Learn Robust Pedestrian Representation within Minimal Modality Discrepancy for Visible-Infrared Person Re-Identification

Yujie Liu, Wenbin Shao*, Xiaorui Sun
China University of Petroleum(East China)
Qingdao,Shandong Province,China
liuyujie@upc.edu.cn,{wbShao,sun}@s.upc.edu.cn

Abstract

Visible-Infrared person re-identification has attracted extensive attention from the community due to its potential great application prospects in video surveillance. There is huge modality discrepancy between visible and infrared images caused by different imaging mechanisms. Existing works alleviate modality discrepancy by aligning modality distribution or extracting modality-shared features on the original image. However, they ignore a key solution——converting visible images to gray images directly——which is computing-saving and effective to reduce modality discrepancy. In this paper, we transform the cross-modality person re-identification task from Visible-Infrared to Gray-Infrared which is named as minimal modality discrepancy. In addition , we propose a Pyramid Feature Integration Network (PFINet), which mines the discriminative refined features of pedestrian images and fuses high-level semantically stronger features to build a robust pedestrian representation. Specifically, PFINet first performs feature extraction from concrete to abstract and top-down semantic transfer to obtain multi-scale feature maps. Then, the multi-scale feature maps are input to Discriminative-Region Response(DRR) module to emphasize the identity-discriminative region by the spatial attention mechanism. Finally, Pedestrian representation is obtained by feature integration. Extensive experiments on the two public datasets SYSU-MM01 and RegDB demonstrate the effectiveness of PFINet that outperforming the state-of-the-arts dramatically.

1. Introduction

Person re-identification is an image retrieval problem, which aims to retrieve images with the same identity from gallery for a given query image. These images are usually captured from multiple non-overlapping cameras. The discrepancy caused by the variations of viewpoints, body poses, and occlusion pose great challenges to person re-identification. Despite the above-mentioned difficulties, the visible image based person re-identification has made great progress and achieved high accuracy with the development of deep learning. However, visible cameras cannot capture clear pedestrian images under low light conditions, which limits the application scenarios of single-modality person re-identification.

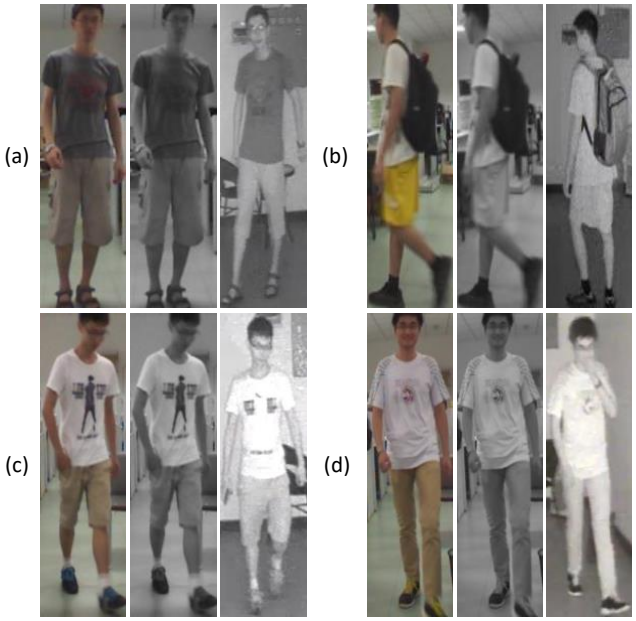

Figure 1. (a), (b), (c) and (d) are comparison of different modalities,from left to right are visible, gray and infrared image in each group. Gray image has smaller modality discrepancy with infrared image than visible image.

The intelligent surveillance camera can automatically switch between visible mode and infrared mode according to lighting conditions. Cross-modality person re-identification has attracted extensive attention from the community. The existing single-modality person re-identification method is not suitable for eliminating the modality discrepancy between visible images and infrared

images. Modality discrepancy is currently the key issue in this field.

In order to alleviate the modality discrepancy, many elaborated methods have been proposed for pixel-level and feature-level alignment. GAN based methods [4, 24, 25, 35] generate corresponding infrared modalities for visible modalities and vice versa. The cross-modality image pair is mapped into an unified image space to achieve pixel-level alignment. Due to the lack of paired-images, it is difficult to generate high quality images. In particular, The identity-related information can not been transfed into opponent modality correctly. GAN based methods cannot maintain the consistency of identity information between pairs of images, that inevitably brings some noises, especially generating visible modality for infrared modality. Further more, GAN is computing-intensive and structure-complicated. A simple and effective solution to reduce modality discrepancy is ignored: convert visible images to gray images directly by an acknowledged algorithms. The gray image has higher quality than the pseudo-infrared image generated by GAN based method. At the same time, it can retain the identity cues of pedestrians effectively. In this paper, we transform the cross-modality person re-identification task from Visible-Infrared to Gray-Infrared. In principle, the infrared image reflects the electromagnetic wave intensity of the object reflecting the near-infrared wavelength. It is a single-channel image. The gray image converted from three-channel visible image is more similar to the infrared image. Moreover, compared with visible image, the gray image looks less different from infrared image, as shown in Fig. 1. So we call it the minimal modality discrepancy. Extensive experiments have proved that this transformation can effectually improve the performance of person re-identification.

Existing methods of representation learning and metric learning [3, 18, 38, 29] mostly regard the features of the last layer of CNN as the final representation, which cannot fully contain the clues that determining pedestrian identity between gray image and infrared image. To address this problem, we propose a Pyramid Feature Integration Network (PFINet), which mines the discriminative refined features of pedestrian images and fuses high-level semantically stronger features to build a robust pedestrian representation. PFINet consists of three modules including pyramid information extraction(PIE), discriminative-region response(DRR), and multi-scale feature integration (MSFI). Specifically, PIE first extracts multi-scale features from concrete to abstract, and then perform top-down semantic transmission. DRR exploits the discriminative areas of pedestrian images by the spatial attention to give them higher responsiveness and improve the discriminativeness of pedestrian features. Multi-scale features are integrated by MSFI to obtain the robust pedestrian representation. PFINet combines high-scale features which are semantically strong and fine-grained low-scale features, and discovers identity-related information. In addition, PFINet can efficiently deal with the the variations of viewpoints, body poses and occlusion. In summary,our main contributions include the following aspects:

- This paper proves that converting visible images to gray images can effectively alleviate modality discrepancy, which is called minimal modality discrepancy. We transform the cross-modality person re-identification task from Visible-Infrared to Gray-Infrared that improving the accuracy substantially.

- In order to extract the identity-related information in pedestrian images, we proposes the PFINet, that extract more robust pedestrian representation with pyramid network structure and spatial attention mechanism.

- The proposed method outperforms the state-of-the-art methods significantly,which demonstrate its effectiveness.

2. Related work

2.1. Single-modality person re-identification

Visible image based person re-identification has made great progress and high accuracy, with the advancement of deep learning. Metric learning based methods aimed to improve the intra-class similarity and reduced the inter-class similarity through continuous improvement of the metric function,*e.g.*,Ding *et al.* [5] proposed triplet loss, Chen *et al.* [2] proposed quatruplet loss, Hermans *et al.* [13] proposed hard triplet loss. Representation learning based methods, such as IDE[36], MuDeep[20], SVDNet[21], *etc.*, focused on learning the discriminative feature representation of pedestrians, and regarded person re-identification as a multi-category classification task, which treated each identity as a separate category. Horizontal stripe based on methods(*e.g.*,PCB[22], Alignedreid[33], Beyond Human Parts[10]) and pose estimation based methods(*e.g.*,SpindleNet[34], PVPM[7]), performed local feature aggregation. In recent years, unsupervised methods have made great progress, and provided new perspective. A number of excellent works have appeared, such as Spcl[9], MMT[8], GCL[1], *etc.* Although the single-modality person re-identification method has made great achievements, it is not suitable for this cross-modality task.

2.2. Visible-Infrared person re-identification

Many excellent methods have been proposed to alleviate the modal discrepancy. GAN based methods either generate

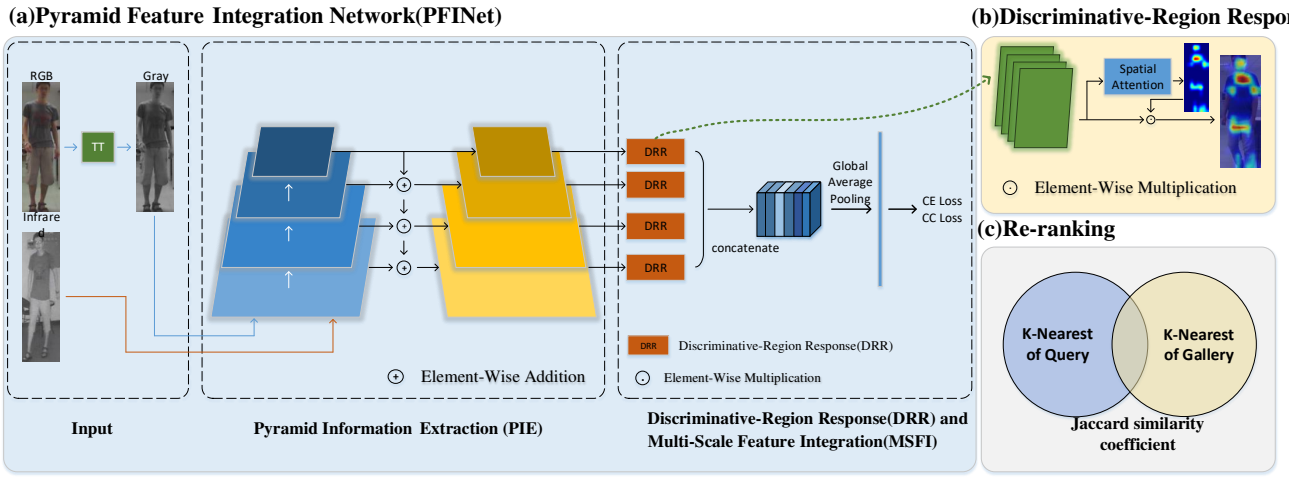

Figure 2. (a)The overview of proposed Pyramid Feature Integration Network. PFINet consist of Pyramid Information Extraction, Discriminative-Region Response and Multi-Scale Feature Integration.(b)The details of Discriminative-Region Response Module.(c)Re-ranking strategy.

the corresponding infrared image (visible image) for visible image (infrared image), or eliminate modality discrepancy with generative adversarial training. cmGAN[4] projected pedestrian features into a common space with generative adversarial training. D^{2}RL[26] generated corresponding modality information, and mapped the original image into an unified four-channel space, which contains both the visible and infrared information. AlignGAN[24] converted visible images into infrared images, and achieved pixel-level and feature-level alignment in an unified generative adversarial network. JSIA-ReID[25] performed instance-level feature alignment to eliminate modality discrepancy. CICL[35] generated pedestrian images with different color of clothes, and implemented color-independent consistent learning to eliminate the influence of color. However, these methods can not ensure that the discriminative identity-related features are transfed to the corresponding modality.

Apart from GAN, there are many excellent methods based on representation learning or metric learning. Zero-paddind[28] explored domain-specific nodes in a single network, and responded to different modality inputs selectively. eBDTR[31] learnt modality-shared features by a two-stream network, and propoed a dual-constrained top-ranking loss. HSME[11] proposed a hypersphere manifold embedding model, which maps cross-modality images into an unified feature space. MACE[30] proposed a modality-aware collaborative ensemble learning method, that reduce modality discrepancy at feature-level and classification-level in a dual-stream network. SIM[15] presented a similarity inference metric that probe the intra-modality sample similarities. DDAG[32] explored the relationship between local-parts of intra-modality, and obtained graph-level cross-modality context with graph-based method. X Modality[16] introduced an auxiliary modality to narrow the gap between infrared image and visible image, and converted the original task into a three modality matching task. NFS[3] introduced a new feature search space, and proposed a neural feature selection method to select identity-related information automatically in the channel and space dimension. MPANet[29] discovered the nuances from pedestrian images to improves the distinguishability.

2.3. Attention mechanism

Attention mechanism is a commonly used method. It was inspired by a common experience that people always recognize objects by grasping key information, rather than all the information. In recent years, researchers have proposed many methods of attention mechanism to exploit the key information of research objectives from different aspects. SENet[14] proposed channel attention to examine the dimension-wise relationship. Cbam[27] combined spatial attention and channel attention to explored the relationship of channel-level and spatial-level. Ashish *et al.* [23] proposed self-attention and transformer models, which are widely used in the fields of natural language processing and computer vision. DANet[6] drew on the idea of self-attention to design attention in channel and spatial dimensions, and proposed a dual attention network. In this paper, we use spatial attention mechanism to find identity correlation regions in pedestrian images to improve the ability of feature discrimination.

3. Methodology

3.1. Task Transformation

The modality discrepancy between infrared image and RGB image is the biggest challenge for cross-modality person re-identification. Huge modality discrepancy are not conducive to the extraction of shared features, and often lead to larger intra-class gap and lower accuracy. In this paper, we transform the cross-modality person re-identification task from Visible-Infrared to Gray-Infrared. Specifically, we convert the visible image into gray image directly which can be formulated as:

$$Gray = R * 0.299 + G * 0.587 + B * 0.114, \tag{1}$$

where $Gray$ denotes the converted gray image, and R, G, and B represent the red, green, and blue channels of the visible image, respectively.

The gray image has smaller discrepancy with infrared image than visible image. In principle, the infrared image reflects the electromagnetic wave intensity of the object reflecting the near-infrared wavelength. It is a single-channel image. The gray image converted from three-channel visible image is more similar to the infrared image. So, we call it the minimal modality discrepancy.

Compared with GAN, our method is computing-saving and simple which needs only to perform three times multiplication operations. Further more, the converted image is more natural and has higher quality, which can effectively transform the identity-related infomation.

3.2. Pyramid Feature Integration

We propose a Pyramid Feature Integration Network (PFINet), which mines the discriminative refined features of pedestrian images and fuses high-level semantically stronger features to build a robust pedestrian representation. The overview of the proposed method is illustrated in Fig. 2. After task transformation, PFINet first performs feature extraction from concrete to abstract and top-down semantic transfer to obtain multi-scale feature maps. Then the multi-scale feature maps are input to Discriminative-Region Response(DRR) module to emphasize the identity-discriminative region by the spatial attention mechanism. Pedestrian representation is obtained by feature integration. More details are described in the ensuing subsections.

Pyramid Information Extraction. For a given input image $X \in \mathbb{R}^{c*h*w}$, We define Z to represent the result of the task transformation. we strive to extract rich identity-related features by well-designed structure.

Inspired by FPN[17], we employ a pyramid network to construct multi-scale feature maps. Like PFN[17], we obtains the feature map C_i layer by layer of each scale to perform the feature extraction from concrete to abstract,which can be represented by

$$C_i = Conv_i(C_{i-1}), (i = 1, 2, 3, 4), \tag{2}$$

where $Conv_i$ denotes operation of convolution of the ith layer. Note that,C_0 equals Z.

In this paper, we choose ResNet-50[12] as backbone of feature extraction network. $\{C_1, C_2, C_3, C_4\}$ represent the output of four residual blocks of the backbone. The resolution of these feature mpas is from high to low, semantics is from low to high, and features are from concrete to abstract.

Following FPN[17], to make low-scale features have high semantic while having fine features,we conduct top-down semantic transfer by

$$P_i = BI(C_{i+1}) + Conv(C_i), (i = 1, 2, 3), \tag{3}$$

where BI denotes bilinear interpolation up-sampling, $Conv$ is 3*3 convolution. Particularly, $P_4 = C_4$.

Discriminative-Region Response. Identity-related information is usually hidden in the body parts, such as head-and-neck, torso, arms, legs, and feet, *etc*. However, not all parts are identity-related. The Discriminative Regional Response(DRR) module aims to emphasize the identity-discriminative region in pedestrian images. We generate a response map $M \in^{h*w}$ for the feature map of each scale, where h and w are equal to the height and width of the corresponding feature map. M_i is formulated as:

$$M_i = \sigma(Att(P_i)), \tag{4}$$

where σ is sigmod activation function, and Att denotes spatial attention mechanism.

With these attention maps,the Responsed feature map is calculated by:

$$F_i = M_i \bigodot P_i, \tag{5}$$

where $\bigodot$ denotes element-wise multiplication.

Multi-Scale Feature Integration. After obtaining the responsed feature maps at each scale, the representation of k-th scale $f_k \in \mathbb{R}^d$ is generated by global average pooling. In the end, we fuse the multi-scale representation as the final pedestrian representation.

$$f_k = GAP(F_k) \tag{6}$$

$$Rep = [f_1, f_2, f_3, f_4], \tag{7}$$

where GAP is global average pooling, $[...]$ denotes the operation of concatenat, Rep is the final pedestrian representation.

Features at multiple scales contain different information. These features can complement each other. This provides sufficient information for pedestrian representation.

The pedestrian representation obtained by this method has strong robustness. It can overcome the obstacle of modality gap within minimal modality discrepancy, and contain identity-discriminative information by Discriminative-Region Response module. Moreover, fusing the features of each scale from a global perspective can effectively deal with the variations of viewpoints, body poses, and occlusion.

3.3. Re-ranking

Re-ranking is a widely used strategy in single-modality person re-identification. It is based on an assumption: if a query image and a gallery image have the same identity, the K nearest neighbors of the two images in gallery should overlap. We apply re-ranking strategy to cross-modality person re-identification. It can still improve the accuracy greatly in cross-modality person re-identification. More details of re-ranking please refer to [37].

3.4. Objective Functions

This paper uses cross-entropy loss and center cluster loss as the objective function to train the entire network in an end-to-end manner.

Cross-entropy loss is widely used in classification tasks. It encourages samples with the same identity to be distributed in adjacent regions in feature space, which can be formulated as:.

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{P} y_{ic}\log(P_{ic}), \tag{8}$$

where N is the number of samples, P is the quantity of identity. y_{ic} denotes the ground-truth that whether sample i belongs to identity c, and P_{ic} is prediction.

We use a fully connected layers with shared parameters and an unified cross-entropy loss function for the input image of different modaliy, which can reduce the modality discrepancy to a certain extent.

Center cluster loss[29] aims at gathering the features to their identity-center and forcing the centers of feature of different identities be far away from each other by

$$\begin{aligned} L_{cc} &= \frac{1}{V+I}\sum_{i=1}^{V+I}\|f_i - h_{yi}\|_2 \\ &+ \frac{2}{P(P-1)}\sum_{k=1}^{P-1}\sum_{j=k+1}^{P}[\rho - \|h_{yk} - h_{yj}\|], \end{aligned} \tag{9}$$

where V and I denote the number of samples of visible and infrared image respectively, P is the quantity of identity, h_{yi}, h_{yj}, h_{yk} are the center of their corresponding identity, ρ denotes the mini-threshold between different identities.

4. Experiments

4.1. Datasets and Settings

Datasets. We evaluate the proposed method on two public datasets SYSU-MM01[28] and RegDB[19].

SYSU-MM01[28] is a large-scale dataset, which is captured by six non-overlapping cameras including four visible cameras and two infrared cameras. And, it is collected in different scenarios including indoor or outdoor. This dataset contains a total of 419 identities. The training set contains 22258 visible images and 11909 infrared images of 395 identities , and the test set contains 3803 infrared images as query and 3010(301) randomly selected visible images as gallery of 96 identities.

RegDB[19] is another widely used dataset which is captured by a pair of visible and infrared cameras. This dataset includes a total of 412 identities, each of which contains 10 visible images and 10 infrared images. It is randomly divided into two parts for training and testing.

Evaluation metrics and strategies. The acknowledged Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) are adopted as evaluation metrics.

For the SYSU-MM01, two evaluation strategies are employed: all-search and indoor-search, each of which includes single-shot and multi-shot modes. Indoor-search means that the gallery only contains visible images captured by indoor cameras, while all-search means that the gallery contains all visible images. One-shot means that each identity in the gallery contains only one image, while multi-shot 10.

For the RegDB, two evaluation strategies are employed: visible-to-infrared and infrared-to-visible. Visible-to-infrared signifies visible image as query while infrared image as gallery, and vice versa.

Implementation details. We choose ResNet-50 pretrained on the ImageNet as the backbone, and train on two Tesla P100 GPUs. The batchsize is set to 128, consisting of 16 randomly selected pedestrian identities, and 8 images for each identity including 4 visible images and 4 infrared images. We adopt Adam optimizer, and the initial learning rate is set to 0.00035, warmup 10 epochs at the beginning. At the 90th, 150th, and 200th epoch, the learning rate is decayed with a factor of 0.1, and the weight decay is set to 0.0005. The total number of training epochs is set to 240. Input image is first resized to 384*128, and enhanced with a probability of 50. The data enhancement strategies includes random flip, random crop, and random erasure. We train the network in an end-to-end manner.

4.2. Comparison with State-of-the-Arts

The proposed method is compared with some state-of-the-art methods, including GAN based methods (*e.g.*, cmGAN[4], D^{2}RL[26], AlignGAN[24], JSIA-ReID[25],

Table 1. Comparison with the state-of-the-art methods over the dataset SYSU-MM01

method	venue and year	All-seaerch								Indoor-search							
		single-shot				Multi-shot				single-shot				Multi-shot			
		R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
Zero-Padding[28]	ICCV-2017	14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
D-HSME[11]	AAAI-2019	20.68	62.74	77.95	23.12	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN[4]	IJCAI-2018	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	2.1	32.8
eBDTR[31]	TIFS-2020	27.82	67.34	81.34	28.42	-	-	-	-	-	-	-	-	-	-	-	-
D^2RL[26]	CVPR-2019	28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
JSIA-ReID[25]	AAAI-2020	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
AlignGAN[24]	ICCV-2019	42.40	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
X modality[16]	AAAI-2020	49.92	89.79	95.96	50.73	-	-	-	-	-	-	-	-	-	-	-	-
MACE[30]	TIP-2020	51.64	87.25	94.44	50.11	-	-	-	-	57.35	93.02	97.47	64.79	-	-	-	-
DDAG[32]	ECCV-2020	54.75	90.39	95.81	53.02	-	-	-	-	61.02	94.06	98.41	67.98	-	-	-	-
SIM[15]	IJCAI-2020	56.93	-	-	60.88	-	-	-	-	-	-	-	-	-	-	-	-
NFS[3]	CVPR-2021	56.91	91.34	96.52	55.45	63.51	94.42	97.81	48.56	62.79	96.53	99.07	69.79	70.03	97.70	99.51	61.45
CICL[35]	AAAI-2021	57.2	94.3	98.4	59.3	60.7	95.2	98.6	52.6	66.6	98.8	99.7	74.7	73.8	99.4	99.9	68.3
cm-SSFT[18]	CVPR-2020	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
MPANet[29]	CVPR-2021	70.58	96.21	98.80	68.24	75.58	97.91	99.43	62.91	76.74	98.21	**99.57**	80.95	84.22	99.66	99.96	75.11
PFINet(ours)$_{w/o\ rr}$	CVM-2022	68.53	**96.73**	**98.89**	65.51	**76.28**	97.71	99.36	60.13	**76.88**	98.10	99.49	80.26	**85.04**	99.48	99.89	74.68
PFINet(ours)$_{w/rr}$	CVM-2022	**80.83**	**99.93**	**99.98**	**79.03**	**81.95**	**99.36**	**99.81**	**74.49**	**88.54**	**98.67**	99.56	**89.97**	**91.53**	**99.94**	**99.98**	**86.62**

Note: w/rr denotes re-ranking was applied. $w/o\ rr$ denotes re-ranking was removed.

Table 2. Comparison with the state-of-the-art methods over the dataset RegDB

method	venue and year	Visible-to-Infrared		Infrared -to- Visible	
		Rank-1	mAP	Rank-1	mAP
Zero-Padding[28]	ICCV-2017	17.8	18.9	16.6	17.8
eBDTR[31]	TIFS-2020	34.62	33.46	34.21	32.49
D^2RL[26]	CVPR-2019	43.4	44.1	-	-
JSIA-ReID[25]	AAAI-2020	48.5	49.3	48.1	48.9
D-HSME[11]	AAAI-2019	50.85	47.00	-	-
AlignGAN[24]	ICCV-2019	57.9	53.6	56.3	53.4
X modality[16]	AAAI-2020	-	-	62.3	60.2
DDAG[32]	ECCV-2020	69.34	63.46	68.06	61.80
cm-SSFT[18]	CVPR-2020	72.3	72.9	71.0	71.7
MACE[30]	TIP-2020	72.37	69.09	72.12	68.57
SIM[15]	IJCAI-2020	74.7	75.2	75.2	78.3
CICL[35]	AAAI-202	78.8	69.4	77.9	69.4
NFS[3]	CVPR-2021	80.54	72.10	77.95	69.79
MPANet[29]	CVPR-2021	83.7	80.9	**82.8**	80.7
PFINet(ours)$_{w/o\ rr}$	CVM-2022	81.13	**81.49**	80.21	**81.02**
PFINet(ours)$_{w/rr}$	CVM-2022	**85.09**	**85.75**	82.64	**84.90**

Note: w/rr denotes re-ranking was applied. $w/o\ rr$ denotes re-ranking was removed.

Table 3. Ablation Study on SYSU-MM01 in the Single-shot and All-search mode

method	Single-shot and All-search mode	
	Rank-1	mAP
Baseline	58.17	56.69
Baseline+TT	65.67	62.50
PFI	63.43	62.41
PFI+TT	68.53	65.52

Table 4. Performance comparison with other TT methods on SYSU-MM01

method	Single-shot and All-search mode	
	Rank-1	mAP
Fixed Parameter	68.53	65.52
Learnable Single Parameter	63.33	61.49
Learnable Multi-Parameter	57.96	57.42

Table 5. Performance comparison of different layers on SYSU-MM01

Integrated Layers	Single-shot and All-search mode	
	Rank-1	mAP
One Layer	66.40	63.12
Two Layers	66.61	63.58
Three Layers	67.16	64.02
All Layers	68.53	65.52

CICL[35]), and metric learning based method (*e.g.*, zero-padding[28], eBDTR[31], HSME[11], MACE[30], SIM[15], DDAG[32], X modality[16], MPANet[29], NFS[3], cm-SSFT[18]).

Comparison on SYSU-MM01. As shown in Table 1, the PFINet surpasses the existing state-of-the-art methods dramatically. In multi-shot and indoor-search mode,PFINet has reached the level of single-modality person re-identification, achievs the rank-1 accuracy of 91.53% and mAP of 86.82%. This is the first time that the rank-1 accuracy exceed 90%, which proves the significance of the proposed method. For fair comparison, we also show the experiment result which is obtained without re-ranking. Our proposed method can still outperforms the most state-of-the-art methods. And compared with SOTA MPANet, PFINet get comparable performance.

Comparison on RegDB. As shown in Table 2, the proposed method also achieves comparable performance with the state-of-the-art method in RegDB dataset. Compared with SOTA MPANet, PFINet improves the mAP by 0.59%

and 0.32% in the infrared-to-visible mode and visible-to-infrared mode separately. And comparing the last two rows, we found that re-ranking strategy can not bring dramatically performance improvement in this dataset.

4.3. Ablation Study

To demonstrate the effectiveness of each module of the proposed method, we conduct ablation study on SYSU-MM01. Baseline indicates that only the ResNet-50 backbone network is applied. TT indicates Task Transformation. PFI represents the backbone was replaced by PFINet. Due to the effect of re-ranking in different experimental settings is inconsistent, the re-ranking strategy is not adopted in ablation study. So, the performance is inferior to Table 1.

As shown in Table 3, the experimental results reveal the effectiveness of each module of the proposed method. Comparing the row 1 and the row 2, the performance of the baseline has been greatly improved by task transformation, where the rank-1 accuracy increased by 7.5%, and the mAP increased by 5.81%. It can be proved that TT plays an important role in cross-modality person re-identification, and the modality discrepancy between visible image and infrared image is greatly reduced. Comparing the row 1 and the row 3 can prove that the proposed PFINet is more effective for extracting the identity information in the pedestrian image. Comparing the row 2 and the row 4, it can be found that with the minimal modality discrepancy, the propoed PFINet can more accurately extract the identity-discriminative information in the pedestrian image, and make up for the weakness of the baseline.

4.4. Discussions

Task Transformation. We employ a widely used fixed formula to convert visible image to gray image s shown in Formula 1. In the context of deep learning, high expectations are placed on the learning ability of the model. So, we compare it with the method of setting learnable conversion parameters, including (1) LSP:each channel sets learnable single parameter; (2) LMP: Set Learnable Multi-Parameters for each position of each channel, the size of parameters is the same as input image. As shown in Table 4, the fixed parameter formula has achieved the best results. LSP is slightly inferior, but LMP drops drastically. We empirically attribute it to insufficient module complexity of learnable task transformation. It proves that the acknowledged method of converting visible image into gray image has the ability to effectively reduce modality discrepancy.

Visualization of Discriminative Regions. In order to prove that the Discriminative-Region Response(DRR) module can effectively locate the identity-related region in pedestrian images, we select 3 visible images and 3 infrared images with different identities to visualize the responsed attention maps. As shown in Fig. 3, the upper line is the input image, and the lower line is the responded map. We can observe that under the efforts of DRR, some discriminative regions in pedestrian images have received more attention. And, it can be seen from the figure that the discriminative feature of pedestrians mainly exist in the head-neck, upper-torso, legs, feet. Usually, the identity of pedestrians is also judged by these areas in reality.

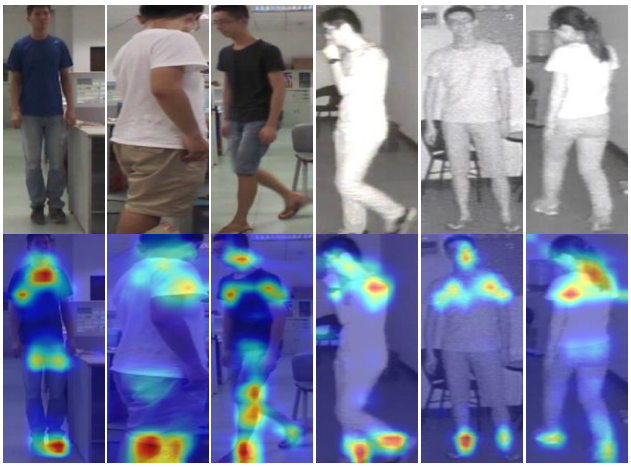

Figure 3. Visualization of Discriminative-Region.We show the original and responsive images of six pedestrians with different identities.

Moreover, the poses and viewpoints in these images are different, and there are occlusions. The DRR can still accurately locate the discriminative areas. This also proves that the proposed PFINet is useful for solving the variations of viewpoints, body poses, and occlusion.

Multi-scale Feature Integration. The pyramid network is applied to extract multi-scale features. The features of different scales obtained in each layer are integrated together as the final robust pedestrian representation. This kind of method can extract more sufficient information. We designs comparative experiments to evaluate the effects of using only the features of the top level, the first two layers, the first three layers, and all layers of pyramid to prove that the features of each layer in the pyramid structure play a key role in pedestrian representation. The experimental results in Table 5 prove that the feature of each layer are indispensable. From row 1 to row 4, it can be seen that the performance of the model gradually improves with the increase of the number of feature fusion layers.

5. Conclusion

Cross-modality person re-identification is an interesting and meaningful task. The severe challenge it faces is modality discrepancy. In this paper, we propose a method for learning robust pedestrian representation within minimal modality discrepancy for visible-infrared person re-identification. Specifically, We first transform the

cross-modality person re-identification task from Visible-Infrared to Gray-Infrared which is named as minimal modality discrepancy. Then, Pyramid Feature Integration Network (PFINet) is proposed to mine the discriminative refined features of pedestrian images and fuses high-level semantically stronger features to build a robust pedestrian representation. The pedestrian representation obtained by this method has strong robustness. It can overcome the obstacle of modality gap, and contain identity-discriminative information. Especially the minimal modality discrepancy, it has always been ignored by researchers, but in fact it is computing-saving and also can greatly alleviates the modality discrepancy. Extensive experiments on the two public datasets SYSU-MM01 and RegDB demonstrate the effectiveness of the proposed method.

References

[1] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2004–2013, 2021. 2

[2] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2

[3] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun. Neural feature search for rgb-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 587–597, 2021. 2, 3, 6

[4] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 2, 2018. 2, 3, 5, 6

[5] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 3

[7] S. Gao, J. Wang, H. Lu, and Z. Liu. Pose-guided visible part matching for occluded person reid. pages 11744–11752, 2020. 2

[8] Y. Ge, D. Chen, and H. Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 2

[9] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *Advances in Neural Information Processing Systems*, 2020. 2

[10] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3642–3651, 2019. 2

[11] Y. Hao, N. Wang, J. Li, and X. Gao. Hsme: hypersphere manifold embedding for visible thermal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8385–8392, 2019. 3, 6

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[14] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[15] M. Jia, Y. Zhai, S. Lu, S. Ma, and J. Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1026–1032, 2020. 3, 6

[16] D. Li, X. Wei, X. Hong, and Y. Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020. 3, 6

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[18] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu. Cross-modality person re-identification with shared-specific feature transfer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13376–13386, 2020. 2, 6

[19] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 5

[20] X. Qian, Y. Fu, Y. G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architectures for person re-identification. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[21] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision*, pages 3800–3808, 2017. 2

[22] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 2

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[24] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019. 2, 3, 5, 6

[25] G.-A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, and Z.-G. Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12144–12151, 2020. 2, 3, 5, 6

[26] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–626, 2019. 3, 5, 6

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[28] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 3, 5, 6

[29] Q. Wu, P. Dai, J. Chen, C.-W. Lin, Y. Wu, F. Huang, B. Zhong, and R. Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 2, 3, 5, 6

[30] M. Ye, X. Lan, Q. Leng, and J. Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020. 3, 6

[31] M. Ye, X. Lan, Z. Wang, and P. C. Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2020. 3, 6

[32] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 229–247. Springer, 2020. 3, 6

[33] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 2

[34] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017. 2

[35] Z. Zhao, B. Liu, Q. Chu, Y. Lu, and N. Yu. Joint color-irrelevant consistency learning and identity-aware modality adaptation for visible-infrared cross modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3520–3528, 2021. 2, 3, 6

[36] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 2

[37] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 5

[38] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao. Hetero-center loss for cross-modality person re-identification. *Neurocomputing*, 2019. 2