Journal of Computers Vol. 37 No. 4, August 2026, pp. 29-42
https://doi.org/10.63367/19911599202608370400 3

Multi Scale Feature Alignment Network for Cross-Modal Visible-Infrared Person Re-Identification

Bailiang Huang[1], Bin Chen[1], Tianran Sun[1],
Linlin Zhang[1], Zhujun Yang[1*], and Bing Zhang[2]

[1] Changchun Technical University of Automobile, Changchun 130000, Jilin Province, China
ww5171351@126.com

[2] Sergeant School of Army Academy of Armored Force, Changchun 130000, Jilin Province, China
13174312160@163.com

Received 21 June 2025; Revised 13 December 2025; Accepted 6 January 2026

Abstract. To address the performance degradation of person re-identification (ReID) under complex lighting and day-night conditions, this study proposes a novel dual-path convolution based multi-scale feature alignment (DCMFA) network. The network mainly focuses on addressing the challenges of modality discrepancy and feature alignment between visible and infrared images. First, to cope with feature deformation in persons caused by external factors in both modalities, we design a dual-path convolution (DPCon) layer. Secondly, to further alleviate feature loss caused by scale variation, we construct a multi-scale feature aggregation (MSFA) module by stacking DPCon layers of different depths and incorporating an attention mechanism to effectively aggregate key multi-scale information and suppress redundancy. Finally, to enable more effective alignment of multi-scale features across the two modalities, we propose a feature mapping and alignment operations (FMAO) module. Experimental results on three publicly available cross-modal visible-infrared person re-identification (VI-ReID) datasets demonstrate that our DCMFA network significantly outperforms existing mainstream methods in terms of recognition accuracy. Specifically, on the SYSU-MM01 dataset, our method achieves a Rank-1 accuracy of 83.74% in the single-shot setting and 89.47% in the multi-shot setting.

Keywords: machine vision, visible and infrared, deep learning, person re-identification

1 Introduction

ReID is a crucial task in the field of computer vision, with the core objective of matching and retrieving target individuals across cameras and scenes. With the widespread adoption of intelligent surveillance systems, ReID technology has shown great potential in applications such as public security and smart cities. However, traditional ReID research primarily relies on visible images. In real-world scenarios, limited lighting conditions such as nighttime or low illumination often prevent visible cameras from capturing clear person images, which severely hinders the performance of ReID systems in complex lighting environments. In contrast, infrared cameras can operate reliably under low-light conditions. Thus, achieving cross-modal ReID between visible and infrared images has become a hot and challenging research topic.

The main challenge of cross-modal VI-ReID lies in the significant modality discrepancy between visible and infrared images. Visible images contain rich color and texture information, while infrared images mainly capture thermal radiation features of the target and lack color information, resulting in a considerable distribution gap between the two modalities. Moreover, the appearance of a person in different modalities may vary due to factors such as ambient temperature and clothing material, further increasing the difficulty of cross-modal matching. Existing single-modal ReID methods often rely on color features or local details [1-2], making them difficult to transfer directly to cross-modal scenarios. Thus, more robust network models are needed to learn inter-modality differences.

To address the challenge of modality discrepancy, researchers have proposed two commonly used approaches to improve VI-ReID performance. First, to handle modality differences, some studies focus on designing more refined feature extraction networks that align features by separately extracting raw features from visible and

* Corresponding Author

infrared images [3-6]. These methods typically adopt convolution-based network models to learn shared features across modalities, enabling the extraction of unified and discriminative feature representations. However, in real-world scenarios, variations in camera installation conditions, shooting angles, human poses, and imaging distances lead to noticeable deformations in person images, accompanied by significant scale changes. Limited by the fixed structure and receptive field of conventional convolutions, these methods often fail to effectively capture shared features across the two modalities. Consequently, critical personal information can easily become compromised or lost, which affects the accuracy of cross-modal re-identification. In addition, some studies attempt to reduce modality gaps using image generation techniques by converting visible and infrared images into each other's domain [7-9]. In this way, the cross-modal task is transformed into a single-modal task, simplifying the feature alignment process. However, this approach still faces significant challenges. Due to insufficient cross-modal training data, image translation often suffers from low quality and introduces substantial noise. The presence of noise not only degrades image quality but also hinders feature extraction, resulting in poor ReID performance. Although image generation techniques show promise in reducing modality gaps, achieving effective feature alignment remains a critical and unresolved challenge.

To address the above issues and further enhance the network's sensitivity and capability in feature extraction for more effective person feature alignment, we propose the DCMFA network. In this network, we first design the DPCon layer, a basic feature extraction unit that combines deformable convolution and standard convolution. In DPCon, deformable convolution is used to extract the person appearance deformation features caused by external factors, while standard convolution focuses on extracting the local underlying structural information in the image. The combination of the two convolution types enhances the network's sensitivity to person feature information. Considering the non-linear nature of information variation in person images, we introduce a hyperbolic interpolation method in deformable convolution to improve feature extraction under deformation. Secondly, since external factors may cause changes in image scale, it is necessary to fully extract person features at multiple scales to ensure accurate subsequent feature alignment. In this regard, we construct a MSFA module based on multi-scale structure by stacking multiple DPCon modules layer by layer. By stacking varying numbers of DPCon layers, the MSFA module forms receptive fields of multiple sizes, enabling better focus on pedestrian features at different scales and enhancing feature extraction capability. In the MSFA module, we also design a special attention aggregation mechanism to enhance or suppress extracted features based on their importance, enabling more effective feature aggregation and further improving the network's sensitivity to features. Finally, based on the principle of identity and feature consistency between visible and infrared images, we design the FMAO module. The aim of this module is to map features from both modalities into a shared feature space, where more effective alignment can be performed to improve cross-modal VI-ReID accuracy.

In summary, our contributions are as follows:

To address feature and modality discrepancies in cross-modal VI-ReID, we construct the DCMFA framework to improve recognition accuracy.

To strengthen the network's capability of extracting pedestrian features and reduce deformation interference, we design the DPCon layer by combining deformable and standard convolutions, and propose a hyperbolic interpolation method to optimize deformable convolution.

To alleviate the feature loss caused by image scale variation, we construct the MSFA module. This module obtains multiple sets of different receptive fields by stacking DPCon layers with different depths, and utilizes the attention mechanism to dynamically enhance the key information to suppress the redundant information, so as to improve the robustness of feature extraction.

Based on the principle of identity and feature consistency, we design the FMAO module. It maps visible and infrared features into a unified space, and leverages multi-scale information to achieve more efficient feature alignment, thereby reducing the impact of modality discrepancy on re-identification performance.

We conduct rigorous experiments on three challenging cross-modal VI-ReID datasets. In the multi-shot test of SYSU-MM01, our method improves the Rank-1 score from 84.34% to 89.47%. In the bidirectional testing of RegDB, the Rank-1 score reaches 94.19% in both directions. In the bidirectional evaluation of LLCM, the mAP score increases by more than 1.81. The experiments demonstrate that the DCMFA framework achieves state-of-the-art performance.

2 Related Work

VI-ReID is a highly challenging cross-modal retrieval task, with its core objective being the accurate matching of people between daytime visible images and nighttime infrared images. This task not only inherits common challenges from traditional single-modal person re-identification, such as pose variation and occlusion, but also faces the core challenge of significant modality discrepancy caused by differing imaging spectral characteristics. To tackle these challenges, current research primarily focuses on person features and modality-specific image representations.

Current research on person features mainly aims to acquire more effective and discriminative representations. On the one hand, some methods focus on learning modality-invariant features by employing shared-weight networks or feature alignment mechanisms to reduce modality discrepancies [10-12]. On the other hand, person body structure information, local region alignment strategies, or auxiliary information are introduced to enhance the robustness and discriminability of feature representations [5, 13-15]. For instance, Ye et al. [3] proposed a dynamic dual-attentive aggregation learning method that enhances cross-modal alignment by mining part-level features and contextual relationships from images. Qiu et al. [4] modeled various local features of pedestrian images to construct high-order structural relationships. This method effectively enhances the representation capability of person features under cross-modal conditions. Ye et al. [16] performed image augmentation by randomly swapping color channels, thereby improving the network's robustness to color variations and reducing cross-modal discrepancies. Moreover, some approaches jointly optimize classification loss and modality adversarial loss to improve the consistency and discriminative ability of cross-modal features [6, 17]. However, these methods often neglect the impact of external factors such as appearance deformation and scale variation, which can cause the network to lose critical features and impair cross-modal re-identification accuracy. Nevertheless, these methods typically assume that person poses and scales remain relatively stable. Therefore, when deformations or scale variations caused by external factors are encountered, the network tends to lose key features, which affects the accuracy of cross-modal VI-ReID.

For handling images from two modalities, current research mainly focuses on using generative adversarial network (GAN) techniques to transform data between the two modalities [7-9, 18]. For example, Zhang et al. [19] employed a GAN to synthesize infrared pedestrian images, and introduced an infrared teacher module to guide the student model, effectively narrowing the modality gap between visible and infrared images. Cheng et al. [20] explored a cross-modal VI-ReID method based on hybrid dual-path neural networks, where intermediate modality images were generated via GANs to alleviate modality differences. Wang et al. [21] utilized a GAN to perform image modality transformation and incorporated an attention mechanism to enhance the feature extraction process. In addition to image-level transformation, Yu et al. [22] attempted to convert images from both modalities into textual descriptions, leveraging textual information to reduce modality discrepancies. However, the image generation process inevitably introduces noise, which may result in pseudo-features that are inconsistent with the original modality information, thereby affecting the stability and generalizability of downstream recognition models. Moreover, the modality transformation process may lead to detail loss or structural distortion in person images, further complicating semantic alignment between modalities.

We construct the DCMFA network to address the loss of key information caused by person deformation and scale variation, as well as the alignment challenges introduced by modality discrepancies. The network improves local information extraction adaptability through deformable convolutions. It also integrates multi-scale feature extraction with a feature mapping mechanism, achieving efficient cross-modal feature extraction and alignment. Our network provides an effective solution for improving the accuracy and generalization capability of cross-modal VI-ReID.

3 Model and Methods

In this chapter, we provide a systematic description of the overall architecture and core methods of the model. We first introduce the design principles of DPCon and explain its advantages in extracting structural and deformation features. We then describe the MSFA module constructed on top of DPCon. Finally, we present the specific operational procedure of FMAO.

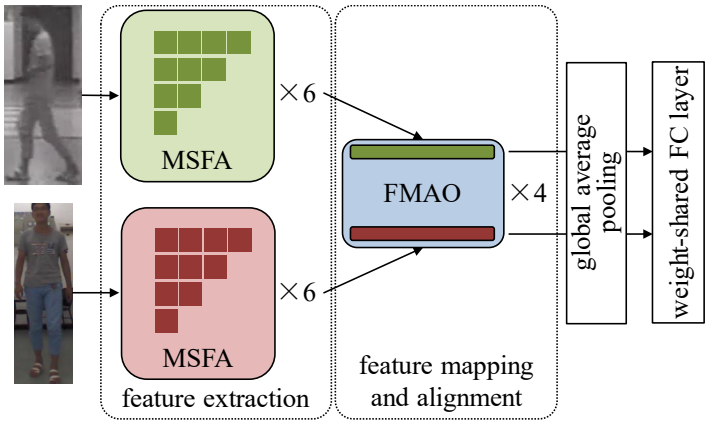

Fig. 1. Schematic diagram of the overall network structure

The overall network architecture we constructed is shown in Fig. 1. The network first employs the MSFA module based on DPCon to extract multi-scale feature information from visible and infrared images, respectively. Subsequently, based on the principle of information consistency, the FMAO module is used to map feature information from both modalities into a common feature space, where feature alignment is performed. Finally, a fully connected layer is used to classify person identities, thereby achieving the cross-modal VI-ReID task.

3.1 Dual-path Convolution

In deep learning–based VI-ReID tasks, convolutional structures typically extract locally stable structural features from person images, such as contour edges and basic textures. Such information exhibits strong consistency across modalities and serves as the fundamental basis for constructing discriminative cross-modal features. However, in real-world scenarios, factors such as pose variations, camera angle differences, and imaging distance often cause deformations or positional shifts in local body regions. Conventional convolutions with fixed sampling positions struggle to adapt to these spatial variations, resulting in insufficient capture of detailed features. To address this limitation, we introduce deformable convolutions in DPCon, enabling the convolution kernel to dynamically adjust sampling positions based on image content. This allows for more flexible focusing on critical regions significantly affected by deformation.

In DPCon, we follow the design by Dai et al. [23] and define the elements based on their coordinate positions. For the 3×3 convolution kernel, each element x can be represented as $x(p_0+p_n)$, where p_0 is the center position of the kernel, and $p_n \in R$ [(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)]. Accordingly, the computation expression for the standard convolution can be derived as *Con*:

$$Con(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{1}$$

Here, w represents the weight of the corresponding element. Based on standard convolution, an offset Δp_n is added to each element to enable deformable convolution. We follow the design of Zhu et al. [24] by introducing a modulation offset Δm_n to enhance the adaptability of deformable convolution to person images, thereby enabling more accurate focus on informative features. The deformable convolution *DCon* can be expressed as:

$$DCon(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \triangle p_n) \cdot \triangle m_n \tag{2}$$

In this context, deformable convolution predicts offsets so that the sampling locations are no longer fixed, and the new coordinates $x(p_0+p_n+\Delta p_n)\cdot\Delta m_n$ after the offset are typically floating-point values, while the feature values are defined only at integer coordinates. To address this, we employ hyperbolic interpolation to compute and

obtain feature values at non-integer positions.

To preserve the invariance of local structural features of the person, while enhancing the ability to capture fine-grained local details, we integrate *Con* and *DCon* to form a DPCon module, as illustrated in Fig. 2.

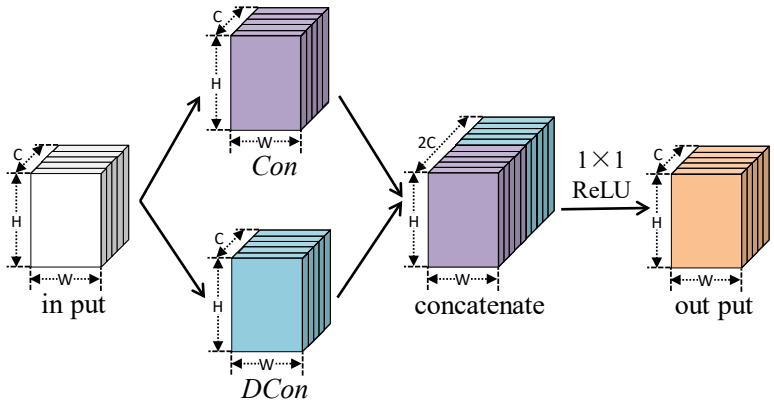

Fig. 2. The overall structure of dual-path convolution

By combining standard and deformable convolutions in parallel, we preserve the stability of local structures while enhancing adaptability to spatial variations. The result is a complementary feature representation model that integrates 'human structural information + local detail information'. We use channel concatenation to fuse the two feature streams, thereby maximally preserving their semantic integrity and avoiding information overwriting or detail loss during fusion. Subsequently, we use 1×1 convolutions with ReLU activation to reorganize the channels, enabling the network to further model inter-channel correlations within the fused feature space. This enhances the overall expressive power of the features and improves cross-modal consistency.

In DPCon, the deformable convolution extends standard convolution by providing adaptive capability to capture cross-modal discrepancy regions. It dynamically adjust sampling positions to precisely focus on discriminative features, particularly in scenarios involving pose variations, limb position shifts, and misalignment of key semantic regions. These regions often exhibit significant differences between visible and infrared modalities, serving as core factors affecting cross-modal matching performance. Therefore, introducing a dynamic sampling mechanism enhances the network's robustness to body deformation and modality gaps while strengthening the consistency of feature representations.

3.2 Hyperbolic Interpolation

Variations between pixels often exhibit nonlinear characteristics. This nonlinearity is particularly pronounced in detailed regions such as edges and textures of person images. Traditional linear interpolation methods often struggle to accurately capture the complex relationships between pixels when dealing with such nonlinear variations, resulting in loss or distortion of person image information. Especially in deformable convolution, the limitations of linear interpolation methods are more prominent because the shape and position of the convolution kernel can be dynamically adjusted according to the image content. To address this issue, we propose a hyperbolic interpolation method designed to make the sampled values closer to the true values, thereby better preserving the original features of person images. This method is inspired by a thorough observation of pixel variation characteristics in images. In real images, pixel value changes are not uniformly distributed but exhibit a distance-dependent nonlinear relationship. Specifically, in regions close to the original pixel, value changes are usually gradual and minor; as the distance increases, the magnitude of variation intensifies in a nonlinear fashion. This property closely aligns with the mathematical characteristics of the Sigmoid curve, which effectively models the intensification of changes with increasing distance. Therefore, we adopt the Sigmoid curve as the foundation to construct a hyperbolic interpolation method, aiming to better accommodate deformable convolution operations.

For four adjacent elements on the plane, we denote their values as Q_{11}, Q_{21}, Q_{12}, and Q_{22}, with corresponding coordinates (x_1,y_1), (x_2,y_1), (x_1,y_2), and (x_2,y_2), respectively. Additionally, let Q represent the interpolated value computed via hyperbolic interpolation, and (x, y) denote its coordinate. For clarity, the spatial arrangement of the five elements is illustrated in Fig. 3.

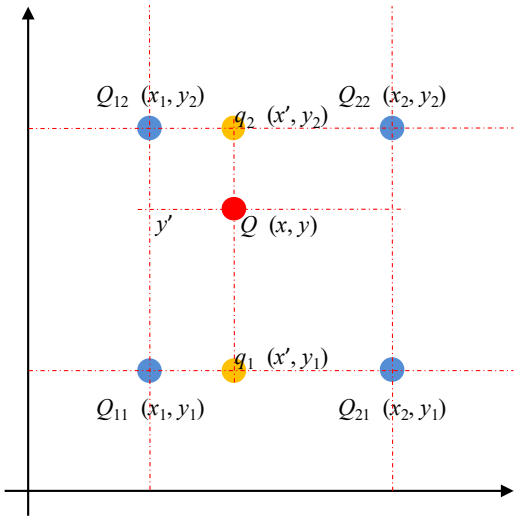

Fig. 3. Positional relationship among the elements

In the process of hyperbolic interpolation, we first normalize the coordinates of the elements to obtain the normalized coordinate (x', y') of Q.

$$x' = \frac{x - x_1}{x_2 - x_1} \quad y' = \frac{y - y_1}{y_2 - y_1} \tag{3}$$

Next, we need to define the nonlinear interpolation function S. Considering that the spacing between adjacent elements in spatial coordinates is 1 unit and the normalized coordinate values fall within the 0 to 1 range, function S must satisfy the following boundary conditions: $S(1) = 1$, $S(0.5) = 0.5$, and $S(0) = 0$. By adjusting the Sigmoid curve, we derive the following nonlinear interpolation function S that satisfies the above conditions.

$$S(t) = \begin{cases} 0 & t = 0 \\ \dfrac{1}{1 + e^{-10(t-0.5)}} & 1 > t > 1 \\ 1 & t = 1 \end{cases} \tag{4}$$

Then, we use the nonlinear interpolation function S to perform the first interpolation along the x-axis, resulting in q_1 and q_2, as calculated below:

$$q_1 = (Q_{21} - Q_{11}) \cdot S(x') + Q_{11} \tag{5}$$

$$q_2 = (Q_{22} - Q_{12}) \cdot S(x') + Q_{12} \tag{6}$$

After obtaining the results from the first interpolation, we proceed to apply function S along the y-axis for the

second interpolation, thereby obtaining the final interpolated value Q, as shown below.

$$Q = (q_2 - q_1) \cdot S(y') + q_1 \tag{7}$$

Compared to traditional linear interpolation methods, the hyperbolic interpolation approach offers significant advantages in handling image details. Linear interpolation assumes uniform variation of pixel values. However, in real images, pixel variations are nonlinear, which can lead to substantial discrepancies between the interpolated and true values. Especially, this deviation is further amplified in edge and texture regions of the image, causing blurring or distortion of the image detail information. In contrast, the hyperbolic interpolation method introduces a nonlinear weight distribution mechanism that more accurately simulate the change trend of pixel values, thereby preserving detailed information more effectively during interpolation.

3.3 Multi-scale Feature Aggregation Module Based on Dual-path Convolution

We stack multiple DPCon units to construct the MSFA module. In this module, considering the hierarchical differences in person features, we design a feature aggregation mechanism within the multi-scale feature aggregation module, which can be formulated as:

$$F(X)^i = f(X)^i \odot g[f(X)^i] \tag{8}$$

Here, X denotes the input features. $f(X)^i$ represents the feature information extracted by the i-th branch in the MSFA module. $g[f(X)^i]$ indicates the aggregation weight for the corresponding channel, obtained via a small network composed of a shared-parameter convolutional channel attention layer and a non-parametric global average pooling layer. $\odot$ denotes the Hadamard product. Then we concatenate the features extracted from each branch in sequence for aggregation. Considering the size of person images in the dataset, we set the number of branches to i=4. The overall structure is illustrated in Fig. 4. The 1×1 convolution on the input X is used to increase the number of channels, facilitating residual computation within the MSFA module. Between adjacent MSFA modules, we add depthwise convolutions, pooling layers, and other related operations to progressively compress the feature representations. With each compression step, the number of channels in the feature tensor is doubled, enhancing its representational capacity.

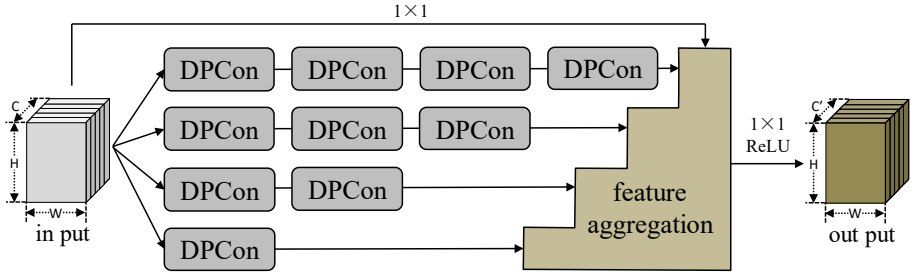

Fig. 4. Multi-scale feature aggregation module

The MSFA module progressively expands the receptive field in a hierarchical manner and aggregates person features across different scales. In this process, shallow DPCon layers focus on extracting local detail features of the person image, while deeper DPCon layers leverage a larger receptive field to capture global contextual information. This cascaded structure effectively compensates for the limitations of standard convolutions in capturing global contextual features. Moreover, the proposed aggregation mechanism efficiently fuses local detail features and global semantic information, enhancing key features while suppressing redundancy, thereby significantly improving the completeness and discriminability of person feature representation.

3.4 Feature Mapping and Alignment Operations

After six layers of MSFA-based feature extraction, the two branches respectively acquire multi-scale information of the person in visible and infrared images. Considering the principle that an individual's identity should remain consistent across different modalities, we further propose a feature-mapping and alignment method, FMAA, based on a multilayer perceptron architecture. The core idea of this module is to progressively map the multi-scale features of the visible and infrared modalities into a shared representation space through a hierarchical feature-transformation mechanism. This enables the network to eliminate modality discrepancies while preserving stable identity discrimination information. For the visible multi-scale features F_{VIS}, the input needs to be flattened to accommodate the mapping and alignment operation. However, due to the strong structural integrity of person images, we introduce structural features LOC_{VIS} during the flattening process. LOC_{VIS} is extracted from F_{VIS} using a 3×3 DPCon. After summing F_{VIS} and LOC_{VIS}, the result is flattened using FLA(), and then processed by a multilayer perceptron for mapping and alignment (MLP[]). The first FMAA operation for the visible data F_{VIS} can be expressed as:

$$F_{\mathrm{VIS\text{-}1}} = \mathrm{MLP}[\mathrm{FLA}(F_{\mathrm{VIS}} + LOC_{\mathrm{VIS}})] \tag{9}$$

The first FMAA operation for the infrared data F_{IR} can be expressed as:

$$F_{\mathrm{IR\text{-}1}} = \mathrm{MLP}[\mathrm{FLA}(F_{\mathrm{IR}} + LOC_{\mathrm{IR}})] \tag{10}$$

Where, LOC_{IR} represents the structural feature corresponding to F_{IR}. The $F_{\mathrm{VIS+1}}$ and $F_{\mathrm{IR+1}}$ denote the computed visible and infrared feature information, respectively. They are used in the next FMAA operation.

After four iterations of FMAA, the module gradually compresses the feature distribution gap between the two modalities and highlights consistent identity-related features across them. Particularly in the final mapping stage, we introduce a parameter-sharing strategy. This enables the network to handle the feature distributions of both modalities in a unified manner, fundamentally improving the consistency and discriminability of cross-modality representations. This stage-wise feature alignment mechanism enables FMAA to effectively capture stable semantic components across modalities. It mitigates interference from lighting conditions, pose variations, and viewing angles. Ultimately, it generates more discriminative feature representations.

4 Experiments

In this chapter, we first introduce the datasets and experimental settings. Then, we compare the performance with several mainstream methods and analyze the contribution of each module through ablation studies. In addition, visualized results are presented to illustrate the feature extraction capability of the model. This chapter aims to comprehensively evaluate the performance of the proposed method.

4.1 Datasets and Evaluation Metrics

To evaluate the overall performance of the DCMFA network, we conducted comprehensive comparative experiments on three cross-modal VI-ReID datasets, SYSU-MM01 [25], RegDB [26], and LLCM [5]. SYSU-MM01 was collected and organized by Wu A et al. from Sun Yat-sen University and officially released in 2017. It contains over 30,000 images of 491 identities captured by four RGB cameras and two infrared cameras. It is currently the largest and most challenging dataset in the field of cross-modal VI-ReID. RegDB was collected and curated by the U.S. Army Research Laboratory, comprising images of 412 different identities. Each identity has 10 RGB images and 10 thermal infrared images, with 254 females and 158 males. In terms of capture angles, 156 subjects were shot from the front and 256 from the back. The dataset has fewer images with lower resolution, and each identity's RGB and infrared images are one-to-one aligned in pose with minimal variation, making the re-identification task relatively easy. LLCM was collected and compiled by Zhang et al. from the School of Informatics Xiamen University and was released in 2023. The dataset is divided into training and testing sets in a roughly 2:1 ratio, and the two are mutually exclusive. The training set contains 30,921 images from 713

identities, including 16,946 visible images and 13,975 infrared images. The test set comprises 13,909 images of 351 identities.

To evaluate and compare the performance of the network models, we adopt cumulative matching characteristic curve (Rank-1) and mean average precision (mAP) as evaluation metrics. The final results are reported as the average of 10 independent test runs.

4.2 Experimental Setup

The overall network build is done using python 3.7 in the pytorch 1.8.0 environment. The hardware platform primarily consists of a workstation equipped with two NVIDIA GEFORCE RTX 3090 GPUs. At the early stage of training, we pre-train the feature extraction branch on ImageNet to enhance the model's generalization and feature representation capabilities. For the input image data, we uniformly resize them to 288×144. Data augmentation techniques include random erasing, random cropping, and random flipping. The batch size is set to 120, consisting of 10 different identities with 12 randomly selected images per identity. We use ID loss [27] and Similarity Distribution Matching loss [28] to train the network model. We use the Adam optimizer with an initial learning rate of 4×10^{-4} and train the model for 180 epochs. The learning rate decays by a factor of 0.1 at the 80th, 120th, and 150th epochs. After the 150th epoch, we freeze the parameters of the feature extraction branch.

4.3 Comparison with State-of-the-art Methods

To validate the advancement of the DCMFA network model, we select several representative works proposed in recent years for comparison. These include DDAG [3] and HOS-Net [4], which employ feature fusion techniques. DEEN [5], IDKL [6], and MUN [17] improve re-identification accuracy by mining deep discriminative information from each modality. CSDN [22] achieves cross-modal information alignment through modality transformation. CAJ [16] enhances network robustness by exchanging image channel data, while SDCL [15] employs unsupervised learning techniques. The comparative data are summarized in Table 1 and Table 2.

Table 1. Comparison of performance on the SYSU-MM01 and RegDB datasets

Methods	SYSU-MM01				RegDB			
	Single-shot		Multi-shot		VIS to IR		IR to VIS	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
DDAG	54.75	53.02	-	-	69.34	63.46	68.06	61.80
HOS-Net	75.6	74.2			94.7	90.4	93.3	89.2
DEEN	74.7	71.8	-	-	91.1	85.1	89.5	83.4
IDKL	81.42	79.85	84.34	78.22	94.72	90.19	94.22	90.43
MUN	76.24	73.81	-	-	95.19	87.15	91.86	85.01
CSDN	75.2	71.8	80.6	66.3	89.0	84.7	88.2	82.8
CAJ	69.88	66.89	-	-	85.03	77.82	84.75	77.82
SDCL	64.49	63.24	-	-	86.91	78.92	85.76	77.25
DCMFA	83.74	80.93	89.47	80.64	95.33	89.87	94.19	90.57

Table 2. Comparison of performance on the LLCM dataset

Methods	LLCM			
	VIS to IR		IR to VIS	
	Rank-1	mAP	Rank-1	mAP
DDAG	40.3	48.4	48.0	52.3
HOS-Net	64.9	67.9	56.4	63.2
DEEN	62.5	65.8	54.9	62.9
IDKL	72.22	66.43	70.72	65.19
MUN	-	-	-	-
CSDN	-	-	-	-
CAJ	56.5	59.8	48.8	56.6
SDCL	-	-	-	-
DCMFA	76.63	69.71	69.98	67.46

The comparative data in the table indicate that the DCMFA network achieves state-of-the-art overall performance. In the multi-shot evaluation on the large-scale SYSU-MM01 dataset, our method improves the Rank-1 accuracy from 84.34% to 89.47% compared with the high-performing IDKL. Meanwhile, in both single-shot and multi-shot settings, the mAP score also increases by more than 3.08. This is because the IDKL enhances cross-modality feature consistency through implicit discriminative knowledge learning. Its core idea is to mine deep discriminative relations in the feature space, to reduce the semantic gap between visible and infrared modalities. As a result, IDKL focuses more on global discriminative information while paying insufficient attention to feature instability caused by pose changes, local deformations, and scale variations. In contrast, DCMFA incorporates the DPCon-based MSFA at the feature extraction stage. It adapts to local deformations using deformable convolution and achieves cross-modality feature alignment by processing multi-scale information with FMAA. Therefore, DCMFA exhibits stronger feature robustness and alignment capability on large-scale datasets, where substantial appearance differences and unstable modality characteristics are present. On the large-scale LLCM dataset, the mAP scores in both visible-to-infrared and infrared-to-visible tests increase by more than 1.81. This is primarily due to the considerable differences in shooting distance and viewpoints in the LLCM dataset, resulting in more complex variations in image scale and appearance. Meanwhile, the multi-scale architecture of DCMFA can stably extract discriminative features under different receptive fields, thereby effectively adapting to these variations and achieving performance gains.

To further validate the effectiveness of the DCMFA network, we perform activation visualization on two sets of re-identification images from the LLCM dataset, as shown in Fig. 5.

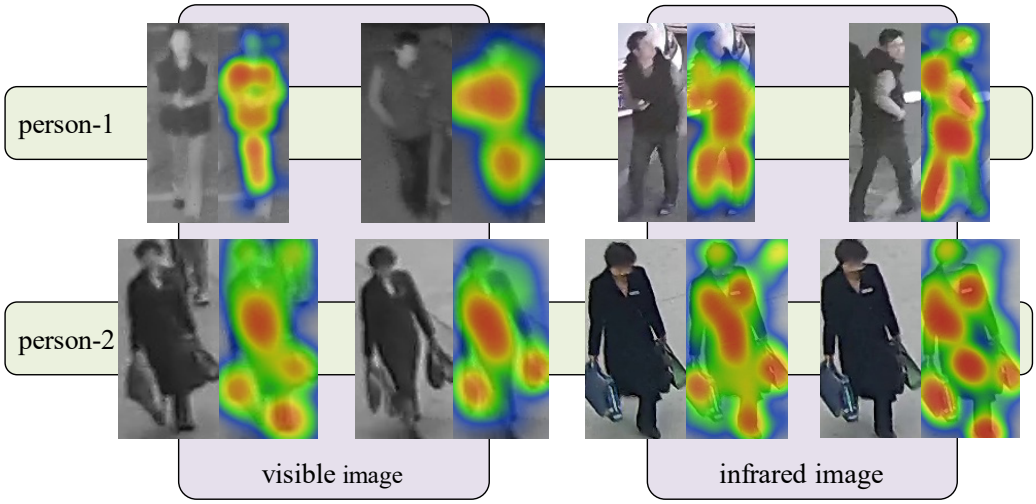

Fig. 5. Activation visualization of person images

As shown in the activation maps, the DCMFA network effectively extracts person features from both visible and infrared images, achieving cross-modal feature alignment. For instance, in the case of person-2, the network is able to extract the distinguishing feature of a handbag even under low resolution and occlusion conditions. This is attributed to the use of DPCon as the basic feature extraction unit in DCMFA, which alleviates geometric distortions caused by viewpoint changes and maintains the ability to capture fine-grained details. In addition, by introducing a multi-scale structure, DCMFA can focus on both local details and global semantic information of pedestrians at different spatial scales. Thanks to its multi-level receptive field design, the network can capture fine texture variations in small regions as well as the overall layout of large structures. Consequently, it maintains stable feature representation even under complex pose changes, viewpoint shifts, and scale fluctuations. This effectively prevents discriminative features from being weakened or overlooked during cross-modal matching.

4.4 Ablation Study

To demonstrate the effectiveness of the overall network structure, we conduct five ablation studies on the SYSU-

MM01 dataset, and the results are shown in Table 3. Setting-1 corresponds to the original structure of the DCMFA network. In Setting-2, we replace hyperbolic interpolation with conventional bilinear interpolation, resulting in a slight decrease in re-identification accuracy. Because bilinear interpolation tends to cause loss of detail and edge blurring during feature extraction, while hyperbolic interpolation better preserves spatial structure and fine details, which are crucial for capturing subtle differences in Re-ID. In Setting-3, replacing DPCon with standard convolution results in a 10.87% drop in Rank-1 accuracy under multi-shot testing. Because DPCon captures geometric deformation features more effectively than standard convolution, and its deformable sampling mechanism enhances the model's adaptability to complex spatial transformations. In Settings 4 and 5, we replace MSFA and FMAA respectively, and the results show significant decreases in re-identification accuracy. Because the multi-scale structure enhances the network's sensitivity to feature information and prevents the loss of discriminative person features. After replacing MSFA, the network's ability to perceive feature information is compromised, resulting in an mAP drop of more than 6.51 under both shots. FMAA effectively aligns feature representations from different modalities, facilitating cross-modal VI-ReID. Replacing FMAA therefore results in a decrease in multi-view Rank-1 accuracy, from 89.47% to 62.56%.

Table 3. Structural ablation experiments of DCMFA

Settings	Hyperbolic interpolation	DPCon	MSFA	FMAA	Single-shot		Multi-shot	
					Rank-1	mAP	Rank-1	mAP
1		√	√	√	83.74	80.93	89.47	80.64
2	√	√	√	√	80.12	80.45	83.78	78.03
3			√	√	79.77	76.19	78.6	76.84
4		√		√	68.5	74.42	73.34	71.91
5		√	√		67.23	66.89	62.56	62.67

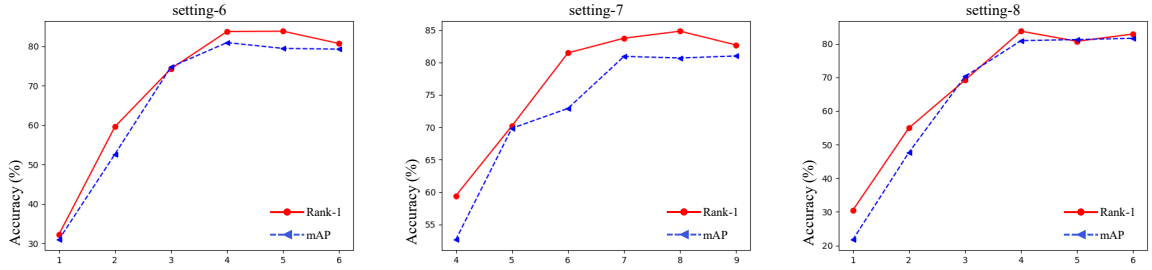

Fig. 6. Curve of ablation experiment results

Furthermore, we conduct three groups of ablation studies on MSFA and FMAA in the single-shot mode of the SYSU-MM01 dataset, and the results are shown in Fig. 6. In setting-6, we modify the number of channels in MSFA. The results show that re-identification accuracy is low when the channel number is insufficient. This is because insufficient channel dimensionality weakens the representation capacity of multi-scale features. Conversely, too many channels introduce redundant information, reduce the effectiveness of feature aggregation, and cause the re-identification accuracy to decline slightly after reaching its peak. Therefore, we set the number of MSFA channels to 4 to achieve optimal performance. In setting-7, we modify the computational layers of MSFA. The results show that gradually increasing the number of layers improves re-identification accuracy. But performance begins to deteriorate once a certain depth is exceeded. This is because insufficient layers restrict the expansion of the multi-scale structure, limiting the receptive field and hindering effective fusion of global and local features. Conversely, too many layers cause excessive smoothing and redundant accumulation during progressive aggregation. This weakens fine-grained differences and harms discriminative ability. To obtain the best multi-scale representation, we adopt seven stacked layers for MSFA. In setting-8, we modify the number of computational layers of FMAA. The results show that gradually increasing the computational layers of FMAA improves re-identification performance. However, the performance drops slightly after reaching the peak. Because too few layers lead to insufficient feature mapping and alignment across modalities, which prevents the

network from effectively learning the difference in distribution between visible and infrared features. Conversely, too many layers can introduce noise into the mapping space, causing the network to overfit the intra-modal distribution and reducing its ability to learn cross-modality consistency. Considering the final performance, we set the number of FMAA layers to 4.

5 Conclusion and Future Work

In this study, we propose the DCMFA network framework for cross-modal Re-ID based on visible and infrared images. The DCMFA framework mainly consists of two modules, the MSFA module and the FMAO module. The MSFA module is composed of multiple stacked DPCon units to extract person features from both modalities. The FMAO module maps features from both modalities into a shared feature space to achieve cross-modal feature alignment. In DPCon, we leverage the ability of standard convolution to capture local structural information and the adaptability of deformable convolution to geometric deformations. By combining standard and deformable convolutions, the feature extraction unit achieves enhanced sensitivity to person features. In addition, to better adapt deformable convolution to cross-modal VI-ReID, we introduce a hyperbolic interpolation method that addresses the nonlinear nature of image information variation. To validate the effectiveness of the DCMFA framework, we conduct extensive experiments on three datasets, SYSU-MM01, RegDB, and LLCM. The results show that DCMFA achieves state-of-the-art performance, and the ablation studies confirm the effectiveness of each module and method.

In future work, we continue to investigate the potential of deformable convolutions in cross-modal re-identification tasks. First, we plan to incorporate learnable parameters into the hyperbolic interpolation method so that the interpolation process can dynamically adapt to image content. This enables a more accurate modeling of the nonlinear appearance variations of pedestrians across different modalities. It also improves the ability of deformable convolutions to capture deformation-sensitive regions. Moreover, based on the performance of DCMFA in cross-modal VI-ReID, we also intend to apply this method to more intricate real-world surveillance scenarios, such as ReID at night or under low-illumination conditions. We expect to improve the model's recognition stability and generalization ability under low light, occlusion, and severe viewpoint variations by combining low-light enhancement strategies with multi-scale feature representation. Overall, future studies will focus on enhancing the model's adaptability and robustness in complex scenarios and promoting the use of cross-modal ReID in real-world surveillance environments.

This research was funded by the Science and Technology Research Program of the Education Department of Jilin Province (No. JJKH20251913KJ), and Changchun Technical University of Aautomobile.

Competing Interests Statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, Y. Shan, Dual cross-attention learning for fine-grained visual categorization and object re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. https://doi.org/10.1109/CVPR52688.2022.00465

[2] Y. Zhai, Y. Zeng, Z. Huang, Z. Qin, X. Jin, D. Cao, Multi-prompts learning with cross-modal alignment for attribute-based person re-identification, in: Proc. AAAI Conference on Artificial Intelligence, 2024. https://doi.org/10.1609/aaai.v38i7.28524

[3] M. Ye, J. Shen, D.J. Crandall, L. Shao, J. Luo, Dynamic dual-attentive aggregation learning for visible-infrared person re-identification, in: Proc. European Conference on Computer Vision, 2020. https://doi.org/10.1007/978-3-030-58520-4_14

[4] L. Qiu, S. Chen, Y. Yan, J.H. Xue, D.H. Wang, S. Zhu, High-order structure based middle-feature learning for visible-infrared person re-identification, in: Proc. AAAI Conference on Artificial Intelligence, 2024. https://doi.org/10.1609/aaai.v38i5.28259

[5] Y. Zhang, H. Wang, Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. https://doi.org/10.1109/CVPR52729.2023.00214

[6] K. Ren, L. Zhang, Implicit discriminative knowledge learning for visible-infrared person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
https://doi.org/10.1109/CVPR52733.2024.00045
[7] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, S.-I. Satoh, Learning to reduce dual-level discrepancy for infrared-visible person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
https://doi.org/10.1109/CVPR.2019.00071
[8] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
https://doi.org/10.1109/CVPR42600.2020.01027
[9] G.A. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, X. Liang, Z.-G. Hou, Cross-modality paired-images generation for RGB-infrared person re-identification, in: Proc. AAAI Conference on Artificial Intelligence, 2020.
https://doi.org/10.1609/aaai.v34i07.6894
[10] K. Kansal, A.V. Subramanyam, Z. Wang, S.-I. Satoh, SDL: Spectrum-disentangled representation learning for visible-infrared person re-identification, IEEE Transactions on Circuits and Systems for Video Technology 30(10)(2020) 3422–3432.
https://doi.org/10.1109/TCSVT.2019.2963721
[11] J. Feng, A. Wu, W.-S. Zheng, Shape-erased feature learning for visible-infrared person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
https://doi.org/10.1109/CVPR52729.2023.02179
[12] P. Zhang, Y. Wang, Y. Liu, Z. Tu, H. Lu, Magic tokens: Select diverse tokens for multi-modal object re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
https://doi.org/10.1109/CVPR52733.2024.01620
[13] M. Ye, C. Chen, J. Shen, L. Shao, Dynamic tri-level relation mining with attentive graph for visible infrared re-identification, IEEE Transactions on Information Forensics and Security 17(2021) 386–398.
https://doi.org/10.1109/TIFS.2021.3139224
[14] A. Zheng, P. Pan, H. Li, C. Li, B. Luo, C. Tan, R. Jia, Progressive attribute embedding for accurate cross-modality person re-id, in: Proc. 30th ACM International Conference on Multimedia, 2022.
https://doi.org/10.1145/3503161.3548336
[15] B. Yang, J. Chen, M. Ye, Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.
https://doi.org/10.1109/CVPR52733.2024.01596
[16] M. Ye, W. Ruan, B. Du, M.Z. Shou, Channel augmented joint learning for visible-infrared recognition, in: Proc. IEEE/CVF International Conference on Computer Vision, 2021.
https://doi.org/10.1109/ICCV48922.2021.01331
[17] H. Yu, X. Cheng, W. Peng, W. Liu, G. Zhao, Modality unifying network for visible-infrared person re-identification, in: Proc. IEEE/CVF International Conference on Computer Vision, 2023.
https://doi.org/10.1109/ICCV51070.2023.01027
[18] G.A. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, RGB-infrared cross-modality person re-identification via joint pixel and feature alignment, in: Proc. IEEE/CVF International Conference on Computer Vision, 2019.
https://doi.org/10.1109/ICCV.2019.00372
[19] Z. Zhang, S. Jiang, C. Huang, Y. Li, R.Y.D. Xu, RGB-IR cross-modality person ReID based on teacher-student GAN model, Pattern Recognition Letters 150(2021) 155–161.
https://doi.org/10.1016/j.patrec.2021.07.006
[20] D. Cheng, Y. Hao, J.Y. Zhou, N.N. Wang, X.B. Gao, Cross-modality person re-identification utilizing the hybrid two-stream neural networks, Journal of Xidian University 48(5)(2021) 190–200.
https://doi.org/10.19665/j.issn1001-2400.2021.05.022
[21] Y. Wang, Multi-granularity modality specific attention algorithm for visible-infrared person re-identification, Computer Science and Application 13(1)(2023) 83–92.
https://doi.org/10.12677/CSA.2023.131009
[22] X. Yu, N. Dong, L. Zhu, H. Peng, D. Tao, Clip-driven semantic discovery network for visible-infrared person re-identification, IEEE Transactions on Multimedia 27(2025) 4137-4150.
https://doi.org/10.1109/TMM.2025.3535353
[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proc. IEEE International Conference on Computer Vision, 2017.
https://doi.org/10.1109/ICCV.2017.89
[24] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
https://doi.org/10.1109/CVPR.2019.00953
[25] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, RGB-infrared cross-modality person re-identification, in: Proc. IEEE International Conference on Computer Vision, 2017.
https://doi.org/10.1109/ICCV.2017.575

[26] D.T. Nguyen, H.-G. Hong, K.-W. Kim, K.-R. Park, Person recognition system based on a combination of body images from visible light and thermal cameras, Sensors 17(3)(2017) 605. https://doi.org/10.3390/s17030605

[27] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, ACM Transactions on Multimedia Computing, Communications, and Applications 16(2)(2020) 1–23. https://doi.org/10.1145/3383184

[28] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. https://doi.org/10.1109/CVPR52729.2023.00273