Adversarial Decoupling and Modality-invariant Representation Learning for Visible-Infrared Person Re-identification

Weipeng Hu, Bohong Liu, Haitang Zeng, Yanke Hou, Haifeng Hu

***Abstract*—Visible-infrared person re-identification (RGB-IR ReID) has now attracted increasing attention due to its surveillance applications under low-light environments. However, the large intra-class variations between different domains are still a challenging issue in the field of computer vision. To address the above issue, we propose a novel adversarial Decoupling and Modality-invariant Representation learning (DMiR) method to explore potential spectrum-invariant yet identity-discriminative representations for cross-modality pedestrians. Our model consists of three key components, including Domain-related Representation Disentanglement (DrRD), Modality-invariant Discriminative Representation (MiDR) and Representation Orthogonal Decorrelation (ROD). First, two subnets named Identity-Net and Domain-Net are designed to extract identity-related features and domain-related features, respectively. Given this two-stream structure, the DrRD is introduced to achieve adversarial decoupling against domain-specific features via a min-max disentanglement process. Specifically, the classification objective function on Domain-Net is minimized to extract spectrum-specific information while maximizing it to reduce domain-specific information. Second, in Identity-Net, we introduce MiDR to enhance intra-class compactness and reduce domain variations by exploring positive and negative pair variations, semantic-wise differences, and pair-wise semantic variations. Finally, the correlation between the two decomposed features, i.e., identity-related features and domain-related features, may lead to the introduction of modal information in identity representations, and vice versa. Therefore, we present the ROD constraint to make the two decomposed features unrelated to each other, which can more effectively separate the two-component features and enhance feature representations. Practically, we construct ROD at the feature-level and parameter-level, and finally select feature-level ROD as the decorrelation strategy because of its superior decorrelation performance. The whole scheme leads to disentangling spectrum-dependent information, as well as purifying identity information. Extensive experiments are carried out on two mainstream RGB-IR ReID datasets, and the results demonstrate the effectiveness of our method.**

***Index Terms*—Visible-infrared person re-identification, modality-invariant representations, orthogonal decorrelation.**

I. Introduction

PERSON re-identification (ReID) aims to match pedestrian images across camera views, which is a vital topic in intelligent video surveillance [1], [2], [3]. Person re-identification helps people analyze surveillance data efficiently and relieves people from the tedious task of watching video. Most existing studies focus on ReID based on visible light images; however, visible light images are not suitable for low-light environments or at night, when criminal incidents often occur. In contrast, infrared images can reflect the night environment, and most surveillance cameras can automatically switch from RGB mode to InfraRed (IR) mode in a low-light environment [4], [5]. Based on the above facts, it is of great significance to realize visible-infrared person re-identification (VIS-IR ReID, or RGB-IR ReID). Visible-infrared person re-identification is a cross-modality image retrieval problem that can match pedestrian images between visible light and infrared images. Visible-infrared person re-identification is more challenging than single-modality person re-identification but more adaptable to the real world. Compared with single-modality person re-identification, in addition to the noise caused by pose, background, illumination, and shielding, visible-infrared person re-identification faces considerable modality differences [6], [7], [8].

Feature learning and domain adaptation are two methods usually used for RGB-IR ReID task. Feature learning directly extract common identity features from images for identification. Many researchers focus on narrowing the gap between the RGB and infrared images via feature alignment. Their basic idea is to match two kinds of images in a shared feature space. [9] proposes an improved two-stream Convolutional Neural Network (CNN) to learn deep multi-modality sharable feature representations. [6] pays attention to the design of loss function and further presents a novel feature learning framework named Hard Pentaplet and Identity Loss Network (HPILN). However, the accuracy of feature learning is limited under current situation of RGB-IR ReID, i.e., less data and large modality differences. As for domain adaptation, researchers often combine it with Generative Adversarial Networks (GANs) [10], which is widely used in RGB-IR ReID and achieves high accuracy. Domain adaptation converts source-domain image to the target domain, and transforms the multi-modal problem into a single-modality problem for processing. In [11], authors utilize generative adversarial learning approach. [12] proposes an end-to-end Alignment Generative Adversarial Network (AlignGAN), in which they generate fake infrared images based on real RGB images by a pixel alignment module and then match fake infrared images and real infrared ones by a feature alignment module. However, domain adaptation method combined with GANs technology converges slowly and requires a huge amount of

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: huwp5@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

computation. Synthetically analysing the above content, it is not hard to discover that existing methods based on feature learning or domain adaptation cannot effectively solve RGB-IR ReID problems.

Faced with this dilemma, recently, some researchers start with a new perspective via a disentanglement mechanism [7], [13]–[17]. By switching between identity and identity-unrelated information [7], [16], pose information [17] or structural information [18], the generation module is able to generate pseudo-pedestrian images of specific attributes, which can enhance identity-related encoder. However, it is difficult to generate photo-realistic images and the poor quality of the synthesized images may negatively effect the disentanglement process. Moreover, these models require additional computational consumption for image-level generation module except for the identity encoder. Some competitors introduce disentanglement techniques at the feature level, thus avoiding the synthetic pseudo-image problem described above. Specifically, to disentangle unrelated factors, many effective techniques have been explored, including conditional-invariant deep modality generalization [19], camera invariance representation disentanglement [20] and spectrum-unrelated feature learning [21]. In [22], global features are disentangled to multiple semantic-attribute features to enhance discriminative representation learning. Spectrum-Disentangled representation Learning (SDL) [13] presents a two-branch network with disentanglement loss, distilling identity features while dispelling spectrum features. However, they lack consideration of feature decomposition and feature decorrelation, and thus the decomposed identity features may be intermixed with modal or identity-unrelated information. Obviously, RGB-IR person ReID remains a problem demanding a prompt solution.

In this paper, we propose a novel approach adversarial Decoupling and Modality-invariant Representation learning (DMiR), for RGB-IR ReID, which has high accuracy and is demonstrated to be effective. Our DMiR network adopts an end-to-end two-stream structure to extract the representations of invariant modality and discriminative identity. It consists of three parts, Domain-related Representation Disentanglement(DrRD), Modality-invariant Discriminative Representation (MiDR) and Representation Orthogonal Decorrelation (ROD). Firstly, the designed two-stream structure contains two subnets, i.e., Identity-Net and Domain-Net, which are utilized to extract identity-related features and domain-related features, respectively. Based on this two-stream structure, the DrRD is introduced to Domain-Net for the purpose of reducing modality gap between visible images and infrared images via a min-max adversarial disentanglement process, where the classification objective function on domain-related features is minimized to extract spectrum-specific information, and maximized to reduce the domain-specific information. Secondly, in Identity-Net, MiDR is introduced to further enhance identity-discriminative learning by exploring positive and negative pair variations, semantic-wise differences, and pair-wise semantic variations. Finally, the ROD constraint is presented to make the two decomposed features unrelated to each other, which can more effectively separate the two-component information and enhance feature representations. In this part, we implement ROD at the feature-level and parameter-level and select Feature-level ROD (FROD), the more effective level, as our final decorrelation strategy. Through the abovementioned processes, domain-related information is effectively eliminated, resulting in the purification of identity discriminative features. Extensive experiments are carried out in two popular cross-modality person re-identification datasets (i.e., SYSU-MM01 [4] and RegDB [23] datasets), which demonstrates the effectiveness of our method. In summary, the main contributions of this paper are listed as follows.

- The proposed DMiR network integrates DrRD, MiDR, and ROD into an end-to-end two-stream architecture to extract pure identifiable features by decoupling modal information. Extensive experiments verify the effectiveness of this orthogonal decoupling method.
- Identity-related features and domain-related features are mapped into different subspaces by Identity-Net and Domain-Net, respectively. In Domain-Net, the DrRD is introduced to eliminate spectrum variations and purify identity-related features via a min-max adversarial disentanglement process, effectively avoiding the negative impact from mixed spectrum information in identity-related features.
- The MiDR is introduced to Identity-Net to enhance intra-class compactness and reduce domain variations. To be specific, we explore positive and negative pair variations, semantic-wise differences, and pair-wise semantic variations in MiDR, which aims to learn potential domain-invariant yet identity-discriminative representations.
- The ROD constraint is presented to make the two decomposed features (identity-related features and domain-related features) unrelated to each other, which can effectively separate the two-component information and enhance feature representations.

II. Related work

A. RGB-RGB Person ReID

The task of person ReID is to filter out images of specific pedestrians from a large number of candidate query [7], [24], [25], [26], [27], [28]. Recent work is mainly based on deep learning methods to obtain more prominent features. Some researchers use part-level feature learning to obtain a more fine-grained representations. They focus on locating regions with specific pre-defined semantics to learn local representations. Wang *et al.* [29] construct an end-to-end feature learning network to combine features with different granularities. Sun *et al.* [30] horizontally cut original feature tensor into 6 parts and then concatenate the processed column vectors to represent final features for the input image. Some are dedicated to designing appropriate loss functions based on metric learning, such as contrast loss, triple loss, and so on. Facing the inevitable unavoidable occluded person images in real scenarios, occluded ReID methods are designed for eliminating the negative impact of occlusion. Wang *et al.* [31] model human-topology information and high-order relation to learn well-aligned robust features. In addition, pose-based ReID attracts lots of interest. Zhu *et al.* [32] simultaneously

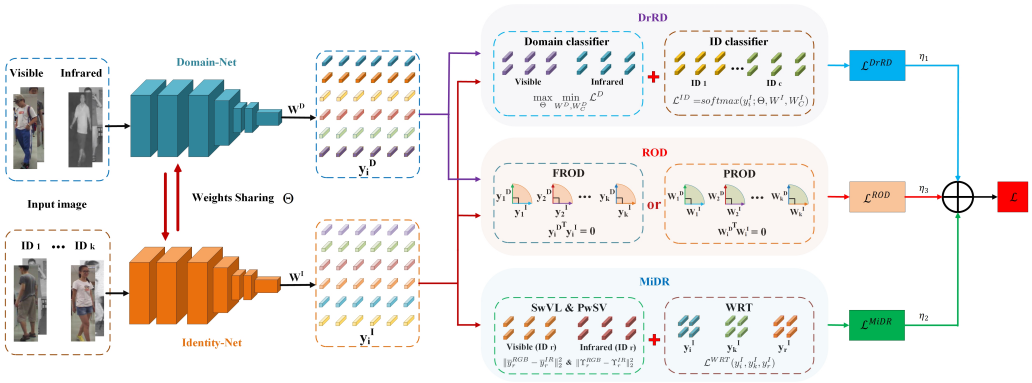

Fig. 1. The flowchart of DMiR model. Firstly, visible and infrared images are input into Identity-Net and Domain-Net to obtain identity-related features and domain-related features, respectively. In Domain-Net, DrRD is introduced to reduce modality variations and purify identity-related features via a min-max adversarial disentanglement process. Secondly, in Identity-Net, MiDR is designed to enhance intra-class compactness and reduce domain variations. Finally, ROD is elaborated to further eliminate the correlation between identity-related features and domain-related features.

utilize pose and appearance features to guide the pose transfer process. The above-mentioned methods reduce noise interference in pedestrian images, such as illumination, view point, and attitude.

B. RGB-IR Person ReID

Conventional person ReID usually refers to RGB-RGB ReID, the most general single-modality ReID problem [13], [30], [33], [34], [35]. Nevertheless, RGB-RGB ReID is easy to suffer from the surveillance when lighting is poor or even unavailable. Visible-infrared person re-identification, more suitable for real scenarios, is based on RGB images or infrared images acquired under different illumination conditions for retrieval or matching. In other words, it attempts to match RGB and infrared pedestrian images across disjoint cameras. Compared to traditional Re-ID task, due to cross-modality variations and intra-modality variations, RGB-IR ReID task is more challenging [6]. Intra-modality variations imply that the quality of image influenced by viewpoint, illumination, occlusion, human body pose, etc., can be notably discrepant even within the same RGB or infrared modality. In addition to intra-domain differences, RGB-IR person re-identification also suffers from huge modal variations between images in different domains, which makes it still a challenging issue in the field of computer vision.

Feature Learning Based Method. To solve the RGB-IR person ReID problem, many researchers focus on designing network structure for invariant feature extraction, which can be classified as feature learning based method. For example, Wu *et al.* [4] first construct a large RGB-IR pedestrian re-recognition dataset named SYSU-MM01 and further propose a deep zero-padding method for evolving domain-specific structure automatically in one-stream network for RGB-IR ReID task. Ye *et al.* [9] propose a two-stream CNN network by sharing the backbone network. Zhao *et al.* [6] make innovative design of loss function and put forward a novel structure named HPILN, which helps the network perform better feature learning. On the basis of the two-stream CNN network, Zhu *et al.* [36] introduce hetero-center loss to narrow distribution gap between different modes. Ye *et al.* [38] solve the cross-domain person ReID by introducing weighted tri-directional ranking loss and homogeneous-heterogeneous identification loss. Fu *et al.* [39] present cross-domain deep network architecture search to explore BN layers and search the optimal architecture. In short, existing representation learning method mainly focuses on obtaining better feature expression while learning information shared between modalities.

Domain Adaptation Based Method. Some other mainstream cross-modality ReID methods are dedicated to reducing spectrum variations through domain adaptation or GANs to achieve style transfer. To reduce the domain gap, generative adversarial learning [11] and alignment generative adversarial networks [12] are presented for cross-modality ReID. AlignGAN [12] firstly generates fake infrared images based on visible images via a pixel alignment module and then matches fake infrared images and real ones via a feature alignment module. Wang *et al.* [37] introduce a novel Dual-level Discrepancy Reduction Learning (D2RL), in which an image-level subnet is trained to translate a visible image into its infrared counterpart and an infrared image to its visible version. In [40], a model named Joint Set-level and Instance-level Alignment ReID (JSIA-ReID) is proposed to generate cross-modality pair images. Nevertheless, domain adaptation based method combined with GANs usually converges slowly and consumes a huge amount of computation.

Disentangled Representation Learning. Abovementioned two methods may confuse specific information of each modal, and identity-unrelated information will also be learned by the network. How to disentangle identity-irrelated information (such as spectrum, background information, etc.) and enhance learning of identity-related information has become the focus of recent research, which yielding a new research area called disentangled representation learning [7], [13], [18], [22], [47], [49]. Recently, many works have been focused on learning disentangled representations in various fields [41], [42], [43], [44], [45], [46], [49]. Also, in ReID task, several researchers have made efforts to disentangle the pedestrian image into identity related and irrelevant information (e.g., viewpoint, background, modal information and other attributes) [13], [47], [48], [49], [50], [51]. Choi *et al.* [7] propose a Hier-

archical Cross-Modality Disentanglement (Hi-CMD) method, which disentangles identity-discriminative factors and identity-excluded factors from visible-infrared images. By switching between identity information and identity-independent information [16], pose information [17] or structural information [18], the model can synthesize pseudo-pedestrian images of specific attributes and enhance identity representation learning. It is worth noting that synthesizing images requires additional computational consumption and the poor quality of the synthesized images may negatively affect the disentanglement process. Some methods do not introduce the image synthesis process and design decoupling modules at the feature level. For instance, [13] presents a two-branch network with disentanglement loss to distill identity features while dispelling spectrum features, and [19] integrates four sub-networks to decouple unrelated representations. Delorme *et al.* [20] design a camera classifier and identity classifier to enforce camera invariance and learn identity representations. Zhao *et al.* [22] disentangle features into groups of sub-features, each of which corresponds to a specific semantic property.

Our proposed approach is based on the framework of two-stream CNN network, which integrates DrRD, MiDR and ROD in an end-to-end architecture. Specifically, DrRD, MiDR and ROD are designed for modality information adversarial decoupling, representation enhancement and features decorrelation, respectively. The whole scheme leads to the disentanglement of spectrum-specific information, as well as the purification of identity information. There exists some apparent distinctions between our method and the ones based on feature learning [4], [9], [38], [39] and domain adaptation [11], [37], [40]. On the one hand, our method focuses on feature adversarial decoupling and feature enhancement rather than simply feature learning. On the other hand, domain adaptation based method regards RGB-IR ReID task as a single-modality problem after cross-modality image conversion, which may result in the loss of pedestrian identity information during the cross-spectrum synthesis. Unlike domain adaptation method, our method does not include a synthesis process, and we utilize feature-level orthogonal adversarial disentanglement to eliminate domain variations.

The aforementioned disentanglement-based techniques [7], [16]–[18], [46] disentangle identity-related factors and unrelated factors to reconstruct pseudo-pedestrian images. However, it is difficult to generate photographs-realistic images, and the poor quality of the composite images may have a negative impact on the disentanglement process. In addition, the design of the these networks is very complex, containing multiple encoders and decoders, which may encounter slow convergence, massive computation and over-fitting problems. On the contrary, our DMiR model does not incorporate the image generation process, and it only contains two shared parameter encoders, i.e., Identity-Net and Domain-Net, which has stable training and easy convergence behavior. Although some methods [13], [20]–[22] use the disentanglement mechanism to enhance representation learning, they lack representation constraints on the disentanglement of identity-related and identity-unrelated information, while DMiR network considers representation decorrelation to avoid introduction of spectrum information to identity representations. Different from methods at [15], [44], DMiR is designed to solve RGB-IR ReID task by exploring different optimization objectives and de-correlation techniques. In particular, the DMiR extracts pure identifiable features and decouples modal information by introducing DrRD, MiDR and ROD into the network.

III. The Proposed Adversarial Decoupling and Modality-invariant Representation Model

In this section, we provide details of the proposed adversarial decoupling and modality-invariant representation learning network. Due to the dramatic modality variations between visible-spectrum pedestrian images and infrared-spectrum pedestrian images, pedestrian representations extracted by deep networks may contain both identity-related information and spectrum-related information, which reduces the performance of RGB-IR ReID. To reduce the spectrum discrepancies, a feasible idea is to separate the modal information and identity information, and then polish the identity information while suppressing and decoupling the modality information. As illustrated in Fig. 1, the DMiR model mainly consists of three key components, comprising domain-related representation disentanglement, modality-invariant discriminative representation and representation orthogonal decorrelation. First, the DrRD is presented to achieve adversarial decoupling against domain-specific information through min-max adversarial disentanglement. Specifically, two subnets, i.e., Identity-Net and Domain-Net, are designed to learn identity-related features and domain-related features, respectively, and the adversarial process is introduced to Domain-Net to disentangle modality-related information. Second, MiDR is introduced to Identity-Net to enhance intra-class compactness and reduce domain variations by exploring positive and negative pair variations, semantic-wise differences and pairwise semantic variations. Finally, considering that the identity-dependent features and modality-specific features are unrelated to each other, we introduce ROD to ensure the orthogonal irrelevant relationships between the two decomposed representations, which can purify identity-related features and improve cross-domain ReID performance. The details of each component of DMiR are given below.

A. Problem Modeling

A cross-modality person ReID dataset consists of infrared-spectrum person images $X^{IR} = \{(x_1^{IR}, l_{I1}^{IR}, l_{D1}^{IR}), (x_2^{IR}, l_{I2}^{IR}, l_{D2}^{IR}), ...\}$ and visible-spectrum ones $X^{RGB} = \{(x_1^{RGB}, l_{I1}^{RGB}, l_{D1}^{RGB}), (x_2^{RGB}, l_{I2}^{RGB}, l_{D2}^{RGB}), ...\}$, where x_i^{IR} and x_i^{RGB} indicate image data, $l_{Ii} \in \{l_{Ii}^{IR}, l_{Ii}^{RGB}\}$, $l_{Ii}^{IR} \in \{1, 2, .., c\}$ and $l_{Ii}^{RGB} \in \{1, 2, .., c\}$ represent identity tags, and $l_{Di}^{RGB} \in \{0\}$ and $l_{Di}^{IR} \in \{1\}$ denote modality tags. In the training stage, we optimal networks to learn identity-discriminative feature embedding using training dataset $\{X^{IR}, X^{RGB}\}$. In the testing stage, we use the trained model to encode features for both RGB and IR images, and a ranking list within the gallery set is calculated.

B. Domain-related Representation Disentanglement

Given the input images, we first construct an Identity-Net to extract identity information. However, identity-related feature may inevitably contain spectrum information, which brings about performance deterioration [13]. To decouple modal information from the network, we introduce a Domain-Net to implement the domain decoupling process, thereby purifying the identity information.

As shown in Fig. 1, our two-stream feature extractor includes Identity-Net and Domain-Net. In particular, Identity-Net is developed to learn identity-dependent representations for both RGB image and IR image. Specifically, each input image $x_i \in \{X^{IR}, X^{RGB}\}$ passes through the convolutional structures and fully connected layer to generate ID features $y_i^I \in R^{d\times 1}$, taking the following form:

$$y_i^I = FC^{ID}(Conv^{ID}(x_i;\Theta);W^I) \tag{1}$$

where $Conv^{ID}(\cdot)$ and $FC^{ID}(\cdot)$ denote convolutional operation and fully connected operation, and Θ and W^I are parameters of $Conv^{ID}$ and FC^{ID} structures. To extract identity-related information, the common softmax is used as classification objective, which can be expressed as:

$$\begin{aligned}\mathcal{L}^{ID} =& -\frac{1}{n^{IR}+n^{RGB}}\Sigma_{i=1}^{n^{IR}+n^{RGB}}\mathcal{L}_i^{ID}(y_i^I;\Theta,W^I,W_C^I)\\ =& -\frac{1}{n^{IR}+n^{RGB}}(\Sigma_{j=1}^{n^{IR}}\Sigma_{i=1}^{c}\mathbf{1}\{l_{Ij}^{IR}=i\}log\widehat{p}_{ij}\\ &+\Sigma_{j=1}^{n^{RGB}}\Sigma_{i=1}^{c}\mathbf{1}\{l_{Ij}^{RGB}=i\}log\widehat{p}_{ij})\end{aligned} \tag{2}$$

where $\widehat{p}_{ij}$ indicates the predicted probability for the identity classifiers in Identity-Net, W_C^I is the parameters in the classifier layer, and n^{IR} and n^{RGB} are the number of samples in X^{IR} and X^{RGB}. $\mathbf{1}\{\cdot\}$ represents the indicator function, with value of 1 or 0.

In Identity-Net, both RGB and IR images are adopted to train the network with the classification objective, which can enhance classification ability of the network. However, due to dramatic variations between different modalities, it is hard to learn identity-discriminative representations, resulting in the introduction of domain-related information and unsatisfied performance of cross-modal pedestrian re-identification. To reduce modality-dependent components, an intuitive idea is to separate modality information from the network while retaining identity-related information. To achieve the above aim, Domain-Net is designed to reduce the modality gap between infrared image and visible image via an adversarial decoupling process. Specifically, we design the Domain-Net to distill domain-specific features $y_i^D \in R^{d\times 1}$ as follow:

$$y_i^D = FC^D(Conv^D(x_i;\Theta);W^D) \tag{3}$$

where $Conv^D(\cdot)$ and $FC^D(\cdot)$ represent the operations of convolutional structures and fully connected layer, and W^D denotes parameters in FC^D. It is worth noting that $Conv^{ID}$ and $Conv^D$ share the same parameters Θ. The softmax is used to train Domain-Net with modality tags $\{l_{Di}^{IR}, l_{Di}^{RGB}\}$, taking the following form:

$$\begin{aligned}\mathcal{L}^D =& -\frac{1}{n^{IR}+n^{RGB}}\Sigma_{i=1}^{n^{IR}+n^{RGB}}\mathcal{L}_i^D(y_i^D;\Theta,W^D,W_C^D)\\ =& -\frac{1}{n^{IR}+n^{RGB}}(\Sigma_{j=1}^{n^{IR}}\Sigma_{i\in\{0,1\}}\mathbf{1}\{l_{Dj}^{IR}=i\}log\widehat{p}_{ij}\\ &+\Sigma_{j=1}^{n^{RGB}}\Sigma_{i\in\{0,1\}}\mathbf{1}\{l_{Dj}^{RGB}=i\}log\widehat{p}_{ij})\end{aligned} \tag{4}$$

where W_C^D represents parameters in the classifier layer, and $l_{Dj}^{RGB} \in \{0\}$ and $l_{Dj}^{IR} \in \{1\}$ denote modality tags for x_j^{RGB} and x_j^{IR} images. In this way, two sub-networks, including Identity-Net and Domain-Net, aim to learn identity-related features and domain-related features, respectively. Obviously, Identity-Net has a positive effect on learning identity-related features. On the contrary, Domain-Net may introduce modal-related information, which may reduce the performance of cross-spectrum pedestrian re-identification. In order to switch the goal of Domain-Net from learning modal information to eliminating modal information, we introduce a min-max adversarial decoupling process [10] to Domain-Net, which can be expressed as:

$$\Theta^*, {W^D}^*, {W_C^D}^* = \max_{\Theta}\min_{W^D,W_C^D}\mathcal{L}^D \tag{5}$$

On the one hand, the domain-related discriminator (i.e., $\{W^D, W_C^D\}$) is updated by minimizing $\mathcal{L}^D$, and thus the spectrum-specific information can be extracted in the modality-related layer. On the other hand, domain-related generator (i.e., $\{\Theta\}$) is trained by maximizing the loss $\mathcal{L}^D$, which aims to disentangle domain-specific information from the network.

The joint loss in DrRD can be expressed as:

$$\mathcal{L}^{DrRD} = \mathcal{L}^{ID} + \kappa\mathcal{L}^D \tag{6}$$

where κ is a hyper-parameter that balances these two objectives.

To summarize, we propose a min-max adversarial training process in DrRD to decouple modal information and purify identity information, which is committed to solving the problem of performance degradation caused by unpurified identity information mixed with modal information.

C. Modality-invariant Discriminative Representation

Features y_i^I extracted in Identity-Net are used as the pedestrian representations for both RGB image and IR image. Eq. (2) shows that the identity classification objective L^{ID} is introduced to Identity-Net to learn identity-related features. However, due to the huge discrepancies between the heterogeneous pedestrian images, including spectrum, occlusion, view point, illumination, person's pose, etc, it is difficult for Identity-Net to learn discriminative representations using only classification loss. Thus, we aim to enhance discriminative representations of Identity-Net and improve the performance for RGB-IR ReID by exploring potential spectrum-invariant yet identity-discriminative features learning.

Similar to [52], a modified Triplet, called Weighted Regularization Triplet (WRT), is introduced to Identity-Net to improve the identity representation ability, taking the following form:

$$\begin{aligned}\mathcal{L}^{WRT} =& \frac{1}{n^{WRT}}\Sigma_{i=1}^{n^{WRT}} log(exp(\delta_i^P \mathbf{Eud}(y_i^I, y_k^I) \\ & - \delta_i^N \mathbf{Eud}(y_i^I, y_r^I)) + 1)\end{aligned} \tag{7}$$

where

$$\delta_i^P = \frac{exp(\mathbf{Eud}(y_i^I, y_k^I))}{\sum_{i,k} \mathbf{1}\{l_{Ii} = l_{Ik}\}\mathbf{Eud}(y_i^I, y_k^I)}$$

$$\delta_i^N = \frac{exp(\mathbf{Eud}(y_i^I, y_r^I))}{\sum_{i,r} \mathbf{1}\{l_{Ii} \neq l_{Ir}\}\mathbf{Eud}(y_i^I, y_r^I)}$$

where y_i^I, y_k^I and y_r^I represent anchor, positive and negative features, $\mathbf{Eud}(\cdot,\cdot)$ denotes the Eucilidean distance, δ_i^P and δ_i^N indicate the weighting coefficient for positive pair and negative pair, and n^{WRT} denotes the number of selected triplets. The WRT aims to minimize positive pair variations and maximize negative pair variations, which can learn discriminative representations for RGB-IR ReID.

The Identity-Net is designed to map heterogeneous pedestrian images to a high-level semantic representations. Intuitively, the semantic variations between RGB image and IR image of the same pedestrian should be as small as possible. Thus, we introduce a Semantic-wise Variations Learning (SwVL) to reduce the domain differences, taking the following form:

$$\mathcal{L}^{SwVL} = \frac{1}{c}\sum_{r=1}^{c} \|\overline{y}_r^{RGB} - \overline{y}_r^{IR}\|_2^2 \tag{8}$$

where

$$\overline{y}_r^{RGB} = \frac{1}{n_r^{RGB}} \sum_{i=1}^{n_r^{RGB}} \mathbf{1}\{l_{Ii}^{RGB} = l_{Ir}^{RGB}\} y_i^I$$

$$\overline{y}_r^{IR} = \frac{1}{n_r^{IR}} \sum_{i=1}^{n_r^{IR}} \mathbf{1}\{l_{Ii}^{IR} = l_{Ir}^{IR}\} y_i^I$$

where $\overline{y}_r^{RGB}$ and $\overline{y}_r^{IR}$ denote RGB mean features and IR mean features of the r^{th} class, and n_r^{RGB} and n_r^{IR} are the number of RGB images and IR images in the r^{th} class. The SwVL aims to minimize semantic-wise discrepancies of the same pedestrian in two different modalities. In addition to semantic-wise variations, we further introduce Pair-wise Semantic Variations learning (PwSV) to the network, which explores the inter-semantic relationship of cross-domain pedestrian images. The objective in PwSV can be expressed as:

$$\mathcal{L}^{PwSV} = \frac{1}{c}\sum_{r=1}^{c} \|\Upsilon_r^{RGB} - \Upsilon_r^{IR}\|_2^2 \tag{9}$$

where

$$\begin{aligned}\Upsilon_r^{RGB} =& \mathbb{E}\left[\mathbf{1}\{l_{Ii}^{RGB} = l_{Ir}^{RGB}\}(y_i^{RGB} - \overline{y}_r^{RGB})(y_i^{RGB} - \overline{y}_r^{RGB})^T\right] \\ \Upsilon_r^{IR} =& \mathbb{E}\left[\mathbf{1}\{l_{Ii}^{IR} = l_{Ir}^{IR}\}(y_i^{IR} - \overline{y}_r^{IR})(y_i^{IR} - \overline{y}_r^{IR})^T\right]\end{aligned} \tag{10}$$

where y_i^{RGB} and y_i^{IR} denote extracted features of RGB image x_i^{RGB} and IR image x_i^{IR} in Identity-Net, and Υ_r^{RGB} and Υ_r^{IR} represent the covariance matrices of RGB image and IR image in the r^{th} class. The covariance (i.e., Υ_r^{RGB} and Υ_r^{IR}) is used to measure the correlation of pair-semantic variables, and minimizing the difference between Υ_r^{RGB} and Υ_r^{IR} can ensure the consistency of pair-semantic relationship between different modalities [46], [53], thereby effectively reducing spectrum variations.

The joint loss in MiDR can be expressed as:

$$\mathcal{L}^{MiDR} = \mathcal{L}^{WRT} + \varphi_1\mathcal{L}^{SwVL} + \varphi_2\mathcal{L}^{PwSV} \tag{11}$$

where φ_1 and φ_2 are hyper parameters that balance the three objective items. The proposed MiRD enhances potential domain-invariant yet identity-discriminative representation learning, by exploring positive pair and negative variations, semantic-wise differences as well as pair-wise semantic variations.

D. Representation Orthogonal Decorrelation

As mentioned in Section 3.2, y_i^I and y_i^D are obtained by linearly mapping from two different fully connected layer FC^{ID} and FC^D, and thus they may linearly correlate to each other. However, the correlation between modal information and identity information may lead to the introduction of modal information in identity representations [44], [45], and vice versa. Therefore, we present the representation orthogonal decorrelation that is effective to reduce the correlation between the two learning features. In particular, we come up with two different decorrelation strategies, including Feature-level Representation Orthogonal Decorrelation (FROD) as well as Parameter-level Representation Orthogonal Decorrelation (PROD), and conduct simple analysis for these two decorrelation ideas.

Feature-level ROD To eliminate representational relevance, an intuitive idea is to impose orthogonal constraint to two decomposed features y_i^I and y_i^D, taking the following form:

$$y_i^{I^T} y_i^D = 0 \tag{12}$$

Considering that the vector modulus value (i.e., $\|y_i^I\|$ and $\|y_i^D\|$) has no effect on the orthogonality, we limit the representation vector to the unit vector, and Eq. (12) can be rewritten as:

$$\frac{y_i^{I^T} y_i^D}{\|y_i^I\|\|y_i^D\|} = 0 \tag{13}$$

Eq. (13) imposes an orthogonal constraint between unit-vector $\frac{y_i^I}{\|y_i^I\|}$ and unit-vector $\frac{y_i^D}{\|y_i^D\|}$, to make them irrelevant to each other, which can enhance modality-dependent representations and identity-dependent representations, simultaneously. Based on Lagrange theory [54], the equality constraints in Eq. (13) can be transformed into the following form:

$$\mathcal{L}^{FROD}(y_i^I, y_i^M; \Theta, W^I, W^D) = \sum_{i=1}^{n} \tau_i \left\| \frac{y_i^I \left(y_i^M\right)^T}{\|y_i^I\| \, \|y_i^M\|} \right\|_F^2 \tag{14}$$

where τ_i is the Lagrange multiplier, and n is the number of samples in a mini-batch. In Feature-level ROD, we aims to minimize the objective $\mathcal{L}^{FROD}$ and update the whole network parameters $\{\Theta, W^I, W^D\}$, and thus domain-related features and identity-related features can be effectively purified by removing their correlation.

Parameter-level ROD In addition to directly making the two decomposed features irrelevant (i.e., FROD) to each other, it can also be done in an indirect way, that is, to force the parameters (i.e., W^I and W^D) to be irrelevant to achieve the purpose of de-correlation, which can be expressed as:

$$W_i^{I\,T} W_i^D = 0 \tag{15}$$

where W_i^I and W_i^D denote the i^{th} mapping vector of W^I and W^D. Similar to FROD, we constrain the mapping vector to a unit-vector, to reduce the influence of vector modulus, and Eq. (15) can be rewritten as:

$$\frac{W_i^{I\,T} W_i^D}{\|W_i^I\| \|W_i^D\|} = 0 \tag{16}$$

The imposed parameter-level orthogonal constraint can make pair-wise mapping vectors (i.e., W_i^I and W_i^D) unrelated to each other, and thus the two mapping matrices project domain-specific features and identity-dependent features into two irrelevant subspaces, which indirectly reduces the correlation of these two decomposed representations. The Lagrange theory is applied to optimize the objective, taking the following form:

$$\mathcal{L}^{PROD}(W^I, W^D) = \sum_{i=1}^{n} \mu_i \left\| \frac{W_i^I \left(W_i^M\right)^T}{\|W_i^I\| \, \|W_i^M\|} \right\|_F^2 \tag{17}$$

where μ_i is the Lagrange multiplier. In PROD, the objective $\mathcal{L}^{PROD}$ is minimized to eliminate the correlation of the two mapping matrices W^I and W^D, thereby enhancing the representation ability of domain-related features and identity-related features.

Analysis of FROD and PROD We propose two alternative representation orthogonal decorrelation strategies, i.e., FROD and PROD, with the common purpose of reducing the correlation between domain-private features and identity-dependent features. The main differences between these two strategies are as follows. Firstly, FROD directly imposes anti-correlation constraint on the two decomposed representations, while PROD indirectly suppresses the correlation of the two decomposed features by constraining the orthogonality of the two mapping matrices. Secondly, the objective $\mathcal{L}^{FROD}$ in FROD is minimized by updating the whole network $\{\Theta, W^I, W^D\}$. In contrast, PROD is minimized by only optimizing the two mapping matrices $\{W^I, W^D\}$.

Algorithm 1: DMiR model

Input: RGB training set X^{RGB} and IR training set X^{IR}, iterations number $iter$.
Output: Identity-Net parameters $\{\Theta, W^I\}$.
1: **for** $k = 1, 2, ..., iter$ **do**
2: Update $\{\Theta, W^I, W_C^I\}$ by minimizing $\mathcal{L}^{ID} + \mathcal{L}^{MiDR} + \mathcal{L}^{ROD}$;
3: Update $\{W^D, W_C^D\}$ by minimizing $\mathcal{L}^D + \mathcal{L}^{ROD}$;
4: Update $\{\Theta\}$ by maximizing $\mathcal{L}^D$;
5: **end for**
6: **Return** $\{\Theta, W^I\}$

We conduct performance comparison experiment of FROD and PROD in RegDB and SYSU-MM01 datasets, and corresponding results and analysis are given in Subsection 4.3. The FROD is finally selected as the decorrelation strategy because of its superior decorrelation performance. For simplicity, the ROD below refers to the FROD, and $\mathcal{L}^{ROD}$ is equivalent to $\mathcal{L}^{FROD}$, i.e., $\mathcal{L}^{ROD} = \mathcal{L}^{FROD}$.

E. Training Process

The proposed DMiR approach consists of three parts: DrRD, MiDR and ROD. In particular, DrRD reduces modality variations and purifies identity-related features via a min-max adversarial process. MiDR enhances intra-class compactness and reduces domain variations. Moreover, ROD constraint is presented to make the two decomposed features unrelated to each other, which can more effectively separate the two-component features and enhance feature representations. The final objective loss of our method can be expressed as:

$$\mathcal{L} = \eta_1 \mathcal{L}^{DrRD} + \eta_2 \mathcal{L}^{MiDR} + \eta_3 \mathcal{L}^{ROD} \tag{18}$$

where η_1, η_2 and η_3 represent the trade-off weights for each part. Parameters in DMiR include Θ, W^D, W^I, W_C^D and W_C^I. Algorithm 1 illustrates the optimization details of the DMiR model. We adopt the total loss $\mathcal{L}$ to optimize the whole network in an end-to-end manner. Specifically, we alternatively train parameters in the identity-related extractor (i.e., $\{\Theta, W^I, W_C^I\}$), domain-related discriminator (i.e., $\{W^D, W_C^D\}$) and domain-related generator (i.e., $\{\Theta\}$) by minimizing $\mathcal{L}^{ID} + \mathcal{L}^{MiDR} + \mathcal{L}^{ROD}$, minimizing $\mathcal{L}^D + \mathcal{L}^{ROD}$ and maximizing $\mathcal{L}^D$, respectively. Notably, we integrate DrRD, MiDR and ROD into an end-to-end architecture and train the whole network in an end-to-end iterative manner.

IV. Experiments Evaluation

In this section, extensive experiments are conducted on two popular RGB-IR ReID datasets, including SYSU-MM01 dataset and RegDB dataset, to illustrate the effectiveness of the DMiR model. First, we analyze the performance of different decorrelation strategies, i.e., FROD and PROD. Second, we investigate the effects of different components, including DrRD, ROD and MiDR. Finally, we compare of our method with other state-of-the-art methods, such as eBDTR [55], D2RL [37], AlignGAN [12], DFE [56], SDL [13], Hi-CMD [7],

(a) RegDB dataset

(b) SYSU-MM01 dataset

Fig. 2. Example images of different identities from (a) RegDB dataset and (b) SYSU-MM01 dataset. In each sub-graph, infrared images are shown in the top row and RGB ones are given in the bottom row.

TABLE I
PERFORMANCE COMPARISON OF FROD AND PROD ON REGDB AND SYSU-MM01 DATASETS.

PROD	FROD	RedDB		SYSU-MM01	
		Rank-1	mAP	Rank-1	mAP
✗	✗	74.59	69.06	48.27	47.69
✔	✗	74.89	69.18	49.08	48.34
✗	✔	**75.79**	**69.97**	**50.54**	**49.29**
✔	✔	75.16	69.80	49.27	48.71

JSIA-ReID [40], to illustrate the effectiveness of the DMiR method.

A. Datasets and Protocols

RegDB Dataset This database contains 4,120 visible light images and 4,120 infrared images, and it is provided by Donggu University in South Korea [23]. Visible images and infrared images are captured by visible light webcam camera and Tau2 camera, respectively. There are 412 persons in the dataset, and 254 of them are females and 158 of them are males. Every person contains 10 visible light images and 10 infrared images. Because the 412 persons are moving while capturing, there exist many differences, e.g, illumination, capturing distance and pose variations. We follow the evaluation protocol described in [52]. The RegDB dataset is randomly divided into two halves for training and testing. There are two evaluation modes, including visible to thermal and thermal to visible. We repeat evaluation procedure for 10 times and calculate the average performance. Example images from the RegDB are given in Figure 2 (a).

SYSU-MM01 Dataset This dataset is a cross-modality person re-identification dataset provided by Zheng's group at Sun Yat-sen University [4]. It is a large-scale dataset including 287,628 visible light images and 15792 infrared images from 6 cameras. In these 6 cameras, two of them are infrared cameras while others are visible light cameras. These cameras capture 491 identities from indoor or outdoor environments with dark and bright conditions. SYSU-MM01 is a challenge dataset since there are great differences between indoor environments and outdoor environments. Similarly, all images are resized to 144×288 before experiments. 395 person identities in this dataset are for training, which conclude 22,258 visible light images and 11,909 infrared images. The other 96 identities are for testing. It is important to notice that SYSU-MM01 contains single-shot and all-search mode, in which only one image of every identity in each camera is selected to form the gallery set. Hence, when testing, 301 visible light images are selected randomly as gallery set while 3803 infrared images are selected for query. We conduct this test for 10 times and calculate the average performance. Example images from the SYSU-MM01 are given in Figure 2 (b).

TABLE II
SIMILARITY BETWEEN ID FEATURES y_i^I AND DOMAIN-SPECIFIC FEATURES y_i^D ON REGDB DATASET.

PROD	FROD	Similarity
✗	✗	3.5e-2
✔	✗	5.4e-3
✗	✔	**8.7e-4**
✔	✔	9.0e-4

B. Implementation Details

We evaluate the DMiR network in Pytorch platform with a GPU of TITAN Xp. The ResNet-50 [52], [57] is used as the backbone network Θ, and the dimensions d of y_i^I and y_i^D are set to 2048. In the training process, We set κ, φ_1, φ_2, η_1, η_2 and η_3 to 1.0, 0.5, 0.0001, 1.0, 1.0, 0.05 in RegDB dataset and 0.5, 0.1, 0.001, 1.0, 1.0, 0.1 in SYSU-MM01 datasets, respectively. According to the experimental results, the above parameters setting can obtain better performance for RGB-IR ReID.

C. Performance Analyses of FROD and PROD

We analyze the performance of the DMiR model with different representation orthogonal decorrelation strategies including FROD and PROD. As shown in Table I, it is clear to find that: **(a)** FROD obtains better performance than PROD, with improvement up to 0.90% (75.79−74.89) in Rank-1 and 0.79% (69.97−69.18) in mAP on RegDB, and up to 1.46% (50.54−49.08) in Rank-1 and 0.95% (49.29−48.34) in mAP on SYSU-MM01. **(b)** Either PROD or FROD can bring performance improvement to the model. Take PROD as an example, it makes the Rank-1 raise 0.3% (74.89−74.59) and 0.81% (49.08−48.27) on RegDB and SYSU-MM01, respectively. **(c)** Utilizing PROD and FROD simultaneously can not bring further improvement. For instance, it trails 1.27% (50.54−49.27) and 0.58% (49.29−48.71) in Rank-1 and mAP on SYSU-MM01. The results indicate that only adopting FROD can more effectively make the two decomposed features irrelevant

TABLE III
ABLATION STUDY ON REGDB DATASET.

Method	Rank-1(%)	Rank-5(%)	Rank-10(%)	mAP
BL	66.94	78.47	84.84	62.24
BL+DrRD	68.26	79.29	85.40	63.56
BL+DrRD+MiDR	74.59	84.17	89.02	69.06
BL+DrRD+MiDR+ROD	**75.79**	**85.26**	**89.86**	**69.97**

TABLE IV
ABLATION STUDY ON SYSU-MM01 DATASET.

Method	Rank-1(%)	Rank-5(%)	Rank-10(%)	mAP
BL	42.65	71.65	83.18	43.13
BL+DrRD	44.86	75.22	85.88	45.31
BL+DrRD+MiDR	48.27	76.88	86.76	47.69
BL+DrRD+MiDR+ROD	**50.54**	**78.82**	**88.12**	**49.29**

TABLE V
EFFECT OF DIFFERENT LOSSES ON SYSU-MM01 DATASET.

$\mathcal{L}^{ID}$	$\mathcal{L}^{WRT}$	$\mathcal{L}^{D}$	$\mathcal{L}^{SwVL}$	$\mathcal{L}^{PwSV}$	$\mathcal{L}^{ROD}$	Rank-1	Rank-5	Rank-10	mAP
✗	✔	✗	✗	✗	✗	39.08	70.05	82.97	41.31
✔	✔	✗	✗	✗	✗	42.65	71.65	83.18	43.13
✔	✗	✔	✗	✗	✗	44.37	73.41	83.73	44.10
✔	✔	✔	✗	✗	✗	44.86	75.22	85.88	45.31
✔	✗	✗	✗	✔	✗	44.06	71.51	81.62	43.13
✔	✗	✗	✔	✗	✗	46.35	74.27	84.09	45.29
✔	✗	✗	✔	✔	✗	46.99	74.59	84.56	45.40
✔	✔	✔	✔	✔	✗	48.27	76.88	86.76	47.69
✔	✔	✔	✔	✔	✔	**50.54**	**78.82**	**88.12**	**49.29**

to each other and enhance their representation ability. Importantly, FROD optimizes the entire network $\{\Theta, W^I, W^D\}$ towards the goal of representation decorrelation, while PROD optimizes the partial network $\{W^I, W^D\}$. Obviously, FROD can schedule more network parameters to achieve this goal, thereby obtaining a better decorrelation effect than PROD. Moreover, the FROD performs better than FROD+PROD, which is mainly due to the fact that redundant constraints in FROD+PROD will increase the burden of network optimization.

To illustrate the effectiveness of our representation orthogonal decorrelation, we further analyze the feature similarity between ID features y_i^I and domain-specific features y_i^D on RegDB dataset. We calculate the absolute value cosine distance between y_i^I features and y_i^D features for each sample, and take the average of all samples as the similarity, as shown in Table II. Obviously, the introduction of PROD or FROD can effectively reduce the similarity, indicating that representation orthogonal decorrelation is effective to reduce the correlation between the two learning features. The smaller similarity scores obtained by FROD (8.7e-4) compared to PROD (5.4e-3) indicates that FROD has better representation decorrelation performance. Compared to PROD (5.4e-3), PROD+FROD (9.0e-4) has smaller similarity scores, which is mainly attributed to the introduction of FROD.

From the analysis of the results in Table I and Table II, the FROD is more effective for representation decorrelation. In the following, we adopt FROD as representation orthogonal decorrelation to reduce the correlation between domain-private features and identity-related features. For simplicity, ROD below refers to FROD.

D. Various Properties Analysis of the DMiR on NIR-VIS Databases

In order to verify the effectiveness of each component of the DMiR algorithm, including DrRD, MiDR and ROD, we carry out extensive experiments on two datasets.

Ablation study We present the ablation study results on two datasets, as shown in Tables III-IV, to justify the validity of DrRD, MiDR and ROD method. In particular, IdentityNet trained with $\mathcal{L}^{ID} + \mathcal{L}^{WRT}$ is represented as BL method. As shown in Table III, in RegDB dataset, DrRD, MiDR and ROD all have improved accuracy on all evaluation metrics (Rank-1, Rank-5, Rank-10 and mAP) for RGB-IR ReID tasks. Specifically, **(1)** BL+DrRD has outperformed BL by 1.32% both in Rank-1 (68.26−66.94) and mAP (63.56−62.24), which proves the effectiveness of DrRD. **(2)** Based on BL+DrRD, BL+DrRD+MiDR has further increased by 6.33% (74.59−68.26) in Rank-1 and 5.50% (69.06−63.56) in mAP, which demonstrates that introducing MiDR helps to improve performance. **(3)** Compared with BL+DrRD+MiDR, BL+DrRD+MiDR+ROD has improved the performance by 1.20 % (75.79−74.59) in Rank-1 and 0.91% (69.97−69.06) in mAP, which verifies the effectiveness of our proposed ROD. In particular, similar results can be obtained in SYSU-MM01

dataset (see Table IV), and DrRD, ROD and MiDR are also effective in improving the accuracy of the experiment on all evaluation metrics. Take the results from Table III and IV into comprehensive consideration, it is not hard to draw following conclusions:

- No matter on RegDB dataset or SYSU-MM01 dataset, our proposed components (DrRD, MiDR and ROD) all show superiority in improving RGB-IR ReID performance. To be specific, our DrRD method effectively implements spectrum information decoupling and identity information purification, which is benefited from a min-max adversarial training strategy. Besides, the comparative experiment between BL+DrRD and BL+DrRD+MiDR has proven the validity of our MiDR method, which helps further eliminate domain differences. In addition, BL+DrRD+MiDR+ROD outperforms BL+DrRD+MiDR indicates that reducing correlation between identity-related features and domain-related features is beneficial to performance improvement.
- It is obvious that BL+DrRD+MiDR+ROD achieves the best improvement on all evaluation metrics, which strongly confirms the feasibility and validity of our DMiR algorithm for RGB-IR ReID task.

Loss-level performance analysis In order to verify the effectiveness of the proposed method, we further conducted experimental analysis to show the loss-level importance of each module. Table V records the effect of different losses, including $\mathcal{L}^{ID}$, $\mathcal{L}^{WRT}$, $\mathcal{L}^{D}$, $\mathcal{L}^{SwVL}$, $\mathcal{L}^{PwSV}$, $\mathcal{L}^{ROD}$, on SYSU-MM01. According to the results, we can get the following observations. **(a)** Compared to $\mathcal{L}^{WRT}$, $\mathcal{L}^{ID}+\mathcal{L}^{WRT}$ obtains better performance, with improvement up to 3.57% (42.65−39.08) and 1.82% (43.13−41.31) on Rank-1 and mAP, respectively. $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}$ has outperformed $\mathcal{L}^{ID}+\mathcal{L}^{D}$ by 0.49% (44.86−44.37), 1.81% (75.22−73.41), 2.15% (85.88−83.73) and 1.21% (45.31−44.10) on Rank-1, Rank-5, Rank-10 and mAP, respectively. In addition, $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}$ performs better than $\mathcal{L}^{ID}+\mathcal{L}^{WRT}$, with Rank-1 of 44.86% and mAP of 45.31%. These results indicate that the introduction of $\mathcal{L}^{ID}$, $\mathcal{L}^{WRT}$, $\mathcal{L}^{D}$ are all necessary for improving the RGB-IR ReID performance. **(b)** $\mathcal{L}^{ID}+\mathcal{L}^{SwVL}+\mathcal{L}^{PwSV}$ is superior to $\mathcal{L}^{ID}+\mathcal{L}^{SwVL}$ and $\mathcal{L}^{ID}+\mathcal{L}^{PwSV}$, with Rank-1, Rank-5, Rank-10 and mAP of 46.99%, 74.59%, 84.56% and 45.40%, respectively. Moreover, $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}+\mathcal{L}^{SwVL}+\mathcal{L}^{PwSV}$ outperforms $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}$, with improvement up to 3.41% (48.27−44.86) and 2.38% (47.69−45.31) on Rank-1 and mAP, respectively. The results demonstrate that the combination $\mathcal{L}^{PwSV}$ and $\mathcal{L}^{SwVL}$ can effectively improve cross-modality ReID performance. **(c)** Compared with $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}+\mathcal{L}^{SwVL}+\mathcal{L}^{PwSV}$, $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}+\mathcal{L}^{SwVL}+\mathcal{L}^{PwSV}+\mathcal{L}^{ROD}$ obtains better performance, which indicates the introduction of ROD is effective to improve the network performance. **(d)** The performance of $\mathcal{L}^{ID}+\mathcal{L}^{WRT}+\mathcal{L}^{D}+\mathcal{L}^{SwVL}+\mathcal{L}^{PwSV}+\mathcal{L}^{ROD}$ is better than all other variants, with Rank-1, Rank-5, Rank-10 and mAP of 50.54%, 78.82%, 88.12% and 49.29%, respectively. Therefore, the whole scheme leads to disentanglement of spectrum-specific information, as well as purification of identity information.

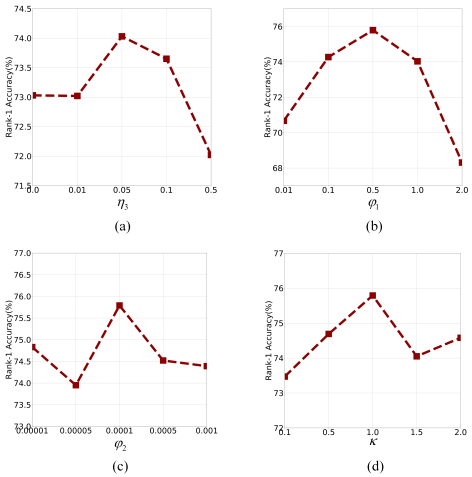

Fig. 3. Hyperparameter analysis in RegDB dataset. Sub-graphs (a)(b)(c)(d) reveal how the Rank-1 accuracy changes with η_3, φ_1, φ_2, κ, respectively.

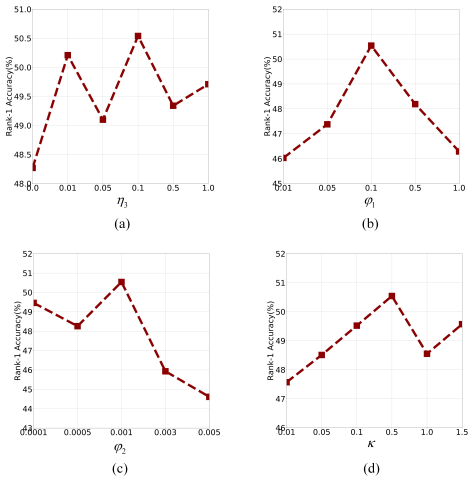

Fig. 4. Hyperparameter analysis in SYSU-MM01 dataset. Sub-graphs (a)(b)(c)(d) represent how the Rank-1 accuracy changes with η_3, φ_1, φ_2, κ, respectively.

Hyperparameter analysis For the hyperparameters η_3, φ_1, φ_2 and κ, we implement ablation within a certain range, and take the optimal value as our final experimental results. The hyperprameter analysis are conducted in both RegDB dataset and SYSU-MM01 dataset, and the Rank-1 accuracy curves of η_3, φ_1, φ_2, κ, are shown in Figures 3-4. As illustrated in Figure 3, in the RegDB dataset, **(a)** When η_3 is set to 0, the Rank-1 accuracy is lower than some situations where η_3 is 0.05 or 0.1, similar to that where η_3 is 0.01, and higher than that where η_3 is 0.5. Evidently, $\eta_3 = 0.5$ has poor ReID performance while $\eta_3 = 0.05$ achieves the best performance with accuracy of 74.03%. Besides, compared to $\eta_3 = 0$, an appropriate η_3 value, i.e., $\eta_3 = 0.05$ and $\eta_3 = 0.1$, can further improve cross-modality ReID performance, which directly demonstrates the necessity of our ROD method. **(b)** As φ_1 increases, the RGB-IR ReID performance rises and then falls, reaching a peak at $\varphi_1 = 0.5$ with the accuracy of 75.79%. It also tells that

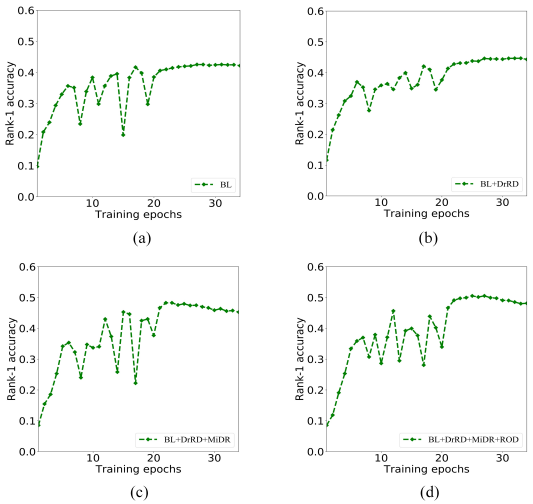

Fig. 5. Training curves on SYSU-MM01 dataset. Sub-graphs (a)(b)(c)(d) represent BL, BL+DrRD, BL+DrRD+MiDR and BL+DrRD+MiDR+ROD, respectively.

φ_1 should not be set too large or too small. **(c)** The Rank-1 accuracy changes irregularly with φ_2 ranges from 0.00001 to 0.001. $\varphi_2 = 0.0001$ leads to the best accuracy of 75.79%. On the contrary, $\varphi_2 = 0.00005$ gives rise to the lowest Rank-1 accuracy of 73.95%. **(d)** When κ is set to 0.1, the corresponding Rank-1 accuracy is lower than all the other settings. The most appropriate value of κ is 1.0, resulting in 2.32% (75.79−73.47) of the advantage. Figure 4 reveals the trend in SYSU-MM01 dataset: **(a)** $\eta_3 = 0.1$ yields the highest Rank-1 accuracy of 50.54% while $\eta_3 = 0$ generates the lowest one (48.27%). Similar to the case in the RegDB dataset, it demonstrates again that our ROD method is indispensable. **(b)** The DMiR method obtains the best performance when setting an appropriate value $\varphi_1 = 0.1$. **(c)** The curve first falls, then rises and falls again with φ_2 increases from 0.0001 to 0.005. The network obtains the best performance when setting φ_2 to 0.001, with the accuracy of 50.54%. **(d)** The curve is almost an escalating trend except at $\kappa = 1.0$. $\kappa = 0.5$ brings about the best performance of 50.54% and $\kappa = 0.01$ results in the worst one with 47.57%. In order to better reflect how such hyperparameters affect the Rank-1 accuracy, we briefly make a summary as follows:

- In most cases, the corresponding curves show a similar trend in both RegDB and SYSU-MM01 datasets.
- Our ROD method is an important part of DMiR algorithm and is indispensable for improving performance.
- The MiDR can effectively enhance identity-discriminative representation learning when adopting an appropriate value for φ_1 and φ_2.
- The trade-off parameter κ may affect the performance of DrRD, and a proper κ can effectively train Domain-Net to decouple modality-related information.

Convergence analysis In order to analyse the convergence trend of the proposed method, we carry out experiment on SYSU-MM01 dataset and plot the training curves under different settings, including BL, BL+DrRD, BL+DrRD+MiDR and BL+DrRD+MiDR+ROD. As shown in Figure 5, the training curves concretely describe how the Rank-1 accuracy changes with training epochs (mainly range from 1 to 34). We find that both BL (see Fig. 5 (a)) and BL+DrRD (see Fig. 5 (b)) have a stable training process, which indicates that the introducing of min-max adversarial disentanglement strategy in DrRD will not affect the network training stability. Moreover, it is not hard to discover that no matter which setting, the training curve totally tends to converge around the 25th epoch and remains steady as the epoch increases. Take BL+DrRD+MiDR+ROD (see Fig. 5 (d)) as an example, the Rank-1 accuracy reaches 50.54% and keeps stable even though the training epochs continue to increase. Therefore, we can easily draw a conclusion that the proposed DMiR approach has stable training and easy convergence behavior.

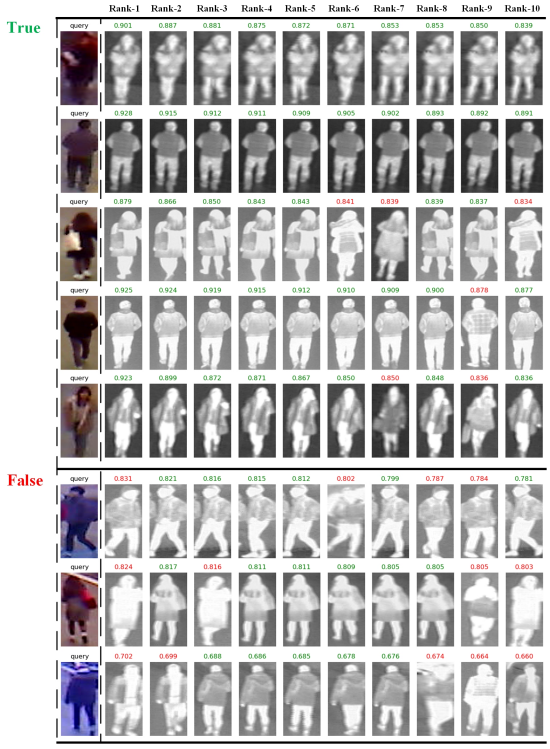

Fig. 6. Top-10 ranking list of some RGB query images on RegDB dataset. The images with value in green belong to the same identity as the given query, while those in red is opposite. If the Rank-1 image matches the given query, we define this re-identification process as 'True' in green. On the contrary, if they does not match, we classify it into 'False' in red.

Visualization analysis For the sake of showing the performance of our DMiR algorithm vividly, we provide matching results between probe and gallery on RegDB dataset, and correct classification results and misclassification results are shown in Figure 6. We first take a visible light image as a query, and arrange them in the first column. The retrieved infrared images are sorted according to similarity scores from left to right (from Rank-1 to Rank-10). We define correct match as 'True' in green and wrong match as 'False' in red. It is clear that the proposed DMiR model gets good shots under most situations, which means that our method has excellent performance.

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE REGDB DATASET. OUR METHOD OUTPERFORMS THE EXISTING METHODS ON RANK-1, RANK-10, RANK-20 AND MAP METRICS.

Method	Visible to Infrared				Infrared to Visible			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad (2017)	17.74	34.21	44.35	18.90	16.63	34.68	44.25	17.82
BDTR (2018)	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
HCML (2018)	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
TONE (2018)	16.87	34.03	-	14.92	-	-	-	-
HSME (2019)	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
eBDTR (2019)	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
D2RL (2019)	43.4	66.1	76.3	44.1	-	-	-	-
AlignGAN (2019)	57.9	-	-	53.6	56.3	-	-	53.4
DFE (2019)	70.13	86.32	91.96	69.14	67.99	85.56	91.41	66.70
SDL (2020)	26.47	51.34	61.22	23.58	25.74	50.23	59.66	22.89
JSIA-ReID (2020)	48.5	-	-	49.3	48.1	-	-	48.9
Hi-CMD (2020)	70.93	86.39	-	66.04	-	-	-	-
HAT (2021)	71.83	87.16	92.16	67.56	70.02	86.45	91.61	66.30
DMiR (our)	**75.79**	**89.86**	**94.18**	**69.97**	**73.93**	**89.87**	**93.98**	**68.22**

TABLE VII
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE SYSU-MM01 DATASET. OUR METHOD EXCEEDS THE EXISTING METHODS ON RANK-1, RANK-10, RANK-20 AND MAP METRICS.

Method	All-Search				Indoor-Search			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Zero-Pad (2017)	14.80	54.12	71.33	15.95	20.58	68.38	85.79	26.92
BDTR (2018)	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
cmGAN (2018)	26.97	67.51	80.56	27.80	31.63	77.23	89.18	42.19
HCML (2018)	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
TONE (2018)	12.52	50.72	68.60	14.42	20.82	68.86	84.46	26.38
HSME (2019)	20.68	62.74	77.95	23.12	-	-	-	-
eBDTR (2019)	27.82	67.34	81.34	28.42	32.46	77.42	89.62	42.46
D2RL (2019)	28.9	70.6	82.4	29.2	-	-	-	-
AlignGAN (2019)	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
DFE (2019)	48.71	88.86	95.27	48.59	52.25	89.86	95.85	59.68
SDL (2020)	28.12	70.23	83.67	29.01	32.56	80.45	90.67	39.56
JSIA-ReID (2020)	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
Hi-CMD (2020)	34.94	77.58	-	35.94	-	-	-	-
HAT (2021)	**55.29**	**92.14**	**97.36**	**53.89**	**62.10**	**95.75**	**99.20**	**69.37**
DMiR (our)	50.54	88.12	94.86	49.29	53.92	92.50	97.09	62.49

E. Comparison to State-of-the-Art Approaches

RegDB We compare our method with some state-of-the-art methods, including Zero-Pad [4], BDTR [55], HCML [9], TONE [9], HSME [58], eBDTR [55], D2RL [37], AlignGAN [12], DFE [56], SDL [13], JSIA-ReID [40], Hi-CMD [7], and HAT [38], are shown in Table VI. From the results in Table VI, the following conclusions can be drawn:

- Our DMiR has outperformed Zero-Pad by 58.05% (75.79−17.74) and 51.07% (69.97−18.90) in Rank-1 and mAP respectively (Visible to Infrared). The increasement in Infrared to Visible setting is 57.3% (73.93−16.63) in Rank-1 and 50.4% (68.22−17.82) in mAP. Evidentlly, our DMiR is definitely a feasible scheme for RGB-IR ReID task.
- Our DMiR has better performance than methods such as BDTR, HCML, TONE, HSME, eBDTR, DFE, HAT, etc. For example, there exists an improvement compared to HAT: 3.96% (75.79−71.83) and 2.41% (69.97−67.56) in Rank-1 and mAP (Visible to Infrared), and 3.91% (73.93−70.02) and 1.92% (68.22−66.30) in Rank-1 and mAP (Infrared to Visible). These results indicate that the DMiR approach can learn identity-discriminative features for cross-domain person ReID.
- Compared with some disentanglement-based methods, such as SDL and Hi-CMD, our DMiR has surpassed SDL by 49.32% (75.79−26.47), 46.39% (69.97−23.58) in Rank-1 and mAP (Visible to Infrared), and 48.19% (73.93−25.74), 45.33% (68.22−22.89) in Rank-1 and mAP (Infrared to Visible). When compared with Hi-CMD, the increment is 4.86% (75.79−70.93), 3.93% (69.97−66.04) in Rank-1 and mAP in Visible to Infrared setting. The above comparison has verified the validity of our idea of modal decoupling.

SYSU-MM01 We evaluate our DMiR method against 13 previous methods, including Zero-Pad [4], BDTR [55], cmGAN [11], HCML [9], TONE [9], HSME [58], eBDTR [55], D2RL [37], AlignGAN [12], DFE [56], SDL [13], Hi-CMD [7], JSIA-ReID [40], and HAT [38], on SYSU-MM01 dataset, as shown in Table VII. According to the results, we make several typical comparisons in detail and draw corresponding conclusions as follows:

- Compared with methods based on domain adaptation

and GANs, e.g., cmGAN and AlignGAN, our DMiR shows a huge advantage. In particular, our DMiR has surpassed AlignGAN by 8.14% (50.54−42.4) in Rank-1 and 8.59% (49.29−40.7) in mAP, 8.02% (53.92−45.9) in Rank-1 and 8.19% (62.49−54.3) in mAP, in all-search and indoor-search scenarios, respectively. With regard to cmGAN, the improvement mainly includes 23.57% (50.54−26.97) in Rank-1, 21.49% (49.29−27.80) in mAP (all-search), and 22.29% (53.92−31.63) in Rank-1, 20.3% (62.49−42.19) in mAP (indoor-search). The above detailed results prove that our DMiR is more effective than domain adaptation and GANs based methods.

- When compared with Hi-CMD, our DMiR has increased by 15.60% (50.54−34.94), 13.35% (49.29−35.94) in Rank-1 and mAP in all-search scenario. Besides, our method also obviously outperforms SDL model. Therefore, it is convincing that orthogonal modality adversarial decoupling is necessary and valid.
- Note that our DMiR performs better than HAT on VIS-Thermal RegDB dataset, while HAT outperforms ours on VIS-NIR SYSU-MM01 dataset. Since NIR images contain more high-frequency detail than thermal images, HAT can effectively preserve structure information in VIS-NIR tasks, while it does not work well in the larger domain gaps of VIS-Thermal tasks. In DMiR, the DrRD, ROD and MiDR are specifically designed for cross-domain disentanglement and can effectively decouple modal differences despite on large modal difference tasks.

Totally speaking, the design of our DMiR, including DrRD, ROD and MiDR, is effective for cross-domain ReID.

F. Generalization Analyses of the Proposed Method

The above experiments demonstrate the effectiveness of our DMiR method in cross-modal pedestrian re-identification. To verify the generality of the model in other cross-domain image identification tasks, we further experiment in the challenging NIR-VIS face database CASIA NIR-VIS 2.0 (first fold) [59], which contains 17,580 images of 725 people. We follow the training and testing protocol in [59], and approximately select 360 individuals as training set. As shown in Table VIII, the BL+DrRD exceeds the BL on both rank-1 accuracy and VR@FAR=0.01%, indicating that the DrRD can eliminate spectrum variations via a min-max adversarial disentanglement process. BL+DrRD+ROD performs better than BL+DrRD, and BL+DrRD+MiDR+ROD performs better than BL+DrRD+ROD, which shows that both ROD and MiDR can improve network performance. In particular, the BL+DrRD+MiDR+ROD obtains the best performance with rank-1 accuracy and VR@FAR=0.01% of 99.2% and 96.6%. These results indicate that our proposed components (i.e., DrRD, MiDR and ROD) all show superiority in improving network performance.

TABLE VIII
RANK-1 ACCURACY AND VERIFICATION RATE ON CASIA NIR-VIS 2.0 DATASET.

Method	Rank-1(%)	VR@$FAR = 0.01\%$(%)
BL	98.2	92.0
BL+DrRD	98.5	93.5
BL+DrRD+ROD	98.9	94.9
BL+DrRD+MiDR+ROD	**99.2**	**96.6**

V. CONCLUSION

To solve the RGB-IR ReID problems, in this paper, we propose a novel adversarial decoupling and modality-invariant representation learning method to explore potential domain-independent discriminative representations for cross-modality pedestrian images. Our model consists of three key components, including domain-related representation disentanglement, modality-invariant discriminative representation and representation orthogonal decorrelation. Identity-related features and domain-related features are extracted by two subnets, i.e., Identity-Net and Domain-Net, respectively. Specifically, in Domain-Net, DrRD is firstly introduced to reduce the modality gap between visible image and infrared image via a min-max adversarial disentanglement process. Secondly, in Identity-Net, MiDR is introduced to enhance identity-discriminative learning by exploring potential domain-invariant features. Finally, ROD constraint is presented to make the two decomposed features unrelated to each other, which can more effectively separate the two-component information and enhance feature representations. The whole scheme successfully accomplishes the disentanglement of spectrum-dependent information as well as purification of identity information. Extensive experiments are conducted on two challenging RGB-IR ReID datasets, and the results demonstrate the effectiveness of our method.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," arXiv preprint arXiv: 1610.02984, 2016.

[2] Y. C. Chen, X. Zhu, et al, "Person Re-identification by Camera Correlation Aware Feature Augmentation," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 392-408, 2018.

[3] Z. J. Zhuang, L. H. Wei, L. X. Xie, H. Z. Ai, and Q. Tian, "Camera-based Batch Normalization: An Effective Distribution Alignment Method for Person Re-identification," IEEE Trans. on Circuits and Systems for Video Technology, Early Access Article, 2021.

[4] A. Wu, W. S. Zheng, H. X. Yu, S. Gong, and J. Lai, "RGB-Infrared Cross-Modality Person Re-Identification," IEEE Int. Conf. on Computer Vision, 2017, pp. 5390-5399.

[5] X. Z. Xiang, N. Lv, Z. T. Yu, M. L. Zhai, and A. E. Saddik, "Cross-modality Person Re-identification Based on Dual-path Multi-branch Network," IEEE Sensors, vol. 19, no. 23, pp. 11706-11713, 2019.

[6] Y. B. Zhao, J. W. Lin, Q. Xuan, and X. Xi, "HPILN: a Feature Learning Framework for Cross-Modality Person Re-identification," IET Image Processing, vol. 13, no. 14, pp. 2897-2904, 2019.

[7] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, "Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2020, pp. 10257-10266.

[8] N. Tekeli, A. B. Can, "Distance Based Training for Cross-modality Person Re-identification," IEEE Int. Conf. on Computer Vision Workshop, 2019, pp. 4540-4549.

[9] M. Ye, X. Y. Lan, J. W. Li, and P. C. Yuen, "Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification," AAAI Conf. on Artificial Intelligence, 2018, pp. 7501-7508.

[10] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Int. Conf. on Neural Information Processing Systems, 2014.

[11] P. Y. 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, 2018, pp. 677-683.

[12] G. A. Wang, T. Z. Zhang, et al, "RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment," IEEE Int. Conf. on Computer Vision, 2019, pp. 3623-3632.

[13] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "SDL: Spectrum-Disentangled Representation Learning for Visible-Infrared Person Re-Identification," IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3422-3432, 2020.

[14] W. P. Hu and H. F. Hu, "Adversarial Disentanglement Spectrum Variations and Cross-modality Attention Networks for NIR-VIS Face Recognition," IEEE Trans. on Multimedia, vol.23, no. 1, pp.145-160, 2021.

[15] W. P. Hu and H. F. Hu, "Orthogonal Modality Disentanglement and Representation Alignment Network for NIR-VIS Face Recognition," IEEE Trans. on Circuits and Systems for Video Technology, Early Access, 2021.

[16] E. Chanho, and H. Bumsub, "Learning Disentangled Representation for Robust Person Re-identification," Int. Conf. on Neural Information Processing Systems, 2019, pp. 5298-5309.

[17] Y. J. Li, C. S. Lin, Y. B. Lin, and Y. C. F. Wang, "Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation," IEEE Int. Conf. on Computer Vision, 2019, pp. 7918-7928.

[18] Z. D. Zheng, X. D. Yang, Z. D. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint Discriminative and Generative Learning for Person Re-Identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 2133-2142.

[19] Y. Li, X. M. Tian, M. M. Gong, Y. J. Liu, T. L. Liu, K. Zhang, and D. C. Tao, "Deep domain generalization via conditional invariant adversarial networks," European Conf. on Computer Vision, 2018, pp. 624-639.

[20] G. Delorme, Y. H. Xu, et al, "CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification," IEEE Int. Conf. on Pattern Recognition, 2021, pp. 4428-4435.

[21] W. P. Hu and H. F. Hu, "Disentangled Spectrum Variations Networks for NIR-VIS Face Recognition," IEEE Trans. on Multimedia, vol. 22, no. 5, pp. 1234-1248, 2020.

[22] Y. R. Zhao, et al, "Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 4908-4917.

[23] 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. 605-634, 2017.

[24] S. Karanam, Z. Wu, and R. J. Radke, "Learning Affine Hull Representations for Multi-Shot Person Re-Identification," IEEE Trans. on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2500-2512, 2018.

[25] M. Jia, Y. Zhai, S. Lu, S. Ma, and J. Zhang, "A Similarity Inference Metric for Rgb-infrared Cross-modality Person Re-identification," arXiv preprint arXiv:2007.01504v1, 2020.

[26] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style Normalization and Restitution for Generalizable Person Re-identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2020, pp. 3143-3152.

[27] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention Driven Person Re-identification," Pattern Recognition, vol. 86, pp. 143-155, 2019.

[28] Y. Huang, J. Xu, Q. Wu, Y. Zhong, P. Zhang and Z. Zhang, "Beyond Scalar Neuron: Adopting Vector-Neuron Capsules for Long-Term Person Re-Identification," IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3459-3471, 2020.

[29] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning Discriminative Features with Multiple Granularities for Person Re-identification," ACM International Conference on Multimedia, 2018, pp. 274-282.

[30] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline)," European Conf. on Computer Vision, 2018, pp. 480-496.

[31] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2020, pp. 6449-6458.

[32] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive Pose Attention Transfer for Person Image Generation," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 2347-2356.

[33] J. F. Song, Y. X. Yang, et al, "Generalizable Person Reidentification by Domain-invariant Mapping Network," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 719-728.

[34] Q. Z. Yang, H. X. Yu, et al, " Patch-based Discriminative Feature Learning for Unsupervised Person Re-identification. IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 3633-3642.

[35] M. Zheng, S. Karanam, Z. Y. Wu, and R. J. Radke, "Re-identification with Consistent Attentive Siamese Networks," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 5735-5744.

[36] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, "Hetero-Center Loss for Cross-Modality Person Re-identification," arXiv preprint arXiv: 1910.09830v1, 2019.

[37] Z. X. Wang, Z. Wang, Y. Q. Zheng, Y. Y. Chuang, and S. Satoh, "Learning to Reduce Dual-level Discrepancy for Infrared-visible Person Re-identification," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 618-626.

[38] M. Ye, J. B. Shen, and L. Shao, "Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning," IEEE Trans. on Information Forensics and Security, vol. 16, pp. 728-739, 2021.

[39] C. Y. Fu, Y. B. Hu, X. Wu, H. L. Shi, T. Mei, and R. He, "CM-NAS: Cross-Modality Neural Architecture Search for Visible-Infrared Person Re-Identification," IEEE Int. Conf. on Computer Vision, 2021.

[40] G. A. Wang, T. Z. Zhang, Y. Yang, J. Cheng, J. L. Chang, X. Liang, and Z. G. Hou, "Cross-modality Paired-images Generation for Rgb-infrared Person Re-identification," AAAI Conf. on Artificial Intelligence, 2020.

[41] X. Huang, M. Y. Liu, et al, "Multimodal Unsupervised Image-to-image Translation," European Conf. on Computer Vision, 2018, pp. 172-189.

[42] K. K. Singh, U. Ojha, and Y. J. Lee, "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 6490-6499.

[43] A. G. Garcia, J. V. deWeijer, and Y. Bengio, "Image-to-image Translation for Cross-domain Disentanglement," Int. Conf. on Neural Information Processing Systems, 2018, pp. 1287-1298.

[44] W. P. Hu, H. F. Hu, "Dual Adversarial Disentanglement and Deep Representation Decorrelation for NIR-VIS Face Recognition," IEEE Trans. on Information Forensics and Security, vol. 16, no. 1, pp. 70-85, 2020.

[45] R. He, X. Wu, Z. N. Sun, and T. N Tan, "Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1803-1816, 2019.

[46] W. P. Hu, W. J. Yan, and H. F. Hu, "Dual Face Alignment Learning Network for NIR-VIS Face Recognition," IEEE Trans. on Circuits and Systems for Video Technology, Early Access Article, 2021.

[47] J. M. Bao, D. Chen, F. Wen, H. Q. Li, and G. Hua, "Towards Open-set Identity Preserving Face Synthesis," IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 6713-6722.

[48] L. Tran, X. Yin, and X. M. Liu, "Disentangled Representation Learning Gan for Pose-invariant Face Recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 1415-1424.

[49] H. Wang, D. H. Gong, Z. F. Li, and W. Liu, "Decorrelated Adversarial Learning for Age-invariant Face Recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 3527-3536.

[50] Z. Y. Zhang, L. Tran, X. Yin, Y. Atoum, X. M. Liu, J. Wan, and N. X. Wang, "Gait Recognition via Disentangled Representation Learning," IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 4710-4719.

[51] H. T. Zeng, W. P. Hu, D. H. Chen, and H. F. Hu, "Two-way constraint network for RGB-Infraredperson re-identification," ELECTRONICS LETTERS, vol. 57, no. 17, pp. 653-655, 2021.

[52] M. Ye, J. B. Shen, G. J. Lin, T. Xiang, L. Shao, and C. H. Hoi, "Deep Learning for Person Re-identification: A Survey and Outlook," IEEE Trans. on Pattern Analysis and Machine Intelligence, Early Access Article, 2021.

[53] X. Mestre, "Improved Estimation of Eigenvalues and Eigenvectors of Covariance Matrices Using Their Sample Estimates," IEEE Trans. on Information Theory, vol. 54, no. 11, pp. 5113-5129, 2008.

[54] S. Boyd, and L. Vandenberghe, "Convex Optimization," Cambridge University Press, 2009, pp. 215-287.

[55] M. Ye, X. Y. Lan, Z. Wang, et al, "Bi-directional Center-constrained Top-ranking for Visible Thermal Person Re-identification," IEEE Trans. on Information Forensics and Security, vol. 15, pp. 407-419, 2020.

[56] 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, 2019, pp. 57-65.

[57] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[58] Y. Hao, N. N. Wang, J. Li, X. B. Gao, "Hsme: Hypersphere Manifold Embedding for Visible Thermal Person Re-identification," AAAI Conf. on Artificial Intelligence, 2019, pp. 8385-8392.

[59] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2013, pp. 348-353.