future internet

MDPI

Article

MFCNet: Mining Features Context Network for RGB–IR Person Re-Identification

Jing Mei *, Huahu Xu, Yang Li, Minjie Bian and Yuzhe Huang

School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; huahuxu@staff.shu.edu.cn (H.X.); liyang162@shu.edu.cn (Y.L.); bianmj0302@aliyun.com (M.B.); yzhuang@shu.edu.cn (Y.H.)

* Correspondence: meijing9615@shu.edu.cn

Abstract: RGB–IR cross modality person re-identification (RGB–IR Re-ID) is an important task for video surveillance in poorly illuminated or dark environments. In addition to the common challenge of Re-ID, the large cross-modality variations between RGB and IR images must be considered. The existing RGB–IR Re-ID methods use different network structures to learn the global shared features associated with multi-modalities. However, most global shared feature learning methods are sensitive to background clutter, and contextual feature relationships are not considered among the mined features. To solve these problems, this paper proposes a dual-path attention network architecture MFCNet. SGA (Spatial-Global Attention) module embedded in MFCNet includes spatial attention and global attention branches to mine discriminative features. First, the SGA module proposed in this paper focuses on the key parts of the input image to obtain robust features. Next, the module mines the contextual relationships among features to obtain discriminative features and improve network performance. Finally, extensive experiments demonstrate that the performance of the network architecture proposed in this paper is better than that of state-of-the-art methods under various settings. In the all-search mode of the SYSU and RegDB data sets, the rank-1 accuracy reaches 51.64% and 69.76%, respectively.

Keywords: person re-identification; RGB–IR; MFCNet; SGA module

check for updates

Citation: Mei, J.; Xu, H.; Li, Y.; Bian, M.; Huang, Y. MFCNet: Mining Features Context Network for RGB–IR Person Re-Identification. *Future Internet* **2021**, *13*, 290. https://doi.org/10.3390/fi13110290

Academic Editors: Mustafa Atay, Ajita Rattani and Kaushik Roy

Received: 9 October 2021
Accepted: 15 November 2021
Published: 18 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Person re-identification (Re-ID) is regarded as a subproblem of image retrieval. The goal is to retrieve a cross-camera image of a person from a recorded video. This approach is widely used in video surveillance, social security, and smart city applications and is challenging to implement. Due to different postures, environmental lighting changes, view changes, and occlusions, few distinguishing features may be available among different modalities, and many interference features may exist [1,2]. Many traditional methods focus on images taken by visible-light cameras and express the person re-identification problem as an RGB–RGB single-modality matching problem. These methods have achieved success in the field of person re-identification, especially with deep learning.

However, under poor lighting conditions (such as those during the night), it is difficult for visible-light cameras to capture effective appearance information, which limits the practical application of person re-identification methods [3–8].

Therefore, other types of vision sensors, such as infrared cameras, are now widely used as supplements to visible-light cameras to overcome lighting problems. As a result, it is important to study person re-identification in real environments as a cross-modality matching problem [9,10].

For RGB–IR Re-ID [9], cross-modality discrepancies and intra-modality variations are two problems that urgently need to be solved. Cross-modality discrepancies are due to the different imaging principles of RGB cameras and IR cameras. In Figure 1, the IR images in the first row were taken with an IR camera at night and only contain a single

channel of heat information; the RGB images in the second row were captured by an RGB camera under good lighting conditions and contain the most important three-channel color information related to pedestrian appearance. As a result, color information that is very important in RGB-RGB Re-ID cannot be used in RGB–IR Re-ID. However, similar to those in single-modality Re-ID, intra-modality variations are caused by multiple factors, such as changes in perspective, person posture and lighting conditions. These issues make RGB–IR Re-ID extremely challenging.



Figure 1. Four people are depicted: (**a**–**d**). The RGB images in the first row were taken with an RGB camera under good lighting conditions and contain the most important three-channel color information related to pedestrian appearance.The IR images in the second row were taken by an IR camera at night and only contain single-channel heat information. There is a one-to-one correspondence between infrared image and RGB image.

These challenges make it difficult to use the most advanced single-modality Re-ID methods [11–14] to reliably learn to distinguish part-level features. As a compromise, the existing RGB–IR Re-ID methods mainly focus on learning multi-modality-sharable global features through networks. Some works also integrate modality discriminant supervision or GAN-generated images to deal with modality differences. However, the global feature learning method is very sensitive to background clutter and cannot deal with modality differences clearly. In addition, part-based feature learning methods for single-modality Re-ID generally cannot capture reliable part features under large cross-domain gaps. In addition, when the appearance of the two modalities is very different, the learning is easily contaminated by noisy samples. All these challenges will lead to reduced discrimination of cross-modality features.

However, the global shared feature learning method is sensitive to background clutter and does not consider the contextual relationships among mined features. Therefore, we propose a new dual-path attention network. To solve the interference problem caused by background clutter, we apply a method based on an attention mechanism [12,15–23]. Based on this approach, MFCNet, a novel dual-path attention network, is established. Through learning, the key parts of an image are weighted, and those parts that do not contribute to or interfere with the recognition task are ignored. This module obtains discriminative features by mining the contextual relationships among features. Quantitative experiments verify that the method proposed in this paper is superior to other state-of-the-art methods.

The contributions of this article can be summarized as follows:

- A new end-to-end deep learning network structure is proposed to enhance the discriminant ability of feature representations for RGB–IR Re-ID problems. This approach provides a strong baseline for future research in this field.

- This paper introduces a new attention module combining spatial attention and global attention into the field of RGB–IR Re-ID. Based on this design, the SGA module combined with the RestNet50 backbone network can extract discriminative features by mining the context of features and can effectively avoid the effect of background clutter.
- Our proposed SGA module, in an end-to-end manner with the two-stream CNN structure, outperforms the state-of-the art approaches by large margins on two public RGB–IR Re-ID datasets. We also experimentally verified the improvement of the network by embedding different positions and numbers of SGA modules.

The remainder of this paper is structured as follows: Section 2 outlines related works on person Re-ID, cross modality retrieval and generative adversarial networks. Section 3 describes the proposed framework for learning discriminative representations for RGB–IR person Re-ID. Experimental results of our approach are demonstrated in Section 4. We discuss our SGA module in Section 5. Finally, we conclude this paper in Section 6.

2. Related Works

2.1. RGB–RGB Re-ID

The purpose of RGB–RGB Re-ID is to match person RGB images from disjoint visible-light cameras. Re-ID involves specific person information retrieval, which usually includes two steps: feature learning and metric learning. Feature learning methods focus on learning distinguishing and robust feature representations to represent a person's image, and metric learning methods usually apply a loss function to measure the similarity of learned features. The existing Re-ID methods use a CNN network for end-to-end learning to provide powerful discrimination capabilities, some of which even exceed the level of humans. Most of the existing methods have been developed for RGB–RGB Re-ID, and person re-identification in RGB–IR is ignored, thus limiting the use of person re-identification methods in actual monitoring [3–8].

2.2. RGB–IR Re-ID Based on CNN Networks

RGB–IR Re-ID based on CNN networks has been a long-time research focus of scholars. Since RGB–IR Re-ID was formally proposed in 2017, many excellent works based on the CNN network architecture have been proposed. Ye et al. [24] proposed a two-way convolutional network to extract the features of different modalities and identify common features through parameter sharing. Additionally, the author improved upon the triple loss function and proposed the top-ranking loss within and between modalities. Hao et al. [25] applies sphere softmax loss, a variant of the classic softmax loss, to RGB–IR Re-ID by converting two-dimensional coordinates to spherical coordinates and relating the spherical coordinates to the angle between the vectors when performing classification tasks. In this approach, the mode is irrelevant. Liu et al. [26] use two-way convolution to extract features. In the process of extracting features, skip connections are used to fuse the middle layers of the CNN model and enhance the robustness and non-descriptiveness of the extracted features. Zhao et al. [27] expanded the triple loss function to pentaplet loss, and the cross-modality problem was considered on the basis of the original triple loss function; additionally, a method for mining difficult samples was introduced. Zhu et al. [28] involves feature centers of the same category and the same modality, and hetero center loss was proposed on the basis of center loss, with a focus on the differences among feature centers of different modalities in the same category. Feng et al. [29] uses mapping loss over two Euclidean distances to replace the triple loss function, and shared features are considered that reduce cross-modality differences by minimizing the distance for features with the same ID. Lu et al. [30] separates the features extracted in the common space and uses a confrontation strategy to classify the information associated with modal features; notably, a shared feature does not contain modal information, and a unique feature is a non-shared feature. Li et al. [31] uses X modalities to connect RGB and IR images, and the purpose is to reduce the differences between modalities and the RGB and IR images.

2.3. RGB–IR Re-ID Based on GAN Networks

Another way to solve RGB–IR Re-ID problems is to use a generative adversarial network (GAN) to generate cross-modality images. This method can reduce the differences among modalities and provide data enhancement capability. Since Goodfellow et al. [32] proposed GAN, GAN-based methods have gained popularity in the field of deep learning. Additionally, GAN has been applied in RGB–IR Re-ID tasks. Dai et al. [33] uses a GAN to distinguish the features of different modalities, a generator to extract shared features, and a discriminator to distinguish the unique features of each modality and different modalities to improve modal classification. Wang et al. [34] uses cycleGAN to generate cross-modality images and applies the generated images at the same pixel and feature levels to reduce the differences among modalities. Wang et al. [35] uses a GAN to generate cross-modality paired images and then uses modality paired images for feature extraction. Choi et al. [36] combines a feature unwrapping representation with a GAN, separates ID-discriminative features and ID-excluded features, and then performs separate learning tasks. The above use of GANs to solve RGB–IR Re-ID problems involves set-level alignment, but misalignment will occur in some cases, thus reducing the performance of the network. Wang et al. [37] uses a GAN to first generate paired images and align them at the instance level. Then, the paired images are input into the set-level encoder for further alignment. Moreover, the learning is easily contaminated by noisy samples and destabilized when the appearance discrepancy is large across the two modalities. All of these challenges result in less discriminative cross-modality features.

In this field, the use of GAN to solve RGB–IR Re-ID problems involves two steps: distinguishing the characteristics of different modalities through a GAN and generating cross-modality images by producing adversarial images.These methods perform well for RGB-generated IR images but perform poorly in the one-to-many mapping of IR-generated RGB images. Additionally, background interference, noise interference, and the contextual relationships between image pairs are not considered.

2.4. Attention Mechanisms

An important feature of the human visual system is that humans selectively focus on the salient parts of a series of glimpses in order to capture valuable information. With reference to the human visual system, there have been several attempts to use the attention mechanism to improve the performance of CNNs. Hu et al. [17] introduced SENet to exploit the dimension-wise relationship. They propose the Squeeze-and-Excitation module to apply attention mechanisms on the dimensions. Considering the relationship between any two positions, Wang et al. [19] proposed non-local neural network to capture the relationships among them. To broaden horizon, namely to make it can see "what" and "where" at the same time, CBAM [16] was proposed, which exploits both spatial and dimension-wise attention. Following these methods, we propose the SGA modules in the context of mining features.

3. Our Method

In this section, we first present a general framework of our network and then introduce the two attention modules which capture long-range contextual information in spatial and global dimensions, respectively.

3.1. Problem Description

The difficulty of RGB–IR Re-ID is related to the spatial misalignment associated with differences in person posture, lighting, and viewing angle conversion; additionally, image modality misalignment can be associated with variations in imaging principles. In this paper, $V = \{v_i | i = 1, 2, \ldots, N_1\}$ represents the RGB image data set, $T = \{t_i | i = 1, 2, \ldots, N_2\}$ represents the IR image data set, and $Y = \{y_i\}$ represents the identity labels. $D = \{D_{train}, D_{test}\}$ represents a multimodality dataset, where $D_{train} = \{V_{train}, T_{train}\}$ represents a training set and $D_{test} = \{V_{test}, T_{test}\}$ represents a test set. RGB–IR Re-ID

aims to automatically learn a feature extraction function $f(.)$ that can convert the input RGB image into a feature vector f_i^v. Similarly, the input IR image can be converted into a feature vector f_i^t. The superscripts ^{v} and ^{t} represent different modalities, and the subscript $_i$ represents the index of the current image. The core step in solving RGB–IR Re-ID problem is to obtain feature representations that are as similar as possible for similar features corresponding to different modalities.

3.2. The Proposed MFCNet

The framework MFCNet used in this paper is shown in Figure 2. This framework includes an end-to-end network architecture, including image input, feature extraction and loss measurement tasks. Due to the differences among the modalities of input images, this paper uses an infrared and visible-light dual-path network as the basic structure. Based on [15,25,26], the backbone network for feature extraction in this paper is Resnet50 [38], and SGA modules are embedded between the second layer and the third layer and the third layer and the fourth layer. The SGA module designed in this paper can extract the features of key parts of images and mine the relationships among contextual features. The experiment in Section 4 verifies that embedding the SGA module in the convolutional layers can improve network performance and improve image retrieval capabilities. This paper uses generalized mean (GeM) pooling [15] after the final convolutional layer to focus on areas with features of different scales. This paper uses WRT loss to reduce the differences among modalities, and ID loss is used to learn identity-invariant features.

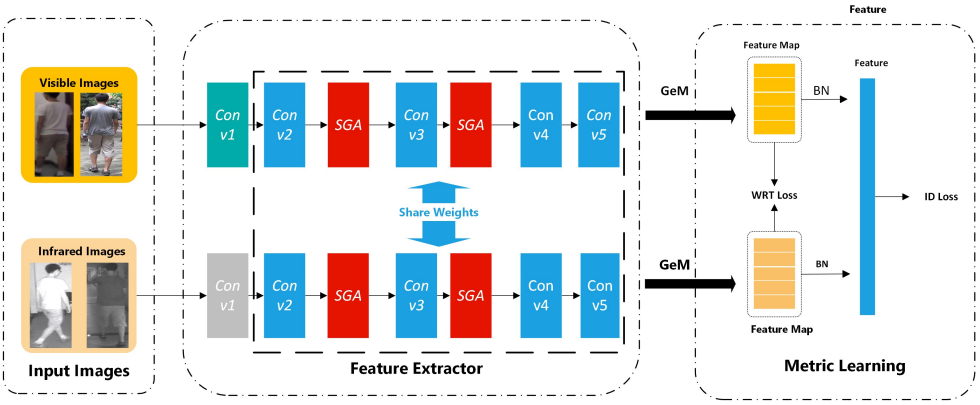


Figure 2. MFCNet. The model in this paper is a two-stream (one RGB image and one infrared image) end-to-end architecture that includes image inputs, feature extraction, and loss measurement. The feature extraction process uses ResNet50 as the backbone network, and the convolutional layer parameters from the first layer are not shared. The remaining convolutional layer parameters are shared to obtain middle features. In this paper, MFCNet with the same shared parameters is embedded into the network to capture key discriminative features, and WRT loss and ID loss are used for network training to improve network performance. This figure is best viewed in color.

Feature Extraction. The model in this paper uses a dual-path convolutional network architecture to learn feature expressions through partial parameter sharing to resolve modality differences at the feature level, as shown in Figure 2. Inspired by [12], the RGB and IR images of the same person are different based on low-level features but the same when middle-level features are considered. Thus, the dual-path network architecture in this paper does not share the parameters of the first convolution block to extract low-level features of a specific modality, and the convolution block of the deep network learns the shared mid-level features of the two modalities through shared parameters. After feature extraction, GeM pooling is used to obtain additional feature information. Compared with other dual-path architectures, this kind of backbone network can better mine the intermediate feature information and obtain more features for identification. The SGA module proposed in this paper plays a key role in extraction, and we will introduce this module next.

SGA Module. In the task of person re-identification, we use convolutional networks to extract image features. Since convolution operations would lead to a local receptive field and the intra-class variations are caused by pose and view e.g., the features corresponding to the pixels with the same label may have some differences. These differences introduce intra-class inconsistency and affect the recognition accuracy. To address this issue, we explore global contextual information by building associations among features with the attention mechanism. Our method could adaptively aggregate long-range contextual information, thus improving feature representation for RGB–IR Re-ID [39–41]. Therefore, this paper designs a key part attention module that automatically learns the key parts of each image through training. The attention module designed in this paper can mine discriminative features and is robust to image occlusion. We will introduce this module in detail.

The detailed structure of the SGA module is shown in Figure 3. For the input features $F \in \mathbb{R}^{C\times H\times W}$, the attention map $A(F) \in \mathbb{R}^{C\times H\times W}$ is obtained after learning by the attention module. We use the following formula to calculate the resulting feature set F':

$$F' = F + F \otimes A(F) \tag{1}$$

where $\otimes$ represents elementwise multiplication. Inspired by residual learning, we add the features learned by the attention module to the original features to retain information.

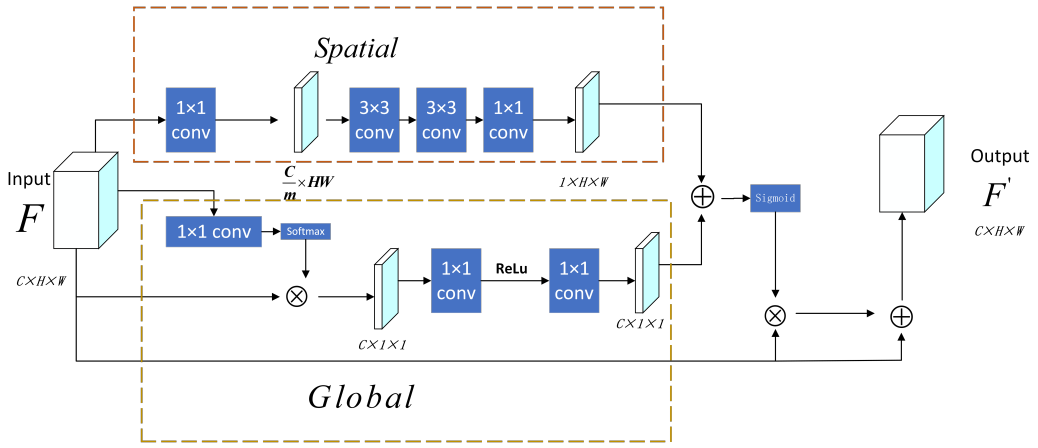


Figure 3. SGA framework diagram. In this paper, MFCNet includes two parts: spatial attention and global attention parts. Given the feature map F, the module calculates the relevant attention matrix through two attention branches: the spatial attention branch $A_S(F)$ and the global attention branch $A_G(F)$. There is a hyperparameter, the dimensionality reduction rate m in the spatial attention branch; it is set to 16 in this paper.

The attention module designed in this paper is divided into two branches to learn more the most critical areas of images. The upper part of the figure is the spatial attention branch, and the lower part is the global attention branch. The attention map learned by the spatial attention branch is $A_S(F) \in \mathbb{R}^{1\times H\times W}$, and the attention map learned by the global attention branch is $A_G(F) \in \mathbb{R}^{C\times 1\times 1}$. Therefore, we get can obtain $A(F)$:

$$A(F) = \delta(A_S(F) + A_G(F)) \tag{2}$$

where δ is the sigmoid activation function.

Spatial attention branch. The spatial attention [39,41] branch aims to improve feature expression in key areas. In essence, it converts the spatial information in the original image to another space to preserve the key information, generates a weight mask for each position and weights the output to enhance key areas and weaken noncritical background areas. Compared with other attention mechanisms, the spatial attention mechanism focuses on "where", which requires a larger receptive field to obtain more information. For the input

feature $F \in \mathbb{R}^{C\times H\times W}$, a 1 × 1 convolution kernel is used to compress the cross-channel dimension and obtain a feature with a dimension of $\mathbb{R}^{\frac{C}{m}\times H\times W}$. Then, two 3 × 3 small-scale convolution kernels are used to obtain a larger receptive field and capture key areas in this paper. Finally, the channel features are integrated and compressed through the 1 × 1 convolution kernel, and a spatial attention matrix of dimension $\mathbb{R}^{1\times H\times W}$ is obtained. After passing the BN layer, the obtained matrix is output for the next step in the fusion process. The formula for the spatial attention mechanism is as follows:

$$A_S(F) = BN(conv^{1\times1}(conv^{3\times3}(conv^{3\times3}(conv^{1\times1}(F))))) \tag{3}$$

where BN denotes a batch normalization operation and $conv$ denotes a convolution operation.

Discriminant feature representations are essential for RGB–IR Re-ID and can be obtained by capturing long-range contextual information. In order to model rich contextual relationships over local features, we introduce a global attention module. The global attention module encodes a wider range of contextual information into local features, thus enhancing their representation capability. Next, we elaborate the process to adaptively aggregate spatial contexts.

Global Attention Branch. In this paper, the global attention [17,23,42] mechanism uses global contextual information to recalibrate the weights of different channels and adjust the channels' dependence levels. In general, this mechanism can be summarized into three steps: (1) context modeling, (2) feature conversion, and (3) feature fusion. For the input feature set $F \in \mathbb{R}^{C\times H\times W}$, 1 × 1 convolution and the softmax function are first used to obtain attention weights; then, the original input features are fused to obtain global context features. We use $F_1 \in \mathbb{R}^{C\times1\times1}$ to represent the features obtained by completing context modeling. The corresponding calculation is as follows:

$$F_1 = \sum_{j=1}^{N_p} \frac{exp(conv^{1\times1}(F_j))}{\sum_{q=1}^{N_p} exp(conv^{1\times1}(F_q))} F_j \tag{4}$$

where N_p represents the number of positions in the feature vector, F_j represents the feature corresponding to position index j, and $conv^{1\times1}$ denotes the convolution operation, which uses a linear conversion matrix. Additionally, $w_j = \frac{exp(conv^{1\times1}(F_j))}{\sum_{m=1}^{N_p} exp(conv^{1\times1}(F_m))}$ is expressed in embedded Gaussian form to normalize the positions and encompass the weights used in global attention pooling.

After completing the modeling process, feature conversion is performed on the obtained features to capture the dependence among channels. We complete feature conversion by reducing the number of parameters using operations similar to those in a bottleneck layer. After feature conversion, we obtain the feature set $A_G(F) \in \mathbb{R}^{C\times1\times1}$, which is obtained with the following formula:

$$A_G(F) = conv^{1\times1}(ReLU(LN(conv^{1\times1}F_1))) \tag{5}$$

where $conv^{1\times1}$ represents that the convolution operation, which uses a linear conversion matrix. F_1 is obtained from formula (4) and expresses the relevant features after modeling. LN denotes layer normalization, in which different channels are used to reduce the difficulty of optimization and improve the generalization ability of the algorithm.

After modeling and conversion, the features are merged with the features obtained by the spatial attention branch to increase feature diversity. Because the dimensions of the feature sets from the two branches are different, they are expanded to the dimension $\mathbb{R}^{C\times H\times W}$ to obtain the spatial attention matrix $A_S(F) \in \mathbb{R}^{C\times H\times W}$ and the global context attention matrix $A_G(F) \in \mathbb{R}^{C\times H\times W}$ before fusion. Then, fusion is performed through an element-wise addition operation.

Previous experiments have verified that the addition operation can achieve better performance than other operations. After fusion, the attention matrix is obtained through

the sigmoid function with weights, and the original input $F \in \mathbb{R}^{C\times H\times W}$ is multiplied by the weights to obtain the weighted feature matrix. Finally, the result is added to the original input set $F \in \mathbb{R}^{C\times H\times W}$ to improve the representation ability of the features. The corresponding calculation is based on formula (1).

GeM Pooling. Unlike the commonly used maximum pooling and average pooling methods, GeM pooling focuses on areas of different scales to capture the characteristics of specific modalities. The corresponding formula is as follows:

$$\boldsymbol{f} = [f_1 ... f_k ... f_K]^T, f_k = (\frac{1}{|\chi_k|} \sum_{x\in\chi_k} x^{p_k})^{\frac{1}{p_k}} \tag{6}$$

where χ_k represents each feature map and p_k is a hyperparameter. When $p_k \to \infty$, formula (6) becomes the expression for maximum pooling. Correspondingly, when $p_k \to 0$, formula (6) becomes the expression for average pooling.

Loss Function. To learn the feature representations, two types of loss functions are utilized to supervise the training of our network. One is identity loss, employed for identity information learning, and the other is triplet loss, used for similarity learning. However, contrastive loss and triplet loss are both common methods for similarity learning. The effect of triplet loss is better than that of contrastive loss, because it considers the distance between the positive and negative samples and the anchor point. Therefore, the combination of triplet loss and identity loss is one of the most popular solutions for deep person Re-ID model learning.

Identity Loss. The identity loss aims to learn the discriminative feature representations by using modality-specific information, which could distinguish different persons within each modality. It treats the training process of person Re-ID as an image classification problem, where each identity is a distinct class. Given an input image v_i with the identity label y_i, the probability of v_i being recognized as class y_i encoded with a softmax function, represented by $p(y_i|v_i)$. The identity loss is then computed by the cross-entropy

$$L_{id} = -\frac{1}{n} \sum_{i=1}^{n} log(p(y_i|v_i)) \tag{7}$$

where n represents the number of training samples within each batch. Similarly, we could get the identity loss for the thermal image t_i.

Weighted Regularization Triplet (WRT) Loss. The triplet loss is used for similarity learning. It tries to reduce the feature distances between images of the same person and expand the distances between images of different people.The triplet loss with a margin parameter is represented by

$$L_{tri}(i,j,k) = max(\rho + d_{ij} - d_{ik}, 0) \tag{8}$$

where $d(.)$ measures the Euclidean distance between two samples. ρ is the pre-defined margin. ij is a positive sample pair, and ik is a negative sample pair.

The weighted regularized triple loss function is an improvement of the triple loss function. It not only inherits the advantages of the original method but also avoids the pre-defined parameter.

$$L_{wrt}(i,j,k) = log(1 + exp(w_i^p d_{ij}^p - w_i^n d_{ik}^n)) \tag{9}$$

$$w_i^p = \frac{exp(d_{ij}^p)}{\sum_{d^p\in P} exp(d^p)}, w_i^n = \frac{exp(-d_{ik}^n)}{\sum_{d^n\in N} exp(-d^n)} \tag{10}$$

Notably, i, j, and k are the sample indexes. Additionally, p represents the positive sample data set in each training batch, and n represents the negative sample data set. ij is a positive sample pair, and ik is a negative sample pair. d represents the Euclidean distance between two samples.

In this paper, weighted regularized triple loss function L_{wrt} and identity loss function L_{id} are combined for network training to learn discriminative features [12,15,24–29]. Therefore, the loss function is defined as:

$$L = L_{wrt} + L_{id} \tag{11}$$

4. Results

In this section, a series of experiments is conducted to verify the effectiveness of the proposed method.

4.1. Experimental Settings

The experimental data in this paper are from open-source datasets: SYSU-MM01 and RegDB. For both RegDB and SYSU-MM01, we use a standard cumulated matching characteristics (CMC) curve and mean average precision (mAP) to evaluate our algorithm. Specifically, we adopt the cumulative Matching Characteristics (CMC) at rank-1, rank-10, and rank-20. The rank-k calculates the percentage of testing samples that find the correct result in the top k search outcomes of the query sample. The metric mAP is an average of the maximum recalls for each class in multiple types of tests.

SYSU-MM01. It is a large-scale benchmark data set [9] used to solve RGB–IR Re-ID problems. The dataset contains a total of 303,420 images of 491 pedestrians taken by six cameras (four visible light and two infrared). Among them, the visible-light cameras captured 287,628 pedestrian RGB images, and the infrared cameras took 15,792 pedestrian IR images. The images taken include both outdoor and indoor shots. Indoor photographs were taken by one infrared and two visible-light cameras, and outdoor photographs were taken by the remaining cameras. The training set includes 395 pedestrians in 22,258 RGB images and 11,909 IR images. The test set contains 96 pedestrians; 3803 IR images are used for querying, and 301 selected RGB images are selected as the gallery set.

RegDB. It is a relatively small data set captured by a dual-camera system consisting of visible-light and infrared cameras. The dataset contains 8240 images of 412 pedestrians. Each pedestrian is included in 10 visible-light images and 10 far-infrared images.

For fair comparison with other methods, the data set used in this paper was randomly divided into two halves: one half was used for training, and the other half was used for testing. In the test, the RGB images were used as the query set, and the IR images were used as the gallery set.

4.2. Experimental Parameters

The PyTorch framework was used to write the experimental source code. We trained our network on a single NVIDIA V100 GPU with 32 GB of memory. To compare the proposed method with other methods, the experiment used ResNet50 as the backbone network and the parameters of a pretrained ImageNet to initialize the network proposed in this paper. All input images were adjusted to 288 × 144. Zero padding and horizontal flipping with random cropping were used for data expansion, and the SGD optimizer with a parameter set to 0.9 was used to optimize the network. A total of 80 training cycles were considered.

According to [12], the initial learning rate was set to 0.1; the learning rate decayed by 0.1 after the 30th iteration and 0.01 after the 50th iteration. Eight different pedestrian images, including four visible-light images and four infrared images, were randomly selected from each training batch.

In the feature extraction stage, the first convolution block does not share parameters to extract low-level features associated with different modalities, and the remaining four convolution blocks share parameters with the attention module to obtain discriminative features. To produce a fine-scale feature matrix, the last stride of the last convolution block is set to 1. In the SGA module, we set m to 16 to obtain a larger receptive field and capture key areas.

In Equation (6), there is a different pooling parameter p_k for each feature map χ_k. However, one can also use a shared parameter p for all feature maps, and it turns out that this simpler option achieves slightly better results than multiple parameters, as demonstrated in the experimental section of [43]. Consequently, we shall also adopt this option for our experiments. We set p = 3 in our code.

4.3. Comparison with State-of-the-Art Methods

In this section, the proposed method is compared with other state-of-the-art methods based on the SYSU and RegDB data sets. Previous works [12,44] provided results for most of the considered methods. The methods compared in this paper include manual feature design methods (HOG [45] and LOMO [3]) and deep learning methods [9,15,24–27,29,33–35,46–50]. The results of the compared methods were obtained from the original text. "-" means that this method has not been tested.

Table 1 shows the experimental results based on the SYSU data set. This study conducted experiments in all-search and indoor-search modes. From the superficial analysis of the results listed in the table, we can conclude that compared with manual design methods, deep learning algorithms perform better in all considered aspects. Notably, deep learning methods can learn more than manually designed features. Additionally, the indoor-search mode yields better results than the all-search mode. Our analysis suggests that the background interference in photos taken indoors is less influential in learning key features.

In all-search mode, the experimental results of the method proposed in this paper are 51.64%, 86.80%, 93.66%, and 49.99% for r = 1, r = 10, r = 20, and the mAP, respectively. Among them, the hit rate for r = 1 is 10.28% and 9.24% higher than those of HPLIN and AlignGAN. Compared with the latest baseline [15], the proposed method yields 4.14%, 2.59%, 1.52%, and 2.34% improvements for r = 1, r = 10, r = 20 and mAP, respectively. Compared to the FBP-AL, we achieve about 1 point improvement on r = 10 and r = 20.

In the indoor search mode, the results of this paper are 57.29%, 92.89%, 96.38%, and 64.91% for r = 1, r = 10, r = 20, and mAP, respectively. Among them, the accuracy for r = 1 is 11.52% and 11.39% higher than those of HPLIN and AlignGAN, respectively. Additionally, the proposed method performs better than the AGW method in all aspects.

The experimental results of RegDB are shown in Table 2. The RegDB experiment in this paper is in Visible2Infrared mode, using visible images as the query set and infrared as the gallery set. It can be seen from Table 2 that the experimental results of the method proposed in this paper are 69.76%, 86.07%, 90.73%, and 52.52% respectively. Among them, the experiments at r = 1 and r = 10 not only show the best performance but also have better performance than AlignGAN at r = 1.

The above impressive performance suggests that our method can learn better modality-sharable features.

Combining the experimental results of SYSU and RegDB, the method proposed in this paper can capture more discriminative features and improve the accuracy. Experiments demonstrate our method is better than the latest baseline algorithm AGW. This shows that the method proposed in this paper will provide a more powerful baseline algorithm for RGB–IR Re-ID.

Table 1. Comparison of the results of the proposed method and other state-of-the-art methods based on the SYSU dataset. Bold numbers are the best.

Settings	**All-Search**				**Indoor Search**			
Methods	**r = 1**	**r = 10**	**r = 20**	**mAP**	**r = 1**	**r = 10**	**r = 20**	**mAP**
HOG [45]	2.76	18.3	32	4.24	3.22	24.7	44.6	7.25
LOMO [3]	3.64	23.2	37.3	4.53	5.75	34.4	54.9	10.2

Table 1. *Cont.*

Settings	All-Search				Indoor Search			
Methods	**r = 1**	**r = 10**	**r = 20**	**mAP**	**r = 1**	**r = 10**	**r = 20**	**mAP**
One-stream [9]	12.04	49.68	66.74	13.67	16.94	63.55	82.1	22.95
Two-stream [9]	11.65	47.99	65.5	12.85	15.6	61.18	81.02	21.49
Zero-padding [9]	14.8	54.12	71.33	15.95	20.58	68.38	85.79	26.92
Tone [46]	12.52	50.72	68.6	14.42	20.82	68.86	84.46	26.38
HCML [46]	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
cmGAN [33]	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
BDTR [24]	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
HSME [25]	20.68	32.74	77.95	23.12	-	-	-	-
D2RL [35]	28.9	70.6	82.4	29.2	-	-	-	-
MAC [47]	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
MSR [29]	37.35	83.4	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN [34]	42.4	85	93.7	40.7	45.9	87.6	94.4	54.3
HPLIN [27]	41.36	84.78	94.31	42.95	45.77	91.82	**98.46**	56.52
AGW [15]	47.5	84.39	92.14	47.65	54.17	91.14	95.98	62.97
LCRF [48]	43.23	82.78	90.91	43.09	50.07	90.63	96.99	58.88
H2H [49]	45.47	72.78	82.28	47.99	-	-	-	-
FBP-AL [50]	**54.14**	86.04	93.03	**50.20**	-	-	-	-
Ours	51.64	**86.8**	**94.66**	49.99	**57.29**	**92.89**	96.38	**64.91**

Table 2. Comparison of the results of the proposed method and other state-of-the-art methods based on the RegDB dataset. Bold numbers are the best.

Settings	Visible to Infrared			
Methods	**r = 1**	**r = 10**	**r = 20**	**mAP**
HOG [45]	13.49	33.22	43.66	10.31
LOMO [3]	0.85	2.47	4.10	2.28
Zero padding [9]	17.75	34.21	44.35	18.90
HCML [46]	24.44	47.53	56.78	20.08
BDTR [24]	33.56	58.61	67.43	32.76
HSME [25]	50.85	73.36	81.66	47.00
D2RL [35]	43.4	66.1	76.3	44.1
MAC [47]	36.43	62.36	71.63	37.03
MSR [29]	48.43	70.32	79.95	48.67
EDFL [26]	52.58	72.10	81.47	52.98
AlignGAN [34]	57.9	-	-	53.6
Xmodal [31]	62.21	83.13	91.72	60.18
H2H [49]	62.27	77.56	83.72	**61.90**
Ours	**69.76**	**86.07**	**91.73**	52.52

5. Discussion

In this section, the effectiveness of each component of the network designed in this paper is verified based on ablation experiments. We also experimentally verified the improvement of the network by embedding different positions and numbers of SGA

modules. It proves that the network proposed in this paper can extract more discriminative features and is optimal.

5.1. Ablation Experiment

This section verifies each model component based on the SYSU data set in all-search and indoor-search modes. A dual-path network without SGA module is used as B, S represents the spatial attention branch, and G represents the global attention branch in the SGA module.

The results are shown in Table 3. We can draw the following conclusions.

1. In the two modes, compared to the baseline method, MFCNet achieves an improvement in the rank-1 accuracy and mAP. Notably, the rank-1 accuracy increased by 5.76% and 7.56% between the proposed method and other methods. The results show that the SGA module designed in this paper can identify more discriminative features than can other methods to improve the hit rate.
2. The spatial attention branch in SGA module helps improve network performance. In all-search and indoor-search modes, the spatial attention mechanism increases the hit rate for r = 1 by 2–5%. Specifically, the spatial attention mechanism can retain the key information form an image while ignoring noncritical information and interference;
3. The global attention branch in SGA module yields a 3–5% improvement in the hit rate for r = 1 in the two modes. This result verifies that the global attention mechanism can mine feature context, which helps the network extract robust features and improve network performance.
4. As shown in Tables 1–3, for the same method, indoor-search performs better than all-search. This is because images of indoor have less background variation and a more stable light condition and person pose, which makes matching easier. In the first row of the Figure 1, the first two images were taken indoors, and the second two were taken outdoors. We can see that the photos taken indoors are purer than those taken outdoors.

Table 3. SYSU ablation experiment results.

Settings	All-Search				Indoor Search			
Methods	**r = 1**	**r = 10**	**r = 20**	**mAP**	**r = 1**	**r = 10**	**r = 20**	**mAP**
B	45.88	82.96	90.38	46.36	49.73	90.90	96.97	59.21
B+S	48.51	84.41	91.69	46.12	54.30	90.62	97.10	61.10
B+G	49.65	84.12	92.72	49.01	54.57	91.44	96.24	62.21
B+S+G	51.64	86.80	94.66	49.99	57.29	92.89	96.38	64.91

5.2. Supplementary Experiment

We want to explore the influence of the position and number of SGA modules on the network. We designed some ablation experiments. In this part, we compare the performance of SGA conducted at different convolution stages (Table 4). The ranks 1, 10, 20, and mAP tend to increase when SGA is conducted on more layers. The best result appears when we perform SGA on Stage 2 and 3. This is probably because searching more stages expands the search space, which allows us to explore more varied selections of features. However, blindly extending the search scope will pose great challenges to the discovery of an optimal feature set. The underlying reason is that STE (straight-through estimator) will generate more and more gradient estimation errors when backpropagating through too many layers [51].

Table 4. Comparison on SGA at different convolution stages. Bold numbers are the best.

Setting					All-Search			
Design	**stage1**	**stage2**	**stage3**	**stage4**	**r = 1**	**r = 10**	**r = 20**	**mAP**
M1	×	×	×	×	45.88	82.96	90.38	46.36
M2	√	×	×	×	46.23	82.16	90.14	45.39
M3	×	√	×	×	46.70	83.64	90.66	46.47
M4	×	×	√	×	47.15	84.72	92.53	46.75
M5	×	×	×	√	47.75	85.46	93.29	44.73
M6	√	√	×	×	48.28	83.91	91.87	47.12
M7	√	×	√	×	48.46	84.30	92.06	44.95
M8	√	×	×	√	48.09	85.35	92.35	47.50
M9	×	√	√	×	**51.64**	**86.80**	**94.66**	**49.99**
M10	×	√	×	√	48.83	85.43	92.01	48.13
M11	√	√	√	×	50.17	85.46	92.30	49.08
M12	√	×	√	√	49.65	86.41	93.48	49.16
M13	×	√	√	√	49.33	85.25	92.82	48.55
M14	√	√	√	√	49.22	84.30	91.11	48.66

Based on the results in Table 4, we can draw the following conclusions.

1. Compared with the baseline (M1), the SGA module could improve the performance of the network. At r = 1, the results of M2-M10 are better than the baseline.
2. The network of two SGA modules gives better results. For long-range dependency, two SGA modules can better preserve this relationship, which could mine the discriminative features.
3. Experiments prove that the extracted features of the network proposed in this paper are more discriminative. The rank 1 accuracy reaches 51.64%. The framework in this paper is optimal.

5.3. Visualization Experiment

One of the keys to RGB–IR person Re-ID is to improve the discriminability of features. To further illustrate the effectiveness of the SGA module, which can focus on the parts of interest, we visualize the features maps learned by MFCNet. We apply Grad-Cam [52] to visualize these areas by highlighting them on the image.

Figure 4 illustrates the key area for the four pedestrian images of one person. In the RGB–IR person Re-ID, the information such as color and face cannot be used, so the upper limbs and lower limbs of the human body are the most critical features. As shown in Figure 4, our network is more interested in these discriminative parts. This proves the SGA module proposed in this paper focuses on the key parts of the input image to obtain robust feature.

5.4. Comparison of Different Attention Mechanisms

The attention mechanism plays an important role in the proposed MFCNet. In order to prove the advantages of the proposed SGAs, we compared them with existing attention methods, including CBAM [16], SE [17], and non-local block [19]. For a fair comparison, the different attention mechanism modules that work on channels or spatially are used to replace the modules that work in the same way. As shown in Table 5, compared with SGA module, the SE block and CBAM drop the r = 1 accuracy by 4.66% and 4.29%, respectively. When a non-local block is replaced by SGA, the model can not only maintain the long-range

dependency of features but mines the relationship in the spatial domain, and the r = 1 accuracy drop 4.14%, respectively. The result indicates that SGA module outperform the other attention methods for visible-infrared person Re-ID.

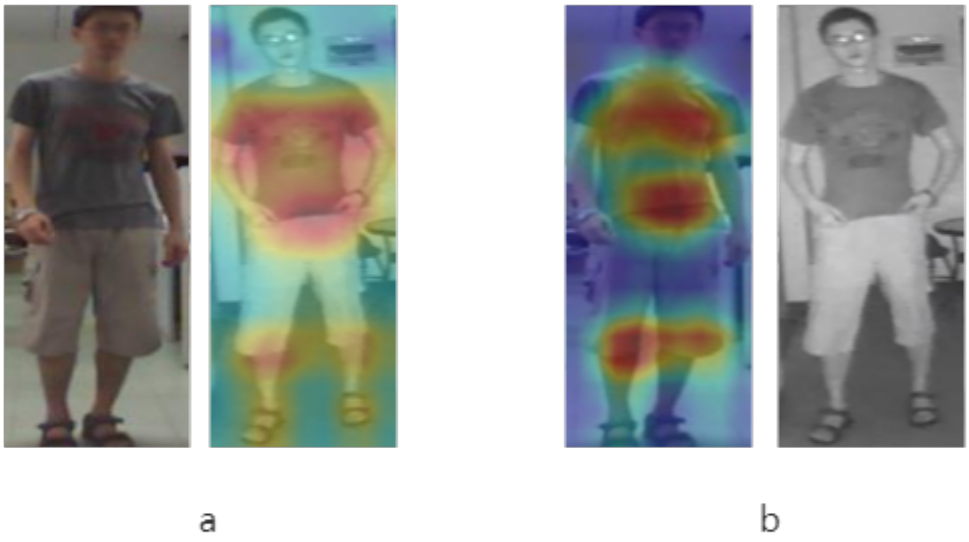


Figure 4. Feature Map Visualization. The picture shows two images of one identity. (**a**) uses IR images to search RGB image. (**b**) is the opposite

Table 5. Comparison of different attention mechanisms.

Settings	**All-Search**			
Methods	**r = 1**	**r = 10**	**r = 20**	**mAP**
SE	46.98	82.68	90.82	46.76
CBAM	47.35	84.54	91.87	46.97
Non-local	47.50	84.39	92.14	47.65
SGA	51.64	86.80	94.66	49.99

6. Conclusions

This paper studies the problem of cross-modality person re-identification, which is an important problem in many specific practical monitoring applications. This paper proposes a dual-path attention network architecture MFCNet to learn discriminative shared features and improve the accuracy. An attention module called SGA is designed, and the SGA module is embedded into the backbone network to focus on the key information in input images and mine the feature context. Experiments based on the public SYSU and RegDB datasets verify that the model designed in this paper performs better than the existing state-of-the-art methods with various settings.

Comprehensive experimental results have demonstrated our approach outperforms state-of-the-art methods, but the accuracy is far away from other recognition tasks such as facial recognition. This paper uses self-attention to mine the pixel–pixel feature context and obtain the shared feature, but ignore the relationship in part within each modality. In the future, we will explore the not only learning of part-aggregated features by mining the contextual part relations within each modality, but also more efficient attention network architectures, to solve RGB–IR Re-ID. The current methods for domain adaptation person re-identification [53–56] are mostly for the RGB–RGB datasets, and there is little research on RGB–IR datasets. Hence, we will pay attention to the RGB–IR domain adaptation and try to solve this problem.

Author Contributions: Conceptualization, J.M. Methodology, J.M. Writing—original draft preparation, J.M. Writing—review and editing, Y.L. and Y.H. Supervision, H.X. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, L.; Yang, Y.; Tian, Q. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1224–1244. [CrossRef] [PubMed]
2. Ye, M.; Liang, C.; Yu, Y.; Wang, Z.; Leng, Q.; Xiao, C.; Chen, J.; Hu, R. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Trans. Multimed.* **2016**, *18*, 2553–2566. [CrossRef]
3. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
4. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
5. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
6. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Vrstc: Occlusion-free video person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
7. Yu, H.; Zheng, W.; Wu, A.; Guo, X.; Gong, S.; Lai, J. Unsupervised person re-identification by soft multilabel learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
8. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.-S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
9. Wu, A.; Zheng, W.; Yu, H.; Gong, S.; Lai, J. Rgb-infrared cross-modality person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
10. Xiang, X.; Lv, N.; Yu, Z.; Zhai, M.; Saddik, A. Cross-Modality Person Re-Identification Based on Dual-Path Multi-Branch Network. *IEEE Sens. J.* **2019**, *19*, 11706–11713. [CrossRef]
11. Chen, D.; Li, H.; Liu, X.; Shen, Y.; Shao, J.; Yuan, Z.; Wang, X. Improving deep visual representation for person re-identification by global and local image-language association. In Proceedings of the 15th European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018.
12. Gong, Y.; Zhang, Y.; Poellabauer, C. Second-order non-local attention networks for person re-identification. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
13. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Interaction-and-aggregation network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019.
14. Wang, J.; Zhu, X.; Gong, S.; Li, W. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
15. Fang, P.; Zhou, J.; Roy, S.; Petersson, L.; Harandi, M. Bilinear attention networks for person retrieval. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
16. Li, S.; Bak, S.; Carr, P.; Wang, X. Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018.
17. Liu, C.-T.; Wu, C.-W.; Wang, Y.-C.F.; Chien, S.-Y. Spatially and temporally efficient non-local attention network for video-based person re-identification. In Proceedings of the 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, 9–12 September 2019.
18. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-Identification: A Survey and Outlook. Available online: https://arxiv.org/pdf/2001.04193 (accessed on 14 November 2021).
19. Ye, M.; Shen, J.; Crandall, D.J.; Shao, L.; Luo, J. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-identification. In Proceedings of the 2020 European Conference on Computer Vision, ECCV 2020, Glasgow, UK, 23–28 August 2020.
20. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Chua, T. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
21. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]

22. Jaderberg, M.; Simonyan, K.; Andrew Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the Neural Information Processing Systems, Montreal, QB, Canada, 7–12 December 2015.
23. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
24. Ye, M.; Lan, X.; Wang, Z.; Yuen, P.C. Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 407–419. [CrossRef]
25. Hao, Y.; Wang, N.; Li, J.; Gao, X. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
26. Liu, H.; Cheng, J.; Wen, W.; Su, Y.; Bai, H. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing* **2020**, *398*, 11–19. [CrossRef]
27. Zhao, Y.; Lin, J.; Xuan, Q.; Xi, X. HPLIN: A feature learning framework for cross-modality person re-identification. *IET Image Process.* **2020**, *13*, 2897–2904. [CrossRef]
28. Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; Tao, D. Hetero-Center Loss for Cross-Modality Person Re-Identification. *Neurocomputing* **2020**, *386*, 97–109. [CrossRef]
29. Feng, Z.; Lai, J.; Xie, X. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 579–590. [CrossRef] [PubMed]
30. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-Modality Person Re-Identification With Shared-Specific Feature Transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
31. Li, D.; Wei, X.; Hong, X.; Gong, Y. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QB, Canada, 8–13 December 2014.
33. Dai, P.; Ji, R.; Wang, H.; Wu, Q.; Huang, Y. Cross-Modality Person Re-Identification with Generative Adversarial Training. In Proceedings of the 2018 International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, 13–19 July 2018.
34. Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; Hou, Z. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
35. Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.; Satoh, S. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
36. Choi, S.; Lee, S.; Kim, Y.; Kim, T.; Kim, C. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
37. Wang, G.; Yang, T.; Yang, Y.; Yang, T.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
39. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
40. Park, J.; Woo, S.; Lee, J.; Kweon, I. BAM: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
41. Park, J.; Woo, S.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018.
42. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the International Conference on Computer Vision Workshop, Seoul, Korea, 27 October–2 November 2019.
43. Filip, R.; Giorgos, T.; Ondrej, C. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 1655–1668.
44. Zhang, S.; Yang, Y.; Wang, P.; Zhang, X.; Zhang, Y. Attend to the Difference: Cross-Modality Person Re-identification via Contrastive Correlation. *arXiv* **2019**, arXiv:1910.11656.
45. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005.
46. Ye, M.; Lan, X.; Li, J.; Yuen, P.C. Hierarchical discriminative learning for visible thermal person re-identification. In Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018.
47. Ye, M.; Lan, X.; Leng, Q.; Shen, J. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Trans. Image Process.* **2020**, *29*, 9387–9399. [CrossRef] [PubMed]
48. Gao, G.; Shao, H.; Yu, Y.; Wu, F.; Yang, M. Leaning Compact and Representative Features for Cross-Modality Person Re-Identification. *arXiv* **2021**, arXiv:2103.14210.

49. Liang, W.; Wang, G.; Lai, J.; Xie, X. Homogeneous-to-Heterogeneous: Unsupervised Learning for RGB-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 6392–6407. [CrossRef] [PubMed]
50. Wei, Z.; Yang, X.; Wang, N.; Gao, X. Flexible Body Partition-Based Adversarial Learning for Visible Infrared Person Re-Identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–12. [CrossRef] [PubMed]
51. Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; Sun, Z. Neural Feature Search for RGB-Infrared Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
52. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
53. Ma, A.J.; Yuen, P.C.; Li, J. Domain transfer support vector ranking for person re-identification without target camera label information. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
54. Li, H.; Chen, Y.; Tao, D.; Yu, Z.; Qi, G. Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1480–1494. [CrossRef]
55. Li, H.; Dong, N.; Yu, Z.; Tao, D.; Qi, G. Triple Adversarial Learning and Multi-view Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, C2. [CrossRef]
56. Zhu, Z.; Luo, Y.; Chen, S.; Qi, G.; Mazur, N.; Zhong, C.; Li, Q. Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification. *J. Vis. Commun. Image Represent.* **2021**, *80*, 103303. [CrossRef]