Learning to Reduce Dual-level Discrepancy for Infrared-Visible Person Re-identification

Zhixiang Wang[1] Zheng Wang[2*] Yinqiang Zheng[2] Yung-Yu Chuang[1] Shin'ichi Satoh[2,3]
[1]National Taiwan University [2]National Institute of Informatics [3]The University of Tokyo

Abstract

Infrared-Visible person RE-IDentification (IV-REID) is a rising task. Compared to conventional person re-identification (re-ID), IV-REID concerns the additional modality discrepancy originated from the different imaging processes of spectrum cameras, in addition to the person's appearance discrepancy caused by viewpoint changes, pose variations and deformations presented in the conventional re-ID task. The co-existed discrepancies make IV-REID more difficult to solve. Previous methods attempt to reduce the appearance and modality discrepancies simultaneously using feature-level constraints. It is however difficult to eliminate the mixed discrepancies using only feature-level constraints. To address the problem, this paper introduces a novel Dual-level Discrepancy Reduction Learning (D^{2}RL) scheme which handles the two discrepancies separately. For reducing the modality discrepancy, an image-level sub-network is trained to translate an infrared image into its visible counterpart and a visible image to its infrared version. With the image-level sub-network, we can unify the representations for images with different modalities. With the help of the unified multi-spectral images, a feature-level sub-network is trained to reduce the remaining appearance discrepancy through feature embedding. By cascading the two sub-networks and training them jointly, the dual-level reductions take their responsibilities cooperatively and attentively. Extensive experiments demonstrate the proposed approach outperforms the state-of-the-art methods.

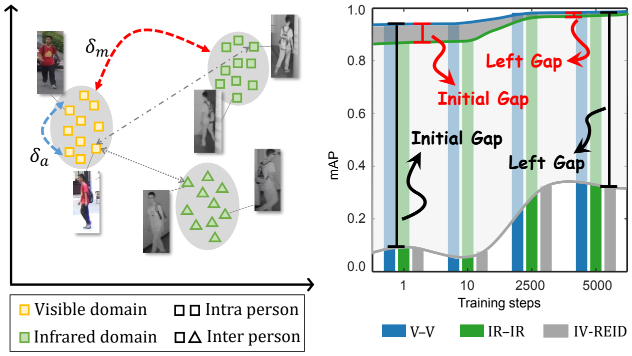


Figure 1. Illustration of the difficulty in the IV-REID task. (Left) The modality discrepancy δ_m is much larger than δ_a, the appearance discrepancy. Thus, the modality discrepancy could result in the intra-person distance $\mathcal{D}$(🟨, 🟩) being larger than the inter-person distance $\mathcal{D}$(🟨, ▲). (Right) The performance obtained by [23] on the RegDB dataset using only the feature-level constraints. "V-V" and "IR-IR" respectively denote the performance of visible-visible and infrared-infrared re-ID. The red gap indicates the performance gap between visible-visible and infrared-infrared single-modality re-ID. The black gap denotes the performance gap between cross-modality re-ID and single-modality re-ID. It is clear that the re-ID problem across modalities is much more difficult than the one with the same modality.

1. Introduction

Person re-identification (re-ID) has recently received increasing attention in the computer vision community [18, 17, 6, 7, 20, 28] because of its great importance in video surveillance. Most current re-ID methods rely on person's appearance under good visible light conditions. Under poor illumination conditions, due to poor appearance, conventional re-ID models could become "blind in dark". In practice, for dealing with poor illumination, most surveillance cameras automatically switch from the visible mode to the infrared mode in the dark [21]. Consequently, this raises a new task in which, given a visible (or infrared) image of a specific person, the goal is to find the corresponding infrared (or visible) images of the person captured by other spectrum cameras [21, 22, 23, 2]. This cross-modality image matching task is named Infrared-Visible person RE-IDentification (IV-REID).

Compared to the conventional re-ID task, IV-REID encounters additional modality discrepancy resulting from the differences between imaging processes of different spectrum cameras, in addition to the person's appearance discrepancy caused by viewpoint changes, pose variations, scale changes and deformations. The modality discrepancy is often more significant than the appearance discrepancy.

[*] Corresponding Author

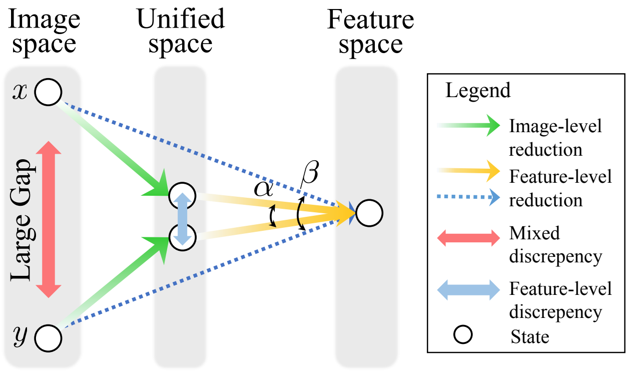


Figure 2. A high-level overview of our approach. Previous methods reduce the total discrepancy by converting images to the feature space and using only the feature-level constraint (the blue dashed line). Our dual-level solution first converts images into a unified space (the green arrows) and then embeds them into the feature space (the orange arrow). After the image-level modality reduction by unifying image representations, the gap becomes much smaller than the one in the original image space. Thus, the feature-level embedding can be effective in reducing the remaining appearance discrepancy.

The intra-person (images from the same person) distance across visible and infrared cameras is often larger than the inter-person (images from different persons) distance of the same type of cameras. While the main goal of IV-REID is still to maximize the inter-person distance and meanwhile to minimize the intra-person distance. The co-existed modality and appearance discrepancies make IV-REID difficult (as Figure 1 shows). As far as we know, all previous methods [21, 22, 23, 2] regard the modality discrepancy δ_m as a part of the appearance discrepancy δ_a and attempt to reduce the mixed discrepancy δ_m+δ_a using feature-level constraints employed by most conventional re-ID methods. Note that the performance gap between single-modality re-ID ("V-V" or "IR-IR") and cross-modality IV-REID is extremely large. It indicates that the modality discrepancy cannot be effectively eliminated using only feature-level constraints.

Figure 2 gives the main idea of the proposed method. Because they are taken with different modalities, infrared images and visible images have quite different appearances. Thus, direct mappings of them into the feature space can not be effective. For alleviating the problem, we propose to reduce the modality discrepancy first by unifying the image representations using image-level conversion. More specifically, we form a multi-spectral image by augmenting an infrared image with its visible counterpart or a visible image with its infrared version. In the unified space, the appearance discrepancy is greatly alleviated. Thus, conventional re-ID methods can be effective in reducing the remaining appearance discrepancy by feature-level constraints.

With the idea in mind, we propose a novel *dual-level* solution, named Dual-level Discrepancy Reduction Learning (D^{2}RL). We separate the modality discrepancy apart, and alleviate it through the image-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{I}$ which unifies image representations by synthesizing multi-spectral images from the given visible or infrared images. The appearance discrepancy is then handled by the feature-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{F}$, where feature embedding is more effective with the unified representation. These two sub-networks are cascaded and jointly optimized in an end-to-end manner. To this end, $\mathcal{T}_\mathcal{F}$ benefits $\mathcal{T}_\mathcal{I}$ to generate spectral images more discriminatively, and meanwhile $\mathcal{T}_\mathcal{I}$ provides $\mathcal{T}_\mathcal{F}$ with more translated samples. The contributions of this paper are summarized below:

- A novel dual-level discrepancy reduction learning scheme is introduced. We are the first to decompose the mixed modality and appearance discrepancies and handle them separately.
- Our end-to-end scheme enforces these two sub-networks benefit each other. The balance between them affects the performance.
- Extensive experiments on two datasets demonstrate the superior performance of our proposed approach compared to the state-of-the-art methods.

2. Related Work

Single-modality re-ID. The conventional re-ID researches mainly focus on the challenges of appearance changes in the single visible modality, such as image misalignment [6], viewpoint variations [11] and scale changes [16, 19]. Li *et al.* [6] formulated a harmonious attention CNN model for joint learning of pixel and regional attention to optimize re-ID performance with misaligned images. Liu *et al.* [11] proposed a pose transferrable framework for generating samples with rich pose variations. Wang *et al.* [16] combined effective embedding schemes built on multiple layers from high- and low-level details. Wang *et al.* [19] cascaded multiple super-resolution networks to overcome the resolution misalignment problem. Existing state-of-the-art single-modality re-ID methods are very effective in reducing the appearance discrepancy as their retrieval accuracy has already surpassed the accuracy of human [24].

Infrared-visible re-ID. For the IV-REID problem, in addition to the appearance discrepancy, the modality discrepancy needs to be addressed. Existing methods attempt to reduce the mixed appearance and modality discrepancies using feature embedding frameworks similar to conventional re-ID methods. Wu *et al.* [21] proposed a deep zero-padding framework for shared feature learning under two different modalities. Ye *et al.* [22] introduced a two-step framework for feature learning and metric learning. They [23] also proposed an end-to-end dual-path network to learn common representations. Dai *et al.* [2] designed a network to learn discriminative representations from different modalities. Due to the modality discrepancy brought in by different spectrum with dramatic image-level unbalance, all these feature-level methods cannot obtain satisfactory results.

Image generation meets re-ID. Recently developed GANs provide a powerful tool for image translation [29, 12, 1]. A lot of researches attempted to utilize GANs to generate more training samples and then facilitate solving the conventional re-ID problem. Ma *et al.* [13] manipulated the foreground, background and pose information, and generated images based on manipulated information. Li *et al.* [7] used the GAN to generate target-like images. There is another category of researches trying to deal with the problem of the domain gap using GANs. For the conventional re-ID task, the domain gap mainly lies in the camera style or illumination differences. Zhong *et al.* [28, 27] utilized the CycleGAN with label smooth regularization to generate person images with different camera styles. Deng *et al.* [4] also exploited the CycleGAN with self-similarity and domain-dissimilarity constraints. Li *et al.* [7] exploited the CycleGAN to generate images under different illumination conditions. With the similar idea, Wei *et al.* [20] proposed a person transfer GAN to bridge the domain gap. However, working towards the problem of different poses, illuminations, and camera styles, all these methods focus on generating visible images based on visible images.

3. Proposed Method

Let $X = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}\}$ and $Y = \{\boldsymbol{y} \mid \boldsymbol{y} \in \mathbb{R}^{H\times W\times 1}\}$ denote the visible image set and the infrared image set respectively, where H and W are the height and the width of images. Each image $\boldsymbol{x} \in X$ or $\boldsymbol{y} \in Y$ corresponds to a label $l \in \{1, 2, \ldots, N_p\}$, and N_p is the number of person identities. Given an infrared (or visible) query image $\boldsymbol{y}$ (or $\boldsymbol{x}$) and a visible (or infrared) gallery set X (or Y), the goal of the IV-REID task is to propose a ranking list R of the gallery set, in which the images with the same identity as the query image should be ranked to the top. A common strategy is projecting $\boldsymbol{x}$ and $\boldsymbol{y}$ to a feature space through feature embedding, $\boldsymbol{f}_x = h_x(\boldsymbol{x})$ and $\boldsymbol{f}_y = h_y(\boldsymbol{y})$, where $\boldsymbol{f}_x \in \mathbb{R}^d$ and $\boldsymbol{f}_y \in \mathbb{R}^d$, and then generating the ranking list R using the distance between them, $\boldsymbol{f}_x^T \boldsymbol{f}_y$.

We propose a new strategy to replace the direct mapping function h_x and h_y. Figure 3 shows the framework of our proposed method. It consists of two sub-networks: (1) the image-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{I}$ for reducing the modality discrepancy, and (2) the feature-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{F}$ for reducing the appearance discrepancy. These two sub-networks are cascaded and jointly optimized in an end-to-end manner. In the following, we describe their details.

3.1. Image-level discrepancy reduction — $\mathcal{T}_\mathcal{I}$

To reduce the modality discrepancy, $\mathcal{T}_\mathcal{I}$ exploits two variational autoencoders (VAEs) for style disentanglement followed by two GANs for domain specific image generation. $\mathcal{T}_\mathcal{I}$ translates the visible (infrared) image $\boldsymbol{x}$ ($\boldsymbol{y}$) to its infrared (visible) counterpart $\hat{\boldsymbol{x}}$ ($\hat{\boldsymbol{y}}$). Together, they form the multi-spectral image $[\boldsymbol{x}, \hat{\boldsymbol{x}}]$ (or $[\hat{\boldsymbol{y}}, \boldsymbol{y}]$) to provide a unified representation for reducing the modality discrepancy.

Style disentanglement. It consists of two encoder-decoder pairs: $\text{VAE}_v = \{E_v, G_v\}$ and $\text{VAE}_i = \{E_i, G_i\}$, that are responsible for visible and infrared modality disentanglement respectively. For VAE_v, given a visible input $\boldsymbol{x} \in X$, the encoder E_v first maps $\boldsymbol{x}$ to the latent vector $\boldsymbol{z}$, and then the decoder G_v reconstructs the input from the latent vector $\boldsymbol{z}$. The reconstructed image is $\hat{\boldsymbol{x}}^{v\to v} = G_v(\boldsymbol{z}_v \sim q_v(\boldsymbol{z}_v|\boldsymbol{x}))$, where $q_v(\boldsymbol{z}_v|\boldsymbol{x})$ is the distribution of latent information $\boldsymbol{z}_v$. The loss for VAE_v is defined as

$$\mathcal{L}_{\text{VAE}_v}(E_v, G_v) = \lambda_0 \text{KL}(q_v(\boldsymbol{z}_v|\boldsymbol{x})||p_\eta(\boldsymbol{z})) - \lambda_1 \mathbb{E}_{\boldsymbol{z}_v \sim q_v(\boldsymbol{z}_v|\boldsymbol{x})} \left[\|\boldsymbol{x} - G_v(\boldsymbol{z}_v)\|_1 \right] , \quad (1)$$

where the hyper-parameters λ_0 and λ_1 control the weights of the objective terms, and the Kullback-Leibler divergence term (KL) penalizes deviation between the distribution of the latent information and the prior $p_\eta(\boldsymbol{z})$ which is a zero-mean Gaussian distribution. The ℓ_1 loss penalizes dissimilarity between the image and the reconstructed image, and also encourages sharp output images.

Domain specific image generation. Two generative adversarial networks $\text{GAN}_v = \{G_v, D_v\}$ and $\text{GAN}_i = \{G_i, D_i\}$ are employed to generate domain specific images from the style-free latent vector $\boldsymbol{z}$. In GAN_v, the generator G_v is expected to generate realistic visible images from the latent vector $\boldsymbol{z}$ that can fool the discriminator D_v, while the discriminator D_v is expected to discriminate real and synthetic visible images. Adversarial losses are utilized to play the minimax game, which can be expressed as

$$\mathcal{L}_{\text{GAN}_v}(E_i, G_v, D_v) = \lambda_2 \mathbb{E}_{\boldsymbol{x}\sim P_X}[\log D_v(\boldsymbol{x})] + \lambda_2 \mathbb{E}_{\boldsymbol{z}_i \sim q_i(\boldsymbol{z}_i|\boldsymbol{y})}[\log(1 - D_v(G_v(\boldsymbol{z}_i)))] , \quad (2)$$

where the hyper-parameter λ_2 controls the impact of GAN. The discriminator is trained to maximize Equation (2) while the generator tries to minimize it. The loss is used to ensure the translated images resemble images in the visible domain.

Cycle-consistency. The cycle consistency is used to further regularize the ill-posed unsupervised image-to-image translation problem. Similar to CycleGAN [29], our cycle-consistency loss is defined as

$$\mathcal{L}_{\text{CC}_v}(E_v, G_v, E_i, G_i) = \lambda_3 \mathbb{E}_{\boldsymbol{z}_i \sim q_i(\boldsymbol{z}_i|\boldsymbol{x}^{v\to i})} \left[\|\boldsymbol{x} - G_v(\boldsymbol{z}_i)\|_1 \right] , \quad (3)$$

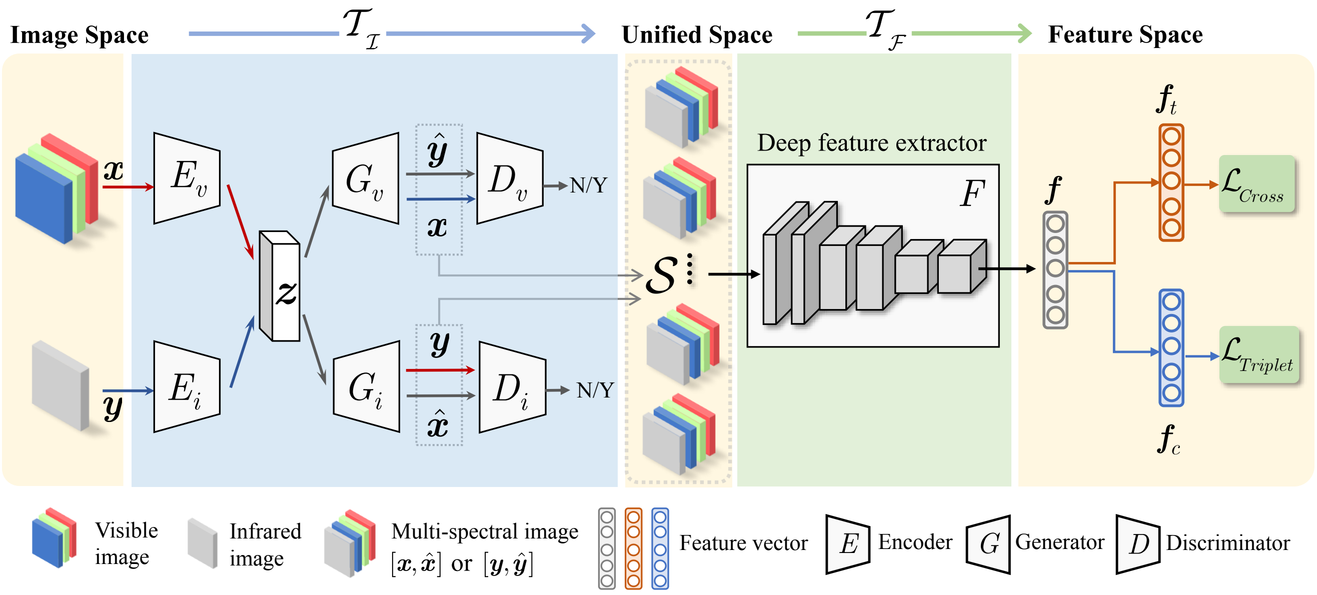


Figure 3. The framework of our proposed method. The image-level discrepancy reduction sub-network $\mathcal{T}_{\mathcal{I}}$ first projects inputs from the image space (visible or infrared modality) to the unified space, where the modality discrepancy is alleviated. Then, the feature-level discrepancy reduction sub-network $\mathcal{T}_{\mathcal{F}}$ is utilized to eliminate the remaining appearance discrepancy. The two sub-networks are cascaded and jointly optimized in an end-to-end manner.

where the negative log-likelihood objective term ensures a twice translated image resembles the input. The hyper-parameter λ_3 controls the weight of this objective term.

The losses $\mathcal{L}_{\text{VAE}_i}(E_i, G_i)$, $\mathcal{L}_{\text{GAN}_i}(E_v, G_i, D_i)$ and $\mathcal{L}_{\text{CC}_i}(E_i, G_i, E_v, G_v)$ can be similarly defined. More specifically, they are defined by substituting the subscript i for v, v for i, and $\boldsymbol{y}$ for $\boldsymbol{x}$ in Equation (1), Equation (2) and Equation (3).

Modality unification. There are three possible options for modality unification, *i.e.*, unifying images to the infrared modality, the visible modality or the multi-spectral modality. We choose to generate the multi-spectral images for modality unification for two reasons. First, the infrared and visible images are two representations of the same reflected light of the same person due to different imaging processes. They are potentially related and it is likely for them to reconstruct each other. Second, if we unify the images to the visible or infrared modality, some distinctive information in the visible or infrared modality may be lost.

Objective for training $\mathcal{T}_{\mathcal{I}}$. The total loss is a combination of the VAE loss, the GAN loss and the CC loss:

$$\mathcal{L}_{\mathcal{I}} = \mathcal{L}_{\text{VAE}_v} + \mathcal{L}_{\text{VAE}_i} + \mathcal{L}_{\text{GAN}_v} + \mathcal{L}_{\text{GAN}_i} + \mathcal{L}_{\text{CC}_v} + \mathcal{L}_{\text{CC}_i}\,. \quad (4)$$

By optimizing the above loss, we obtain a network $\mathcal{T}_{\mathcal{I}}$ which is able to translate a visible image $\boldsymbol{x}$ into its infrared counter part $\hat{\boldsymbol{x}}$ and translate an infrared image $\boldsymbol{y}$ into its visible counter part $\hat{\boldsymbol{y}}$. Thus, we can form a training set $\mathcal{S}$ by constructing multi-spectral images, $\boldsymbol{u}_v = [\boldsymbol{x}, \hat{\boldsymbol{x}}]$ and $\boldsymbol{u}_i = [\hat{\boldsymbol{y}}, \boldsymbol{y}]$, as the unified representations. This way, all images including both query and gallery images are represented in the same way and the modality discrepancy is greatly reduced.

3.2. Feature-level discrepancy reduction — $\mathcal{T}_{\mathcal{F}}$

Since $\mathcal{T}_{\mathcal{I}}$ has unified all the images to the same modality, a feature embedding network could be sufficient for reducing the appearance discrepancy. For each batch, we use $\mathcal{T}_{\mathcal{I}}$ to generate a sample set $\mathcal{S}$. The feature-level discrepancy reduction network $\mathcal{T}_{\mathcal{F}}$ plays a role of feature learning on the unified multi-spectral images generated by $\mathcal{T}_{\mathcal{I}}$. Given a multi-spectral image $\boldsymbol{u}$, sampled from $\mathcal{S}$, the deep feature extractor $F : \boldsymbol{u} \rightarrow \boldsymbol{f}$ maps it to the person descriptor $\boldsymbol{f}$. In particular, we use ResNet-50 as the backbone network of F and follow the training strategy in [27]. The last 1000-d fully connected (FC) layer is replaced with a new layer named as “FC-1024”. The person descriptor $\boldsymbol{f} \in \mathbb{R}^{1024}$ uses the output feature vector of “FC-1024” followed by Batch Normalization, ReLU and Dropout. The output $\boldsymbol{f}$ of the FC-1024 layer is then fed to two independent FC layers $\mathcal{H}_t$ and $\mathcal{H}_c$ for generating two feature vectors $\boldsymbol{f}_t \in \mathbb{R}^{128}$ and $\boldsymbol{f}_c \in \mathbb{R}^{N_p}$. Two types of loss functions are utilized to supervise the training of $\mathcal{T}_{\mathcal{F}}$. One is the triplet loss, employed for identity information learning, and the other is the cross-entropy loss, used for similarity learning. The triplet loss is coupled with $\boldsymbol{f}_t$ while the cross-entropy loss is tied with $\boldsymbol{f}_c$.

Triplet loss. It is used for similarity learning. It tries to reduce the feature distances between images of the same person and expand the distances between images of different people. The triplet loss can be formulated as following:

$$\mathcal{L}_{\mathcal{F}}^{T} = \sum_{\boldsymbol{f}_t^a, \boldsymbol{f}_t^p, \boldsymbol{f}_t^n \in \mathcal{S}} [\mathcal{D}(\boldsymbol{f}_t^a, \boldsymbol{f}_t^p) - \mathcal{D}(\boldsymbol{f}_t^a, \boldsymbol{f}_t^n) + \xi]_+ \,, \quad (5)$$

where $\boldsymbol{f}_t^a$ is the anchor point; $\boldsymbol{f}_t^p$ is a positive sample with the same identity with $\boldsymbol{f}_t^a$; and $\boldsymbol{f}_t^n$ is a negative sample with the different identity from $\boldsymbol{f}_t^a$. Note that $\boldsymbol{f}_t^a \neq \boldsymbol{f}_t^p$. ξ is a margin parameter. $\mathcal{D}(\cdot)$ calculates the Euclidean distance, and $[\boldsymbol{d}]_+ = \max(d, 0)$ truncates negative numbers to zero while keeping positive numbers the same.

Cross-entropy loss. It is employed for identity learning and is written as

$$\mathcal{L}_{\mathcal{F}}^{C} = -\frac{1}{N_b} \sum_{j=1}^{N_b} \log \boldsymbol{p}_j \,, \quad (6)$$

where $N_b = |\mathcal{S}|$ is the number of images in the training mini-batch; $\boldsymbol{p}$ is the predicted probability of the input belonging to the ground-truth class with $\boldsymbol{p} = \texttt{softmax}(\mathbf{W}\boldsymbol{f} + \mathbf{b})$, where $\mathbf{W}$ and $\mathbf{b}$ are the trainable weight and bias of $\mathcal{H}_c$.

Objective for training $\mathcal{T}_{\mathcal{F}}$. The loss is a combination of the cross-entropy and triplet losses as follows:

$$\mathcal{L}_{\mathcal{F}} = \lambda_4 \mathcal{L}_{\mathcal{F}}^{C} + \lambda_5 \mathcal{L}_{\mathcal{F}}^{T} \,. \quad (7)$$

3.3. End-to-end joint training

We optimize our network in an end-to-end manner, by cascading $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$ and minimizing the combined loss:

$$\arg\min_{\theta_{\mathcal{T}_{\mathcal{I}}}, \theta_{\mathcal{T}_{\mathcal{F}}}} (1-\gamma)\mathcal{L}_{\mathcal{I}} + \gamma \mathcal{L}_{\mathcal{F}} \,, \quad (8)$$

where $0 < \gamma < 1$ and it is a trade-off parameter for balancing the contributions of two sub-networks $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$.

4. Experiments

This section reports the experiment settings, implementation details, comparisons with other methods, the ablation study and analysis of our method.

4.1. Experiment settings

Datasets. We evaluated our method on two publicly available datasets: RegDB [15] and SYSU-MM01 [21].

- **RegDB** [15]. It was collected from two aligned cameras (one visible and one *far-infared*). It contains totally 412 persons. Each person has 10 visible images and 10 *far-infrared* images. We follow the evaluation protocol in [23] to randomly split the dataset into two halves, which are used for training and testing respectively.
- **SYSU-MM01** [21]. It is a large-scale dataset collected by six cameras (four visible and two *near-infared*), including both indoor and outdoor environments. It contains in total 491 persons, and each person was captured by at least two different cameras. Following [21], we adopt the most challenging single-shot all-search mode evaluation protocol. The training set contains 395 persons, with 22,258 visible images and 11,909 infrared images. The testing set contains 96 persons, with 3,803 infrared images for query and 301 randomly selected visible images as the gallery set.

Evaluation metrics. The standard Cumulative Matching Characteristics (CMC) curve and mean Average Precision (mAP) are adopted to evaluate the performance. Note that there is a slight difference with the conventional re-ID problem [25]. Images from one modality are used as the gallery set while the ones from the other modality as the probe set during testing.

4.2. Implementation details

Network architecture. The architecture of our proposed method is shown in Figure 3. The sub-network $\mathcal{T}_{\mathcal{I}}$ is based on UNIT[1]. The size of inputs and outputs is resized to $228 \times 228 \times 3$, for both visible and infrared images. For infrared images, the three channels are the same. $\mathcal{T}_{\mathcal{F}}$ is based on Open-reid[2] with the difference that our inputs have four channels.

Training strategy. In order to avoid mode collapse and over-fitting, we pretrained the sub-network $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$ respectively with the Market-1501 dataset [25], where we used the original images as the visible input, and the decomposed illuminations as the infrared input. Then, we jointly trained them in an end-to-end manner. Note that the SYSU-MM01 [21] dataset includes outdoor and indoor scenes. We trained them separately. We set weight parameters of the losses in $\mathcal{T}_{\mathcal{I}}$ as $\lambda_0 = 0.1$, $\lambda_1 = 100$, $\lambda_2 = 10$, $\lambda_3 = 100$ by following [12]. For the sub-network $\mathcal{T}_{\mathcal{F}}$, we set $\lambda_4 = \lambda_5 = 10$. The pre-defined margin for the triplet loss is set as $\xi = 0.8$. The model is optimized using Adam [5] with a learning rate of 0.0002 and the momentum terms $\beta_1 = 0.5$, $\beta_2 = 0.999$.

4.3. Comparison with the state-of-the-art methods

To demonstrate the effectiveness of our method, we compare our method with most of the related methods for IV-REID. These methods include Zero-Padding [21], TONE [22], HCML [22], BDTR [23] and *cm*GAN [2]. In addition, several other learning-based methods are also included for comparisons. The additional competing methods contain some feature learning methods including HOG [3], LOMO [8], one-stream and two-stream networks [21]. The one-stream and two-stream networks are modifications of the IDE method [26] under IV-REID settings. Their detailed descriptions can be found in [21]. In addition, two matching model learning methods, MLAPG [9] and GSM [10],

[1] UNIT code: `https://github.com/mingyuliutw/UNIT`

[2] Open-reid code: `https://github.com/Cysu/open-reid`

Table 1. Comparison with the state-of-the-art IV-REID methods on two different datasets, RegDB and SYSU-MM01.

Approach	**Constraints**		**RegDB**				**SYSU-MM01**			
	Feature-level	Image-level	CMC-1	CMC-10	CMC-20	mAP	CMC-1	CMC-10	CMC-20	mAP
LOMO [8]	✗	✗	0.85	2.47	4.10	2.28	1.75	14.14	26.63	3.48
MLBP [9]	✗	✗	2.02	7.33	10.90	6.77	2.12	16.23	28.32	3.86
HOG [3]	✗	✗	13.49	33.22	43.66	10.31	2.76	18.25	31.91	4.24
GSM [10]	✗	✗	17.28	34.47	45.26	15.06	5.29	33.71	52.95	8.00
One-stream [21]	✓	✗	13.11	32.98	42.51	14.02	12.04	49.68	66.74	13.67
Two-stream [21]	✓	✗	12.43	30.36	40.96	13.42	11.65	47.99	65.50	12.85
Zero-Padding [21]	✓	✗	17.75	34.21	44.35	18.90	14.80	54.12	71.33	15.95
TONE [22]	✓	✗	16.87	34.03	44.10	14.92	12.52	50.72	68.60	14.42
HCML [22]	✓	✗	24.44	47.53	56.78	20.80	14.32	53.16	69.17	16.16
BDTR [23]	✓	✗	33.47	58.42	67.52	31.83	17.01	55.43	71.96	19.66
*cm*GAN [2]	✓	✗	–	–	–	–	26.97	67.51	80.56	27.80
Proposed D^{2}RL	✓	✓	**43.4**	**66.1**	**76.3**	**44.1**	**28.9**	**70.6**	**82.4**	**29.2**

Table 2. Ablation study on the RegDB dataset.

Method	**Components**				**RegDB**	
	VAE	CC	CE	triplet	CMC-1 (%)	mAP (%)
Baseline	✓	✓	✗	✗	28.5	23.8
D^{2}RL (no VAE)	✗	✓	✓	✓	34.8	31.3
D^{2}RL (no CC)	✓	✗	✓	✓	33.7	29.9
D^{2}RL (no CE)	✓	✓	✗	✓	41.7	40.6
D^{2}RL (no triplet)	✓	✓	✓	✗	39.5	37.4
D^{2}RL	✓	✓	✓	✓	**43.4**	**44.1**

are also included for comparisons. Table 1 presents the results of all the methods. The methods specially designed for IV-REID generally perform much better than the ones that are not designed for IV-REID. Our method significantly outperforms the state-of-the-art IV-REID methods on both the RegDB and SYSU-MM01 datasets.

4.4. Ablation study

Our method consists of two sub-networks, the image-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{I}$ and the feature-level discrepancy reduction sub-network $\mathcal{T}_\mathcal{F}$, respectively taking GAN and ResNet-50 as their backbones. $\mathcal{T}_\mathcal{I}$ is mainly configured with the VAE and cycle-consistency (CC) losses while $\mathcal{T}_\mathcal{F}$ is optimized with the cross-entropy (CE) and triplet losses. For the ablation study, Table 2 reports the resultant CMC-1 and mAP values on the RegDB dataset by removing one loss at a time. Note that the baseline is obtained by only using $\mathcal{T}_\mathcal{F}$ without image-level modality unification.

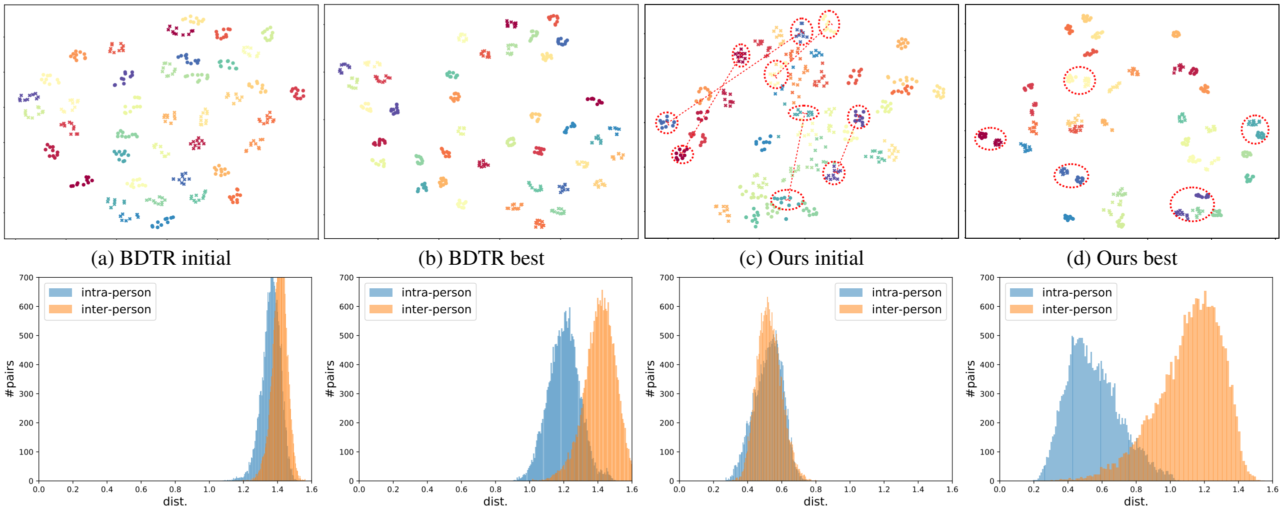
Note that the first two losses, VAE (for modality disentanglement) and cycle-consistency (for modality transfer), are responsible for the image-level modality unification. Removing either of them affects the image generation, thus degrading the performance more significantly. When removing both of them (the baseline), the performance drops dramatically to 28.5% in CMC-1 as it can only rely on feature embedding across modalities. The triplet loss is slightly more effective than the cross-entropy loss.

4.5. Discussions

Why reducing discrepancy separately? For IV-REID, previous methods try to reduce the appearance and modality discrepancies together from the feature-level view. Our method aims at reducing the appearance and modality discrepancies separately. We compare a feature-level method BDTR [23] with our proposed dual-level discrepancy reduction method D^{2}RL for investigating which strategy is more effective. We evaluate both BDTR and our method on the RegDB dataset. First, we plot the 1024-d person descriptor in the 2D feature space for visualization using the t-SNE method [14]. Testing samples of 20 persons were randomly selected from the RegDB dataset. Figure 4(a) and Figure 4(b) stand for visualizations of the initial and best results of the BDTR model respectively. We can observe that single-modality intra-person samples get closer to each other after training, but cross-modality intra-person samples relatively do not change too much. Figure 4(c) and Figure 4(d) are visualizations of the initial and best results of our proposed D^{2}RL network respectively. From these figures, we can find that not only the single-modality intra-person samples get closer to each other after training, but also some cross-modality intra-person samples move closer (as indicated by the red cycles in Figure 4(d)).

For further validating the effectiveness of reducing discrepancy separately, we conduct experiments to see how effective the embedded features are by again comparing BDTR and our method on the RegDB dataset. Figure 4(e) shows the initial distribution of distances of the inter-person and intra-person pairs for BDTR. Figure 4(f) shows the distribution after 4,800 training steps. Figure 4(g) and (h) show the distributions before and after training for our method. It is clear that, after training, our method can separate the inter-person and intra-person pairs more further apart than BDTR. It indicates that feature embedding in the unified space is more effective than the one in the image space.

Which modality to unify? We evaluate three options for modality unification, *i.e.*, the visible modality, the infrared



(a) BDTR initial (b) BDTR best (c) Ours initial (d) Ours best

(e) BDTR initial (f) BDTR best (g) Ours initial (h) Ours best

Figure 4. (Top) Visualization of the feature space. A total of 20 persons are randomly selected from the testing set. Samples with the same color indicate they are of the same person. The markers "dot" and "cross" denote images from the visible and infrared domain respectively. (a-b) are obtained using BDTR [23] on the RegDB datasetm which uses only feature-level constraints; (c-d) are obtained by our method. (Bottom) Histogram of the feature distances. (e-f) are obtained by BDTR; (g-h) are obtained by our method.

Table 3. Comparison of different modality unification options.

	RegDB		**SYSU-MM01**	
Metrics (%)	mAP	CMC-1	mAP	CMC-1
D^{2}RL(v)	36.4	39.1	28.4	28.1
D^{2}RL(i)	43.6	42.9	27.8	27.4
D^{2}RL	**44.1**	**43.4**	**29.2**	**28.9**

Table 4. Comparison between joint and separate training.

	RegDB		**SYSU-MM01**	
Metrics (%)	mAP	CMC-1	mAP	CMC-1
Separate	40.7	39.9	25.7	26.1
Joint	**44.1**	**43.4**	**29.2**	**28.9**

modality and the multi-spectral modality. We respectively use D^{2}RL(v), D^{2}RL(i) and D^{2}RL to denote these three options. Table 3 shows the results and there are several observations. First, unification to the multi-spectral modality, D^{2}RL, performs the best. Second, compared with other methods in Table 1, modality unification helps no matter which modality is chosen for unification. Finally, we find that D^{2}RL(i) performs better than D^{2}RL(v) on the RegDB dataset, while performing worse on the SYSU-MM01 dataset. We attribute this phenomenon to the setting for dataset evaluation. For the RegDB dataset, the gallery consists of infrared images, implying that the infrared modality plays an important role. The majority of original infrared images makes D^{2}RL(i) more effective on the dataset. As for the SYSU-MM01 dataset, the gallery consists of visible images, the results go the other way round. Our unification to the multi-spectral modality takes advantages of both domains and is thus more robust.

Why joint training? The whole framework consists of an image-level discrepancy reduction sub-network and a feature-level discrepancy reduction sub-network. They play different roles. They can be trained separately or jointly. Table 4 compares these two options. First, joint training provides significant performance boost as the two sub-networks benefits each other. Second, when comparing with other methods in Table 1, even with the separate training, our method outperforms the state-of-the-art methods.

How to balance sub-networks $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$? In the total loss of the proposed method defined in Equation (8), we use the weight γ to balance the contributions of $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$. As $\mathcal{T}_{\mathcal{I}}$ focuses on the modality discrepancy reduction and $\mathcal{T}_{\mathcal{F}}$ pays attention to the appearance discrepancy reduction, the larger γ is, the more contribution will attribute to the appearance reduction, in other words, the feature-level discrepancy reduction sub-network.

Figure 5 shows the results of the mAP and CMC-1 values on RegDB dataset by varying the weight γ. We can find that the re-identification accuracy varies when the weight γ changes, and there exists a suitable value to balance the contributions of $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$. Although $\mathcal{T}_{\mathcal{I}}$ alleviates the modality discrepancy, it could also brings in noisy information. Thus, the balance between $\mathcal{T}_{\mathcal{I}}$ and $\mathcal{T}_{\mathcal{F}}$ is important.

4.6. Visualization of results

The ability of modality unification. To demonstrate the effectiveness of our image-level discrepancy reduction sub-network $\mathcal{T}_{\mathcal{I}}$, we show some visual results of image translation in Figure 6. For each of the RegDB and SYSU-MM01 datasets, we show six groups of images. Each group has

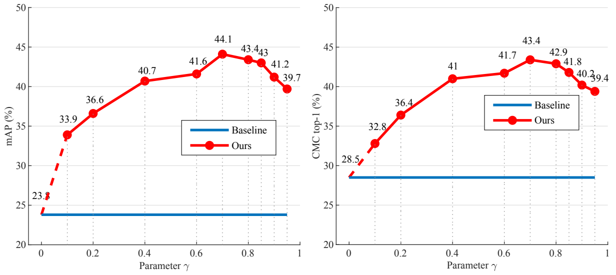
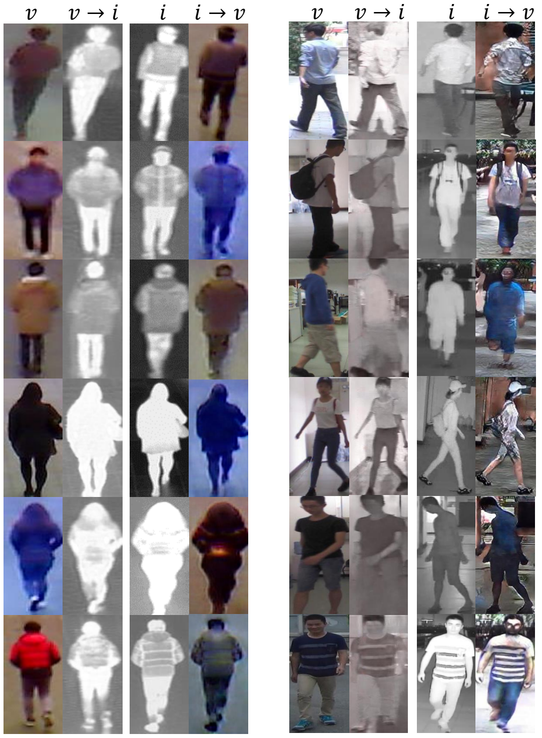


Figure 5. Curves of mAP (left) and CMC-1 (right) with respect to the hyper-parameter γ on RegDB dataset.



(a) RegDB

(b) SYSU-MM01

Figure 6. Examples of translated images generated by $\mathcal{T}_{\mathcal{I}}$ on (a) RegDB and (b) SYSU-MM01. For each dataset, from left to right, the four images of a row are the original visible image, the generated infrared image, the original infrared image, and the generated visible image respectively. The original visible and infrared images of the same row have the same identity.

four images: the original visible image, the generated infrared image, the original infrared image and the generated visible images. From the visual examples, we can observe that the sub-network $\mathcal{T}_{\mathcal{I}}$ is good at translating visible images to infrared ones, and the effectiveness of translating infrared images to visible ones is acceptable. However, some generated images could have color distortion, such as the sixth person of Figure 6(a). We can also find that the translation results of the RegDB dataset looks better than those of the SYSU-MM01 dataset. It is because the SYSU-MM01 dataset is more colorful, and the person images are not aligned well with different postures and scales. It may lead to difficulties for the image-level discrepancy reduction sub-network to be trained well. However, note that the final goal is not to generate images with good visual appearances, but to have good retrieval results. From the results in Table 1, the translated images do help a lot on IV-REID.

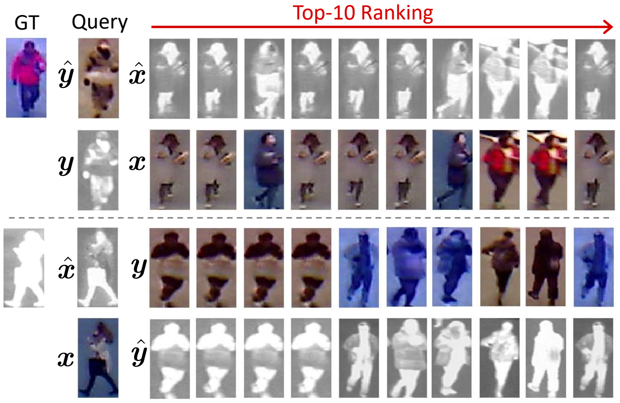


Figure 7. Visualization of failure cases. (Top) The query set is $[\hat{\boldsymbol{y}}; \boldsymbol{y}]$ and the gallery set is $[\boldsymbol{x}; \hat{\boldsymbol{x}}]$; (Bottom) The query set is $[\boldsymbol{x}; \hat{\boldsymbol{x}}]$ and the gallery set is $[\hat{\boldsymbol{y}}; \boldsymbol{y}]$.

Failure cases. We select the two worst query results (in which none of the top-10 results is correct) for illustrating the failure cases. For each query, the two rows respectively show the ranking list of generated images and the list of corresponding original images. It shows that, for some cases, the generated images could be bad and image-level modality unification can not work well on these queries.

5. Conclusions

In this paper, we present Dual-level Discrepancy Reduction Learning network (D^{2}RL) for the IV-REID task that exhibits both modality discrepancy and appearance discrepancy. Unlike previous IV-REID methods, instead of handling the mixed discrepancy with feature embedding, we propose to handle the discrepancies separately. We propose an image-level sub-network for modality unification, which generates a unified multi-spectral representation by image translation. With the unified representations, a feature-level sub-network can better reduce the appearance discrepancy by feature embedding. The proposed method shows significant improvement against the state-of-the-art methods.

Acknowledgements. This work was finished when Zhixiang Wang visited the Optical Sensing and Camera System Laboratory, National Institute of Informatics, Japan, through the NII International Internship Program. This work was supported by JST CREST (JPMJCR1686), Grant-in-Aid for JSPS Fellows (18F18378), NSFC (61801335), MOST (107-2221-E-002-147-MY3), and Microsoft Research Asia.

References

[1] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[2] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of International Joint Conferences on Artificial Intelligence*, 2018.

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2005.

[4] Weijian Deng, Liang Zheng, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[6] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[7] Xiang Li, Ancong Wu, and Wei-Shi Zheng. Adversarial open-world person re-identification. In *Proceedings of European Conference on Computer Vision*, 2018.

[8] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2015.

[9] Shengcai Liao and Stan Z. Li. Efficient PSD constrained asymmetric metric learning for person re-identification. In *Proceedings of International Conference on Computer Vision*, 2015.

[10] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1089–1102, 2017.

[11] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proceedings of Neural Information Processing Systems Conference*, 2017.

[13] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

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

[16] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q. Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Zheng Wang, Ruimin Hu, Chen Chen, Yi Yu, Junjun Jiang, Chao Liang, and Shin'ichi Satoh. Person reidentification via discrepancy matrix and matrix metric. *IEEE Transactions on Cybernetics*, 2018.

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

[19] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In *Proceedings of International Joint Conferences on Artificial Intelligence*, 2018.

[20] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[21] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-Infrared cross-modality person re-identification. In *Proceedings of International Conference on Computer Vision*, 2017.

[22] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2018.

[23] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C. Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *Proceedings of International Joint Conferences on Artificial Intelligence*, 2018.

[24] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv:1711.08184*, 2017.

[25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of International Conference on Computer Vision*, 2015.

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

[27] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of European Conference on Computer Vision*, 2018.

[28] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2018.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision*, 2017.