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

Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, Zengguang Hou

1Institute of Automation, Chinese Academy of Sciences, Beijing, China
2University of Chinese Academy of Sciences, Beijing, China
3Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
4University 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 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 outperforms 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.

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 (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 \( G \) to generate cross-modality paired-images and a feature alignment module \( F \) to learn both set-level and instance-level aligned features. The generation module \( G \) 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 \( F \), 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 \( G \) and feature alignment module \( F \) 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. 2018) 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. All those methods focus on single-modality RGB Re-ID and cannot deal with cross-modality RGB-IR Re-ID. Different from them, ours 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 $G$ to generate cross-modality paired-images and a feature alignment module $F$ 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.(1).

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

(1)

**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 $M^c_{ir}$ 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).

$$X_{ir2rgb} = D_{rgb}(M^c_{ir}, M^s_{rgb}), X_{rgb2ir} = D_{rgb}(M^c_{rgb}, M^s_{ir})$$

(3)
Reconstruction Loss. A simple supervision is to force the disentangled features to reconstruct their original images. Thus, we can formulate the reconstruction loss $L_{\text{recon}}$ as below, where $|| \cdot ||_1$ is L1 distance.

$$L_{\text{recon}} = ||X_{rgb} - D_{rgb}(E^i(X_{rgb}), E^s_{rgb}(X_{rgb}))||_1 + ||X_{ir} - D_{ir}(E^i(X_{ir}), E^s_{ir}(X_{ir}))||_1$$ (4)

Cycle-Consistency Loss. The reconstruction loss $L_{\text{recon}}$ 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:

$$L_{\text{cyc}} = ||X_{rgb} - X_{rgb2ir2gb}||_1 + ||X_{ir} - X_{ir2rgb2ir}||_1$$ (5)

where $X_{ir2rgb2ir}$ and $X_{rgb2ir2gb}$ are the cycle-reconstructed images as in Eq.(6).

$$X_{ir2rgb2ir} = D_{ir}(E^i_{rgb}(X_{ir2rgb}), E^s_{rgb}(X_{ir2rgb}))$$

$$X_{rgb2ir2gb} = D_{rgb}(E^i_{ir}(X_{rgb2ir}), E^s_{ir}(X_{rgb2ir}))$$ (6)

GAN loss. The reconstruction loss $L_{\text{recon}}$ and cycle-consistency loss $L_{\text{cyc}}$ 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:

$$L_{\text{gan}} = E[\log Dis_{rgb}(X_{rgb}) + \log(1 - Dis_{rgb}(X_{ir2rgb}))] + E[\log Dis_{ir}(X_{ir}) + \log(1 - Dis_{ir}(X_{rgb2ir}))]$$ (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 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 $T$, i.e. $T = E^il(M)$. Then, based on the feature space $T$, 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).

$$
L_{align} = E_{(x_1,x_2) \in (X_{ir},X_{ir},2irgb)}[KL(p_1||p_2)] + E_{(x_1,x_2) \in (X_{rgb2ir},X_{rgb})}[KL(p_1||p_2)]
$$

(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 $T$, $C$ 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 $T$ to corresponding feature vectors $V$. Given real images $X$, we optimize their feature vectors $V$ with a classification loss $L_{cls}$ of a classifier $C$ and a triplet loss $L_{triplet}$. $L_{cls} = Eve \log p(v)$

$$
L_{triplet} = E_{v \in V}[m - D_{v,v} + D_{v,v}]
$$

(10)

where $p(.)$ is the predicted probability predicted by the classifier $C$ that the input feature vector belongs to the ground-truth, $v_0$ and $v_p$ are a positive pair of feature vectors belonging to the same person, $v_n$ 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 be formulated as below:

$$
L = \lambda_{cye}L_{cye} + \lambda_{gan}L_{gan} + \lambda_{align}L_{align} + \lambda_{reid}(L_{cls} + L_{triplet})
$$

(11)

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

During the test stage, only feature learning module $F$ 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 $G$, 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 $F$, 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.

Please see our code\textsuperscript{1} for more implementation details.

**Comparison with State-of-the-arts**

**Results on SYSU-MM01 Datasets.** We compare our model with 10 methods including hand-crafted 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 divergence of features (cmGAN (Dai et al. 2018), D²RL (Wang et al. 2019c)). The 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

\textsuperscript{1}https://github.com/wangguanan/JSIA-ReID
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 outperforms 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.

**Results on SYSU-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.

**Ablation Study.** To further analyze the 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 demonstrates 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 alignment (index-5), Rank-1 score drops by 1.3%. This illustrates that disentanglement is helpful for 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)
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. w/ means with and w/o means without.

Figure 5: Rank-1 and mAP scores with different $\lambda_{align}$ 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.

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.

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. 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,
References


