Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4330804, 13 pages
https://doi.org/10.1155/2022/4330804

Research Article

Crossmodality Person Reidentification Based on Global and Local Alignment

Qiong Lou, Junfeng Li, Yaguan Qian, Anlin Sun, and Fang Lu

School of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China

Correspondence should be addressed to Yaguan Qian; qianyaguan@zust.edu.cn

Received 7 July 2021; Accepted 9 December 2021; Published 6 January 2022

Academic Editor: Hasan Ali Khattak

RGB-infrared (RGB-IR) person reidentification is a challenge problem in computer vision due to the large crossmodality difference between RGB and IR images. Most traditional methods only carry out feature alignment, which ignores the uniqueness of modality differences and is difficult to eliminate the huge differences between RGB and IR. In this paper, a novel AGF network is proposed for RGB-IR re-ID task, which is based on the idea of global and local alignment. The AGF network distinguishes pedestrians in different modalities globally by combining pixel alignment and feature alignment and highlights more structure information of person locally by weighting channels with SE-ResNet-50, which has achieved ideal results. It consists of three modules, including alignGAN module (A), crossmodality paired-images generation module (G), and feature alignment module (F). First, at pixel level, the RGB images are converted into IR images through the pixel alignment strategy to directly reduce the crossmodality difference between RGB and IR images. Second, at feature level, crossmodality paired images are generated by exchanging the modality-specific features of RGB and IR images to perform global set-level and fine-grained instance-level alignment. Finally, the SE-ResNet-50 network is used to replace the commonly used ResNet-50 network. By automatically learning the importance of different channel features, it strengthens the ability of the network to extract more fine-grained structural information of person crossmodalities. Extensive experimental results conducted on SYSU-MM01 dataset demonstrate that the proposed method favorably outperforms state-of-the-art methods. In addition, we evaluate the performance of the proposed method on a stronger baseline, and the evaluation results show that a RGB-IR re-ID method will show better performance on a stronger baseline.

1. Introduction

Person reidentification (re-ID) is a process of retrieving the same target person from multiple different camera perspectives. It is widely used in video surveillance, security, and intelligent city applications and is an important problem in video surveillance. Due to its importance, reID has attracted more and more attention in computer vision [1–6]. However, re-ID depends on good lighting conditions, which will not always be satisfied in real word. For example, in night or dark environment, the visible cameras cannot capture effective appearance. Fortunately, most surveillance cameras can automatically switch from visible (RGB) mode to near infrared (IR) mode, which provides the possibility to study the RGB-IR crossmodality matching problems in real scenes.

Although the research of RGB-IR re-ID in the real world is very meaningful, it also has the same challenges as previous work. First, there is a large difference between RGB images and IR images in channel nature. RGB images are three-channel, while IR images are single-channel. Second, the wave length range of RGB and IR images is different, which means that it is difficult to identify the same person according to the color information. In addition, different poses, illumination, and viewpoint change may even lead intraclass distances larger than interclass distances, and this is also a great challenge in RGB-IR re-ID.

The idea of global and local alignment is to obtain the global information first and then highlight more fine-grained information as a supplement through local alignment. They complement each other and can achieve better results. Inspired by the global and local alignment method

in re-ID, we introduce the idea of global and local alignment into the research of RGB-IR re-ID, so as to better solve the great challenges in RGB-IR re-ID.

In order to reduce the large crossmodality difference between RGB and IR images, the existing RGB-IR re-ID mainly uses feature alignment. However, only by matching RGB and IR images directly in the shared feature space, it is difficult to eliminate the huge difference between these two modalities. In order to solve this problem, this paper uses the Alignment Generative Adversarial Network (AlignGAN) [7], which combines pixel alignment and feature alignment, to generate the input images needed by our framework. AlignGAN consists of three components: pixel generator, feature generator, and joint discriminator. It can reduce the cross-modality difference in the pixel space and the intramodality difference in the feature space. At the same time, because of the existence of the joint discriminator, identity-consistency features can also be maintained. Using the output images of AlignGAN as the input images of our framework, we can reduce the crossmodality difference between RGB and IR images and the intra-modality difference caused by different poses, lighting, viewpoint change, and occlusion.

After considering the pixel alignment in pixel space, it is necessary to align the features in the feature space. In order to solve the problem that most existing works only focus on the global set-level alignment between the entire RGB and IR sets when learning feature alignment in feature space, which leads to some instance misalignment, we use the method of joint set-level and instance-level alignment Re-ID (JSIA-ReID) [8]. Firstly, the modality-specific features and modality-invariant features of RGB and IR images are separated by two modality-specific encoders and a modality-invariant encoder, and the modality-invariant encoder is used to map the images of different modalities into the shared feature space to perform set-level alignment. Then, new paired images are generated by exchanging modality-specific features, and the instance-level alignment is directly performed by minimizing the distance between paired images. In this way, feature alignment between RGB and IR images is well performed in feature space.

In addition, ResNet-50 is usually used in the research of RGB-IR crossmodality re-ID. However, considering that in the current deep network, more features are fused in space; with the deepening of network layers, the receptive field of feature map will gradually become larger, which makes the deep network obviously insufficient in fusing nonlocal information. In view of the above reasons, we use SE-ResNet-50 network [9] as the backbone network to automatically learn the importance of different channel features by focusing on the relationship between channels and effectively using contextual information. In essence, it is to do attention operation on the channel dimension. This attention mechanism enables the model to pay more attention to the channel features with most information, while suppressing unimportant channel features.

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

(1) We propose a network framework (AGF) based on pixel alignment and feature alignment and jointly model two alignment strategies for RGB-IR re-ID task. In the aspect of pixel alignment, we use the AlignGAN method to generate the input images required by our framework, which reduces the gap between crossmodalities. In the aspect of feature alignment, we adopt a feature alignment method combining set level and instance level to generate paired images by separating and exchanging the modality-specific features of RGB images and IR images. At the same time, we solve the problem of set-level and instance-level alignment

(2) In the aspect of network, we innovatively use SE-ResNet-50 network to replace the commonly used ResNet-50 network. It can not only obtains global features to better focus on contextual information but also automatically learns the importance of different channel features, which improves the performance of the network. To the best of our knowledge, it is the first time that SENet has been successfully applied to the research of RGB-IR re-ID

(3) Extensive experimental results on the SYSU-MM01 dataset demonstrate that the proposed model performs favorably against the state-of-the-art methods as we know and achieves a significant improvement of 5.7% rank 1 and 4.0% mAP, respectively

2. Related Works

2.1. Person Reidentification. Person re-ID aims to solve the problem of matching pedestrian images on disjoint visible cameras. It plays a great role in real-world video surveillance, public safety, and intelligent city; so, it has attracted more and more attention recently. According to different ideas, it can be divided into representation learning-based re-ID method, metric learning-based re-ID method, and local feature-based re-ID method. Representation learning-based re-ID method [10] can automatically extract robust person representation features from the original images by using convolutional neural network (CNN), so as to better verify whether person in different images are the same. The goal of metric learning-based re-ID method is to make the distance of positive sample pairs less than that of negative sample pairs. The commonly used methods to metric learning loss include contrast loss [11], triple loss [12–14], quadruplet loss [15], and trihard loss [16]. However, both representation learning-based re-ID method and metric learning-based re-ID method directly carry out feature extraction and metric distance for image retrieval globally. Although the effect of re-ID has been improved, it is difficult to make a further breakthrough. In order to solve this problem, local feature-based re-ID method is gradually emerging. Its main idea is that global features can identify different person globally, while local features can highlight more detailed information locally. The combination of the two can achieve better results. The local feature-based re-ID method mainly reidentifies by introducing image segmentation [17], pose estimation [18–20] attribute description [21], and so on. Miao et al. [20] solved the occlusion problem in re-ID by

introducing pose estimation under the assumption that both probe and gallery images may be occluded. Lin et al. [21] proposed an attribute-person recognition (APR) network, which effectively combined local features and global features by adding attribute description to re-ID and then improved the performance of re-ID. With the rapid development of deep learning, re-ID has made more and more breakthroughs, and there are gradually appeared video sequence-based re-ID method, GAN-based re-ID method, crossmodality re-ID method, and so on. The video sequence-based re-ID method considers not only the content information of images but also the motion information between frames in the video, so as to improve the accuracy of re-ID. Wu et al. [22] proposed the exploit the unknown gradually (EUG) method, which gradually selected unlabeled samples with most reliable pseudolabels to be added to the labeled data to update CNN continuously, so as to better solve the problem of one-shot video-based person re-ID. Subsequently, Wu et al. [23] set a large number of data with insufficient confidence as index data on the basis of EUG and introduced them into it, completed further optimization and achieved better results. GAN-based re-ID method can solve the problems caused by camera changes [24], datasets differences [25], different pedestrian postures [26], and so on, which is very powerful. Crossmodality re-ID method is used to study person re-ID in different modalities. At present, RGB-IR re-ID is widely studied. This paper is based on the research on RGB-IR re-ID and inspired by the combined global and local feature alignment in the local feature-based re-ID method, the person re-ID carried out globally through the joint pixel alignment and feature alignment, and the person structure details are highlighted locally through the operation of adding attention mechanism to the channel through SENet.

2.2. RGB-IR Person Reidentification. This paper focuses on the crossmodality alignment, which has already been widely studied in the general computer vision field. For example, for the retrieval between text and image, Zhao et al. [27] first transformed the multiview problem into a single view hash problem through an end-to-end deep learning framework. For the retrieval between vision and audio, Wu et al. [28] proposed a dual attention matching (DAM) module, which queries the local feature of another modality in a bidirectional way through the global feature of one modality, and pays attention to the global and local feature alignment at the same time. However, these studies are retrieval in other fields of computer vision and cannot be directly applied to RGB-IR re-ID. RGB-IR person re-ID attempts to match RGB and IR images of pedestrians under disjoint cameras. In addition to the recognition difficulties caused by different poses, illumination, viewpoint change, and occlusions in traditional re-ID, the crossmodality difference between RGB and IR images brings new challenges to RGB-IR re-ID. In [29], Wu et al. collected a large RGB-IR crossmodality dataset named SYSU-MM01. This paper not only discussed three different network structures but also proposed a deep zero-padding method, which was used to train one-stream network towards automatically evolving domain-specific nodes in the network. Besides one-stream method, two-stream method is also very effective. In [30], Ye et al. proposed a two-stage framework including feature learning and metric learning (TONE+HCML). Firstly, the features of two modalities were extracted separately, and the shared layer was used to obtain unified features, and then metric learning was used to further improve the performance. In [31], Ye et al. proposed a dual-path end-to-end feature learning framework, which consists of two parts, one is a dual-path network for feature extraction, and the other is a bidirectional dual-constrained top-ranking loss for feature learning. Compared with HCML, it has the advantage of direct end-to-end learning without additional metric learning. In [32], Ye et al. further improved the bidirectional dual-constrained top-ranking loss based on [31] to bidirectional center-constrained top-ranking loss. Using anchor to centers instead of anchor to samples comparison can not only reduce the computational cost, but also preserve the properties to handle both crossmodality and intramodality variations. In [33], Ye et al. proposed a novel modality-aware collaborative ensemble (MACE) learning method, which handled the modality-discrepancy in both feature level and classifier level. In feature level, MSTN learns better features by mining shareable information in middle-level convolution blocks, which is very important for fine-grained recognition task. In classifier level, on the one hand, both modality-sharable and modality-specific classifiers are introduced to guide the feature learning; on the other hand, in order to make better collaborative ensemble learning among different classifiers, ensemble learning strategy and collaborative learning strategy are introduced. Recently, many studies started from GAN, which provided a new idea for RGB-IR re-ID. In [34], Dai et al. introduced a crossmodality generative adversarial network (cmGAN), which reduces the crossmodality difference between RGB and IR images. Most methods mainly use feature alignment to make up the gap between RGB and IR images. Recently, new researches have taken into account both pixel alignment and feature alignment. In [35], Wang et al. used image-level subnetworks to convert visible (infrared) images into infrared (visible) images to reduce modality discrepancy and then used feature-level sub-networks to embed features to reduce appearance difference. In [7], Wang et al. innovatively proposed an end-to-end alignment generative adversarial network (AlignGAN) based on pixel-level and feature-level constraints. This model is composed of a pixel generator, a feature generator, and a joint discriminator. By playing min-max game among the three components, it can not only alleviate crossmodality and intramodality differences but also maintain identity consistency. Inspired by the idea of [7], we combined pixel alignment and feature alignment to generate new IR images from RGB images in the dataset SYSU-MM01 through AlignGAN. The newly generated IR images and the original SYSU-MM01 dataset constitute a new dataset, which is the dataset to be used in our entire framework. So, starting from the dataset, we carry out pixel alignment and feature alignment to make up for the gap between crossmodalities. Recently, Ye et al. [36] proposed a Homogeneous Augmented Tri-Modal (HAT) learning

method, which solved the trimodal feature learning from both multimodal classification and multiview retrieval perspectives. They put forward that learning from grayscale images generated from visible images can effectively enforces the network to mine structure relations across multiple modalities. As far as we know, this work has achieved the best results at present. This provides a new idea for bridging the modality gap in VI-ReID, which is worth learning.

2.3. Disentangled Representation Learning. Disentangled representation learning aims to extract the necessary parts from the data to form a more meaningful representation. In the single-modality re-ID task, the application of disentangled representation learning is generally to extract illumination-invariant features [37] or to separate foreground, background, and posture factors [38]. However, in the task of RGB-IR crossmodality re-ID, due to both crossmodality and intramodality discrepancies, this brings particular challenges to disentangle common identity information and the remaining attributes from RGB and IR images. In [39], Choi et al. proposed a hierarchical crossmodality disentanglement (Hi-CMD) method, which automatically disentangles ID-discriminative factors and ID-excluded factors from RGB and IR images to reduce crossmodality difference and intramodality difference. In [40], considering that the existing research embedded different modalities into a common feature space to reduce crossmodality difference, but ignored the specific features of different modalities, Lu et al. proposed a crossmodality shared-specific feature transfer algorithm (cm-SSFT) to solve the above problems. Firstly, input the images into the two-stream feature extractor to obtain the shared and specific features. Then, the intramodality and intermodality affinities are modeled based on the shared-specific transfer network (SSTN) and exploit the potential of specific characteristics of different modalities. In [8], Wang et al. proposed to decompose RGB images and IR images into modality-specific features and modality-invariant features in order to solve the problem that some instances are out of alignment between RGB images and IR images. By separating and exchanging modality-specific features between them, paired images that remain the same modality-invariant but have different modality-specific are generated, so that instance-level alignment can be directly performed by minimizing the distance between each pair of paired images. In recent years, the disentangled representation learning method has been widely used in RGB-IR re-ID. The specific information in different modalities can be mined better by disentangle representation, which provides a powerful weapon for reducing crossmodality difference and intramodality difference. The whole framework of this paper is based on [8], which generates crossmodality paired images and simultaneously executes feature alignment of global set-level and fine-grained instance-level, which is combined with previous pixel alignment of the data to reduce crossmodality and intramodality difference more effectively. Moreover, the recent disentangled representation learning is also very popular in audio-visual events. Wu and Yang [41] solved the problem of modality uncertainty caused by audio-visual asynchrony by exchanging cross-modality signals of different video and audio, and introduced contrastive learning to introduce temporal difference into aggregated features, so as to better temporal localization performances. The expansion of disentangled representation learning in audio and video events further illustrates the effectiveness and popularity of separating and exchanging modality-specific features.

2.4. Deep Architectures. Convolutional neural networks have achieved great success in visual recognition, and many attempts have been made on the basis of the original convolutional neural networks. VGGNets [42] and Inception [43] show that increasing the depth of network can significantly improve the ability of network learning. ResNet [44] proved that by introducing residual blocks, we can learn deeper and stronger network, and the effects of the network will become better. ResNeXt [45] and exception [46] used block convolution to increase cardinality. Deformable convolution [47, 48] designed deformable convolution to enhance geometric modeling ability. SENet [9] learns the importance of different channels by adding different weights to the channels. Since we consider the global context information and the connection among different channels, thus SE-ResNet-50 is adopted as our CNN backbone.

3. AGF NetWork

In this section, we introduce the details of AGF network proposed for RGB-IR crossmodality re-ID. As shown in Figure 1, our proposed AGF consists of an AlignGAN module (A), a crossmodality paired-image generation module (G), and a feature alignment module (F). Firstly, according to [7], we can generate IR images which can be confused with the real ones. The purpose of this is to eliminate the huge crossmodality gap between RGB-IR crossmodality images directly from the pixel level perspective. This work is the preliminary work of our entire framework, which can be simply understood as image preprocessing. Then, the IR images generated by AlignGAN and all the images in SYSU-MM01 dataset are transmitted to the framework as input images. Through the Generation Module (G) and the Feature Alignment Module (F), the unpaired images are finally generated into paired images by separating and exchanging features, so as to simultaneously perform global set-level and instance-level alignment [8]. In addition, we also use SE-ResNet-50 network [9] as the backbone of CNN, paying attention to channels, learning the importance of different channels, and improving the network performance.

3.1. AlignGAN Module. In this module, our aim is to generate the required images as the input images of the whole framework according to the method AlignGAN provided by [7], reducing the crossmodality gap from the dataset source. As shown in Figure 2, AlignGAN consists of three parts: pixel alignment module (P), feature alignment module (F), and joint discriminator module (D). The function of the pixel alignment module is to reduce the crossmodality difference between RGB and IR images and to convert real RGB

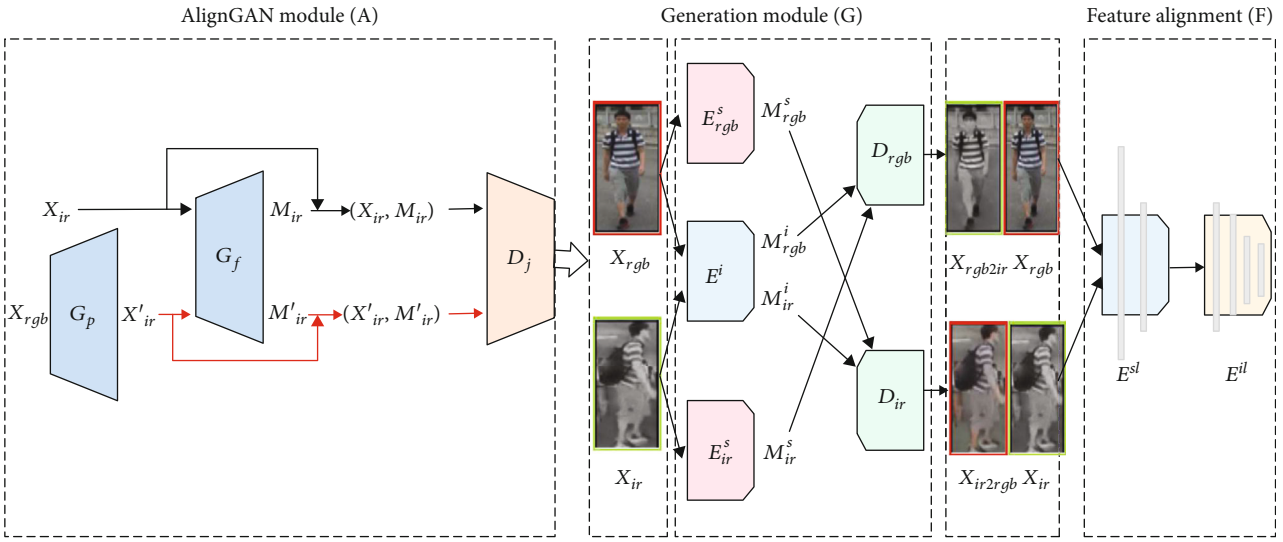

FIGURE 1: Our proposed framework (AGF) consists of a AlignGAN module (A), a crossmodality paired-images generation module (G), and a feature alignment module (F). A generates fake IR images from RGB images to reduce the crossmodality gap, and then we transfer it to the framework as input images together with all the images in the SYSU-MM01 dataset. G disentangles images to modality-specific and modality-invariant features and then decodes from the exchanged features. F uses an encoder whose weights are shared with modality-invariant encoder to perform set-level alignment and then further performs instance-level alignment by minimizing distance between each pair images.

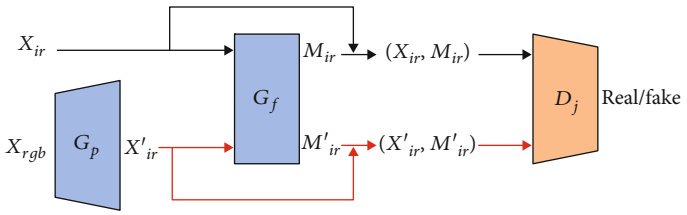

FIGURE 2: Framework of the AlignGAN model. It consists of a pixel alignment module (P), a feature alignment module (F), and a joint discriminator module (D). The P can generate fake IR images X'_{ir} to alleviate the crossmodality difference, the F can alleviate the intramodality difference, and the D can obtain identity-consistent features by making P and F learn from each other and penalizing negative pairs which are not real or belong to different identities.

images into fake IR images by pixel alignment module, so as to reduce the gap between RGB and IR crossmodality. The generated fake IR images can keep the original RGB identity information unchanged while having IR style. The function of feature alignment module is to reduce the intramodality difference and encode the real infrared images X_{ir} and the fake infrared images X'_{ir} generated by pixel alignment module into the shared space, so as to reduce the intramodality difference caused by different poses, viewpoint change, lighting, and so on. The function of the joint discriminator module is to discriminate the authenticity of the input pictures while maintaining identity consistency. Taking the image-feature pair (X, M) as the input of the discriminator, the output is either 0 or 1, where 0 means false and 1 means true, and only when the paired real infrared images and infrared features are taken as the input, 1 will be output.

3.2. Crossmodality Paired-Image Generation Module. In this module, images are decomposed into modality-specific features and modality-invariant features, and paired images are generated by exchanging modality-specific features of unpaired images. Two paired images have the same modality-invariant features such as pose, but with different modality-specific features such as clothing colors. The crossmodality paired-images generation module is composed of three encoders E^i, E^s_{rgb}, and $E^s_{ir}$and two decoders D_{rgb} and D_{ir}.

The encoders are responsible for disentangling the features of RGB and IR images. Specifically, the modality-invariant encoder E^i responsible for learning the content information of RGB images and IR images, the modality-specific encoders E^s_{rgb}, and E^s_{ir} are responsible for learning the style information of RGB images and IR images, respectively. The modality-specific features M^s_{rgb} and M^s_{ir} of RGB images and IR images are shown in equation (1), and the modality-invariant features $M^i_{\text{rgb}}$and $M^i_{ir}$are shown in equation (2).

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

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

And the decoders are responsible for generating paired images by exchanging modality-specific features. Specifically, IR images $X_{\text{rgb2}ir}$ are generated by using the content features M^i_{rgb} of real RGB images and the style features M^s_{ir} of real IR images, which contain both the content information from RGB images and the style information from IR images and are paired with real RGB images. Similarly, we can also generate RGB images $X_{ir2\text{rgb}}$ to be paired with real IR images X_{ir}. The whole process can be expressed by equation (3).

$$X_{ir2\text{rgb}} = D_{ir}\left(M^i_{ir}, M^s_{\text{rgb}}\right), X_{\text{rgb2}ir} = D_{\text{rgb}}\left(M^i_{\text{rgb}}, M^s_{ir}\right). \tag{3}$$

In order to generate more realistic paired images, [8] has made three efforts, including the following:

Firstly, construct a reconstruction loss to make the disentangled features can reconstruct their original images, where $\|\bullet\|_1$ is L1 distance.

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

Secondly, a cycle-consistency loss is introduced to guarantee that the generated images can keep the original modality-invariant features and be translated back to the original version. The cycle-consistency loss is shown in equation (5), where the $X_{ir2\text{rgb}2ir}$ and $X_{rgb2ir2\text{rgb}}$ are cycle-reconstructed images, which are specifically expressed in equation (6).

$$L_{\text{cyc}} = \left\|X_{\text{rgb}-}X_{rgb2ir2\text{rgb}}\right\|_1 + \left\|X_{ir} - X_{ir2\text{rgb}2ir}\right\|_1, \tag{5}$$

$$\begin{aligned} X_{ir2\text{rgb}2ir} &= D_{ir}\left(\mathrm{E}^{\mathrm{i}}_{\text{rgb}}\left(X_{ir2\text{rgb}}\right), E^s_{ir}\left(X_{\text{rgb}2ir}\right)\right)X_{\text{rgb}2ir2\text{rgb}} \\ &= D_{\text{rgb}}\left(E^i_{ir}\left(X_{\text{rgb}2ir}\right), E^s_{\text{rgb}}\left(X_{ir2\text{rgb}}\right)\right). \end{aligned} \tag{6}$$

Finally, due to the introduction of reconstruction loss and cycle-consistency loss, the images will be blurred; so, adversarial loss is applied to make the images more realistic. Specifically, the discriminators Dis_{rgb} and Dis_{ir} are used to distinguish the real images and the generated images on RGB and IR modalities, while encoders and decoders are used to make the real images indistinguishable from the generated images, so as to achieve the purpose of making the generated images more realistic. The expression of GAN loss is shown in equation (7).

$$\begin{aligned} L_{\text{gan}} = &E\left[\log Dis_{\text{rgb}}\left(X_{\text{rgb}}\right) + \log\left(1 - Dis_{\text{rgb}}\left(X_{ir2\text{rgb}}\right)\right)\right] \\ &+ E\left[\log Dis_{ir} + \log\left(1 - Dis_{ir}\left(X_{rgb2ir}\right)\right)\right]. \end{aligned} \tag{7}$$

3.3. Feature Alignment Module

3.3.1. Set-Level Feature Alignment. In the crossmodality paired-image generation module, modality-invariant encoder E^i is trained to explicitly remove modality-specific features. The weight of set-level encoder E^{sl} is shared with modality-invariant encoder E^i; thus, it can map different modality images with removed modality-specific features into the shared feature space and reduce modality difference between set level.

3.3.2. Instance-Level Feature Alignment. The instance-level encoder E^{il} aligns the paired images in pairs to directly solve the problem of instance imbalance. Specifically, the instance-level encoder E^{il} maps the set-level aligned features M into a new feature space T and then aligns every two crossmodality paired-images by minimizing their Kullback-Leibler Divergence. The loss of instance-level feature alignment is shown in equation (8).

$$\begin{aligned} L_{\text{align}} = &E_{(X_1,X_2)\in\left(X_{ir},X_{ir2\text{rgb}}\right)}[KL(p_1\|p_2)] \\ &+ E_{(X_1,X_2)\in\left(X_{rgb2ir},X_{\text{rgb}}\right)}[KL(p_1\|p_2)], \end{aligned} \tag{8}$$

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

In addition, an identity-discriminative feature learning includes classification loss and triplet loss to overcome the intramodality difference.

$$L_{\text{cls}} = E_{v\in V}(-logp(v)), \tag{9}$$

$$L_{\text{triplet}} = E_{v\in V}\left[m - D_{v_a,v_p} + D_{v_a,v_n}\right]_+, \tag{10}$$

where V represents feature vectors $V = E^{il}(E^{sl}(X))$, $p(\cdot)$ is the predicted probability predicted by the classifier C that the input feature vector belongs to the groundtruth, 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, and m is a margin parameter and $[X]_+ = \max(0, x)$.

Thus, the overall loss can be formulated as in equation (11).

$$L = \lambda_{\text{cyc}}L_{\text{cyc}} + \lambda_{\text{gan}}L_{\text{gan}} + \lambda_{\text{align}}L_{\text{align}} + \lambda_{\text{reid}}\left(L_{\text{cls}} + L_{\text{triplet}}\right). \tag{11}$$

We set $\lambda_{cyc} = 10$, $\lambda_{\text{gan}} = 1$, and $\lambda_{\text{rid}} = 1$, and λ_{align} is decided by grid search.

3.4. Network Module. In the traditional convolutional neural network, each convolution kernel only operates on the local receptive field; so, each unit of convolution output cannot use the contextual information outside the region. In fact, every pixel of a picture may relate to each other, and the network of local receptive field ignores the related information between global pixels, which makes the experimental results unsatisfactory. However, SENet cannot only obtain global features and make effective use of contextual information but also make information interaction among channels

possible by adding processing between two adjacent layers, so that the model can automatically learn the importance of different channel features and further improve the network accuracy. A detailed description of SE-ResNet-50 is given in Table 1 for a specific example of the SENet architecture.

The main purpose of SE module is to improve the sensitivity of the model to the characteristics of channel. This module is lightweight and can be applied to the existing network structure, which can improve the performance with only increase a small amount of calculation. It consists of two parts: squeeze part and excitation part. Squeeze compresses the feature map with the size of $H \times W \times C$ to $1 \times 1 \times C$, which represents the corresponding global distribution of feature channels, so that the lower layers can also obtain the global receptive field, thus obtaining global features. With the information of channels, it is necessary to establish the correlation between channels. The excitation part predicts the importance of each channel and adds it to the corresponding channel, so as to weight different channels. Essentially, SE module does attention or gating operation on the channel dimension. This attention mechanism allows the model to pay more attention to the channel features with the most of information, while suppressing the unimportant channel features. A diagram illustrating the SE block structure is shown in Figure 3 [9].

In the RGB-IR crossmodality scene, the color information of pedestrian clothing is almost unavailable. Thus, the model should pay more attention to the structure information of person so as to better complete the RGB-IR re-ID task. SE module can obtain the global features of RGB and IR images through Squeeze operation, and weighting channels through excitation operation can enlarge the dependence on texture/shape features and restrain the dependence of model on color features. Therefore, adding SE module to the network is helpful to extract finer-grained contour features of pedestrians and enforce the ability of the network to mine structure relations across RGB and IR modalities, making it robust to color variations. In view of the above reasons, the SE-ResNet-50 network [9] obtained by applying SE module to ResNet-50 network is used as CNN backbone, so as to obtain the global understanding and channel relationship of images in RGB-IR re-ID.

Table 1: (Left) ResNet-50. (Right) SE-ResNet-50. The shapes and operations with specific parameter settings of a residual building block are listed inside the brackets, and the number of stacked blocks in a stage is presented outside. The inner brackets following by fc indicates the output dimension of the two fully connected layers in an SE module.

Output size	ResNet-50	SE-ResNet-50
112×112	Conv, 7×7, 64, stride 2	
	Max pool, 3×3, stride 2	
56×56	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \\ f_c, [16, 256] \end{bmatrix} \times 3$
	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \\ f_c, [32, 512] \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 6, 256 \\ \text{conv}, 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \\ f_c, [64, 1024] \end{bmatrix} \times 6$
7×7	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \\ f_c, [128, 2048] \end{bmatrix} \times 3$
1×1	Global average pool, 1000-d f_c, softmax	

4. Experiment

4.1. Dataset and Evaluation Protocol

4.1.1. Dataset. We evaluate our model on the standard benchmark SYSU-MM01.

SYSU-MM01 [29] is a popular RGB-IR re-ID dataset, which includes 491 identities from 4 RGB cameras and 2 IR cameras. The training set includes 19659 RGB images and 12792 IR images of 395 persons, and the test set includes 96 persons. According to [29], there are two test modes, namely, 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, and 6th cameras are used. For both modes, the single-shot and multishot settings are adopted, respectively, in which single-shot setting randomly selects one image of a person to form the gallery set, while multishot setting randomly selects ten images of a person to form the gallery set. In both modes, IR images are used as probe set and RGB images as gallery set.

In this paper, the dataset used in our training consists of two parts, including all the images in SYSU-MM01 dataset and IR images converted from RGB images taken by the four RGB cameras including 1st, 2nd, 4th, and 5th in SYSU-MM01 dataset by the AlignGAN model. Our dataset is shown in Table 2.

4.1.2. Evaluation Protocols. The cumulative matching characteristic (CMC) and mean average precision (mAP) are used as evaluation metrics. After [29], the results of SYSU-MM01 are evaluated with the official code, which was based on the average of 10 times repeated random split of gallery and probe set.

4.2. Implementation Details

4.2.1. Network Architecture. In the generation module G, [8] constructs modality-specific encoders, which has two strided convolution layers; then, a global average pooling layer and a fully connected layer. For decoders, four residual blocks with

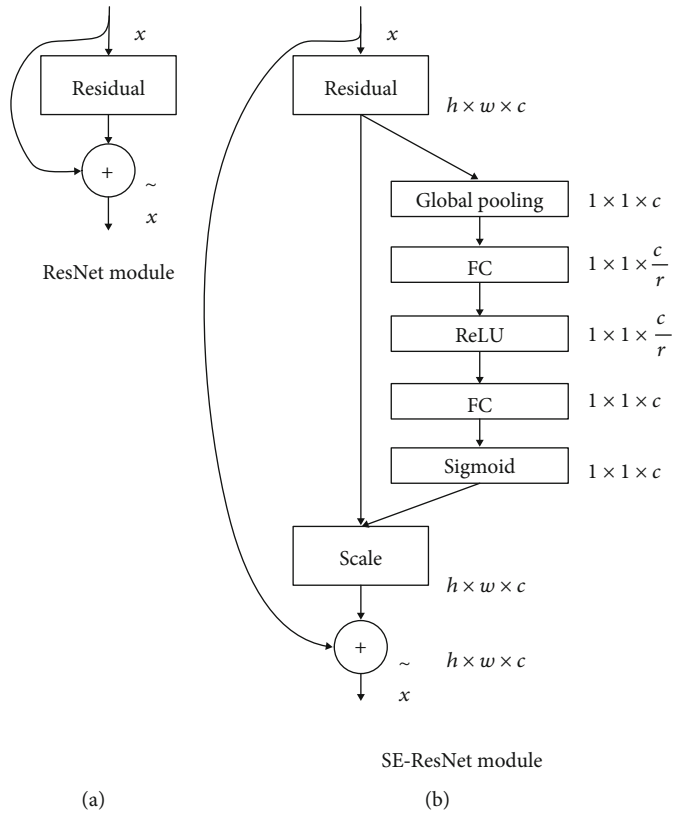

Figure 3: The schema of the original residual module (a) and the SE-ResNet module (b). (this picture is derived from [9]).

Table 2: Overview of our dataset. The red part represents IR images converted from RGB images taken by 1st, 2nd, 4th, and 5th RGB cameras in SYSU-MM01 dataset through the AlignGAN model, while the black part represents all images in SYSU-MM01 dataset, including RGB images taken by 1st, 2nd, 4th, and 5th RGB cameras and IR images taken by 3rd and 6th IR cameras.

Cam	Location	(in/out) door	Lighting	ID#	RGB#/ID	IR#/ID
1	Room 1	Indoor	Bright	259	400+	—
2	Room 2	Indoor	Bright	259	400+	—
4	Gate	Outdoor	Bright	493	20	—
5	Garden	Outdoor	Bright	502	20	—
1′	Room 1	Indoor	Dark	259	—	400+
2′	Room 2	Indoor	Dark	259	—	400+
4′	Gate	Outdoor	Dark	493	—	20
5′	Garden	Outdoor	Dark	502	—	20
3	Room 2	Indoor	Dark	486	—	20
6	Passage	Outdoor	Dark	299	—	20

adaptive instance normalization (AdaIN) and two upsampling with convolution layers are used. Here, the parameters of AdaIN are dynamically generated by modality-specific features. In GAN loss, discriminator and LSGAN [49] are used to stabilize training. In the feature learning module F, we use SE-ResNet-50 as our CNN backbone, the first two layers of the SE-ResNet-50 as our set-level encoder, and the remaining layers as our instance-level encoder.

4.2.2. Training Strategies. Our model is implemented with Pytorch. GAN's input images size is set to [128, 64], and Reid's input images size is set to [256,128]. Application of random horizontal flip for data enhancement: the whole training process is set to 649 epochs, the PK sampling parameters of GAN are set to $p = 3$ and $k = 3$, and those of Reid are set to $p = 16$ and $k = 4$. Adam optimizer with hyperparameters of GAN's $\beta = [0.5, 0.999]$, weight_decay = 0.0001, Reid's $\beta = [0.9,0.999]$, and weight_decay = 0.0005 is adopted for optimization. In the crossmodality paired-image generation module, the learning rates of the generator and the discriminator are both set to 0.0001, and in the feature alignment module, the learning rates of set-level alignment and instance-level alignment are set to 0.00045.

4.3. Comparison with State-of-the-Arts. To prove the effectiveness of our method, we compare it with most related methods, including most advanced RGB-IR crossmodality re-ID methods, zero padding [29], BCTR [30], BDTR [31], eBDTR [32], cmGAN [34],D^2RL [35], MAC [50], and AlignGAN [7], which considering both pixel-level differences and feature-level differences and JSIA-ReID [8], which performs set-level and instance-level alignment simultaneously, and some feature learning methods, HOG [51], LOMO [52], single-stream, and double-stream networks [29]. The experimental results are shown in Table 3. Our method is obviously superior to most of the most existing state-of-the-arts. On SYSU-MM01 dataset, compared with JSIA-ReID [8], we always outperform it, with a matching rate of over 5.7% on rank 1 and over 4% on mAP. Specifically, we

Table 3: Comparison with the state-of-the-arts on SYSU-MM01 dataset. The R1, R10, and 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				Multishot				Single-shot				Multishot			
	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	3.53
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
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
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
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	—	—	—	—	—	—	—	—	—	—	—	—
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
eBDTR	27.8	67.3	81.3	28.4	—	—	—	—	32.5	77.4	89.6	42.5	—	—	—	—
D^2RL	28.9	70.6	82.4	29.2	—	—	—	—	—	—	—	—	—	—	—	—
MAC	33.3	79.0	90.1	36.2	—	—	—	—	36.4	62.4	71.6	37.0	—	—	—	—
AlignGAN	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
JSIA-ReID	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
Ours	43.8	86.2	93.4	40.9	53.8	92.1	96.7	33.8	46.4	90.5	96.7	55.3	58.2	95.1	98.3	44.9

Table 4: Comparison with different variants of CNNs on SYSU-MM01 dataset and our dataset under the all-search mode. The R1, R10, and R20 denote rank 1, rank 10, and rank 20 accuracies (%), respectively. The mAP denotes mean average precision score (%).

Dataset	CNNs	All-search							
		Single-shot				Multishot			
		R1	R10	R20	mAP	R1	R10	R20	mAP
SYSU-MM01	ResNet-50	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5
	IBN-ResNet-50	38.8	81.5	89.9	36.5	48.4	88.0	94.1	29.6
	ResNext-50	37.9	81.0	89.7	36.0	47.1	86.9	94.3	29.1
	SE-ResNet-50	39.6	82.6	91.1	37.1	48.5	88.6	95.0	30.1
Our dataset	ResNet-50	41.2	83.6	91.5	38.9	50.7	89.2	95.2	31.8
	IBN-ResNet-50	40.3	81.8	90.3	37.5	50.6	88.0	94.1	30.5
	ResNext-50	40.9	83.1	90.9	38.7	49.8	88.5	94.9	31.6
	SE-ResNet-50	43.8	86.2	93.4	40.9	53.8	91.9	96.7	33.7

achieved rank 1 = 43.8% and mAP = 40.9% on SYSU-MM01 dataset. This demonstrates the effectiveness of our model for the RGB-IR re-ID task.

4.4. Model Analysis

4.4.1. CNN Analysis. For a fair comparison, most frameworks adopt the ResNet-50 as CNN backbone in the research of RGB-IR re-ID task. However, because CNNs play an important role in the research and development of person reidentification, we believe that there exist still a large space for the study of CNN backbone network in the field of crossmodality person reidentification. Therefore, we choose ResNext-50, IBN-ResNet-50, SE-ResNet-50, and ResNet-50 for comparative analysis. The reason for choosing them is that these networks are not only closely related to ResNet-50 but also have their own characteristics; so, it is worth exploring and analyzing. Firstly, ResNext-50 introduces cardinality into the original ResNet to change one-way convolution into multiway parallel convolution and achieves the goal of improving the accuracy of the model without increasing the complexity of parameters by grouped convolutions. Secondly, IBN-ResNet-50 is obtained by applying IBN-Net to ResNet-50. IBN-Net inherits the advantages of IN and BN and can easily learn the visual expression of shallow network and the content information of deep network, and it is very suitable for crossdomain transfer learning. Finally, SE-ResNet-50 is generated by embedding SE module into ResNet-50. Unlike ordinary convolutional neural networks, which only pay attention to spatial information and ignore channel information, SENet adds attention mechanism to channels by modeling the correlation between feature channels and enhances the accuracy by strengthening important features. The comparison results are shown in Table 4.

Table 5: Comparison with different variants of our methods on SYSU-MM01 dataset under the single-shot and all-search mode. The R1, R10, and R20 denote rank 1, rank 10, and rank 20 accuracies (%), respectively. The mAP denotes mean average precision score (%).

Methods			Single-shot			
Index	AlignGAN	SE-ResNet-50	R1	R10	R20	mAP
1	×	×	38.1	80.7	89.9	36.9
2	×	√	39.6	82.6	91.1	37.1
3	√	×	41.2	82.2	91.1	38.9
4	√	√	43.8	86.2	93.4	40.9

The comparison results in Table 4 show that the experimental results of SE-ResNet-50 are obviously superior to the other three networks. Taking all-search and single-shot mode as an example, the rank 1 of SE-ResNet-50 is 39.6%, and the mAP is 37.1% on SYSU-MM01 dataset, and the rank 1 of SE-ResNet-50 is 43.8%, and mAP is 40.9% on our dataset. The results on two datasets show that SE-ResNet-50 has the best effect. This shows that SE-ResNet-50 as the CNN backbone can make our framework play the best effect, and it also shows that SE-ResNet-50 adds different weights to different channels, which is very effective for RGB-IR re-ID research.

4.4.2. Ablation Study. In order to further analyze the effectiveness of our proposed methods, we conducted an ablation experiment; that is, without adding any modules we wanted, we added the methods we wanted to study one by one. For example, the AlignGAN module is added separately to preprocess the data, and the SE-ResNet-50 network is added separately to pay attention to the connection between different channels. Finally, all modules are added to the basic network for experiments, and the effectiveness of the research method is verified by the experimental results.

As shown in Table 5, when all modules are removed, that is, baseline is the framework network of JSIA-ReID [8], and rank 1 score is 38.1%. The rank 1 score is 39.6% after adding SE-ResNet-50 module alone. The AlignGAN module is used alone to convert RGB images into IR images, which are added to the classic SYSU-MM01 dataset to form a new dataset, and the rank 1 score is 41.2%. It is proved that the addition of these two modules has obvious effects on RGB-IR crossmodality re-ID. Finally, using both the AlignGAN module and the SE-ResNet-50 module at the same time, our method obtained a rank 1 score of 43.8%, which shows that the AlignGAN module and the SE-ResNet-50 can be effectively combined, and it has a good effect on the matching of RGB-IR crossmodality re-ID.

4.5. Baseline Analysis. In order to verify whether a RGB-IR re-ID method will show stronger performance on a stronger baseline, in this subsection, we evaluate the performance of the RGB-IR re-ID method proposed in this paper configured with a stronger AGW baseline [53]. The evaluation results are shown in Table 6.

Table 6: Evaluation of the method proposed in this paper configured with different baseline on the large-scale SYSU-MM01 dataset. The R1, R10, and R20 denote rank 1, rank 10, and rank 20 accuracies (%), respectively. The mAP denotes mean average precision score (%).

Methods	All-search (single-shot)				Indoor-search (single-shot)			
	R1	R10	R20	mAP	R1	R10	R20	mAP
JSIA-ReID	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
AGW	47.5	—	—	47.7	54.2	—	—	63.0
Ours (JSIA-ReID)	43.8	86.3	93.4	40.9	46.4	90.5	96.7	55.3
Ours (AGW)	50.9	88.3	95.0	49.8	57.1	91.5	97.1	64.2
MACE	51.6	87.3	94.4	50.1	57.4	93.0	94.5	64.8

Figure 4: We show the fake IR images ir' generated from RGB images by AlignGAN, which not only has IR style but also can maintain identities and contents (such as views, poses) of the corresponding real RGB images. It also shows the achievements of the crossmodality paired-image generation module, which can not only stably generate paired images with given real RGB and IR images but also stably generate paired images with fake IR images ir'.

The results in Table 6 show that the method proposed in this paper configured with a stronger baseline can achieve more significant improvement. On the large-scale SYSU-MM01 dataset, compared with the method proposed in this paper (baseline is JSIA-ReID), the method proposed in this paper (baseline is AGW) achieves 8.9% mAP improvement under al-search query setting and 8.3% mAP improvement under indoor-search query setting. Our method (baseline is AGW) has even achieved comparable performance as MACE [33], and our method is relatively simple and easy, without the need to design complex networks and elaborate classifier ensemble learning. According to the results in Table 6, we can draw two important conclusions: (1) the RGB-IR re-ID method proposed in this paper is very effective. RGB images are converted into fake IR images by the

AlignGAN method for pixel alignment, which can effectively make up the crossmodality gap between RGB and IR images. Adding SE module to the network can not only obtain the global features of the images but also weight the channels, which is helpful to extract the structural features of pedestrians and plays an important role in RGB-IR re-ID. (2) RGB-IR re-ID method will show stronger performance on a stronger baseline. AGW baseline is very powerful, which provides important insights for the further study of RGB-IR re-ID.

4.6. Visualization of Images. In order to better show the effects of pixel alignment module (A) and crossmodality paired-image generation module (G), in this part, we visualize the fake IR images generated by pixel alignment module and crossmodality paired images generated by crossmodality paired-image generation module. From Figure 4, firstly, we can see that the fake infrared images generated by the pixel alignment module have infrared style and keep the content information (view, posture, etc.) of the corresponding real RGB image. Therefore, the generated fake infrared images can reduce the huge crossmodality changes between RGB and infrared images. Secondly, we can see that when a person's crossmodality unpaired image is given, whether it is a real RGB image and IR image, or a fake IR image generated by the AlignGAN module, our method can stably generate crossmodality paired images.

But objectively speaking, the generated images are not clear enough. As shown in Figure 4, for example, the legs of the ir' image generated by person A under the umbrella are very blurry, and the contour of the whole person is very blurry in the ir' image generated by person B under the background of steps, which intuitively shows the influence of occlusion and complex background on RGB-IR re-ID. It also shows that in the research of RGB-IR re-ID, besides the huge crossmodality differences among different modalities, the problems of occlusion, viewpoint change, and complex background faced by traditional re-ID are still huge problems to be solved urgently. We can learn from the mature methods in the field of re-ID, such as attribute-person recognition (APR) proposed by Lin et al. [21] and re-ID based on pose estimation proposed by Miao et al. [20], and apply these methods to RGB-IR re-ID, so as to solve the problems of occlusion and complex background in RGB-IR re-ID. In a word, we still have a long way to go in the research of RGB-IR re-ID.

5. Conclusions

In this paper, we propose a new method based on global and local alignment: AGF. Firstly, the RGB images are converted into IR images by using the AlignGAN model, which reduces the crossmodality difference. Then, the newly generated IR images and all the images in SYSU-MM01 are introduced into the framework, and the paired images are generated by separating and exchanging the modality-specific features between the unpaired images, and the set level and instance level are aligned at the same time. Finally, we first tried to apply SE-ResNet-50 network to the research of RGB-IR crossmodality re-ID and made a major breakthrough. The experimental results on SYSU-MM01 dataset show the effectiveness of our proposed method. In addition, we also verify that a RGB-IR re-ID method will show better performance on a stronger baseline, which shows the importance of designing a strong baseline in the region of RGB-IR re-ID.

Data Availability

The data related in our work is publicly available which is from reference [29] as we cited. Thus we think our description is suitable.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 11801511).

References

[1] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Piscataway, NJ, USA, 2018.

[2] X. Li, A. Wu, and W. Zheng, "Adversarial open-word person re-identification," in *in Proceedings of European Conference on Computer Vision*, pp. 280–296, Berlin, Germany, 2018.

[3] Z. Wang, R. Hu, and C. Chen, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 3006–3020, 2018.

[4] Z. Wang, R. Hu, C. Liang et al., "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[5] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person reidentification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 79–88, Piscataway, NJ, USA, 2018.

[6] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person reidentification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, Piscataway, NJ, USA, 2018.

[7] G. Wang, T. Zhang, and J. Cheng, "Rgb-infrared cross-modality person re-Identification via joint pixel and feature alignment," in *in Proceedings of the IEEE International Conference on Computer Vision*, pp. 3623–3632, Piscataway, NJ, USA, 2019.

[8] G. Wang, Y. Yang, and T. Zhang, "Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification," *Neural Networks*, vol. 128, pp. 294–304, 2020.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Piscataway, NJ, USA, 2018.

[10] M. Geng, Y. Wang, and T. Xiang, "Deep transfer learning for person reidentification," 2016, arXiv preprint arXiv:1611.05244.

[11] R. R. Varior and M. Haloi, "Gated Siamese convolutional neural network architecture for human re-identification," in *in Proceedings of European Conference on Computer Vision*, pp. 791–808, Berlin, Germany, 2016.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: a unified embedding for face recognition and clustering," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 815–823, Piscataway, NJ, USA, 2015.

[13] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[14] D. Cheng, Y. Gong, and S. Zhou, "Person re-identification by multichannel parts-based cnn with improved triplet loss function," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, Piscataway, NJ, USA, 2016.

[15] W. Chen, X. Chen, and J. Zhang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 403–412, Piscataway, NJ, USA, 2017.

[16] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person reidentification," 2017, arXiv preprint arXiv:1703.07737.

[17] R. R. Varior and B. Shuai, "A Siamese long short-term memory architecture for human re-identification," in *in Proceedings of European Conference on Computer Vision*, pp. 135–153, Berlin, Germany, 2016.

[18] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person reidentification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4500–4509, 2019.

[19] H. Zhao, M. Tian, and S. Sun, "Spindle net: person re-identification with human body region guided feature decomposition and fusion," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1077–1085, Piscataway, NJ, USA, 2017.

[20] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[21] Y. Lin, L. Zheng, Z. Zheng et al., "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019.

[22] Y. Wu, Y. Lin, and X. Dong, "Exploit the unknown gradually: one-shot video-based person re-odentification by stepwise learning," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, Piscataway, NJ, USA, 2018.

[23] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2872–2881, 2019.

[24] Z. Zhong, L. Zheng, and Z. Zheng, "Camera syle adaptation for person re-identification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, Piscataway, NJ, USA, 2018.

[25] L. Wei, S. Zhang, and W. Gao, "Person transfer GAN to bridge domain gap for person re-identification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 79–88, Piscataway, NJ, USA, 2018.

[26] X. Qian, Y. Fu, and W. Wang, "Pose-normalized image generation for person re-identification," in *in Proceedings of the European Conference on Computer Vision*, pp. 650–667, Berlin, Germany, 2018.

[27] X. Zhao, G. Ding, and Y. Guo, "TUCH: turning cross-view hashing into single-view hashing via generative adversarial nets," in *in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3511–3517, San Mateo, CA, USA, 2017.

[28] Y. Wu, L. Zhu, and Y. Yan, "Dual attention matching for audio-visual event localization," in *in Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6292–6300, Piscataway, NJ, USA, 2019.

[29] A. Wu, W. S. Zheng, and H. X. Yu, "Rgb-infrared cross-modality person re-identification," in *in Proceedings of the IEEE International Conference on Computer Vision*, pp. 5390–5399, Piscataway, NJ, USA, 2017.

[30] M. Ye, X. Lan, and J. Li, "Hierarchical discriminative learning for visible thermal person re-identification," in *in Proceedings of the AAAI conference on Artificial Intelligence (AAAI)*, pp. 7501–7508, Palo Alto, CA, USA, 2018.

[31] M. Ye, Z. Wang, and X. Lan, "Visible thermal person re-identification via dual-constrained top-ranking," in *in Proceedings of conference on International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1092–1099, San Francisco, CA, USA, 2018.

[32] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2020.

[33] M. Ye, X. Lan, Q. Leng, and J. Shen, "Cross-modality person re-identification via modality-aware collaborative ensemble learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 9387–9399, 2020.

[34] P. Dai, R. Ji, and H. Wang, "Cross-modality person re-identification with generative adversarial training," in *in Proceedings of conference on International Joint Conference on Artificial Intelligence*, pp. 677–683, San Francisco, CA, USA, 2018.

[35] Z. Wang, Z. Wang, and Y. Zheng, "Learning to reduce dual-level discrepancy for infrared-visible person reidentification," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 618–626, Piscataway, NJ, USA, 2019.

[36] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 728–739, 2021.

[37] Z. Zeng, Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3064–3074, 2019.

[38] L. Ma, Q. Sun, and S. Georgoulis, "Disentangled person image gGeneration," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 99–108, Piscataway, NJ, USA, 2018.

[39] S. Choi, S. Lee, and Y. Kim, "Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *in Proceedings of the IEEE conference on*

Computer Vision and Pattern Recognition, pp. 10257–10266, Piscataway, NJ, USA, 2020.

[40] Y. Lu, Y. Wu, and B. Liu, "Cross-modality person re-identification with shared-specific feature transfer," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 13379–13389, Piscataway, NJ, USA, 2020.

[41] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1326–1335, Piscataway, NJ, USA, 2021.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *in Proceedings of conference on International Conference on Learning Representations*, pp. 1–14, San Diego, CA, USA, 2015.

[43] C. Szegedy, W. Liu, and Y. Jia, "Going deeper with convolutions," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1–9, Piscataway, NJ, USA, 2015.

[44] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, Piscataway, NJ, USA, 2016.

[45] S. Xie, R. Girshick, and P. Dollar, "Aggregated residual transformations for deep neural networks," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, Piscataway, NJ, USA, 2017.

[46] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, Piscataway, NJ, USA, 2017.

[47] J. Dai, H. Qi, and Y. Xiong, "Deformable convolutional networks," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 764–773, Piscataway, NJ, USA, 2017.

[48] X. Zhu, H. Hu, and S. Lin, "Deformable convnets v2: more deformable, better results," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, Piscataway, NJ, USA, 2019.

[49] X. Mao, Q. Li, and H. Xie, "Least Squares Generative Adversarial Networks," in *in Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, Piscataway, NJ, USA, 2016.

[50] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *in ACM Multimedia (ACM MM)*, pp. 347–355, Nice, France, 2019.

[51] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 886–893, Piscataway, NJ, USA, 2005.

[52] S. Liao, Y. Hu, and X. Zhu, "Person reidentification by local maximal occurrence representation and metric learning," in *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, Piscataway, NJ, USA, 2015.

[53] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: a survey and outlook," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, p. 1, 2021.