2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

A Base-Derivative Framework for Cross-Modality RGB-Infrared Person Re-Identification

Hong Liu, Ziling Miao*, Bing Yang and Runwei Ding
Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School
Shenzhen, China
{hongliu, zilingmiao}@pku.edu.cn, bingyang@sz.pku.edu.cn, dingrunwei@pku.edu.cn

***Abstract*—Cross-modality RGB-infrared (RGB-IR) person re-identification (Re-ID) is a challenging research topic due to the heterogeneity of RGB and infrared images. In this paper, we aim to find some auxiliary modalities, which are homologous with the visible or infrared modalities, to help reduce the modality discrepancy caused by heterogeneous images. Accordingly, a new *base-derivative* framework is proposed, where *base* refers to the original visible and infrared modalities, and *derivative* refers to the two auxiliary modalities that are derived from base. In the proposed framework, the *double-modality* cross-modal learning problem is reformulated as a *four-modality* one. After that, the images of all the base and derivative modalities are fed into the feature learning network. With the doubled input images, the learned person features become more discriminative. Furthermore, the proposed framework is optimized by the enhanced intra- and cross-modality constraints with the assistance of two derivative modalities. Experimental results on two publicly available datasets SYSU-MM01 and RegDB show that the proposed method outperforms the other state-of-the-art methods. For instance, we achieve a gain of over 13% in terms of both Rank-1 and mAP on RegDB dataset.**

I. Introduction

Person re-identification (Re-ID) aims to retrieve a target person across several non-overlapping cameras [1]. Given the query images of a person-of-interest, Re-ID targets to search the gallery set to find the images which own the same identity with the query. Person Re-ID using RGB images have achieved great success with the development of deep learning [2], [3], [4]. It provides favorable performance at good illuminations, but fails to capture valid appearance information of pedestrians under poor lighting conditions such as in the dark. Instead, infrared images can offer better pedestrian information at poor illuminations. The complementarity of the two modalities is exploited by RGB-infrared (RGB-IR) Re-ID which matches the day-time visible images with the night-time infrared images.

One challenge for RGB-IR person Re-ID is the large heterogeneity of RGB and IR images. Various methods are proposed to reduce the modality discrepancy. Similar to Re-ID using visible images, one-stream RGB-IR Re-ID method directly changes the inputs from single-modality images to modality-mixed images [5], and the two heterogeneous images are sent into a common feature learning network, as shown in Fig. 1 (a). To further reduce the cross-modality discrepancy, a two-stream method [6], [7] is proposed to separately model the modality-sharable and modality-specific information, as shown in Fig. 1 (b). Asymmetric method [8], [9] uses shared embedding learning network and tries to bridge the two modalities in the classification subspace, as shown in Fig. 1 (c). Moreover, Hao et al. [10] explored the correlation between the classification subspace and embedding subspace. Recently, Generative Adversarial Networks (GAN) [11] based Re-ID methods have attracted a lot of attention due to its powerful ability of image generation. Some works [12], [13] generate heterogeneous images and combine the discriminators with the Re-ID network, as shown in Fig. 1 (d).

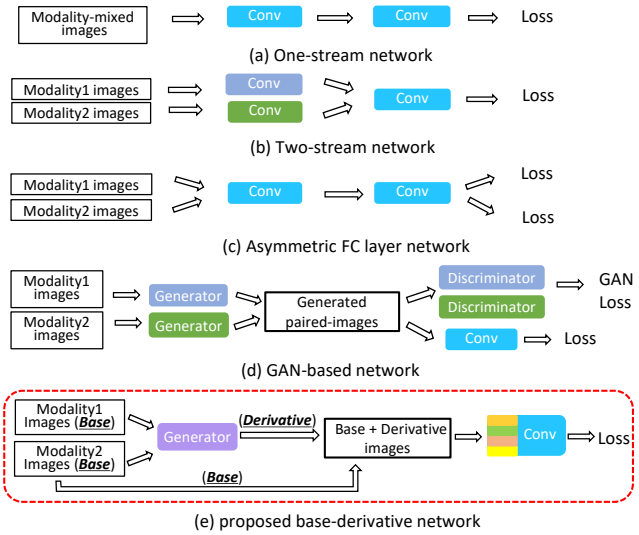

Fig. 1. Different networks for RGB-IR person Re-ID. Components with the same name and color are modality-sharing in each subgraph. Our base-derivative method constructs a generator-sharing network, which is different from the existing RGB-IR Re-ID methods. More details are in Section II.

Above all, some methods focus on the embedding and classification subspaces, however, the discrepancy in image subspace also exists [14], [15]. The GAN based methods reduce the modality discrepancy in image subspace by reconstructing the original RGB and IR images, however, they are usually devoted to construct heterogeneous images for visible/infrared modalities. In this paper, we try to construct some new modalities which are not heterogeneous but homologous with visible and infrared modalities.

In this paper, a novel lightweight base-derivative framework is proposed for cross-modality RGB-infrared person Re-ID. We generate two new auxiliary modalities that are homologous

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9413029

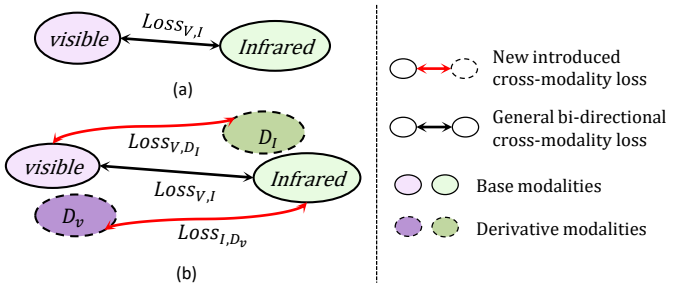

Fig. 2. Cross-modality constraints. (a) the general bi-directional cross-modality loss. (b) the base-derivative framework with an enhanced multi-directional loss (each modality is restrained by two bi-directional losses).

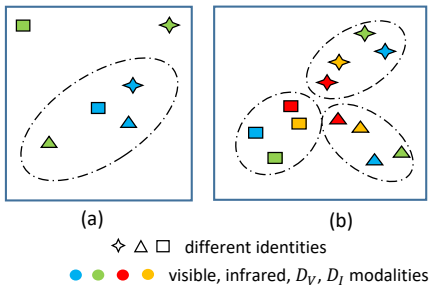

Fig. 3. Cross-modality learning. Different colors represent different modalities. (a) the results of general cross-modality learning. (b) the results of cross-modality learning with derivative modalities. Samples in dashed circle mean they are possible to be matched together during prediction (**Best viewed in color**).

with the base visible and infrared modalities, which we call derivative modalities. The main components of the proposed base-derivative framework are stated as following:

- **Base modalities:** The visible and infrared modalities.
- **Derivative modalities:** Totally two.
  D_V: the homologous modality with visible modality, which inherits appearance information (such as pose, clothing category and carrying) from visible modality.
  D_I: the homologous modality with infrared modality, which inherits appearance information from infrared modality.

As shown in Fig. 1 (e), the derivative modalities are constructed by a common generator, which introduces guidance from both base modalities with a lightweight network. With the help of derivative modalities, more connections between visible and infrared modalities can be established. Consequently, the general bi-directional cross-modality loss are enhanced into multi-directional loss (MDL), which contains three bi-directional cross-modality losses between *visible-infrared*, *visible-*D_I and *infrared-*D_V modalities, as shown in Fig. 2. With the base and derivative modalities, the RGB-IR person Re-ID is reformulated as a *four-modality* cross-modal learning problem, which becomes easier with the proposed MDL as demonstrated by Fig. 3. Take the D_V as an example, the cross-modality loss between D_V*-infrared* and *visible-infrared* force the positive visible images and D_V images to approach the infrared ones. Experimental results on SYSU-MM01 and RegDB demonstrate the superiority of the proposed method.

The contributions of our method can be described as following:

- The modality discrepancy is reduced with the assistance of two generated homologous modalities, instead of using heterogeneous modalities as in popular methods.
- The *double-modality* cross-modality learning problem is reformulated as a *four-modality* one with a *base-derivative* framework, through which the cross-modality learning becomes easier with a devised multi-directional loss (MDL).
- The proposed method achieves a significant improvement compared with the state-of-the-arts on two popular datasets. Especially on the RegDB dataset, it achieves a gain of over 13%.

II. Base-Derivative Networks

We assume the cross-modality discrepancy contains two parts: First is the appearance discrepancy caused by various viewpoints, poses and illuminations in conventional Re-ID task. Second is the modality discrepancy originated from different imaging processes of visible and infrared cameras [12]. In this paper, we mainly focus on the modality discrepancy and try to reduce it in channel dimension as in [16].

The framework of proposed base-derivative method is shown in Fig. 4. First, two new derivative modalities are generated from the captured visual and infrared images. Then, the base and derivative images are fed into a feature learner to extract more discriminative person features. Finally, the whole network is trained in an end-to-end manner with designed multi-directional loss (MDL), multi-mode loss (MML) and identity (ID) loss.

A. Derivative Modality Generation

Because the appearance information of RGB and IR images are actually diverse, which makes it difficult to fuse them together to construct only one new modality. Thus, two derivative modalities are generated separately with inherited appearance information from the two base modalities. In this part, the 'Generator' is responsible for introducing assistance from base modalities and constructing the auxiliary derivative images. X_V, X_I, X_{D_V} and X_{D_I} are images of different modalities.

A two-step benchmark is designed for the generation of D_V and D_I, as the green and blue stream shown in Fig. 4. Firstly, an encoder E_V is used to compress three-channel RGB images into one-channel, which is the same format as IR images. Secondly, the encoded one-channel RGB images together with initial IR images are fed into a weight-sharing generator, which is used to siphon off knowledge from two base modalities (assistance from the infrared modality to construct D_V images and vice versa). Above all, a mutual-assistance modality generation network is defined as:

$$\begin{aligned} X_{D_V} &= G_S(E_V(X_V)), \\ X_{D_I} &= G_S(X_I), \end{aligned} \tag{1}$$

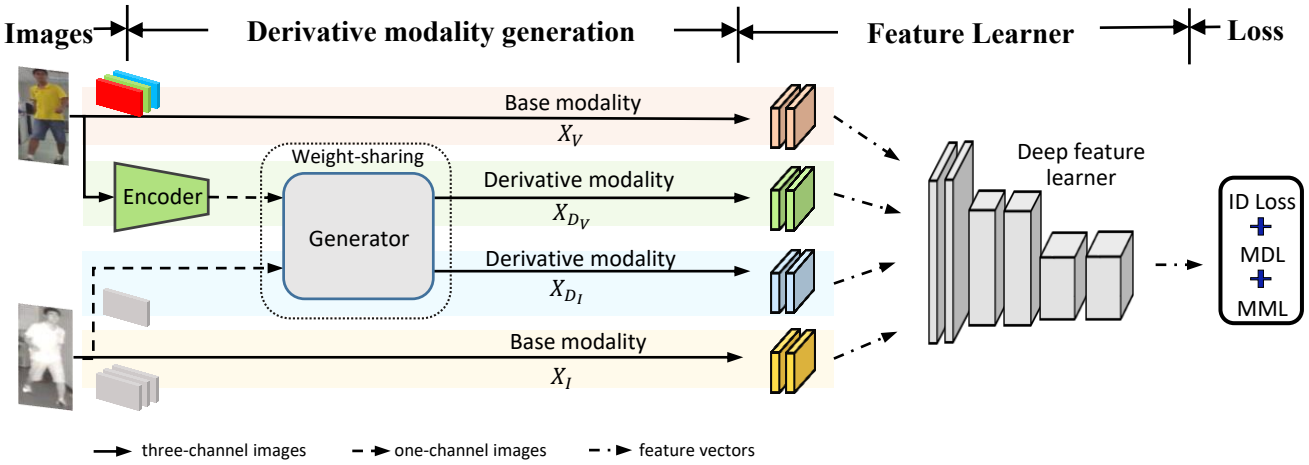

Fig. 4. The overall framework of our base-derivative network. X_V, X_I, X_{D_V} and X_{D_I} are images of different modalities. The four modalities take different streams (shown in different colors). The feature learner is divided into two parts: four modality-independent feature learners and a modality-sharing deep feature learner. The whole network is optimized by three loss functions. (**Best viewed in color**).

E_V only encodes images in the channel level (from three channels to one), which is achieved by using a simple convolutional layer with an 1×1 kernel and a ReLU activation layer as in [16]. G_S is the weight-sharing generator, which aims to map the one-channel images (including the infrared images and encoded visible images) into three-channel images . It also works in the channel level merely and is achieved by a convolutional layer with three 1×1 kernels. The generation of D_V and D_I is a self-supervised process, where X_{D_V} and X_{D_I} still keep the consistent identity with X_V and X_I (without additional manual annotations).

B. Feature Learner

The feature learner is split into two parts. One is the modality-independent feature extractor F_I, which aims to learn the appearance-level information. The other is the modality-sharing feature extractor F_S, which aims to learn the semantic-level information.

Modality-independent feature extractor. The discrepancy of different modalities is still obvious in image space, so F_I is customized for each of the four modalities. They take the original images as input and output the low-level features. The first several layers of ResNet50 [17] are used to achieve F_I, that is, a convolutional layer, a BatchNorm layer, a ReLU layer and a maxpooling layer.

Modality-sharing feature extractor. Images from different modalities with the same labels should confirm the consistent distribution in high-level feature space [18]. So a weight-sharing extractor F_S is introduced following F_I. F_S takes the low-level modality-independent features as input and encodes them in semantic-level. It is achieved by the rest part of ResNet50 except layers in F_I.

C. Loss Functions

Three loss functions are used to guide the cross-modal learning. First is the multi-directional loss (MDL), which is applied across the four modalities to promote the circulation of cross-modality information. Second is the proposed multi-mode loss (MML), which is applied on each modality to reinforce the convergence of the four modalities. Third is the identity (ID) loss, which is applied on each modality to learn more discriminative features. In this part, the symbol M is used to represent any one of the four modalities, $M \in [visible, infrared, D_V, D_I]$.

Multi-directional cross-modality loss (MDL). As shown in Fig. 2, the proposed MDL can be divided into several bi-directional losses. Moreover, one bi-directional loss can be split into two unidirectional cross-modality losses. So MDL can be defined as:

$$L_{MDL} = \sum_{\substack{m1,m2 \in M \\ m_1 \neq m_2}} L_{m_1 \to m_2} + L_{m_1 \leftarrow m_2}, \tag{2}$$

here the groups of $visible \leftrightarrow D_V$ and $infrared \leftrightarrow D_I$ are not taken into consideration because they are homologous modalities. In addition, the group of $D_V \leftrightarrow D_I$ is also ignored considering that there are only visible and infrared images in testing stage. Therefore, main attention is paid to reduce the discrepancy between visible/infrared modalities and other modalities. Corresponding discussions are in Section III-E.

The unidirectional cross-modality loss is achieved by an improved cross-modality triplet loss. Triplet loss [19] treats images with the same identity as positive pairs, and images with different identities as negative pairs. We improve the triplet loss with hard sample mining, which requires the Euclidean distances of all positive pairs to be smaller than those of all negative pairs. The step-by-step calculation is explained in the following. First, a vizard matrix $V_{m_1 \leftrightarrow m_2}$ is calculated by :

$$v^{i,j}_{m_1 \leftrightarrow m_2} = \begin{cases} 0, & y_i \neq y_j \\ 1, & y_i = y_j, \end{cases} \tag{3}$$

where $v^{i,j}_{m_1 \leftrightarrow m_2}$ is the value at row i and column j in $V_{m_1 \leftrightarrow m_2}$. $i \in [1, N_{m_1}]$, $j \in [1, N_{m_2}]$, and N_{m_1} and N_{m_2} are the number

TABLE I
COMPARISON RESULTS(%) AT RANK r WITH THE STATE-OF-THE-ART CROSS-MODALITY RE-ID METHODS ON THE SYSU-MM01 DATASET.

Methods	All-search				Indoor-search			
	R1	R10	R20	mAP	R1	R10	R20	mAP
Zero-Padding [20] $_{ICCV17}$	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
HCML [7] $_{AAAI18}$	14.32	53.16	69.17	16.16	24.52	-	-	30.08
D-HSME [10] $_{AAAI19}$	20.68	62.74	77.95	23.12	-	-	-	-
eBDTR [6] $_{TIFS17}$	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
cmGAN [24] $_{IJCAI18}$	26.97	67.51	80.56	27.8	31.63	77.23	89.18	42.19
D^{2}RL [12] $_{CVPR19}$	28.9	70.60	82.40	29.20	-	-	-	-
MAC [8] $_{MM19}$	33.26	79.04	90.09	36.22	33.37	82.49	93.69	44.95
MSR [9] $_{TIP19}$	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN [13] $_{ICCV19}$	42.4	85.00	93.70	40.7	45.9	87.6	94.4	54.3
Hi-CMD [23] $_{CVPR20}$	34.94	77.58	-	35.94	-	-	-	-
JSIA [25] $_{AAAI20}$	38.10	80.7	89.9	36.9	43.8	86.2	94.2	52.9
Ours	**51.05**	**87.75**	**94.43**	**49.63**	**55.93**	**91.55**	**96.95**	**63.38**

Fig. 5. Images from the two datasets to explain why the improvements on RegDB are more significant. Details are in Section III-D.

of images from modality m_1 and m_2 in every batch. y_i and y_j are corresponding identity labels. Then $L_{m_1\to m_2}$ and $L_{m_2\to m_1}$ can be calculated by:

$$\begin{aligned} L_{m_1\to m_2} &= \sum_{i=1}^{N_{m_1}}[\alpha_2 + \max D^i_{m_1,m_2} - min\widetilde{D}^i_{m_1,m_2}]_+, \\ L_{m_2\to m_1} &= \sum_{i=1}^{N_{m_2}}[\alpha_2 + \max D^i_{m_2,m_1} - min\widetilde{D}^i_{m_2,m_1}]_+, \end{aligned} \tag{4}$$

where

$$\begin{aligned} D_{m_1,m_2} &= S_{m_1\leftrightarrow m_2} * V_{m_1\leftrightarrow m_2}, \\ \widetilde{D}_{m_1,m_2} &= S_{m_1\leftrightarrow m_2} * \widetilde{V}_{m_1\leftrightarrow m_2}, \end{aligned} \tag{5}$$

here $\widetilde{V}_{m_1\leftrightarrow m_2}$ is calculated by reversing the element values in the $V_{m_1\leftrightarrow m_2}$, that is, value '0' is reset to '1' and value '1' is reset to '0'. i is the row index of matrix D_{m_1,m_2} and $\widetilde{D}_{m_1,m_2}$. The $*$ means the dot product between two matrixes. α_2 is a margin parameter and $[z]_+ = max(z, 0)$. $S_{m_1\leftrightarrow m_2}$ is the distance matrix calculated using the Euclidean distance. The D_{m_2,m_1} and $\widetilde{D}_{m_2,m_1}$ can be calculated using functions similar to (5).

Multi-mode intra-modality loss (MML). MML enlarges the general two-mode intra-modality triplet loss into multi-mode (two base modalities and two derivative modalities) with an improved strategy of hard example sample mining. The MML is defined as:

$$L_{MML} = \sum_M L_M, \tag{6}$$

where

$$\begin{aligned} L_M = \sum_i^B [\alpha_1 + \max_{\substack{j=1,\cdots,B \\ y_i=y_j}} S(f(M_i), f(M_j)) - \\ \min_{\substack{k=1,\cdots,B \\ y_i\neq y_k}} S(f(M_i), f(M_k))]_+, \end{aligned} \tag{7}$$

$S(\cdot)$ is the Euclidean distance and $f(\cdot)$ is the feature vectors. i is the index of images from modality M. α_1 is a margin parameter and $[z]_+ = max(z, 0)$. B is the batch-size and $i, j, k \in [1, B]$. yi, yj and yk are the identities of images.

ID loss. In the cross-modality person Re-ID task, each identity is a distinct class. Based on that, the task can be treated as an image classification problem and the identity loss can be used for network optimization. We apply the ID loss on the four modalities, so the weight and bias vectors of all layers can be optimized during training. The total ID loss can be formulated by

$$L_{id} = -\frac{1}{N}\sum_M\sum_{i=1}^N y_i^M log(p(y_i^M|x_i^M)) \tag{8}$$

where N is the number of samples in every training batch, and i is the index of N. y_i is the label of image x_i.

Overall loss function. Above all, the overall objective loss function can be formulated as:

$$L_{total} = \alpha L_{id} + \beta L_{MML} + \gamma L_{MDL}, \tag{9}$$

where α, β and γ are weights of corresponding losses. α and β are both set to 1 for each modality empirically. γ is decided by grid-search.

III. EXPERIMENTS AND ANALYSIS

A. Datasets

Our experiments are performed on two popular and public available datasets, SYSU-MM01 [20] and RegDB [21].

SYSU-MM01. It is a challenging and large-scale RGB-IR Re-ID dataset, which contains 30,071 RGB images and 15,792 IR images from 491 identities captured by four RGB cameras and two IR cameras. The dataset is divided into a training set and a testing set. The training set contains 22,580 RGB images and 11,909 IR images from 395 persons. There are two test modes for SYSU-MM01, *i.e. all-search* and *indoor-search*. In each mode, there are two settings can be chosen, *i.e. single-shot* and *multi-shot*.

TABLE II
COMPARISON RESULTS(%) WITH THE STATE-OF-THE-ART CROSS-MODALITY RE-ID METHODS ON THE REGDB DATASET.

Methods	visible2thermal			
	R1	R10	R20	mAP
Zero-Padding [20] $_{ICCV17}$	17.75	34.21	44.35	18.90
HCML [7] $_{AAAI18}$	24.44	47.53	56.78	20.08
eBDTR [6] $_{TIFS19}$	34.62	58.96	68.72	33.46
MAC [8] $_{MM19}$	36.43	62.36	71.63	37.03
D^{2}RL [12] $_{CVPR19}$	43.4	66.10	76.30	44.1
MSR [9] $_{TIP19}$	48.43	70.32	79.95	48.67
D-HSME [10] $_{AAAI19}$	50.85	73.36	81.66	47.00
AlignGAN [13] $_{ICCV19}$	57.9	-	-	53.6
XIV [16] $_{AAAI20}$	62.21	83.13	91.72	60.18
Hi-CMD [23] $_{CVPR20}$	70.93	86.39	-	66.04
Ours	**80.67**	**87.72**	**90.45**	**78.83**

TABLE III
COMPARISON RESULTS(%) WITH THE BASELINE AND THE AGW USING THE SAME BACK-BONE ON THE SYSU-MM01 AND REGDB DATASETS.

Methods	RegDB		SYSU-MM01			
			all-search		indoor-search	
	R1	mAP	R1	mAP	R1	mAP
Baseline	65.79	64.69	42.83	41.97	47.66	56.50
AGW [22] $_{2020}$	70.05	66.37	47.50	47.65	54.17	62.97
Ours	**80.67**	**78.83**	**51.05**	**49.63**	**55.93**	**63.38**

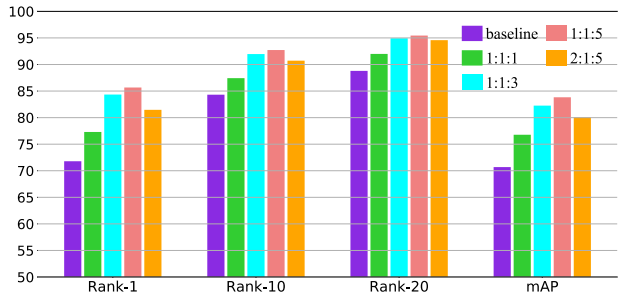

Fig. 6. Performance of baseline and our method with respect to the hyperparameters over RegDB dataset. Different colors mean the different proportion of α, β and λ.

RegDB. There are totally 412 identities in the dataset, where each person has 10 RGB images and 10 thermal images. All the identities are randomly divided into two halves, one for training and the other for testing. The training set thus has 2,060 visible images and 2,060 thermal images, and the same goes for the testing set.

B. Evaluation Metrics

In this paper, The Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP), which are widely used in cross-modality person Re-ID tasks, are applied on our experiments. For more stable results, the testing stage is run ten trials with randomly choosing query and gallery images every time.

C. Implementation Details

The experiments are implemented with PyTorch. Images are resized to 288×144 following [22]. The ResNet50 with pre-trained parameters on ImageNet is used as the backbone of the feature extractor. We use the stochastic gradient descent to optimize the network, and the momentum parameter is set to 0.9. The training epochs are set to 60 and 80 separately for the RegDB dataset and SYSU-MM01 dataset. The initial learning rate is set to 0.01 with a warm-up strategy. The batch size is set to six and an identity-balanced sampling strategy [6] is applied at each training step. In experiments, we randomly select six identities, and then four RGB and four IR images for each identity in each batch. The two margin parameters in (4) and (7) are set to 0.5 and 0.3 respectively. In SYSU-MM01 dataset, the trade-off hyperparameters α, β and γ are all set to 1. In RegDB, the proportion of α, β and γ is set to 1:1:5.

D. Comparison with State-of-the-art Methods

The results over SYSU-MM01 dataset are shown in Table I. The *all-search* and *indoor-search* modes both with *single-shot* choice are applied on our experiments. Our method outperforms the state-of-the-arts by 8.65% and 8.93% with regard to Rank-1 and mAP scores.

The results over RegDB dataset are shown in Table II. The *visible2thermal* mode is applied, which means the visible images are taken as the query images, and the thermal images form the gallery set. Our method still achieves improvements of 9.74% and 12.79% compared with the state-of-the-arts with regard to Rank-1 and mAP scores.

The improvements on RegDB dataset are more significant than the SYSU-MM01 dataset. For RegDB, RGB and thermal images of the same person contain appearance information with high similarity. For SYSU-MM01, it has low similarity, as shown in Fig. 5. The proposed base-derivative framework mainly focuses on the channel-level modality discrepancy, thus it works better on the RegDB dataset, which has images with a tiny discrepancy in appearance-level.

E. Model Analysis

Ablation study. To further analyze the effectiveness of the proposed framework, comparisons with the baseline and the AGW method [22] are performed. The results are shown in Table III. Here, 'baseline' means the network only contains the visible and infrared modality, and is trained using ID loss, intra-modality triplet loss and cross-modality triplet loss, just like the index 1 in Table IV. And 'AGW' uses the same backbone as us. The comparisons illustrate that our base-derivative framework can significantly improve the performance of baseline by 6.88%- 13.88% on Rank-1 and mAP score. In addition, our method also outperforms the AGW method with improvements of 10.62% and 12.46% in terms of Rank-1 and mAP score on RegDB dataset.

Moreover, ablation study about the loss functions are shown in Table IV. 'B1' means the network is only trained with ID loss and 'B2' contains the triplet loss intra- and cross-modality. Results show that the combination of ID loss and triplet loss (MDL + MML) is more effective.

Then, more evaluations are performed to verify the effectiveness of the two auxiliary derivative modalities and each

TABLE IV
ABLATION STUDY ON THE SYSU AND REGDB DATASET. 'B' MEANS 'BASELINE' AND THE SUPERSCRIPT REPRESENTS DIFFERENT TYPES OF MML AND MDL LOSS.

Methods	Modality				Loss						RegDB		SYSU-MM01			
					ID	MML	MDL						all-search		indoor-search	
	I	V	D_I	D_V			I,V	V,D_I	I,D_V	D_I,D_V	R1	mAP	R1	mAP	R1	mAP
B1	✓	✓	×	×	✓	×	×				45.34	39.79	33.49	33.69	36.82	47.14
B2	✓	✓	×	×	✓	✓	✓	×	×	×	65.79	64.69	42.83	41.97	47.66	56.50
Ours																
B2+MML+MDL^1	✓	✓	✓	×	✓	✓	✓	✓	×	×	79.98	77.76	41.58	42.33	46.02	55.11
B2+MML+MDL^2	✓	✓	×	✓	✓	✓	✓	×	✓	×	75.42	74.17	48.14	47.06	53.97	62.15
B2+MML+MDL^3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	79.21	76.85	50.66	**49.64**	**56.57**	**64.11**
B2+MML+MDL^4	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	**80.67**	**78.83**	**51.05**	49.63	55.93	63.38

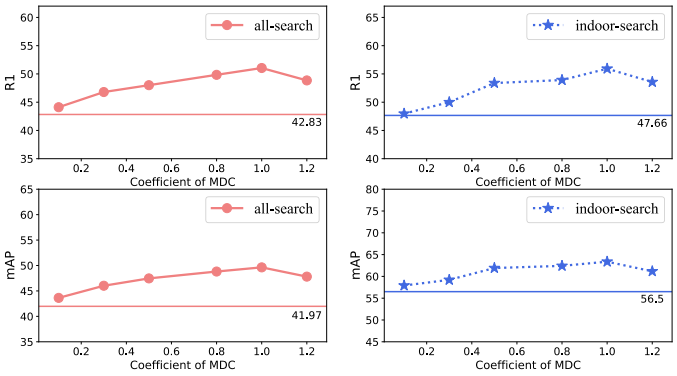

Fig. 7. Performance of our method with respect to different values of hyperparameters λ over SYSU-MM01 dataset. The straight lines in each figure stand for the performance of baseline.

Fig. 8. Images reference to D_V. Different identities are separated by dotted lines. Group (a) contains the original visible images and generated D_V images. Group (b) contains the single-channel visible (the upper row) and D_V (the lower row) images in R,G,B mode. Group (c) contains the single-channel visible (the upper row) and D_V images (the lower row) in gray mode. Details are in Section III-E

single bi-directional cross-modality loss of MDL on RegDB and SYSU-MM01 dataset as in Table IV . For SYSU-MM01 dataset, the experiments are performed with *indoor-search* and *all-search* modes. The experiment with superscript 1 is set to verify the existence of the D_I and its correlative cross-modality loss. The same goes for superscript 2 and D_V. Results show that either of the derivative modalities benefits the cross-modality learning and the combination of four modalities achieves the best performance. The experiment with superscript 4 confirms the settings in our results mentioned before. The experiment with superscript 3 adds the constraint between D_V and D_I. Because in the testing stage, the query images are RGB or IR, so general operations try to build the constraints at least containing visible or infrared modalities for better results. Based on that, the index 5 is set to explore the effect of constraint between modalities not used in the testing stage. Results show that the index 5 has close performances to index 4, which indicates that MDL provides a unified cohesive force for the four modalities and fundamentally promotes cross-modal learning.

Parameters analysis. Two experiments are performed to evaluate the values of α, β and λ in (9) on the two datasets. First, experiments with four different proportions of weights are performed on RegDB dataset. Results are shown in Fig. 6, which prove our method is vastly superior to the baseline. In addition, the superiority appears more obvious with the increasing weight of MDL. Second, experiments with six different weights of MDL under two modes are performed on SYSU-MM01 dataset. Results are shown in Fig. 7, the accuracies increase along with the coefficient of MDL, and the best result appears when the weight is 1. Both R1 and mAP scores are always better than the baseline no matter in the *indoor-search* or the *all-search* mode.

Discussions. We try to visualize D_V and D_I on RegDB dataset. Fig. 8 shows three groups of visualizations. Group (a) shows that the 'Red' channel owns a larger proportion in generated D_V images. Group (b) is set to observe the color information. Results show that D_V images still keep the similar color information with the RGB ones. In group (c), single-channel RGB images are shown in gray mode (best viewed), which aims to contrast the pixel value in three channels. Darker color means larger value. As we can see, the three channels of visible images are not exactly the same, and

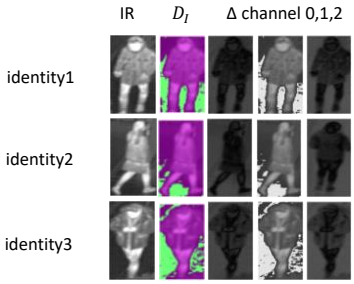

Fig. 9. Images reference to D_I. There are the original thermal images (1^{st} column), generated D_I images (2^{nd} column), and the single-channel divergence images in gray mode, which are calculated using subtraction pixel-by-pixel (3^{th}-5^{th} columns). Details are in Section III-E

differences are more obvious on the images surrounded by a red frame. But the three channels of D_V images are almost the same.

The visualizations about D_I are shown in Fig. 9. For 3^{th}-5^{th} columns, the darker color means the higher similarity. As we can see, the D_I images show obvious discrepancy with the IR images especially in the channel 1. The channel 0 and 2 also show the differences mainly in the background.

IV. Conclusions

This paper proposes a base-derivative framework for cross-modality person Re-ID. To reduce the discrepancy between the visible and infrared modalities, two auxiliary derivative modalities, which inherit the identity information from the base modalities respectively, are generated using a common lightweight generator. The derived images are taken as the input of feature learning network together with base images, which provide more discriminative information. Accordingly, multi-mode intra-modality loss and multi-directional cross-modality loss are designed to promote the reduction of intra- and cross-modality discrepancy. Experimental results on two publicly available datasets SYSU-MM01 and RegDB demonstrate the superiority of the proposed approach.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No.U1613209,61673030), National Natural Science Foundation of Shenzhen (No.JCYJ20190808182209321).

References

[1] S. Gong, M. Cristani, S. Yan, and C. Loy, "Person re-identification," Springer, 2014.

[2] H. Liu, W. Shi, W. Huang, and Q. Guan, "A discriminatively learned feature embedding based on multi-loss fusion for person search," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1668-1672.

[3] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normneuralization neck for deep person reidentification," arXiv preprint arXiv: 1906.08332, 2019.

[4] W. Shi, H. Liu, F. Meng, and W. Huang, "Instance enhancing loss: Deep identity-sensitive feature embedding for person search," Proceedings of the IEEE International Conference on Image Processing (ICIP), 2018, pp. 4108–4112.

[5] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5390-5399.

[6] M. Ye, X. Lan, Z. Wang, and P. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person reidentification," IEEE Transactions on Information Forensics and Security (TIFS), vol. 15, pp. 407–419, 2020.

[7] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," the Association for the Advance of Artificial Intelligence (AAAI), 2018, pp. 7501–7508.

[8] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," ACM International Conference on Multimedia (ACM MM), 2019, pp. 347–355.

[9] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," IEEE Transactions on Image Processing (TIP), vol. 29, pp. 579–590, 2020.

[10] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: Hypersphere manifold embedding for visible thermal person re-identification," the Association for the Advance of Artificial Intelligence (AAAI), 2019, pp. 8385–8392.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Neural Information Processing Systems Conference (NIPS), 2014, pp.2672-2690.

[12] Z. Wang, Z. Wang, Y. Zheng, Y. Chuang, and S. Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," Internaltional Conference on Computer Vision and Pattern Recogintion (CVPR), 2019, pp. 618–626.

[13] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," IEEE International Conference on Computer Vision (ICCV), 2019, pp. 3623–3632.

[14] Z. Zhong, L.Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 598–607.

[15] K. Jungling, and M. Arens, "Local feature based person reidentification in infrared image sequences," IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 448–455.

[16] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," the Association for the Advance of Artificial Intelligence (AAAI), 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[18] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," ACM International Conference on Multimedia (ACM MM), 2019, pp. 57–65.

[19] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv: 1703.07737, 2017.

[20] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5380–5389.

[21] 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, pp. 605, 2017.

[22] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and Hoi, Steven C. H., "Deep learning for person re-identification: A survey and outlook", arXiv preprint arXiv: 2001.04193, 2020, pp.1-9.

[23] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared Person Re-Identification," arXiv preprint arXiv: 1912.01230, 2019.

[24] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 677–683.

[25] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z. Hou, "Cross-modality paired-images generation for RGB-infrared person re-identification," the Association for the Advance of Artificial Intelligence (AAAI), 2020.