Neural Networks 128 (2020) 294–304

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification

Guan'an Wang [a,b,1], Yang Yang [a,1], Tianzhu Zhang [d], Jian Cheng [a,b,c], Zengguang Hou [a,b,c,*], Prayag Tiwari [e], Hari Mohan Pandey [f]

[a] Institute of Automation, Chinese Academy of Sciences, No.95 Zhongguancun East Road, Beijing 100190, PR China
[b] University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, PR China
[c] Center for Excellence in Brain Science and Intelligence Technology, 320 Yue Yang Road, Shanghai, 200031, PR China
[d] University of Science and Technology of China, No.96 JinZhai Road, Hefei, Anhui, 230026, PR China
[e] Department of Information Engineering, University of Padova, Italy
[f] Department of Computer Science, Edge Hill University, Ormskirk, United Kingdom

ARTICLE INFO

Article history:
Received 27 December 2019
Received in revised form 19 April 2020
Accepted 7 May 2020
Available online 19 May 2020

Keywords:
Person re-identification
Cross-modality
Feature disentanglement
Image generation
Adversarial learning

ABSTRACT

RGB-Infrared (IR) person re-identification is very challenging due to the large cross-modality variations between RGB and IR images. Considering no correspondence labels between every pair of RGB and IR images, most methods try to alleviate the variations with set-level alignment by reducing marginal distribution divergence between the entire RGB and IR sets. However, this set-level alignment strategy may lead to misalignment of some instances, which limit 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 features. With them, we can directly perform instance-level alignment by minimizing distances of every pair of images. Third, our method learns a latent manifold space. In the space, we can random sample and generate lots of images of unseen classes. Training with those images, the learned identity feature space is more smooth can generalize better when test. Finally, extensive experimental results on two standard benchmarks demonstrate that the proposed model favorably against state-of-the-art methods.

1. Introduction

Person Re-Identification (Re-ID) (Gong, Cristani, Yan, & Loy, 2014; Zheng, Yang, & 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 is very challenging due to the large intra-class and small inter-class variations caused by different poses, illuminations, views, and occlusions. To tackle the above issue, lots of methods have been proposed, which can be grouped into hand-crafted descriptors (Liao, Hu, Zhu, & Li, 2015; Ma, Su, & Jurie, 2014; Yang, et al., 2014), metric learning (Koestinger, Hirzer, Wohlhart, Roth, & Bischof, 2012; Liao & Li, 2015; Zheng, Gong, & Xiang, 2013), and deep learning (Hermans, Beyer, & Leibe, 2017; Schmidhuber, 2015; Sun, Zheng, Yang, Tian, & Wang, 2018; Tavanaei, Ghodrati, Kheradpisheh, Masquelier, & Maida, 2018; Zheng et al., 2016). 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.

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.

* Corresponding author.
E-mail addresses: wangguanan2015@ia.ac.cn (G. Wang), yang.yang@nlpr.ia.ac.cn (Y. Yang), tzzhang@ustc.edu.cn (T. Zhang), jcheng@nlpr.ia.ac.cn (J. Cheng), zengguang.hou@ia.ac.cn (Z. Hou), prayag.tiwari@dei.unipd.it (P. Tiwari), pandeyh@edgehill.ac.uk (H.M. Pandey).
[1] Equal Contribution.

https://doi.org/10.1016/j.neunet.2020.05.008
0893-6080/

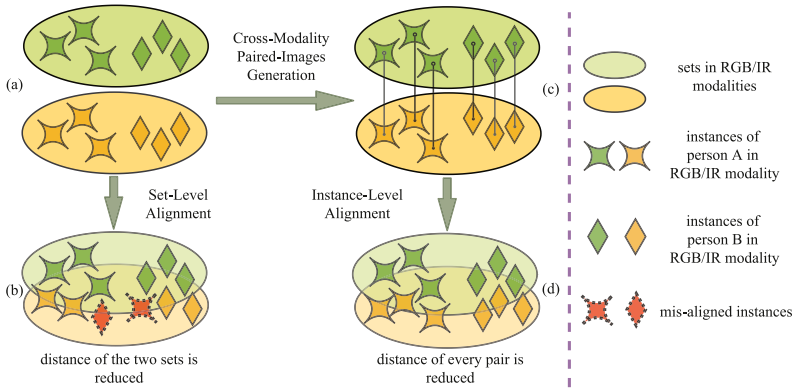

Fig. 1. Illustration of set-level and instance-level alignment (please view in color). (a) There is a significant gap between the RGB and IR sets. (b) Existing methods perform set-level alignment by minimizing distances between the two sets, which may lead to misalignment of some instances. (c) Our method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

As shown in Fig. 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. As a result, even human can hardly recognize the person cross the two modalities.

The main challenge is modality-gap between RGB and IR images, making intra-class distance large. However, due to the lack of correspondence labels between every pair of images in different modalities like in Fig. 2(a), existing RGB–IR Re-ID methods (Dai, Ji, Wang, Wu, & Huang, 2018; Hao, Wang, Li, & Gao, 2019; Wu, Zheng, Yu, Gong, & Lai, 2017; Ye, Lan, Li, & c Yuen, 2018a; Ye, Wang, Lan, & Yuen, 2018b) 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 Fig. 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, Liu, Wang, Wassell, & Chetty, 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 Fig. 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 Fig. 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 the other by using image translation models, such as CycleGAN (Zhu, Park, Isola, & Efros, 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 for Re-ID task. As shown in Fig. 2(d, e), the generated images by CycleGAN and StarGAN are unsatisfying. Besides, those image translation methods (Choi, et al., 2018; Zhu et al., 2017) use conditional generative models, *i.e.* given an image from a IR mode, they can only generate one image with RGB style. This kind of conditional generative models cannot generate unseen ID and fails to enlarge the heterogeneous datasets.

In this paper, we propose a novel framework, named Joint Set-level and Instance-Level Alignment Re-ID (JSIA-ReID) which enjoys several merits. (1) 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. (2) Given cross-modality unpaired-images of a person, our method can generate cross-modality paired-images. With them, we can perform instance-level alignment by minimizing distance between the two images. The instance-level alignment can reduce the modality-gap and avoid misalignment of instances. (3) Our method can synthesize lots of images of unseen IDs from random noise. This is important for person ReID, a zero-shot image matching task. Because IDs of training and test sets are unshared, ReID methods more easily overfit training IDs. With those synthesized images of unseen ID, better generalization can be obtained.

Specifically, as shown in Fig. 3, our framework consists of three modules, *i.e.* a generation module $\mathcal{G}$, a variation module $\mathcal{V}$ and a feature alignment module $\mathcal{F}$. $\mathcal{G}$ generates cross-modality paired-images from unpaired-ones, $\mathcal{V}$ learns a continuous manifold space, and $\mathcal{F}$ learns 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 Fig. 2(c). The variation module $\mathcal{V}$ includes an encoder and a decoder, both of which play the same roles as VAEs (Kingma & Welling, 2014). The goal is to learn a low-dimensional continuous manifold space for modality-invariant features. In the manifold feature space, we can sample and decode lots of unseen and meaningful modality-invariant features. Then, using the generation module $\mathcal{G}$, we can get more images of unseen IDs. In 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 $\mathcal{G}$, variation module $\mathcal{V}$ and feature alignment module $\mathcal{F}$ with the re-id loss, we can learn both modality-aligned, identity-discriminative and generalizable features.

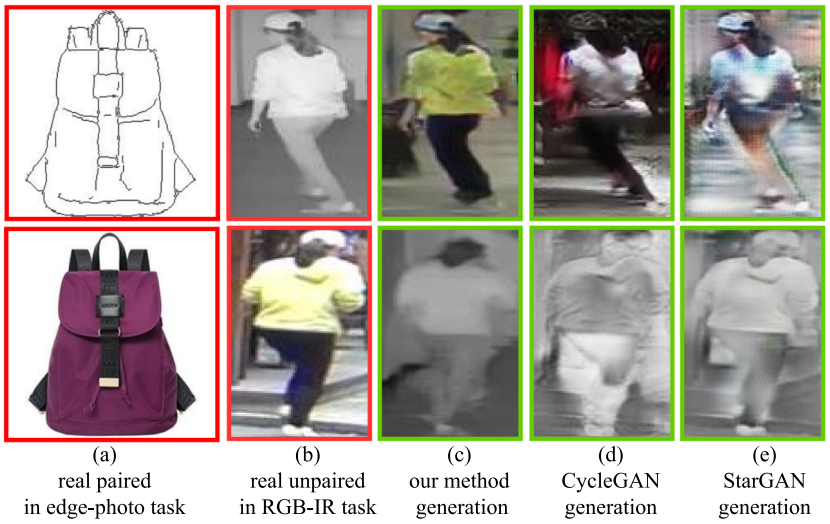

Fig. 2. (a) In the edge-photo task, we can get cross-modality paired-images. By minimizing their distances in a feature space, we can easily reduce the cross-modality gap. (b) In RGB–IR Re-ID task, we have only unpaired-images. The appearance variation caused by the cross-modality gap makes the task 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

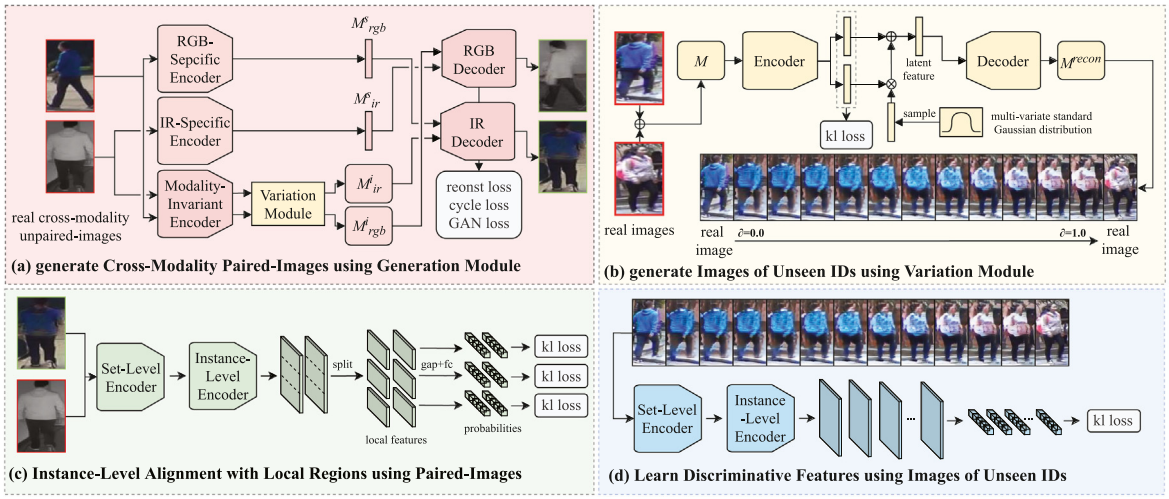

Fig. 3. Our proposed method consists of 4 steps. **(a)** Given cross-modality unpaired-images, generate cross-modality paired-images via the Generation Module. **(b)** Given any two real images, generate images of unseen IDs using the Variation Module. **(c)** Instance-Level Feature Alignment using the generated cross-modality paired-images from *(a)*. **(d)** Discriminative Feature Learning using both seen IDs and unseen IDs from *(b)*. Finally, all the four parts are jointly trained in an end-to-end way.

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 set-level and instance-level modality-variation. The Instance-level alignment can not only reduce modality-gap but also guarantees identity-consistency of features.

(3) We propose a new variation module to map modality-invariant features to a latent manifold feature space. With it, we can sample and generate large-scale images of unseen IDs from noise. Those new images help the learned identity feature space more smooth and generalize better when test.

(4) We improve the instance-level alignment by applying local features. This is based on our generated cross-modality paired-images. Because our paired-images have absolutely same structure (views, poses), we can accurately reduce modality-gap of local regions meanwhile avoiding noise.

(5) Extensive experimental results on two standard benchmarks demonstrate that the proposed model performs against state-of-the-art methods.

2. Related works

In this section, we briefly overview methods that are related to RGB–RGB person re-identification, RGB–IR person re-identification and generative adversarial networks.

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 (Hermans et al., 2017; Wang, Yang, Cheng, Wang, & Hou, 2019b; Zheng et al., 2016) have been proposed. The key challenges lie in the large intra-class variation caused by different views, poses, illuminations, and occlusions. Existing methods can be grouped into hand-crafted descriptors (Liao et al., 2015; Ma et al., 2014; Yang, et al., 2014), metric learning methods (Koestinger et al., 2012; Liao & Li, 2015; Zheng et al., 2013) and deep learning algorithms (Hermans et al., 2017; Sun et al., 2018; Zheng et al., 2016). The goal of hand-crafted descriptors is to design robust features. For example, Ma et al. (2014) handle the background and illumination variations by combining biologically inspired features with covariance descriptors. Yang, et al. (2014) explore color information by using salient color names. In Liao et al. (2015), Liao et al. propose an effective feature representation called local maximal occurrence, which can analyze the horizontal occurrence of local features, and maximize the occurrence to make a stable representation against viewpoint changes. Metric learning methods are designed to make a pair of true matches have a relatively smaller distance than that of a wrong match pair in a discriminant manner. Zheng et al. (2013) formulate person RE-ID as a relative distance comparison learning problem in order to learn the optimal similarity measure between a pair of person images. The model is formulated to maximize the likelihood of a pair of true matches having a relatively smaller distance than that of a wrong match pair in a soft discriminant manner. Deep learning algorithms adopt deep neural networks to straightly learn robust and discriminative features in an end-to-end manner. For example, Zheng et al. (2016) learn identity-discriminative features by fine-tuning a pre-trained CNN to minimize a classification loss. In Hermans et al. (2017), Hermans et al. show that using a variant of the triplet loss outperforms most other published methods by a large margin. In Sun et al. (2018), a network named Part-based Convolutional Baseline (PCB) is proposed to learn fine-grained part-level features with a uniform partition strategy. Most of existing 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. The proposed method explores three different network structures and uses deep zero-padding for training one-stream network toward automatically evolving domain-specific nodes in the network for cross-modality matching. 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. (2018a) propose a hierarchical cross-modality matching model by jointly optimizing the modality-
specific and modality-shared metrics. The modality-specific metrics transform two heterogeneous modalities into a consistent space that modality-shared metric can be subsequently learnt. In Ye et al. (2018b), a dual-path network is proposed with a new bi-directional dual-constrained top-ranking loss to learn discriminative feature representations. 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. (2019) achieve visible thermal person re-identification via a hyper-sphere manifold embedding model. In Wang, et al. (2019) and Wang, Wang, Zheng, Chuang, and Satoh (2019a), 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, Zheng, and Yang (2017) use a GAN model to generate unlabeled images as data augmentation. Huang, et al. (2018) first assign pseudo labels to generated pedestrian images and then learn them in a supervision manner. Zhong, Zheng, Luo, Li, and Yang (2019), Zhong, Zheng, Zheng, Li, and Yang (2018b) and Zhong, Zheng, Li, and Yang (2018a) 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., 2017, 2018) use a cGAN to generate pedestrian images with different poses to learn features free of influences of pose variation. 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.

Generative Models. Variational autoencoders (VAEs) (Kingma & Welling, 2014) and generative adversarial networks (GANs) (Goodfellow, et al., 2014) are the most popular generative models. VAEs includes an encoder and a decoder network. The encoder maps an input image to the latent variables which matches a prior distribution, and the decoder samples images from the latent variable. In this paper, we use VAEs idea to map features to a latent manifold feature space for continuous and smooth representation. GANs (Goodfellow, et al., 2014) learns data distribution in a self-supervised way via the adversarial training. According to Huang, zhihang li, He, Sun, and Tan (2018), VAEs have nice manifold representations, while GANs are better at generating sharper images. GANs has been widely used in image translation. Choi, et al. (2018), Isola, Zhu, Zhou, and Efros (2017) and Zhu et al. (2017) and domain adaptation (Ganin, et al., 2016; Hoffman, et al., 2018) 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. In Zhu et al. (2017), with unpaired data, CycleGAN simultaneously learns two reciprocal image translations between two domains and enforces the translated images to reconstruct their original images. Further, StarGAN (Choi, et al., 2018) learns multi-domain image translations by making the generator take both images and domain labels as inputs, and improving the discriminator to simultaneously distinguish image sources and classify their domains. 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.

3. 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.

3.1. Cross-modality paired-images generation module

As shown in Fig. 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 Eqs. (1) and (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. 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 do not 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:

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

Overall Loss. The overall loss of the cross-modality paired-images generation module can be formulated as below:

$$\mathcal{L}^G = \mathcal{L}_{recon} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{cyc}\mathcal{L}_{gan} \tag{8}$$

3.2. Variation module

Although the generation module above generate cross-modality paired-images by disentangling and exchanging features, it cannot enlarge the heterogeneous ReID dataset. Given an image from a modality, it only generates one image of the other modality. The weakness becomes worse when training data is limited. In this section, we propose a novel variation module $\mathcal{V}$ to generate lots of cross-modality paired-images with the same identity from random noise. The main idea is inspired by VAEs (Kingma & Welling, 2014) to learn a low-dimensional continuous manifold space for modality-invariant features. In the manifold feature space, we can sample and decode lots of unseen and meaningful modality-invariant features. Then, using the cross-modality paired-images generation module, we can generate more paired-images and further enhance ReID features. Different from VAEs which deal with images, ours takes feature map as input and is joined trained with our generation module.

Latent Space Learning. As shown in Fig. 3, the variation module consist of an encoder E^v and a decoder D^v. The encoder E^v maps modality-invariant feature maps M^i from Eq. (1) to a latent space z by a reparameterization trick: $z = u + \tau \cdot \epsilon$, where $u = E^v_u(M)$ and $\tau = E^v_\tau(M)$ denote mean and standard deviation of feature maps, respectively. The decoder D^v takes z as input and reconstruct corresponding feature map.

$$z = E^v(M),\ \ M^{rcon} = D^v(z) \tag{9}$$

Reconstruction and Distribution Loss. This module includes two losses, they are reconstruction loss and distribution loss. The former guarantees the decoder to be able to reconstruct input feature maps M^i from their latent features z. The latter makes sure that latent features z satisfy multi-variant standard Gaussian distributions. The losses are formulated in Eq. (10), where $\|\cdot\|_p$ is p norm, $p \in \{1, 2\}$.

$$\begin{aligned}\mathcal{L}^v &= \mathcal{L}^v_{recon} + \mathcal{L}^v_{distri} \\ &= \|M - M^{recon}\|_1 + \|\mu\|_2 + \|\tau - 1\|_2\end{aligned} \tag{10}$$

3.3. Feature alignment module

Set-Level (SL) 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, Wang, Peng, & Chiu, 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 (IL) 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 misalignment 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. (11).

$$\begin{aligned}\mathcal{L}^{il} &= E_{(x_1,x_2)\in(X_{ir},X_{ir2rgb})}[KL(p_1 \parallel p_2)] \\ &+ E_{(x_1,x_2)\in(X_{rgb2ir},X_{rgb})}[KL(p_1 \parallel p_2)]\end{aligned} \tag{11}$$

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 global average pooling and a fully-connected layer.

Improving IL with Local Regions (LR). In cross-modality paired-images, two images contain absolutely the same contents such as poses, views. This allows us to finely align every local region without worrying about importing noise. Following Sun et al. (2018), we horizontally split feature maps T to n blocks $\{T^i\}_i^n$, and then align every block as in Eq. (12), where $p^n = C^n(t^n)$ are the predicted probabilities, t^n is the nth local feature of image x, C^n is nth classifier, different classifiers do not share weights.

$$\begin{aligned}\mathcal{L}^{il+lr} &= \frac{1}{n}\sum_{i=1}^{n} E_{(x_1,x_2)\in(X_{ir},X_{ir2rgb})}[KL(p_1^n \parallel p_2^n)] \\ &+ \frac{1}{n}\sum_{i=1}^{n} E_{(x_1,x_2)\in(X_{rgb2ir},X_{rgb})}[KL(p_1^n \parallel p_2^n)]\end{aligned} \tag{12}$$

Identity-Discriminative Feature Learning. To overcome the intra-modality variation, following Hermans et al. (2017) and Zheng et al. (2016), 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 $\mathcal{L}_{triplet}$.

$$\begin{aligned}\mathcal{L}^{id} &= \mathcal{L}_{cls} + \mathcal{L}_{triplet} \\ &= E_{v\in V}(-log\ p(v)) + E_{v\in V}[m - D_{v_a,v_p} + D_{v_a,v_n}]_+\end{aligned} \tag{13}$$

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 $[x]_+ = max(0, x)$, D is the L2 distance.

Enhancing Features with Unseen IDs (UI). Given any two images x_1 and x_2 from the same modality, we first use the variation module $\mathcal{V}$ to learn their features z_1 and z_2 in latent space, and their modality-specific features m_1^s and m_2^s. Then we mix their latent and modality-specific features with a ratio α as below.

$$\begin{aligned}z_{mix} &= \alpha z_1 + (1-\alpha)z_2,\ s.t.\ \alpha \in [0,1] \\ m_{mix}^s &= \alpha m_1^s + (1-\alpha)m_2^s,\ s.t.\ \alpha \in [0,1]\end{aligned} \tag{14}$$

Algorithm 1 Overview of Proposed Method

Input: (Train) Cross-Modality Unpaired-Images X_{rgb} and X_{ir}. (Test) Query and Gallery Images X_q and X_g.

Train:

1: Generate cross-modality paired-images (X_{rgb}, X_{rgb2ir}) and (X_{ir}, X_{ir2rgb}) via Eq. (3)
2: Generate images of unseen IDs X_{mix} via Eq. (14)
3: Instance-Level Alignment using Paired-Images via Eq. (11)
4: Learn discriminative features with both seen and unseen IDs via Eqs. (13) and (15)
5: Train the framework in an end-to-end way via Eq. (16)

Test:

1: Compute features of query and gallery images V_q and V_g via Eq. (17)
2: Compute cosine similarities between V_q and V_g
3: Rank by sorting according to cosine similarities

Then, we can reconstruct the mixed modality-invariant feature map m_{mix}^i via Eq. (9). Finally, through Eq. (3), we can get a new person image x_{mix} which contain information from both x_1 and x_2. We use classification loss to train x_{mix}, whose ground truth probability is $\alpha p_1 + (1-\alpha)p_2$. Here, p_1 and p_2 are the predicted probability of images x_1 and x_2.

$$\mathcal{L}^{ui} = KL(p_{mix} \parallel \alpha p_1 + (1-\alpha)p_2) \tag{15}$$

3.4. Overall objective function and test

The overall objective function of our method is formulated as below:

$$\mathcal{L} = \mathcal{L}^g + \mathcal{L}^v + \mathcal{L}^{id} + \lambda_{il}\mathcal{L}^{il} + \lambda_{lr}\mathcal{L}^{lr} + \lambda_{ui}\mathcal{L}^{ui} \tag{16}$$

where λ_* are weights of corresponding terms. They are decided by cross-validation.

During the test stage, only feature learning module $\mathcal{F}$ is used. Given query and gallery images X_q and X_g, we use the set-level alignment encoder E^{sl} and the instance-level encoder E^{il} to extract features as in Eq. (17). Then, compute cosine similarities of query and gallery feature vectors V_q and V_g. Finally, the results are returned via nearest neighbor search on the similarities.

$$V_q = E^{il}(E^{sl}(X_q)),\ \ V_g = E^{il}(E^{sl}(X_g)) \tag{17}$$

4. Experiment

4.1. 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 1*st*, 2*nd*, 3*rd*, 6*th* 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, Hong, Kim, & Park, 2017) contains 412 persons, where each person has 10 images from a visible camera and 10 images from a thermal camera.

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^2RL	28.9	70.6	82.4	29.2	–	–	–	–	–	–	–	–	–	–	–	–
Ours	**43.4**	**81.3**	**90.5**	**38.0**	**49.7**	**87.0**	**94.2**	**30.0**	**46.8**	**88.2**	**94.7**	**54.7**	**57.8**	**92.1**	**97.1**	**44.3**

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.

4.2. 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, Liang, Zhang, Yeung, & Xing, 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 LSGAN 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, Zhang, Ren, & Sun, 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.

4.3. Comparison with state-of-the-arts

Results on SYSU-MM01 Datasets We compare our model with 10 methods including hand-crafted features (HOG Dalal & 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 divergence of features (cmGAN Dai et al., 2018, D^2RL Wang et al., 2019a). The experimental results are shown in Table 1.

Table 2
Comparison with state-of-the-arts on the RegDB dataset under different query settings. thermal2visible means use thermal images as query and visible images as gallery, vice versa. mAP denotes mean average precision scores (%).

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^2RL	43.4	44.1	43.4	44.1
Ours	51.3	52.0	52.1	51.9

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^2RL 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^2RL (Wang et al., 2019a). We adopt two settings, *i.e* 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 7.9% and 8.7% in terms of Rank-1 scores with thermal2visible and visible2thermal modes, respectively. Overall, the results verify the effectiveness of our model.

4.4. Model analysis

Ablation Study. To further analyze effectiveness of the set-level alignment (SL) the instance-level alignment (IL), paired-images of unseen IDs (UI) and feature learning with local regions

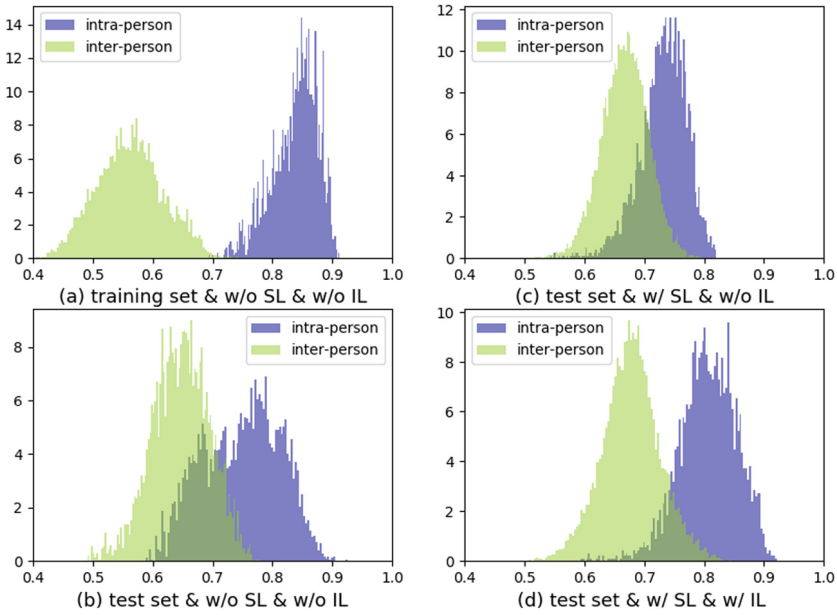

Fig. 4. Distribution of cross-modality similarities of intra-person and inter-person. The instance-level alignment (IL) can enhance intra-person similarity while keep inter-person similarity unchanged, which improves performance. Please note that w/ means with and w/o means without. Please see text for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

Table 3
Analysis of set-level (SL), instance-level (IL) alignment, feature learning with local regions (LR) and paired-images of unseen IDs (UI). Please see text for more details.

index	SL	IL	LR	UI	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	✓	✓	✓	×	40.5	82.2	90.5	38.3
6	✓	✓	✓	✓	43.2	83.5	91.1	39.9
7	–	✓	✓	✓	40.0	81.9	90.0	37.5
8	✓	✓	–	✓	41.9	82.5	91.1	38.9

(LR), we evaluate our method under 6 different settings. Specifically, when removing set-level alignment (SL), we use separate set-level encoder E^{sl}, *i.e.* we do not share weights of set-level encoder E^{sl} with modality-invariant encoder E^{i}. When removing instance-level alignment (IL), learning with local regions (LR) or paired-images of unseen IDs (UI), we set corresponding weights in Eq. (16)λ_{il}, λ_{lr} or λ_{ui} as 0. Moreover, to analyze whether the feature disentanglement strategy contributes to set-level alignment, we use a degraded set-level encoder by do not sharing it weight with modality-invariant encoder and train it with a GAN loss as in Dai et al. (2018). To shows the importance of generated paired-images for local region learning, we conduct a degraded version, that is use original unpaired-images. Please note that symbol × means use no the module, – means using the degraded module, and ✓represents using the module.

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 better performance at 38.1% Rank-1 score, which demonstrates that SL and IL can be complementary with each other. Further, when adding LR (index-5) and UI (index-6), the Rank-1 score increases to 40.5% and 43.2%, showing the effectiveness of LR and UI. Finally, when removing the disentanglement from set-level alignment (index-7), Rank-1 score drops by 3.2%. This illustrates that disentanglement strategy is helpful for learning set-level alignment. When using unpaired-images for LR (index8), the performance also drop by 1.4% Rank-1 score. This shows that paired-images are important local feature learning.

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 Fig. 4(a) and (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 Fig. 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 Fig. 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 Fig. 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 parameters, they are weights in Eq. (16) including λ_{il}, λ_{ir}, λ_{ul} and local region number n. As shown in Fig. 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, at most parameters, our method can stably have an significant improvement. The experimental results show that our method is robust to different parameters.

4.5. Analysis of generated images

In our framework, the generated cross-modality paired-images play an important roles, and is necessary to be evaluated. Here, we evaluate the synthetic images from real images (FRI) and from latent space (FLS) with two evaluation protocols, *i.e.* mean distance (MD) and Frechet Inception Distance (FID) (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017). The former represents identity consistency with L2 distance of images in a pair, the latter is distribution consistency between real and synthetic

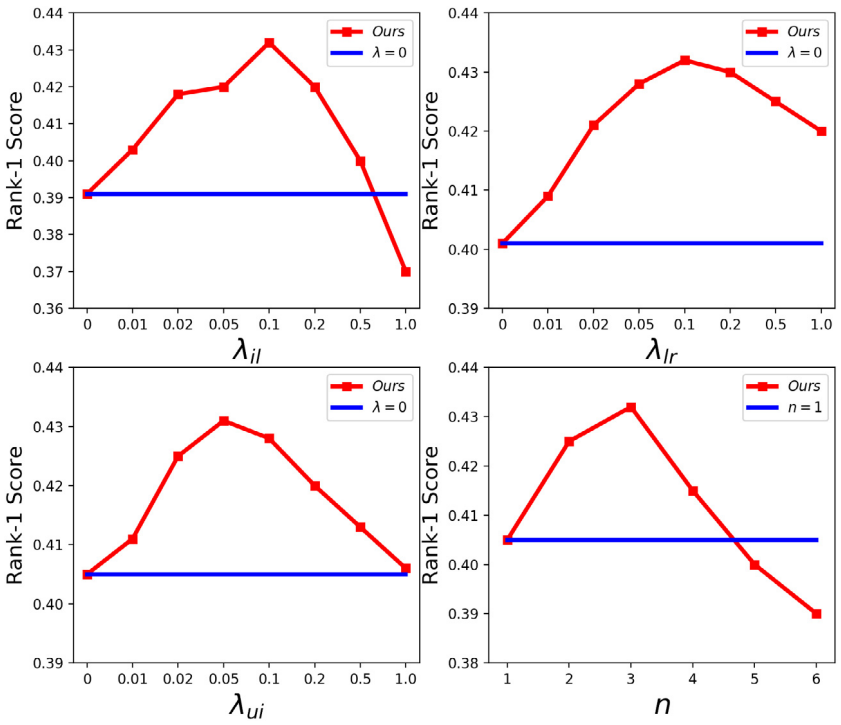

Fig. 5. Parameter analysis. λ_{il}, λ_{ir}, λ_{ul} are weights in Eq. (16), n is local region number in Eq. (12). Rank-1 score is evaluated on SYSU-MM01 under *single-shot&all-search* mode.

Table 4
Identity and distribution consistency analysis of synthetic images from ours, CycleGAN (Zhu et al., 2017), StarGAN (Choi, et al., 2018) and VAEs (Kingma & Welling, 2014) on SYSU-MM01 dataset. FRI means synthesizing images from real images, FLS from latent space. MD and FID are metrics for identity and distribution consistency, respectively. Smaller is better. Please see text for more details.

Methods	MD(identity)		FID(distribution)	
	FRI	FLS	FRI	FLS
CycleGAN	0.52	×	9.9	×
StarGAN	0.55	×	9.5	×
VAE	0.64	0.69	9.6	8.9
Ours	0.41	0.56	7.5	8.2

images. For both items, smaller values means better consistency. The experimental results are reported in Table 4.

Identity Consistency. Identity consistency is important for instance-alignment. That is, where the synthetic paired-images are the same person. As we can see in Table 4 (left-column), when synthesizing images from real ones, VAEs performs worst, followed by StarGAN and CycleGAN, and ours is best. The reason is that ours utilize disentanglement strategy and overcome one-to-multi difficulty. For images generated from latent space (FLS), their FLS score is worse than that of FRI, this is because the features are sampled from low-dimensional latent space, which loses some details and lead to fuzzy images. Please see Figs. 6 and 7 for visualization of the images.

Distribution Consistency. Distribution consistency shows how realistic a synthetic image is. As we can see in Table 4(right-column), when generating images from real , FRI is Please note that FRN performs worse then FRN. The reason is that the latter sample images from a latent space, leading to fuzzy images. While the former translate a real image to another style, which is more easier.

Visualization of Synthetic Images FRI. We display the generated cross-modality paired-images from ours, CycleGAN (Zhu et al., 2017) and StarGAN (Choi, et al., 2018). From Fig. 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 Fig. 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 Fig. 6(c) and (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.

Visualization of Synthetic Images FLS. To better understand the synthetic images of unseen classes, we display them in Fig. 7. Specifically, the most left and right images are real ones, middle ones are generated by mixing corresponding real images in the latent feature space. We can observe that the synthetic images contain information of the two real images, *e.g.* color, body type, clothes and so on. The preference can be continuously and smoothly controlled by the parameter α. Those images are realistic and unseen in original training set. Thus we can enlarge the training set with those unseen class by mixing any two real images with different parameter α.

5. 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 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. To learn fine-grained features, we perform instance-level alignment on

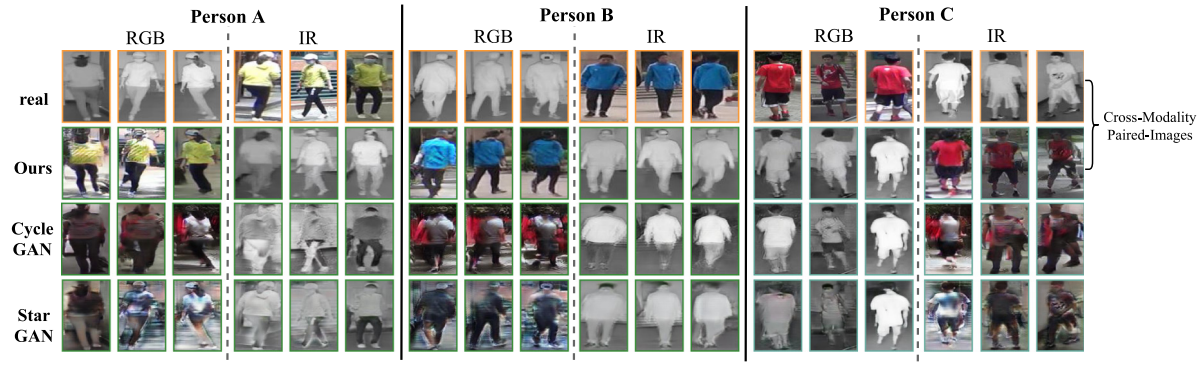

Fig. 6. Visualization of synthetic cross-modality paired-images generated from real unpaired-ones. We compare ours with CycleGAN (Zhu et al., 2017) and StarGAN (Choi, et al., 2018). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

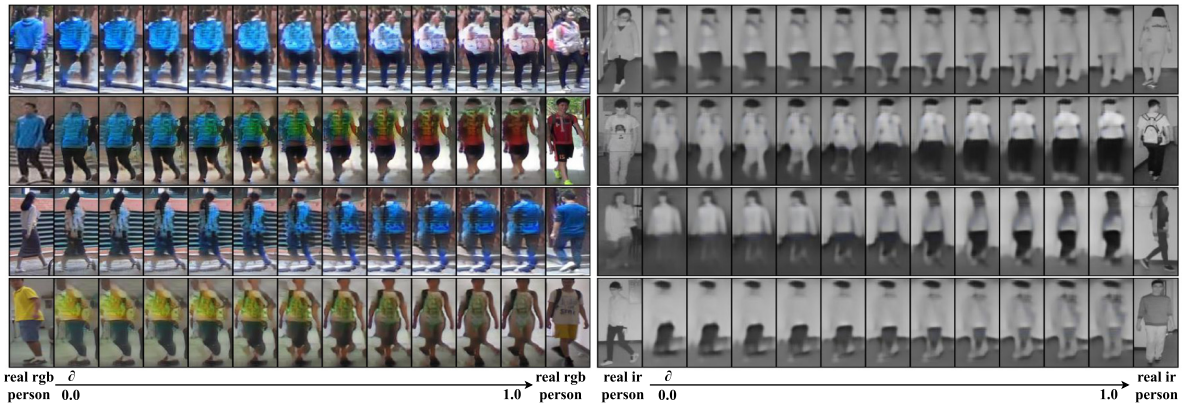

Fig. 7. Visualization of synthetic images generated from the latent space. As we can see, our latent space is continuous and smooth. The synthetic images contain characters of both images. The parameter α can control preference of the new image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using both global and local features. Considering that every image in a pair contains absolutely the same content information such as pose and view, the local-feature alignment does not import extra noise. Besides, we can straightly synthesize cross-modality paired images of unseen IDs from random noise. Those unseen IDs can further enhance feature learning and achieve better generalization. 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

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., & 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* (pp. 1900–1909). URL: https://academic.microsoft.com/paper/2932414082.

Chen, Q., Liu, Y., Wang, Z., Wassell, I. J., & Chetty, K. (2018). Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7976–7985). URL: https://academic.microsoft.com/paper/2798377719.

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & 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* (pp. 8789–8797).

Dai, P., Ji, R., Wang, H., Wu, Q., & Huang, Y. (2018). Cross-modality person re-identification with generative adversarial training. In *IJCAI 2018: 27th international joint conference on artificial intelligence* (pp. 677–683).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *International conference on computer vision & pattern recognition (vol. 1)* (pp. 886–893). IEEE Computer Society.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, *17*(1), 1–35.

Gong, S., Cristani, M., Yan, S., & Loy, C. C. (2014). *Person re-identification* (pp. 301–313).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems (vol. 27)* (pp. 2672–2680).

Hao, Y., Wang, N., Li, J., & 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., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

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

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANS trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in neural information processing systems* (pp. 6626–6637). URL: https://academic.microsoft.com/paper/2963981733.

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., et al. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989–1998).

Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z., & Zhang, J. (2018). Multi-pseudo regularized label for generated samples in person re-identification. arXiv preprint arXiv:1801.06742. URL: https://academic.microsoft.com/paper/2785286326.

Huang, H., zhihang li, He, R., Sun, Z., & Tan, T. (2018). IntroVAE: Introspective Variational autoencoders for photographic image synthesis. In *NIPS 2018: The 32nd annual conference on neural information processing systems* (pp. 52–63). URL: https://academic.microsoft.com/paper/2884581909.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 5967–5976).

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *ICLR 2014 : International conference on learning representations*. URL: https://academic.microsoft.com/paper/1959608418.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2288–2295). IEEE.

Liao, S., Hu, Y., Zhu, X., & 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* (pp. 2197–2206).

Liao, S., & Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3685–3693).

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Gool, L. V. (2017). Pose guided person image generation. In *31st Annual conference on neural information processing systems* (pp. 406–416). URL: https://academic.microsoft.com/paper/2962819541.

Ma, B., Su, Y., & Jurie, F. (2014). Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, *32*(6–7), 379–390.

Ma, L., Sun, Q., Georgoulis, S., Gool, L. V., Schiele, B., & Fritz, M. (2018). Disentangled person image generation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 99–108). URL: https://academic.microsoft.com/paper/2771558241.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2016). Least squares generative adversarial networks. arXiv preprint arXiv:1611.04076. URL: https://academic.microsoft.com/paper/2949496494.

Nguyen, D. T., Hong, H. G., Kim, K. W., & 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., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.*, *61*, 85–117.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision* (pp. 480–496).

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2018). Deep learning in spiking neural networks. *Neural Networks*.

Wang, H., Liang, X., Zhang, H., Yeung, D. Y., & Xing, E. P. (2017). ZM-NEt: Real-time zero-shot image manipulation network. arXiv preprint arXiv:1703.07255. URL: https://academic.microsoft.com/paper/2605028456.

Wang, Z., Wang, Z., Zheng, Y., Chuang, Y. Y., & Satoh, S. (2019). Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 618–626). URL: https://academic.microsoft.com/paper/2954773727.

Wang, G., Yang, Y., Cheng, J., Wang, J., & Hou, Z. (2019). Color-sensitive person re-identification. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 933–939). URL: https://academic.microsoft.com/paper/2964438507.

Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., & Hou, Z. (2019). RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. In *The IEEE international conference on computer vision*.

Wu, A., Zheng, W. S., Yu, H. X., Gong, S., & Lai, J. (2017). RGB-infrared cross-modality person re-identification. In *2017 IEEE international conference on computer vision* (pp. 5390–5399).

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., & Li, S. Z. (2014). Salient color names for person re-identification. In *European conference on computer vision* (pp. 536–551). Springer.

Ye, M., Lan, X., Li, J., & c Yuen, P. (2018). Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI-18 AAAI conference on artificial intelligence* (pp. 7501–7508).

Ye, M., Wang, Z., Lan, X., & Yuen, P. C. (2018). Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI 2018: 27th International joint conference on artificial intelligence* (pp. 1092–1099).

Zheng, W. S., Gong, S., & Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(3), 653–668.

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

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., & Kautz, J. (2019). Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2138–2147). URL: https://academic.microsoft.com/paper/2963049565.

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. arXiv preprint arXiv:1701.07717. URL: https://academic.microsoft.com/paper/2949257576.

Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018). Generalizing A person retrieval model hetero- and homogeneously. In *Proceedings of the European conference on computer vision* (pp. 176–192). URL: https://academic.microsoft.com/paper/2896016251.

Zhong, Z., Zheng, L., Luo, Z., Li, S., & Yang, Y. (2019). Invariance matters: Exemplar memory for domain adaptive person re-identification. In *2019 IEEE conference on computer vision and pattern recognition*.

Zhong, Z., Zheng, L., Zheng, Z., Li, S., & Yang, Y. (2018). Camera style adaptation for person re-identification. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 5157–5166). URL: https://academic.microsoft.com/paper/2963289251.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision* (pp. 2242–2251).