, which limits the performance of their proposed methods for VT-REID task.

For text-to-image retrieval, several end-to-end deep learning methods have been introduced on top of dual-path network. They usually contains one text CNN path and one image CNN path [45], [52]. The dual-path network bridges the modality gap between the visual images and text descriptions with partially sharable parameters. Following this direction, we introduce a dual-path learning framework to learn the feature representations for VT-REID. In particular, a bi-directional dual-constrained top-ranking loss is introduced to train the network, which addresses the cross-modality variations and intra-modality variations simultaneously.

III. Proposed Method

This paper proposes a dual-path end-to-end learning framework for VT-REID, which optimizes the feature representations and distance metrics in an end-to-end manner. The framework of the proposed method is shown in Fig. 4. In particular, the proposed method contains two main components: feature learning with dual-path network and metric learning with bi-directional dual-constrained top-ranking loss. On one hand, the parameters of the dual-path network are partially shared to model the multi-modality sharable information and independent parameters mine the modality-specific information. On the other hand, the dual-constrained top-ranking loss tries to learn a low-dimensional feature embedding which is discriminative to distinguish different persons identities from two heterogenous modalities. In addition, we further integrate the identity loss to learn identity invariant information, which stabilizes and accelerates the learning process.

A. Dual-Path Network

We firstly introduce the dual-path network structure for feature learning, which mainly contains two parts: feature extractor and feature embedding, as shown in Fig. 4. The former feature extractor focuses on modeling the modality-specific information for two heterogenous modalities. The latter feature embedding aims at learning a low-dimensional multi-modality sharable embedding space for cross-modality re-identification. The details are described in the following.

1) Feature Extractor: We utilize the off-the-shelf feature extractors designed for visible images to extract the features for two heterogenous modalities. In particular, we utilize the same network structure for both modalities, because both of them are designed to capture extract feature representation of the person images. However, the network parameters of two paths are optimized separately to capture the modality-specific information. The general image classification network parameters pre-trained on the large-scale ImageNet are adopted to initialize the feature extractor due to the limited training data in VT-REID task.

In particular, we use AlexNet [53] as the backbone network[2] in our model for both visible and thermal paths. The pre-trained five convolutional layers ($conv1 \sim conv5$) and one fully connected layer (with size of 4096) are used to initialize the network as feature extractor for fine-tuning. The rationale is that the shallow convolutional layers usually capture the low-level visual patterns which are shared across two modalities. Meanwhile, we add a batch normalization layer after the pre-trained FC layer. The main reason is that ImageNet contains 1,000 classes which contain "person", while person re-identification are all "person" images. The features of all person images would be collapsed into a small region of the feature space if without batch normalization. Empirically, the batch normalization significantly improves the performance for person re-identification.

2) Feature Embedding: To learn a discriminative low-dimensional embedding space between two heterogenous modalities, a shared fully connected layer is introduced after the dual-path feature extractor. Note that the network parameters of this fully connected layer are shared for two modalities to model the modality-sharable information. Experimental results demonstrate that shared structure improves the performance for cross-modality person re-identification. The shared layer acts as a projection function to project features from two different modalities into a common embedding space. We also add a L2 normalization layer for the output embedding features, which smooths the feature representation as done in [54]. To simplify the description, we denote the feature representation function as $\mathcal{F}_v(\cdot)$ for visible images and $\mathcal{F}_t(\cdot)$ for thermal images. Given an input visible image I_v and an input thermal image I_t, the corresponding features (x and z) are represented by

$$x = \mathcal{F}_v(I_v), \quad z = \mathcal{F}_t(I_t) \tag{1}$$

B. Bi-Directional Dual-Constrained Top-Ranking

In this subsection, we introduce the designed learning objective function, namely bi-directional dual-constrained top-ranking loss for the feature embedding learning. The learning objective mainly contains the cross-modality and intra-modality constraints to handle both cross-modality and intra-modality variations as shown in Fig. 5. Firstly, we will revisit the general ranking loss.[3]

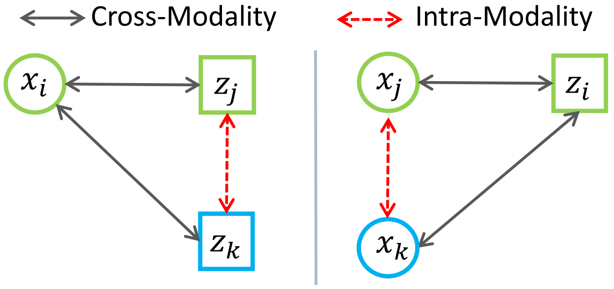


Fig. 5. Illustration of the bi-directional dual-constrained top-ranking loss with cross-modality and intra-modality constraint. Rectangles denote the thermal domain while circles denote the visible domain. The color represents the person identity. Left: visible-thermal top-ranking loss. Right: thermal-visible top-ranking loss.

1) Ranking Loss Revisit: Given a training batch, it contains K visible and K thermal images. For an anchor visible image x_i with annotated label y_i, the distance of its positive thermal image z_j should be smaller than the distance between x_i and the negative thermal image z_k by a pre-defined margin ρ_1:

$$D(x_i, z_j) < D(x_i, z_k) - \rho_1, \quad \forall y_i \neq y_k, \ \forall y_i = y_j, \tag{2}$$

Note that all the input feature vectors x and z are ℓ_2 normalized for stable convergence. $D(\cdot)$ represents the squared Euclidean distance, which is represented by

$$D(x_i, z_j) = \frac{1}{2}\|x_i - z_j\|_2^2. \tag{3}$$

For cross-modality person re-identification, we further employ a bi-directional training strategy to constrain the overall learning process. In particular, the bi-directional ranking loss considers two kinds of relationships: *visible to thermal relationship* (one anchor visible image, two thermal images) and *thermal to visible relationship* (one anchor thermal image, two visible images). The bi-directional ranking loss is represented by

$$\begin{aligned}\mathcal{L}_{bi} = &\sum_{\forall y_i = y_j, y_i \neq y_k} \max[\rho_1 + D(x_i, z_j) - D(x_i, z_k), 0] \\ &+ \sum_{\forall y_i = y_j, y_i \neq y_k} \max[\rho_1 + D(z_i, x_j) - D(z_i, x_k), 0]\end{aligned} \tag{4}$$

where the subscripts i and j represent the same person identity, while i and k are different person identities.

2) Cross-Modality Top-Ranking Constraint: To deal with the problem that large amounts of intra-class distances are larger than the inter-class distances caused by cross-modality variations, we use a top-ranking constraint [28] with online batch mining to improve the discriminability. The basic idea is that we compare the distance of a positive visible-thermal pair and the minimum distance of all related negative visible-thermal pairs within each training batch. The cross-modality constrained top-ranking loss is defined by:

$$\begin{aligned}\mathcal{L}_{cross} = &\sum_{\forall y_i = y_j} \max[\rho_1 + D(x_i, z_j) - \min_{\forall y_i \neq y_k} D(x_i, z_k), 0] \\ &+ \sum_{\forall y_i = y_j} \max[\rho_1 + D(z_i, x_j) - \min_{\forall y_i \neq y_k} D(z_i, x_k), 0]\end{aligned} \tag{5}$$

[2] Other networks such as the VggNet, GoogLeNet and ResNet architectures can also be configured without any limitation.

[3] It is also called triplet loss in [28].

The bi-directional cross-modality top-ranking loss has two main advantages: (1) The top-ranking constraint reduces the cross-modality variations while guarantees high discriminability with hard mining. It does not need off-line hard triplet sampling, and all the triplet comparison are conducted online within the training batch. (2) The bi-directional training strategy ensures that the learned feature representation is modality invariant. It improves the robustness for different query settings (*i.e.*, *visible to thermal and thermal to visible*) as verified in Sec. IV-C.

3) Intra-Modality Top-Ranking Constraint: As discussed in Section I, the VT-REID task also suffers from large intra-modality variations caused by camera environment changes, pose and viewpoint difference, etc. Motivated by contrastive loss, we introduce another intra-modality similarity constraint to address this issue. The main idea is that, for each identity i within the training batch, the distance between its positive sample j and negative sample k within each modality should be larger than a pre-defined margin ρ_2. On top of the sampled batch in cross-modality top-ranking loss, the intra-modality constrained top-ranking loss is computed by

$$\mathcal{L}_{intra} = \sum \max[\rho_2 - \min_{\forall y_i \neq y_k} D(z_j, z_k), 0] + \sum \max[\rho_2 - \min_{\forall y_i \neq y_k} D(x_j, x_k), 0] \quad (6)$$

where ρ_2 is a pre-defined margin. j and k represent the index which represent different person identities within the training batch for each anchor i. The intra-modality constraint guarantees that the images of different persons within each modality should also be distinguishable with a margin parameter. It works especially when the inequality criteria Eq. 2 does not hold for large-scale training. Therefore, it enhances the robustness of the learned feature representation to intra-modality variations.

4) Identity Loss: Above dual-constrained ranking loss aims at learning discriminative features with the underly relationship among different person identities. However, the features of visible and thermal person images might be totally different due to cross-modality variations. Then the ranking loss would be trapped into convergence problem due to incorrect relationship measurements, it is hard to converge for large-scale dataset. Meanwhile, the learnt feature representation cannot address the intra-class variations by simply using the relationship information. Therefore, we integrate the identity information to the overall loss function by treating each person identity as a class. In particular, the widely used softmax cross-entropy loss in image classification is adopted. In this manner, the identity loss ($\mathcal{L}_{id}$) models the identity specific information to enhance the robustness of the feature learning process.

The overall learning objective is a weighted sum of three components, the bi-directional cross-modality and intra-modality top-ranking constraints and identity loss, which is represented by

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cross} + \lambda_1 \mathcal{L}_{intra} + \lambda_2 \mathcal{L}_{id} \quad (7)$$

where λ_1 and λ_2 are pre-defined weighting parameters to combine the ranking loss and identity loss.

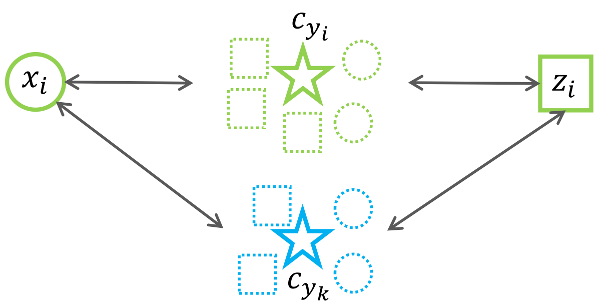


Fig. 6. Illustration of the bi-directional center-constrained top-ranking loss. We compare the *sample to center similarity* rather than *sample to sample similarity*. Color demonstrates the identity. Left: visible-thermal top-ranking loss. Right: thermal-visible top-ranking loss.

C. Bi-Directional Center-Constrained Top-Ranking

In this subsection, we introduce a bi-directional center-constrained top-ranking loss, which combines the cross-modality and intra-modality constraint into a single formula. The basic idea is to replace the comparison of *anchor to samples* by *anchor to centers* [55]. An illustration of the refined bi-directional center-constrained top-ranking loss is shown in Fig. 6. Meanwhile, it reduces the forward computational cost by half in each training batch at each step.

We firstly review the original center loss in [56], which was firstly introduced in face verification to learn a center for the features of each class. It aims at compacting the deep features of the same class to their corresponding center. Mathematically, the center loss is formulated by:

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^{K} D(x_i, c_{y_i}) \quad (8)$$

where $c_{y_i} \in \mathcal{R}^d$ is a d-dimensional vector representing the center of class y_i, and K is the batch size. For the cross-modality re-identification problem, we also assume that each person identity with label y_i should also be compacted to the center c_{y_i} in both heterogenous modalities. Under this situation, we define the bi-directional center-constrained top-ranking loss by

$$\mathcal{L}_{bicenter} = \sum_{i=1}^{K} \max[\rho_1 + D(x_i, c_{y_i}) - \min_{\forall y_i \neq y_k} D(x_i, c_{y_k}), 0] + \sum_{i=1}^{K} \max[\rho_1 + D(z_i, c_{y_i}) - \min_{\forall y_i \neq y_k} D(z_i, c_{y_k}), 0] \quad (9)$$

1) Discussion: In center-constrained top-ranking loss, we replace the comparison of *anchor to samples* by *anchor to centers*. The modification has two major advantages: 1) It reduces the computational cost. For an input training batch with K visible and K thermal images, the bi-directional dual-constrained top-ranking loss requires to calculate pairwise distance $2 * K^2$ for cross-modality constraint and $2 * K^2$ for intra-modality constraint. Meanwhile, it needs four times ranking operation to select the top-ranked samples. In comparison, the center-constrained top-ranking only need to compute the pairwise distance between the samples and the centers ($2 * K^2$) for each training batch at each step. It reduces the computational cost compared to the dual-constrained top-ranking loss. 2) It preserves the property to handle both cross-modality and intra-modality variations. On one hand, the triplet comparison with the centers ensures that the person identities from different modalities are distinguishable. On the other hand, the center itself constrains the samples belong to the same identity either from visible modality or thermal modality are concentrated, which addresses the intra-modality variations simultaneously.

2) Backward Propagation Analysis: To calculate the back-propagation gradients of the input feature representations from two heterogenous modalities and their corresponding centers in the current step, we firstly simplify the representation of the center-constrained ranking loss. The minimum sample to inter-class center distance $\min_{\forall y_i \neq y_k} D(x_i, c_{y_k})$ and $\min_{\forall y_i \neq y_k} D(z_i, c_{y_k})$ is calculated by $D(x_i, c_{p_i})$ and $D(z_i, c_{q_i})$, where p_i and q_i represent the class label of the centers with minimum distance. Therefore, the center-constrained ranking loss for visible sample x_i and thermal sample z_i is formulated by

$$\tilde{\mathcal{L}}_{x_i} = \max[\rho_1 + D(x_i, c_{y_i}) - D(x_i, c_{p_i}), 0] \qquad (10)$$

$$\tilde{\mathcal{L}}_{z_i} = \max[\rho_1 + D(z_i, c_{y_i}) - D(z_i, c_{q_i}), 0] \qquad (11)$$

Therefore, the gradients of the center-constraint in Eq. 9 with respect to x_i and z_i are computed by:

$$\begin{aligned} \frac{\partial \mathcal{L}_{bicenter}}{\partial x_i} &= (\frac{\partial D(x_{\rm i}, c_{y_i})}{\partial x_i} - \frac{\partial D(x_{\rm i}, c_{p_i})}{\partial x_i}) \delta(\tilde{\mathcal{L}}_{x_i} > 0) \\ &= (c_{p_i} - c_{y_i}) \delta(\tilde{\mathcal{L}}_{x_i} > 0) \qquad (12) \\ \frac{\partial \mathcal{L}_{bicenter}}{\partial z_i} &= (\frac{\partial D(z_{\rm i}, c_{y_i})}{\partial z_i} - \frac{\partial D(x_{\rm i}, c_{q_i})}{\partial z_i}) \delta(\tilde{\mathcal{L}}_{z_i} > 0) \\ &= (c_{q_i} - c_{y_i}) \delta(\tilde{\mathcal{L}}_{z_i} > 0) \qquad (13) \end{aligned}$$

where $\delta(condition)$ is an indicator function, *i.e.*, $\delta(condition) = 1$ when $condition$ is satisfied and $\delta(condition) = 0$ otherwise. Within each training batch, the gradient of c_j is calculated by

$$\begin{aligned} \frac{\partial L_{bicenter}}{\partial c_j} = & \frac{\sum_{i=1}^{K} (x_i - c_j) \delta(\tilde{\mathcal{L}}_{x_i} > 0) \delta(y_i = j)}{1 + \sum_{i=1}^{K} \delta(\tilde{\mathcal{L}}_{x_i} > 0) \delta(y_i = j)} \\ & - \frac{\sum_{i=1}^{K} (x_i - c_j) \delta(\tilde{\mathcal{L}}_{x_i} > 0) \delta(p_i = j)}{1 + \sum_{i=1}^{K} \delta(\tilde{\mathcal{L}}_{x_i} > 0) \delta(p_i = j)} \\ & + \frac{\sum_{i=1}^{K} (z_i - c_j) \delta(\tilde{\mathcal{L}}_{z_i} > 0) \delta(y_i = j)}{1 + \sum_{i=1}^{K} \delta(\tilde{\mathcal{L}}_{z_i} > 0) \delta(y_i = j)} \\ & - \frac{\sum_{i=1}^{K} (z_i - c_j) \delta(\tilde{\mathcal{L}}_{z_i} > 0) \delta(q_i = j)}{1 + \sum_{i=1}^{K} \delta(\tilde{\mathcal{L}}_{z_i} > 0) \delta(q_i = j)} \qquad (14) \end{aligned}$$

where K represents the number of visible/thermal images in each training batch.

We adopt the Stochastic Gradient Descent (SGD) with the learning rate α to update the centers. Specifically, the updating procedure of c_j is denoted by

$$c_j^{t+1} = c_j^t - \alpha \Delta c_j^t \qquad (15)$$

The overall learning objective function is the combination of bi-directional center-constrained top-ranking loss and identity loss, which is represented by

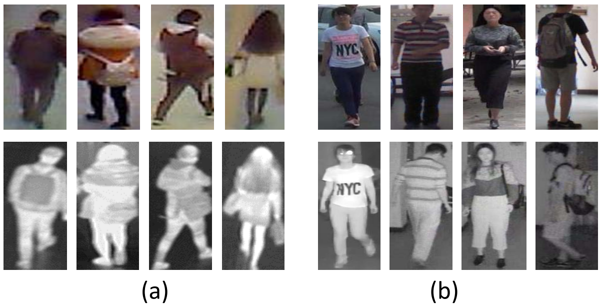


Fig. 7. Example images from (a) RegDB [36] and (b) SYSU-MM01 [11] datasets.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{bicenter} + \lambda_2 \mathcal{L}_{id} \qquad (16)$$

where λ_1 and λ_2 are pre-defined weighting parameters to combine the ranking loss and identity loss.

D. Online Batch Sampling Training Strategy

This subsection introduces the online batch sampling training strategy for cross-modality person re-identification. In particular, n person identities are firstly randomly selected at each iteration, and then k visible and k thermal images are randomly selected for each sampled identity. In total, each training batch contains $n * k$ visible images and $n * k$ thermal images. And totally $2 * n * k$ images are fed into the network for training at each iteration. To compute the dual-constrained top-ranking loss, we mine the top-ranked visible-to-thermal triplet and thermal-to-visible triplet to calculate Eq. 5 and Eq. 6 online within the training batch. This strategy could fully utilize the relationship between all the sample within the batch [57]. It is different from general triplet network, which needs to construct the triplets offline. Due to the randomly sampling mechanism, all the possible assemblies will be traversed to get the global optimal solution.

IV. Experimental Results

A. Experimental Settings

1) Datasets and Settings: We adopt two publicly available cross-modality person re-identification datasets (RegDB [36] and SYSU-MM01 datasets [11]) for evaluation. Some example images from two datasets are shown in Fig. 7.

RegDB dataset [36] is collected by dual camera systems. It contains 412 person identities from two different cameras (one visible camera and one thermal camera). For each person identity, it contains 10 different visible light images and 10 different thermal images. We follow the evaluation protocol described in [18], where the dataset is randomly divided into two halves, one for training and one for testing. For testing, the images from one modality (thermal domain) are used as the gallery set while the ones from the other modality (visible domain) as the probe set. This procedure is repeated for 10 times and the average performance is reported.

SYSU-MM01 dataset [11] is a large-scale cross-modality dataset collected in a campus environment. It is collected by 6 cameras, including four RGB (visible) cameras and two near-infrared (thermal) cameras. It is challenging because some of

TABLE I

COMPARISON WITH BASELINE METHODS INCLUDING DIFFERENT NETWORK STRUCTURES AND DIFFERENT LOSS FUNCTIONS AS LEARNING OBJECTIVES. RE-IDENTIFICATION RATES (%) AT RANK r AND mAP (%)

Datasets		RegDB				SYSU-MM01 (*Single-Shot All Search*)			
Strucutre	Loss Type	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
Evaluation of Different Network Structures Designs									
Two Paths	Softmax	1.56	8.02	10.42	3.16	2.47	16.42	29.41	3.92
Single-Path	Softmax	24.74	47.82	57.70	22.48	15.64	53.23	71.96	16.72
Dual-Path (S)	Softmax	25.72	50.23	60.86	23.36	16.13	54.64	72.46	17.79
Dual-Path (D)	Softmax	26.24	51.75	62.08	23.74	16.42	55.62	73.56	18.10
Evaluation of Different Loss Functions									
Dual-Path (D)	Softmax	26.24	51.75	62.08	23.74	16.42	55.62	73.56	18.10
Dual-Path (D)	Softmax + Center	32.65	56.12	66.94	30.94	17.04	56.73	73.98	18.65
Dual-Path (D)	Triplet	26.44	49.01	58.79	24.11	16.80	57.24	74.32	19.03
Dual-Path (D)	Triplet (Hard)	31.96	56.54	67.24	30.64	17.68	59.22	76.35	20.58
Dual-Path (D)	BiTriplet	32.62	57.32	66.86	31.03	18.32	60.12	77.11	21.07
Dual-Path (D)	BDTR	33.56	58.61	67.43	32.76	20.84	63.81	79.14	22.86
Dual-Path (D)	eBDTR	34.62	58.96	68.72	33.46	22.42	64.61	78.62	24.11

the person images are captured in the indoor environments and some of them are in outdoor environments. Specifically, it contains 491 persons, each person is captured by at least two different modality cameras. The single-shot *all-search* mode evaluation protocol is adopted for evaluation by default, since it is the most challenging setting as described in [11]. This dataset contains 395 person identities for training, which contains 22,258 visible images and 11,909 thermal images. The remaining 96 persons are adopted for testing, which contains 3803 thermal images for query and 301 randomly selected visible images as gallery set. The selection of gallery set is also conducted for 10 times following [11].

2) Evaluation Metrics : To evaluate the person re-identification performance, the standard cumulated matching characteristics (CMC) and mean average precision (mAP) are adopted. CMC measures the probability of the correct match occurs at the top-k retrieved results while mAP measures the retrieval performance of all correct samples since one query person may have multiple groundtruths in the gallery set. Since all the output features are ℓ_2 normalized, we directly calculate the cosine similarity with inner products for testing, which is quite efficient and suitable for real applications.

3) Implementation Details: We adopt the AlexNet as the backbone. The network parameters are initialized with pre-trained model on ImageNet. The dimension of the embedding fully connected layer is set as 512 and the training batch size is set as 32 for both datasets. Each batch contains 32 randomly selected person identities. Dropout rate is set as 0.5. Random cropping is utilized for data argumentation, where images are firstly resized to 256 × 256, and then a random cropped 227 × 227 image is fed into the network. We set the trade-off parameters $\lambda_1 = 1$ and $\lambda_2 = 0.1$ on RegDB dataset, and $\lambda_1 = 0.1$ and $\lambda_2 = 1$ on the SYSU-MM01 dataset. SGD optimizer is utilized for optimization, and the momentum is set to 0.9. The initial learning rate is set as 0.001 on RegDB dataset and 0.01 on SYSU dataset. The learning rate is decayed by 0.1 when the loss is plateaued. The predefined cross-modality margin ρ_1 is set to 0.5 while the intra-modality margin ρ_2 is set to 0.1. For the center-constrained top-ranking, the margin is also set to 0.5. The center updating parameter α is firstly set to 0.1 and decayed by 0.1 after every 40 epochs.

B. Comparison With Baseline Methods

In this subsection, we compare the proposed method with baseline methods in two different aspects: *Different Network Structure Designs* and *Different Loss Functions*. The experimental results on RegDB and the large-scale SYSU-MM01 datasets are listed in Table I.

1) Different Network Structure Designs: Four different network structures designs are included for comparison: 1) *Two Paths*, we learn the features of two modalities separately with two independent paths; 2) *Single-Path*, we learn the features of two modalities together in a single path network without considering the modality difference; 3) *Dual-Path (S)*, we learn the features with dual-path network, while the parameters of the shallow layers are shared, and the parameters of the fully connected layers are independent for two different modalities; 3) *Dual-Path (D)*, we learn the features with the proposed dual-path network structure, *i.e.*, the parameters of the shallow layers are specific and the parameters of fully connected layers are shared for two different modalities. An illustration is shown in Fig. 8. We adopt the general softmax loss as the learning objectives by treating each person identity as a class (also known as identity loss in Re-ID tasks).

Results shown in Table I demonstrate that the proposed *Dual-Path (D)* achieves the best performance on both datasets compared to other network structure designs. Note that two separate paths network would learn good feature representation for each modality, but it totally fails in this cross-modality re-identification task. It demonstrates that the modality-specific parameters in shallow layers is better than shared parameters for VT-REID. Compared to single-path network, dual-path

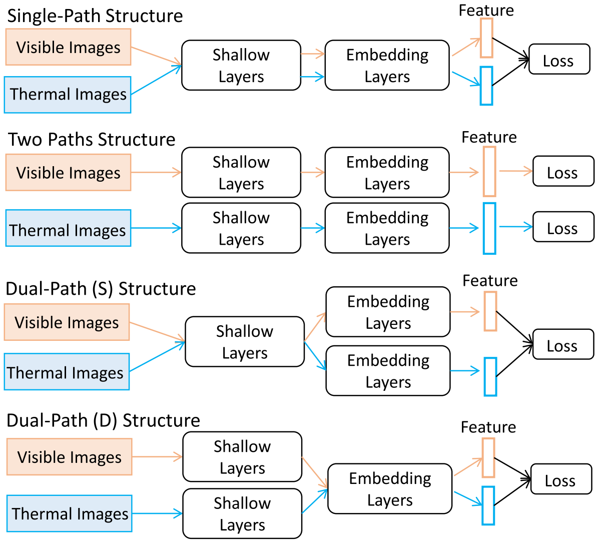


Fig. 8. Illustration of different network structure designs.

network also consistently achieves better performance by modeling the modality-specific information.

2) Different Loss Functions: We also conduct the experiments with different baseline loss functions to verify the effectiveness of the proposed method. Five different loss functions are evaluated: Softmax loss, Softmax loss combined with center loss, triplet loss, triplet loss with hard mining and bi-directional triplet loss with hard mining. Note that the baseline network structure is Dual-Path (D) as described in Section III-A.

We can observe that the proposed method achieves the best performance compared to other baseline loss functions. The triplet loss with hard mining performs much better than the general triplet loss without hard mining, which demonstrate the importance of hard mining strategy for cross-modality person re-identification. Meanwhile, bi-directional training strategy can also improve the performance compared to general triplet loss with hard mining. The proposed bi-directional dual-constrained top-ranking loss (BDTR) improves the performance further with additional intra-modality constraint. After being extended with center-constrained top-ranking loss (eBDTR), the re-identification performance on two datasets is consistently improved. This experiment verifies that eBDTR preserves the discriminability by considering both cross-modality and intra-modality constraints into a single formula.

C. Comparison With the State-of-the-Arts

Competing methods. Only three published works have studied the cross-modality person re-identification in visible and thermal domains. Their methods in their original papers are only evaluated on one dataset.

- **Zero-Padding** [11]. A one-stream network with deep zero-padding to capture the domain-specific information to learn the feature representations. We re-implement their method and evaluate it on the RegDB dataset.
- **TONE + HCML** [18]. A two-stage framework which contains two separate parts: feature learning (TONE) and metric learning (HCML). We adopt the authors' released code to evaluate this method on SYSU-MM01 dataset.
- **cmGAN** [17]. A cross-modality generative adversarial network is proposed to jointly discriminate the modalities and person identities.

In addition, we also list several cross-modality learning methods for comparison. Most of the results are taken from [18] on the RegDB dataset and [11] on the SYSU-MM01 dataset. In particular, some feature extraction methods (HOG, MLBP, LOMO [23]) are included for comparison. In addition, some matching model learning methods (XQDA [23], MLAPG [59], GSM [58] SCDL [60] and rCDL [61]) are also included for comparison. We also compare state-of-the-art methods (SVDNet [30] and PCB [31]) which are originally designed for single modality person Re-ID. The results on two datasets are shown in Table II.

Results shown in Table II demonstrate that the proposed method achieves the best or at least comparable performance on two datasets. On the one hand, we can observe that the state-of-the-art methods (SVDNet [30] and PCB [31]) designed for single modality with visible images are unsatiable for the cross-modality person re-identification problem. Their performance drops dramatically for VT-REID since they do not consider the modality difference between two modalities. On the other hand, when compared to three state-of-the-art cross-modality person re-identification methods, the proposed method also performs better in different aspects. In particular, our proposed method also outperforms the two-stage learning method (TONE + HCML) [18] usually by a large margin for both rank-1 accuracy and mAP on two datasets. Compared to the state-of-the-art deep zero-padding [11], the proposed method is also a clear winner. We achieve comparable performance with cmGAN [17] on the SYSU-MM01 dataset with the same ResNet50 baseline, but this method needs *more than 2,000* epochs to converge due to the adversarial training process, which is unsuitable for large-scale training. In comparison, the proposed method can get comparable performance with *less than 100* epochs, which is much more efficient.

In terms of the different performances of AlexNet and ResNet50 on two datsets as shown in Table II, we have the following analysis. Compared to SYSU-MM01 dataset, thermal images of RegDB dataset are less similar to the visible ones (Fig. 7). Therefore, the less deep AlexNet (7 layers) might be able to more easily adapt the pre-trained ImageNet weights to the thermal input. Secondly, another possible reason is that the size of the RegDB dataset is much smaller than the that SYSU-MM01 dataset (4K vs. 34K), and it is easier to adapt the pre-trained ImageNet weights on small RegDB dataset for less deep AlexNet. In comparison, we have much more data for the SYSU-MM01 dataset to fine-tune the deeper ResNet50 (50 layers).

D. Further Evaluation and Analysis

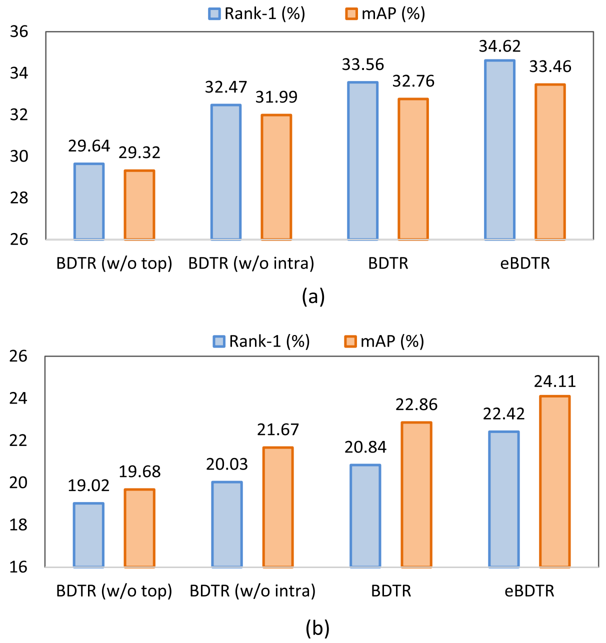
1) Ablation Study: This subsection evaluates the proposed method with different variants to verify the effectiveness of each component. In particular, four different variants are evaluated. *BDTR (w/o top)* represents the proposed method without

TABLE II

COMPARISON WITH THE STATE-OF-THE ARTS ON TWO DATASETS. RE-IDENTIFICATION RATES (%) AT RANK r AND MAP (%)

Datasets	RegDB				SYSU-MM01 (*Single-Shot All Search*)			
Methods	$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
HOG	13.49	33.22	43.66	10.31	2.76	18.25	31.91	4.24
MLBP	2.02	7.33	10.90	6.77	2.12	16.23	28.32	3.86
LOMO [23]	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
GSM [58]	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
SVDNet [30]	17.24	34.12	44.51	19.04	14.64	53.28	64.24	15.17
PCB [31]	18.32	36.42	46.51	20.13	16.43	54.06	65.24	16.26
TONE [18]	16.87	34.03	44.10	14.92	12.52	50.72	68.60	14.42
TONE + XQDA	21.94	45.05	55.73	21.80	14.01	52.78	69.06	15.97
TONE + MLAPG	17.82	40.29	49.73	18.03	12.43	50.64	68.72	14.61
TONE + SCDL	8.06	22.09	28.89	10.03	6.58	35.62	56.32	10.32
TONE + rCDL	9.47	22.96	29.42	10.26	7.02	37.31	57.64	10.46
TONE + HCML	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
Zero-Padding [11]	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
cmGAN [17]†	-	-	-	-	26.97	**67.51**	80.56	**27.80**
BDTR (AlexNet)	**33.56**	**58.61**	**67.43**	32.76	20.84	63.81	79.14	22.86
eBDTR (AlexNet)	**34.62**	**58.96**	**68.72**	**33.46**	22.42	64.61	78.62	24.11
BDTR (ResNet50)	30.56	54.62	65.42	32.45	**27.32**	66.96	**81.07**	27.32
eBDTR (ResNet50)	31.83	56.12	66.81	**33.18**	**27.82**	**67.34**	**81.34**	**28.42**

† cmGAN [17] is originally configured with ResNet50 [47] as the backbone. It trains more than 2,000 epochs to converge while the proposed method is usually less than 100 epochs. For ResNet50, we resize the image to 384 × 128 as suggested in [31].



(a)

(b)

Fig. 9. Evaluation of different variants of proposed method on two datasets. Re-identification rates (%) at rank 1 and mAP (%). (a) Results on the RegDB dataset. (b) Results on the SYSU-MM01 dataset.

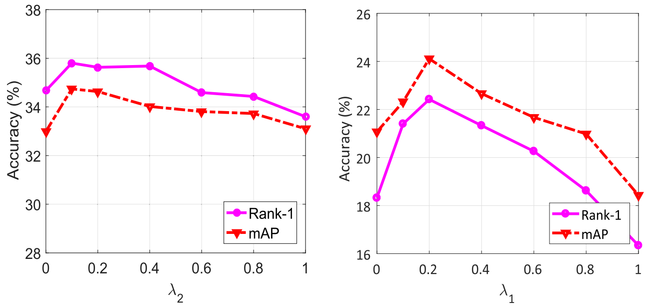


Fig. 10. Evaluation of the weighting parameter λ_1 and λ_2 on two datasets (left: $\lambda_1 = 1$ and $\lambda_2 \in [0, 1]$ for RegDB dataset, right: $\lambda_2 = 1$ and $\lambda_1 \in [0, 1]$ for SYSU-MM01 dataset). Rank-1 matching accuracy (%) and mAP (%) are reported.

top-ranking (hard mining) constraint. *BDTR (w/o intra)* means the results without intra-modality constraint. *BDTR* denotes the performance with bi-directional dual-constrained top-ranking loss. *eBDTR* is the extended version of BDTR with center-constrained bi-directional top-ranking loss. The results on the RegDB and SYSU-MM01 datasets are shown in Fig. 9. Note that the baseline network is AlexNet.

Results shown in Fig. 9 demonstrate that the performance would be decreased by about 3% for both rank-1 accuracy and mAP if without the cross-modality top-ranking constraint. This experiment verifies that the top-ranking constraint increases the discriminability of the learned feature representation by handling the large cross-modality variations. If without the intra-modality constraint, the performance will drop 1-2% on both datasets. This observation demonstrates that the intra-modality constraint which addresses the intra-modality variations would improve the feature learning consistently. Another observation is that the extended eBDTR performs slightly better than the proposed BDTR. eBDTR preserves the properties of handling cross-modality and intra-modality variations simultaneously but it has less constraint.

2) Weighting Parameters: This subsection evaluates the weighting parameter λ_1 and λ_2 of the identity loss and the ranking loss. In particular, λ_1 is set to 1 on the small-scale RegDB dataset, and we adjust the weighting parameter $\lambda_2 \in [0, 1]$. And λ_2 is set to 1 on the large-scale SYSU-MM01 dataset, and we adjust the weighting parameter $\lambda_1 \in [0, 1]$. The results on two datasets are shown in Fig. 10.

TABLE III
DIFFERENT MARGIN PARAMETERS ON TWO DATASETS. RE-IDENTIFICATION RATES (%) AT RANK 1 AND MAP (%)

Datasets		RegDB		SYSU-MM01	
ρ_1	ρ_2	Rank 1	mAP	Rank 1	mAP
0.3	0.1	33.82	32.96	21.96	23.96
0.5	0.1	33.62	33.46	22.42	24.11
1	0.1	31.68	31.38	18.64	20.32
0.5	0.1	33.62	33.46	22.42	24.11
0.5	0.3	34.02	33.98	22.14	24.06
0.5	0.5	33.96	34.06	21.86	23.96

As demonstrated in Fig. 10, we could observe that integrating identity loss could consistently improve the performance of cross-modality person re-identification. In general, the identity loss can achieve better performance on the large-scale dataset with enough training samples while the ranking loss usually performs better with limited samples on small datasets. This phenomenon has also been verified in cross-view re-ID [62]. Therefore, for the small-scale RegDB dataset, we assign a small value for the identity loss, where the weighting parameter λ_2 is set to 0.1 in our experiments. A larger λ_2 will damage the feature learning process, and the performance drops dramatically. For the large-scale-SYSU-MM01 dataset, we assign a small value λ_2 for the ranking loss. We could observe that a proper $\lambda_2 \in [0, 0.5]$ improves the performance, while a large λ_2 would damage the performance. Note that the ranking loss training directly from the pre-trained ImageNet model is hard to converge, where we need to use the pre-trained parameters with identity loss for initialization, it further shows the importance of the identity loss for cross-modality person re-identification.

3) Margin Parameter: It is commonly recognized that the ranking loss usually suffers from the convergence issue for different applications. This subsection evaluates the margin parameters ρ_1 and ρ_2 in the proposed ranking loss. In particular, three different parameters are evaluated, 0.3, 0.5 and 1 for ρ_1, 0.1 0.3 and 0.5 for ρ_2. The results on two datasets are shown in Table III. We could observe that the margin parameter is also quite important for cross-modality person re-identification. In particular, a small margin parameter for ρ_1 usually performs good on both dataset. The main reason is that we adopt the normalized feature with Euclidean distance, if we use too large margin (e.g., 1), the inequality constraint is Eq. 2 is hard to satisfy, which makes the network hard to optimize. We select a relative small value for ρ_2 for fast convergence. Empirically, we choose 0.5 as the margin parameter ρ_1 in our method.

4) Indoor Search on SYSU-MM01 Dataset: In this subsection, we also evaluate the proposed method under the *indoor search* mode on the SYSU-MM01 dataset. In particular, the images of two outdoor cameras are excluded for the gallery set, and the same probe set from Cam 3 and Cam 6 is adopted. Detailed description about this evaluation protocol can be found in [11]. This protocol is less challenging than previous single-shot *all search* mode. We compare the competing methods as mentioned in Section IV-C. The results are shown in Table IV.

TABLE IV
RESULTS ON THE SYSU-MM01 DATASET WITH INDOOR-SEARCH SINGLE-SHOT MODE. RE-IDENTIFICATION RATES (%) AT RANK r AND MAP (%)

Methods	$r = 1$	$r = 10$	$r = 20$	mAP
HOG	3.22	24.68	44.52	7.25
MLBP	3.43	26.42	45.36	7.72
LOMO	2.24	22.53	41.53	6.64
GSM	9.46	48.98	72.06	15.57
One-stream	16.94	63.55	82.10	22.95
Two-stream	15.60	61.18	81.02	21.49
SVDNet [30]	20.24	64.32	83.62	28.74
PCB [31]	22.63	65.24	83.92	30.46
Zero-Padding [11]	20.58	68.38	85.79	26.92
TONE [18]	20.82	68.86	84.46	26.38
cmGAN [17]	31.63	77.23	89.18	42.19
BDTR (AlexNet)	26.83	73.26	87.68	37.68
eBDTR (AlexNet)	27.62	74.95	88.12	38.40
BDTR (ResNet50)	31.92	77.18	89.28	41.86
eBDTR (ResNet50)	**32.46**	**77.42**	**89.62**	**42.46**

Results shown in Table IV demonstrate that the proposed method still outperforms the competing baseline under this evaluation protocol. The advantages of the proposed method can be summarized as three folds: 1) End-to-end learning without any human intervention can obtain discriminative features. 2) The dual-constrained top-ranking loss on top of dual-path network provides a good solution to simultaneously consider the large cross-modality and intra-modality variations for VT-REID. Meanwhile, the extended version eBDTR also preserves the property of handling the two variations simultaneously. 3) Integrating the identity loss helps to stabilize the training process by modeling the identity-specific information. It also improves the discriminability of the learned feature representation of ranking loss.

5) Retrieved Examples: We show the top ten retrieved results of 5 randomly selected query examples on the SYSU-MM01 dataset in Fig. 11. The corresponding cosine similarity score is reported on top of each image. First, we could observe that it is even hard for human to tell which person is the correct matching of the query by simply using the color information, which makes it an extremely challenging task, but it plays an important role in night-time surveillance applications. Second, even there are some wrongly retrieved examples in the ranking list, but the top-ranked images are usually quite close to the query image due to similar texture or structure information.

6) Different Query Settings: We also test the performance under different query settings as done in [18]. Results on the RegDB dataset are shown in Table V. We can observe that the proposed method achieves close performance for visible-to-thermal matching and thermal-to-visible matching, where the difference is less than 1%. The rank-1 matching accuracy is about 34% and the mAP is about 32% on both settings.The robustness of the learned feature demonstrates

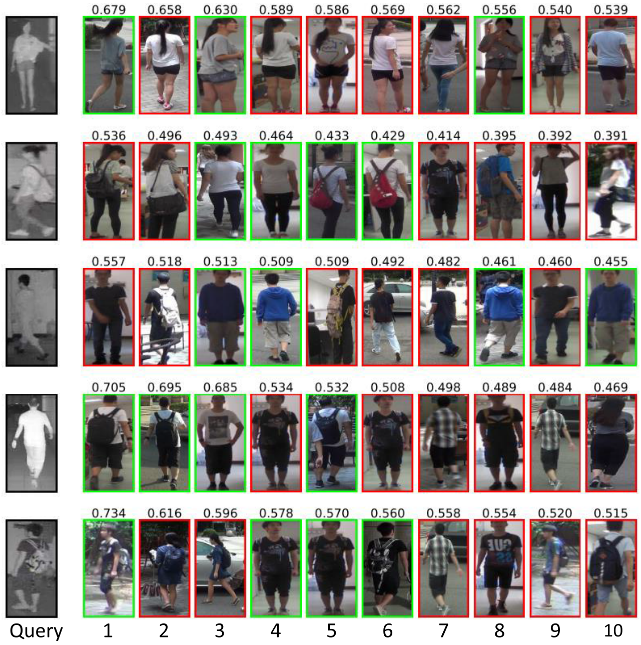


Fig. 11. The top-10 retrieved results of some example queries with the proposed method on the SYSU-MM01 dataset. The green bounding boxes indicate the correct matchings and red bounding boxes represent wrong matchings (best viewed in color.)

TABLE V

EVALUATION OF DIFFERENT QUERY SETTINGS ON THE REGDB DATASET. RE-IDENTIFICATION RATES (%) AT RANK r AND mAP (%)

Method	$r = 1$	$r = 10$	$r = 20$	mAP
Setting	*Visible to Thermal*			
TONE	16.87	34.03	44.10	14.92
TONE + HCML	24.44	47.53	56.78	20.08
Zero-Padding	17.75	34.21	44.35	18.90
BDTR	33.56	58.61	67.43	32.76
eBDTR	34.62	58.96	68.72	33.46
Setting	*Thermal to Visible*			
TONE	13.86	30.08	40.05	16.98
TONE + HCML	21.70	45.02	55.58	22.24
Zero-Padding	16.63	34.68	44.25	17.82
BDTR	32.92	58.46	68.43	31.96
eBDTR	34.21	58.74	68.64	32.49

the effectiveness of the bi-directional relationship training strategy, which enhances the modality invariant property of the feature learning. Meanwhile, we achieve the best performance compared to the competing methods on both settings. The robustness of the proposed method demonstrate that the proposed method suitable for practical night-time applications with different query settings.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose an end-to-end learning framework on top of dual-path network for visible thermal person re-identification (VT-REID). The network structure captures the modality-specific and modality-sharable information. To guide the training process, a novel dual-constrained top-ranking loss is introduced. The proposed learning objective simultaneously considers the intra-modality and cross-modality variations, which guarantees the discriminability of the learned feature representation for VT-REID. Meanwhile, we also adopt a bi-directional training strategy to enhance the robustness of the feature learning. We further show that integrating the identity loss is very important for cross-modality person re-identification which models the identity specific information. In addition, we present a variant with center-constrained top-ranking loss, which preserves the property to handle the intra-modality and cross-modality variations. Extensive experiments on two public cross-modality datasets demonstrate the superiority of the proposed method.

Since the performance of cross-modality person re-identification is quite low compared to the single modality person re-identification, there is a large space for improvement. How to transfer the state-of-the-art single modality re-ID methods to VT-REID is still an open issue. Meanwhile, labeling the thermal images is more difficult than labeling the visible images. Due to abundant labeled datasets in visible domain, it is interesting to investigate how to utilize the labeled datasets with visible images to improve the performance of VT-REID.

REFERENCES

[1] L. Zheng, Y. Yang, and Q. Tian, “SIFT meets CNN: A decade survey of instance retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[2] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, “Joint sparse representation and robust feature-level fusion for multi-cue visual tracking,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.

[3] X. Lan, S. Zhang, P. C. Yuen, and R. Chellappa, “Learning common and feature-specific patterns: A novel multiple-sparse-representation-based tracker,” *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2022–2037, Apr. 2018.

[4] Z. Wang *et al.*, “Person reidentification via discrepancy matrix and matrix metric,” *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[5] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, “Fast open-world person re-identification,” *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2286–2300, May 2018.

[6] 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, Jun. 2017.

[7] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, “Dynamic graph co-matching for unsupervised video-based person re-identification,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.

[8] J. Huo, Y. Gao, Y. Shi, W. Yang, and H. Yin, “Heterogeneous face recognition by margin-based cross-modality metric learning,” *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1814–1826, Jun. 2018.

[9] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, “Learning modality-consistency feature templates: A robust RGB-infrared tracking system,” *IEEE Trans. Ind. Electron.*, to be published.

[10] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, “Modality-correlation-aware sparse representation for RGB-infrared object tracking,” *Pattern Recognit. Lett.*, to be published.

[11] 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)*, Oct. 2017, pp. 5380–5389.

[12] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, “Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition,” *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.

[13] Y. Jin, J. Lu, and Q. Ruan, “Coupled discriminative feature learning for heterogeneous face recognition,” *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015.

[14] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, “Unconstrained face recognition: Identifying a person of interest from a media collection,” *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2144–2157, Dec. 2014.

[15] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[16] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for cross-modal face recognition," *Int. J. Comput. Vis.*, vol. 122, no. 2, pp. 426–438, 2017.

[17] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 677–683.

[18] 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.

[19] L. Wu, Y. Wang, X. Li, and J. Gao, "What-and-where to match: Deep spatially multiplicative integration networks for person re-identification," *Pattern Recognit.*, vol. 76, no. 3, pp. 727–738, 2018.

[20] 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.

[21] M. Ye *et al.*, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

[22] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.

[23] 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.

[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian descriptor for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1363–1372.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[26] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 717–732, Mar. 2018.

[27] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *Proc. ICCV*, Jun. 2017, pp. 5142–5150.

[28] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

[29] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[30] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNET for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.

[31] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," Tech. Rep., 2018.

[32] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[33] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 170–186.

[34] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[35] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[36] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[37] 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. Workshops (ECCVW)*, 2012, pp. 433–442.

[38] A. Møgelmose, C. Bahnsen, T. B. Moeslund, A. Clapés, and S. Escalera, "Tri-modal person re-identification with RGB, depth and thermal features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2013, pp. 301–307.

[39] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," 2017, *arXiv:1703.07220*. [Online]. Available: https://arxiv.org/abs/1703.07220

[40] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2275–2284.

[41] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5187–5196.

[42] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1890–1899.

[43] M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, and J. Liu, "Specific person retrieval via incomplete text description," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, 2015, pp. 547–550.

[44] 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, Feb. 2018.

[45] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Cross-modal deep variational hashing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4097–4105.

[46] G.-S. J. Hsu, Y.-L. Liu, H.-C. Peng, and P.-X. Wu, "RGB-D-based face reconstruction and recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2110–2118, Dec. 2014.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6210–6219.

[49] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Trans. Multimedia*, to be published.

[50] R. He, X. Wu, Z. Sun, and T. Tan, "Learning invariant deep representation for NIR-VIS face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 2000–2006.

[51] X. Wu, L. Song, R. He, and T. Tan, "Coupled deep learning for heterogeneous face recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1679–1686.

[52] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4107–4116.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[54] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3733–3742.

[55] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1945–1954.

[56] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 499–515.

[57] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[58] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.

[59] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3685–3693.

[60] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2216–2223.

[61] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2496–2503.

[62] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," 2017, *arXiv:1710.00478*. [Online]. Available: https://arxiv.org/abs/1710.00478

Mang Ye received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Hong Kong Baptist University. His research interests focus on multimedia content analysis and retrieval, computer vision, and pattern recognition.

Xiangyuan Lan received the B.Eng. degree in computer science and technology from the South China University of Technology, China, in 2012, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2016. He is currently a Research Assistant Professor with Hong Kong Baptist University. His current research interests include intelligent video surveillance and biometric security.

Zheng Wang (M'19) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, in 2017. He is currently a JSPS Fellowship Researcher with Shin'ichi Satoh's Lab, National Institute of Informatics, Japan. His research interests focus on person re-identification and instance search. He received the Best Paper Award at the 15th Pacific-Rim Conference on Multimedia (PCM) in 2014 and the 2017 ACM Wuhan Doctoral Dissertation Award.

Pong C. Yuen received the B.Sc. degree (Hons.) in electronic engineering from the City University of Hong Kong, in 1989, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, in 1993. He joined Hong Kong Baptist University in 1993, where he served as the Head of the Department of Computer Science from 2011 to 2017 and he is currently a Professor with the Department of Computer Science and also an Associate Dean with the Science Faculty.

He was a recipient of the University Fellowship to visit The University of Sydney in 1996. In 1998, he spent a six-month sabbatical leave with The University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland at College Park. From 2005 to 2006, he was a Visiting Professor with INRIA Rhone Alpes, France. From 2017 to 2018, he was a Visiting Faculty with ETH, Zürich. He was the Director of the Croucher Advanced Study Institute (ASI) on biometric authentication in 2004 and the Director of Croucher ASI on Biometric Security and Privacy in 2007. He has been serving as the Director for IAPR/IEEE Winter School on Biometrics since 2017. His current research interests include video surveillance, human face recognition, and biometric security and privacy.

Dr. Yuen serves as a member for the Hong Kong Research Grant Council Engineering Panel. He is a fellow of IAPR. He is an Editorial Board Member of *Pattern Recognition*, and a Senior Editor of the *SPIE Journal of Electronic Imaging*. He received the Outstanding Editorial Board Service Award in 2018. He has been actively involved in many international conferences and professional community. He was the Track Co-Chair of the International Conference on Pattern Recognition (ICPR) 2006, the Program Co-Chair of the IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS) 2012, the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA) 2016, and the International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI) 2018. He serves as the Program Co-Chair for the International Workshop on Information, Forensics and Security (WIFS) 2018. He served as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY from 2014 to 2018. He is currently the Vice President (Technical Activities) of the IEEE Biometrics Council.