Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking

Mang Ye[1], **Zheng Wang**[2], **Xiangyuan Lan**[1], **Pong C. Yuen**[1]
[1] Department of Computer Science, Hong Kong Baptist University, Hong Kong
[2] National Institute of Informatics, Japan
{mangye, pcyuen}@comp.hkbu.edu.hk, wangz@nii.ac.jp, xiangyuanlan@life.hkbu.edu.hk

Abstract

Cross-modality person re-identification between the thermal and visible domains is extremely important for night-time surveillance applications. Existing works in this filed mainly focus on learning sharable feature representations to handle the cross-modality discrepancies. However, besides the cross-modality discrepancy caused by different camera spectrums, visible thermal person re-identification also suffers from large cross-modality and intra-modality variations caused by different camera views and human poses. In this paper, we propose a dual-path network with a novel bi-directional dual-constrained top-ranking loss to learn discriminative feature representations. It is advantageous in two aspects: 1) end-to-end feature learning directly from the data without extra metric learning steps, 2) it simultaneously handles the cross-modality and intra-modality variations to ensure the discriminability of the learnt representations. Meanwhile, identity loss is further incorporated to model the identity-specific information to handle large intra-class variations. Extensive experiments on two datasets demonstrate the superior performance compared to the state-of-the-arts.

1 Introduction

Person re-identification (REID) aims at searching a specific person from a gallery of images captured by disjoint surveillance cameras [Zheng *et al.*, 2017; Ye *et al.*, 2016]. It has gained increasing attention in the research community due to its importance in various video surveillance and intelligent applications. Recent progresses mainly focus on visible cameras module, i.e., given a query image/video of a person and search it out from a gallery set of images/videos captured by other non-overlapping cameras [Zhu *et al.*, 2018; Wang *et al.*, 2017]. However, the visible cameras cannot capture valid appearance information under poor illumination environments (e.g. during the night), which limits the applicability in practical surveillance applications [Lan *et al.*, 2015; 2018; Jiang *et al.*, 2017]. Therefore, we address a cross-modality problem named visible thermal person re-identification (VT-REID) in this paper. It provides a good supplement for night-time surveillance applications.

Figure 1: Visible thermal person re-identification (VT-REID). Person images captured by different spectrum cameras should be matched. Note that the cross-modality discrepancy make the visual characteristic of images from two modalities are entirely different.

Given a visible (thermal) image of a specific person, VT-REID[1] tries to search out the corresponding thermal (visible) images from a gallery set captured by other spectrum cameras as illustrated in Figure 1. To our best knowledge, two pioneer works exist in the literature. Wu *et al.* [Wu *et al.*, 2017b] proposed a one-stream deep zero-padding network for shared feature learning where only identity information is utilized, which limits the discriminability of the learnt representation. Contemporarily, a two-stage framework containing feature learning and metric learning steps is introduced in [Ye *et al.*, 2018]. However, the two-stage training needs human intervention which is unsuitable for practical large-scale applications [Wu *et al.*, 2018b]. Therefore, we try to investigate an end-to-end learning framework to learn invariant shared features while preserving high discriminability for VT-REID.

In this paper, we propose a dual-path network to learn the feature representations for VT-REID, which contains a visible path and a thermal path. Specifically, the parameters of the shallow layers are independent to extract the modality-specific information, which addresses the cross-modality discrepancy problem caused by different sensor spectrums. Then a shared fully connected layer is further leveraged to learn the embedding space. Thus the multi-modality sharable feature is learnt by simultaneously considering the modality

[1] It's also called RGB-infrared person re-id in [Wu *et al.*, 2017b].

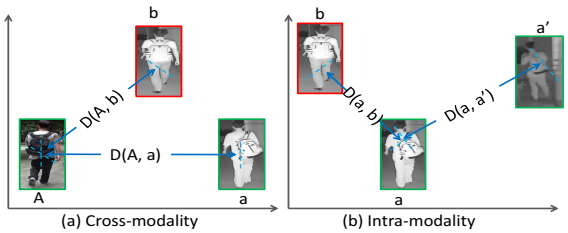

Figure 2: Intra-class distance might be even larger than the inter-class distance caused by (a) Cross-modality variation and (b) Intra-modality variation. The box color represents the identity.

commonality and discrepancy. However, besides the cross-modality discrepancy issue, VT-REID also suffers from 1) large cross-modality variations caused by the different cross-camera views and hard negatives ($D(A,a)>D(A,b)$ in Figure 2) and 2) large intra-modality variations caused by different human poses and viewpoints($D(a,a')>D(a,b)$ in Figure2). Consequently, large amount of intra-class distances might be even larger than that of inter-class distances [Wang *et al.*, 2016b]. Existing methods in visible re-ID cannot simultaneously handle the cross- and intra- modality variations, which limit the discriminability of the learnt feature representation.

To address the aforementioned issues, we design a novel bi-directional dual-constrained top-ranking loss to guide the training process. The designed loss simultaneously considers the following two aspects: 1) cross-modality top-ranking constraint, which aims at addressing the large cross-modality variations. The main idea is that the distance of *anchor to its furthest cross-modality positive* should be smaller than the *anchor to its nearest cross-modality negative* by a predefined margin. 2) intra-modality top-ranking constraint, which focuses on handling the intra-modality variations. Under the same framework of 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 distinguishable. Furthermore, a bi-directional training strategy (*visible to thermal and thermal to visible*) is employed to enhance the robustness.

In addition, since the large intra-class variations also exist in VT-REID as shown in Figure 3, it's hard to ensure the discriminability by simply exploiting the relationships among persons with the ranking loss. In light of early success of human annotated labels of each person image, we further aggregate the identity loss into the dual-constrained top-ranking framework to extract the identity-specific information. It treats the same person identity across heterogenous modalities as the same class. It guarantees the learnt feature representation is identity invariant to address the large intra-class variations. Meanwhile, it also helps to stabilize the overall training process since two paths of the network may have totally different parameters caused by heterogenous modalities.

The main contributions can be summarized as follows:

- We present an end-to-end dual-path feature and metric learning framework, which is the first attempt for VT-REID. It provides a superior baseline in this research field for future improvements.

Figure 3: Intra-class variations. Images represent the same person.

- We introduce a novel bi-directional dual-constrained top-ranking loss to simultaneously consider the cross-modality and intra-modality variations, which provide new insights to enhance the discriminability of the learnt representation. Meanwhile, identity loss is incorporated to model the identity-specific information.

2 Related Work

Multi-Modality Person Re-identification. A detailed overview about person re-identification in visible domains can be found in [Zheng *et al.*, 2016]. Here we mainly discuss the multi-modality person re-identification.

Previously, several multi-modal fusion models have been proposed for person re-identification in visible-thermal modules [Nguyen *et al.*, 2017] and RGB-D modules [Barbosa *et al.*, 2012; Wu *et al.*, 2017a], where additional modality information captured by other spectral cameras (depth camera, thermal camera) is integrated with standard RGB images to improve the person re-identification performance [Mogelmose *et al.*, 2013]. Meanwhile, Lin *et al.* [Lin *et al.*, 2017b] aggregated semantic attributes information with visible images for person re-identification. In comparison, we focus on cross-modality person re-identification problem in this paper.

For cross-modality person re-identification, some text-to-image person retrieval methods [Li *et al.*, 2017a; 2017b; Ye *et al.*, 2015] have been proposed, but their approaches cannot be directly adopted for VT-REID. In VT-REID, a two-stage framework with feature learning and metric learning is introduced in [Ye *et al.*, 2018]. Additionally, Wu *et al.* [Wu *et al.*, 2017b] presented a deep zero-padding network [Chen *et al.*, 2018] to learn the invariant feature representations. In contrast, we present an end-to-end dual-path learning framework for feature and metric learning.

Deep Cross-modality Matching. Cross-modality matching has been widely investigated in the literature, especially in heterogenous face recognition and text-to-image retrieval. Here, we mainly discuss the deep learning based techniques due to their superior performances in various vision tasks [Zhou *et al.*, 2018; Song *et al.*, 2018; Ye *et al.*, 2017].

For heterogenous face recognition, deep invariant feature representation learning has been investigated in [He *et al.*, 2017; Wu *et al.*, 2018c] for NIR-VIS face recognition, and Sarfra *et al.* [Sarfraz and Stiefelhagen, 2017] presented a deep matching method via a two-layer non-linear function with hand-crafted features. In comparison, the VT-REID task also suffers from large intra-class variations besides the cross-modality variations compared with the face recognition, which makes their methods unsuitable for VT-REID.

For text-to-image matching, several dual-path based networks have been proposed to bridge the gap between the vi-

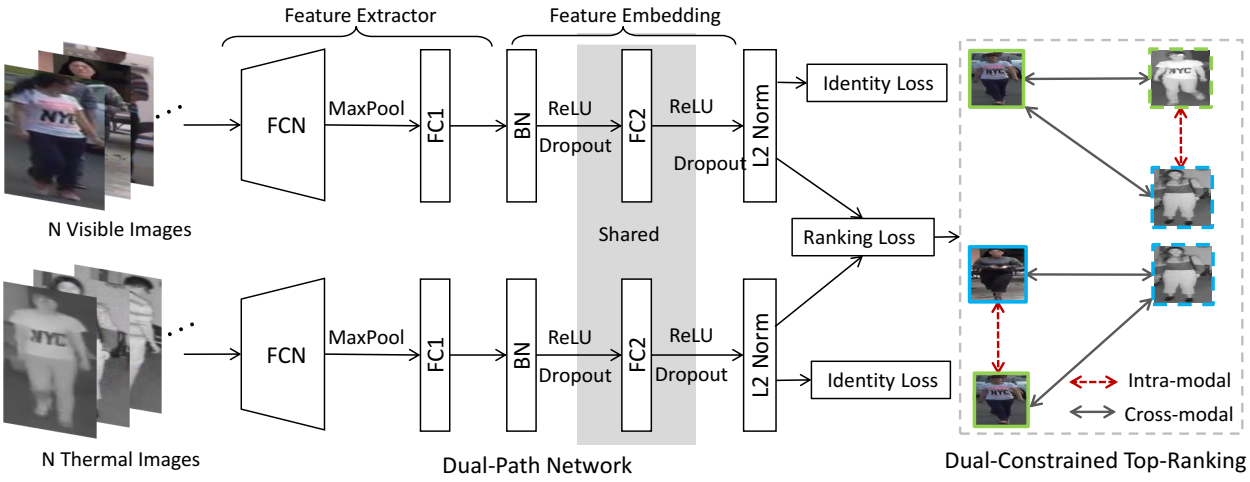

Figure 4: The proposed dual-path end-to-end learning framework for VT-REID. N represents the batch size, while totally $2*N$ images are fed into the network for training. It comprises two main components: dual-path network for feature extraction (one path for visible images and the other for thermal images) and bi-directional dual-constrained top-ranking loss for feature learning. Note that the weights of the shallow layers (*feature extractor*) are different to extract the modality-specific information while the weights of the embedding FC layer (*feature embedding*) are shared for multi-modality sharable feature learning. After L2 normalization, we introduce a bi-directional dual-constrained top-ranking loss for network training. Meanwhile, the identity loss is further integrated with the ranking loss to improve the performance.

sual images and text descriptions [Cao *et al.*, 2017; Liu *et al.*, 2017; Liong *et al.*, 2017]. Typically, the network contains one text CNN path and one image CNN path. Under this pipeline, we design a dual-path learning framework for our cross-modality person re-identification. Specifically, to address the cross-modality variations and intra-modality variations existing in VT-REID, a novel dual-constrained top-ranking loss on top of the ranking loss is introduced.

3 Proposed Method

This paper proposes a dual-path end-to-end feature learning framework for VT-REID as shown in Figure 4. The framework learns the feature representations and distance metrics in an end-to-end manner while preserving high discriminability. It comprises two main components: dual-path network for feature extraction and bi-directional dual-constrained top-ranking loss for feature learning. Specifically, the dual-path network utilizes partially shared structures to learn the multi-modality sharable features by simultaneously modeling the modality specific and modality shared information. The dual-constrained top-ranking loss ensures the learnt feature representations are discriminative enough to distinguish different persons from two heterogenous modalities. Identity loss is integrated to facilitate the feature learning process.

3.1 Dual-path Network

We propose a dual-path network to extract the features for visible and thermal domains. Specifically, the dual-path feature learning network contains two parts: feature extractor and feature embedding. The former feature extractor aims at capturing modality specific information for different image modalities. The latter feature embedding focuses on learning a multi-modality sharable space to bridge the gap between two heterogenous modalities.

Feature extractor. We adopt the off-the-shelf image feature extractors to extract the features from two heterogenous modalities. Due to the limited training data, the general image classification networks pre-trained on ImageNet are adopted for initialization to boost the training procedure for fast convergence. Note that both thermal-path and visible-path share similar network structures in our cross-modality person re-identification task. The main reason is that we assume the low-level visual patterns (eg. texture, corner) of thermal images are similar to general visible images. However, the parameters of two streams are optimized separately to capture the modality-specific information.

In our model, we adopt the AlexNet [Krizhevsky *et al.*, 2012] as the baseline network[2] for both visible and thermal paths. Specifically, we adopt the pre-trained five convolutional layers ($conv1 \sim conv5$) and one fully connected layer (with size of 4096) as the initialized feature representation. The main reason is that the shallow convolutional layers mainly capture the low-level visual patterns which might be shared among all images. Meanwhile, we add another batch normalization layer after the FC layer.

Feature embedding. To learn a discriminative embedding space of two heterogenous modalities, we introduce a shared fully connected layer on top of the dual-path feature extractors. Note that the weights of the fully connected layer are shared to model the modality shared information. If not, the learnt visible and thermal image features may lie in totally different subspaces [Wang *et al.*, 2016a; Wu *et al.*, 2018a]. Experimental results in the following section show that shared structure could achieve better performance for VT-REID, where it acts as a projection function to

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

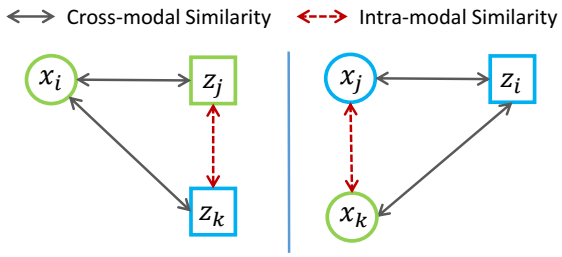

Figure 5: Illustration of the bi-directional dual-constrained top-ranking loss. Rectangles represent the thermal domain while circles represent the visible domain. Color demonstrates the identity. Left: visible-thermal top-ranking loss. Right: thermal-visible top-ranking loss.

project two different modalities into the common space. For simplicity in presentation, we denote the embedded function together with the feature extractor as $\mathcal{F}_v(\cdot)$ for visible images while $\mathcal{F}_t(\cdot)$ for thermal images. Given a visible image I_v and a thermal image I_t, the extracted features (x and z) are represented by

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

3.2 Dual-Constrained Top-Ranking

After transforming the visible and thermal images into a shared embedding space, we propose a novel bi-directional dual-constrained top-ranking loss to guide the feature learning objectives. The learning objective mainly contains the cross-modality and intra-modality constraints as shown in Figure 5. Firstly, we will revisit the general ranking loss.

Ranking Loss Revisit. Given a mini-batch, it contains N visible images and N thermal images. For an anchor visible image x_i with its label denoted by y_i, we want 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:

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

Note that all the input feature vectors x and z are l_2 normalized for stable convergence. In our proposed method, Euclidean distance is utilized as the similarity measurement, in which we empirically find that it achieves slightly better performance than other measurements for VT-REID. Furthermore, we further employ a bi-directional ranking loss strategy to constrain the overall learning for the cross-modality person re-identification problem. The bi-directional ranking loss contains two kinds of relationships: *visible to thermal triplet* (one anchor visible image, two thermal images) and *thermal to visible triplet* (one anchor thermal image, two visible images). The bi-directional ranking loss is formulated by

$$\begin{aligned}\mathcal{L}_{bi\_rank} = &\sum_{\forall y_i=y_j, y_i\neq y_k} \max[\rho_1 + D(x_i, z_j) - D(x_i, z_k), 0] \\ &+ \sum_{\forall y_i=y_j, y_i\neq y_k} \max[\rho_1 + D(z_i, x_j) - D(z_i, x_k), 0]\end{aligned} \tag{3}$$

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

Cross-modality Top-Ranking Constraint. To address the issue that large amounts of intra-class distances might be even larger than the inter-class distances caused by cross-modality variations, we employ a top-ranking constraint following [Hermans *et al.*, 2017] to enhance the discriminability. The underlying idea is that we compare the distance of a positive visible-thermal pair and the minimum distance of all related negative visible-thermal pairs, rather than each of the negative pairs. The cross-modality constrained top-ranking loss is further developed as:

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

The bi-directional cross-modality top-ranking loss has two main advantages: (1) The top-ranking constraint ensures that the closest cross-modality negative sample is far from the farthest cross-modality positive sample, thus helps to reduce the cross-modality variations while preserve high discriminability. (2) The bi-directional training strategy makes sure that the learnt feature representation is modality invariant. It improves the robustness for different query settings (*i.e.*, *visible to thermal and thermal to visible*) as illustrated in Sec. 4.3.

Intra-modality Top-Ranking Constraint. As discussed in the Section 1, VT-REID also suffers from the intra-class intra-modality variations due to different poses, viewpoints and etc. To address this issue, we introduce another intra-modality similarity constraint to enhance the robustness of the learnt feature representation to intra-modality variations. On top of the cross-modality top-ranking loss, the intra-modality constrained loss is computed by

$$\begin{aligned}\mathcal{L}_{intra} = &\sum \max[\rho_2 - D(z_j, z_k), 0] \\ &+ \sum \max[\rho_2 - D(x_j, x_k), 0]\end{aligned} \tag{5}$$

where ρ_2 is a pre-defined margin. j and k represent the same index as in the cross-modality top-ranking within the mini-batch for each anchor i. This intra-modality top-ranking constraint ensures that the hardest cross-modality negative sample should also be far from to its corresponding cross-modality positive samples. It guarantees that the images of different persons within each modality should also be distinguished with additional constraint, especially when the cross-modality equality Eq. 2 does not hold for large-scale training.

Overall Embedding Loss. Since above ranking loss constrains the feature learning process with their underlying relationships among persons, it's hard to learn a robust feature representation to reduce the intra-class variations by simply exploiting the relationship cues. Meanwhile, the visible and thermal image features may exist in totally different feature spaces, the ranking loss may also be trapped into convergence problem due to incorrect relationship measurements. Therefore, we integrate the identity information to the overall loss function. For the sake of feasibility and effectiveness in classification, the general softmax loss is utilized by treating each person identity as a class. In this manner, the identity specific information is integrated to enhance the robustness.

The final loss function is a weighted summation three components, the bi-directional cross-modality and intra-modality top-ranking constraints and identity loss, defined by

$$\mathcal{L} = \mathcal{L}_{cross} + \lambda_1 \mathcal{L}_{intra} + \lambda_2 \mathcal{L}_{id} \quad (6)$$

where λ_1 and λ_2 are predefined weighting parameters.

Batch Sampling. Since our dual-constrained ranking loss is slightly different with general person re-identification tasks, it's essential to introduce the mini-batch sampling strategy. Specifically, N person identities are firstly randomly selected at each iteration, where N is the batch size. Then we randomly select one visible image and one thermal image of the selected identity from two different modalities to construct the mini-batch, in which totally $2*N$ images are fed into the network for training. In this manner, within the mini-batch, we can select N anchor visible images to calculate the visible-thermal top-ranking loss, and N corresponding anchor thermal images for the thermal-visible top-ranking loss. Due to the randomly sampling mechanism, all the possible assemblies will be traversed to get the global optimum.

4 Experimental Results

4.1 Experimental Settings

Datasets and settings. Two publicly available RegDB dataset [Nguyen *et al.*, 2017] and SYSU-MM01 [Wu *et al.*, 2017b] are adopted for evaluation. RegDB is collected by dual camera systems, and contains 412 persons. For each person, 10 different visible light images are captured by a visible camera, and 10 different thermal images are obtained by a thermal camera. We follow the evaluation protocol in [Ye *et al.*, 2018], where the dataset is randomly split into two halves, one for training and one for testing. For testing, the images from one modality were used as the gallery set while the ones from the other modality as the probe set. The procedure is repeated for 10 trials to achieve statistically stable results.

SYSU-MM01 [Wu *et al.*, 2017b] is a large-scale dataset collected by 6 cameras, including four visible and two thermal cameras. This dataset is challenging since some of the person images are captured in the indoor environments and some are in outdoor environments. It contains 491 persons, each person is captured by at least two different cameras. We adopt the single-shot *all-search* mode evaluation protocol, since it's the most challenging case as mentioned in [Wu *et al.*, 2017b]. The training set contains 395 persons, with 22258 visible images and 11909 thermal images. The testing set contains 96 persons, with 3803 thermal images for query and 301 randomly selected visible images as gallery set.

Evaluation metrics. To indicate the performance, the standard cumulated matching characteristics (CMC) curve and mean average precision (mAP) are adopted, since one person has multiple groundtruths in the gallery set.

Implementation details.[3] We implement our algorithm with Tensorflow. The size of the embedding fully connected layer is set as 1024 and the batch size is set as 64 for both datasets. 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 as $\lambda_1 = 0.1$ and $\lambda_2 = 1$. Momentum optimizer is utilized for optimization, and the momentum is set to 0.9. The predefined cross-modality margin ρ_1 is set to 0.5 while the intra-modality margin ρ_2 is set to 0.1. The initial learning rate is set as 0.001. The training step for RegDB dataset is 5000 and SYSU-MM01 dataset is 50000.

[3]Code is available on the first author's website

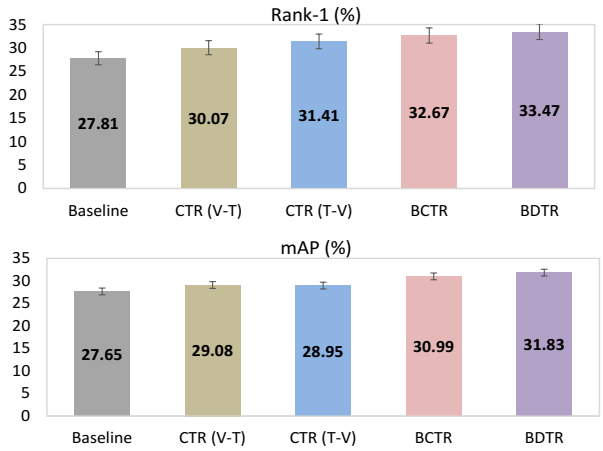

Figure 6: Evaluation of different variants of proposed method on the RegDB dataset. Re-identification rates (%) at rank r and mAP (%).

4.2 Ablation Study

Variants evaluation. This subsection evaluates the proposed end-to-end learning framework with different variants, where the results on the RegDB dataset are shown in Figure 6. *"Baseline"* means the results when general ranking loss integrated with the identity loss is used. *"CTR (V-T)"* and *"CTR (T-V)"* denote a cross-modality top-ranking constraint is further incorporated in the ranking loss. Two kinds of ranking strategies are evaluated, *i.e.*, *visible to thermal and thermal to visible.* "BCTR" *represents the results when bi-directional training strategy is employed.* "BDTR" *expresses the performance with a further aggregated intra-modality constraint,* which demonstrates the overall bi-directional dual-constrained top-ranking loss.

Results shown in Figure 6 illustrate that a cross-modality top-ranking constraint could consistently improve the performance of ranking loss by 2-3%. It verifies the idea that top-rank constraint helps to handle large intra-class cross-modality variations, which ensures the discriminability of the learnt feature representation. After employing the bi-directional training strategy (BCTR), the overall performances of rank-1 and mAP are enhanced further with about 2%. Improvements show that bi-directional relationship training helps to improve the robustness of the learnt feature representation for two heterogenous modalities. Furthermore, both the rank-1 matching rates and mAP are further improved by integrating an intra-modality similarity constraint, which addresses the intra-modality variations caused by different poses or viewpoints. Overall, the proposed BDTR improves the rank-1 matching rate from 27.81% to 33.47%, and mAP from 27.65% to 31.83% on the RegDB dataset.

Datasets	RegDB				SYSU-MM01			
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	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
GSM	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
One-stream	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
TONE	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
Ours (Baseline)	27.81	51.41	60.99	27.65	12.96	51.80	71.00	16.11
Ours (BCTR)	32.67	57.64	66.58	30.99	16.12	54.90	71.47	19.15
Ours (BDTR)	**33.47**	**58.42**	**67.52**	**31.83**	**17.01**	**55.43**	**71.96**	**19.66**

Table 2: Comparison with the state-of-the arts on the RegDB and SYSU-MM01 datasets. Re-identification rates (%) at rank r and mAP (%).

SYSU-MM01	$r = 1$	$r = 10$	r=20	mAP
Only ranking loss	7.99	36.31	53.17	10.11
Only identity loss	13.52	47.62	65.08	15.86
Full model	17.01	55.43	71.96	19.66
RegDB	$r = 1$	$r = 10$	r=20	mAP
Only ranking loss	32.46	58.65	66.86	31.23
Only identity loss	18.60	49.35	65.08	17.68
Full model	33.47	58.42	67.52	31.83

Table 1: Effectiveness of identity loss on the RegDB and SYSU-MM01 datasets. Re-identification rates (%) at rank r and mAP (%).

Effectiveness of identity loss. We also conduct the experiments to verify the effectiveness of the identity loss and the ranking loss on both the RegDB and the large-scale SYSU-MM01 datasets. We report the results with only ranking loss, only identity loss and our full model as shown in Table 1.

As demonstrated in Table 1, the rank-1 matching rate is about 7.99% for the ranking loss while the mAP is about 10.11% on the SYSU-MM01 dataset. After integrating the identity loss, our final full model could achieve rank-1 = 17.01%, and mAP = 19.66%. The results illustrate that the integration could improve the performance by aggregating the identity-specific information (identity loss) to the ranking loss. Meanwhile, for the results on the small scale RegDB dataset, we could observe that ranking loss achieves rank-1 accuracy about 32.46%, 18.60% for the identity loss and the overall model is 33.47%. Although the improvement is not that significant, the combination still improves the performance of raking loss. It verifies that the fusion of two different losses work well for the cross-modality person re-identification. Another observation is that the ranking loss performs much better on small RegDB dataset while identity loss could achieve better performance with abundant training samples on the SYSU-MM01 dataset. This phenomenon has also been verified in cross-view re-ID [Xiao *et al.*, 2017].

4.3 Comparison with the State-of-the-arts

Competing methods. Since only two published works have investigated the visible thermal person re-identification, and they only evaluate their method on one dataset.

- **Zero-Padding** [Wu *et al.*, 2017b]. A deep zero-padding method to utilizes a one-stream network to capture the domain specific information. We re-implement it their method to evaluate on the RegDB dataset.
- **TONE + HCML** [Ye *et al.*, 2018]. A two-stage framework for feature learning (TONE) and metric learning (HCML) is proposed. Previous method has only been evaluated on RegDB dataset. We adopt the authors' released code to evaluate on SYSU-MM01 dataset.

In addition, several other cross-modality learning methods are also included for comparison. Most of the results are originated from [Ye *et al.*, 2018] on the RegDB dataset and [Wu *et al.*, 2017b] on the SYSU-MM01 dataset. The competing methods contain some feature learning methods (HOG, LOMO [Liao *et al.*, 2015], one-stream and two-stream networks). Note that one-stream and two-stream networks are the modifications of the IDE method [Zheng *et al.*, 2016] under our cross-modality re-identification settings, detailed description can be found in [Wu *et al.*, 2017b].In addition, some matching model learning methods (XQDA [Liao *et al.*, 2015], MLAPG [Liao and Li, 2015], GSM [Lin *et al.*, 2017a] SCDL [Wang *et al.*, 2012] and rCDL [Huang and Frank Wang, 2013]) are also included for comparison. The results are shown in Table 2.

The results shown in Table 2 demonstrate that the proposed end-to-end learning framework outperforms existing state-of-the-art methods usually by a large margin on the RegDB dataset. Compared to the two-stage feature learning and metric learning method (TONE + HCML), we consistently outperform them with nearly 10% for both rank-1 matching rate and mAP. For the large-scale SYSU-MM01 dataset, the proposed method also achieves the best performance compared

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
Ours (BDTR)	33.47	58.42	67.52	31.83
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
Ours (BDTR)	32.72	57.96	68.86	31.10

Table 3: Evaluation of different query settings on the RegDB dataset. Re-identification rates (%) at rank r and mAP (%).

to the competing methods. Specifically, we achieve rank-1 = 33.47% and mAP = 31.83% on the RegDB dataset, and rank-1 = 17.01% and mAP = 19.66% on the SYSU-MM01 dataset. The advantages of our proposed method can be summarized as two folds: 1) End-to-end learning could learn discriminative features without any human intervention. 2) The proposed dual-constrained top-ranking loss with the dual-path network provides a good solution to address the large cross-modality and intra-modality variations for VT-REID.

Different query settings. We also evaluate the performance of different query settings on the RegDB dataset as done in [Ye *et al.*, 2018]. Results shown in Table 3 illustrate that the proposed method is robust to different query settings, the performance of visible-to-thermal matching is close to the results of thermal-to-visible matching with less than 1% difference. We could achieve about 33% rank-1 matching accuracy and 31% for mAP on both settings. Moreover, the proposed method outperforms the competing methods consistently by a large margin on both settings. The results demonstrate the flexibility and applicability of the proposed method in real applications. The superiority of the proposed method is attributed to the designed bi-directional training strategy.

5 Conclusion

In this paper, a well-designed end-to-end learning framework via dual-constrained top-ranking loss is proposed for visible thermal cross-modality person re-identification. To address the large cross-modality variations, a bi-directional cross-modality constrained top-ranking loss is employed to enhance the discriminability of the learnt feature representation. Meanwhile, to address the large intra-class intra-modality variations, an intra-modality constraint is incorporated to train the network. Additionally, identity loss is further integrated to learn identity-specific information. Extensive experiments illustrate the superiority of the proposed method when compared with the state-of-the-arts.

Acknowledgments

This work is partially supported by Hong Kong RGC General Research Fund HKBU (12202514), and National Natural Science Foundation of China (61562048).

References

[Barbosa *et al.*, 2012] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCVW*, pages 433–442, 2012.

[Cao *et al.*, 2017] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, pages 3974–3980, 2017.

[Chen *et al.*, 2018] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, and Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE T-PAMI*, 40(2):392–408, 2018.

[He *et al.*, 2017] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *AAAI*, pages 2000–2006, 2017.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[Huang and Frank Wang, 2013] De-An Huang and Yu-Chiang Frank Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *ICCV*, pages 2496–2503, 2013.

[Jiang *et al.*, 2017] Junjun Jiang, Chen Chen, Jiayi Ma, Zheng Wang, Zhongyuan Wang, and Ruimin Hu. Srlsp: A face image super-resolution algorithm using smooth regression with local structure prior. *IEEE Transactions on Multimedia*, 19(1):27–40, 2017.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Lan *et al.*, 2015] Xiangyuan Lan, Andy J Ma, Pong C Yuen, and Rama Chellappa. Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. *IEEE TIP*, 24(12):5826–5841, 2015.

[Lan *et al.*, 2018] Xiangyuan Lan, Shengping Zhang, Pong C Yuen, and Rama Chellappa. Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker. *IEEE Transactions on Image Processing*, 27(4):2022–2037, 2018.

[Li *et al.*, 2017a] Shuang Li, Tong Xiao, and et al. Person search with natural language description. In *CVPR*, pages 1345–1353, 2017.

[Li *et al.*, 2017b] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, pages 1890–1899, 2017.

[Liao and Li, 2015] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal

occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Lin *et al.*, 2017a] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE TPAMI*, 39(6):1089–1102, 2017.

[Lin *et al.*, 2017b] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[Liong *et al.*, 2017] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Cross-modal deep variational hashing. In *ICCV*, pages 4077–4085, 2017.

[Liu *et al.*, 2017] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, pages 4107–4116, 2017.

[Mogelmose *et al.*, 2013] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert Clapés, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *CVPRW*, pages 301–307, 2013.

[Nguyen *et al.*, 2017] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

[Sarfraz and Stiefelhagen, 2017] M Saquib Sarfraz and Rainer Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *IJCV*, 122(3):426–438, 2017.

[Song *et al.*, 2018] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *CVPR*, 2018.

[Wang *et al.*, 2012] S. Wang, L. Zhang, Liang Y., and Q. Pan. Semi-coupled dictionary learning with applications in image super-resolution and photo-sketch synthesis. In *CVPR*, pages 2216–2223, 2012.

[Wang *et al.*, 2016a] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[Wang *et al.*, 2016b] Zheng Wang, Ruimin Hu, Chao Liang, Yi Yu, Junjun Jiang, Mang Ye, Jun Chen, and Qingming Leng. Zero-shot person re-identification via cross-view consistency. *IEEE Transactions on Multimedia*, 18(2):260–272, 2016.

[Wang *et al.*, 2017] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics*, 2017.

[Wu *et al.*, 2017a] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE TIP*, 26(6):2588–2603, 2017.

[Wu *et al.*, 2017b] Ancong Wu, Wei-shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianghuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017.

[Wu *et al.*, 2018a] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition (PR)*, 73:275–288, 2018.

[Wu *et al.*, 2018b] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. *Pattern Recognition (PR)*, 76:727–738, 2018.

[Wu *et al.*, 2018c] Xiang Wu, Lingxiao Song, Ran He, and Tieniu Tan. Coupled deep learning for heterogeneous face recognition. In *AAAI*, 2018.

[Xiao *et al.*, 2017] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017.

[Ye *et al.*, 2015] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete text description. In *ICMR*, pages 547–550, 2015.

[Ye *et al.*, 2016] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[Ye *et al.*, 2017] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, pages 5142–5150, 2017.

[Ye *et al.*, 2018] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.

[Zheng *et al.*, 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[Zheng *et al.*, 2017] Liang Zheng, Yi Yang, and Qi Tian. SIFT meets CNN: A decade survey of instance retrieval. *IEEE TPAMI*, 2017.

[Zhou *et al.*, 2018] Joey Tianyi Zhou, Heng Zhao, Xi Peng, Meng Fang, Zheng Qin, Zheng Qin, and Rick Siow Mong Goh. Transfer hashing: From shallow to deep. *IEEE Transaction on Neural Network and Learning Systems*, 2018.

[Zhu *et al.*, 2018] Xiatian Zhu, Botong Wu, Dongcheng Huang, and Wei-Shi Zheng. Fast open-world person re-identification. *IEEE Transactions on Image Processing (TIP)*, 27(5):2286 – 2300, 2018.