Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification

Guan-An Wang,[12] Tianzhu Zhang,[4] Yang Yang,[1] Jian Cheng,[123]
Jianlong Chang,[12] Xu Liang,[12] and Zengguang Hou[123]*
[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[4] University of Science and Technology of China, Beijing, China
{wangguanan2015,liangxu2013,zengguang.hou}@ia.ac.cn, tzzhang@ustc.edu.cn,
{yang.yang,jcheng,jianlong.chang}@nlpr.ia.ac.cn

Abstract

RGB-Infrared (IR) person re-identification is very challenging due to the large cross-modality variations between RGB and IR images. The key solution is to learn aligned features to the bridge RGB and IR modalities. However, due to the lack of correspondence labels between every pair of RGB and IR images, most methods try to alleviate the variations with set-level alignment by reducing the distance between the entire RGB and IR sets. However, this set-level alignment may lead to misalignment of some instances, which limits the performance for RGB-IR Re-ID. Different from existing methods, in this paper, we propose to generate cross-modality paired-images and perform both global set-level and fine-grained instance-level alignments. Our proposed method enjoys several merits. First, our method can perform set-level alignment by disentangling modality-specific and modality-invariant features. Compared with conventional methods, ours can explicitly remove the modality-specific features and the modality variation can be better reduced. Second, given cross-modality unpaired-images of a person, our method can generate cross-modality paired images from exchanged images. With them, we can directly perform instance-level alignment by minimizing distances of every pair of images. Extensive experimental results on two standard benchmarks demonstrate that the proposed model favourably against state-of-the-art methods. Especially, on SYSU-MM01 dataset, our model can achieve a gain of 9.2% and 7.7% in terms of Rank-1 and mAP. Code is available at https://github.com/wangguanan/JSIA-ReID.

Introduction

Person Re-Identification (Re-ID) (Gong et al. 2014; Zheng, Yang, and Hauptmann 2016) is widely used in various applications such as video surveillance, security and smart city. Given a query image of a person, Re-ID aims to find images of the person across disjoint cameras. It's very challenging due to the large intra-class and small inter-class variations caused by different poses, illuminations, views, and occlusions. Most of existing Re-ID methods focus on visible cameras and RGB images, and formulate the person Re-ID as a single-modality (RGB-RGB) matching problem.

*corresponding author
Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

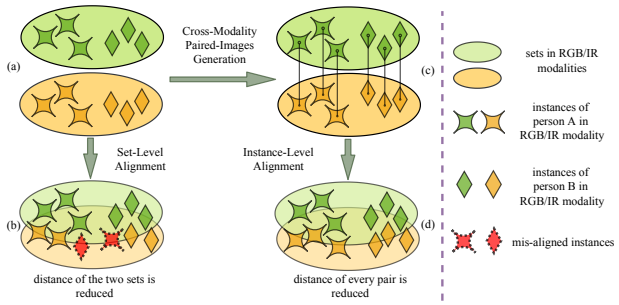

Figure 1: Illustration of set-level and instance-level alignment (please view in color). (a) There is a significant gap between the RGB and IR sets. (b) Existing methods perform set-level alignment by minimizing distances between the two sets, which may lead to misalignment of some instances. (c) Our method first generates cross-modality paired-images. (d) Then, instance-level alignment is performed by minimizing distances between each pair of images.

However, the visible cameras are difficult in capturing valid appearance information under poor illumination environments (e.g. at night), which limits the applicability of person Re-ID in practical. Fortunately, most surveillance cameras can automatically switch from visible (RGB) to near-infrared (IR) mode, which facilitates such cameras to work at night. Thus, it is necessary to study the RGB-IR Re-ID in real-world scenarios, which is a cross-modality matching problem. Compared with RGB-RGB single-modality matching, RGB-IR cross-modality matching is more difficult due to the large variation between the two modalities. As shown in Figure 2(b), RGB and IR images are intrinsically distinct and heterogeneous, and have different wavelength ranges. Here, RGB images have three channels containing color information of visible light, while IR images have one channel containing information of invisible light.

The key solution is to learn aligned features to bridge the two modalities. However, due to the lack of correspondence labels between every pair of images in different modalities like in Figure 2(a), existing RGB-IR Re-ID methods

Figure 2: (a) In the edge-photo task, we can get cross-modality paired-images. By minimizing their distances in a feature space, we can easily reduce the cross-modality gap. (b) In RGB-IR Re-ID task, we have only unpaired-images. The appearance variation caused by the cross-modality gap makes the task more challenging. (c) Our method can well generate images paired with given ones, which help us to improve RGB-IR Re-ID. (d,e) Vanilla image translation models such as CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018) fail to deal with this issue.

(Wu et al. 2017; Ye et al. 2018a; 2018b; Dai et al. 2018; Hao et al. 2019) try to reduce the marginal distribution divergence between RGB and IR modalities, while cannot deal with their joint distributions. That is to say, as shown in Figure 1(b), they only focus on the global set-level alignment between the entire RGB and IR sets while neglecting the fine-grained instance-level alignment between every two images. This may lead to misalignment of some instances when performing the global alignment (Chen et al. 2018). Although we can alleviate this issue by using label information, in Re-ID task, labels of training and test sets are unshared. Thus, simply fitting training labels may not perform very well for unseen test labels.

Different from the existing approaches, a heuristic method is to use cross-modality paired-images in Figure 2(a). With the paired images, we can directly reduce the instance-level gap by minimizing the distance between every pair of images in a feature space. However, as in Figure 2(b),all images are un-paired in RGB-IR Re-ID task. This is because the two kinds of images are captured at different times. RGB images are captured at daytime while IR ones at night. We can also translate images from one modality to another by using image translation models, such as CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018). But these image translation models can only learn one-to-one mappings, while mapping from IR to RGB images are one-to-many. For example, gray in IR mode can be blue, yellow even red in RGB mode. Under this situation, CycleGAN and StarGAN often generate some noisy images and cannot be used in the following Re-ID task. As shown in Figure 2(d,e), the generated images by CycleGAN and StarGAN are unsatisfying.

To solve the above problems, in this paper, we propose a novel Joint Set-level and Instance-Level Alignment Re-ID (JSIA-ReID) which enjoys several merits. First, our method can perform set-level alignment by disentangling modality-specific and modality-invariant features. Compared with encoding images with only one encoder, ours can explicitly remove the modality-specific features and significantly reduce the modality-gap. Second, given cross-modality unpaired-images of a person, our method can generate cross-modality paired-images. With them, we can directly perform instance-level alignment by minimizing the distances between the two images in a feature space. The instance-level alignment can further reduce the modality-gap and avoid misalignment of instances.

Specifically, as shown in Figure 3, our proposed method consists of a generation module $\mathcal{G}$ to generate cross-modality paired-images and a feature alignment module $\mathcal{F}$ to learn both set-level and instance-level aligned features. The generation module $\mathcal{G}$ includes three encoders and two generators. The three encoders disentangle a RGB(IR) image to modality-invariant and RGB(IR) modalities-specific features. Then, the RGB(IR) decoder takes a modality-invariant feature from an IR(RGB) image and a modality-specific feature from an IR(RGB) image as input. By decoding from the across-feature, we can generate cross-modality paired-images as in Figure 2(c). In the feature alignment module $\mathcal{F}$, we first utilize an encoder whose weights are shared with modality-invariant encoder. It can map images from different modalities into a shared feature space. Thus, set-level modality-gap can be significantly reduced. Then, we further import an encoder to refine the features to reduce the instance-level modality-gap by minimizing distance between feature maps of every pair of cross-modality images. Finally, by jointly training the generation module $\mathcal{G}$ and feature alignment module $\mathcal{F}$ with the re-id loss, we can learn both modality-aligned and identity-discriminative features.

The major contributions of this work can be summarized as follows. (1) We propose a novel method to generate cross-modality paired-images by disentangling features and decoding from exchanged features. To the best of our knowledge, it is the first work to generate cross-modality paired-images for the RGB-IR Re-ID task. (2) Our method can simultaneously and effectively reduce both set-level and instance-level modality-variation. (3) Extensive experimental results on two standard benchmarks demonstrate that the proposed model performs favourably against state-of-the-art methods.

Related Works

RGB-RGB Person Re-Identification. RGB-RGB person re-identification addresses the problem of matching pedestrian RGB images across disjoint visible cameras (Gong et al. 2014). Recently, many deep ReID methods (Zheng, Yang, and Hauptmann 2016; Hermans, Beyer, and Leibe 2017; Wang et al. 2019a) have been proposed. Zheng *et al.* (Zheng, Yang, and Hauptmann 2016) learn identity-discriminative features by fine-tuning a pre-trained CNN to minimize a classification loss. In (Hermans, Beyer, and Leibe 2017), Hermans *et al.* show that using a variant of the triplet loss

outperforms most other published methods by a large margin. Most of exiting methods focus on the RGB-RGB Re-ID task, and cannot perform well for the RGB-IR Re-ID task, which limits the applicability in practical surveillance scenarios.

RGB-IR Person Re-Identification. RGB-IR Person re-identification attempts to match RGB and IR images of a person under disjoint cameras. Besides the difficulties of RGB-RGB Re-ID, RGB-IR Re-ID faces a new challenge due to cross-modality variation between RGB and IR images. In (Wu et al. 2017), Wu *et al.* collect a cross-modality RGB-IR dataset named SYSU RGB-IR Re-ID and explores three different network structures with zero-padding for automatically evolve domain-specific nodes in the network. Ye *et al.* utilize a dual-path network with a bi-directional dual-constrained top-ranking loss (Ye et al. 2018a) and modality-specific and modality-shared metrics (Ye et al. 2018b). In (Dai et al. 2018), Dai *et al.* introduce a cross-modality generative adversarial network (cmGAN) to reduce the distribution divergence of RGB and IR features. Hao *et al.* (Hao et al. 2019) achieve visible thermal person re-identification via a hyper-sphere manifold embedding model. In (Wang et al. 2019b) and (Wang et al. 2019c), they reduce modality-gap in both image and feature domains. Most above methods mainly focus on global set-level alignment between the entire RGB and IR sets, which may lead to misalignment of some instances. Different from them, our proposed method performs both global set-level and fine-grained instance-level alignment, and achieves better performance.

Person Re-Identification with GAN. Recently, many methods attempt to utilize GAN to generate training samples for improving Re-ID. Zheng *et al.* (Zheng, Zheng, and Yang 2017) use a GAN model to generate unlabeled images as data augmentation. Zhong *et al.* (Zhong et al. 2018b; 2018a; 2019) translate images to different camera styles with CycleGAN (Zhu et al. 2017), and then use both real and generated images to reduce inter-camera variation. Ma *et al.* (Ma et al. 2018) use a cGAN to generate pedestrian images with different poses to learn features free of influences of pose variation. Zheng *et al* (Zheng et al. 2019) propose joint learning framework that end-to-end couples re-id learning and image generation in a unified network. All those methods focus on single-modality RGB Re-ID and cannot deal with cross-modality RGB-IR Re-ID. Different from them, our method can generate cross-modality paired-images and learn both set-level and instance-level aligned features.

Image Translation. Generative Adversarial Network (GAN) (Goodfellow et al. 2014) learns data distribution in a self-supervised way via the adversarial training, which has been widely used in image translation. Pix2Pix (Isola et al. 2017) solves the image translation by utilizing a conditional generative adversarial network and a reconstruction loss supervised by paired data. CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018) learn images translations with unpaired data using cycle-consistency loss. Those methods only learn one-to-one mapping among different modalities and cannot be used in RGB-IR Re-ID, where the mapping from IR to RGB is one-to-many. Different from them, our method first disentangles images to modality-invariant and modality-specific features, and then generates cross-modality paired-images by decoding from exchanged features.

The Proposed Method

Our method includes a generation module $\mathcal{G}$ to generate cross-modality paired-images and a feature alignment module $\mathcal{F}$ to learn both global set-level and fine-grained instance-level aligned features. Finally, by training the two modules with re-id loss, we can learn both modality-aligned and identity-discriminative features.

Cross-Modality Paired-Images Generation Module

As shown in Figure 2(b), in RGB-IR task, the training images from two modalities are unpaired, which makes it more difficult to reduce the gap between the RGB and IR modalities. To solve the problem, we propose to generate paired-images by disentangling features and decoding from exchanged features. We suppose that images can be decomposed to modality-invariant and modality-specific features. Here, the former includes content information such as pose, gender, clothing category and carrying, *etc*. Oppositely, the latter has style information such as clothing/shoes colors, texture, *etc*. Thus, given unpaired-images, by disentangling and exchanging their style information, we can generate paired-images, where the two images have the same content information such as pose and view but with different style information such as clothing colors.

Features Disentanglement. We disentangle features with three encoders. The three encoders are the modality-invariant encoder E^i of learning content information from both modalities, the RGB modality-specific encoder E^s_{rgb} of learning RGB style information, and the IR modality-specific encoder E^s_{ir} of learning IR style information. Given RGB images X_{rgb} and IR images X_{ir}, their modality-specific features M^s_{rgb} and M^s_{ir} can be learned in Eq.(2). Similarly, their modality-invariant features M^i_{rgb} and M^i_{ir} can be learned in Eq.(1).

$$M^s_{rgb} = E^s_{rgb}(X_{rgb}),\ M^s_{ir} = E^s_{ir}(X_{ir}) \tag{1}$$

$$M^i_{rgb} = E^i(X_{rgb}),\ M^i_{ir} = E^i(X_{ir}) \tag{2}$$

Paired-Images Generation. We generate paired-images using two decoders including a RGB decoder D_{rgb} of generating RGB images and an IR decoder D_{ir} of generating IR images. After getting the disentangled features in Eq.(1) and Eq.(2), we can generate paired-images by exchanging their style information. Specifically, to generate RGB images X_{ir2rgb} paired with real IR images X_{ir}, we can use the content features M^i_{ir} from the real IR images X_{ir} and the style features M^s_{rgb} from the real RGB images X_{rgb}. By doing so, the generated images will contain content information from the IR images and style information from the RGB image. Similarly, we can also generate fake IR images X_{rgb2ir} paired with real RGB images X_{rgb}. Note that to ensure that the generated images have the same identities with their original ones, we only exchange features intra-person.

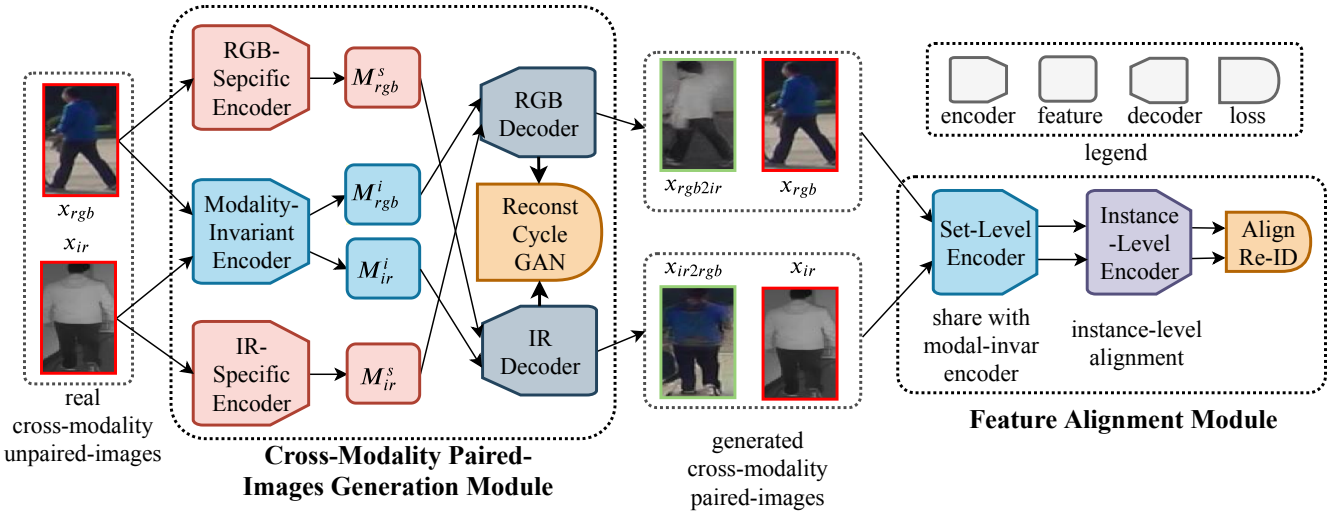

Figure 3: Our proposed framework consists of a cross-modality paired-images generation module $\mathcal{G}$ and a feature alignment module $\mathcal{F}$. $\mathcal{G}$ first disentangle images to modality-specific and modality-invariant features, and then decode from the exchanged features. $\mathcal{F}$ first use the modality-invariant encoder to perform set-level alignment, then further perform instance-level alignment by minimizing distance of each pair images. Finally, by training the two modules with re-id loss, we can learn both modality-aligned and identity-discriminative features.

This processes can be formulated in Eq.(3).

$$X_{ir2rgb} = D_{ir}(M^i_{ir}, M^s_{rgb}), X_{rgb2ir} = D_{rgb}(M^i_{rgb}, M^s_{ir}) \tag{3}$$

Reconstruction Loss. A simple supervision is to force the disentangled features to reconstruct their original images. Thus, we can formulate the reconstruction loss $\mathcal{L}_{recon}$ as below, where $||\cdot||_1$ is L1 distance.

$$\begin{aligned}\mathcal{L}_{recon} = &||X_{rgb} - D_{rgb}(E^i(X_{rgb}), E^s_{rgb}(X_{rgb}))||_1 \\ &+ ||X_{ir} - D_{ir}(E^i(X_{ir}), E^s_{ir}(X_{ir}))||_1\end{aligned} \tag{4}$$

Cycle-Consistency Loss. The reconstruction loss $\mathcal{L}_{recon}$ in Eq.(4) cannot supervise the cross-modality paired-images generation, and the generated images may not contain the expired content and style information. For example, when translating IR images X_{ir} to its RGB version X_{ir2rgb} via Eq(3), the translated images X_{ir2rgb} may not keep the poses (content information) from X_{ir}, or don't have the right clothing color (style information) with X_{rgb}. This is not the case we want and will harm the feature learning module. Inspired by CycleGAN (Zhu et al. 2017), we introduce a cycle-consistency loss to guarantee that the generated images can be translated back to their original version. By doing so, the consistency loss further limits the space of the generated samples. The cycle-consistency loss can be formulated as below:

$$\mathcal{L}_{cyc} = ||X_{rgb} - X_{rgb2ir2rgb}||_1 + ||X_{ir} - X_{ir2rgb2ir}||_1 \tag{5}$$

where $X_{ir2rgb2ir}$ and $X_{rgb2ir2rgb}$ are the cycle-reconstructed images as in Eq.(6).

$$\begin{aligned}X_{ir2rgb2ir} &= D_{ir}(E^i_{rgb}(X_{ir2rgb}), E^s_{ir}(X_{rgb2ir})) \\ X_{rgb2ir2rgb} &= D_{rgb}(E^i_{ir}(X_{rgb2ir}), E^s_{rgb}(X_{ir2rgb}))\end{aligned} \tag{6}$$

GAN loss. The reconstruction loss $\mathcal{L}_{recon}$ and cycle-consistency loss $\mathcal{L}_{cyc}$ lead to blurry images. To make the generated images more realistic, we apply the adversarial loss (Goodfellow et al. 2014) on both modalities, which have been proved to be effective in image generation tasks (Isola et al. 2017). Specifically, we import two discriminators Dis_{rgb} and Dis_{ir} to distinguish real images from the generated ones on RGB and IR modalities, respectively. In contrast, the encoders and decoders aim to make the generated images indistinguishable. The GAN loss can be formulated as below:

$$\begin{aligned}\mathcal{L}_{gan} = &E[logDis_{rgb}(X_{rgb}) + log(1 - Dis_{rgb}(X_{ir2rgb}))] \\ &+E[logDis_{ir}(X_{ir}) + log(1 - Dis_{ir}(X_{rgb2ir}))]\end{aligned} \tag{7}$$

Feature Alignment Module

Set-Level Feature Alignment. To reduce the modality-gap, most methods attempt to learn a shared feature-space for different modalities by using dual path (Ye et al. 2018a; 2018b), or GAN loss (Dai et al. 2018). However, those methods do not explicitly remove the modality-specific information, which may be encoded into the shared feature-space and harms the performance (Chang et al. 2019). In our method, we utilize a set-level encoder E^{sl} to learn set-level aligned features. The weights E^{sl} are shared with the modality-invariant encoder E^i. As we can see, in the cross-modality paired-images generation module, our modality-invariant encoder E^i is trained to explicitly remove modality-specific features. Thus, given images X from any modality, we can learn their set-level aligned features $M = E^{sl}(X)$.

Instance-Level Feature Alignment. Even so, as we discuss in the introduction, only performing global set-level alignment between the entire RGB and IR sets may lead to mis-

alignment of some instances. To overcome this problem, we propose to perform instance-level alignment by using the cross-modality paired-images generated by the generation module. Specifically, we first utilize instance-level encoder E^{il} to map the set-level aligned features M to a new feature space $\mathcal{T}$, *i.e.* $T = E^{il}(M)$. Then, based on the feature space $\mathcal{T}$, we align every two cross-modality paired-images by minimizing their Kullback-Leibler Divergence. Thus, the loss of the instance-level feature alignment can be formulated in Eq.(8).

$$\begin{aligned}\mathcal{L}_{align} = & E_{(x_1,x_2)\in(X_{ir},X_{ir2rgb})}[KL(p_1||p_2)] \\ & + E_{(x_1,x_2)\in(X_{rgb2ir},X_{rgb})}[KL(p_1||p_2)]\end{aligned} \tag{8}$$

where $p_1 = C(t_1)$ and $p_2 = C(t_2)$ are the predicted probabilities of x_1 and x_2 on all identities, t_1 and t_2 are the features of x_1 and x_2 in the feature space $\mathcal{T}$, C is a classifier implemented with a fully-connected layer.

Identity-Discriminative Feature Learning. To overcome the intra-modality variation, following (Zheng, Yang, and Hauptmann 2016; Hermans, Beyer, and Leibe 2017), we averagely pool the feature maps T in instance-level aligned space $\mathcal{T}$ to corresponding feature vectors V. Given real images X, we optimize their feature vectors V with a classification loss $\mathcal{L}_{cls}$ of a classifier C and a triplet loss $\mathcal{L}_{triplet}$.

$$\mathcal{L}_{cls} = E_{v\in V}(-log\ p(v)) \tag{9}$$

$$\mathcal{L}_{triplet} = E_{v\in V}[m - D_{v_a,v_p} + D_{v_a,v_n}]_+ \tag{10}$$

where $p(\cdot)$ is the predicted probability predicted by the classifier C that the input feature vector belongs to the ground-truth, v_a and v_p are a positive pair of feature vectors belonging to the same person, v_a and v_n are a negative pair of feature vectors belonging to different persons, m is a margin parameter and $[x]_+ = max(0, x)$.

Overall Objective Function and Test

Thus, the overall objective function of our method can formulated as below:

$$\begin{aligned}\mathcal{L} = & \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{gan}\mathcal{L}_{gan} \\ & + \lambda_{align}\mathcal{L}_{align} + \lambda_{reid}(\mathcal{L}_{cls} + \mathcal{L}_{triplet})\end{aligned} \tag{11}$$

where λ_* are weights of corresponding terms. Following (Zhu et al. 2017), we set $\lambda_{cyc} = 10$ and $\lambda_{gan} = 1$. λ_{reid} is set 1 empirically and λ_{align} is decided by grid search.

During the test stage, only feature learning module $\mathcal{F}$ is used. Given images X, we use the set-level alignment encoder E^{sl} and the instance-level encoder E^{il} to extract features, *i.e.* $V = E^{il}((E^{sl}(X))$. Finally, matching is conducted by computing cosine similarities of feature vectors V between the probe images and gallery ones.

Experiment

Dataset and Evaluation Protocol

Dataset. We evaluate our model on two standard benchmarks including SYSU-MM01 and RegDB. (1) SYSU-MM01 (Wu et al. 2017) is a popular RGB-IR Re-ID dataset, which includes 491 identities from 4 RGB cameras and 2 IR ones. The training set contains 19,659 RGB images and 12,792 IR images of 395 persons and the test set contains 96 persons. Following (Wu et al. 2017), there are two test modes, *i.e. all-search* mode and *indoor-search* mode. For the *all-search* mode, all images are used. For the *indoor-search* mode, only indoor images from $1st, 2nd, 3rd, 6th$ cameras are used. For both modes, the *single-shot* and *multi-shot* settings are adopted, where 1 or 10 images of a person are randomly selected to form the gallery set. Both modes use IR images as probe set and RGB images as gallery set. (2) RegDB (Nguyen et al. 2017) contains 412 persons, where each person has 10 images from a visible camera and 10 images from a thermal camera.

Evaluation Protocols. The Cumulative Matching Characteristic (CMC) and mean average precision (mAP) are used as evaluation metrics. Following (Wu et al. 2017), the results of SYSU-MM01 are evaluated with official code based on the average of 10 times repeated random split of gallery and probe set Following (Ye et al. 2018a; 2018b), the results of RegDB are based on the average of 10 times repeated random split of training and testing sets.

Implementation Details

In generation module $\mathcal{G}$, following (Radford, Metz, and Chintala 2016), we construct our modality-specific encoders with 2 strided convolutional layers followed by a global average pooling layer and a fully connected layer. For decoders, following (Wang et al. 2017), we use 4 residual blocks with Adaptive Instance Normalization (AdaIN) and 2 upsampling with convolutional layers. Here, the parameters of AdaIN are dynamically generated by the modality-specific features. In GAN loss, we use discriminator and LS-GAN as in (Mao et al. 2016) to stable the training.

In feature learning module $\mathcal{F}$, for a fair comparison, we adopt the ResNet-50 (He et al. 2016) pre-trained with ImageNet (Russakovsky et al. 2015) as our CNN backbone. Specifically, we use the first two layers of the ResNet-50 as our set-level encoder E^{sl}, and use the remaining layers as our instance-level encoder E^{il}. For the classification loss, the classifier C takes the feature vectors V as inputs, followed by a batch normalization, a fully-connected layer and a soft-max layer to predict the inputs' labels.

We implement our model with open-source deep learning framework Pytorch The training images are resized to 256×128 and augmented with horizontal flip. The batch size is set to 128 (16 person, 4 RGB images and 4 IR images). We optimize our framework using Adam with learning rate 0.0002 and betas $[0.5, 0.999]$. The generation module is first pre-trained for 100 epochs. Then the overall framework is jointly optimized for 50 epochs, where the learning rate is decayed to its 0.1 at 30 epochs.

Results on SYSU-MM01 Datasets

We compare our model with 10 methods including handcrafted features (HOG (Dalal and Triggs 2005), LOMO (Liao et al. 2015)), feature learning with the classification loss (One-Stream, Two-Stream, Zero-Padding) (Wu et al. 2017), feature learning with both classification and ranking losses (BCTR, BDTR) (Ye et al. 2018a), metric learning (D-HSME (Hao et al. 2019)), and reducing distribution diver-

Table 1: Comparison with the state-of-the-arts on SYSU-MM01 dataset. The R1, R10, R20 denote Rank-1, Rank-10 and Rank-20 accuracies (%), respectively. The mAP denotes mean average precision score (%).

Methods	*All-Search*								*Indoor-Search*							
	Single-Shot				*Multi-Shot*				*Single-Shot*				*Multi-Shot*			
	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG	2.76	18.3	32.0	4.24	3.82	22.8	37.7	2.16	3.22	24.7	44.6	7.25	4.75	29.1	49.4	3.51
LOMO	3.64	23.2	37.3	4.53	4.70	28.3	43.1	2.28	5.75	34.4	54.9	10.2	7.36	40.4	60.4	5.64
Two-Stream	11.7	48.0	65.5	12.9	16.4	58.4	74.5	8.03	15.6	61.2	81.1	21.5	22.5	72.3	88.7	14.0
One-Stream	12.1	49.7	66.8	13.7	16.3	58.2	75.1	8.59	17.0	63.6	82.1	23.0	22.7	71.8	87.9	15.1
Zero-Padding	14.8	52.2	71.4	16.0	19.2	61.4	78.5	10.9	20.6	68.4	85.8	27.0	24.5	75.9	91.4	18.7
BCTR	16.2	54.9	71.5	19.2	-	-	-	-	-	-	-	-	-	-	-	-
BDTR	17.1	55.5	72.0	19.7	-	-	-	-	-	-	-	-	-	-	-	-
D-HSME	20.7	62.8	78.0	23.2	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.7	77.2	89.2	42.2	37.0	80.9	92.3	32.8
D^{2}RL	28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
Ours	**38.1**	**80.7**	**89.9**	**36.9**	**45.1**	**85.7**	**93.8**	**29.5**	**43.8**	**86.2**	**94.2**	**52.9**	**52.7**	**91.1**	**96.4**	**42.7**

Table 2: Comparison with state-of-the-arts on the RegDB dataset under different query settings.

Methods	thermal2visible		visible2thermal	
	Rank-1	mAP	Rank-1	mAP
Zero-Padding	16.7	17.9	17.8	31.9
TONE	21.7	22.3	24.4	20.1
BCTR	-	-	32.7	31.0
BDTR	32.8	31.2	33.5	31.9
D^{2}RL	43.4	44.1	43.4	44.1
Ours	48.1	48.9	48.5	49.3

Table 3: Analysis of set-level (SL) and instance-level (IL) alignment. Please see text for more details.

index	SL	IL	R1	R10	R20	mAP
1	×	×	32.1	75.7	87.0	31.9
2	✓	×	35.1	78.6	88.2	33.8
3	×	✓	36.0	79.8	89.0	35.5
4	✓	✓	38.1	80.7	89.9	36.9
5	-	✓	36.8	80.2	89.4	36.0

gence of features (cmGAN (Dai et al. 2018), D^{2}RL (Wang et al. 2019c)). The experimental results are shown in Table 1.

Firstly, LOMO only achieves 3.64% and 4.53% in terms of Rank-1 and mAP scores, respectively, which shows that hand-crafted features cannot be generalized to the RGB-IR Re-ID task. Secondly, One-Stream, Two-Stream and Zero-Padding significantly outperform hand-crafted features by at least 8% and 8.3% in terms of Rank-1 and mAP scores, respectively. This verifies that the classification loss contributes to learning identity-discriminative features. Thirdly, BCTR and BDTR further improve Zero-Padding by 1.4% in terms of Rank-1 and by 3.2% in terms of mAP scores. This shows that the ranking and classification losses are complementary. Additionally, D-HSME outperforms BDTR by 3.6% Rank-1 and 3.5% mAP scores, which demonstrates the effectiveness of metric learning. In addition, D^{2}RL outperform D-HSME by 8.1% Rank1 and 6.0% mAP scores, implying the effectiveness of adversarial training. Finally, Our method outperforms the state-of-the-art method by 9.2% and 7.7% in terms of Rank-1 and mAP scores, showing the effectiveness of our model for the RGB-IR Re-ID task.

Results on RegDB Dataset

We evaluate our model on RegDB dataset and compare it with Zero-Padding (Wu et al. 2017), TONE (Ye et al. 2018b), BCTR (Ye et al. 2018a), BDTR (Ye et al. 2018b) and D^{2}RL (Wang et al. 2019c). We adopt visible2thermal and thermal2visible modes. Here, the visible2thermal means that visible images are query set and thermal images are gallery set, and so on. As shown in Table 2, our model can significantly outperform the state-of-the-arts by 4.7% and 5.1% in terms of Rank-1 scores with thermal2visible and visible2thermal modes, respectively. Overall, the results verify the effectiveness of our model.

Model Analysis

Ablation Study. To further analyze effectiveness of the set-level alignment and the instance-level alignment, we evaluate our method under four different settings, *i.e.* with or without set-level (SL) and instance-level (IL) alignment. Specifically, when removing set-level alignment, we use separate set-level encoder E^{sl}, *i.e.* we don't share weights of set-level encoder E^{sl} with modality-invariant encoder E^{i}. When removing instance-level alignment, we set $\lambda_{align} = 0$. Moreover, to analyze whether the feature disentanglement contributes to set-level alignment, we remove the disentanglement strategy by using separate set-level encoder E^{sl} and training it with a GAN loss as in (Dai et al. 2018).

As shown in Table 3, when removing both SL and IL (index-1), our method only achieve 32.1% Rank-1 score. By adding SL (index-2) or IL (index-3), the performance is improved to 35.1% and 36.0% Rank-1 score, which demonstrate the effectiveness of both SL and IL. When using both SL and IL (index-4), our method achieves the best performance at 38.1% Rank-1 score, which demonstrates that SL and IL can be complementary with each other. Finally, when removing the disentanglement from set-level align-

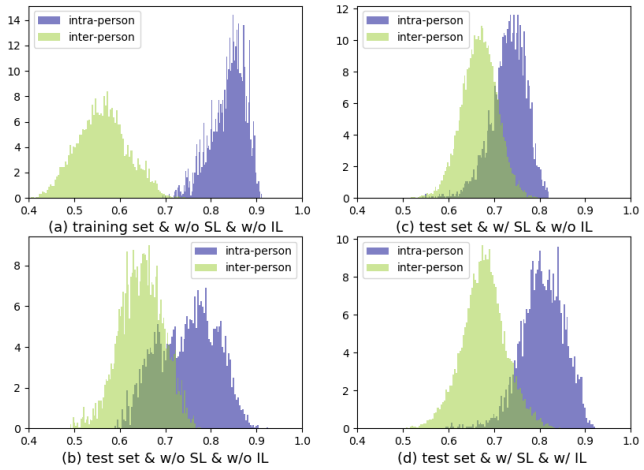

Figure 4: Distribution of cross-modality similarities of intra-person and inter-person. The instance-level alignment (IL) can enhance intra-person similarity while keep inter-person similarity unchanged, which improves performance. Please note that w/ means with and w/o means without. Please see text for more details.

ment (index-5), Rank-1 score drops by 1.3%. This illustrates that disentanglement strategy is helpful for learning set-level alignment.

To better understand set-level alignment (SL) and instance-level alignment (IL), we visualize the distribution of intra-person similarity and inter-person similarity under different variants. The similarity is calculated with cosine distance. Firstly, when comparing with Figure 4(a) and Figure 4(b), we can find that even using no SL and IL, model can easily fit training set, while fails to generalize to test set. As we can see in Figure 4(b), the two kind of similarities are seriously overlapped. This shows that the cross-modality variation cannot be well reduced by simply fitting identity information in training set. Secondly, in Figure 4(c), we find that although the similarity of intra-person becomes more concentrated, the similarity of inter-person also become larger. This shows that SL imports some misalignment of instances which may harm the performance. Finally, in Figure 4(c) we can see that, IL boosts intra-person similarity, meanwhile keeps the inter-person similarity unchanged. This illustrate that the IL explicitly reduce . In summary, experimental results and analysis above show the importance and effectiveness of instance-level alignment.

Parameters Analysis. We evaluate the effect of the weights, *i.e.* λ_{align}. As shown in Figure 5, we analyze our method with respect to the λ_{align} on SYSU-MM01 dataset under *single-shot&all-search* mode. We can see that, with different λ_{align}, our method can stably have an significant improvement. The experimental results show that our method is robust to different weights.

Visualization of Images

In this part, we display the generated cross-modality paired-images from ours, CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018). From Figure 6(a), we can see that, images of a person in the two modalities are significant different, even human beings cannot easily identify them. In Figure 6(b), our method can stably generate fake images when given cross-modality unpaired-images from a person. For example, in person A, ours can translate her IR images to RGB version with right colors (yellow upper and black bottom clothes). However, in Figure 6(c) and Figure 6(d), CycleGAN and StarGAN cannot learn the right colors even poses. For example, person B should have blue upper clothing. However, images generated by CycleGAN and StarGAN are red and black, respectively. Those unsatisfying images cannot be used to learn instance-level aligned features.

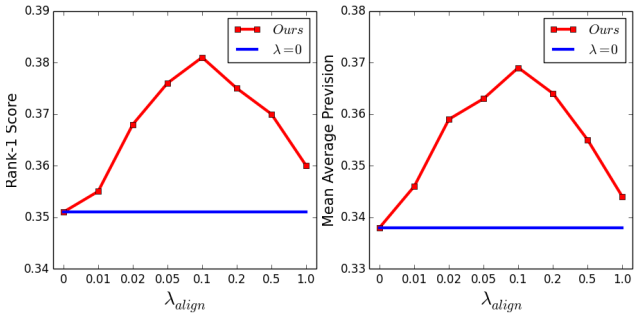

Figure 5: Rank-1 and mAP scores with different λ_{align} on SYSU-MM01 under *single-shot&all-search* mode.

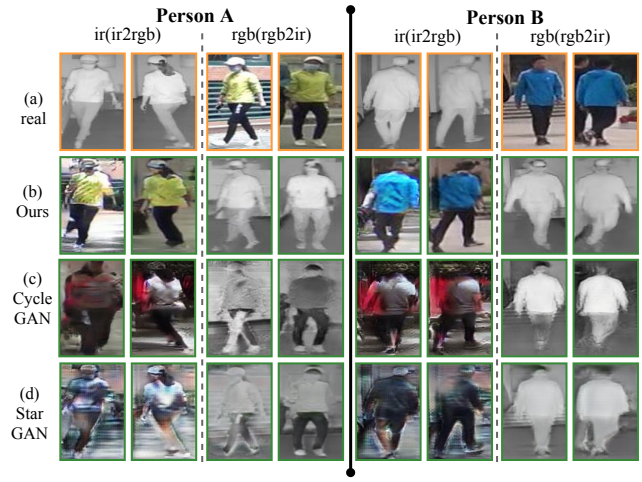

Figure 6: Comparision among generated images from ours, CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018). Ours can stably generate paired-images with given real ones, while CycleGAN and StarGAN fail.

Conclusion

In this paper, we propose a novel Joint Set-Level and Instance-Level Alignment Re-ID (JSIA-ReID). On the one hand, our model performs set-level alignment by disentangling modality-specific and modality-invariant features. Compared with vanilla methods, ours can explicitly remove

the modality-specific information and significantly reduce the modality-gap. On the other hand, given cross-modality unpaired images, we we can generate cross-modality paired-images by exchanging their features. With the paired-images, instance-level variations can be reduced by minimizing the distances between every pair of images. Finally, together with re-id loss, our model can learn both modality-aligned and identity-discriminative features. Experimental results on two datasets show the effectiveness of our proposed method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61720106012, 61533016 and 61806203, the Strategic Priority Research Program of Chinese Academy of Science under Grant XDBS01000000, and the Beijing Natural Science Foundation under Grant L172050.

References

Chang, W.-L.; Wang, H.-P.; Peng, W.-H.; and Chiu, W.-C. 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1900–1909.

Chen, Q.; Liu, Y.; Wang, Z.; Wassell, I. J.; and Chetty, K. 2018. Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7976–7985.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *IJCAI 2018: 27th International Joint Conference on Artificial Intelligence*, 677–683.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, 886–893. IEEE Computer Society.

Gong, S.; Cristani, M.; Yan, S.; and Loy, C. C. 2014. *Person Re-Identification*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2672–2680.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. Hsme hypersphere manifold embedding for visible thermal person re-identification. In *AAAI-19 AAAI Conference on Artificial Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2197–2206.

Ma, L.; Sun, Q.; Georgoulis, S.; Gool, L. V.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 99–108.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2016. Least squares generative adversarial networks. *arXiv preprint arXiv:1611.04076*.

Nguyen, D. T.; Hong, H. G.; Kim, K.-W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17(3):605.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *international conference on learning representations*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Wang, H.; Liang, X.; Zhang, H.; Yeung, D.-Y.; and Xing, E. P. 2017. Zm-net: Real-time zero-shot image manipulation network. *arXiv preprint arXiv:1703.07255*.

Wang, G.; Yang, Y.; Cheng, J.; Wang, J.; and Hou, Z. 2019a. Color-sensitive person re-identification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 933–939.

Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; and Hou, Z. 2019b. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019c. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 618–626.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5390–5399.

Ye, M.; Lan, X.; Li, J.; and c Yuen, P. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI-18 AAAI Conference on Artificial Intelligence*, 7501–7508.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI 2018: 27th International Joint Conference on Artificial Intelligence*, 1092–1099.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2138–2147.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*.

Zhong, Z.; Zheng, L.; Li, S.; and Yang, Y. 2018a. Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 176–192.

Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; and Yang, Y. 2018b. Camera style adaptation for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5157–5166.

Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; and Yang, Y. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251.