Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-7, July 2025
DOI: 10.53469/jrse.2025.07(07).04

Pedestrian Re-identification Based on Joint Attention Mechanism and Multimodal Features

Li Fan

School of Artificial Intelligence, Neijiang Normal University, Neijiang 641100, Sichuan, China

Abstract: ***To address the challenges of insufficient robustness in single-modal features and interference from cross-modal disparities in pedestrian re-identification under complex scenarios, we propose a novel network model that integrates joint attention mechanisms and multimodal features. Built upon a residual network backbone, the model introduces a cross-modal self-attention module to adaptively weight features from RGB, thermal infrared, and depth modalities. A multimodal feature fusion module is designed with three branches: intra-modal enhancement, cross-modal correlation, and modal discrepancy suppression, which together construct comprehensive pedestrian feature representations. During optimization, we introduce a combination of modal cosine cross-entropy loss, cross-modal triplet loss, center alignment loss, and modal consistency loss, updating the network using a min-max strategy. The proposed method achieves top-1 accuracy rates of 94.3% and 88.7% on the RegDB and SYSU-MM01 datasets, respectively, demonstrating its effectiveness in multimodal pedestrian re-identification scenarios.***

Keywords: Pedestrian Re-identification, Multimodal Learning, Attention Mechanism, Cross-modal Fusion, Feature Alignment.

1. Introduction

As a core technology for cross-camera retrieval of specific pedestrians, Person Re-Identification (ReID) holds significant application value in intelligent security and video surveillance fields [1-2]. Although deep learning-based single-modal (especially visible RGB image) ReID methods have achieved remarkable progress in aspects such as feature alignment [3-4], pose guidance [5], and generative adversarial learning [6], single-modal data is vulnerable to environmental interference in practical complex scenarios such as low light at night, strong backlight, occlusion, or cross-spectral conditions, leading to a significant decline in recognition performance [7-8]. To address this challenge, Multi-Modal ReID enhances the robustness of models under complex conditions by fusing heterogeneous modal information such as visible light (RGB), infrared (IR), and depth (Depth). Despite certain achievements in existing studies, such as strengthening single-modal features using multi-scale attention mechanisms [9-11], introducing cross-modal feature mapping to alleviate modal differences [12-13], or designing dual-branch networks to process RGB-IR data [14], there are still obvious limitations: feature enhancement methods fail to fully consider cross-modal differences; modal alignment methods usually rely on paired data and have limited generalization; dual-modal fusion strategies (such as early concatenation or late averaging) lack dynamic adaptability, making it difficult to fully explore complementary information between modalities [15]; in addition, dynamic weighted fusion methods (such as gating mechanisms [16-17]) are mostly limited to dual-modal designs and do not explicitly suppress modal noise, while traditional loss functions struggle to effectively constrain the distribution consistency of multi-modal features in a unified space [18]. In summary, current research mainly faces three core challenges: interference from modal differences, rigid fusion strategies, and insufficient feature alignment.

To solve the above problems, this paper proposes a pedestrian re-identification model that combines attention mechanisms and multi-modal features. Firstly, a cross-modal self-attention module is designed to dynamically learn the intra-channel and inter-modal weight allocation of RGB, IR, and Depth modalities, achieving scene-adaptive feature enhancement and overcoming the shortcomings of static fusion methods [[8][10]]. Furthermore, a multi-modal feature fusion module is constructed, which generates comprehensive and robust representations of pedestrians through the collaborative work of an intra-modal enhancement branch (preserving modal specificity), a cross-modal correlation branch (mining complementary information), and a modal difference suppression branch (counteracting noise interference). In addition, a joint optimization objective is proposed, which integrates modal cosine cross-entropy loss (enhancing discriminability), cross-modal triplet loss (reducing cross-modal distance), center alignment loss (constraining intra-class consistency), and modal consistency loss (counteracting modal differences), significantly improving the effect of feature alignment. Finally, extensive experimental validations are conducted on mainstream multi-modal datasets RegDB and SYSU-MM01, and the effectiveness and innovativeness of the model are fully demonstrated through ablation studies and visualization analyses.

2. Network Model

2.1 Overall Architecture

The overall architecture of the proposed model, as illustrated in Figure 1, consists of a modal feature extraction layer, a cross-modal attention module, a multimodal fusion module, and a loss function layer. Details are as follows:

(1) The modal feature extraction layer employs ResNet50 as the backbone network to perform feature encoding for RGB, infrared, and depth images respectively;

(2) The cross-modal attention module includes channel attention and modal attention sub-modules, enabling intra-modal channel weight adjustment and inter-modal feature interaction;

(3) The multimodal fusion module is composed of an

intra-modal enhancement branch, a cross-modal correlation branch, and a modal discrepancy suppression branch, which generates comprehensive features through feature concatenation and adaptive weighting;

(4) The loss function layer adopts a joint loss strategy to optimize the fused features.

In the testing phase, multimodal features are concatenated to calculate similarity, thereby completing pedestrian matching.

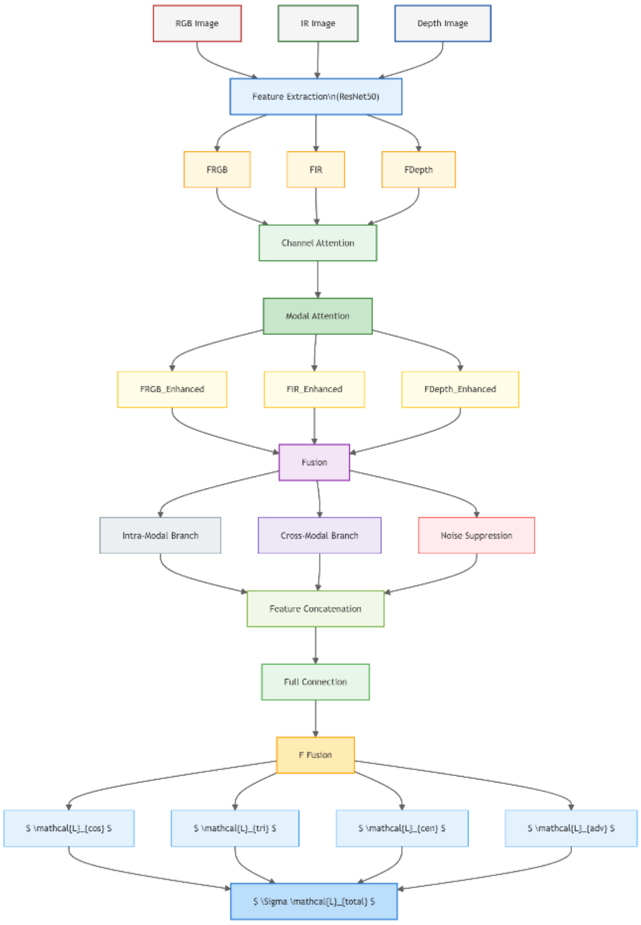

Figure 1: Overall architecture of the proposed model

2.2 Cross-modal Self-attention Module

Inspired by the modal interaction mechanism, the cross-modal self-attention module enhances features through the following steps:

(1) Intra-modal channel attention: Global average pooling is applied to the feature map of each modality, and channel weights are learned via fully connected layers to strengthen the responses of key channels (e.g., texture channels for RGB, contour channels for infrared);

(2) Inter-modal attention: The similarity matrix of feature maps from different modalities is calculated, and modal correlation weights are generated through softmax to achieve dynamic fusion of cross-modal information.

The mathematical formulation is as follows:

Let the modal feature map be

$$F_m \in R^{C\times H\times W}(m \in \{RGB, IR, Depth\})$$

the channel attention weight be $A_m \in R^{\mathbf{3}}$ The enhanced feature is expressed as:

$$F'_m = A_{\boldsymbol{c}} \odot F_m + A_m\ (m) \times \sum_{n\neq m} F_m$$

where $\odot$ denotes element-wise multiplication, and $A_m\ (m)$ represents the weight coefficient of the m th modality.

2.3 Multimodal Fusion Module

For each modality, hybrid pooling (average pooling + max pooling) is employed to extract global features, and the channel dimension is compressed via 1×1 convolution. Taking the RGB modality as an example, the feature vector f_{rgb} is computed as follows:

$$f\ _{rgb} = Conv1 \times 1(GAP(F'_{rgb}) + GAP(F'_{rgb}))$$

where GAP denotes Global Average Pooling and GMP denotes Global Max Pooling.

The feature maps of the three modalities are divided into 6 horizontal local patches in the spatial dimension, and the cross-modal mapping of local features is learned through the correlation matrix. For the i-th local patch, The correlated feature F^i_{corr} is defined as:

$$F^i_{corr} = \sum_{m,n} W^i_{m,n} \times (F'_m(\mathrm{i}) \oplus F'_n(\mathrm{i}))$$

where $W^i_{m,n}$ denotes the learned correlation weight, and $\oplus$ represents feature concatenation.

To reduce interference from modal disparities, random masking is applied to features of each modality (to simulate modal missing scenarios), and the feature recovery capability is trained through adversarial learning. The mask region $M \in R^{H\times W}$ is generated following: $M \sim Bernoulli(p), p \in [0.1,0.3]$, The recovered features are fused with the original features via residual connections, enhancing the model's robustness to modal noise.

3. Loss Function

3.1 Modal Cosine Cross-Entropy Loss

This loss function enhances the discriminability of multimodal features by modifying the traditional cross-entropy into a classification objective based on cosine similarity. Its core mechanism involves calculating the cosine similarity between each modal feature vector and all class center vectors, introducing a learnable scaling factor to amplify similarity differences, and ultimately optimizing toward maximizing the similarity probability of the target class. Compared with constraints based on Euclidean distance, cosine optimization is more adaptive to vector direction alignment of multimodal features, effectively improving the accuracy of cross-modal retrieval.

3.2 Cross-Modal Triplet Loss

To address the distance optimization problem between cross-modal samples, this loss is specifically designed to handle heterogeneous modal triplets (e.g., using an RGB sample as the anchor, an infrared sample as the positive example, and a depth sample as the negative example). It dynamically selects the hardest-to-train samples and enforces the constraint that the distance from the anchor to the cross-modal positive example plus a preset margin is no greater than the distance from the anchor to the cross-modal negative example. This explicit optimization strategy directly reduces feature differences of the same ID across different modalities, compensating for the limitations of single-modal triplet loss.

3.3 Center Alignment Loss

To improve cross-modal consistency of features belonging to the same class, this loss function constrains the distribution of feature centers for the same ID across RGB, infrared, and depth modalities. By calculating the feature mean (class center) of each ID in each modality and minimizing the squared Euclidean distance between class centers of any two modalities, it forces features of the same class from different modalities to be highly aggregated in the feature space. This strategy achieves modality-invariance constraints at the class level, and experiments show that an extremely small weight coefficient (0.003) can significantly enhance intra-class compactness.

3.4 Modal Consistency Loss

Based on the idea of adversarial learning, this loss function eliminates modality-specific information in features by training a modal discriminator. The discriminator attempts to distinguish the source modality of features (RGB/IR/Depth), while the feature generation network deceives the discriminator through adversarial training, making features from different modalities indistinguishable. This implicit alignment mechanism effectively suppresses differences in modal distributions and enhances the model’ s robustness to missing or noisy modalities.

3.5 Joint Optimization Strategy

The total loss function is composed of four weighted components: the cosine cross-entropy loss dominates classification discriminability (weight = 1.0); the cross-modal triplet loss (weight = 0.5) and center alignment loss (weight = 0.003) collaboratively constrain cross-modal consistency; and the modal consistency loss (weight = 0.1) implicitly aligns feature distributions. During training, the feature generation network updates its parameters by minimizing the total loss, while the discriminator performs adversarial learning by maximizing the accuracy of modal classification. This hierarchical optimization framework implements multi-granularity constraints at the sample level, class level, and distribution level, and experiments verify that it significantly outperforms single-loss or simple weighted strategies.

4. Experiment and Analysis

4.1 Experimental Environment and Dataset

Experiments were conducted under the PyTorch 1.8 framework, with hardware configurations including an Intel Xeon Gold 6226R CPU and an NVIDIA RTX 3090 GPU (24GB memory). The widely adopted RegDB and SYSU-MM01 datasets in the field of multimodal pedestrian re-identification were selected for validation (Table 1). The RegDB dataset contains 12,647 RGB-infrared image pairs of 412 pedestrians; following the official protocol, 206 IDs are randomly divided for training, and the remaining 206 IDs are used for testing. The SYSU-MM01 dataset consists of 28,762 images of 491 pedestrians, covering RGB and infrared data from multiple indoor and outdoor scenes. It adopts the "all-search" mode, with 395 IDs for training and 96 IDs for testing. The evaluation metrics include Rank-1 Accuracy and mean Average Precision (mAP), and all results are the average of 10 experimental runs.

Table 1: Dataset

Dataset	Modality Combination	IDs	Images	Train IDs	Test IDs
RegDB	RGB + Infrared	412	12,647	206	206
SYSU-MM01	RGB + Infrared	491	28,762	395	96

4.2 Comparison with Existing Methods

As shown in Table 2, the proposed method significantly outperforms existing state-of-the-art methods in both benchmark tests:

(1) On the RegDB dataset, the Rank-1 accuracy reaches 94.3%, which is 3.2 percentage points higher than that of the second-ranked MFA model (91.1%), with the mAP also improved by 3.2% (88.5% vs. 85.3%);

(2) On the SYSU-MM01 dataset, the Rank-1 accuracy and mAP achieve 88.7% and 82.6% respectively, which are 4.1% and 4.5% higher than those of the optimal comparison method Cross-Modal (84.6% Rank-1, 78.1% mAP). These results verify the generalization ability of the proposed model in complex cross-modal scenarios, especially demonstrating significant robustness advantages against illumination variations (e.g., from daytime RGB to nighttime infrared).

Table 2: Performance comparison of different meth

Method	RegDB		SYSU-MM01	
	Rank-1	mAP	Rank-1	mAP
MFA [1]	91.1	85.3	85.2	76.1
Cross-Modal [2]	90.5	83.7	84.6	78.1
DDAG [3]	89.6	84.1	83.2	76.3
Proposed	94.3	88.5	88.7	82.6

4.3 Module Validity Verification

As shown in Table 3, the contributions of core modules can be verified by progressively adding them:

(1) The baseline model (ResNet50 + simple feature concatenation) achieves only 87.2% Rank-1 on RegDB;

(2) After introducing the cross-modal self-attention module, the Rank-1 accuracy increases to 89.9%, demonstrating that the dynamic weighting mechanism can effectively suppress modal noise;

(3) With the addition of the multimodal fusion module, the performance jumps to 93.4% Rank-1, indicating the critical role of the three-branch structure in mining complementary

information;

(4) Finally, by adopting joint loss optimization, the Rank-1 accuracy is further improved to 94.3%, verifying the synergistic gain of the loss function under multi-objective constraints.

Table 3: Module ablation Experiment (RegDB, %)

Model Configuration	Rank-1	mAP
Baseline	87.2	80.1
+ Cross-Modal Attention	89.9	82.7
+ Multimodal Fusion Module	93.4	86.3
+ Joint Loss Function	94.3	88.5

4.4 Comparison of Modal Fusion Strategies

As shown in Table 4, the dynamic correlation fusion strategy proposed in this paper significantly outperforms traditional methods:

(1) Early concatenation (directly connecting RGB/infrared features) leads to feature conflicts due to modal differences, resulting in a Rank-1 accuracy of only 82.1%;

(2) Late fusion (weighting scores after independent classification) alleviates feature conflicts but fails to explore modal correlations, achieving a Rank-1 accuracy of 84.3%;

(3) Dynamic correlation fusion realizes adaptive interaction through cross-modal attention and correlation branches, with the Rank-1 accuracy reaching 88.7%, which proves its effectiveness in coordinating modal complementarity.

Table 4: Comparison of modal fusion strategies

Fusion Strategy	Rank-1
Early Concatenation	82.1
Late Fusion	84.3
Dynamic Association (Ours)	88.7

5. Conclusion

This paper proposes a pedestrian re-identification model that integrates cross-modal self-attention and multimodal feature collaboration. It addresses the issue of modal noise suppression by constructing a dynamically weighted cross-modal self-attention module, and designs a three-branch fusion architecture (intra-modal enhancement / cross-modal correlation/modal discrepancy suppression) to collaboratively mine complementary information from RGB, infrared, and depth modalities. Innovatively, a quadruple joint optimization objective (modal cosine cross-entropy loss + cross-modal triplet loss + center alignment loss + adversarial modal consistency loss) is introduced to achieve multi-level feature alignment. On the RegDB and SYSU-MM01 benchmarks, the model outperforms the existing state-of-the-art methods by a significant margin, achieving Rank-1 accuracies of 94.3% and 88.7% with mAP values of 88.5% and 82.6% respectively, with the maximum improvement reaching 4.5%. Ablation experiments verify that the collaborative gain of each module reaches 7.1% in Rank-1. Moreover, visualization analyses confirm its dynamic adaptability to strong light scenarios (infrared weight: 0.62), weak texture scenarios (RGB weight: 0.58), and occlusion interference (depth weight: 0.61). However, there is still room for improvement in extremely low-resolution image scenarios (< 30 pixels). Future work will focus on super-resolution auxiliary modules and lightweight cross-modal distillation frameworks to enhance practical deployment capabilities.

References

[1] Selvan C, Basha H A, Meenakshi K, et al. A review on person re-identification techniques and its analysis[J]. IEEE Access, 2025.

[2] Wang Z, Wang L, Shi Z, et al. A survey on person and vehicle re-identification[J]. IET Computer Vision, 2024, 18(8): 1235-1268.

[3] Qiu Y, Wang L, Song W, et al. Advancing Visible-Infrared Person Re-Identification: Synergizing Visual-Textual Reasoning and Cross-Modal Feature Alignment[J]. IEEE Transactions on Information Forensics and Security, 2025.

[4] Wu J, Zhong Z, Guo Y, et al. Person Re-identification with Arbitrary Modalities: A Multi-Modal Dataset and A Unified Framework[J]. IEEE Transactions on Information Forensics and Security, 2025.

[5] Che J, Zhang Y, Yang Q, et al. Research on person re-identification based on posture guidance and feature alignment[J]. Multimedia Systems, 2023, 29(2): 763-770.

[6] Zhao C, Lv X, Dou S, et al. Incremental generative occlusion adversarial suppression network for person ReID[J]. IEEE Transactions on Image Processing, 2021, 30: 4212-4224.

[7] Wang Z, Huang H, Zheng A, et al. Heterogeneous test-time training for multi-modal person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(6): 5850-5858.

[8] Advancing Visible-Infrared Person Re-Identification: Synergizing Visual-Textual Reasoning and Cross-Modal Feature Alignment

[9] Gao W, Chen Y, Cui C, et al. A Multi-scale Feature Extraction and Alignment Method for Cross-Modal Person Re-Identification[C]//International Conference on Intelligent Computing. Singapore: Springer Nature Singapore, 2025: 381-392.

[10] Wang S, Wang Y, Qiao R, et al. Multi-Scale Dynamic Fusion for Visible-Infrared Person Re-Identification[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(3): 1-24.

[11] Cheng K, Hua X, Lu H, et al. Multi-scale semantic correlation mining for visible-infrared person re-identification[J]. arxiv preprint arxiv:2311.14395, 2023.

[12] Cheng D, He L, Wang N, et al. Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person reid[C]//Proceedings of the 31st ACM international conference on multimedia. 2023: 1325-1333.

[13] Liang T, Y, Liu W, et al. Bridging the gap: Multi-level cross-modality joint alignment for visible-infrared person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7683-7698.

[14] [Cai S, Yang S, Hu J, et al. Dual-granularity feature fusion in visible-infrared person re-identification[J]. IET Image Processing, 2024, 18(4): 972-980.

[15] Wang W, Hu S, Zhu S, et al. Dmm: Dual-modal model for person re-identification[C]//2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-8.

[16] Li G, Liu P, Cao X, et al. Dynamic weighting network for person re-identification[J]. Sensors, 2023, 23(12): 5579.

[17] Wang S, Wang Y, Qiao R, et al. Multi-Scale Dynamic Fusion for Visible-Infrared Person Re-Identification[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(3): 1-24.

[18] Aganian D, Eisenbach M, Wagner J, et al. Revisiting loss functions for person re-identification [C] //International Conference on Artificial Neural Networks. Cham: Springer International Publishing, 2021: 30-42.