ISSN: 2664-9640

Cross-modal Pedestrian Re-identification based on Spatially Enhanced Dual-stream Network

Zheyu Fu, Xiaoguang Hu, Xu Wang

School of people's public security of University, Beijing 100038, China

Abstract

In the current research on cross-modal pedestrian re-identification, the main difficulty lies in the low recognition accuracy caused by the difference between modalities, in order to solve this problem. In this paper, we propose a new spatially enhanced dual-stream network, which is called Multivariate Extended Network (DEN). This method can embed different learning features to reduce the gap between modalities. The network is composed of a spatial embedding module (CPM) and a multi-feature aggregation module (SEM), in which the spatial embedding module can embed diversified information to improve performance, and the multi-feature aggregation module can aggregate features at different stages to mine channel and spatial information, thereby improving the ability of the network to mine different embeddings at different levels.

Keywords

Pedestrian Re-identification; Deep Learning; Cross-modality; Feature Learning; Convolutional Neural Network (CNN); Space Enhancement.

1. Introduction

Pedestrian re-identification is widely regarded as a sub-problem of image retrieval, that is, whether the same person can be identified under different surveillance, which is very important for the construction of Ping An China and Xueliang Project. With the development of computer science, single-modal pedestrian re-identification technology has developed rapidly, but with the needs of public security, the visibility at night is relatively low, and the monitoring needs to convert the pedestrian features captured by visible light during the day into visible features at night with infrared light. Therefore, in order to solve the above problems, this paper proposes a new spatially enhanced dual-stream network structure, which can reduce the difference between the two modes by mining and embedding multiple information.

Person re-identification is designed to identify the image of a pedestrian under the camera with the picture in the gallery. Most of the existing methods only focus on pedestrian re-identification during the daytime, but at night, when the visible light camera cannot capture effective pedestrian information, infrared cameras will be used to shoot and use the cross-modal pedestrian re-identification method for detection.

Fig. 1 Visible image (left) and infrared image (right)

ISSN: 2664-9640

However, there is a big modal difference between the picture taken under visible light and the picture taken by the infrared camera. In general, there are two methods to reduce these differences, one is the feature sharing method, in which infrared and visible features are projected into a common embeddable space, in which modal differences can be reduced, for which some Gaussian-based methods are proposed [1,2,3,4] to identify the modal transformation of character images across modal images to alleviate the problem of limited data. These methods often design complex generative models to align cross-modal images. However, due to the lack of visible infrared visible image pairs, the resulting image is inevitably accompanied by some noise. It has also been used to assist cross-modal search tasks by introducing auxiliary intermediate modalities with lightweight networks through the infrared modality [5] and its variants [6,7]. However, due to the large modal differences, it is difficult for these methods to project cross-modal images directly into the common feature space. To this end, MAUM [8] attempts to learn transmodal attitude quantities in both directions, further enhancing them through memory-based enhancement.

RFM [9] introduces cross-center losses to explore a more compact intra-class distribution. DCLNet [10] decreases the distance of positive pixels with the same semantic information and increases the distance of negative pixels. cmGAN [5] designed a cutting-edge discriminator to learn discriminant representations from different modalities. However, due to the large modal gap between VIS and IR images, it is difficult to project cross-modal images directly into public spaces [11,12,13,14].

Some scholars have converted it into a counterpart to a visible (or infrared) image through a GAN network, but due to the lack of available infrared-visible image pairs, the resulting cross-modal images are usually accompanied by some differences. In this paper, a new spatially enhanced dual-flow model is proposed, which adds a spatial embedding module (SEM) and a multi-feature aggregation module (CPM) on the basis of the Resnet50 dual-flow model, which improves the performance of the model through spatial enhancement, and adds a new gradient of the model parameters of the loss function to reduce the loss of effective weights.

1.1. Benchmark Model

The benchmark network used in this article is the Resnet50 dual-stream network, which is a type of residual network, which is called Residual Network with 50 layers. The network is widely used in areas such as object classification, computer vision tasks, and serves as the backbone network for these tasks. The network structure of Resnet50 is divided into multiple layers, including conv1, conv2_x, conv3_x, conv4_x, and conv5_x. Each layer is composed of multiple residual blocks (Bottleneck or Identity Block), and the residual learning is realized through skip connection, which is helpful to solve the problem of gradient vanishing or gradient explosion in deep network training. Compared with a single network architecture, the Resnet50 dual-stream network can make full use of the spatial and temporal information in the video, and improve the accuracy and robustness of the model. At the same time, because Resnet50 itself has good performance and scalability, this combination also has high practical value and research significance.

Some scholars have converted it into a counterpart to a visible (or infrared) image through a GAN network, but due to the lack of available infrared-visible image pairs, the resulting cross-modal images are usually accompanied by some differences. In this paper, a new spatially enhanced dual-flow model is proposed, which adds a spatial embedding module (SEM) and a multi-feature aggregation module (CPM) on the basis of the Resnet50 dual-flow model, which improves the performance of the model through spatial enhancement, and adds a new gradient of the model parameters of the loss function to reduce the loss of effective weights.

ISSN: 2664-9640

1.2. Benchmark Model

The benchmark network used in this article is the Resnet50 dual-stream network, which is a type of residual network, which is called Residual Network with 50 layers. The network is widely used in areas such as object classification, computer vision tasks, and serves as the backbone network for these tasks. The network structure of Resnet50 is divided into multiple layers, including conv1, conv2_x, conv3_x, conv4_x, and conv5_x. Each level is composed of multiple residual blocks (Bottleneck or Identity Block), and the residual learning is realized through hop connections, which is helpful to solve the problem of gradient vanishing or gradient explosion in deep network training. Compared with a single network architecture, the Resnet50 dual-stream network can make full use of the spatial and temporal information in the video, and improve the accuracy and robustness of the model. At the same time, because Resnet50 itself has good performance and scalability, this combination also has high practical value and research significance.

The main innovations in this paper are as follows: we propose a novel Multivariate Extension (DEN) network, which consists of a spatial embedding module (SEM) and a multi-feature aggregation module (CPM) to generate more embeddings to learn different feature representations. We were the first to add embedding to Virid's embedding space. In addition, an effective Multi-Stage Feature Aggregation (MFA) block is proposed to mine potential channel and spatial feature representations. Combining SEM, CPM loss, and MFA into the end-to-end learning framework, an effective Multivariate Embedding Extension Network (DEEN) is proposed, which can effectively reduce the modal differences between VIS and IR images.

2. Methods of this Article

2.1. Dual-stream Network Structure

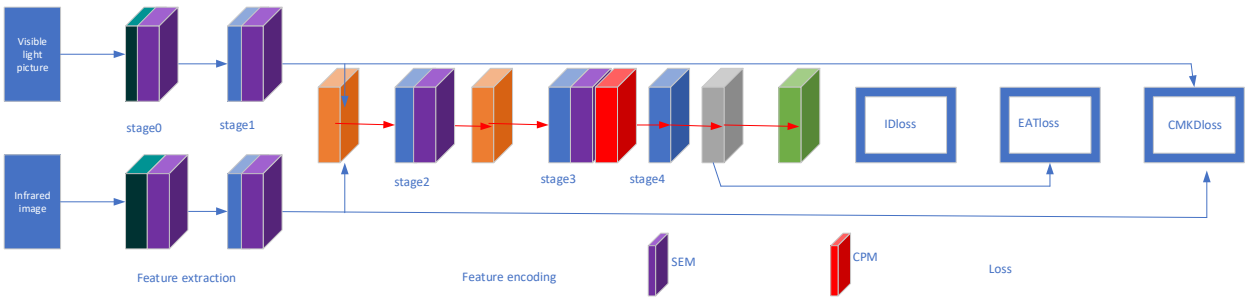

Fig. 2 DEEN network structure

In cross-modal pedestrian re-identification, the most commonly used method is the two-stream model structure, which was first proposed in 2018 [15]. The network structure used in this paper is shown in the figure, using Resnet50 as the benchmark network, after inputting visible light images and infrared light images respectively, the feature extraction part is composed of two independent branches that do not share parameters, and the common features of the two methods are learned. stage0 and stage1 are used as feature extraction parts, and stage2, stage3, and stage4 are used as feature embedding parts to be detected by a public network, and finally three loss functions are used to optimize to improve the accuracy of the results. The first layer and the second layer are extracted for features, and the last three layers are for feature embedding, and the features are input into the CPM module to generate more features for embedding, so that the network can share Features.

In addition, the model incorporates an effective SEM module to aggregate features from different stages to mine different channel and spatial features. In the training phase, all features

ISSN: 2664-9640

before and after the normalized layer are input into different loss functions to jointly optimize the network model.

2.2. Introduction to the CPM module

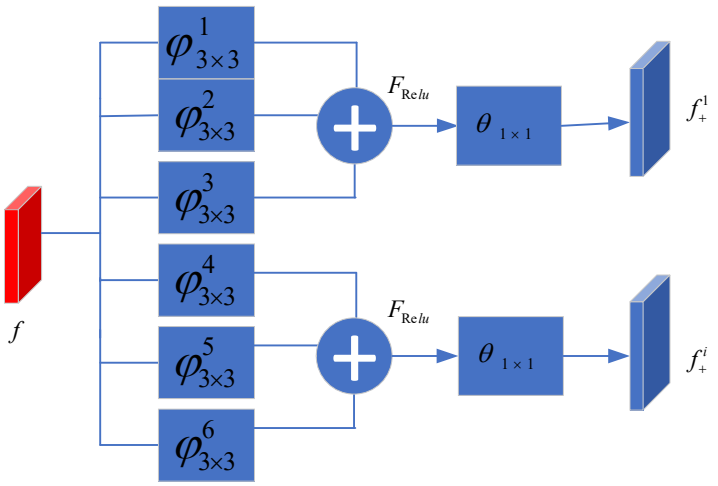

Fig. 3 Structure diagram of the CPM module

The CPM module uses a multi-branch convolutional generation structure to generate more embeddings to alleviate the problem of insufficient training data. Specifically, for each branch of the SEM, we first use 3 3×3 extended convolutional layers for convolution, $\varphi^1_{3\times3}, \varphi^2_{3\times3}, \varphi^3_{3\times3}$, different expansion ratios (1,2,3) reduce the number of feature maps to 1/4 of their own size, then combine them into a single feature map, and then use the ReLU activation layer to improve the nonlinear representation of CPM. Then, another convolutional layer is applied to the resulting feature map, with a kernel size of 1×1, varying its dimensionality as in the same way as f. In this way, the embedding fi+ of the generated i-th branch can be written as:

$$\Re^{\Upsilon}_{+} = \theta_{1\times1}(F_{Relu}(\phi^1_{3\times3}(f) + \phi^2_{3\times3}(f) + \phi^3_{3\times3}(f)) \tag{1}$$

Then, all the resulting embeddings are connected together as input to the next stage of the backbone. From the above operation, we can see that the DEE module can only generate more embeddings using multi-branch convolution blocks. However, this operation is not effective in obtaining diverse embeddings. Therefore, we apply the following three properties to constrain the generated embeddings to be as diverse as possible in order to effectively reduce the modal difference between the visible and infrared images1: The generated embeddings should be as diverse as possible to effectively learn the feature representation of the information. This means that we need to push the distance between the generated embedding and the original embedding to learn different features and mine different cross-modal cues.

Nature 2: The resulting embedding should help to reduce modal differences between visible and infrared images. This means that we need to close the distance between the embeddings generated by the visible modality and the original IR embeddings. Similarly, we need to close the distance between the embedding generated by the IR modality and the original visible image embedding.

Property 3: The intra-class distance should be less than the inter-class distance. According to property 2, it takes the distance between the generated embedding and the original embedding, which can cause different classes of embeddings to become very close. Therefore, it is necessary to keep the intra-class distance less than the inter-class distance.

ISSN: 2664-9640

For IR modally generated embeddings, the CPM loss can be expressed as:

$$\int(f_v, f_n, f_{v+}^i) = [D(f_n^i, f_{v+}^{i,j}) - D(f_v^i, f_{v+}^{i,j}) - D(f_v^i, f_{v+}^k)]_+ \tag{2}$$

where D (·, ·) is the Euclidean distance between two embeddings. f_v and f_n are the original embedding from the visible and infrared modalities, but f_{v+}^i is an embedding generated by the branch from the visible light mode. j, k denotes different identities in the branch. (2) the first term can pull the generated embedding f_{v+}^i to the original infrared image embedding f_n, to reduce the modal difference between $f_{v+}^{i,j}$ and f_n^i. The second term pushes the resulting embedding f_{v+} away from the embedding f_v of visible light. Enables f_{v+} to learn information feature representations. The third term can make the intra-class distance less than the inter-class distance. Then, we use the embedding hub c_v and the embedding hub c_n for each class The resulting embedding center c_{v+}^i and c_{n+}^i more discriminative and introduce constant term α to balance. (2). Therefore, for embeddings from visible light, the CPM loss is expressed as:

$$\zeta(c_v, c_n, c_{v+}^i) = [D(c_n^j, c_{v+}^{i,j}) - D(c_v^j, c_{v+}^{i,j}) - D(c_v^j, c_v^k) + \partial]_+ \tag{3}$$

Similarly, for the embedded class hub c_{n+}^i generated by the IR, we have:

$$\zeta(c_v, c_n, c_{n+}^i) = [D(c_v^j, c_{n+}^{i,j}) - D(c_n^j, c_{n+}^{i,j}) - D(c_n^j, c_n^k) + \partial]_+ \tag{4}$$

Therefore, the final CPM loss can be expressed as:

$$\zeta_{cpm} = \zeta(c_v, c_n, c_{v+}^i) + \zeta(c_v, c_n, c_{n+}^i) \tag{5}$$

In addition, in order to ensure that the embeddings generated by different branches can capture different representations of information features, we force the different embeddings generated by different branches to be orthogonal to minimize overlapping elements. Therefore, the quadrature loss can be expressed as:

$$\zeta_{ort} = \sum_{m=1}^{i-1} \sum_{n=m+1}^{i} (f_+^{mT} f_+^n) \tag{6}$$

where m and n are the mth and nth embeddings generated from the original embedding, respectively. Orthogonal losses can force the generation of embeddings to learn more about the feature representation.

Multi-stage feature aggregation blocks at different levels have been shown to be useful for semantic segmentation, classification, and detection tasks [16,17,18]. In order to aggregate features from different stages to mine different channel and spatial feature representations, we combine an efficient channel-space multi-stage feature aggregation (SEM) block to aggregate multi-stage features inspired by [19].

ISSN: 2664-9640

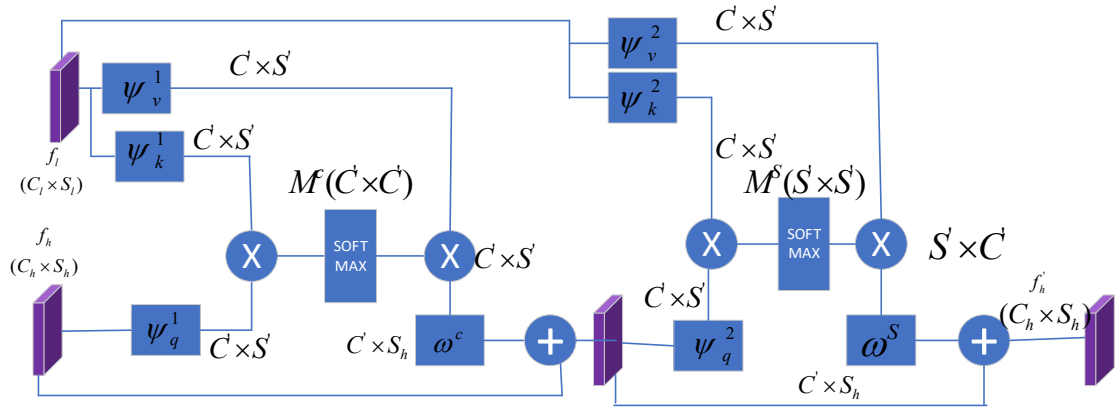

Fig. 4 Structure diagram of the SEM module

As shown in Fig. 4, we consider two types of source features of the channel spatial aggregation blocks at each stage of the backbone network: the low-level feature map before the stage $f_l \in \Re^{C1\times H1\times W1}$ and post-phase high-level feature maps $f_l \in \Re^{C_h\times H_h\times W_h}$ where C, W, and H represent the number of channels, feature width, and feature height, respectively. First, we use three 1×1 convolutional layers, namely: ζ_q^1, ζ_v^1, ζ_k^1, Transform f into three convolutional layers: $\zeta_q^1(f_h)$, $\zeta_v^1(f_l)$ and $\zeta_k^1(f_l)$. Then, we calculate the channel similarity matrix by matrix multiplication and softmax:

$$M^c \in \Re^{C'\times C'}, M^c = F_{soft\max}(\psi_q^1(f_h)\times\psi_k^1(f_l)) \tag{7}$$

Therefore, we multiply by the matrix $\psi_v^1(f_l)$ and M^c to restore the dimensions of the channel, so as to realize the multi-level feature aggregation of the channel. After that, another 1×1 convolutional layer is applied w^c Convert the size of the above feature map to the size of f_h. Finally, we add to the f_h matrix addition to get the output:

$$f_h^c = w^c(\psi_v^1(f_l)\times M^c) + f_h \tag{8}$$

After that, the spatial feature aggregation operation is carried out by using the f_h^c and underlying feature maps f_l obtained by the above operations, which is similar to the channel multi-stage feature aggregation operation. Finally, we get the MFA output as:

$$f_h^s = w^s(\psi_v^2(f_l)\times M^c) + f_h \tag{9}$$

where w^s and ψ_v^2 are two 1 ×× 1 convolutional layers. M^s is a matrix of spatial similarity. In addition to presenting ζ_{cpm} and ζ_{ort}, We also combine cross-entropy loss ζ_{ce} [20] and triplet loss ζ_{tri} [21]. By minimizing the sum of these four losses, the network is optimized together in an end-to-end manner, which is expressed as:

ISSN: 2664-9640

$$\zeta_{total} = \zeta_{ce} + \zeta_{tri} + \Gamma_1 \zeta_{cpm} + \Gamma_2 \zeta_{ort} \tag{10}$$

where Γ_1 and Γ_2 are the coefficient that controls the relative importance of the loss term.

2.3. Experimental Datasets

The SYSU-MM01 dataset [22] contains 491 identities captured by 4 VIS cameras and 2 IR cameras, including All-Search and indoor search modes. For all-search mode, use all images captured by all VIS cameras as the gallery set. For the indoor-search mode, only images captured by two indoor VIS cameras are used as the gallery set. The RegDB dataset [23] consists of 412 identities, each with 10 VIS images and 10 IR images, which are captured by a pair of overlapping cameras.

2.4. Experimental Setup

Firstly, the size of all input images was adjusted to 3 × 384 × 144, and random horizontal flipping and random erase techniques [24] were used in the training stage. The initial learning rate was set to 1 × 10-2, and the learning rate increased to 1 × 10-1 after 10 times through the warm-up strategy. After that, we decayed the learning rate to 1 × 10-2 at 20 epochs and further to 1 × 10-3 and 1 × 10-4 at 60 epochs and 120 epochs, respectively, until a total of 150 epochs. In each mini-batch, we randomly selected 4 VIS images and 4 IR images for 6 identities for training. The training was conducted with an SGD optimizer and the momentum was set to 0.9. For the RegDB dataset, we removed Phase 4 and inserted the proposed CPM module into DEEN after Phase 2. We first compare the proposed DEEN with several state-of-the-art methods to demonstrate the superiority of our approach. The experimental results on the SYSU-MM01 and RegDB datasets are shown in Table 1.

Table 1. Comparison of DEEN with other advanced methods in the SYSU-MM01 and RegDB datasets

METHODS	SYSU-MM01								RegDB							
	All Search				Indoor Search				VIS to IR				IR to VIS			
	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
Hi-CMD [24]	34.9	77.6	—	35.5	—	—	—	—	70.9	86.4	—	66.0	—	—	—	—
JSIA-ReID [25]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9	48.1	—	—	48.9	48.5	—	—	49.3
X-Modality [14]	49.9	89.8	94.0	50.7	—	—	—	—	62.2	83.1	91.7	60.2	—	—	—	—
AlignGAN [28]	42.4	85.0	93.7	40.7	54.9	87.6	94.4	54.3	57.9	—	—	53.6	56.3	—	—	53.4
CM-NAS [26]	60.8	92.1	96.8	58.6	68.0	94.8	97.9	52.4	82.8	95.1	97.9	79.3	81.7	94.1	96.9	77.6
NFS [2]	56.9	91.3	96.5	55.5	62.8	96.5	99.1	69.8	80.5	95.6	99.1	72.5	78.0	90.5	93.1	79.8
LbA [20]	55.4	—	—	54.1	58.5	—	—	66.3	74.2	—	—	67.6	67.5	—	—	72.4
DDAG [27]	54.8	90.4	95.8	53.0	61.0	94.1	98.4	68.0	69.3	86.2	91.5	63.5	68.1	85.2	90.3	61.8
BDTR [28]	17.0	55.4	72.0	19.7	—	—	—	—	33.6	58.6	67.4	32.8	32.9	—	—	32.0
D2RL [16]	28.9	70.6	82.4	29.2	—	—	—	—	43.4	66.1	67.3	44.1	—	—	—	—
GLGCN[15]	57.8	90.7	96.1	55.0	63.5	94.0	97.4	68.4	73.4	87.5	92.2	66.9	71.8	87.0	90.1	65.5
AlignGAN[4]	42.4	85.0	93.7	42.9	54.9	87.6	94.4	54.3	57.9	—	—	53.6	56.3	—	—	53.4
HA T[23]	55.2	92.1	97.3	53.8	62.1	95.1	97.2	69.3	71.8	87.1	92.1	67.5	70.0	86.4	91.6	66.3
DDAG[13]	54.7	90.3	95.8	53.0	61.0	94.0	98.4	67.9	69.3	86.1	91.4	63.4	68.0	85.1	90.3	61.8
PSDRL	59.7	92.7	93.7	58.1	67.1	—	—	72.9	76.4	—	—	68.3	74.3	—	—	66.5
OURS	61.3	93.1	98.2	57.8	65.8	95.2	98.8	73.5	92.0	94.1	99.1	84.1	91.1	94.5	97.2	82.9

SYSU-MM01 and RegDB: From Table 1, we can see that the results on both datasets show that the proposed DEEN achieves the best performance compared to all other state-of-the-art methods. Specifically, for the All-Search mode on the SYSU-MM01, DEEN achieves a Rank-1 accuracy of 61.3 and an mAP accuracy of 57.8. For the Indoor-Search mode, DEEN achieves a Rank-1 accuracy of 65.8 and a MAP accuracy of 73.5. For the VIS to IR mode on RegDB, DEEN achieves a Rank-1 accuracy of 92.0 and an mAP accuracy of 84.1. For the IR to VIS mode, the proposed DEEN also obtains a Rank-1 accuracy of 91.1 and a MAP accuracy of 82.9. The results

ISSN: 2664-9640

verify the effectiveness of the method. In addition, the results also show that the proposed DEEN can effectively reduce the modal difference between VIS and IR modes.

2.5. Ablation Experiments

Table 2. Effect of each component on the performance of the proposed DEEN

CPM	ζ_{cpm}	ζ_{ort}	SEM	R-1	MAP
				57.3	56.2
√				58.2	56.3
√	√			58.4	56.5
√		√		59.6	56.8
			√	58.2	56.3
√	√	√		60.3	57.5
√	√	√	√	61.3	57.8

To validate the effectiveness of each component: In order to evaluate the contribution of each component to DEEN, we performed some studies on the SYSU-MM01 dataset, as shown in Table 2, removing some modules from DEEN and evaluating the impact on performance. The overall settings remain the same, changing the single variable, while only the modules being evaluated are used in or removed from DEEN.

Table 3. The influence of different branches on the model is studied

Methods	SYSU-MM01	
	R-1	MAP
Two branches	59.0	54.0
Three branches	61.0	56.2
Four branches	59.1	54.2

As shown in Table 2, while the CPM module can generate more embeddings using multi-branch convolutional blocks, which slightly improves the performance of the baseline, the results are mediocre.

DEE can greatly improve the performance of the model and effectively reduce the modal difference between VIS and IR images after generating different embeddings under the proposed CPM loss constraints. In addition, the proposed MFA block can improve the performance of the baseline by aggregating features from different stages to mine different channel and spatial feature representations. By integrating DEE, CPM, and MFA into an end-to-end learning framework, DEEN has achieved impressive performance improvements on two challenging VIReID datasets, demonstrating that CPM and SEM can benefit each other in generating different embeddings.

Affects which level of ResNet-50 is plugged into the CPM module. The proposed CPM can be inserted at any stage of the backbone. In our experiments, we used ResNet-50 as the backbone, which has 5 stages: stage0 to stage4. We insert DEE after different phases of ResNet-50 to see how it affects the performance of DEEN.

As can be seen from Table 4, the performance of CPM is gradually improved after stage 0 to stage 3, indicating that the modal gap of CPM becomes smaller and the generation capacity is enhanced at the deeper level of the network. It achieves the best results on the SYSU-MM01 when the CPM is inserted after level 3.

ISSN: 2664-9640

Table 4. The impact of placing CPMs at different Resnet50 levels

Methods	SYSU-MM01	
	R-1	MAP
CPM after stage-0	53.3	52.0
CPM after stage-1	53.6	53.2
CPM after stage-2	55.3	54.2
CPM after stage-3	60.0	55.8
CPM after stage-4	51.2	50.1

However, when the DEE is inserted after stage 4, the performance degrades rapidly as the CPM loss acts directly on the embedding, resulting in an increase in the distance between the generated embedding and the original embedding.

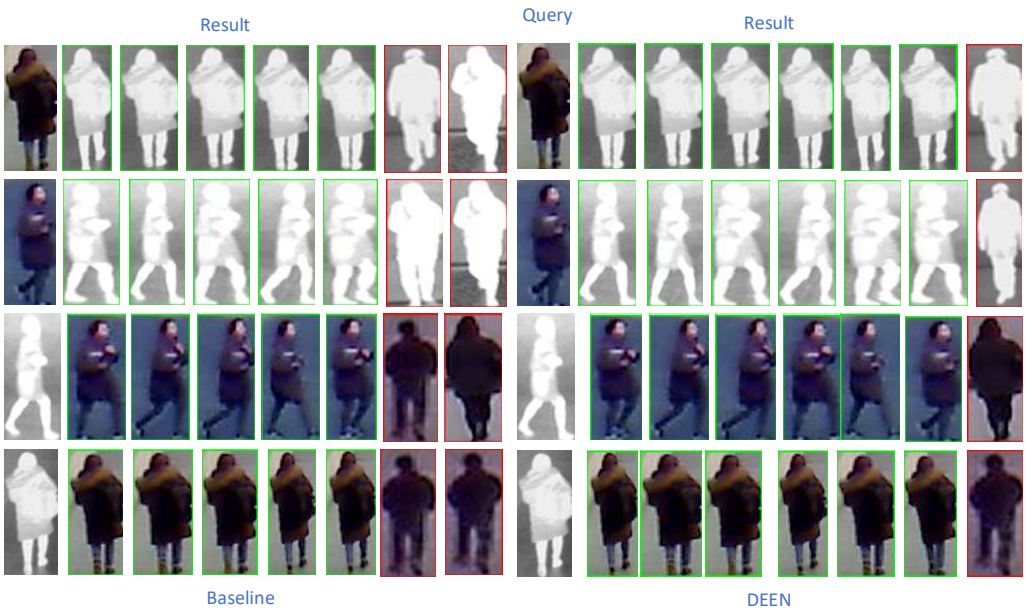

Fig. 5 Some experimental results: basic network (left) and improved network (right)

3. Conclusion

In this paper, a cross-modal pedestrian re-identification method based on excavation channel and spatial enhancement is proposed. In the cross-modal pedestrian re-identification model, the residual shrinkage module is applied to reduce the extraction of redundant features, and the problem of cross-modal local feature matching confusion can be better solved by obtaining the shortest path of the pedestrian space local information distance matrix. The performance of the model has been further improved. The effectiveness of the proposed method is well verified on the public SYSUMMO1 dataset and the RegDB dataset. The next step of research will focus on solving the problems of pedestrian occlusion and low resolution in visible and infrared images of pedestrians.

References

[1] Seokeon Choi, Sumin Lee, Y oungeun Kim, Taekyung Kim,and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification.In Proceedings of the CVPR, pages 10257–10266, 2020.

ISSN: 2664-9640

[2] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, YangYang, and Zengguang Hou. Rgb-infrared cross-modality per-son re-identification via joint pixel and feature alignment. InProceedings of the ICCV, pages 3623–3632, 2019.

[3] Guan-An Wang, Tianzhu Zhang Yang, Jian Cheng, Jian-long Chang, Xu Liang, Zengguang Hou, et al. Cross-modality paired-images generation for rgb-infrared personre-identification. In Proceedings of the AAAI, pages 12144–12151, 2020.

[4] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Y ung-Y uChuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. InProceedings of the CVPR, pages 618–626, 2019.

[5] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong.Infrared-visible cross-modal person re-identification with an x modality. In Proceedings of the AAAI, pages 4610–4617,2020.

[6] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syn-cretic modality collaborative learning for visible infraredperson re-identification. In Proceedings of the ICCV, pages225–234, 2021.

[7] Y ukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. To-wards a unified middle modality learning for visible-infrared person re-identification. In Proceedings of the ACM MM,pages 788–796, 2021.

[8] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang,and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In Pro-ceedings of the CVPR, pages 19366–19375, 2022.

[9] Lei Tan, Y ukang Zhang, Shengmei Shen, Yan Wang,Pingyang Dai, Xianming Lin, Y ongjian Wu, and RongrongJi.Exploring invariant representation for visible-infraredperson re-identification. ArXiv, 2023.

[10] Hanzhe Sun, Jun Liu, Zhizhong Zhang, Chengjie Wang,Yanyun Qu, Y uan Xie, and Lizhuang Ma. Not all pixels arematched: Dense contrastive learning for cross-modality per-son re-identification. In Proceedings of the ACM MM, page5333–5341, 2022.

[11] Yajun Gao, Tengfei Liang, Yi Jin, Xiaoyan Gu, Wu Liu,Yidong Li, and Congyan Lang. Mso: Multi-featurespace jointoptimization network for rgb-infrared person re-identification. In Proceedings of the ACM MM, pages 5257–5265, 2021.

[12] Ziling Miao, Hong Liu, Wei Shi, Wanlu Xu, and HanrongYe. Modality-aware style adaptation for rgb-infrared personre-identification. In Proceedings of the IJCAI, pages 19–27,2021.

[13] Nan Pu, Wei Chen, Y u Liu, Erwin M Bakker, and Michael SLew. Dual gaussian-based variational subspace disentangle-ment for visible-infrared person re-identification. In Pro-ceedings of the ACM MM, pages 2149–2158, 2020.

[14] Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu,Y uan Xie, and Lizhuang Ma. Farewell to mutual infor-mation: V ariational distillation for cross-modal person re-identification. In Proceedings of the CVPR, pages 1522–1531, 2021.

[15] Ye, M., Wang, Z., Lan, X., Yuen, P.C.: Visible thermal person re-identification via dual-constrained top-ranking. In: International Joint Conference on Artificial Intelligence,vol. 1, p. 2 (2018).

[16] Xuesong Chen, Canmiao Fu, Y ong Zhao, Feng Zheng,Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. InProceedings of the CVPR, pages 3300–3310, 2020.

[17] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang.Discriminative feature learning with consistent attention reg-ularization for person re-identification. In Proceedings of theICCV, pages 8040–8049, 2019. 4

[18] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xi-ang Bai. Asymmetric non-local neural networks for semanticsegmentation. In Proceedings of the ICCV, pages 593–602,2019.

[19] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, andRan He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In Proceedingsof the ICCV, pages 11823–11832, 2021.

[20] Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, andZongyuan Sun. Neural feature search for rgb-infrared person re-identification. In Proceedings of the CVPR, pages 587–597, 2021.

ISSN: 2664-9640

[21] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bum-sub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In Pro-ceedings of the ICCV, pages 12046–12055, 2021.

[22] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, andJiebo Luo. Dynamic dual-attentive aggregation learning forvisible-infrared person re-identification. In Proceedings ofthe ECCV, pages 229–247, 2020.

[23] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Y uen.Visible thermal person re-identification via dual-constrainedtop-ranking. In Proceedings of the IJCAI, pages 1092–1099,2018.

[24] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Y ung-Y uChuang, and Shin'ichi Satoh. Learning to reduce dual-leveldiscrepancy for infrared-visible person re-identification. InProceedings of the CVPR, pages 618–626, 2019.

[25] J. Zhang, X, Li, C. Chen, M. Qi, J. Wu, and J. Jiang, "Global-localgraph convolutional network for cross-modality person re-identification," Neurocomputing, vol. 452, pp. 137-146, 2021.

[26] 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." IEEE, 2019, pp. 3622–3631.

[27] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," IEEE Trans. Inf. Forensics Secur, vol. 16, pp. 728-739, 2021.

[28] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person reidentification," ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12362. Springer, 2020, pp. 229–247.