Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 10, No. 3, 2025

Research on Cross-modal Pedestrian Re-identification Based on Transformer

Zheyu Fu, Xiaoguang Hu, Xu Wang

People's public security university of China, Beijing 100038, China

Abstract: Visible infrared human body recognition (VI ReID) is a challenging task for complex modal change retrieval. Existing methods usually focus on extracting discriminative visual features, while ignoring the reliability and commonness of visual features between different modes. In this paper, we propose a new deep learning framework, called multi-scale local progressive transformers (MLT), for effective VI-ReID. In order to reduce the negative impact of modal gap, we first take the gray image as an auxiliary mode, take the Transformer model as the benchmark, and propose a progressive learning strategy. The sea attention mechanism is fused with the dilateformer to further improve the discrimination ability of reliable features, and its feasibility is increased through ablation experiments.

Keywords: Deep learning; Public safety; Transformer.

1. Introduction

The purpose of pedestrian re identification (ReID) is to detect the same person under different cameras. It can be used in many practical applications, such as video surveillance, intelligent security, etc. Recently, with the progress of deep learning, ReID has achieved great success in performance and deployment. However, most existing ReID methods are for visible environments. Therefore, they can be considered as visible visible ReIDs. In fact, most visual ReID methods do not work well at night. In order to solve this problem, the image captured by the infrared camera is considered in the actual scene, which greatly helps the ReID in different modes, thus realizing the visible infrared personnel recognition (VI ReID).

Pedestrian re recognition aims to identify the pedestrian image under the camera with the picture in the picture library. Most existing methods only focus on pedestrian recognition in the white sky. However, at night, when the visible light camera is unable to capture effective pedestrian information, the infrared camera will be used for shooting and the cross modal pedestrian recognition method will be used for detection.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.

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

However, there is a great modal difference between the pictures taken under visible light and those taken by infrared cameras. Compared with single-mode ReID, VI ReID faces three main challenges: (1) The modal gap is large, and it is difficult to unify the identity related features of dual-mode. (2) Compared with visible images, infrared images are more sensitive to light conditions, resulting in less distinctive features of cross modal matching. (3) Due to the large time span, garment changes based on modeling may occur, which further increases the difficulty of robust feature extraction. In order to reduce the heterogeneity difference between the two models, the existing methods (Y e et al. [1] 2021; Gao et al. [2] 2021; Chen et al. [3] (2021b) mainly adopts the dual stream network structure. Before learning patterns to share features, we first use unshared weights to extract features of specific patterns. Although these methods can effectively use the features of specific modes and deal with the differences between modes, they are difficult to extract effective shared features of modes. At the same time, there are also some methods based on the generation of countermeasures network (GAN) (Li D 2020 [4]; Dai et al. 2018 [5]; Wang et al. 2019 [6]), which generate cross modal images by learning the mode conversion mode. However, these methods usually introduce additional image noise and huge computing costs. Therefore, they are difficult to deploy in actual scenarios.

In addition, some excellent methods aim to extract more discriminant information. For example, (Zhu et al. [7]) 2020; Zhang et al. [8] (2021b) horizontally divided the portrait into multiple areas, aligned independent local features, and focused on extracting fine-grained discriminant features. However, due to the huge challenge, over reliance on these features may lead to matching errors, as shown in Figure 1. Therefore, recent work on VI ReID has focused on reducing feature differences between modes. In this work, we propose a new deep learning framework, called Progressive Mode Sharing Converter (PMT), which is used to extract reliable modal invariant features to achieve effective virus recognition. For this reason, we first proposed a transformer progressive learning strategy (Dosovitskiy et al 2020 [9]) to reduce the gap between visible and infrared modes. More specifically, we improve the hard triplet loss, and introduce the gray image as an auxiliary mode to learn the mode independent mode. In

addition, we also propose a modal sharing enhancement loss (MSEL) to reduce the negative impact of modal differences, and use modal sharing information to enhance features. Finally, we propose a Discriminant Center Loss (DCL) to deal with the large variance within the class, which further enhances the discrimination ability of reliable modal shared features. A large number of experiments on SYSU-MM01 and RegDB datasets show that our framework performs better than most of the most advanced methods.

This article uses Transformer as the backbone network. We propose a new deep learning framework (MLT) for effective VI ReID, focusing on extracting more robust modal sharing features.

Transformer model is a deep learning model based on self attention mechanism, which is widely used in natural language processing (NLP) tasks. Its core structure includes two parts: encoder and decoder. Each part is stacked by multiple identical layers. Through the self attention mechanism, Transformer can process all positions in the sequence at the same time, significantly improving the computational efficiency.

The innovations of this paper are as follows: a new variant SeaAttention is proposed to effectively solve the problem of feature unreliability, a new variant DilateFormer is proposed to deal with large intra class differences, and further enhance the identification of modal invariant features. And a large number of experimental results on SYSU-MM01 and RegDB datasets show that the proposed method achieves the new and most advanced performance. In recent years, researchers have attracted the attention of the Attention Mechanism, which mimics the human visual system to focus on areas of data that are of more interest or value. The popular understanding is that when people look at external objects through their eyes, they will focus their eyes on the place of interest, and people's attention will also be on this part. If the gaze shifts from one place to another, then the person's attention will also shift. In the field of computer vision, the attention mechanism selectively assigns different parameters to the data by integrating small-scale models into the model, so that the model can focus on the important areas, so as to reduce the attention to irrelevant areas such as background and occlusions, and finally enable the model to extract features related to the target object and improve the performance of the model.

2. Methods in This Paper

2.1. MLT network structure

In the cross modal pedestrian recognition, the most commonly used method is the dual flow model structure. The network structure used in this paper is shown in the figure. The visible and infrared images are extracted respectively. The visible images are segmented after the transition of a gray module, and the infrared images are directly divided into features for weight sharing. The variant sea attention mechanism recognition is carried out respectively to enhance the extraction of detail features and the semantic learning ability of the dilateformer multi-scale capture. Finally, feature extraction is carried out, and loss function is introduced to derive.

2.2. SeaAttention

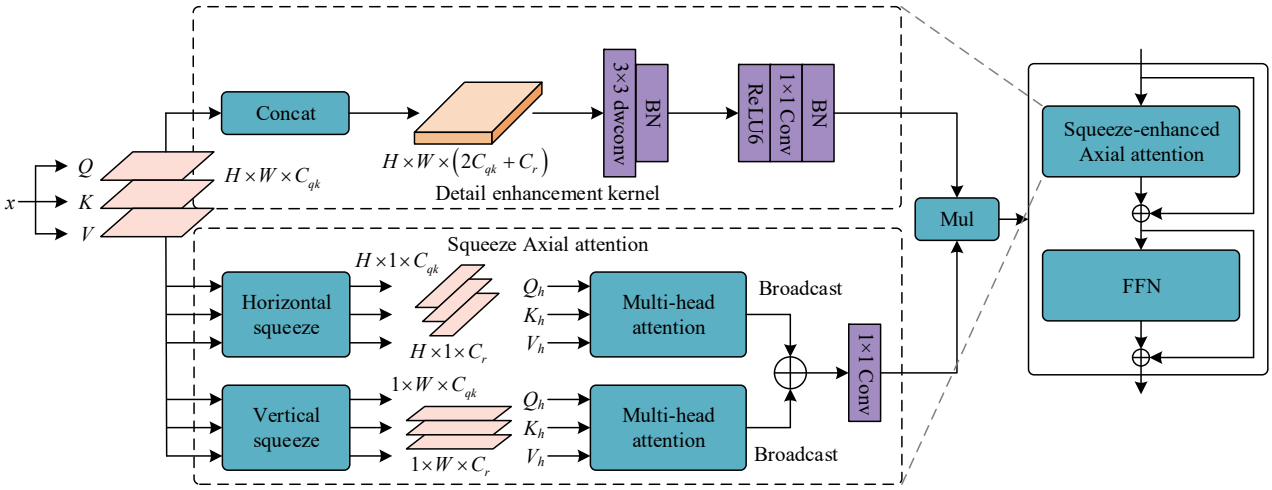

Figure 2. Structure of the Sea module

2.3. Sea Module Introduction

Given input $X \in R^{C\times H\times W}$, through three linear layers Q, K, V:

$$Q \in R^{C_{qk}\times H\times W} = W_q X$$

$$K \in R^{C_{qk}\times H\times W} = W_k X$$

$$V \in R^{C_v\times H\times W} = W_v X$$

Among $W_q, W_k \in R^{C_{qk}\times H\times W}$, $W_v \in R^{C_v\times C}$ is a learnable parameter.

In the horizontal direction (that is, the W direction), horizontal compression is achieved by extracting the average value of Q, K, V feature maps:

$$Q_{(h)} \in R^{C_{qk}\times H} = \frac{1}{W}(Q \in R^{C_{qk}\times H\times W}, I_W)$$

$$K_{(h)} \in R^{C_{qk}\times H} = \frac{1}{W}(K \in R^{C_{qk}\times H\times W}, I_W)$$

$$V_{(h)} \in R^{C_v\times W} = \frac{1}{W}(V \in R^{C_v\times H\times W}, I_W)$$

Among, $I_W \in R^{W\times 1}$ is the total 1 vector of length W, and matrix multiplication is equivalent to adding the elements in the W direction and dividing by the vector($\frac{1}{W}$),get all the

average values of the W characteristic direction.

In the vertical direction (that is, H direction), vertical compression is achieved by extracting the average value of Q, K, V feature maps;

$$Q_{(v)} \in R^{C_v \times W} = \frac{1}{H}(Q \in R^{C_{qk} \times H \times W}, I_H)$$

$$K_{(v)} \in R^{C_v \times W} = \frac{1}{H}(K \in R^{C_{qk} \times H \times W}, I_H)$$

$$V_{(v)} \in R^{C_v \times W} = \frac{1}{H}(V \in R^{C_v \times H \times W}, I_H)$$

Among, $I_W \in R^{H \times 1}$ is the total 1 vector of length H, and matrix multiplication is equivalent to adding the elements in the H direction and dividing by the quantity($\frac{1}{H}$),get the characteristic average value of H direction.

The compression operation retains the global information to a single axis, then applies self attention to model the long-term dependency of the corresponding axis, and finally adds the two:

$$Y_{(h)} \in R^{C_v \times H \times 1} = soft\max(\frac{Q_{(h)} K_{(h)}^T}{\sqrt{d_k}}) V_{(h)}$$

$$Y_{(v)} \in R^{C_v \times 1 \times W} = soft\max(\frac{Q_{(v)} K_{(v)}^T}{\sqrt{d_k}}) V_{(v)}$$

$$Y \in R^{C_v \times H \times W} = Y_{(h)} + Y_{(v)}$$

Compression will cause loss of detail information. You can use convolution to enhance detail information. First, use a set of additional parameters $W_q^{(e)}, W_k^{(e)} \in R^{C_{qk} \times C}, W_v^{(e)} \in R^{C_v \times C}$, apply to Input $X \in R^{C \times H \times W}$, get the corresponding: $Q_{(e)}, K_{(e)}, V_{(e)}$

$$Q_{(e)} = R^{C_{qk} \times H \times W} = W_q^{(e)} X$$

$$K_{(e)} = R^{C_{qk} \times H \times W} = W_k^{(e)} X$$

$$V_{(e)} = R^{C_v \times H \times W} = W_v^{(e)} X$$

Then, the three are spliced and fed into the Conv and BatchNorm layers on the channel dimension to model local spatial dependencies, thus enhancing local detail awareness:

$$X_{qkv} \in R^{(2C_{qk}+C_v) \times H \times W} = Concat(Q_{(e)} + K_{(e)} + V_{(e)})$$

$$X_{qkv} \in R^{(2C_{qk}+C_v) \times H \times W} = BatchNorm(Conv_{3\times3}(X_{qkv}))$$

Then, use ReLU and BatchNorm with order 1 convolution to convert the number of channels $2C_{qk} + C_v$ map to C to generate detail enhancement features:

$$X_{qkv} \in R^{C \times H \times W} = BatchNorm(Conv(\mathrm{Re}\,LU(X_{qkv})))$$

Finally, the feature fusion given by enhanced features and Squeeze Axial attention. The fused output is obtained by generating the weight of the output Y squeezing axial attention through the sigmoid function, and then performing point by point multiplication with the enhancement feature:

$$output = sigmoid(Y) \odot X_{qkv}$$

2.4. Dilateformer module

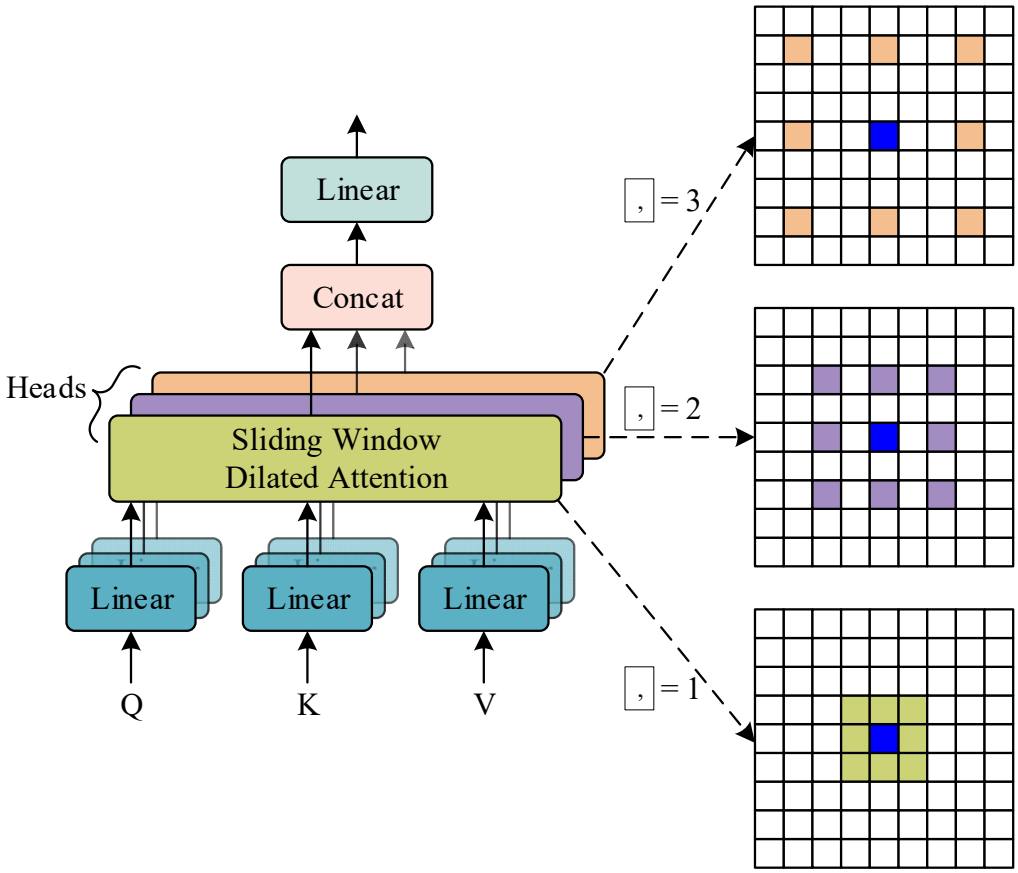

Figure 3. Structure of the Dilateformer module

Vision Transformer (ViT) can model the dependencies between any image patches, but it will lead to the calculation cost of the quadratic. Inspired by CNN, another research route of ViT is to explore local attention, that is, to interact with image patches only in small spatial areas. This can reduce the calculation cost, but it will reduce the receptive field and reduce the modeling ability of the model for long-term dependence. Therefore, an efficient ViT structure is proposed to achieve the balance between the calculation amount and the receptive field. There are two key attributes of ViT in the shallow layer: locality and sparsity (that is, for query, highly relevant keys are sparsely distributed around query), which indicates the redundancy of global dependency modeling in the shallow layer of ViT (that is, most long-distance patches are irrelevant, and the calculation between many redundant patch pairs in the global environment should be reduced). Therefore, the author proposes Sliding Window Dilated Attention (SWDA), which selects a small number of patches in the surrounding environment and then executes self attention. In order to further utilize the information in the receptive field, multi-scale dilative attention (MSDA) is proposed, which simultaneously captures semantic dependencies of different scales. MSDA sets different inflation rates for different attention, realizing the ability of multi-scale representation learning.

According to the two key attributes of ViT in shallow layer: locality and sparsity, SWDA is proposed, where keys and values are sparsely selected in the sliding window centered on query, and then self attention is performed on these patches. The operation of SWDA can be described as:

$$X = SWDA(Q, K, V, r)$$

For the query vector at position, SWDA uses (i, j) as the center, when the dimension is $w \times w$ select keys and values sparsely in the sliding window of, and then execute self attention. And defines the expansion rate $r \in N^+$ to control the degree of sparsity. SWDA in (i, j) the operation of is as follows:

$$x_{ij} = Attention(q_{ij}, K_r, V_r)$$

$$= Soft\max(\frac{q_{ij}K_r^T}{\sqrt{d_k}})V_r, 1 \le i \le W, 1 \le j \le H$$

Among K_r and V_r is the height and width of the feature map. K and V represent the given position of keys and values (3) selected from the characteristic graph and (i, j) query vector at, select the coordinate set $\left(i^{'}, j^{'}\right)$ the keys and values of are used for self attention operation:

$$\left\{\left(i^{'}, j^{'}\right) | i^{'} = i + p \times r, j^{'} = j + q \times r\right\}$$

$$-\frac{w}{2} \le p, q \le \frac{w}{2}$$

SWDA performs self attention operation on all query patches in the way of sliding windows. For the edge query of the feature graph, we simply use the zero padding strategy commonly used in the convolution operation to maintain the size of the feature graph. By sparsely selecting keys and values centered on query, the proposed SWDA explicitly satisfies locality and sparsity, and can effectively model long-range dependencies.

Multi Scale Expanded Attention (MSDA)

In order to explore the sparsity of self attention mechanism at different scales, the author further proposes MSDA module to extract multi-scale semantic information. The structure of MSDA is shown in Figure 19. Given characteristic diagram $X \in R^{C \times H \times W}$, the corresponding queries, keys, and values are obtained through the linear layer. Then, we divide the channels of the feature graph into n attention heads, and perform multi-scale SWDA at different expansion rates in different attention heads. Specifically, our MSDA is defined as:

$$h_i = SWDA(Q_i, K_i, V_i, r_i), 1 \le i \le n,$$

$$X = Linear(Concat\left[h_1, ..., h_n\right])$$

Wherein, r_i is the expansion rate of the ith attention head, Q_i, K_i, V_i it represents the feature map fed to the ith attention head. The output of each attention head is spliced on the channel, and then fed to the linear layer for feature aggregation.

2) By setting different inflation rates for different attention heads, our MSDA effectively aggregates semantic information of different scales, and effectively reduces the redundancy of the self attention mechanism without complex operations and additional computing costs.

3. Experimental dataset

The SYSU-MM01 dataset [23] contains 491 identities captured by 4 VIS cameras and 2 IR cameras, including All Search and indoor search modes. For all - search mode, all images captured by all VIS cameras are used as a gallery. For indoor - search mode, only the images captured by two indoor VIS cameras are used as the atlas. The RegDB dataset [24] consists of 412 identities, each of which has 10 VIS images and 10 IR images captured by a pair of overlapping cameras.

4. Experimental setup

Our server is 3090GPU and the overlapping stride is set to 12 to balance the speed and performance. The size of all character images is adjusted to 256 × 128, with horizontal flip and random erase functions to achieve data enhancement. The batch size is set to 64, including a total of 8 different identities. For each identity, four visible light images and four infrared images are used for sampling.

We first compare the DEEN proposed with several most advanced methods to prove the superiority of our method. See Table 1 for experimental results on SYSU-MM01 and RegDB data sets

The experimental results are as follows

Table 1. Comparison of MLT with other advanced methods in the SYSU-MM01 datasets

Methods	All search					Indoor search				
	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-Pad (Wu et al. 2017[10])	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
Hi-CMD (Choi et al. 2020[11])	34.94	77.58	-	35.94	-	-	-	-	-	-
CMSP (Wu et al. 2020[12])	43.56	86.25	-	44.98	-	48.62	89.50	-	57.50	-
expAT Loss (Ye et al. 2020a[13])	38.57	76.64	86.39	38.61	-	-	-	-	-	-
AGW (Ye et al. 2021[14])	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
HAT (Ye, Shen, and Shao 2020[15])	55.29	92.14	97.36	53.89	-	62.10	95.75	99.20	69.37	-
LBA (Park et al. 2021[16])	55.41	91.12	-	54.14	-	58.46	94.13	-	66.33	-
NFS (Chen et al. 2021b[17])	56.91	91.34	96.52	55.45	-	62.79	96.53	99.07	69.79	-
MSO (Gao et al. 2021[18])	58.70	92.06	97.20	56.42	42.04	63.09	96.61	-	70.31	-
CM-NAS (Fu et al. 2021[19])	61.99	92.87	97.25	60.02	-	67.01	97.02	99.32	72.95	-
MID (Huang et al. 2022[20])	60.27	92.90	-	59.40	-	64.86	96.12	-	70.12	-
SPOT (Chen et al. 2022[21])	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
FMCNet (Zhang et al. 2022[22])	66.34	-	-	62.51	-	68.15	-	-	74.09	-
PMT (Ours)	68.03	93.36	98.54	65.38	52.86	72.06	97.23	99.23	75.52	73.74

Table 2. Comparison of MLT with other advanced methods in the and RegDB datasets

Methods	V to I		I to V	
	r=1	mAP	r=1	mAP
DDAG (Ye et al. 2020b)	69.34	63.46	68.06	61.80
expAT Loss (Ye et al. 2020a)	66.48	67.31	67.45	66.51
AGW (Ye et al. 2021)	70.05	66.37	70.49	65.90
HAT (Ye, Shen, and Shao 2020)	71.83	67.56	70.02	66.30
MSO (Gao et al. 2021)	73.6	66.9	74.6	67.5
LBA (Park et al. 2021)	74.17	67.64	72.43	65.46
NFS (Chen et al. 2021b)	80.54	72.10	77.95	69.79
MCLNet (Hao et al. 2021)	80.31	73.07	75.93	69.49
SPOT (Chen et al. 2022)	80.35	72.46	79.37	72.26
PMT (Ours)	85.84	77.45	82.06	74.13

From Table 1and 2, we can see that the results on the two data sets show that the proposed MLT achieves the best performance compared with all other most advanced methods. Specifically, for the All Search mode on SYSU-MM01, MLT achieves the Rank-1 accuracy of 68.03 and the mAP accuracy of 65.38. For the Indoor Search mode, MLT achieves the Rank-1 accuracy of 72.06 and the MAP accuracy of 75.52. For the VIS to IR mode on RegDB, MLT achieves the Rank-1 accuracy of 85.84 and the mAP accuracy of 77.45. For IR to VIS mode, the proposed MLT also obtained the Rank-1 accuracy of 82.06 and the MAP accuracy of 74.13. The results verify the effectiveness of this method. In addition, the results also show that the proposed MLT can effectively reduce the modal difference between VIS and IR modes.

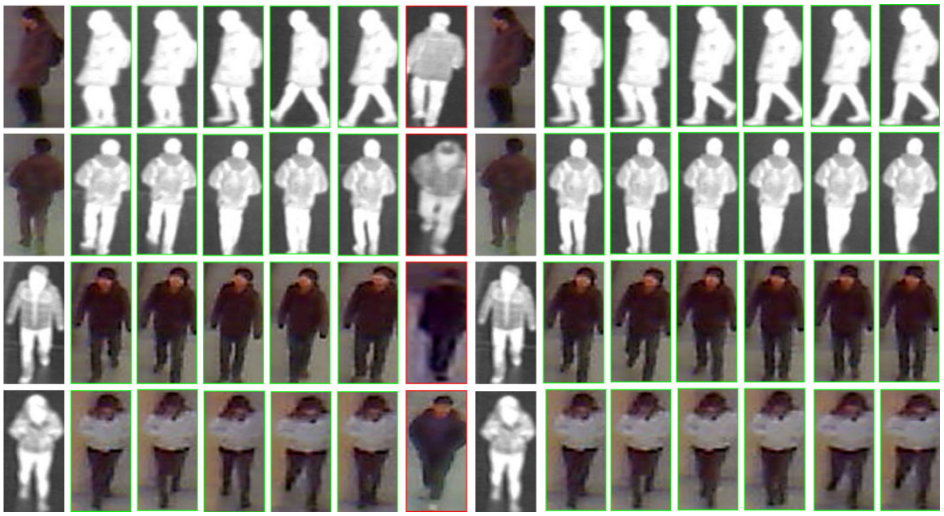

Figure 5. Some experimental results: basic network (left) and improved network (right)

5. Conclusion

In this paper, a Transformer based cross-modal person re-identification method is proposed. Two variant attention mechanisms are proposed to improve it, grayscale images are introduced as auxiliary modalities to learn modal-independent patterns, and two new loss functions are proposed to better solve the problem of cross-modal local feature matching confusion of pedestrians. 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. There is still more room for research and improvement of the loss function

In recent years, researchers have been studying more and more in the field of cross-modal person re-identification of visible light and infrared images, and different types of loss functions have emerged. Therefore, there is still a lot of room for the design of the loss function.

There is still more room for research on network structure

The feature extraction networks experimented in this paper are single-stream ResNet50 networks and dual-stream ResNet50 networks with shared weights, and there are some other types of feature extraction networks in the research field of this paper, such as dual-stream networks with partial weight sharing, dual-stream networks with independent weights, and three-channel networks. Different network architectures often have different identification effects. In the future, different network architectures can be designed to improve the recognition accuracy of the model. (1) There is room for further improvement in the quality of the generated pedestrian images

Although the conversion between the visible image and the infrared image can be realized, and the pedestrian posture in the original image is retained in the generated pedestrian image, the clothing color of the generated pedestrian image may change, which will affect the recognition accuracy of the model. At this point, the generative adversarial network can be improved to improve the quality of the generated pedestrian images, and at the same time, the generated pedestrian images can be screened to remove those images of poor quality to improve the recognition accuracy of the model. (2) There is room for further optimization of the attention mechanism in the recognition model

In this paper, in the research of cross-modal pedestrian re-identification model based on spatial and channel attention mechanism, the spatial and channel attention module of RGA is introduced to improve the recognition accuracy. However, with the development of computer vision in recent years, different types of attention mechanisms have emerged, and other types of attention mechanisms may further improve the recognition effect in this field. Therefore, there is room for further improvement in the attention mechanism.

References

[1] Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C.2021. Deep learning for person re-identification: A surveyand outlook. IEEE Transactions on Pattern Analysis andMachine Intelligence, 44(6): 2872–2893.

[2] Gao, Y.; Liang, T.; Jin, Y.; Gu, X.; Liu, W.; Li, Y.; and Lang,C. 2021. MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, 5257–5265.

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

[4] Li, D.; Wei, X.; Hong, X.; and Gong, Y. 2020. Infraredvisible cross-modal person re-identification with an xmodality. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 4610–4617.

[5] Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Crossmodality person re-identifcation with generative adversarial training. In International Joint Conference on Artifcial Intelligence, volume 1, 6.

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

[7] Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; and Tao, D. 2020. Hetero-center loss for cross-modality person re-identifcation. Neurocomputing, 386: 97–109.

[8] Zhang, L.; Du, G.; Liu, F.; Tu, H.; and Shu, X. 2021b. Global-local multiple granularity learning for cross-modality visible-infrared person reidentifcation. IEEE Transactions on Neural Networks and Learning Systems.

[9] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.

[10] Wu,A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, 5380–5389.

[11] Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. 2020. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10257–10266.

[12] Wu, A.; Zheng, W.-S.; Gong, S.; and Lai, J. 2020. Rgb-ir person re-identification by cross-modality similarity preservation. International Journal of Computer Vision, 128(6): 1765–1785.

[13] Ye, H.; Liu, H.; Meng, F.; and Li, X. 2020a. Bi-directional exponential angular triplet loss for RGB-infrared person re-identification. IEEE Transactions on Image Processing, 30: 1583–1595.

[14] Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872–2893.

[15] Ye, M.; Shen, J.; and Shao, L. 2020. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. IEEE Transactions on Information Forensics and Security, 16: 728–739.

[16] Park, H.; Lee, S.; Lee, J.; and Ham, B. 2021. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 12046–12055.

[17] Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; and Sun, Z. 2021b. Neural feature search for rgb-infrared person re-identification. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 587–597.

[18] Gao, Y.; Liang, T.; Jin, Y.; Gu, X.; Liu, W.; Li, Y.; and Lang, C. 2021. MSO: Multi-feature space joint optimization network for rgb-infrared person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, 5257–5265.

[19] Fu, C.; Hu, Y.; Wu, X.; Shi, H.; Mei, T.; and He, R. 2021. CM-NAS: Cross-modality neural architecture search for visible-infrared person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 11823–11832.

[20] Huang, Z.; Liu, J.; Li, L.; Zheng, K.; and Zha, Z.-J. 2022. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. arXiv preprintarXiv:2203.01735.

[21] Chen, C.; Ye, M.; Qi, M.; Wu, J.; Jiang, J.; and Lin, C.-W. 2022. Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification. IEEE Transactions on Image Processing, 31: 2352–2364.

[22] Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. FMCNet: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7349–7358.

[23] 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.

[24] 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.