Journal of New Media

Tech Science Press

DOI: 10.32604/jnm.2023.038580
Review

Review of Visible-Infrared Cross-Modality Person Re-Identification

Yinyin Zhang*

School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Yinyin Zhang. Email: mkkdzz2428@163.com
Received: 19 December 2022; Accepted: 05 February 2023; Published: 16 June 2023

Abstract: Person re-identification (ReID) is a sub-problem under image retrieval. It is a technology that uses computer vision to identify a specific pedestrian in a collection of pictures or videos. The pedestrian image under cross-device is taken from a monitored pedestrian image. At present, most ReID methods deal with the matching between visible and visible images, but with the continuous improvement of security monitoring system, more and more infrared cameras are used to monitor at night or in dim light. Due to the image differences between infrared camera and RGB camera, there is a huge visual difference between cross-modality images, so the traditional ReID method is difficult to apply in this scene. In view of this situation, studying the pedestrian matching between visible and infrared modalities is particularly crucial. Visible-infrared person re-identification (VI-ReID) was first proposed in 2017, and then attracted more and more attention, and many advanced methods emerged.

Keywords: Person re-identification; cross-modality

1 Introduction

With the ongoing advancement of intelligent surveillance, the traditional way of manually processing and identifying surveillance video has been difficult to adapt to the current development trend. Person re-identification (ReID) is the latest key component in the field of video surveillance research, and its purpose is to accurately and quickly identify the target pedestrians among a large number of pedestrians under cross-devices. ReID is based on pedestrian detection, which can greatly reduce the consumption of human resources, automatically analyze pedestrians and their behaviors, play a positive role in promoting crime prevention and maintaining public order, provide an important guarantee for safety monitoring in public places, and have broad application prospects.

With the vigorous growth of deep learning research, increasingly more researchers are looking towards ReID utilizing deep learning, and achieved a series of successes. At present, most of the ReID methods deal with the matching of pedestrian images in visible-visible scenes. However, in practical application scenes, effective pedestrian features are challenging for RGB cameras to capture at night or in a weak light environment, so more and more infrared cameras are used for monitoring at night or in dim light. Because of the imaging difference between visible camera and infrared camera, there

is a huge visual difference between cross-modality images, so the traditional ReID method is difficult to apply in this scene. In view of this situation, effectively resolving the visible-infrared cross-modality person re-identification (VI-ReID) problem is crucial for public safety.

VI-ReID is one of the difficult issues in computer vision. VI-ReID seeks to match images between visible and infrared modalities, which is distinct from the previous ReID. Its goal is to give a visible (infrared) image of a specific identity and search the corresponding infrared (visible) image from the gallery set. The three categories of VI-ReID methods currently in use are as follows: methods based on feature learning, methods based on metric learning and methods based on image conversion. Method based on feature learning aims to close the gap across heterogeneous images in the same feature space. Method based on metric learning mainly emphasis on designing different measurement methods or loss functions to increase the model's capacity for generalization. Method based on image conversion realizes style conversion from the image level through GAN, closing the distance between the two modalities.

2 Visible-Infrared Cross-Modality Person Re-Identification Datasets

There are two widely used datasets for VI-ReID task: SYSU-MM01 and RegDB. The first substantial dataset for VI-ReID is SYSU-MM01. It was proposed by Wu et al. [1] in 2017, and then it was widely used for the training and testing in VI-ReID task. Fig. 1 displays a variety of pedestrian images captured by several spectrum cameras in SYSU-MM01. The images inside a single column represent the same identity. From 6 cameras, including 4 RGB cameras (cam 1,2,4,5) and 2 infrared cameras (cam 3,6), it includes 30,071 RGB photos and 15,792 IR images. There are 491 accessible IDs in this dataset.

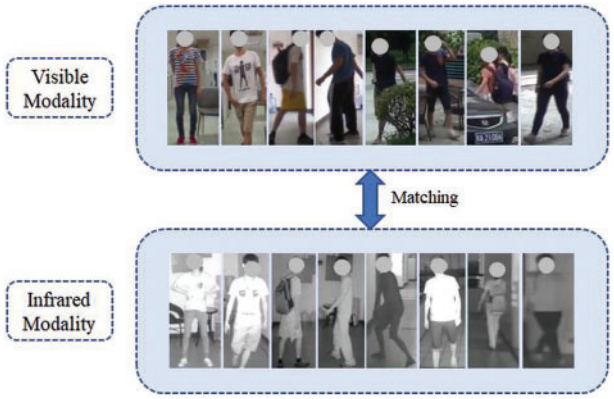

Figure 1: Some examples in SYSU-MM01 dataset

RegDB was not first proposed to solve the cross-modality problem. The author thinks that using thermal image can reduce the negative effects caused by noise, background and pedestrian appearance changes, and it is used to assist the ReID task in RGB images. There are 412 pedestrians in this dataset. 10 visible images and 10 thermal images are taken for each person, and the poses of RGB images and thermal images of each identity are in one-to-one correspondence. 4120 visible images and 4120 associated thermal images make up the dataset. There are 158 men and 254 women among the 412 people. In addition, 156 people were photographed from the front, and the remaining 256 people were photographed from the back. The image of this dataset is small and the definition is poor, but the

posture of pedestrians with the same identity changes little. All these factors reduce the difficulty of VI-ReID task on RegDB.

3 Methods Based on Feature Learning

Studying how to design an appropriate network architecture and extracting robust and discriminative modality-shared features is the cornerstone to feature learning based methods, so as to narrow the differences between modalities.

In 2017, Wu et al. [1] defined the problem of cross-modality person re-identification for the first time, analyzed three network architectures, proposed a data preprocessing method of deep zero-padding, and provided a new dataset named SYSU-MM01 for VI-ReID for the first time. As shown in Fig. 2, the deep zero-padding method is to convert the visible image into a single-channel gray image and place it in the first channel, place its zero-padding image in the second channel, place the infrared image directly in the second channel, and place its zero-padding image in the first channel, to enable flexible learning of the information associated with a specific domain.

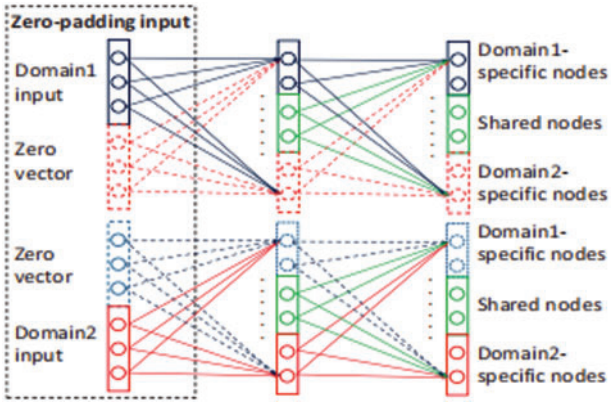

Figure 2: Explanation of deep zero-padding method

Later, a lot of work began to use two-stream network to learn modality-shared features. Ye et al. [2] proposed a hierarchical cross-modality matching model (HCML). As shown in Fig. 3, by jointly improving modality-specific and modality-shared features, the model is accomplished. And the framework is divided into two stages: feature learning and metric learning. A dual-stream network is built in the initial stage to learn the features of the input images from two modalities, and then the similarity is learned by identity loss and contrastive loss. The latter stage focuses on discriminant matching modality training.

In view of the loss of a large amount of modality-specific information due to the concentrated learning of common features of cross-modality images, a new cross-modality shared-specific feature transfer algorithm (cm-SSFT) was proposed by Lu et al. [3] to mine the potential information between common features and unique features between two modalities. As shown in Fig. 4, the algorithm first uses SSTN module to determine the similarity within and between modalities. Then, modality-shared and modality-specific features are spread among different modalities to compensate for the lack of specific information and improve the shared features. A project adversarial learning and a modality adaptation module are added to the feature extractor to obtain distinctive and complementary shared features and specific features. Thereby effectively utilizing the shared information and specific information of each sample.

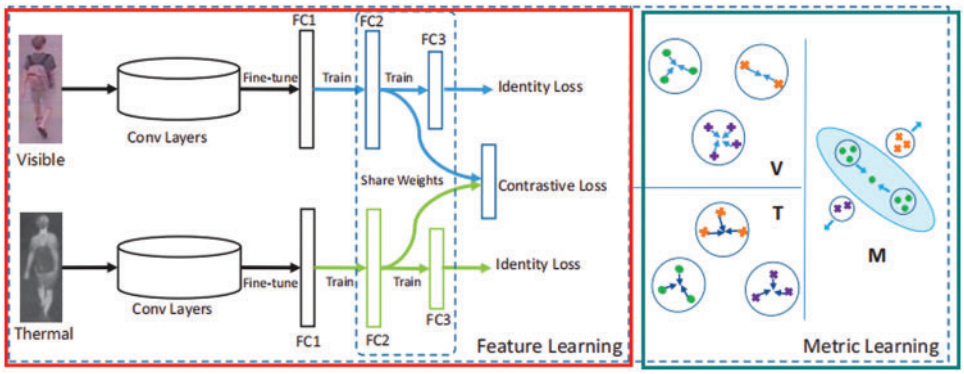

Figure 3: The framework of HCML

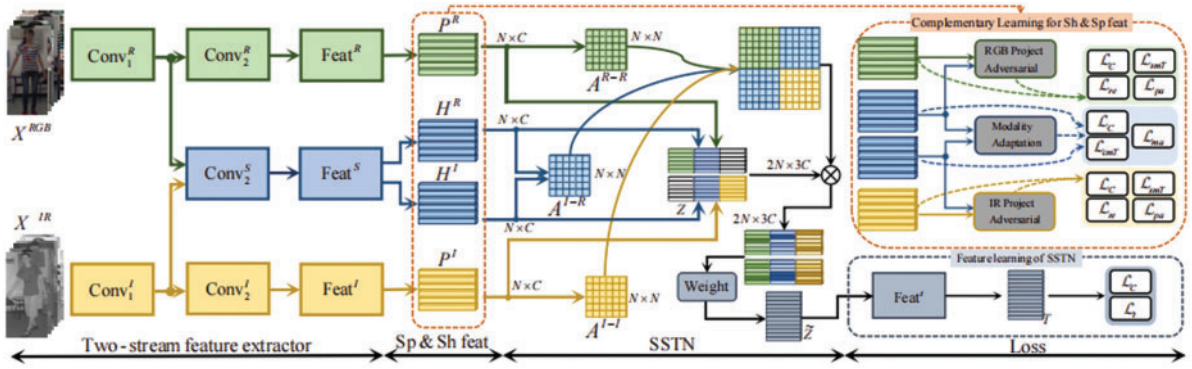

Figure 4: The framework of cm-SSFT

The earlier methods are particularly sensitive to dislocation since they typically extract features from evenly divided parts or learn the global representation. Chen et al. [4] proposed a structure-aware location transformer (SPOT) network, which uses the structure and location information to learn the modality-shared eatures of semantic perception. As shown in Fig. 5, the network is made up of two basic components: transformer-based part interaction (ASR) and attended structure representation (TPI). To deal with the complex background noise in each modality, ASR uses a better attention mechanism to learn the appearance features associated with the structure, and uses the relationship between the heatmaps of key points of human body to represent the structure features. Under the guidance of the structure features, by evaluating each node's relevance, the appearance features of the modality are updated. TPI aggregates differentiated regional features by obtaining context and structure information between different body parts, thus solving the problem of spatial dislocation.

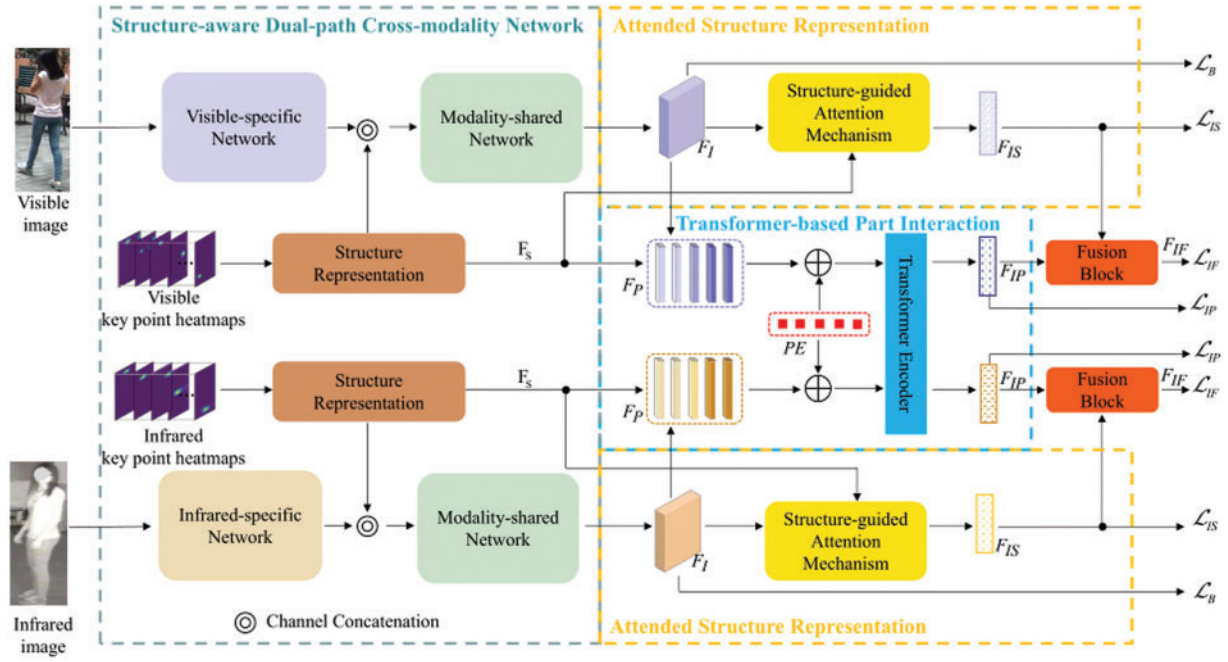

Figure 5: The framework of SPOT

4 Methods Based on Metric Learning

The fundamental goal of the metric learning approach is to develop new metric techniques or loss functions that will enhance the model's capacity for generalization. The objective is to increase the distance between pedestrian photos with various identities across modalities while decreasing the distance between two pedestrian images with the same identity.

In order to fix the issue that the difference of the same pedestrian in different modalities may even be greater than the difference between different pedestrians, Ye et al. [5] suggest a bi-directional ranking loss (BDTR) that considers changes both within and between modalities, shortens the image distance of the same pedestrian in different modalities, and fuses the traditional cross entropy loss, in order to combine the features of the same pedestrian across different modalities. The formula for the bi-directional ranking loss is

$$L_{bi\_rank} = \sum_{\forall y_i=y_j, y_i\neq y_k} \max\left[\rho_1 + D\left(x_i, z_j\right) - \min_{\forall y_i\neq y_k} D\left(x_i, z_k\right), 0\right] + \sum_{\forall y_i=y_j, y_i\neq y_k} \max\left[\rho_1 + D\left(z_i, x_j\right) - \min_{\forall y_i\neq y_k} D\left(z_i, x_k\right), 0\right] \tag{1}$$

Zhao et al. [6] migrates the network of ReID in single-modality to cross-modality scene for the first time, proposes a new feature learning framework, hard pentaplet and identity loss network (HPILN), and designs a new hard pentaplet loss. A hard global triplet loss (HGT) based on a cross-modality batch structure plus a hard cross-modality triplet loss make up the pentaplet loss (HCT), so as to

combine the identity loss and effectively cope with cross- and intra-modality change, and to increase the model's accuracy. This is how the HGT loss and HCT loss are calculated

$$L_{hgt} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\left[\alpha + \overbrace{\max_{p=1,\ldots,2K} d\left(x_i^a, x_i^p\right)}^{\text{hardest global positive}} - \overbrace{\min_{\substack{n=1,\ldots,2K \\ j=1,\ldots,P}} d\left(x_i^a, x_j^n\right)}^{\text{hardest global negative}}\right]_+ \tag{2}$$

$$L_{hct} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\left[\alpha + \overbrace{\max_{cp\in A} d\left(x_i^a, x_i^{cp}\right)}^{\text{hardest cross-modality positive}} - \overbrace{\min_{\substack{cn\in A \\ k=1,\ldots,K}} d\left(x_i^a, x_k^{cn}\right)}^{\text{hardest cross-modality negative}}\right]_+ \tag{3}$$

A dual-stream local feature network (TSLFN) is developed in the paper [7] to lessen intra-class variation and enhance intra-class cross-modality similarity, and the heterogeneous center loss is proposed. Heterogeneous center loss supervises the network to learn the invariable information cross-modality by constraining the distance between the centers within the two modalities, so as to make up for the difference between the two modalities. Eq. (4) below represents the formulation of heterogeneous center (HC) loss.

$$L_{hc} = dist\left(c_v^i - c_t^i\right) \tag{4}$$

where $c_v^i = \frac{1}{K}\sum_{m=1}^{K} x_{i,m}^v$, $c_t^i = \frac{1}{K}\sum_{m=1}^{K} x_{i,m}^t$.

Jia et al. [8] proposed an unique similarity inference metric to overcome the cross-modality discrepancy aiming optimal cross-modality picture matching (SIM), which uses the similarity of sample images in the consistent modality to avoid the cross-modality differences in image matching, trains by continuous similarity graph reasoning and mutual nearest neighbor reasoning, and uses the similarity of sample images to mine the cross-modality sample similarity from two different perspectives, thus narrowing the differences between modalities. SIM loss is formulated as follow

$$\begin{aligned} d_{SIM} &= \alpha d_s + (1-\alpha)\, d \\ d_s\left(q_i, g_j\right) &= \frac{1}{K}\sum_{k=1}^{K} d^{(k)}\left(q_i, g_j\right) \\ d_M\left(q_i, g_j\right) &= 1 - \frac{\left|N_c\left(q_i, k_q, d_s\right) \cap R_i^*\left(g_j, k_g, D_{g,g}\right)\right|}{\left|N_c\left(q_i, k_q, d_s\right) \cup R_i^*\left(g_j, k_g, D_{g,g}\right)\right|} \end{aligned} \tag{5}$$

5 Methods Based on Image Conversion

The method based on modality conversion mainly uses GAN to generate images. Converting visible (infrared) images into corresponding visible (infrared) images is the fundamental goal, and to convert cross-modality ReID problem into single-modality ReID problem, it can successfully lessen modal discrepancies and increase recognition rates.

Wang et al. [9] thinks that the previous methods only rely on the constraint of feature level to deal with cross-modality data is not enough, so the inter-modality differences and intra-modality differences are treated separately, and a dual-level discrepancy reduction learning (D2RL) scheme is proposed. As shown in Fig. 6, the framework is divided into two parts: image-level discrepancy reduction sub-network (TI) and feature-level discrepancy reduction sub-network (TF). The visible (infrared) image is generated into its corresponding infrared (visible) image by TI, forming a unified multispectral image, which reduces the difference between modalities. On the basis of unification, TF uses the traditional ReID method to reduce the appearance difference.

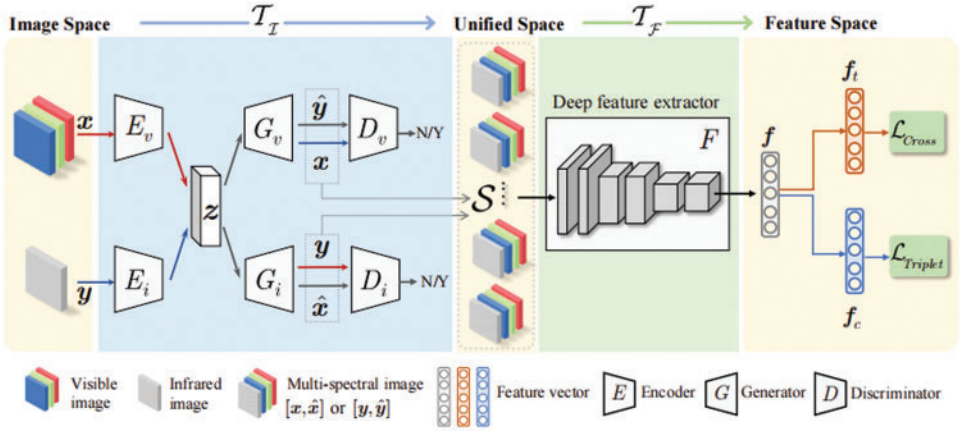

Figure 6: The framework of D²RL

A hierarchical cross-modality disentanglement (Hi-CMD) method was put out by Choi et al. [10]. As shown in Fig. 7, the author introduced an ID-preserving person image generation network to extract the features with invariable posture and illumination, and maintained the identity features of a specific person, so as to automatically separate the ID-discriminative features and the ID-excluded features from visible images, and ID-discriminative features were only used for cross-modality matching.

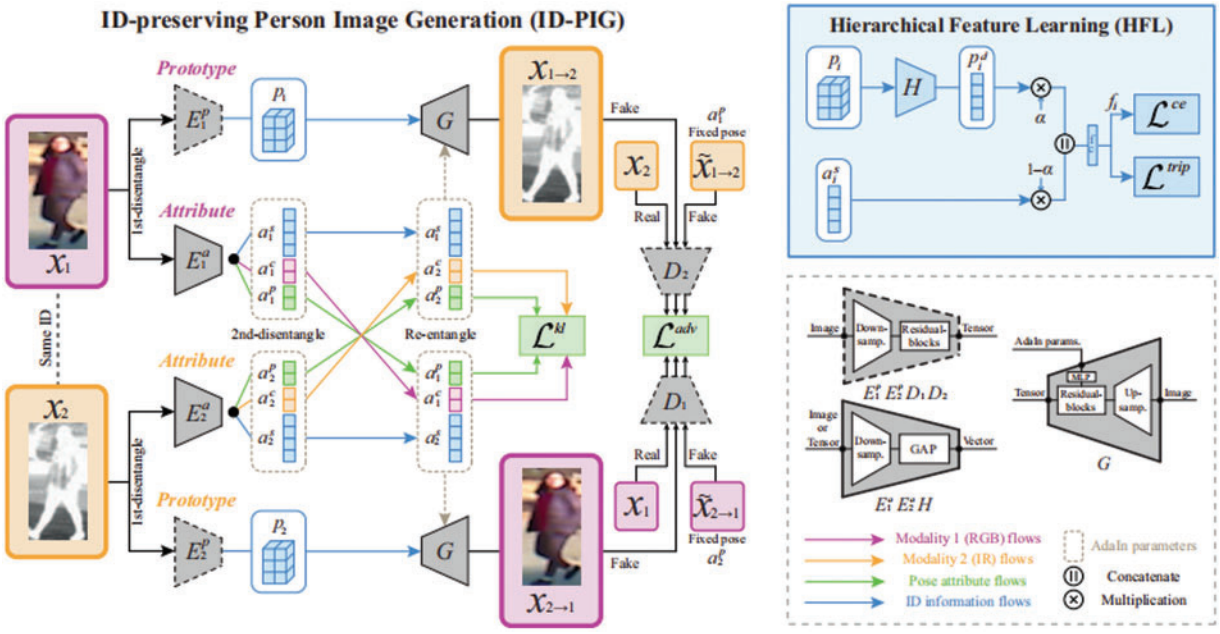

Figure 7: The framework of Hi-CMD

In order to minimizing the discrepancies in modality, Li et al. [11] introduced an auxiliary intermediate modality X and proposed a XIV-ReID method, which reformulated visible-infrared learning as a three-modality learning problem. As shown in Fig. 8, to create X-modality images, the lightweight generator first extract information from visible and infrared images. Then the three

modality images are input into the feature learner. Finally, two modality constraints are designed for regularized feature representation and classification, and in a common space, three modalities' cross-modality knowledge is learned.

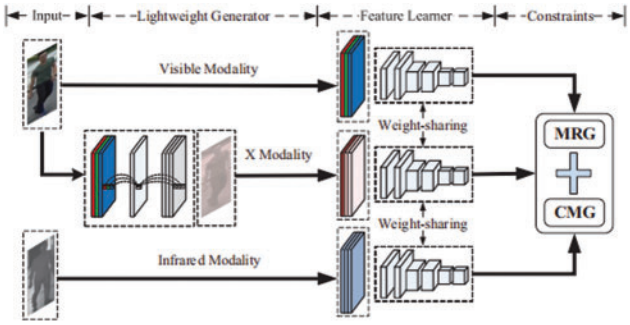

Figure 8: The framework of XIV

6 Conclusion

As a technology of using machine to process video data to identify pedestrians under cross-cameras, person re-identification can quickly and accurately find specific pedestrians, which has strong practicability in daily life and promotes the development of intelligent security and video surveillance. At present, most of the work focuses on the single-modality person re-identification in visible scenes. However, in practical application scenes, RGB cameras are difficult to be used normally at night or in the environment with insufficient light, and cannot capture effective pedestrian features. As science and technology have advanced, cameras that can switch infrared modalities are widely used in monitoring tasks in dark places. Therefore, effectively solving VI-ReID task is of great significance to public safety, crime prevention and criminal investigation. This paper investigates the research progress of VI-ReID. The three primary categories of VI-ReID methods now in use are: methods based on feature learning, methods based on metric learning and methods based on image conversion. The key lies in mapping the information of the two modalities to the same public space, then learning the shared features between the modalities, and minimize the discrepancies between the many manifestations of the same identity.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Wu, W. -S. Zheng, H. -X. Yu, S. Gong and J. Lai, "RGB-infrared cross-modality person re-identification," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 5390–5399, 2017.

[2] M. Ye, X. Y. Lan, J. W. Li and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. of the 32nd AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, pp. 7501–7508, 2018.

[3] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li *et al.,* "Cross-modality person re-identification with shared-specific feature transfer," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 13376–13386, 2020.

[4] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang *et al.,* "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Transactions on Image Processing*, vol. 31, pp. 2352–2364, 2022.
[5] M. Ye, Z. Wang, X. Lan and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, pp. 1092–1099, 2018.
[6] Y. B. Zhao, J. W. Lin, Q. Xuan, X. Xi *et al.,* "HPILN: A feature learning framework for cross-modality person re-identification," *IET Image Processing*, vol. 13, no. 14, pp. 2897–2904, 2020.
[7] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu *et al.,* "Hetero-center loss for cross-modality person re-identification," *Neurocomputing*, vol. 386, pp. 97–109, 2019.
[8] M. Jia, Y. Zhai, S. Lu, S. Ma and J. Zhang, "A similarity inference metric for RGB-infrared cross-modality person re-identification," arXiv preprint arXiv:2007.01504, 2000.
[9] 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 *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 618–626, 2019.
[10] S. Choi, S. Lee, Y. Kim, T. Kim and C. Kim, "Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification," in *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 10254–10263, 2020.
[11] D. Li, X. Wei, X. Hong and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *AAAI Conf. on Artificial Intelligence*, vol. 34, New York, NY, USA, pp. 4610–4617, 2020.