Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-5, May 2025
DOI: 10.53469/jrse.2025.07(05).09

Visible-Infrared Cross-Modal Pedestrian Re-identification based on Two Attention-calibrated Correlation Features

Fan Li

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

Abstract: ***Visible-Infrared Cross-Modal Pedestrian Re-identification faces the challenges of feature misalignment due to viewpoint differences and pose changes, and insufficient robustness to noisy samples. To this end, this paper proposes a cross-modal pedestrian re-identification based on two attention calibrated associative features (Two Attention Calibrated Associative Feature Networks, TACAFNet). Firstly, a data enhancement strategy based on channel relationship is designed to use to generate diverse samples through uniform sampling of channel perception and simulate occlusion scenarios by combining with a random channel erasure technique to improve the model's generalisation ability to cross-modal colour differences. Secondly, the Image Pair Correlation Attention Module (IPCAM) is designed to construct graph structural relationships by exploiting the contextual relevance of intra-modal pedestrian features, to enhance feature discriminative properties and to suppress background noise interference. Further, Cross-modal Alignment Attention Module (CMAAM) is proposed to reduce modal correlation interference through inter-modal component-level feature matching and enhance cross-modal fine-grained feature alignment to reduce intra-class differences. Experiments on SYSU-MM01 and RegDB datasets show that TACAFNet significantly outperforms existing mainstream methods, validating the effectiveness of the proposed model in the cross-modal pedestrian re-identification task.***

Keywords: Visible-Infrared Cross-Modal Pedestrian Re-Identification, Data Augmentation, Attentional Mechanism, Graph Structure**.**

1. Visible-Infrared Cross-Modal Pedestrian Re-identification

Visible-Infrared Cross-modality Person Re-identification (VI-ReID) is a cross-modality retrieval task aiming at matching pedestrian images between visible and infrared modalities. This technology can break through the limitation of light conditions and achieve all-weather pedestrian identification, which has important application value in intelligent surveillance and other fields. Early VI-ReID research was mainly based on manual feature descriptors [1]. With the development of deep learning, researchers have gradually turned to convolutional neural network-based methods, mainly focusing on the two directions of feature representation learning [2] and metric learning [3]. However, the VI-ReID task faces many challenges: firstly, noisy data such as low-resolution images and occluded samples captured in real scenes can seriously affect the recognition performance; secondly, intra-class differences between modalities (e.g., changes in pedestrian poses, background disturbances, differences in viewpoints, etc.) and inter-modal gaps can significantly reduce the matching accuracy. Therefore, how to learn feature representations with cross-modal robustness becomes a core issue in VI-ReID research. Existing methods can be mainly classified into three categories: the first one extracts global features based on single-stream or two-stream networks [4], but such methods are easily interfered by background information; the second one focuses on local features, such as the MAM and PAM modules proposed by Wu et al [5] and the modal obfuscation method of Hao et al [6], but these methods are limited in reducing modal gaps; and the third one tries to learn cross-modal shared features, such as the Hi-CMD method proposed by Choi et al [7], which separates identity discriminators and modal features through feature decoupling techniques. However, these methods often only learn rough image-level features or rigid local features, resulting in inaccurate cross-modal feature alignment and affecting model performance. Data scarcity is another major challenge in VI-ReID research. There are only two public datasets, SYSU-MM01 and RegDB, with insufficient sample diversity. For this reason, researchers have tried to use image conversion methods: 1) GAN-based methods introduce generative noise; 2) RGB to greyscale methods [8] lose colour information; and 3) single-channel image reconstruction methods [9] are difficult to implement. Most of these methods directly follow the paradigm of unimodal ReID and fail to make full use of cross-modal information. Recently, data enhancement techniques [10] have shown advantages in fine-grained visual tasks, such as the random erasure method proposed by Zhong et al [11], which can effectively enhance the robustness of the model to colour differences.

2. TACAFNet and Dataset Enhancement Methodology

Aiming at the problems of feature misalignment and modal differences in cross-modal pedestrian re-recognition, this paper proposes a cross-modal pedestrian re-recognition based on two attention calibrated associative features (Two Attention Calibrated Associative Feature Networks, TACAFNet). The method firstly designs a data enhancement strategy based on channel relationships, and randomly generates enhancement samples with modal invariance by analysing the correlation between multispectral channels, which effectively reduces the matching difficulty caused by modal differences while preserving the cross-modal correlation. In the feature extraction stage, for the intra-modal feature relationships, this paper proposes the Image Pair Correlation Attention Module (IPCAM), which establishes dynamic correlations between local features and between local and global features by constructing graph structure relationships between multi-scale features, effectively solving the problem caused by the rigid division of coarse/ fine-

grained features in the existing methods. /fine-grained feature rigidity, which effectively solves the problem of insufficient discrimination caused by the existing methods, and significantly improves the feature discrimination ability in the presence of annotation noise and large modal differences. To address the cross-modal feature alignment problem, this paper proposes Cross-modal Alignment Attention Module (CMAAM), which establishes the pixel-level correspondence between RGB and IR images based on probabilistic model, and effectively highlights the discriminative region while suppressing the discriminative region caused by the background interference and masking mechanism by the cross-modal feature reconstruction and the parameter-free attention mask mechanism. Through the cross-modal feature reconstruction and parameter-free attention masking mechanism, the noise effect caused by background interference and occlusion is suppressed, and the feature misalignment problem caused by changes in viewpoint and pose is solved.

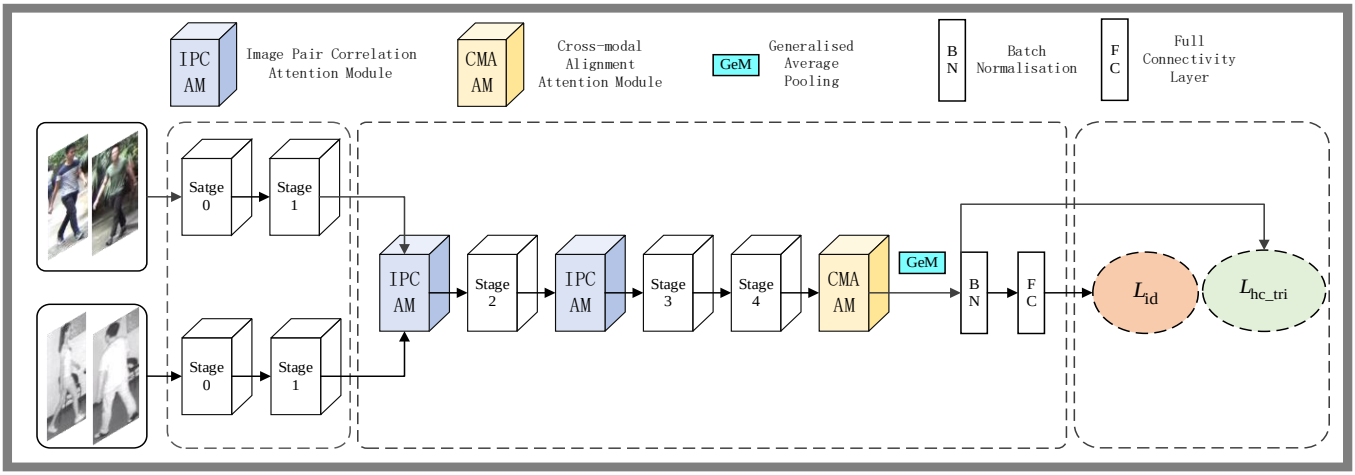

Figure 1: Overall structure of TACAFNet network

One of the important challenges facing current VI-ReID research is the problem of data scarcity, as the existing publicly available datasets contain only two benchmark datasets, SYSU-MM01 [12] and RegDB [13], which are much smaller and less environmentally diverse than the unimodal pedestrian re-identification datasets. This data limitation severely restricts the training and research progress of cross-modal models. Essentially, the core challenge of VI-ReID stems from the significant modal differences between visible (RGB three-channel) and infrared (single-channel) images, which are mainly manifested as significant colour distribution inconsistencies. To address this key issue, this paper analyses the intrinsic correlation between visible three-channel and infrared single-channel, and achieves the reduction of modal differences while preserving the original texture structure. Specifically, the R, G, and B channels of the visible image are randomly exchanged and reorganised to generate three-channel samples with single-channel characteristics, and the images produced by this channel enhancement method not only maintain the rich texture information of the visible image, but also are closer to the infrared image in terms of visual characteristics. The key advantages of this method are (1) effectively narrowing the modal differences through channel-level data enhancement, (2) preserving the original discriminative information, and (3) achieving end-to-end training without modifying the network architecture.

3. Image Pair Correlation Attention Module

In cross-modal pedestrian re-identification tasks, the feature extraction process is often affected by noise interference that affects the effective mining of discriminative features. To address this problem, this paper proposes the IPCAM module, which achieves multi-granularity feature enhancement by constructing graph-structure relationships of pedestrian features. This module breaks through the limitations of the traditional cross-modal approach in which global and local features are split: on the one hand, the initial features extracted by ResNet-50 are divided into horizontal scales, and the low-level features containing rich structural information are retained; on the other hand, the dynamic graph structure modelling mechanism is innovatively adopted to solve the problem of misalignment of key information caused by the traditional horizontal cuts through the multilevel correlation segmentation strategy. Specifically, cyclic superposition feature fusion is implemented for fine-grained features, where neighbouring component features are weighted and aggregated sequentially, and a self-connection mechanism is introduced to avoid structural distortion caused by segmentation errors. In the feature fusion stage, inter-modal alignment relations are constructed by cross-modal feature pair multiplication, combined with the weight squared matrix to maintain channel dimensionality stability, and the ReLu activation function with supervision mechanism is applied to dynamically compress the feature dimensions. This dual-attention calibration mechanism (graph structural relationship modelling and cross-modal feature interaction) not only strengthens the structural relevance of local features, but also achieves the synergistic enhancement of global semantics and local discriminative features, and effectively suppresses the interference of the inter-modal differences and deformation noise on the feature expression.

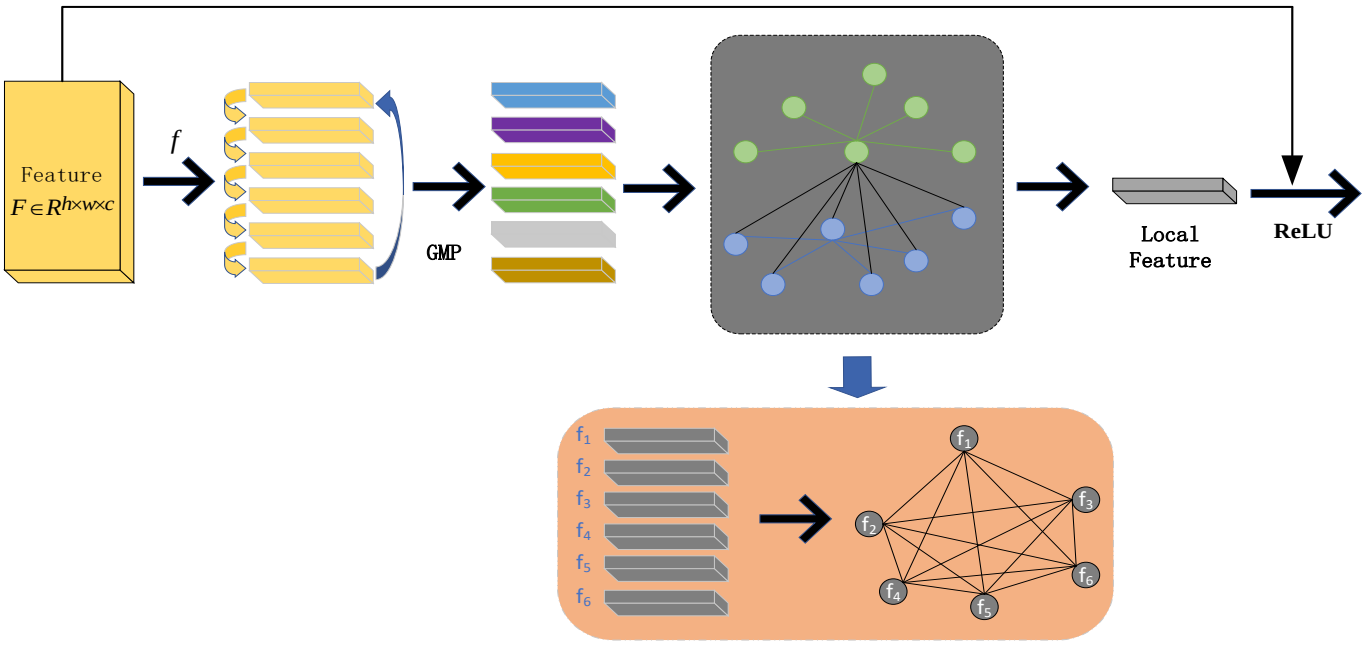

Figure 2: Structure of the IPCAM module

4. Cross-modal Alignment Attention Module

In the field of cross-modal pedestrian re-identification, the feature misalignment problem caused by inter-modal channel differences severely constrains the model performance, especially in the presence of significant structural feature offsets between RGB and infrared modalities. To address the limitations of existing methods that rely on strict image alignment, this paper proposes Cross-modal Alignment Attention Module (CMAAM), which achieves feature space alignment by probabilistically modelling cross-modal dense correspondences. The module breaks through the reliance of traditional methods on explicit alignment and innovatively constructs a feature correspondence ternary function, which jointly exploits cross-modal context complementarity and pedestrian feature self-similarity to establish inter-modal soft correspondences under unsupervised conditions [14]. Specifically, the CMAAM module dynamically approximates the modal feature spacing through a probabilistic propagation mechanism, supports both unidirectional visible to infrared alignment and bi-directional feature mapping, and introduces a parameter-free pedestrian mask constraint reconstruction process that guides the network to focus on the effective pedestrian region. Through pixel-level cross-attention operations in the feature space, the module adaptively suppresses background interference and occlusion noise while eliminating cross-modal channel differences, enabling the network to separate aligned local features with strong discriminative properties from global features. This probabilistic soft alignment-based mechanism not only avoids the error accumulation of explicit geometric alignment, but also significantly enhances the representation consistency of cross-modal features through feature-level co-optimisation, providing a more robust cross-modal feature basis for subsequent identity discrimination.

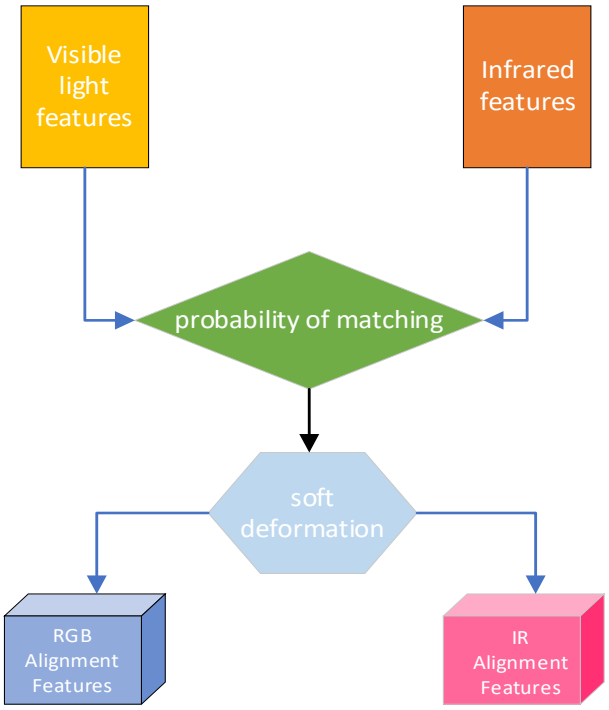

Figure 3: Structure of CMAAM module

5. Analysis of Experimental Results

5.1 Module Ablation Experiment

In this section, the effectiveness of the proposed module is verified through ablation experiments, and the performance is evaluated on the SYSU-MM01 dataset using All Search mode. The experimental setup contains experiments on Image Pair Correlation Attention Module (IPCAM), Cross-modal Alignment Attention Module (CMAAM), and the

experimental results are shown in Table 1. The benchmark experiments show that Rank-1 and mAP are only 50.00% and 52.65% when no module is employed, indicating significant performance limitations of the original model. A breakthrough in system performance is achieved when the two are synergised, with Rank-1 and mAP reaching 77.08% and 80.61%, respectively. It is worth noting that the IPCAM module demonstrates better adaptability in the combined experiments, and its bidirectional feature alignment mechanism effectively mitigates the inter-modal channel bias, providing more significant feature optimisation compared to the unimodal attention mechanism. This experimental sequence verifies the complementary nature of the proposed modules: IPCAM establishes local structural correlations, CMAAM achieves cross-modal spatial alignment, and the three form a synergistic effect through cascading feature optimisation to jointly construct a robust cross-modal representation system.

Table 1: Ablation experiments on the effectiveness of each module

Method		SYSU-MM01	
IPCAM	CMAAM	rank-1	mAP
-	-	50.00	52.65
√	-	54.84	56.77
√	-	69.69	71.61
√	√	77.08	80.61

5.2 Analysis of Model Validity

In this study, we validate the cross-dataset performance of the proposed cross-modal pedestrian re-identification based on two attention calibrated associative features (Two Attention Calibrated Associative Feature Networks, TACAFNet), and carry out systematic experiments on two authoritative datasets SYSU-MM01 and RegDB Evaluation. Under the global search strategy in the SYSU-MM01 dataset, TACAFNet demonstrates significant performance advantages: its mAP index reaches 70.55%, and its Rank-1 matching accuracy reaches 75.98%; in the more challenging indoor restricted scenario, the model performance is further improved, and the Rank-1 index of a single successful match jumps to 79.28%, with a mAP value of 84.62%, verifying the robustness of the method in complex environments. For the RegDB dataset characteristics (including continuously collected pedestrian images), TACAFNet achieves breakthroughs in both visible-infrared and infrared-visible bi-directional tests: in the visible-infrared mode, the model Rank-1 and mAP reach 92.05% and 89.23%, respectively, and in the reverse test, the two metrics are further improved to 94.30% and 90.12%, significantly outperforming FIR. 90.12%, significantly exceeding the baseline level of FMCNet. Comprehensive analysis of the above cross-data set experiments shows that TACAFNet strengthens intra-modal feature correlation through IPCAM and CMAAM, combines with the cross-modal feature alignment strategy, effectively solves the problem of feature space mismatch, and its dual-attention co-optimisation not only achieves the accurate mapping of inter-modal features, but also significantly improves the cross-modal retrieval accuracy through structured feature reconstruction, and provides a better feature representation in the VI-ReID domain. It provides a better feature representation paradigm for VI-ReID.

6. Summary

In this paper, we address the core problems of large modal differences and feature alignment difficulties in the visible-infrared cross-modal pedestrian re-identification task, and propose a cross-modal pedestrian re-identification based on two Attention Calibrated Associative Feature Networks. The method effectively improves the discriminative performance of cross-modal feature representation through three key technological innovations: first, the proposed data enhancement method based on channel relationship significantly reduces modal differences while preserving the original modal features by establishing cross-modal channel correlation. Combined with a random erasure technique to simulate a real occlusion scene, the method maintains a low computational complexity while enhancing data diversity, and has good model compatibility. Secondly, the designed Image Pair Correlation Attention Module breaks through the limitation of traditional rigid feature delineation and achieves adaptive correlation of local features through dynamic graph structure modelling. The module adopts a progressive multi-level feature fusion strategy, which effectively avoids the information loss caused by mechanical division and significantly improves the robustness of the model to noisy samples. Finally, the proposed Cross-modal Alignment Attention Module adopts probabilistic correspondence modelling to achieve accurate alignment of fine-grained features. Compared with existing image generation methods, this module avoids complex generative adversarial network design and achieves better feature alignment while maintaining model simplicity. Extensive experiments on two benchmark datasets, RegDB and SYSU-MM01, show that this paper's method outperforms existing mainstream methods in several evaluation metrics. In particular, the Rank-1 accuracy reaches a significant improvement in the full search mode on the SYSU-MM01 dataset. These experimental results fully validate the effectiveness and sophistication of the TACAFNet model in solving the cross-modal pedestrian re-identification problem. This study provides a new technical idea for cross-modal feature learning, which has important theoretical and application value for promoting the development of all-weather intelligent surveillance system.

References

[1] Liao S, Hu Y, Zhu X, et al. Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197-2206.

[2] Ye M, Lan X, Li J, et al. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification [C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[3] Fan X, Luo H, Zhang C, et al. Cross-Spectrum Dual-Subspace Pairing for RGB-Infrared Cross-Modality Person Re-Identification [J]. arXiv preprint arXiv: 2003.00213, 2020.

[4] Yang M, Huang Z, Hu P, et al. Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14308-14317.

[5] 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.
[6] Hao X, Zhao S, Ye M, et al. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16403-16412.
[7] Choi S, Lee S, Kim Y, et al. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10257-10266.
[8] Duan B, Fu C, Li Y, et al. Cross-Spectral Face Hallucination via Disentangling Independent Factors[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7930-7938.
[9] 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.
[10] Zhong Z, Zheng L, Zheng Z, et al. Camstyle: A Novel Data Augmentation Method for Person Re-Identification [J]. IEEE Transactions on Image Processing, 2018, 28(3): 1176-1190.
[11] Zhong Z, Zheng L, Kang G, et al. Random Erasing Data Augmentation[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 13001-13008.
[12] Wu A, Zheng W S, Yu H X, et al. RGB-Infrared Cross-Modality Person Re-identification [C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 5380-5389.
[13] Nguyen D T, Hong H G, Kim K W, et al. Person Recognition System Based on A Combination of Body Images from Visible Light and Thermal Cameras[J]. Sensors, 2017, 17(3): 605.
[14] Wan L, Sun Z, Jing Q, et al. G2DA: Geometry-Guided Dual-Alignment Learning for RGB-Infrared Person Re-Identification [J]. Pattern Recognition, 2023, 135: 109150.