Attention Mechanism and Feature Enhancement for Visible-Infrared Person Re-Identification

Junli Feng[1,a,*], Jingni Ma[1,b]

[1]School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, 400074, China
[a]3161323382@qq.com, [b]4957232722@qq.com
**Corresponding author*

Abstract: *Visible-Infrared Person Re-identification serves as a core technology in surveillance systems, enabling accurate identification of individuals across different times and locations while breaking through the constraints of lighting conditions. In contrast, traditional methods exhibit poor performance in low-light environments, making it difficult to support the advancement of relevant research. To address the inter-modal and intra-modal differences between infrared and visible light modalities, this paper proposes an Attention and Feature Enhancement Network (AFEN). The network incorporates a median-enhanced spatial-channel attention module, which can effectively capture multi-scale features. The designed feature enhancement module is capable of reducing the distribution gap between modal features, enhancing the discriminative power, robustness, and generalization ability of features, thereby improving the accuracy of cross-modal matching.*

Keywords: *Person Re-identification, Attention Mechanism, Convolutional Neural Network, Feature Learning*

1. Introduction

In the field of pedestrian re-identification (ReID), the core objective is to accurately match pedestrian images captured by different cameras. Most of the current ReID methods focus primarily on RGB images captured by visible light cameras during the day. However, in complex scenarios such as nighttime or low-light conditions, visible light cameras often struggle to effectively capture pedestrian image information, leading to a significant drop in the performance of these methods. To address this, the academic community has proposed Visible-Infrared Pedestrian Re-identification (VIReID) methods, which aim to perform pedestrian retrieval based on infrared (visible) images in the corresponding visible (infrared) images. Compared to traditional pedestrian ReID tasks, VIReID faces a more challenging problem, as there exists a significant cross-modal discrepancy between visible and infrared images.

Currently, there are two main approaches to address the cross-modal discrepancy issue. One approach is feature-level methods, which focus on mapping both visible and infrared features into a shared embedding space in order to minimize the modality gap within that space. However, due to the significant differences between visible and infrared images, directly projecting cross-modal images into the same feature space presents substantial challenges. The other approach is image-level methods, which mainly leverage Generative Adversarial Networks (GANs) to convert infrared (or visible) images into the corresponding visible (or infrared) images, thereby eliminating the modality discrepancy. Although such methods can reduce the modality gap to some extent, the generated cross-modal images often suffer from noise issues due to the lack of large-scale paired visible-infrared image data. However, due to an excessive focus on extracting modality-shared features and reducing modality discrepancies, the aforementioned methods inadequately explore fine-grained details, resulting in difficulty extracting discriminative pedestrian features.

To address the aforementioned challenges, this paper proposes a cross-modal person re-identification framework that leverages attention mechanisms and feature enhancement techniques. The method integrates a Median-Enhanced Spatial–Channel Attention module (MECS) and a Feature Enhancement Module (FEM) to achieve discriminative and robust cross-modal feature extraction. Specifically, the MECS module first exploits fine-grained pedestrian details, thereby strengthening the discriminability of the extracted pedestrian representations. Subsequently, the FEM preserves both high- and low-frequency components within the images, capturing richer cross-modal feature representations while accentuating object boundaries. This dual-frequency preservation enhances feature distinctiveness and robustness against modality discrepancies.

2. Related Works

The primary challenge in visible-infrared (VI) cross-modality person re-identification lies in the significant modality discrepancy between the two image types. Visible images consist of three color channels (red, green, and blue), whereas infrared (IR) images contain only a single channel, and the two are generated based on fundamentally different wavelength spectra. This inherent heterogeneity leads to considerable appearance differences across modalities. To address this issue, most existing approaches aim to reduce the modality gap and learn modality-invariant representations. Currently, convolutional neural network (CNN)-based[1] dual-stream architectures serve as the mainstream backbone, wherein modality-specific features are independently extracted from visible and infrared images, followed by a modality-shared layer with shared weights to align the extracted features and mitigate cross-modality discrepancies. Building upon this foundation, various strategies have been proposed to further bridge the modality gap and enhance retrieval accuracy for cross-modality person re-identification, which can generally be categorized into image-level and feature-level methods. In recent years, Transformer-based[2] approaches have also attracted increasing attention due to their powerful representation capabilities and have demonstrated promising performance in this field.

Image-level approaches aim to unify modalities at the image space by performing modality transformation or fusion to reduce modality discrepancies. Modality transformation methods convert one modality into another to bridge the gap, while modality fusion methods coordinate the relationship between modalities at the pixel level to generate new fused images. For instance, the D2RL framework[3] employs an image-level subnetwork for modality translation, but suffers from high computational cost and the introduction of noise in the generated images. In contrast, the HAT method[4] captures structural information by generating auxiliary grayscale images, avoiding the need for complex image generation processes. The tri-modal learning framework proposed in[5] adopts a lightweight self-supervised network to generate X-modality images, effectively mitigating pathological generation issues. The channel-enhanced joint learning strategy in[6] enhances robustness and reduces overhead by performing color channel exchange and random grayscale transformations. Similarly, the SMCL model[7] promotes feature sharing by generating assimilated modalities, thereby improving performance. However, image-level methods often unify modalities at a coarse pixel level, making them sensitive to noise and prone to introducing new artifacts during the transformation process.

Feature-level methods aim to achieve modality alignment and transformation through strategies such as feature extraction and enhancement, architectural innovations, and contrastive learning. For example, the DDAG approach[8] and MPANet[9] improve performance by enhancing feature representation and mitigating modality discrepancies, respectively. PSFLNet[10] introduces a novel architecture with parameter sharing to integrate modality information from the early stages of feature extraction. Although these methods effectively improve feature discriminability and modality robustness, their performance may still be limited under complex conditions due to the inherent physical differences in imaging between modalities. Recently, contrastive learning-based approaches have emerged as promising alternatives. For instance, [11] enhances modality adaptation and generalization by incorporating modality-aware learning and centroid-based negative sampling, which significantly narrows the modality gap and boosts model performance in challenging scenarios involving illumination variation, occlusion, and viewpoint changes.

Transformer-based cross-modality person re-identification methods leverage the global attention mechanism to extract pedestrian image features and capture complex relationships between different modalities, thereby enhancing model stability under challenging conditions such as occlusion, varying viewpoints, and illumination changes. The self-attention layers in Transformers dynamically adjust their weights according to the input data, enabling flexible adaptation to modality differences. For instance, [12] introduces a modality embedding module along with a modality-aware enhancement loss to learn modality-invariant representations, while[13] employs grayscale images as an auxiliary modality and adopts a progressive learning strategy to reduce modality discrepancies. Both approaches improve the discriminability and robustness of cross-modality features. However, despite the superior performance of Vision Transformers over CNNs in single-modality person re-identification tasks, their effectiveness in cross-modality settings remains limited. This is mainly due to their weaker capability in capturing fine-grained local features and their reliance on large-scale labeled data, which often leads to inferior performance compared to contemporary CNN-based methods in cross-modality scenarios.

3. Proposed method

As shown in the figure1, the proposed Attention Feature Enhancement Network (AFEN) employs a dual-stream ResNet-50 network as the backbone. The AFEN network incorporates the Median Enhanced Spatial Channel Attention Mechanism (MECS), which effectively enhances feature extraction capabilities. The MECS module combines both channel attention and spatial attention mechanisms, enabling the network to capture and integrate features at different scales. Additionally, a Feature Enhancement Module (FEM) is introduced to effectively address the modality discrepancy between visible and infrared images, enhancing the discriminative power, robustness, and generalization ability of the features. This, in turn, significantly improves the performance and reliability of the re-identification system. During the training phase, all features before and after the batch normalization (BN) layers are input into different loss functions to jointly optimize the AFEN network.

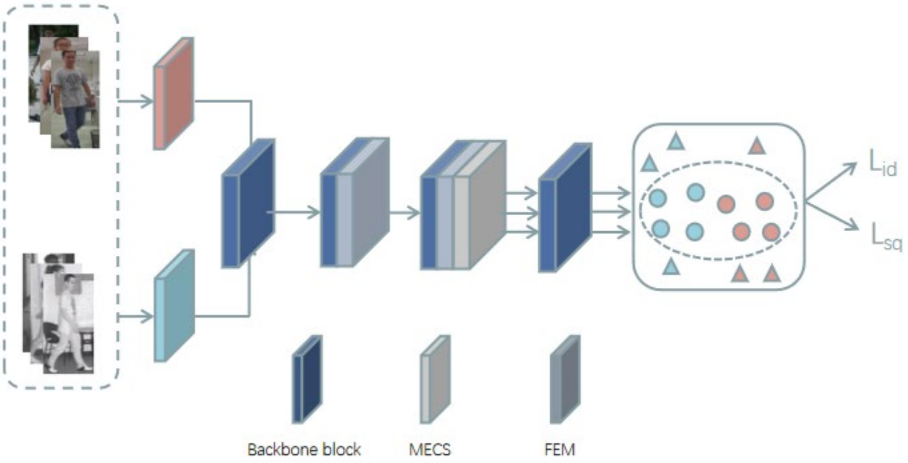

Figure 1: Network architecture.

3.1 Median-Enhanced Spatial and Channel Attention Mechanism

This paper designs a Median-Enhanced Spatial and Channel Attention Module (MECS), which combines both channel attention and spatial attention mechanisms with the goal of improving feature extraction effectiveness and robustness. The channel attention mechanism extracts global statistical information through global pooling operations, while the spatial attention mechanism captures spatial features at different scales through multi-scale deep convolution. The overall design aims to provide richer feature representations, thereby enhancing the model's performance. The structure of this MECS module is shown in Figure 2.

3.1.1 Median-Enhanced Channel Attention

The channel attention module optimizes the channel relationships of features by selecting more meaningful channels in the RGB-IR feature maps. Existing channel attention mechanisms typically use global average pooling and global max pooling to extract global statistical information from feature maps. However, these methods perform inadequately when dealing with noise, especially when significant noise is present in the input feature maps, which may affect the quality of feature extraction. Median pooling is widely used in image processing tasks for noise removal because it can eliminate noise while preserving important feature information. To address the noise issue and enhance the robustness of the channel attention mechanism, we introduce a median pooling operation into the channel attention mechanism, combining it with global average pooling and global max pooling to form a more robust channel attention mechanism. The specific process is as follows:

First, the input feature map undergoes global average pooling (AvgPool), global max pooling (MaxPool), and global median pooling (MedianPool), resulting in three different pooling outputs. The size of each pooled output is $R^{C\times1\times1}$, where C is the number of channels. Each pooling output is then passed through a shared multi-layer perceptron (MLP), which consists of two 1×1 convolutional layers and a ReLU activation function. The first convolutional layer reduces the feature dimension from C to C/r, where r is the reduction ratio, and the second convolutional layer restores the feature dimension back to C. Finally, a Sigmoid activation function is applied to compress the output values within the range of [0, 1], producing three attention maps. The attention maps from the three pooling outputs are then element-wise summed to obtain the final channel attention map. The channel attention map is then element-wise multiplied with the original input feature map to obtain the weighted feature map. The formula is as follows:

$$F_c = \sigma\left(\mathrm{MLP}\big(\mathrm{AvgPool}(F)\big)\right) + \sigma\left(\mathrm{MLP}\big(\mathrm{MaxPool}(F)\big)\right) + \sigma\left(\mathrm{MLP}\big(\mathrm{MedianPool}(F)\big)\right) \tag{1}$$

$$F' = F_c \odot F \tag{2}$$

Here, σ denotes the Sigmoid function, and ⊙ represents element-wise multiplication.

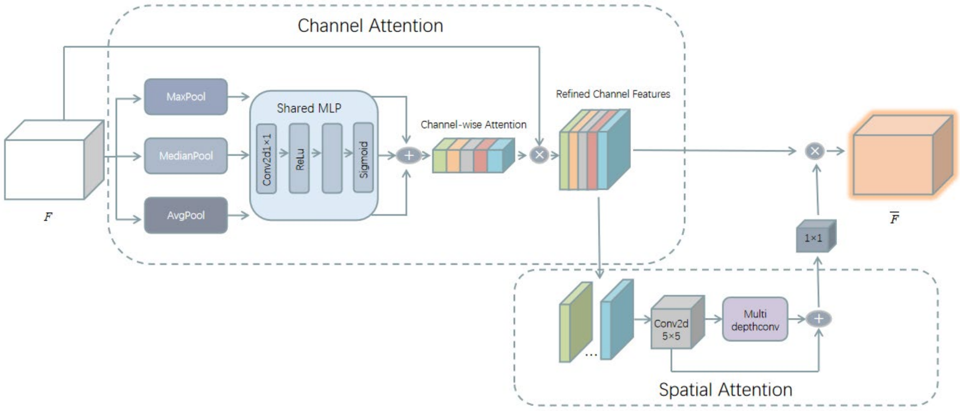

Figure 2: MECS Module.

3.1.2 Spatial Attention

To capture the spatial relationships of features, this paper further adopts a spatial attention module to emphasize feature information, serving as complementary information to the channel attention. First, the input feature map passes through a 5×5 convolutional layer to extract basic features. These basic feature maps are then processed through multiple depthwise convolution layers of varying sizes, including different kernel sizes, to further extract multi-scale features. Finally, these multi-scale features are element-wise summed, and a 1×1 convolutional layer is applied to generate the spatial attention map. The weighted feature map is then element-wise multiplied with the spatial attention map to obtain the final output feature map. The formula is as follows:

$$F_s = \sum_{i=1}^{n} D_i(F') \tag{3}$$

$$F'' = \mathrm{Conv1{\times}1}(F_s) \odot F' \tag{4}$$

Here, n denotes the number of depthwise convolutions, and Conv1×1 represents the 1×1 convolution operation.

3.2 Feature Enhancement Module

This paper proposes a Feature Enhancement Module (FEM), which consists of three key components: Convolutional Embedding (CE), Feature Refinement Module (FRM), and Convolutional Multi-Layer Perceptron (ConvMLP). The focus of this module is to adaptively capture the rich features of cluttered backgrounds, making the object boundaries more distinguishable, and perform feature enhancement to preserve both high-frequency and low-frequency components in the image.

First, the input features are processed through LayerNorm and Convolutional Embedding (CE) to learn generalization and discriminative abilities. The output of CE is passed to a 1x1 convolutional layer, which compresses the channels by half. Channel compression helps reduce computational overhead and encourages the model to mix features based on their shape. The compressed features are then fed into the Feature Refinement Module (FRM) to refine the features. The output of FRM is fused and projected through a 1x1 convolutional layer and ConvMLP to further enhance the representation.

The Feature Refinement Module (FRM) serves as a critical component in our network for enhancing high-frequency details and capturing low-frequency contextual information. Let the input feature map be denoted as $F \in R^{C\times H\times W}$. We first pass F through a deep convolutional layer to obtain a downsampled feature map $P \in R^{C \times H/2 \times W/2}$, simulating a blurred version of the original input. This feature map P is then upsampled to match the spatial resolution of F, resulting in a smoothed feature map Q.

To highlight high-frequency details, we compute the difference between F and Q, yielding a refined feature embedding R. Meanwhile, the FRM incorporates a second branch designed to capture low-frequency components. Specifically, an element-wise multiplication between F and Q is performed to obtain the low-frequency feature representation S.

Next, the high-frequency component R and the low-frequency component S are concatenated along the channel dimension and further processed by a depthwise convolution to extract a fused representation T. Finally, this output is passed through a projection layer to restore the original channel dimension, producing the enhanced feature map $\tilde{F}$.

3.3 Loss Function

This paper adopts a combined optimization model of identity loss L_{id} and enhanced weighted regularized triplet loss L_{sq}. The identity loss L_{id} constrains the gap between image representations of the same person across different scenarios, typically using the cross-entropy function, as shown below:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N} log\left(P\left(l_i \mid C(f_i \mid \theta)\right)\right) \tag{5}$$

In the equation, N represents the number of images in the current batch, l_i denotes the corresponding label value of the feature f_i , and θ represents the parameters of the classifier.

The weighted regularized triplet loss L_{wrt} combines the triplet loss function and a regularization term. By weighting the distances between different samples, the model focuses more on hard samples during training, thereby improving accuracy and generalization ability. Based on this, the optimized squared error is used instead of the l_1 norm difference, which better optimizes the model's efficiency. This is the enhanced weighted regularized triplet loss L_{sq}, as shown below:

$$L_{sq} = \frac{1}{N}\sum_{i=1}^{N} log(1 + exp(\varphi[u_i])) \tag{6}$$

$$u_i = \sum_{ij} w_{ij}^{p} d_{ij}^{p} - \sum_{ik} w_{ik}^{n} d_{ik}^{n} \tag{7}$$

$$\varphi[u_i] = \begin{cases} u_i^2, & u_i > 0 \\ -u_i^2, & u_i > 0 \end{cases} \tag{8}$$

In the equation, (i, j, k) represents a triplet in each training batch, where x_i and p_i correspond to the positive pair, and n_i corresponds to the negative pair, d_{ij}^{p}/ d_{ik}^{n} represents the Euclidean distance between the positive/negative sample pairs.

The total loss function is represented as follows:

$$L = L_{id} + L_{sq} \tag{9}$$

4. Experimental Analysis

4.1 Dataset

The SYSU-MM01 dataset is a large-scale dataset collected using four visible-light cameras and two near-infrared (NIR) cameras, covering both indoor and outdoor environments. It contains images captured under varying camera views, environmental conditions, illumination, and modalities. The training set includes 22,258 RGB images and 11,909 IR images from 395 identities. The query and gallery sets consist of 3,803 IR images and 301 (or 3,010) RGB images randomly sampled from 96 identities under single-shot or multi-shot settings. Specifically, camera 1, 2, 4, and 5 capture RGB images, while camera 3 and 6 capture IR images.

The RegDB dataset is constructed using a pair of aligned cameras (one visible-light camera and one thermal camera). It contains 8,240 images corresponding to 412 identities, with each identity having 10 images captured by the visible camera and 10 images captured by the thermal camera. For training and testing, the dataset is randomly split into two subsets: images of 206 identities are used for training, and the remaining 206 identities are used for testing.

4.2 Evaluation Metrics and Experimental Environment

In this experiment, the model was implemented on the Windows 10 operating system using Python 3.8 and the PyTorch deep learning framework. A NVIDIA RTX 3090 GPU with 24 GB of memory was employed for training and inference. For feature extraction, the ResNet-50 backbone pretrained on ImageNet[14] was adopted. Common data augmentation techniques, including random cropping, horizontal flipping, and channel enhancement, were applied during training. The initial learning rate was set to 0.1, and it was decayed by a factor of 0.1 and 0.01 at the 20th and 50th epochs, respectively. The total number of training epochs was 100. The SGD optimizer was used with a weight decay of 5×10^{-4} and a momentum of 0.9.

For performance evaluation, the experiment adopted standard person re-identification metrics, including the Cumulative Matching Characteristic (CMC) curve, Rank-n accuracy, and mean Average Precision (mAP). The calculation of mAP follows the formula defined as follows.

$$mAP = \frac{1}{N}\sum_{k=1}^{N} AP_k \tag{10}$$

Here, n denotes the total number of query images, and AP_k represents the average precision of the k-th query image.

4.3 Result

The performance of the proposed method is compared with that of current mainstream visible-infrared person re-identification approaches, and the results are presented in Table 1.

Table 1: Comparison results on SYSU-MM01 and RegDB datasets

Model	SYSU-MM01				RegDB			
	All search		Indoor search		V to T		T to V	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
AGW	47.58	47.69	54.29	63.02	70.05	66.37	70.49	65.90
Xmodal	49.92	50.73	-	-	62.21	60.18	-	-
DDAG	53.61	52.02	58.37	65.44	69.34	63.19	64.77	58.54
SPOT	65.34	62.25	69.42	74.63	80.35	72.46	79.37	72.26
PMT	67.53	64.98	71.66	76.52	84.83	76.55	84.16	75.13
DART	68.79	66.55	72.52	78.17	83.78	76.00	81.78	73.64
CAJ	69.88	66.89	76.26	80.37	85.03	79.14	84.75	77.82
AFEN	71.74	68.96	77.86	82.03	86.32	79.60	84.96	78.31

Quantitative results in Table 1 reveal the relative strengths of the different networks: the proposed approach surpasses most existing methods under both the All-Search and Indoor-Search settings on SYSU-MM01. Specifically, under All-Search it improves Rank-1 by 1.86 % and mAP by 2.07 % over CAJ; under Indoor-Search the gains are 1.60 % in Rank-1 and 1.66 % in mAP. Entries marked “–” denote results not reported in the original paper.

5. Conclusions

This paper presents a cross-modal person re-identification framework that integrates attention mechanisms with feature enhancement to extract fine-grained and highly discriminative representations from heterogeneous pedestrian images. The architecture is composed of a median-enhanced spatial–channel attention module (MECS) and a feature enhancement module (FEM), both embedded within a two-stream network. The MECS module selectively amplifies subtle pedestrian details, thereby reinforcing the intra-class consistency and inter-class separability of modality-specific features. Subsequently, the FEM enriches the representation of cluttered backgrounds by jointly preserving high- and low-frequency image components, yielding more comprehensive cross-modal features. The entire model is optimized via a joint supervision of identity loss and an enhanced weighted regularization triplet loss, which enlarges inter-class margins while increasing intra-class similarity across modalities.

Extensive experiments demonstrate the effectiveness of the proposed approach.

References

[1] Zhang Y, Zhao S, Kang Y, et al. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 462-479.

[2] Mukhtar H, Khan M U G. CMOT: A cross-modality transformer for RGB-D fusion in person re-identification with online learning capabilities[J]. Knowledge-Based Systems, 2024, 283: 111155.

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

[4] Wang Z, Wang Z, Zheng Y, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 618-626.

[5] Ye M, Ruan W, Du B, et al. Channel augmented joint learning for visible-infrared recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13567-13576.

[6] Li D, Wei X, Hong X, et al. Infrared-visible cross-modal person re-identification with an x modality[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(04): 4610-4617.

[7] Wei Z, Yang X, Wang N, et al. Syncretic modality collaborative learning for visible infrared person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 225-234.

[8] Ye M, Shen J, J. Crandall D, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020: 229-247.

[9] Wu Q, Dai P, Chen J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4330-4339.

[10] Chan S, Du F, Tang T, et al. Parameter sharing and multi-granularity feature learning for cross-modality person re-identification[J]. Complex & Intelligent Systems, 2024, 10(1): 949-962.

[11] Cheng D, Wang X, Wang N, et al. Cross-modality person re-identification with memory-based contrastive embedding[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(1): 425-432.

[12] Liang T, Jin Y, Liu W, et al. Cross-modality transformer with modality mining for visible-infrared person re-identification[J]. IEEE Transactions on Multimedia, 2023, 25: 8432-8444.

[13] Lu H, Zou X, Zhang P. Learning progressive modality-shared transformers for effective visible-infrared person re-identification[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(2): 1835-1843.

[14] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 248-255.