applied sciences

MDPI

Article

Minimizing Maximum Feature Space Deviation for Visible-Infrared Person Re-Identification

Zhixiong Wu [1] and Tingxi Wen [2,*]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] College of Engineering, Huaqiao University, Quanzhou 362021, China
* Correspondence: t.wen@hqu.edu.cn

Abstract: Visible-infrared person re-identification (VIPR) has great potential for intelligent video surveillance systems at night, but it is challenging due to the huge modal gap between visible and infrared modalities. For that, this paper proposes a minimizing maximum feature space deviation (MMFSD) method for VIPR. First, this paper calculates visible and infrared feature centers of each identity. Second, this paper defines feature space deviations based on these feature centers to measure the modal gap between visible and infrared modalities. Third, this paper minimizes the maximum feature space deviation to significantly reduce the modal gap between visible and infrared modalities. Experimental results show the superiority of the proposed method, e.g., on the RegDB dataset, the rank-1 accuracy reaches 92.19%.

Keywords: deep learning; feature space maximum deviation; visible-infrared person re-identification

Citation: Wu, Z.; Wen, T. Minimizing Maximum Feature Space Deviation for Visible-Infrared Person Re-Identification. *Appl. Sci.* **2022**, *12*, 8792. https://doi.org/10.3390/app12178792

Academic Editors: Xiaobin Zhu and Jianqing Zhu

Received: 25 May 2022
Accepted: 19 August 2022
Published: 1 September 2022

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

1. Introduction

Given a visible (or infrared) query image of a specified person, the goal of Visible-infrared person re-identification [1–6] is to retrieve infrared (or visible) images of the same person from a gallery set, as shown in Figure 1. VIPR has received more and more attention due to its importance to intelligent surveillance systems and intelligent transportation systems in light-less environments. In these application scenarios, visible images are captured by color cameras and infrared images usually are captured by near-infrared [7] or thermal [8] cameras. Just as with the traditional single-modal (i.e., visible) re-identification task [9–11], VIPR meets the great challenges of pose variation, viewpoint variation, and occlusions. Furthermore, VIPR suffers from a huge modal gap between visible and infrared modalities. Therefore, VIPR is a meaningful but challenging topic that is worthy of intensive study.

Figure 1. The schematic diagram of visible-infrared person re-identification.

Existing VIPR works can be roughly divided into three types: (1) metric learning-based methods [3,12–14], (2) feature extraction-based methods [2,15–21], and (3) generation-based methods [22–25]. Many metric learning-based methods usually design specific triplet loss functions to optimize cross-modal samples. However, metric learning-based methods focus on identity-level discrimination, underestimating threats of outliers raised by the gap between visible and infrared modalities. Feature extraction-based methods are often dedicated to designing additional special feature learning architectures (e.g., visual

attention modules) to enhance general deep networks (e.g., residual networks [26]) to map different modal features into a common feature space. However, extra computations caused by additional special feature learning architectures are required.

Generation-based methods use generation adversarial networks (GAN) for modal conversion to eliminate modal gaps. For example, in addition to using ResNet-50 [26], Zhong et al. [22] designed a GAN to produce synthetic color images from infrared and gray images. The GAN contains a generator of five convolutional layers and five convolution-transpose layers, and each layer is followed by a batch normalization layer, a leaky rectified linear unit, and a residual connection. The discriminator is a fully convolutional network, mainly consisting of a series of 3×3 convolutional layers, where each convolutional layer is followed by a max-pooling layer, and the number of channels doubles after each max-pooling layer. The generator is applied to generate synthetic color images, while the discriminator is used to discriminate an image is synthetic or real. Through the antagonism between generator and discriminator, the generator could generate realistic synthetic color images to reduce modal gaps. GANs of generators and discriminators are commonly used, such as [24,25,27,28]. In [24,25,27], multiple GANs are applied to realize more delicate adversarial learning. Leaving aside GAN's architecture complexities, in practice, GAN's training process is complex and involves many hyper-parameters that require stage-wise training or alternate training. Therefore, GAN-based methods require a lot of skill to avoid model collapses.

In this paper, we propose a minimizing maximum feature space deviation (MMFSD) method for VIPR. For modeling the modal gap between visible and infrared modalities, we define feature space deviations (FSD) based on each identity's visible and infrared feature centers. Their centers are calculated by averaging multiple samples on the same modality of the same identity. Furthermore, for reducing the modal gap between visible and infrared modalities, we design a minimizing maximum FSD loss function, which optimizes deviations from the most severe feature dimensions.

The main contributions of this paper can be summarized as follows. (1) A novel feature space deviation measurement is designed to measure the modal gap between visible and infrared modalities. (2) A minimizing maximum feature space deviation loss function is designed to significantly reduce modal gap between visible and infrared modalities. (3) Experimental results on SYSU-MM01 [7] and RegDB [8] datasets demonstrate that our method outperforms many state-of-the-art approaches.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method in detail. Section 4 presents experimental results to analyze our method's superiority. Section 5 concludes this paper.

2. Related Work

2.1. Metric Learning-Based Method

Some works [1,14,16,23,25] directly use the traditional triplet loss function [12,29], which suffers from the huge cross-modal gap in VIPR. Furthermore, many methods [5,13,30] are dedicated to designing new triplet loss functions that optimize cross-modal positive and negative pairs for VIPR. For example, Liu et al. [5] designed a hero-center triplet loss function to pull close positive feature centers from different modalities and push away negative feature centers from different and same modalities.

2.2. Feature Extraction-Based Method

Deep learning feature extraction architectures have been widely used in many computer vision tasks, such as image super-resolution [31], image detection [32], and live searching [33]. For VIPR, the common feature extraction method is to modify general deep learning feature extraction architectures into two-stream networks [1,2,14,16,34,35] whose shallow layers are independent to learn features from different modalities, and whose deep layers are shared to align cross-modal features. For example, Ye et al. [2] first introduced the two-stream network for deep cross-modal shareable feature learning, which contains a visible image stream and infrared image stream. Still, there are some methods [1,34] that

introduced an attention mechanism to the two-stream network, which learned intra-modal discrepancy and inter-modal correlation to achieve better feature alignment. Furthermore, Zhang et al. [18] used an attention module to aggregate part features to global features to enhance discrimination ability of feature learning. Such a part-level feature learning method is also applied in [14,16,17]. Nevertheless, these methods improve feature learning architecture via extra modules (e.g., attention [1,18,34,36] and transformer [35]) that usually consume additional computations.

2.3. Generation-Based Method

The main idea of the generation-based method [6,22,24,25,27,28] is to use generation adversarial network (GAN) for modal conversion, i.e., generating related modality or unified modal information. For example, Hu et al. [24] generated cross-modal paired-images for effective feature alignment. Wang et al. [27] translated the visible and infrared images to their infrared and visible counterpart, respectively, and then combined the original images and generated images to form multi-spectral images for feature learning. Wang et al. [28] directly obtained fake infrared images from visible images by GAN, which alleviated the modality discrepancy. The GAN-based method achieves style transfer for reducing modal variations. However, methods with GAN usually converge slowly and make re-identification module complex.

3. Proposed Method

In this section, we describe our minimizing maximum feature space deviation (MMFSD) method in detail, including: (1) feature space deviation modeling; (2) feature learning architecture; (3) total loss function design.

3.1. Feature Space Deviation Modeling

Let $\{x_{k,i}|x_{k,i} \in \mathbb{R}^{d\times 1}, k \in 1,2,\cdots,K, i = 1,2,\cdots,N\}$ and $\{y_{k,i}|y_{k,i} \in \mathbb{R}^{d\times 1}, k \in 1,2,\cdots,K, i = 1,2,\cdots,N\}$ denote features extracted from visible and infrared images, respectively, where K is the number of classes in the mini-batch, and N is the number of images in each class, and d represents the number of feature dimensions. The feature space deviation modeling progress is described as follows.

(1) Calculating visible and infrared centers of each class to construct the visible feature space $\mathcal{X}$ and the infrared feature space $\mathcal{Y}$ as follows.

$$\mathcal{X} = [\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \ldots, \overline{x}_K]^{\mathrm{T}} \in \mathbb{R}^{K\times d}, \tag{1}$$

$$\mathcal{Y} = [\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \ldots, \overline{y}_K]^{\mathrm{T}} \in \mathbb{R}^{K\times d}, \tag{2}$$

where

$$\overline{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{k,i} \in \mathbb{R}^{1\times d}, \tag{3}$$

$$\overline{y}_k = \frac{1}{N}\sum_{i=1}^{N} y_{k,i} \in \mathbb{R}^{1\times d}. \tag{4}$$

(2) Calculating distances of visible and infrared feature spaces on each single dimension as deviations. Let us re-symbolize the visible feature space $\mathcal{X}$ and the infrared feature space $\mathcal{Y}$ as follows.

$$\mathcal{X} = [p_1, p_2, \ldots, p_j, \ldots, p_d] \in \mathbb{R}^{K\times d}, \tag{5}$$

$$\mathcal{Y} = [q_1, q_2, \ldots, q_j, \ldots, q_d] \in \mathbb{R}^{K\times d}, \tag{6}$$

where $p_j \in \mathbb{R}^{K\times 1}$ and $q_j \in \mathbb{R}^{K\times 1}$ represent the column vectors of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, the distance of visible and infrared feature spaces on each single dimension defined as a feature space deviation as follows.

$$d_j = \|\hat{p}_j - \hat{q}_j\|_2 = \sqrt{\sum_{k=1}^{K} (\hat{p}_{j,k} - \hat{q}_{j,k})^2}, \tag{7}$$

(3) Designing a minimizing maximum feature space deviation (MMFSD) loss function. The MMFSD loss function is calculated as follows:

$$\mathcal{L}_{MMFSD} = \max_{k\in[1,2,\ldots,d]} \|\hat{p}_j - \hat{q}_j\|_2, \tag{8}$$

where $\hat{p}_j = \frac{\hat{p}_j}{\sqrt{\sum_{k=1}^{K} \hat{p}_{j,k}^2}}$ and $\hat{q}_j = \frac{\hat{q}_j}{\sqrt{\sum_{k=1}^{K} \hat{q}_{j,k}^2}}$ are L2 normalized features of p_j and q_j, respectively. Based on the MMFSD loss function, distance of the corresponding column vectors in the cross-modal feature spaces is narrowed, so that the maximum feature space deviation is optimized.

3.2. Feature Learning Architecture

Figure 2 shows the overall framework of VIPR. The two-branch network is used as the feature learning network, which adopts ResNet-50 [26] as a backbone, as performed in many existing works [1,5,37]. As shown in Figure 2, shallow layers of each branch, i.e., the first convolutional layer (Conv), the first residual group (Layer-1), and the second residual group (Layer-2), are un-shared to learning modal-specific features, which means that shallow layers of each modal feature have the same structure but independent parameters. There deep layers, i.e., the third residual group (Layer-3), the forth residual group (Layer-4), the generalized-mean pooling (GeM) layer, and the batch normalization (BN) layer are shared to exploit the modal-shared features. In the inference phase, L2 normalized features from GeM and BN layers are added to be final features for VIPR.

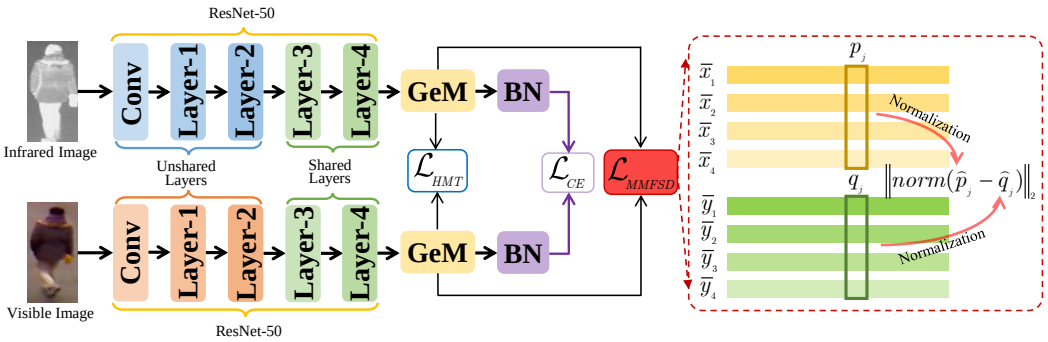

Figure 2. The overall framework of visible-infrared person re-identification.

3.3. Total Loss Function Design

In this paper, we combine the proposed MMFSD loss function with two commonly used loss functions, i.e., the hard-mining triplet (HMT) loss function [38] and the cross-entropy (CE) loss function [39], to supervise the training process. The total loss function is formulated as follows:

$$\mathcal{L}_{Total} = \lambda \mathcal{L}_{MMFSD} + \mathcal{L}_{HMT} + \mathcal{L}_{CE}, \tag{9}$$

where $\lambda > 0$ is a hyper-parameter used to control the contribution of MMFSD loss function; $\mathcal{L}_{HMT}$ and $\mathcal{L}_{CE}$ denotes HMT and CE loss functions, which are formulated as follows:

$$\mathcal{L}_{HMT} = \frac{1}{M}\sum_{m=1}^{M} log[1 + exp(\max_{f_i \in \mathcal{F}_m^+} \|f_m - f_i\|_2 - \min_{f_j \in \mathcal{F}_m^-} \|f_m - f_j\|_2)], \tag{10}$$

$$\mathcal{L}_{CE} = -\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} \epsilon_{m,k} log(p_{m,k}), \tag{11}$$

where $M = 2 \times N \times K$ represents the total number of samples in a mini-batch; K denotes the number of classes; N denotes the number of visible (or infrared) samples of each class; f_m represents the feature of m-th sample; $\mathcal{F}_m^+$ and $\mathcal{F}_m^-$ denotes the positive set and the negative set of the m-th sample f_m; $p_{m,k}$ represents the posterior probability of the m-th sample belongs to the k-th class (i.e., $Class_k$), which is calculated with the Softmax function; $\epsilon_{m,k}$ is the indicator function of smoothed class labels, which is formulated as follows.

$$\epsilon_{m,k} = \begin{cases} 1 - \frac{(K-1)\zeta}{K}, & x_m \in Class_k, \\ \frac{\zeta}{K}, & x_m \notin Class_k, \end{cases} \tag{12}$$

where ζ is the regularized parameter for label smoothing, and it is set to 0.1 as a common setting. Furthermore, as shown in Figure 2, the MMFSD loss function and the HMT loss function are assigned to the GeM layer, while the CE loss function is arranged on the BN layer, which is so-called batch normalization neck (BNNeck) [29].

4. Experiments and Analysis

In this paper, we evaluate our method on two popular datasets, i.e., SYSU-MM01 [7] and RegDB [8] datasets. Following existing works [5,12,34], mean average precision (mAP), mean inverse negative penalty (mINP), and cumulative match characteristic (CMC) curve are applied as the VIPR's performance metrics. Rank1 represents the rank-1 accuracy in a CMC curve.

4.1. Datasets

The SYSU-MM01 dataset [7] contains 491 pedestrian subjects captured by four visible cameras and two infrared cameras. The training set includes 22,258 visible images and 11,909 infrared images of 395 subjects. In the testing processing, there are two test modes, i.e., all-search mode and indoor search mode. In the all-search mode, the query set contains 3803 infrared images of 96 subjects to search from a gallery subset containing 301 visible images of the same subjects. In the indoor-search mode, the query set contains 3803 infrared images of 96 subjects which is same as all-search mode, while the gallery set includes 112 visible images of the same subjects. Following [7], the single-shot evaluation protocol is used, and the final results are based on an average of 10 tests with randomly selected gallery images.

The RegDB dataset [8] includes 412 pedestrian classes, and each class contains ten visible images and ten infrared images captured by the overlapping visible and infrared cameras. Following [1,12,17], this dataset is randomly split into ten trials. In each trial, the training set contains 206 classes, while the testing set includes the non-overlapping 206 classes. The final results are based on an average of 10 tests with 10 trials. Both visible-to-infrared and infrared-to-visible retrieval are evaluated.

4.2. Implementation Details

We employ one GeForce RTX 3090 GPU and the Pytorch [40] deep learning tool to implement our VIPR method. The batch size is set 40, which contains four subjects and each subject has five visible images and five infrared images. All images are resized to 144×288 pixels in the training and testing process. The ImageNet [41] pre-trained ResNet-50 is applied as a backbone. We used the mini-batch stochastic gradient descent (SGD)

optimizer [42] with the weight decays set to 0.0005 and the momentum set to 0.9 for training. There are 90 epochs in the training process. The initial learning rate is set 0.001. In the first 10 epochs, the learning rate is linearly warmed up to 0.01. After warming up, the learning rate is maintained at 0.01 from the 11st to 30th epochs. The learning rate is decayed to 10% every 20 epochs.

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

In this section, we make a comparison with several state-of-the-arts methods to validate the superiority of the proposed method. The comparison results are shown in Tables 1 and 2.

Table 1. The performance comparison on the SYSU-MM01 dataset.

Method	All-Search Mode		Indoor-Search Mode		Reference
	Rank1 (%)	mAP (%)	Rank1 (%)	mAP (%)	
BDTR [3]	17.01	19.66	N/A	N/A	IJCAI 2018
cmGAN [6]	26.97	27.80	31.63	42.19	IJCAI 2018
D-HSME [43]	20.68	23.12	N/A	N/A	AAAI 2019
D^2RL [27]	28.9	29.2	N/A	N/A	CVPR 2019
AlignGAN [28]	42.4	40.7	45.9	54.3	ICCV 2019
eBDTR [13]	27.82	28.42	N/A	N/A	TIFS 2020
CE^2L [4]	29.52	28.4	N/A	N/A	ICPR 2020
Hi-CMD [25]	34.9	35.9	N/A	N/A	CVPR 2020
DGD+MSR [44]	37.35	38.11	39.64	50.88	TIP 2020
JSIA-ReID [23]	38.1	36.9	43.8	52.9	AAAI 2020
MSPAC-MeCen [18]	46.62	47.26	51.63	61.54	ICPR 2020
X modality [20]	49.92	50.73	N/A	N/A	AAAI 2020
MACE [37]	51.64	50.11	57.35	64.79	TIP 2020
LAND [14]	53.6	52.0	57.0	63.2	IOT 2020
DDAG [1]	54.75	53.02	61.02	68.0	ECCV 2020
FBP-AL [17]	43.78	42.91	N/A	N/A	TNNLS 2021
AGW [12]	47.50	47.65	54.17	62.97	TPAMI 2021
LLM [30]	55.25	52.96	59.65	65.46	SPL 2021
ADCNet [24]	55.9	59.6	58.8	65.6	ICME 2021
DMiR [45]	50.54	49.29	53.92	62.49	TCSVT 2022
GECNet [22]	53.37	51.83	60.60	62.89	TCSVT 2022
DFLN-ViT [35]	59.84	**57.70**	62.13	69.03	TMM 2022
MMFSD	**60.64**	57.54	**62.93**	**69.27**	Ours

4.3.1. Comparison on SYSU-MM01

As shown in Table 1, we can see the experimental results on SYSU-MM01 dataset show that the proposed FSMD method obtains the great results in both all-search and indoor-search mode. Specifically, the proposed FSMD method obtains 60.64% Rank1 and 57.54% mAP in all-search mode, while obtains 62.93% Rank1 and 69.27% mAP in indoor-search mode. Compared with the feature extracting method DFLN-ViT [35], the proposed method obtains 0.80% higher Rank1 in both all-search and indoor-search mode. Compared with GAN-based method GECNet [22], the proposed method defeats GECNet [22] by 7.27% in terms of Rank1 and 5.71% in terms of mAP in all-search mode. The improvement demonstrates the superiority of the proposed FSMD method.

Table 2. The performance comparison on RegDB dataset.

Method	Visible-to-Infrared Retrieval		Infrared-to-Visible Retrieval		Reference
	Rank1 (%)	mAP (%)	Rank1 (%)	mAP (%)	
BDTR [3]	33.47	31.83	N/A	N/A	IJCAI 2018
D^2RL [27]	43.3	44.1	N/A	N/A	CVPR 2019
D-HSME [43]	50.85	47.00	50.15	46.16	AAAI 2019
AlignGAN [28]	57.9	53.6	56.3	53.4	ICCV 2019
eBDTR [13]	31.83	33.18	N/A	N/A	TIFS 2020
CE^2L [4]	47.50	44.21	N/A	N/A	ICPR 2020
DGD+MSR [44]	48.43	48.67	N/A	N/A	TIP 2020
JSIA-ReID [23]	48.5	49.3	48.1	48.9	AAAI 2020
MSPAC-MeCen [18]	49.61	53.64	N/A	N/A	ICPR 2020
X modality [20]	62.21	60.18	N/A	N/A	AAAI 2020
DDAG [1]	69.34	63.46	68.06	61.80	ECCV 2020
Hi-CMD [25]	70.93	66.04	N/A	N/A	CVPR 2020
cm-SSFT [15]	72.3	72.9	71.0	71.7	CVPR 2020
MACE [37]	72.37	69.09	72.12	68.57	TIP 2020
LADN [14]	75.7	72.9	75.3	73.0	IoT 2020
AGW [12]	70.05	66.37	N/A	N/A	TPAMI 2021
ADCNet [24]	72.9	66.5	72.4	65.3	ICME 2021
FBP-AL [17]	73.98	68.24	70.05	66.61	TNNLS 2021
LLM [30]	74.85	71.32	N/A	N/A	SPL 2021
SFANet [19]	76.31	68.00	70.15	63.77	TNNLS 2021
GECNet [22]	82.33	78.45	78.93	75.58	TCSVT 2021
MPANet [46]	83.7	80.9	82.8	80.7	CVPR 2021
MSA [21]	84.86	82.16	N/A	N/A	IJCAI 2021
HC-Triplet [5]	91.05	83.28	89.30	81.46	TMM 2021
GLMC [16]	91.84	81.42	91.12	81.06	TNNLS 2021
DMiR [45]	75.79	69.97	73.93	68.22	TCSVT 2022
DTRM [34]	79.09	70.09	78.02	69.56	TIFS 2022
DFLN-ViT [35]	92.10	82.11	**91.21**	81.62	TMM 2022
MMFSD	**92.19**	**85.95**	90.65	**84.38**	Ours

4.3.2. Comparison on RegDB

The experimental results on the RegDB dataset in Table 2 demonstrate the proposed FSMD method achieves competitive performance in both visible-to-infrared and infrared-to-visible retrieval. In visible-to-infrared retrieval, the proposed FSMD method obtains 92.19% Rank1 and 85.95% mAP, defeating attention-based method DTRM [34] by 13.10% Rank1 and 15.68% mAP. In infrared-to-visible retrieval, the proposed FSMD method obtains 90.65% Rank1 and 84.38% mAP, defeating id-level loss function HC-triplet [5] by 1.35% Rank1 and 2.74% mAP. These comparative results validate the effectiveness of the proposed FSMD method.

4.4. Role of Minimizing Feature Space Maximum Deviation Loss Function

We adjust the λ in Equation (9) to evaluate the role of the MFSMD loss function. Table 3 and Figure 3 show the performance on SYSU-MM01 dataset, and Table 4 and Figure 4 show the performance on RegDB dataset.

Table 3. The performance comparison of using different weights of MMFSD loss function (i.e., λ in Equation (9)) on the SYSU-MM01 dataset.

λ	All Search			Indoor Search		
	Rank1 (%)	mAP (%)	mINP (%)	Rank1 (%)	mAP (%)	mINP (%)
0	56.04	54.71	41.96	59.95	67.23	63.35
0.25	**60.64**	**57.54**	**43.31**	62.93	**69.27**	**64.89**
0.5	57.96	54.64	40.15	61.87	67.66	62.82
1	59.52	55.97	41.31	**63.00**	69.08	64.32
2	55.27	51.42	35.66	54.21	61.92	57.09

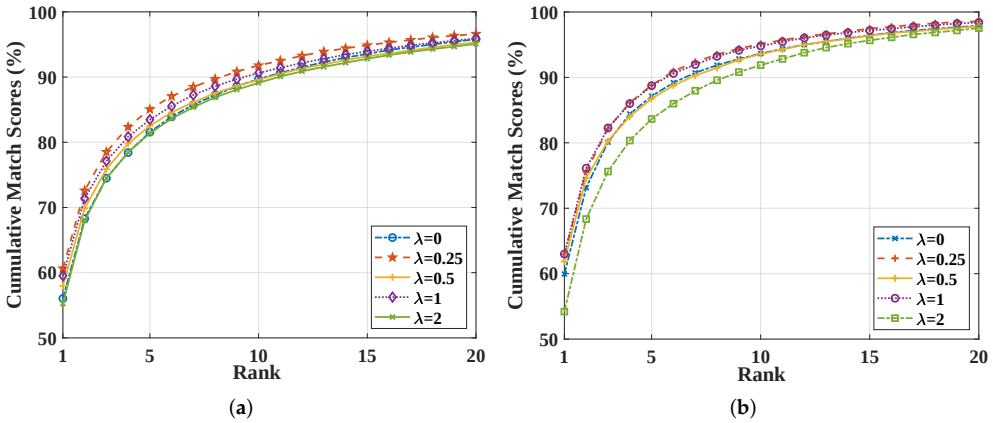

Figure 3. The CMC curves in different MMFSD's weights (i.e., λ in Equation (9)) on SYSU-MM01 dataset of (**a**) all-search mode and (**b**) indoor-search mode.

Table 4. The performance comparison of using different weights of MMFSD loss function (i.e., λ in Equation (9)) on the RegDB dataset.

λ	Visible to Infrared			Infrared to Visible		
	Rank1 (%)	mAP (%)	mINP (%)	Rank1 (%)	mAP (%)	mINP (%)
0	85.12	80.44	68.67	84.22	79.16	65.75
0.25	90.94	85.25	73.26	89.81	83.72	69.58
0.5	**92.19**	**85.95**	**73.86**	**90.65**	**84.38**	**70.73**
1	90.18	81.49	65.46	88.27	79.23	61.65
2	88.02	79.71	63.97	86.45	78.48	61.48

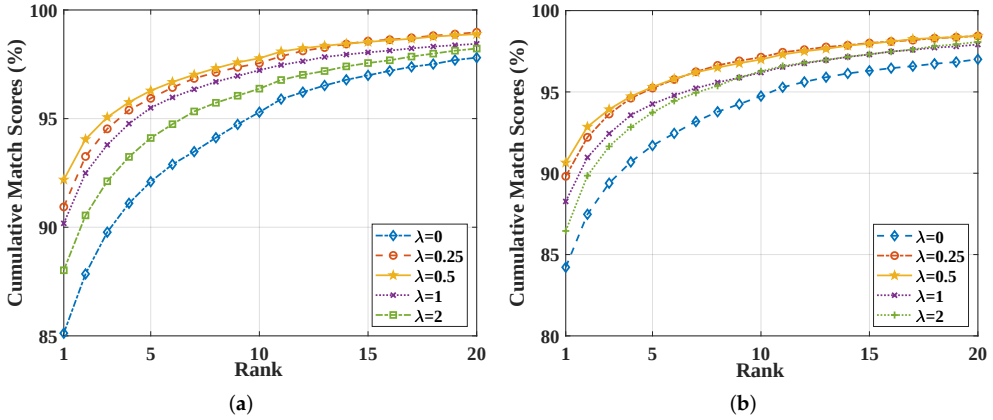

Figure 4. The CMC curves in different MMFSD's weights (i.e., λ in Equation (9)) on RegDB dataset of (**a**) visible-to-infrared retrieval and (**b**) infrared-to-visible retrieval.

On the SYSU-MM01 dataset, from Table 3 and Figure 3, it can be found that the proposed FSMD loss function achieves great performance. For example, compared with the case without FSMD loss function (i.e., $\lambda = 0$), the Rank1, mAP and mINP obtain better results when $\lambda = 0.25$ on all-search mode (i.e., 60.64% Rank1, 57.54% mAP, and 43.31% mINP). It demonstrates the effectiveness of the FSMD loss function.

On the RegDB dataset, Table 4 and Figure 4 also show the effectiveness of the FSMD loss function. Specifically, when adjusting λ from 0.25 to 2, the Rank1, mAP, and mINP are higher than those as $\lambda = 0$ on both visible-to-infrared retrieval and infrared-to-visible retrieval. Among them, when $\lambda = 0.5$, it obtains the best performance, i.e., 92.19% Rank1, 85.95% mAP and 72.86% mINP on visible-to-infrared retrieval, and 90.65% Rank1, 84.38% mAP and 70.73% mINP on infrared-to-visible retrieval.

By comparing with the results on the SYSU-MM01 and RegDB datasets, one can see that the best λ value of RegDB is bigger than that of SYSU-MM01. Furthermore, comparing with Figures 3 and 4, we can find that the CMC curves' variation on RegDB dataset is larger than that on SYSU-MM01 dataset with adjusting the λ. This is because, compared with SYSU-MM01 dataset which suffer from multiple intra-class variation (e.g., pose variation, view variation and modal variation), the modal discrepancy on RegDB dataset is the most significant problem. Therefore, it further demonstrates the FSMD's outstanding performance on deal with modal discrepancy.

5. Conclusions

In this paper, we design a novel feature space deviation measurement for modeling the huge modal gap between visible and infrared images. Furthermore, a minimizing maximum feature space deviation loss function is designed to reduce modal gap between visible and infrared modalities. Experiments on the SYSU-MM01 and RegDB datasets demonstrate that the proposed MMFSD loss function is helpful to improve VIPR performance, outperforming many state-of-the-arts methods.

In the future work, we will study how to deal with partially occluded person images caused by over-saturating infrared sensors in high temperature environments, e.g., we will try to design an automatic data augmentation method to eliminate the adverse effects of high temperatures on infrared sensor imaging. Moreover, we will explore the application of feature space deviation measurements, e.g., extracting modal-common features from infrared and visible images via suppressing feature space deviations and then decoding modal-common features back to image spaces for infrared-visible image fusion to enhance imaging effect.

Author Contributions: Conceptualization, Z.W. and T.W. Writing—original draft preparation, Z.W. Writing—review and editing, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was found by Natural Science Foundation of Fujian under that Grant 2020J01086.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

VIPR	Visible-infrared person re-identification
FSD	Feature space deviation
MMFSD	Minimizing maximum feature space deviation
GAN	Genaration adversarial network
Conv	Convolutional layer
GeM	Generalized-mean pooling
BN	Batch normalization
HMT	Hard-mining triplet
CE	Cross-entropy
BNNeck	Batch normalization neck
mAP	Mean average precision
CMC	Cumulative match characteristic
Rank1	rank-1 accuracy
SGD	Stochastic gradient descent

References

1. Ye, M.; Shen, J.; J. Crandall, D.; Shao, L.; Luo, J. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-identification. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 229–247.
2. Ye, M.; Lan, X.; Li, J.; C., Y.P. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7501–7508.
3. Ye, M.; Wang, Z.; Lan, X.; Yuen, P.C. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 1092–1099.
4. Dai, H.; Xie, Q.; Ma, Y.; Liu, Y.; Xiong, S. RGB-Infrared Person Re-identification via Image Modality Conversion. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 592–598.
5. Liu, H.; Tan, X.; Zhou, X. Parameter Sharing Exploration and Hetero-center Triplet Loss for Visible-Thermal Person Re-Identification. *IEEE Trans. Multimed.* **2021**, *23*, 4414–4425.
6. Dai, P.; Ji, R.; Wang, H.; Wu, Q.; Huang, Y. Cross-Modality Person Re-Identification with Generative Adversarial Training. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 677–683.
7. Wu, A.; Zheng, W.S.; Yu, H.X.; 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; pp. 5390–5399.
8. Nguyen, D.T.; Hong, H.G.; Kim, K.W.; Park, K.R. Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras. *Sensors* **2017**, *17*, 605. [CrossRef] [PubMed]
9. Zhu, J.; Liu, L.; Zhu, X.; Zeng, H. A spatial structural similarity triplet loss for auxiliary vehicle re-identification. *Sciece China Inf. Sci.* **2021**, *64*, 179104. [CrossRef]
10. Zhu, J.; Huang, J.; Zeng, H.; Ye, X.; Li, B.; Lei, Z.; Zheng, L. Object reidentification via joint quadruple decorrelation directional deep networks in smart transportation. *IEEE Internet Things J.* **2020**, *7*, 2944–2954. [CrossRef]
11. Zhu, J.; Zeng, H.; Huang, J.; Zhu, X.; Lei, Z.; Cai, C.; Zheng, L. Body symmetry and part-locality-guided direct nonparametric deep feature enhancement for person reidentification. *IEEE Internet Things J.* **2019**, *7*, 2053–2065.
12. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893.
13. Liu, H.; Tan, X.; Zhou, X. Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 407–419.
14. Liu, S.; Zhang, J. Local Alignment Deep Network for Infrared-Visible Cross-Modal Person Re-identification in 6G-Enabled Internet of Things. *IEEE Internet Things J.* **2021**, *8*, 15259–15266.
15. 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; pp. 13376–13386.
16. Zhang, L.; Du, G.; Liu, F.; Tu, H.; Shu, X. Global-Local Multiple Granularity Learning for Cross-Modality Visible-Infrared Person Reidentification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–11. [CrossRef]
17. 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]
18. Can, Z.; Hong, L.; Wei, G.; Mang, Y. Multi-Scale Cascading Network with Compact Feature Learning for RGB-Infrared Person Re-Identification. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 8679–8686.
19. Liu, H.; Ma, S.; Xia, D.; Li, S. SFANet: A Spectrum-Aware Feature Augmentation Network for Visible-Infrared Person Reidentificatione. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–14. [CrossRef]
20. 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; pp. 4610–4617.
21. Miao, Z.; Liu, H.; Shi, W.; Xu, W.; Ye, H. Modality-aware Style Adaptation for RGB-Infrared Person Re-Identification. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021; pp. 916–922.
22. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C.W. Grayscale Enhancement Colorization Network for Visible-infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1418–1430.
23. Wang, G.A.; Zhang, T.; Yang, Y.; 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; pp. 12144–12151.
24. Hu, B.; Liu, J.; Zha, Z.j. Adversarial Disentanglement and Correlation Network for Rgb-Infrared Person Re-Identification. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–9 July 2021; pp. 1–6.
25. Seokeon, C.; Lee, S.; Kim, Y.; Kim, C. Hi-CMD:Hiererchical 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; pp. 10254–10263.
26. 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; pp. 770–778.

27. Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.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, 16–17 June 2019; pp. 618–626.
28. Wang, G.A.; 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 IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3622–3631.
29. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1487–1495.
30. Feng, Y.; Xu, J.; Ji, Y.m.; Wu, F. LLM: Learning Cross-Modality Person Re-Identification via Low-Rank Local Matching. *IEEE Signal Process. Lett.* **2021**, *28*, 1789–1793.
31. Zhu, X.; Li, Z.; Lou, J.; Shen, Q. Video Super-Resolution Based on a Spatio-Temporal Matching Network. *Pattern Recognit.* **2020**, *110*, 107619.
32. Zhu, X.; Li, Z.; Li, X.; Li, S.; Dai, F. Attention-aware Perceptual Enhancement Nets for Low-Resolution Image Classification. *Inf. Sci.* **2020**, *515*, 233–247.
33. Bianco, V.; Mazzeo, P.L.; Paturzo, M.; Distante, C.; Ferraro, P. Deep learning assisted portable IR active imaging sensor spots and identifies live humans through fire. *Opt. Lasers Eng.* **2020**, *124*, 105818.
34. Ye, M.; Chen, C.; Shen, J.; Shao, L. Dynamic Tri-Level Relation Mining with Attentive Graph for Visible Infrared Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 386–398.
35. Zhao, J.; Wang, H.; Zhou, Y.; Yao, R.; Chen, S.; El Saddik, A. Spatial-Channel Enhanced Transformer for Visible-Infrared Person Re-Identification. *IEEE Trans. Multimed.* **2022**, 1. [CrossRef]
36. Hou, J.B.; Zhu, X.; Liu, C.; Yang, C.; Yin, X.C. Detecting Text in Scene and Traffic Guide Panels with Attention Anchor Mechanism. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 6890–6899.
37. 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.
38. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.
39. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3774–3782.
40. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
41. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, South Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
43. 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; pp. 8385–8392.
44. Feng, Z.; Lai, J.; Xie, X. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 579–590.
45. Hu, W.; Liu, B.; Zeng, H.; Hu, H. Adversarial Decoupling and Modality-invariant Representation Learning for Visible-Infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5095–5109. [CrossRef]
46. Qiong, W.; Pingyang, D.; Jie, C.; Chia-Wei, L.; Yongjian, W.; Feiyue, H.; Bineng, Z.; Rongrong, J. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4328–4337.