Towards a Unified Middle Modality Learning for Visible-Infrared Person Re-Identification

Yukang Zhang, Yan Yan, Yang Lu, Hanzi Wang*
Fujian Key Laboratory of Sensing and Computing for Smart City,
Xiamen University, Xiamen, China.
zhangyk@stu.xmu.edu.cn,{yanyan,luyang,hanzi.wang}@xmu.edu.cn

ABSTRACT

Visible-infrared person re-identification (VI-ReID) aims to search identities of pedestrians across different spectra. In this task, one of the major challenges is the modality discrepancy between the visible (VIS) and infrared (IR) images. Some state-of-the-art methods try to design complex networks or generative methods to mitigate the modality discrepancy while ignoring the highly non-linear relationship between the two modalities of VIS and IR. In this paper, we propose a non-linear middle modality generator (MMG), which helps to reduce the modality discrepancy. Our MMG can effectively project VIS and IR images into a unified middle modality image (UMMI) space to generate middle-modality (M-modality) images. The generated M-modality images and the original images are fed into the backbone network to reduce the modality discrepancy. Furthermore, in order to pull together the two types of M-modality images generated from the VIS and IR images in the UMMI space, we propose a distribution consistency loss (DCL) to make the modality distribution of the generated M-modalities images as consistent as possible. Finally, we propose a middle modality network (MMN) to further enhance the discrimination and richness of features in an explicit manner. Extensive experiments have been conducted to validate the superiority of MMN for VI-ReID over some state-of-the-art methods on two challenging datasets. The gain of MMN is more than 11.1% and 8.4% in terms of Rank-1 and mAP, respectively, even compared with the latest state-of-the-art methods on the SYSU-MM01 dataset.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

VI-ReID, Non-Linear, Middle Modality, Distribution Consistency

ACM Reference Format:
Yukang Zhang, Yan Yan, Yang Lu, Hanzi Wang. 2021. Towards a Unified Middle Modality Learning for Visible-Infrared Person Re-Identification. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475250

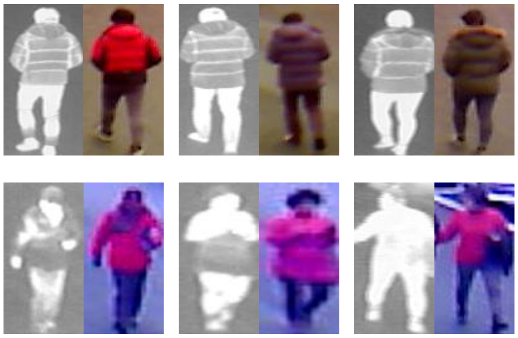


Figure 1: Six sample image pairs of the RegDB dataset, where the images in each pair share the same identity. The relationship between the VIS and IR images is highly non-linear.

1 INTRODUCTION

Given a query person and a database from different modalities, visible-infrared person re-identification (VI-ReID) is a task of retrieving the identities, which are the most relevant to the query person in the database [34, 36, 43, 45]. Due to its important role in the intelligent public security field, VI-ReID has become one of the most popular research areas in both academia and industry [3, 24, 46].

Typically, the existing methods [3, 24, 39, 40] estimate some embedding functions that map the inputs of visible (VIS) and infrared (IR) images into a common embedding space, such that the cross-modality retrieval task becomes the VIS-ReID retrieval task in the Euclidean space. However, because the modality discrepancy between the VIS and IR images is highly non-linear [5, 20], building a common representation space for VI-ReID is challenging, as can be seen from Figure 1.

There are two popular types of methods for the VI-ReID problem. For example, some traditional methods [3, 7, 39, 40, 45] try to find a specific embedding space, where the discrepancy between different modalities can be minimized. These methods typically train an end-to-end network to obtain a modality-invariant embedding. It may be beneficial to pull together samples from the same identities with different modalities as much as possible. Although these methods improve the performance of VI-ReID to a certain extent, relying on the vanilla models does not take into account the non-linear relationship between the VIS and IR images. The other type of methods is the image-level methods (such as [34–36]), which aims to bridge the modality discrepancy by transforming the

[*] Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475250

images from one modality to the other using the DNN-based image processing such as the generative adversarial networks (GANs) [11]. However, this strategy requires complex generative and discriminative networks and there is still a gap between the generated images and the real images. To handle with the above issues, Li *et al.* [20] design an extra lightweight network to generate a novel X-modality from VIS images to treat the VIS-IR dual-modality learning as a three-modality learning problem. Although the spatial information of the original image is preserved, the modality discrepancy between the generated X-modality and IR images still exists. Because of the non-linear relationship, these GAN-based methods can not completely generate the images of one modality from the other modality without changing the identity of a person.

In this paper, we address this problem from the following two perspectives: (a) We introduce a non-linear network to alleviate the modality discrepancy between the VIS and IR images; (b) We project the VIS and IR images into a unified middle modality image (UMMI) space to help to reduce the modality discrepancy between them. Based on the above aims, we design a non-linear middle modality generator (MMG), which respectively encodes the VIS and IR images with two non-linear encoders, and then projects them into a UMMI space through two parameter-shared modality decoders, so that the generated middle-modality (M-modality) images have a unified middle modality. Thus, it greatly alleviates the modality discrepancy between the VIS and IR images. The proposed method reduces the modality discrepancy with an easy-to-implement lightweight network while keeping the inter-identity discriminative power by using the standard cross-entropy loss and the triplet loss [14]. Furthermore, in order to further reduce the discrepancy of the M-modality images generated from the VIS and IR images, we propose a distribution consistency loss (DCL) to minimize the distance between the M-modality images generated from the VIS and IR images. Inspired by the work of PCB [30] in effectively extracting discriminative features, we also use it to improve the performance of our method. With the incorporation of MMG, DCL and PCB into an end-to-end learning framework, the proposed method achieves an impressive performance on two challenging VI-ReID datasets.

The main contributions of this paper can be summarized as follows:

(1) We propose a non-linear middle modality generator to generate middle modality images to assist the VI-ReID task. In particular, the proposed middle modality generator can effectively project the VIS and IR images into a unified middle modality image space.

(2) We propose an effective distribution consistency loss to make the two types of middle-modality images obtained from the VIS and IR images consistent in modality distribution in the UMMI space, which further improves the performance of the proposed method.

(3) Extensive experiments show that the proposed method outperforms the other competing methods significantly on both SYSU-MM01 and RegDB datasets.

2 RELATED WORK

2.1 Single-Modality Person Re-Identification

Recently, some researches [6, 8, 28, 30, 47] focus on discovering discriminative representations, which are invariant to illumination, resolution, human pose, occlusion and other disturbing factors, under the VIS light for the task of person ReID. For this purpose, several part-based methods (such as PCB [30], MGN [33] and Pyramid [48]) horizontally divide an input image or a feature map into several parts by taking advantage of part-level cues. However, this strategy often requires relatively well-aligned body parts. Thus, the authors need to design a complex network to incorporate different levels of features. Compared with the part-based methods that horizontally divide the feature maps, some other methods leverage the person structure priors, such as semantic parsing knowledge [17, 51], person attribute [15, 32] and pose estimation [9, 21, 26], to align human parts for extracting accurate and robust part-level features from human body parts. However, those methods are prone to introduce some additional noise from wrong semantic parsing and pose estimation. Besides, inspired by the powerful generation ability of GANs [11], some other methods [10, 23, 37, 49, 52] utilize them to generate new images to alleviate the shortage of datasets. Although the above methods have achieved state-of-the-art performance, they are not suitable for VI-ReID.

2.2 Visible-Infrared Person Re-Identification

Recently, VI-ReID has become an important topic. The aim of VI-ReID is to match persons with different modalities captured by the VIS and IR cameras. There are two main categories in this direction: one is to extract the modality-invariant features between the VIS-IR images; the other is based on the generative methods aiming at exploring an effective transformation to reduce the modality discrepancy.

The former type of methods attempts to design a network to find a modality-shared feature space, in which the modality discrepancy is minimized. For example, Wu *et al.* [40] propose a deep zero-padding network towards automatically evolving modality-specific nodes. Feng *et al.* [7] utilize the modality-related information and extract modality-specific representations by constructing an individual network for each modality. Inspired by adversarial learning, Dai *et al.* [3] design a cutting-edge generative adversarial training-based discriminator to learn discriminative feature representations from different modalities. Ye *et al.* [45] propose a novel dynamic dual-attentive aggregation learning method by mining both intra-modality and cross-modality contextual cues for VI-ReID. Wu *et al.* [39] exploit the intra-modality similarity to guide the learning of cross-modality similarity along with the alleviation of modality-specific information. Lu *et al.* [24] try to introduce modality-specific features based on the cross-modality near the neighbor affinity model. Pu *et al.* [29] adopt a variational auto-encoder to disentangle an identity-discriminable and an identity-ambiguous cross-modality feature subspace. However, most of these methods mainly focus on reducing the modality discrepancy via the feature alignment, while they ignore the large variations in the image space.

The second type of methods try to reduce the modality discrepancy by transforming one modality into the other one. For example, Wang *et al.* [36] propose to transform IR images into their VIS counterpart and transform VIS images into their IR version. AlignGAN [34] is proposed to transform VIS images into IR images with effective constraints for both images and features alignment, and it achieves significant performance improvement. JSIA-ReID [35] is proposed to generate cross-modality paired-images and perform

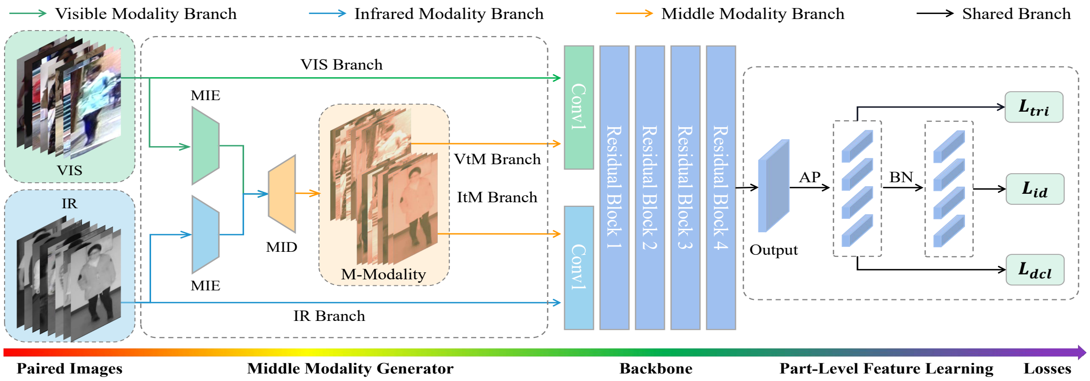


Figure 2: Framework of the proposed MMN, which is built upon a two-stream ResNet50 model. The generated M-modality images and the original images are fed into the backbone network to reduce the modality discrepancy.

both global set-level and fine-grained instance-level alignments. Hi-CMD [2] introduces an effective generator to extract pose- and illumination-invariant features and maintain the identity characteristic of a specific person. Most recently, Li *et al.* [20] propose to generate a novel X-modality from VIS images to treat the VIS-IR dual-modality learning as a three-modality learning problem. Although the network of X-Modality is a lightweight network, there is still a large modality discrepancy between the X-modality and the IR modality. Different from these methods, the proposed method not only adopts a lightweight network as the X-modality, but also effectively projects the VIS and IR modalities into a UMMI space, by which significant improvement can be achieved.

3 METHOD

To reduce the modality discrepancy and improve performance, we propose a unified middle modality generator (MMG) and a distribution consistency loss (DCL). In this section, we first present an overview of the architecture of our middle modality network (MMN). Then we elaborate the design of MMG and the DCL. Finally, we adopt a multi-loss strategy to jointly optimize the proposed end-to-end MMN method.

3.1 Model Architecture

Figure 2 provides an overview of the proposed MMN method. The input of the MMN is a pair of VIS-IR images. The VIS and IR images are fed into the proposed MMG module to generate middle-modality (M-modality) images. The generated M-modality images with the original VIS and IR images are fed into a two-stream ResNet50 [13, 45] backbone to extract the modality-invariant features, where the first convolutional block in each stream is different to capture modality-specific low-level representations while the middle and deep convolutional blocks are shared to learn modality-shared middle- and deep-level representations. Inspired by the work of PCB [30] in extracting discriminative features, which horizontally divides the feature maps into several parts and each part is fed into a classifier to learn local cues, we also use it to improve the performance of the proposed model and the part is set to 4. We further modify the stride of the last convolutional block to 1 in the backbone network to preserve more spatial information [30]. After the convolutional layers with average pooling (AP) layer, we add a batch normalization (BN) layer, which is parameter-shared among all modality images, to make the loss easier to converge [25]. Finally, the features before and after the BN layer are fed into different loss functions to jointly optimize the network.

3.2 Middle Modality Generator (MMG)

For convenience, we first define the VI-ReID task. Our MMG takes an image pair of two modalities of the same identity as the input. Let $I = \left\{I_{VIS}, I_{IR} | I_{VIS}, I_{IR} \in \mathbb{R}^{3\times H\times W}\right\}$ denote the pairs of VIS and IR images from a dataset, respectively. I is a set of image pairs. $3\times H\times W$ correspond to the channel, hight and weight, respectively. All input images are resized to $3 \times 384 \times 192$, so we can treat an IR image as a three-channel image. In fact, it is a gray image with only one channel. We set $I_m = \left\{I_{VtM}, I_{ItM} | I_{VtM}, I_{ItM} \in \mathbb{R}^{3\times H\times W}\right\}$ as the M-modality VtM and ItM images generated from the I_{VIS} and I_{IR} images.

MMG includes the modality information encoder (MIE), which encodes different modality information with two encoders that do not share parameters, and the modality information decoder (MID), which projects the VIS and IR images into a UMMI space with two decoders that share parameters. Figure 3 provides an overview of the proposed MMG module.

3.2.1 *Modality Information Encoder.*

Since the IR images have one channel containing the information of IR light, while the VIS images have three channels containing the colour information of VIS light, we first align the VIS images and the IR images at the channel level. According to [34], it is easier to transform VIS to IR than IR to VIS.

Based on the above analysis, we propose two independent MIEs $\mathcal{F}_{VtC}$ and $\mathcal{F}_{ItC}$ to encode the two modalities, respectively.

For the VIS modality, we have:

$$I_{VtC} = \mathcal{F}_{VtC}(I_{VIS}), \tag{1}$$

For the IR modality, we have:

$$I_{ItC} = \mathcal{F}_{ItC}(I_{IR}), \tag{2}$$

Since the modality discrepancy between the VIS and IR modalities mainly comes from the channel level [20], the proposed MIEs

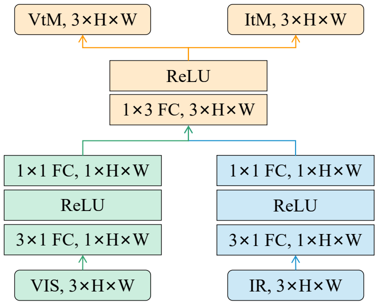


Figure 3: Framework of the proposed MMG, which includes two MIEs that do not share the same parameters, and two MIDs that share the same parameters.

$\mathcal{F}_{VtC}$ and $\mathcal{F}_{ItC}$ first utilize two 3×1 fully-connected layers at the channel level to encode the three-channel image I_{VIS} and I_{IR} into one-channel data $I_C = \left\{I_{VtC}, I_{ItC} | I_{VtC}, I_{ItC} \in \mathbb{R}^{1 \times H \times W}\right\}$. Then, since the relationship between VIS and IR images is highly non-linear, the ReLU activation layer [18] is used to increase the non-linear ability of the network. Finally, two 1×1 fully-connected layers are used to further adjust I_C.

Through the above operations, we get the features of VIS and IR images and perform the alignment between the VIS images and the IR images at the channel level.

3.2.2 Modality Information Decoder. The proposed MID is used to project the data encoded as one channel into a unified three channel image space. In this unified image space, the distance between the VIS and IR images becomes closer, and thus we can reduce the modality discrepancy.

Here, we adopt two parameter-shared MIDs $\mathcal{F}_{CtM}$ to decode the one-channel images I_c into the three-channel images $I_m = \left\{I_{VtM}, I_{ItM} | I_{VtM}, I_{ItM} \in \mathbb{R}^{3 \times H \times W}\right\}$. That is:

$$I_{VtM} = \mathcal{F}_{CtM}(I_{VtC}), \quad I_{ItM} = \mathcal{F}_{CtM}(I_{ItC}), \tag{3}$$

The proposed MID includes a 1×3 fully-connected layer at the channel level followed by a ReLU activation layer to obtain the three-channel middle-modality images I_m. The ReLU activation layer is used to further increase the non-linear relationship. Through the above operations, we can generate the M-modality images.

The generated M-modality images have the same labels as the VIS images and IR images. At last, the M-Modality, VIS and IR modality images are fed into the backbone network together to assist the VI-ReID task for reducing the modality discrepancy. By considering the non-linear relationship between the VIS and IR images, and projecting them into a UMMI space, the proposed MMG module can promote the optimization process of the backbone network. Meanwhile, the optimization of the backbone network can effectively promote the learning of MMG, and further improve the quality of the generated M-modality images. Therefore, the proposed end-to-end MMN scheme can make these two networks benefit from each other.

3.3 Distribution Consistency Loss (DCL)

In order to further pull together the two types of M-modality images generated by I_{VIS} and I_{ir}, we propose a distribution consistency loss (DCL) for the proposed M-modality images. DCL is calculated as:

$$\mathcal{L}_{dcl} = \frac{1}{N} \sum_{i=1}^{N} mean[f(I_{VtM}^{i}) - f(I_{ItM}^{i})], \tag{4}$$

where N is the number of I_{VtM} and I_{ItM} images in a batch during the training phase, $f(\cdot)$ is the output of the proposed network followed by two fully-connected layers. $mean[A - B]$ is the mean operation of the difference of A and B. It is obvious that the optimization of DCL would make two types of M-modality features to be similar.

3.4 Multi-Loss Optimization

The generated M-modality images are fed into a two-stream ResNet50 backbone network [13, 45] together with the original VIS and IR images to assist the optimization of the network. In the proposed MMN, besides the proposed DCL, we also combine the label-smoothed cross-entropy loss [25] and the triplet loss [14] to jointly optimize the MMN.

The label-smoothed cross-entropy loss can prevent the estimated model from overfitting [25], which is a widely used method for a classification task. The label-smoothed cross-entropy loss is formulated as:

$$\mathcal{L}_{id} = \sum_{i=1}^{C} -q_i \log(p_i), q_i = \begin{cases} 1 - \frac{C-1}{C}\varepsilon & i = y \\ \varepsilon / C & i \neq y \end{cases}, \tag{5}$$

where C is the numbers of person IDs in the training set, y represents the person ID label, p_i is the ID prediction logits of class i, ε is a small constant to encourage the model to be less confident on the training set [25, 31]. Following the previous works [25], ε is set at 0.1.

Then, we leverage the triplet loss [14], which helps to minimize the intra-class similarity and maximize the inter-class similarity for metric learning. Because the triplet loss can decrease the distances between positive samples and increase the distances between negative samples, it can be used for the VI-ReID task to greatly reduce the modality discrepancy. Since we use M-modality images together with the original VIS and IR images to assist the training of the MMN, we form a batch with the size of $4M$, where M represent the number of input images per modality. We set the first M is for I_{VIS}, the second M for I_{VtM}, the third M for I_{ItM}, and the fourth M for I_{IR}. For I_{VIS} and I_{IR}, the cross-modality triplet loss between I_{VIS} and I_{IR} is formulated as:

$$\mathcal{L}_{tri}^{(V,I)} = \mathcal{L}_{tri}(V, I) + \mathcal{L}_{tri}(I, V), \tag{6}$$

where $\mathcal{L}_{tri}(V, I)$ represents that the positive sample pairs are from the VIS and IR modalities, and the negative sample pairs are from the VIS modality. Then we have:

$$\begin{aligned} \mathcal{L}_{tri}(V, I) = \sum_{i=1}^{M} [\xi + \max_{\substack{j=3M+1,\cdots,4M \\ y_i = y_j}} D(V_i, I_j) \\ - \min_{\substack{k=3M+1,\cdots,4M \\ y_i \neq y_k}} D(V_i, I_k)]_+, \end{aligned} \tag{7}$$

Table 1: Performance on the RegDB and SYSU-MM01 datasets. R-1, R-10, R-20 denotes the Rank-1, Rank-10, Rank-20 accuracy, respectively. Here, all the results obtained by the competing methods are the results using the single-query mode. The bold font represents the best performance and the underline means the second performance. The “†” symbol represents the feature extraction based methods, the “*” symbol represents the image generation based methods and the “§” symbol represents that the method uses a lightweight network to generate new images.

Model	Pub.	RegDB								SYSU-MM01							
		Visible to Infrared				Infrared to Visible				All Search				Indoor Search			
		R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
†Zero-Padding[40]	ICCV’17	17.8	34.2	44.4	18.9	16.6	34.7	44.3	17.8	14.8	54.1	71.3	15.9	20.6	68.4	85.8	26.9
†HCML[44]	AAAI’18	24.4	47.5	56.8	20.8	21.7	45.0	55.6	22.2	14.3	53.2	69.2	16.2	24.5	73.3	86.7	30.1
†MHM[41]	AAAI’20	31.1	47.0	58.6	32.1	-	-	-	-	35.9	73.0	86.1	38.0	-	-	-	-
†BDTR[46]	IJCAI’18	33.6	58.6	67.4	32.8	32.9	58.5	68.4	32.0	17.0	55.4	72.0	19.7	-	-	-	-
†MAC[43]	MM’19	36.4	62.4	71.6	37.0	-	-	-	-	33.3	79.0	90.1	36.2	33.4	82.5	93.7	45.0
†cmGAN[3]	IJCAI’18	-	-	-	-	-	-	-	-	27.0	67.5	80.6	27.8	31.7	77.2	89.2	42.2
†MSR[7]	TIP’20	48.4	70.3	80.0	48.7	-	-	-	-	37.4	83.4	93.3	38.1	39.6	89.3	97.7	50.9
†HSME[12]	AAAI’19	50.9	73.4	81.7	47.0	50.2	72.4	81.1	46.2	20.7	62.8	78.0	23.2	-	-	-	-
†SNR[16]	CVPR’20	-	-	-	-	-	-	-	-	34.6	75.9	86.6	33.9	40.9	83.8	91.8	50.4
†expAT[42]	TIP’20	66.5	-	-	67.3	67.5	-	-	66.5	38.6	76.6	86.4	38.6	-	-	-	-
†CMM[22]	MM’20	59.8	80.4	88.7	60.9	-	-	-	-	51.8	92.7	97.7	51.2	55.0	94.4	99.4	63.7
†CMSP[39]	IJCV’20	65.1	83.7	-	64.5	-	-	-	-	43.6	86.3	-	45.0	48.6	89.5	-	57.5
†SSFT[24]	CVPR’20	65.4	-	-	65.6	63.8	-	-	64.2	47.7	-	-	54.1	-	-	-	-
†DDAA[45]	ECCV’20	69.3	86.2	91.5	63.5	68.1	85.2	90.3	61.8	54.8	90.4	95.8	53.0	61.0	94.1	98.4	68.0
†CoAL[38]	MM’20	74.1	90.2	94.5	70.0	-	-	-	-	57.2	92.3	97.6	57.2	63.9	95.4	98.8	70.8
†NFS[1]	CVPR’21	80.5	91.6	95.1	72.1	78.0	90.5	93.6	69.8	56.9	91.3	96.5	55.5	62.8	96.5	99.1	69.8
*D^{2}RL[36]	CVPR’19	43.4	66.1	76.3	44.1	-	-	-	-	28.9	70.6	82.4	29.2	-	-	-	-
*JSIA-ReID[35]	AAAI’20	48.1	-	-	48.9	48.5	-	-	49.3	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
*AlignGAN[34]	ICCV’19	57.9	-	-	53.6	56.3	-	-	53.4	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
*Hi-CMD[2]	CVPR’20	70.9	86.4	-	66.0	-	-	-	-	34.9	77.6	-	35.9	-	-	-	-
*DG-VAE[29]	MM’20	73.0	86.9	-	71.8	-	-	-	-	59.5	93.8	-	58.5	-	-	-	-
§X-Modality[20]	AAAI’20	62.2	83.1	91.7	60.2	-	-	-	-	49.9	89.8	96.0	50.7	-	-	-	-
MMN	**MM’21**	**91.6**	**97.7**	**98.9**	**84.1**	**87.5**	**96.0**	**98.1**	**80.5**	**70.6**	**96.2**	**99.0**	**66.9**	**76.2**	**97.2**	**99.3**	**79.6**

where $D(V_i, I_j)$ is the Euclidean distance between the VIS and IR images, ξ is a margin parameter and $[z]_+ = max(z, 0)$, i and j are from the same identity but different with k. Similiarly, $\mathcal{L}_{tri}(I, V)$ represents the positive sample pairs are from IR and VIS modalities, and the negative samples pairs are from the IR modality. That is:

$$\begin{aligned}\mathcal{L}_{tri}(I, V) = \sum_{i=3M+1}^{4M} [\xi + & \max_{\substack{j=1,\cdots,M \\ y_i=y_j}} D(I_i, V_j) \\ & - \min_{\substack{K=1,\cdots,M \\ y_i \neq y_k}} D(I_i, V_k)]_+,\end{aligned} \tag{8}$$

The other modality triplet losses are similar to this one. The final cross-modality triplet loss function is calculated as:

$$\mathcal{L}_{tri} = \mathcal{L}_{tri}^{(V,I)} + \mathcal{L}_{tri}^{(V,ItM)} + \mathcal{L}_{tri}^{(I,VtM)} + \mathcal{L}_{tri}^{(VtM,ItM)}. \tag{9}$$

The proposed MMN can be optimized in an end-to-end manner by minimizing the joint loss $\mathcal{L}$, which consists of the identity loss $\mathcal{L}_{id}$, the cross-modality triplet loss $\mathcal{L}_{tri}$ and the distribution consistency loss $\mathcal{L}_{dcl}$:

$$\mathcal{L} = \mathcal{L}_{id} + \lambda_1 \mathcal{L}_{tri} + \lambda_2 \mathcal{L}_{dcl}, \tag{10}$$

where λ_1 and λ_2 are the weights of $\mathcal{L}_{tri}$ and $\mathcal{L}_{dcl}$.

4 EXPERIMENTS

In this section, we compare the proposed MMN method with several state-of-the-art methods. We also conduct the ablation studies to analyze the key components in the proposed MMN method.

4.1 Implementation Details

Dataset. The proposed MMN method is evaluated on two challenging VI-ReID datasets, including SYSU-MM01 [40] and RegDB [27].

The SYSU-MM01 dataset [40] contains 491 identities captured by four VIS cameras and two IR cameras, including both indoor and outdoor environments. Each identity is captured by at least one VIS camera and one IR camera. The training set contains 19, 659 VIS images and 12, 792 IR images of 395 identities and the test set contains 96 identities with 3, 803 IR images as the query images. The gallery set is determined by the test mode, which includes the all-search mode and the indoor-search mode. In the all-search mode, all the images captured by the VIS cameras are used as the gallery. In the indoor-search mode, only the images captured by the two indoor VIS cameras are used as the gallery.

The RegDB dataset [27] consists of 412 identities, where each identity has ten VIS images and ten IR images from a pair of overlapping VIS and IR cameras. We use the evaluation protocol in [44], which repeats ten trails with a randomly half-half split of the dataset. One half of identities are used for training and the other half identities are used for testing. The final results are based on the average of the ten times test.

Evaluation Protocol. The standard Cumulative Matching Characteristics (CMC) curve and the mean Average Precision (mAP) are used as the performance evaluation metrics in our experiments.

Experimental details. The proposed method is implemented with PyTorch. The model is trained for 80 epochs in total. The

Table 2: Influence of different modality image generation methods on the SYSU-MM01 dataset.

Methods	SYSU-MM01			
	R-1	R-10	R-20	mAP
baseline	61.68	91.83	96.75	58.51
ItM + VtM	63.16	93.88	97.60	58.48
baseline + ItM	62.10	92.44	96.95	58.94
baseline + VtM	64.56	87.72	93.58	60.66
baseline + VtM + ItM	65.37	93.71	97.51	60.89
baseline + M-Modality	**66.12**	**94.34**	**97.75**	**61.92**

Table 3: Influence of different distance measures on the SYSU-MM01 dataset.

Modes	SYSU-MM01			
	R-1	R-10	R-20	mAP
connect	68.28	94.90	97.98	63.95
minimum	67.52	94.56	97.66	63.17
maximum	68.24	**95.06**	**98.13**	63.88
mean	**68.36**	94.96	98.04	**64.00**

Table 4: Influence of different components on the SYSU-MM01 dataset.

Methods	Settings			SYSU-MM01			
	MMG	DCL	PCB	R-1	R-10	R-20	mAP
baseline				61.68	91.83	96.75	58.51
baseline	✓			66.12	94.34	97.75	61.92
baseline	✓	✓		68.36	94.96	98.04	64.00
baseline			✓	67.15	95.19	98.60	65.27
baseline	✓		✓	69.45	95.06	98.22	65.02
baseline	✓	✓	✓	**70.61**	**96.17**	**98.99**	**66.88**

backbone network employed in our method is first pre-trained on the ImageNet dataset [4], and then fine-tuned on the training images. All the input images are first resized to the size of $3 \times 384 \times 192$, and the random horizontal flip and the random erasing [50] are adopted for data augmentation during the training phase. Besides, we adopt a warm-up strategy [25] to make the training gradient smooth. The initial learning rate is set at 1×10^{-2} and then it linearly increases to 1×10^{-1} in 10 epochs. Then, we decay the learning rate to 1×10^{-2} at 20 epoch, and then we further decay it to 1×10^{-3} at epoch 60. Following [45], in each mini-batch, we randomly select 4 identities with 4 VIS images and 4 IR images for training. The SGD optimizer is adopted for optimization, and the momentum parameter is set to 0.9. For the margin parameter ξ in the triplet loss, following the work in [14], we set it to 0.3. For the λ_1 and λ_2 parameters in Eq. (10), we set them to 1 and 0.5, respectively.

4.2 Comparison with State-of-the-Art Methods

We first compare the proposed MMN method with several other state-of-the-art methods to demonstrate the superiority of MMN method. The competing methods include the methods based on feature extraction (including Zero-Padding [40], HCML [44], MHM [41], BDTR [46], MAC [43], cmGAN [3], MSR [7], HSME [12], SNR[16], expAT[42], CMM [22], CMSP [39], SSFT [24], DDAA [45], CoAL [38], NFS [1]) and the methods based on image generation (including D^{2}RL [36], JSIA-ReID [35], AlignGAN [34], Hi-CMD [2], DG-VAE [29], X-Modality [20]). The results on the RegDB and SYSU-MM01 datasets are reported in Table 1.

From Table 1, we have the following observations. (1) The proposed MMN method significantly outperforms the other state-of-the-art methods by a large margin on both the RegDB and SYSU-MM01 datasets. Specifically, the proposed MMN method obtains the Rank-1 accuracy = (70.6%) and mAP = (66.9%) on the SYSU-MM01 dataset for the all-search settings, and it obtains the Rank-1 accuracy = (91.6%) and mAP = (84.1%) on the RegDB dataset for the visible-to-infrared settings. The results suggest that MMN can learn better modality-shared features between VIS and IR by utilizing the auxiliary M-modality images. (2) All the GAN-based methods (such as D^{2}RL and AlignGAN) employ an encoder-decoder architecture to generate new features or new images to reduce the modality discrepancy between the VIS and IR images. On the one hand, these methods ignore the fact that the modality discrepancy between VIS and IR is very complex due to the non-linear relationship. On the other hand, the GAN-based methods reconstruct modality information at both channel and spatial levels but the modality discrepancy mainly remains at the channel level. These methods destroy the spatial structure information in a certain degree, resulting in a large gap between the generated images and the real images. Compared with these methods, our method transforms the VIS and IR images into the unified M-modality images and it can effectively preserve the original spatial information. Therefore, the proposed MMN method has a great performance advantage over the GAN-based methods. (3) Although X-modality also utilizes an auxiliary modality to reconcile the VIS and IR modalities, our method projects these two modalities into a unified M-modality image space, which is a more effective way to reduce the modality discrepancy between the VIS and IR images.

4.3 Ablation Study

Influence of different modalities. To validate the effectiveness of the proposed MMN and show the advantages of projecting VIS and IR into a UMMI space, we conduct an experiment on the SYSU-MM01 dataset (using the all-search mode) [40] to show which generated method is more effective to reduce the modality discrepancy. We report the experimental results in Table 2. The 'ItM + VtM' represents the proposed method using M-modality but without using the original modality. The 'baseline + ItM' and 'baseline + VtM' respectively represent that we only use one encoder and one decoder to generate a new modality to assist the original modality. The 'baseline + VtM + ItM' represents that we use two encoders and two decoders, which do not share the parameters, to generate a new modality to assist the original modality. From Table 2, we can see that: (1) All the auxiliary modalities can improve the performance of the baseline method, which shows that it is an effective method to reduce the modality discrepancy by generating new modalities through a non-linear encoding and decoding module to assist the VI-ReID task. (2) 'Baseline + VtM' is more effective than 'baseline + ItM'. This is because that transforming VIS into a grayscale image can make the distances between VIS and IR closer. (3) After projecting the VIS and IR images to a UMMI space, the

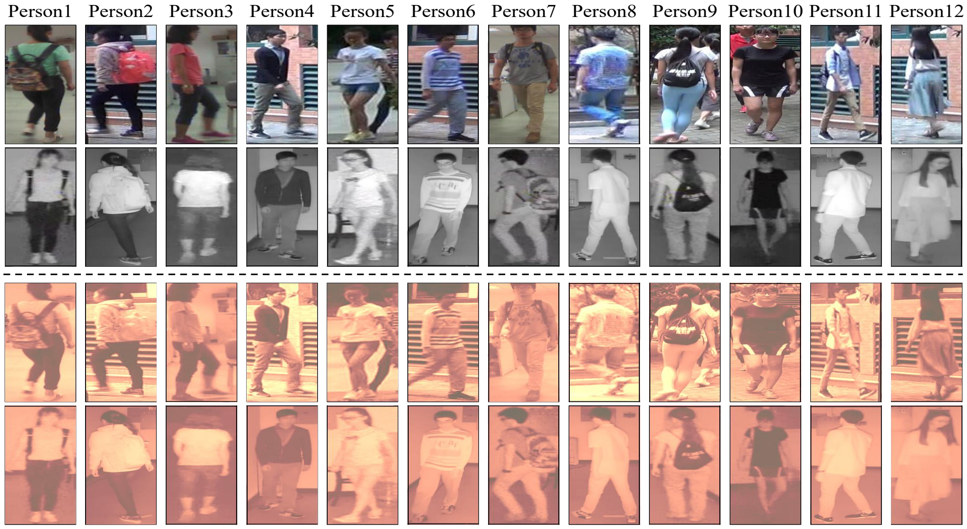


Figure 4: Contrast visualization between the generated M-modality images and the original images. The four images in each column are images of the same person from different modalities. The first rows is VIS images, the second is IR images, the third and the fourth rows are the M-modality images generated from VIS and IR respectively.

performance can be further improved. In terms of Rank-1 and mAP, the proposed M-modality exceeds the baseline by 4.44% and 3.41%, respectively. It shows that the proposed M-modality is an effective data enhancement manner for VI-ReID to reduce the modality discrepancy.

Influence of different distance measures. Since the proposed MMN generated M-modality images to assist the VI-ReID task, we evaluate which distance processing measure is the most proper for MMN. First, we denote the distance between the query images and the gallery images as $D(q, g)$:

$$D(q, g) = [D(V, I), D(V, ItM), D(VtM, I), D(VtM, ItM)], \quad (11)$$

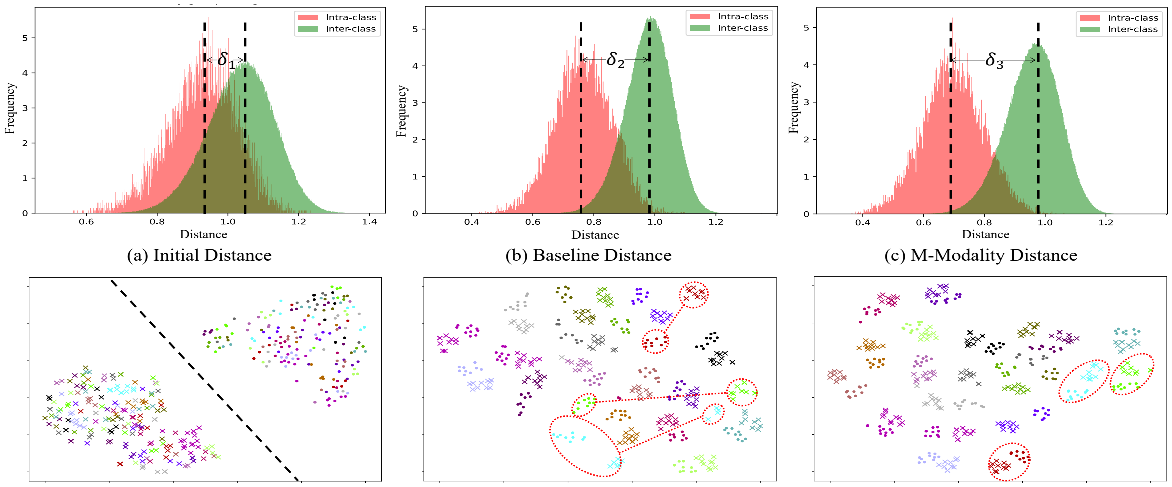
Then, we set up four measures: (1) directly connect the original feature with the generated feature. (2) Taking the minimum value of $D(q, g)$. (3) Taking the maximum value of $D(q, g)$ and (4) taking the mean value of $D(q, g)$. Table 3 shows the experimental results obtained by the four measures. In terms of Rank-1 and mAP, the mean value of $D(q, g)$ has achieved the highest results. Thus, we use it as the final result in our experiment.

Influence of different components. To further demonstrate the influence of each component in the proposed method, we conduct ablation studies by removing certain modules from MMN method and evaluate the variants on the SYSU-MM01 dataset (using the all-search setting). The overall settings remain the same, while only the module under demonstration is added or removed from MMN. Table 4 shows the obtained results. As we can see, (1) the proposed DCL can further improve the performance of MMN, which shows DCL can make the M-modality images generated by MMG consistent in modality distribution. (2) Both MMG and PCB can improve the performance of the baseline. By combining MMG, PCB and DCL, the modality discrepancy is effectively reduced, which makes the proposed MMN achieve better results.

4.4 Visualization

M-modality images. In order to verify whether the proposed MMN can effectively project the VIS images and the IR images into a UMMI space, we visualize some M-modality images. As shown in Figure 4, we can see the following results: (1) compared with the VIS and IR images, the generated M-modality images appear much 'redder', which indicates that the shared feature between the VIS and IR images has been extracted. From the viewpoint of electromagnetic radiation, the wavelength of the R channel of the VIS images is closer with the IR images. Thus, there will be more shared features. (2) As shown in the first column of Figure 4, a person in the green coat in the VIS modality turns into the person with a white coat under the IR modality, which indicates the large modality discrepancy between the VIS and IR images. However, after passing through the proposed MMG module, both the VIS and IR modalities turn into a unified modality. This shows that the proposed MMG module can effectively project the VIS and IR modalities into a unified modality, by which it can make the distances between the two modalities closer. The proposed M-modality is a unified middle modality. It can not only learn the information-shared between the VIS and IR images, but also reduce the modality discrepancy effectively.

Feature distribution. Considering the non-linear relationship between the VIS and IR images, we introduce a novel method to generate unified M-modality images. In order to further analyze the reason why our method is effective, we conduct some experiments on the SYSU-MM01 dataset to compute the frequency of inter-class and intra-class distances. Figure 5 (a) shows the initial distance distributions of the inter-person and intra-person pairs, Figure 5 (b) shows the distance distributions obtained by the baseline and Figure 5 (c) shows the distance distributions obtained by the proposed MMN method. Comparing Figure 5 (c) with Figure 5 (a) and 5 (b), the means of inter-class and intra-class distances are seperated far away by using the proposed method. It shows the proposed MMN



(a) Initial Distance (b) Baseline Distance (c) M-Modality Distance

(d) Initial Distribution (e) Baseline Distribution (f) M-Modality Distribution

Figure 5: (a-c) The distributions of the three types of distances between the cross-modality features. The intra-class and inter-class distances are indicated by red and green color, respectively. (d-f) Visualization of the corresponding feature space. A total of 20 persons are selected from the test set. The samples with the same color are from the same person. The "dot" and "cross" markers denote the images from the VIS and IR modalities, respectively.

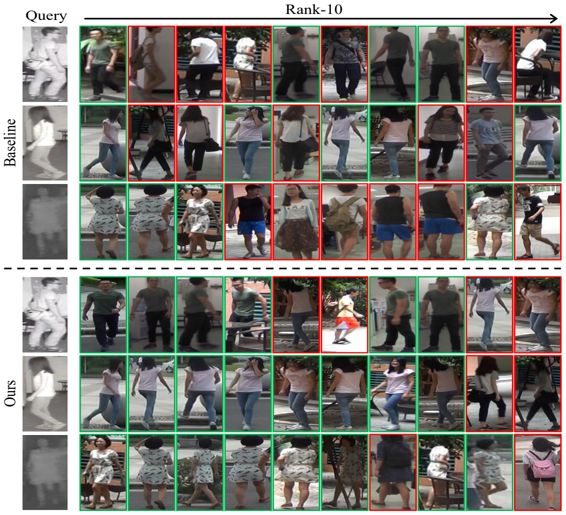


Figure 6: The Rank-10 retrieval results obtained by the baseline and the proposed MMN method on the SYSU-MM01 dataset. For each retrieval case, the query images of the first column are the IR images, and the gallery images are the VIS images. The retrieved VIS images with green bounding boxes have the same identities with the query images, and those with red bounding boxes have different identities with the query images.

method can effectively reduce the intra-class modality discrepancy. For further validating the effectiveness of M-modality, we plot the feature representations in the 2D feature space for visualization using the t-SNE [19] distribution. As shown in Figure 5 (d-f), the results show that the proposed method can greatly shorten the distances between two modality images with the same identity, and effectively reduce the modality discrepancy.

Retrieval result. To further show the benefits of our proposed MMN, we compare MMN method with the baseline on the SYSU-MM01 dataset, using the multi-shot setting and the all-search mode. The obtained Rank-10 ranking results are shown in Figure 6. In general, the proposed MMN method can significantly improve the ranking list with more correctly retrieved images ranked in the top positions.

5 CONCLUSION

In this paper, we propose a non-linear middle modality generator, which helps to reduce the modality discrepancy between the VIS and IR images. The proposed middle modality generator can effectively project VIS and IR images into a unified middle modality image space to generate M-modality images. The generated M-modality images and the original images are fed into the backbone network to reduce the modality discrepancy. Furthermore, to pull together the M-modality images generated from VIS and IR images, we propose a distribution consistency loss to make the modality distribution of the generated M-modality images as consistent as possible. Finally, we propose a middle modality network to further enhance the discrimination and richness of features in an explicit manner. Extensive experiments have demonstrated the superior performance of the proposed middle modality network compared with the state-of-the-art methods.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant 61872307, Grant 62071404, Grant 62002302 and the Natural Science Foundation of Fujian Province under Grant 2020J01001.

REFERENCES

[1] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. 2021. Neural Feature Search for RGB-Infrared Person Re-Identification. In *Proceedings of the IEEE CVPR*. 587–597.

[2] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. 2020. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE CVPR*. 10257–10266.

[3] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. 2018. Cross-Modality Person Re-Identification with Generative Adversarial Training.. In *Proceedings of the IJCAI*. 677–683.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE CVPR*. 248–255.

[5] Zhongying Deng, Xiaojiang Peng, and Yu Qiao. 2019. Residual Compensation Networks for Heterogeneous Face Recognition. In *Proceedings of the AAAI*, Vol. 33. 8239–8246.

[6] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. 2019. Bilinear Attention Networks for Person Retrieval. In *Proceedings of the IEEE ICCV*. 8030–8039.

[7] Zhan-Xiang Feng, Jianhuang Lai, and Xiaohua Xie. 2020. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. In *IEEE TIP*, Vol. 29. 579–590.

[8] Lishuai Gao, Hua Zhang, Zan Gao, Weili Guan, Zhiyong Cheng, and Meng Wang. 2020. Texture Semantically Aligned with Visibility-aware for Partial Person Re-identification. In *Proceedings of the ACM MM*. 3771–3779.

[9] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-Guided Visible Part Matching for Occluded Person ReID. In *Proceedings of the IEEE CVPR*. 11744–11752.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and hongsheng Li. 2018. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification. In *Proceedings of the NeurIPS*, Vol. 31. 1230–1241.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the NeurIPS*, Vol. 27. 2672–2680.

[12] Yi Hao, Nannan Wang, Jie Li, and Xinbo Gao. 2019. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-identification. In *Proceedings of the AAAI*, Vol. 33. 8385–8392.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning For Image Recognition. In *Proceedings of the IEEE CVPR*. 770–778.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of The Triplet Loss for Person Re-identification. *ArXiv* (2017).

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-Aggregation Network for Person Re-identification. In *Proceedings of the IEEE CVPR*. 9317–9326.

[16] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style Normalization and Restitution for Generalizable Person Re-identification. In *Proceedings of the IEEE CVPR*. 3143–3152.

[17] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. 2018. Human Semantic Parsing for Person Re-Identification. In *Proceedings of the IEEE CVPR*. 1062–1071.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the NeurIPS*, Vol. 25. 1106–1114.

[19] Van Der Maaten Laurens and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. In *JMLR*, Vol. 9. 2579–2605.

[20] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. 2020. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In *Proceedings of the AAAI*. 4610–4617.

[21] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. 2019. Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation. In *Proceedings of the IEEE ICCV*. 7919–7929.

[22] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. 2020. Class-Aware Modality Mix and Center-Guided Metric Learning for Visible-Thermal Person Re-Identification. In *Proceedings of the ACM MM*. 889–897.

[23] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. 2020. Unity Style Transfer for Person Re-Identification. In *Proceedings of the IEEE CVPR*. 6887–6896.

[24] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer. In *Proceedings of the IEEE CVPR*. 13379–13389.

[25] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *Proceedings of the IEEE CVPR Workshops*. 0–0.

[26] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-Guided Feature Alignment for Occluded Person Re-Identification. In *Proceedings of the IEEE ICCV*. 542–551.

[27] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person Recognition System Based on a Combination of Body Images From Visible Light and Thermal Cameras. In *Sensors*, Vol. 17. 605.

[28] Peixi Peng, Yonghong Tian, Yangru Huang, Xiangqian Wang, and Huilong An. 2020. Discriminative Spatial Feature Learning for Person Re-Identification. In *Proceedings of the ACM MM*. 274–283.

[29] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. 2020. Dual Gaussian-based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification. In *Proceedings of the ACM MM*. 2149–2158.

[30] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline). In *Proceedings of the ECCV*. 480–496.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE CVPR*. 2818–2826.

[32] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. 2019. AANet: Attribute Attention Network for Person Re-Identifications. In *Proceedings of the IEEE CVPR*. 7134–7143.

[33] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-identification. In *Proceedings of the ACM MM*. 274–282.

[34] Guan’an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. 2019. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In *Proceedings of the ICCV*. 3623–3632.

[35] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. 2020. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In *Proceedings of the AAAI*, Vol. 34. 12144–12151.

[36] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin’ichi Satoh. 2019. Learning to Reduce Dual-level Discrepancy for Infrared-visible Person Re-identification. In *Proceedings of the IEEE CVPR*. 618–626.

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In *Proceedings of the IEEE CVPR*. 79–88.

[38] Xing Wei, Diangang Li, Xiaopeng Hong, Wei Ke, and Yihong Gong. 2020. Co-attentive Lifting for Infrared-visible Person Re-identification. In *Proceedings of the ACM MM*. 1028–1037.

[39] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. 2020. RGB-IR Person Re-identification by Cross-Modality Similarity Preservation. In *IJCV*, Vol. 128. 1765–1785.

[40] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. Rgb-Infrared Cross-Modality Person Re-identification. In *Proceedings of the ICCV*. 5380–5389.

[41] Fan Yang, Zheng Wang, Jing Xiao, and Shin’ichi Satoh. 2020. Mining on Heterogeneous Manifolds for Zero-Shot Cross-Modal Image Retrieval. In *Proceedings of the AAAI*. 12589–12596.

[42] Hanrong Ye, Hong Liu, Fanyang Meng, and Xia Li. 2020. Bi-directional Exponential Angular Triplet Loss for Rgb-infrared Person Re-identification. In *IEEE TIP*, Vol. 30. 1583–1595.

[43] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. 2020. Cross-Modality Person Re-Identification via Modality-aware Collaborative Ensemble Learning. In *IEEE TIP*, Vol. 29. 9387–9399.

[44] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong Yuen. 2018. Hierarchical Discriminative Learning for Visible Thermal Person Re-identification. In *Proceedings of the AAAI*, Vol. 32.

[45] Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. 2020. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification. In *Proceedings of the ECCV*. 229–247.

[46] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. 2018. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In *Proceedings of the IJCAI*. 1092–1099.

[47] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-Aware Global Attention for Person Re-Identification. In *Proceedings of the IEEE CVPR*. 3186–3195.

[48] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. 2019. Pyramidal Person Re-IDentification via Multi-Loss Dynamic Training. In *Proceedings of the IEEE CVPR*. 8514–8522.

[49] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint Discriminative and Generative Learning for Person Re-identification. In *Proceedings of the IEEE CVPR*. 2138–2147.

[50] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random Erasing Data Augmentation. In *Proceedings of the AAAI*, Vol. 34. 13001–13008.

[51] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. In *Proceedings of the ECCV*. 346–363.

[52] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Vijaya Kumar, and Jan Kautz. 2020. Joint Disentangling and Adaptation for Cross-domain Person Re-identification. In *Proceedings of the ECCV*. 87–104.