Hindawi
Mobile Information Systems
Volume 2022, Article ID 4131322, 9 pages
https://doi.org/10.1155/2022/4131322

Research Article

Learning Identity-Consistent Feature for Cross-Modality Person Re-Identification via Pixel and Feature Alignment

Sixian Chan,[1,2] Feng Du,[1] Yanjing Lei,[1] Zhounian Lai,[3] Jiafa Mao,[1] and Chao Li[4]

[1] *Zhejiang University of Technology, Hangzhou, China*
[2] *Hangzhou Xsuan Technology Co. Ltd, Hangzhou, China*
[3] *Huzhou Institute of Zhejiang University, Hangzhou, China*
[4] *Zhijiang College of Zhejiang University of Technology, Hangzhou, China*

Correspondence should be addressed to Yanjing Lei; leiyj@zjut.edu.cn

Received 16 August 2022; Accepted 17 September 2022; Published 10 October 2022

Academic Editor: Wang Wenyong

RGB-IR cross-modality person re-identification (ReID) can be seen as a multicamera retrieval problem that aims to match pedestrian images captured by visible and infrared cameras. Most of the existing methods focus on reducing modality differences through feature representation learning. However, they ignore the huge difference in pixel space between the two modalities. Unlike these methods, we utilize the pixel and feature alignment network (PFANet) to reduce modal differences in pixel space while aligning features in feature space in this paper. Our model contains three components, including a feature extractor, a generator, and a joint discriminator. Like previous methods, the generator and the joint discriminator are used to generate high-quality cross-modality images; however, we make substantial improvements to the feature extraction module. Firstly, we fuse batch normalization and global attention (BNG) which can pay attention to channel information while conducting information interaction between channels and spaces. Secondly, to alleviate the modal difference in feature space, we propose the modal mitigation module (MMM). Then, by jointly training the entire model, our model is able to not only mitigate the cross-modality and intramodality variations but also learn identity-consistent features. Finally, extensive experimental results show that our model outperforms other methods. On the SYSU-MM01 dataset, our model achieves a rank-1 accuracy of 40.83% and an mAP of 39.84%.

1. Introduction

Person ReID can be viewed as a cross-camera image retrieval problem, which aims at matching individual pedestrian images in a query set to ones in a gallery set captured by different cameras. Its main challenge lies in the interclass and intraclass variations caused by different lighting, poses, occlusions, and views. Most existing methods [1–5] mainly focus on matching RGB images captured by visible cameras, which can be formulated as an image matching problem under a single modality. However, these methods cannot be applied to images taken in poor lighting conditions, because the visible camera cannot capture pictures with discriminative features. However, in practical application scenarios, the camera should ensure all-weather operation.

Since the visible camera has limited effect on the security work at night, the camera that can switch the infrared mode is being widely used in the intelligent monitoring system. In visible mode and infrared mode, RGB images and infrared images are collected, respectively, which belong to two different modalities. RGB images have three channels but IR images have only one channel, so the ReID problem in a cross-modality setting becomes extremely challenging, which is essentially a cross-channel retrieval problem. First, infrared images of different identities are difficult to distinguish but are easy to distinguish in visible images. In addition, the same person varies greatly in different modalities. It is known as modality discrepancy.

To address visible-infrared person ReID, several approaches [6–10] have been proposed, aiming to mitigate

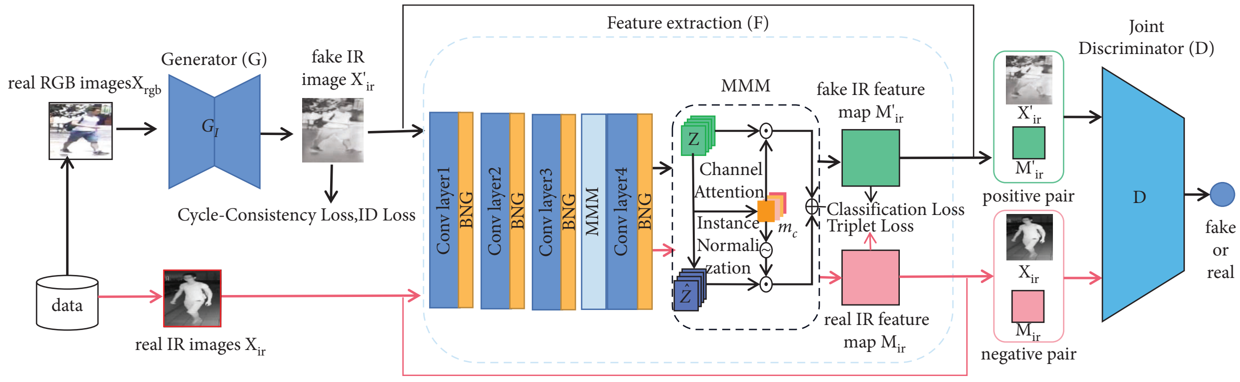

Figure 1: Framework of the proposed model. It consists of an image generation module (G), a joint discriminator module (D), and a feature extraction module (F). The G can generate fake IR images X'_{ir} to mitigate the cross-modality variation, and the F can alleviate the intramodality variation. The F module contains ResNet-50 and BNG attention and MMM module. The BNG module can focus on channel and spatial information, and the MMM module can reduce modality differences.

modal differences by aligning features or pixel distributions. Feature alignment methods [6, 8, 10] mainly focus on bridging the gap between RGB and IR images through features. It is difficult to match RGB and IR images in a shared space due to large cross-modality differences between the two modalities. Different from existing methods that directly match RGB and IR images, we use generative adversarial networks to generate fake IR images based on real RGB images and then match the generated images through a feature alignment network. The generated fake IR images are used to reduce the modality difference between the RGB and IR images. Although the generated fake IR images are very similar to real images, there are still intraclass differences due to pose variations, viewpoint changes, and occlusions.

Inspired by the above discussion, in this paper, we propose a pixel and feature alignment network (PFANet) that simultaneously mitigates cross-modality differences in pixel space and intramodality variation in feature space. As shown in Figure 1, to reduce the modal difference, we apply a generator (G_I) to generate fake IR images. Then, to alleviate the intramodality variation, a feature extraction module (F) is designed to encode fake and real IR images into a shared feature space by exploiting identity-based classification and triplet loss. The batch normalization and global (BNG) attention is added to the feature extraction network (F), which can make the network learn which channel is more important as well as can interact between channels and spaces. Furthermore, to mitigate the modal difference in the feature space, a modal mitigation module (MMM) is proposed, which can significantly mitigate the difference between the two modalities. Finally, to learn identity-consistent recognition, a joint discriminator (D) is utilized. Its input is an image-feature pair.

The major contributions of this work can be summarized as follows:

(i) We propose a generative adversarial network to generate cross-modality images that alleviated modal differences in pixel space. This model consists of a generator and a joint discriminator, by playing a max-min game, our model is able to not only reduce the cross-modality and intramodality variations but also learn identity-consistent features.

(ii) We design a batch normalization and global (BNG) attention, which consists of channel attention and global attention. In the channel attention, we measure the importance of each channel by applying the scale factor of BN to the channel dimension and suppressing insignificant features. As for the global attention module, it can reduce information attenuation and amplify the features of global dimension interaction.

(iii) We apply a modal mitigation module (MMM) to mitigate the modal distribution. The instance normalization (IN) is utilized to mitigate modal differences on a single instance. Moreover, the channel attention is used to guide the learning of IN, which can mitigate modal differences while preserving identity information.

2. Related Works

2.1. RGB-IR Person ReID. RGB-IR cross-modality person ReID can be seen as a multicamera retrieval problem that aims to match pedestrian images captured by visible and infrared cameras, which are widely used in video surveillance, public security, and smart cities. Compared with RGB-RGB single-modality person ReID which only deals with RGB images, the key challenge in this work is to mitigate the large differences between the two modalities. To address the challenge caused by differences in modality distributions, a variety of approaches to cross-modality person re-identification have been proposed. Some early work focused on solving the channel mismatch between RGB images and IR images, due to RGB images having three channels. In contrast, IR images have only one channel. Wu et al. [10] proposed a deep zero-padding network and contributed a new ReID dataset SYSU-MM01. In [11], a dual-path network with a bi-directional dual-constrained top-ranking loss was introduced to learn modality alignment feature representations for RGB-IR ReID. Feng et al. [12] proposed a framework for solving heterogeneous matching problems using modality-specific networks. Ye et al. [13] proposed a dual-stream network with feature learning and metric learning to convert two heterogeneous modalities into a consistent space where the modalities share a metric. Dai et al. [6] introduced a cross-modality generative adversarial network (cmGAN) to reduce the distribution differences between RGB and IR features. Most of the above approaches mostly focus on reducing intermodality differences by feature alignment, while ignoring the large cross-modality differences in pixel space.

Unlike these approaches, the proposed model in this paper is able to combine feature alignment and pixel alignment, effectively reducing intramodality and cross-modality variations. By training the model, the model is able to learn identity consistency features.

2.2. GAN in Person ReID. A generative adversarial network (GAN) consists of a generator and a discriminator, using the idea of game theory, where the generator tries to generate an image to deceive the discriminator, and the discriminator tries to discriminate whether the image is real or generated. Through multiple adversarial training, generative adversarial networks are able to learn deep representations of data in a self-supervised manner. GAN can generate high-quality images, perform image enhancement, generate images from text, and convert images from one domain to another [14, 15]. GAN was first proposed in 2014's [16]. After that, researchers have proposed a variety of task-specific GAN structures, such as CycleGAN [14], Pix2Pix [17], and StarGAN [15]. There are many works in the field of pedestrian re-identification that also apply GAN to improve accuracy. Li et al. [18] proposed a network that allows querying images of different resolutions to process cross-resolution person ReID. Wang et al. [19] designed an end-to-end alignment generative adversarial network (AlignGAN) for the RGB-IR ReID task. JSIA-ReID [20] implemented a two-layer alignment of pixels and features in a unified GAN framework.

In our work, we apply GAN to generate cross-modality images that mitigate modal differences between RGB-IR image data in pixel space.

2.3. Attention Mechanisms. There is an important feature in the human visual system that allows people to selectively focus on things of interest in order to capture valuable information. Inspired by the human visual system, many works have attempted to employ attention mechanisms to improve the performance of CNNs.

Attention mechanisms enable the network to focus on areas of interest to the human body and better extract useful information. SENet [21]integrated spatial information into the channel-level feature responses and computed the corresponding attention with two MLP layers. Later, bottleneck attention module (BAM) [22] built independent space and channel submodules in parallel and embedded them into each bottleneck block. Considering the relationship between any two positions of the feature map, nonlocal feature attention [23] was proposed to capture the relationship between them. The convolution block attention module (CBAM) [24] sequentially cascaded channel attention and spatial attention. However, these works ignored the information about the weights adjusted from the training; therefore, we wanted to highlight the significant features by using the variance of the trained model weights, which also was able to amplify cross-dimensional interactions and captured important features of all three dimensions. We propose new attention (BNG) to solve the above problem. A modal mitigation module (MMM) is designed to mitigate the modal distribution, using channel attention to guide the learning of instance normalization (IN) for mitigating modal differences while preserving identity information.

3. The Proposed Method

In this part, we introduce the proposed PFANet in detail. Our network will be presented in the following three parts, including (1) RGB-IR images generation module, (2) BNG attention module, and (3) modal mitigation module. To reduce cross-modality variation, we apply generative adversarial networks to convert RGB images to fake IR images, which have IR style while maintaining their original identity.

Then, the features of the two modalities are extracted for feature alignment. The BNG attention is designed to make the network focus on channel and spatial information. In addition, the modal mitigation module (MMM) is proposed to mitigate the differences between the two modalities. The main output of the PFAnet during testing is the feature for person ReID.

3.1. RGB-IR Images Generation Module. There is a large cross-modality difference between RGB and IR images, which significantly increases the difficulty of the task of cross-modality pedestrian re-identification. To reduce cross-modality variation, we apply generative adversarial networks to convert RGB images X_{rgb} to fake IR images X'_{ir}, which has IR style while maintaining their original identities. The generated fake IR image X'_{ir} can mitigate the modality differences between RGB and IR images. The module consists of a generator G_I that generates a fake IR image from an RGB image and a joint discriminator D_I that discriminates whether the image is a real image or a generated image. The input of the generator is the real images X_{rgb}, and its output is the fake IR images $X'_{ir} = G_I(X_{rgb})$. The input of the discriminator is the generated fake IR image X'_{ir}; if the image is real, its output is one, and if the image is the generated image, the output is zero. The goal of the generator is to make the generated image as similar as possible to the real image, and the goal of the discriminator is to discriminate as much as possible whether the input image is real or generated. Unlike ordinary discriminators, the input to our discriminator is a pair of IR images and ReID feature maps. The generator and discriminator play the min-max game as [16], and the modal can make the fake IR image X'_{ir} as realistic as possible.

The adversarial loss for generating IR images is defined as follows:

$$\mathscr{L}_{G_I} = \mathbb{E}\left[\log_{D_I}\left(X'_{ir}, f_{map}^{X'_{ir}}\right)\right], \tag{1}$$

$$\mathscr{L}_{D_I} = \mathscr{L}_{D_I}^{real} + \mathscr{L}_{D_I}^{fake}, \tag{2}$$

where

$$\mathscr{L}_{D_I}^{real} = \mathbb{E}_{(x,f)\in\left(X_{ir}, f_{map,R}^{X_{ir}}\right)}\left[\log D_I(x, f)\right], \tag{3}$$

$$\mathscr{L}_{D_I}^{fake} = \mathbb{E}_{(x,f)\in M}\left[\log\left(1 - D_I(x, f)\right)\right], \tag{4}$$

$$M = \left(X'_{ir}, f_{map,R}^{X'_{ir}}\right) \cup \left(X'_{ir}, f_{map,R}^{X_{ir}}\right) \cup \left(X_{ir}, f_{map,R}^{X'_{ir}}\right). \tag{5}$$

Among them, $f_{map,R}^{X_{ir}}$ is the extracted image feature of X_{ir} and $f_{map,R}^{X'_{ir}}$ is the extracted image feature of generated image X'_{ir}. Equation (1) is used to train the generator model; after the constraint of the loss function, the generator will generate a more realistic IR image. Equations (3) and (4) are used to train the discriminator, which differs from traditional discriminators in that the input is a pair of image features. It has two advantages, firstly, the fake IR image X'_{ir} will be closer to the real IR image X_{ir} through the max-min game [16], and the distribution of the features $f_{map,R}^{X'_{ir}}$ of the fake IR image will be more similar to the real image features $f_{map,R}^{X_{ir}}$. Secondly, $f_{map,R}^{X'_{ir}}$ is able to maintain the identity-consistency by the corresponding image X'_{ir} constraint. Although $\mathscr{L}_{G_I}$ loss can ensure that the fake IR image X'_{ir} resembles the real IR image X_{ir}, there is no guarantee that the generated fake IR images retain the structure and content of the original RGB images X_{rgb}. To deal with this problem, we introduce a generator G_R for generating IR images into RGB images and the corresponding discriminator D_R. Also we introduce cycle-consistency loss which is defined as follows:

$$\begin{aligned}\mathscr{L}_{cyc} = {} & E\left[\left\|G_R\left(G_I\left(X_{rgb}\right)\right) - X_{rgb}\right\|_1\right] \\ & + E\left[\left\|G_I\left(G_R\left(X_{ir}\right)\right) - X_{ir}\right\|_1\right].\end{aligned} \tag{6}$$

$\mathscr{L}_{\text{cyc}}$ loss enables the G_I generated IR image to be consistent with the input real RGB image. We use the L1 norm instead of the L2 norm because the L1 norm allows the generator to generate better image edges. Specifically, we input the real RGB image X_{rgb} into the generator G_I to generate the fake IR image X'_{ir} and then use the generator G_R to generate the reconstructed RGB image from the fake IR image. We do something similar with IR images.

Now, the loss of the generator can be defined as follows:

$$\mathscr{L}_G = \mathscr{L}_{G_I} + \omega * \mathscr{L}_{cyc}, \tag{7}$$

where ω is the weight of cycle loss and ω is set to 10 as in [14]. By using this loss during adversarial training, we can generate high-quality IR images.

3.2. *The BNG Attention Module.*

Our proposed BNG attention is an efficient and lightweight attention mechanism. The BNG attention can be embedded at the end of any convolutional neural network, for the residual network ResNet-50; the end of the residual structure can be embedded. The structure of BNG is shown in Figure 2.

BNG attention consists of two submodules, as shown in Figure 2(a); the channel attention submodule can use the weight information of the trained model to highlight salient features. We obtain its scale factor from batch normalized (BN [25]) as shown in

$$B_{out} = BN\left(B_{in}\right) = \gamma\frac{B_{in} - \mu_{\mathscr{B}}}{\sqrt{\sigma_{\mathscr{B}}^2 + \epsilon}} + \beta, \tag{8}$$

where $\mu_{\mathscr{B}}$ and $\sigma_{\mathscr{B}}$ are the mean and standard deviation of mini batch $\mathscr{B}$ and γ and β are the trainable parameters used to fit the data distribution.

The formula for channel attention can be expressed as follows:

$$\mathbf{F}_1 = \text{sigmoid}\left(W_\gamma\left(BN(\mathbf{F})\right)\right), \tag{9}$$

where γ is the scale factor for each channel, and the weights are obtained as $W_\gamma = \gamma_i / \sum_{j=0}\gamma_j$. We measure the importance of each channel by applying the scale factor of BN to the channel dimension and suppressing insignificant features. Since channel attention only focuses on channel information, there is no global space-channel information interaction; to solve this problem, we design a global attention module. It can reduce information attenuation and amplify the features of global dimension interaction. Inspired by CBAM [24], the channel attention and spatial attention are connected in turn. The main structure is shown in Figure 2(b). Given the input feature map $\mathbf{F_1} \in \mathbb{R}^{C\times H\times W}$, the intermediate state F_2 and output F_3 are defined as follows:

$$\begin{aligned}\mathbf{F}_2 &= \mathbf{M}_c\left(\mathbf{F}_1\right) \otimes \mathbf{F}_1, \\ \mathbf{F}_3 &= \mathbf{M}_s\left(\mathbf{F}_2\right) \otimes \mathbf{F}_2,\end{aligned} \tag{10}$$

where $\mathbf{M}_c$ and M_s are the channel and spatial attention maps, respectively. ⊗ denotes element-wise multiplication.

The channel attention submodule uses a 3D arrangement to preserve information across three dimensions and then uses a two-layer MLP layer that amplifies the channel spatial dependencies across dimensions. The channel attention submodule is illustrated in Figure 3.

In the spatial attention submodule, to focus on the spatial information, two convolutional layers are used to fuse the spatial information. The size of the convolution kernel is set to 7 ∗ 7. Since max-pooling reduces information and has a negative influence, we remove the max-pooling operation to retain more features. The same reduction ratio γ is adopted from the channel attention submodule, same as BAM. The spatial attention submodule without group convolution is shown in Figure 4.

3.3. *Modal Mitigation Module (MMM).*

To mitigate the modal distribution, a modal mitigation module (MMM) is

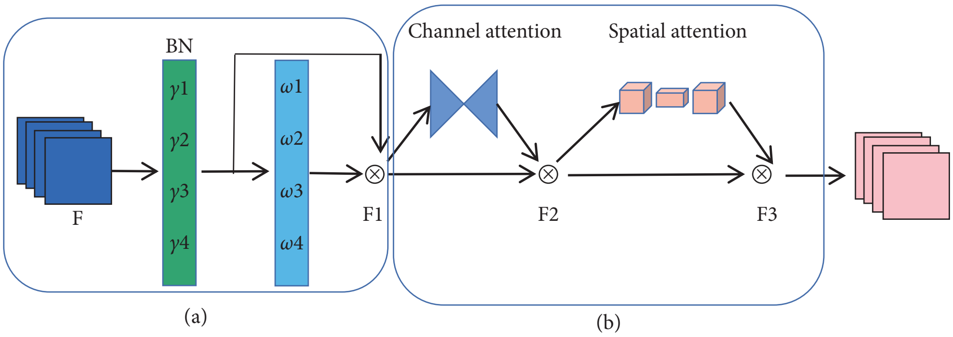

Figure 2: BNG attention.

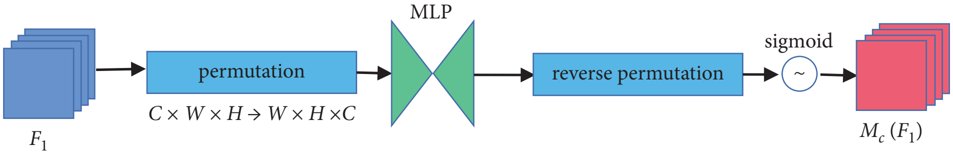

Figure 3: Channel attention submodule.

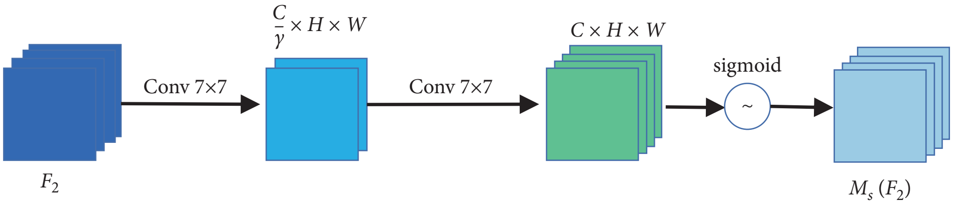

Figure 4: Spatial attention submodule.

designed. For the input image X, we denote the features extracted in the convolution block as $\mathbf{M} \in \mathbb{R}^{h\times w\times c}$ and input it into the MMM, where $h, w,$ an d c represent the height, width, and a number of channels of the feature map $\mathbf{M}$, respectively. The instance normalization (IN) is used to mitigate modal differences on a single instance [27]. Instance normalization (IN) computes the mean and variance in a single instance and reduces the difference between the two data distributions. However, using IN directly may has a negative impact on the ReID task. Because the distribution of image data has changed significantly, some identifying information may be lost.

To overcome these shortcomings, we use channel attention to guide the learning of IN, which mitigates modal differences while preserving identity information. Specifically, we input the feature into a two-layer MLP to downsample the channels and then upsample to the original number of channels and use the activate function to activate the feature as a mask to supervise the IN operation:

$$\mathbf{F} = \mathbf{m}_C \odot \mathbf{M} + \left(1 - \mathbf{m}_C\right) \odot \widehat{\mathbf{M}}, \tag{11}$$

where m_C is the channel mask, representing the identity-related channels, and $\widehat{\mathbf{M}}$ is the instance-normalized result of the input $\mathbf{M}$.

Similar to SENet [21], the method of generating a mask with channel dimension can be expressed as follows:

$$\mathbf{m}_C = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1 g\left(\mathbf{M}\right)\right)\right), \tag{12}$$

where $\mathbf{W}_1 \in \mathbb{R}^{c/r\times c}$ and $\mathbf{W}_2 \in \mathbb{R}^{c\times c/r}$ are learnable parameters in the two bias-free fully connected (FC) layers, which are followed by ReLU activation function $\delta(\cdot)$ and sigmoid activation function $\sigma(\cdot)$. $g(\cdot)$ denotes global average pooling of features. In order to balance performance and reduce the number of parameters, the downsampling ratio is set to $r = 16$.

The formula for instance normalization is defined as follows:

$$\widehat{\mathbf{M}}_j = \mathrm{IN}\left(\mathbf{M}_j\right) = \frac{\mathbf{M}_j - E\left[\mathbf{M}_j\right]}{\sqrt{\mathrm{Var}\left[\mathbf{M}_j\right] + \epsilon}}, \tag{13}$$

where $E[\cdot]$ is to calculate the mean of each dimension and $\mathrm{Var}[\cdot]$ is to calculate the standard deviation of each dimension. To avoid dividing by zero, we add ϵ to the denominator, and $\mathbf{M}_j \in \mathbb{R}^{h\times w}$ is the j-th dimension of the feature map M.

3.4. Loss Function. In this section, we will introduce the loss we used when training the generator to generate a fake IR image X'_{ir}. On the one hand, X'_{ir} should be classified to the same identity class as the corresponding X_{rgb}; on the other hand, X_{ir}' should satisfy the triplet loss [28] of the corresponding X_{rgb} identity constraint. We define these two losses as $\mathscr{L}_{cls}^{gan}$ and $\mathscr{L}_{tri}^{gan}$ and denote them in

$$\begin{aligned}\mathscr{L}_{cls}^{gan} &= \mathscr{L}_{cls}(X'_{ir}) = E_{x\in X'_{ir}}[-\log p(x)],\\ \mathscr{L}_{tri}^{gan} &= \frac{1}{2}\left[\mathscr{L}_{tri}(X'_{ir}, X_{ir}, X_{ir}) + \mathscr{L}_{tri}(X_{ir}, X'_{ir}, X'_{ir})\right],\end{aligned} \tag{14}$$

where $p(\cdot)$ is the predicted probability of belonging to the ground-truth identity; the ground-truth identity of the fake IR image X_{ir}' should be the same as that of the original RGB image X_{rgb}.

Although the generated image X_{ir} can reduce cross-modality differences, there are still large intramodality differences caused by lighting, human pose, and view. We minimize the fake IR image X'_{ir} and the real IR image X_{ir} in a shared space via identity-based classification and triplet loss. We define these two losses as $\mathscr{L}_{cls}^{feat}$ and $\mathscr{L}_{tri}^{feat}$ and denote them in

$$\begin{aligned}\mathscr{L}_{cls}^{feat} &= \mathscr{L}_{cls}(X_{ir} \cup X'_{ir}) = E_{x\in X_{ir}\cup X'_{ir}}[-\log p(x)],\\ \mathscr{L}_{tri}^{feat} &= \mathscr{L}_{tri}(X_{ir}, X'_{ir}, X'_{ir}) + \mathscr{L}_{tri}(X'_{ir}, X_{ir}X_{ir}),\end{aligned} \tag{15}$$

where $p(\cdot)$ represents the predicted probability that the input belongs to the ground-truth identity, and $\cup$ means the union sets. In summary, the overall loss of our module is shown in

$$\begin{aligned}\mathscr{L}_{\mathrm{ReID}} &= \lambda_1\mathscr{L}_G + \lambda_2\mathscr{L}_{D_I} + \lambda_3\mathscr{L}_{cls}^{gan} + \lambda_4\mathscr{L}_{tri}^{gan}\\ &\quad + \lambda_5\mathscr{L}_{cls}^{feat} + \lambda_6\mathscr{L}_{tri}^{feat},\end{aligned} \tag{16}$$

where $\mathscr{L}_G$ and $\mathscr{L}_{D_I}$ are calculated by equations (1) and (2). $\mathscr{L}_{cls}^{gan}$, $\mathscr{L}_{tri}^{gan}$, $\mathscr{L}_{cls}^{feat^I}$, and $\mathscr{L}_{tri}^{feat}$ are calculated by equations (14) and (15), respectively. Among them, $\lambda_1 = 1.0, \lambda_2 = 1.0$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$, $\lambda_5 = 1.0$, and $\lambda_6 = 1.0$.

4. Experiments

4.1. Datasets and Settings. We evaluate our model on SYSU-MM01 [10]. SYSU-MM01 is a very popular RGB-IR ReID dataset; it contains pedestrian images captured by six cameras, including two infrared cameras (camera3 and camera6), and four natural light cameras (camera1, camera2, camera4, and camera5). For each pedestrian, there are at least 400 RGB images and IR images with different poses and viewpoints. Among them, 296 IDs are used for training, 99 IDs are used for verification, and 96 IDs are used for testing. Following [29], 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 use indoor images from 1st, 2nd, 3rd, and 6th cameras. Both modes employ single-shot and multishot settings, in which 1 or 10 images of a person are randomly selected to form a gallery setting. Both modes use IR images as probe sets and RGB images as gallery sets.

Evaluation protocols: we use cumulative matching features (CMC) and mean average precision (mAP) as evaluation metrics. Following [29], the results of SYSU-MM01 are evaluated using the official code based on the mean of 10 repeated random splits of the gallery and probe set.

Implementation details: we use the ResNet-50 [30] pretrained on ImageNet as the CNN backbone, use the output of its pool5 layer as the feature map M, and use the average pooling to obtain the feature vector V. We add BNG-attention to each layer of residual blocks in ResNet-50 and MMM module after the third and fourth layers. For triplet loss, we use the FC layer to map the feature vector V into a 256-dimensional embedding vector. For classification loss, the classifier takes the feature vector V as input and includes a 256-dim fully connected (FC) layer, followed by batch normalization [25], dropout, and RELU as the middle layer, and an FC layer with the identity number logit as the output layer. The dropout rate is set at 0.5. We use PyTorch to implement the model, the images are data augmented by horizontal flipping, and the batch size is set to 72 (9 people, each of which has 4 RGB images and 4 IR images). For the learning rate, the learning rate of the generation module and discriminator module is set to 0.0002 and optimized using the Adam optimizer. We set the classifier and the embedder to 0.2 and the CNN backbone to 0.02 and optimize them by SGD.

4.2. Comparison with the Other Methods. In this section, we compare our method with several different cross-modality person ReID methods including the following methods: (1) with different structures and loss functions, two-stream [10], one-stream [10], zero-padding [10], BCTR [13], BDTR [13], D-HSME [26], and DGD + MSR [12] learned modality-invariant features and align them in feature space and (2) cmGAN [6] and JSIA [20] use the generative adversarial networks (GANs) to generate cross-modality IR images; they mitigate modal differences in pixel space. The experimental results are shown in Table 1.

In Table 1, we can find that there are various evaluation protocols, i.e., all-search/indoor-search and single-shot/multishot; firstly, for the same method, indoor-search performs better than all-search, because the images have less background variation in indoor mode, and matching is easier. Secondly, the rank scores of single-shot are lower than ones of multi-shot, but mAP scores of single-shot are higher than ones of multishot. This is because, in multishot mode, there are ten images in the gallery setting, while in single-shot, there is only one image. As a consequence, under the multishot mode, it is much easier to hit an image but difficult to hit all images. This situation is inverse under the single-shot mode.

The R1, R10, and R20 denote Rank-1, Rank-10, and Rank-20 accuracy (%). The mAP denotes the mean average precision score (%), and our model shows good performance. Compared with JSIA, our model achieves over 2.7% on Rank-1 and 2.49% on mAP in the single-shot setting of all-search mode. In the single-shot setting of indoor-search mode, our model achieves a rank-1 accuracy of 44.0% and an mAP of 52.96%. In the multishot setting of indoor search, our model achieves a rank-1 accuracy of 53.40%, and an mAP of 44.35%, which is higher than JSIA by 0.7% and 1.65%, respectively.

Table 1: Comparison of CMC (%) and mAP (%) performances with other methods on SYSU-MM01.

Methods	All-search								Indoor-search							
	Single-shot multishot								Single-shot multishot							
	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
Two-stream [10]	11.65	47.99	65.50	12.85	16.33	58.35	74.46	8.03	15.60	61.18	81.20	21.49	22.49	72.22	88.61	13.92
One-stream [10]	12.04	49.68	66.74	13.67	16.26	58.14	75.05	8.59	16.94	63.55	82.10	22.95	22.62	71.74	87.82	15.04
Zero-padding [10]	14.80	54.12	71.33	15.95	19.13	61.40	78.41	10.89	20.58	68.38	85.79	26.92	24.43	75.86	91.32	18.86
BCTR [13]	16.20	54.90	71.5	19.2	—	—	—	—	—	—	—	—	—	—	—	—
BDTR [13]	17.1	55.5	72.0	19.7	—	—	—	—	—	—	—	—	—	—	—	—
D-HSME [26]	20.7	62.8	78.0	23.2	—	—	—	—	—	—	—	—	—	—	—	—
cmGAN [6]	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
DGD + MSR [12]	37.35	83.40	93.34	38.11	43.86	86.94	95.68	30.48	39.64	89.92	97.66	50.88	46.56	93.57	98.8	40.08
JSIA-ReID [20]	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
Ours	**40.83**	**83.40**	**92.38**	**39.84**	**48.13**	**86.0**	93.67	**32.54**	**44.0**	**86.8**	**94.87**	**52.96**	**53.40**	90.52	95.70	**44.35**

Table 2: Ablation study in terms of CMC (%) and mAP (%) SYSU-MM01.

Method	SYSU-MM01			
	Single-shot all-search			
	R1	R10	R20	mAP
Baseline	34.13	78.86	90.07	33.54
B + BNG	39.60	81.95	91.60	37.93
B + MMM	39.97	82.38	92.54	39.52
B + BNG + MMM	40.83	83.40	92.38	39.84

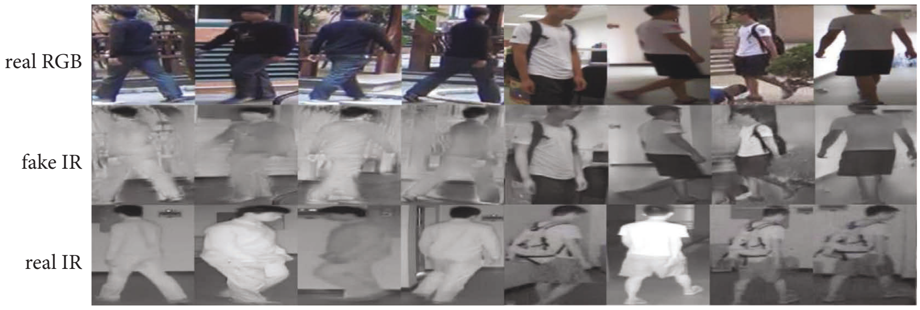

Figure 5: Fake IR images generated by our module. The fake IR images can maintain identities and contents with original real RGB ones and have IR style.

4.3. Ablation Study. In this section, we design ablation experiments to test the effectiveness of the BNG module and MMM module. Our ablation experiments are performed on the dataset SYSU-MM01 and use the single-shot setting of all-search mode.

Influence of BNG module: the results of ablation experiments for BNG attention are shown in Table 2. Compared with the baseline model (B), by adding BNG attention, the rank-1 accuracy and mAP are improved by 5.57% and 4.39%, proving the effectiveness of BNG attention.

Influence of MMM module: as shown in Table 2, the model with MMM (B + MMM) achieves a rank-1 accuracy of 39.97% and an mAP of 39.52%, which are higher than those of the baseline (B) by 5.84% and 5.98%, respectively. It is proved that our proposed MMM module has good performance.

4.4. Visualization of Generated Images. For a more intuitive understanding of the generator model, we show the learned fake IR images in Figure 5. As shown in Figure 5, the first row is the real RGB image, the middle is the fake IR image generated by the generator, and the last row is the real IR image. We can observe that fake IR images have similar content (e.g., pose and view) and maintain the identity of the corresponding real RGB images while having an IR style. Therefore, the generated fake IR images can bridge the gap between RGB and IR images and can reduce cross-modality variation in pixel space.

5. Conclusion

In this paper, we proposed a new pixel and feature alignment network (PFANet) for the RGB-IR ReID task. The model consisted of a feature extractor, a generator, and a joint discriminator. The BNG attention and the MMM module were designed in the feature extraction module. Through these two modules, the model not only mitigated modality differences but also paid attention to channel and global information. The cross-modality IR images were generated by the generator, which could bridge the gap between RGB and IR images and reduce cross-modality variation. Ablation experiments verified the effectiveness of each module. Extensive experiments on the SYSU-MM01 dataset illustrated that our model achieved state-of-the-art performance.

Data Availability

The SYSU-MM01 data used to support the findings of this study have been deposited in the "Rgb-infrared cross-modality person re-identification" repository (http://isee.sysu.edu.cn/project/RGBIRReID.html).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (grant nos. 51906217, 61906168, and 62176237), Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (grant no. LZJWZ22E090001), Zhejiang Provincial Natural Science Foundation of China under (grant no. LQ20F020024), and the Hangzhou AI Major Scientific and Technological Innovation Project (2022AIZD0061).

References

[1] X. Jin, C. Lan, W. Zeng, Z. Chen, and Li Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3143–3152, Seattle, WA, USA, June 2020.

[2] J. Song, Y. Yang, Yi-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 719–728, Long Beach, CA, USA, June 2019.

[3] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2275–2284, Salt Lake City, UT, USA, June 2018.

[4] Z. Zhong, L. Zheng, S. Li, and Yi Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–188, Munich, Germany, August 2018.

[5] Z. Zheng, X. Yang, and Z. Yu, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2138–2147, Long Beach, CA, USA, June 2019.

[6] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person re-identification with generative adversarial training," in *Proceedings of the IJCAI*, vol. 1, p. 6, Stockholm, Sweden, July 2018.

[7] H. Yi, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 57–65, Nice, France, October 2019.

[8] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 4610–4617, 2020.

[9] Z. Wang, Z. Wang, Y. Zheng, Y.-Yu Chuang, and S.'ichi Satoh, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 618–626, Long Beach, CA, USA, June 2019.

[10] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389, Venice, Italy, October 2017.

[11] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proceedings of the IJCAI*, vol. 1, p. 2, Stockholm, Sweden, July 2018.

[12] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 579–590, 2020.

[13] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, Venice, Italy, October 2017.

[15] Y. Choi, M. Choi, and M. Kim, "Stargan: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, Salt Lake City, UT, USA, June 2018.

[16] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, Honolulu, HI, USA, June 2017.

[18] Yu-J. Li, Y.-C. Chen, Y.-Yu Lin, X. Du, and Yu-C. F. Wang, "Recover and identify: a generative dual model for cross-resolution person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8090–8099, Seoul, Korea, October 2019.

[19] G.'an Wang, T. Zhang, and J. Cheng, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3623–3632, Seoul, Korea, October 2019.

[20] G.-An Wang, T. Zhang, Y. Yang et al., "Cross-modality paired-images generation for rgb-infrared person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12144–12151, 2020.

[21] J. Hu, Li Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[22] J. Park, S. Woo, L. Joon-Young, and K. In So, “Bam: bottleneck attention module,” arXiv preprint arXiv:1807.06514, 2018.

[23] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, Salt Lake City, UT, USA, June 2018.

[24] S. Woo, J. Park, L. Joon-Young, and K. In So, “Cbam: convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, Munich, Germany, August 2018.

[25] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International conference on machine learning*, pp. 448–456, PMLR, Lille, France, July 2015.

[26] H. Yi, N. Wang, and J. Li, “Hsme: hypersphere manifold embedding for visible thermal person re-identification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8385–8392, 2019.

[27] X. Pan, P. Luo, J. Shi, and X. Tang, “Two at once: enhancing learning and generalization capacities via ibn-net,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 464–479, Munich, Germany, August 2018.

[28] F. Schroff, D. Kalenichenko, and P. James, “Facenet: a unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, Boston, MA, USA, June 2015.

[29] Y. Yang, Z. Lei, J. Wang, and Z. Stan, “In defense of color names for small-scale person re-identification,” in *Proceedings of the 2019 International Conference on Biometrics (ICB)*, pp. 1–6, IEEE, Crete, Greece, June 2019.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.